UNIVERSITÉ D'ÉVRY VAL D'ESSONNE

ÉCOLE DOCTORALE « *DES GÉNOMES AUX ORGANISMES* »

THÈSE DE DOCTORAT EN SCIENCES

Discipline : Biologie Cellulaire et Moléculaire

CARACTÉRISATION DE L'INTRON DE GROUPE II Pl.LSU/2 EN VUE DE SON UTILISATION EN CIBLAGE GÉNOMIQUE

Présentée en vue de l'obtention du grade de Docteur, le 12/12/2012 par

Madeleine ZERBATO

JURY

Dr. Fulvio MAVILIO ..... Rapporteur

Dr. François MICHEL ..... Rapporteur

Dr. Agnès DELAHODDE ..... Examinateur

Dr. Anne GALY ..... Examinateur

Pr. Javier PEREA ..... Directeur de thèse

INSERM UMR 951, Genethon, Université d'Evry Val d'Essonne, Evry, France

UNIVERSITY OF EVRY VAL D'ESSONNE

DOCTORAL SCHOOL « *DES GÉNOMES AUX ORGANISMES* »

DOCTORAL THESIS

Discipline: Molecular and Cellular Biology

CHARACTERIZATION OF THE Pl.LSU/2 GROUP II INTRON FOR ITS USE IN GENOMIC TARGETING

Submitted for the degree of Doctor in Philosophy, on 12/12/2012 by

Madeleine ZERBATO

JURY

Dr. Fulvio MAVILIO ..... Reviewer

Dr. François MICHEL ..... Reviewer

Dr. Agnès DELAHODDE ..... Examiner

Dr. Anne GALY ..... Examiner

Pr. Javier PEREA ..... Thesis advisor

INSERM UMR 951, Genethon, University of Evry Val d'Essonne, Evry, France

“In all science, error precedes the truth, and it is better it should go first than last.”

- Hugh Walpole -

REMERCIEMENTS

(Section in French)

Pour commencer, il se peut que quelques noms m'échappent tant la liste est longue. Veuillez m'en excuser par avance. Mais au cas où vous seriez dans ce cas, ceci est pour vous : merci !

Je tiens à remercier les membres du jury pour l'intérêt que vous avez porté à mon travail. J'exprime ma profonde reconnaissance au Dr. Fulvio Mavilio et au Dr. François Michel d'avoir accepté sans hésitation d'être rapporteurs de mon travail. Merci au Dr. Agnès Delahodde d'avoir accepté d'examiner mon travail. Et merci pour les levures : elles nous auront bien porté chance ! Je tiens à remercier le Dr. Anne Galy pour son implication lors de ce travail, en tant que directrice de l'unité INSERM au sein de laquelle j'ai effectué ma thèse. Je la remercie également pour sa disponibilité, ses conseils, son expertise, et les discussions scientifiques qui ont permis d'apporter un regard avisé sur ce travail. Merci enfin de m'avoir permis d'effectuer une 4e année de doctorat avec un financement du projet Européen PERSIST.

Je tiens à exprimer ma gratitude au Pr. Javier Perea pour m'avoir donné l'opportunité de réaliser ma thèse sous sa responsabilité. Je le remercie pour ses conseils, sa disponibilité, et la confiance qu'il a su m'accorder. L'autonomie qu'il m'a laissée a été un excellent apprentissage pour la suite.

Peu de personnes m'ont profondément influencée et inspirée dans ma vie, mais Nat, tu es de celles-ci. Je n'étais encore qu'une jeune étudiante (anonyme) lorsque tu m'as fait découvrir la recherche lors de mon tout premier stage, en 2005. Et c'est en grande partie grâce à toi que j'en suis là aujourd'hui. Tu m'as toujours soutenue, motivée, rassurée. Je te l'ai déjà dit, mais je réitère : tu as accumulé un paquet de points « paradis » ! Merci, pour tout.

Je tenais à remercier toutes les personnes avec lesquelles j'ai eu la chance de collaborer. Merci à Sabine Charrier de m'avoir fait confiance en me proposant de participer à un des projets qu'elle a mené. Merci à Fedor Svinartchouk pour son implication dans l'identification de la protéine par spectrométrie de masse, et à Jérôme et Jérémy, qui sont à mes yeux comme Batman et Robin : des super-héros ! Une pensée particulière pour tous ceux et celles qui ont eu Pl.LSU/2 entre leurs mains : Sophie, ma coloc' de bureau, qui a su être à l'écoute et avec qui j'ai partagé de si bons moments au labo que je m'y sentais presque comme chez moi, et Damien et Laura, dont j'ai apprécié leur passage dans l'équipe : j'espère vous avoir fait découvrir les joies du clonage ! Et enfin Cécile : merci d'avoir été là, j'ai partagé mes joies, mon enthousiasme, mais aussi parfois mon désarrois lors de ce travail, et tu as toujours su me remotiver. You rock !

J'adresse mes remerciements les plus chaleureux aux membres (passés et présents) de l'équipe IMBI, qui par leur soutien, leur bonne humeur, leur aide, ont contribué à l'aboutissement de ce travail.

Merci à Laurence (et par extension à Hervé) pour les bons moments passés à essayer de prononcer correctement Shia LeBoeuf, Lud (IMBI de cœur), Hind, tous les immuno et les thésard(e)s.

Je ne saurai comment remercier tous les membres de ma dream team à Généthon : Flo, pour avoir largement contribué à mon développement auditif harmonieux lors de nos soirées concerts ; Fanny (ou Al, ou Maurice : ça dépend des jours…euh…nan des cheveux) pour avoir largement contribué à dégrader mes hépatocytes ; Pierre pour me rassurer personnellement sur mon état psychique relatif, en étant plus fou que moi ; Sèv, pour les soirées entre filles qu'on fait chez toi et qui sont toujours très enrichissantes (héhé) ; Julien, parce qu'on partage beaucoup de choses ; Marina, pour être russe (c'est exotique) ; Karim, pour me divertir ; Peggy, pour être un modèle de gentillesse et de décontraction à surtout suivre !

Merci à l'ensemble des personnes qui font qu'à Généthon, la vie d'une thésarde semble « douce ».

En dehors du labo, il y a une vie (tout de même, un peu…). Alors merci à tous mes amis d'avoir été présents et à l'écoute. Je ne peux pas vous citer tous, mais sachez que vous êtes tous concernés ! Merci à Mago, Julia, Julie, Marilyn, Eva : sans vous, je ne serai qu'un 10^{e} de ce que je suis.

A ma famille, qui m'a toujours soutenue dans ce que j'entreprenais.

Merci à mes parents : la liste des choses que je vous dois est longue, à commencer par mon ADN. Je m'abstiendrai de dire que vous êtes les meilleurs, vous vous y habitueriez trop vite :) Alors, simplement : merci d'être vous.

A mon frère, à ma sœur. Je n'aurai pu espérer être mieux entourée qu'avec vous. Hier comme demain, vous êtes et resterez comme mes âmes sœurs. Quelques mots ne suffiront pas, merci pour tout.

ABSTRACT

Integrating vectors are widely used in gene therapy for stable and long-term transgene expression. In *ex vivo* hematopoietic gene therapy approaches, HIV-1-derived lentiviral vectors can thus be used to transduce hematopoietic progenitors. The biological potency of the vector is expected to correlate positively with the frequency of transduced cells and with the number of integration events (VCN, vector copy number) per cell. However, the use of integrating vectors that cannot target transgene integration into host chromosome may lead to insertional mutagenesis. In this regard, the safety of these vectors remains a significant concern in clinical applications. I collaborated on a study evaluating the level of transduction of hematopoietic progenitor cells at the single-cell level by measuring VCN in individual colony-forming cell units using an adapted quantitative PCR method. It was shown that the frequency of transduced progenitor cells and the distribution of VCN in hematopoietic colonies may depend upon experimental conditions including features of vectors.

On the other hand, the use of vectors that can target the integration of the transgene into a specific-site of the host genome would overcome genotoxicity issues. While site-specific integrative approaches based on engineered nucleases such as Zinc-finger nucleases or Meganucleases are currently developed, I evaluated the use of a group II intron for genomic targeting. Group II introns are self-splicing mobile elements found in prokaryotes and eukaryotic organelles. They can integrate into precise genomic locations by homing, following assembly of a ribonucleoprotein complex containing the intron-encoded protein (IEP) and the spliced intron RNA. Engineered group II introns are commonly used tools for targeted genomic modifications in prokaryotes but not in eukaryotes, probably due limited catalytic activation of currently known group II introns in eukaryotic cells. The brown algae *Pylaiella littoralis* Pl.LSU/2 group II intron is uniquely capable of *in vitro* ribozyme activity at unusually low level of magnesium. As this intron remains poorly characterized, I purified recombinant Pl.LSU/2 IEP expressed in *Escherichia coli* and showed that the protein displays a reverse transcriptase activity either alone or associated with intronic RNA. The Pl.LSU/2 intron could be engineered to splice accurately in *Saccharomyces cerevisiae* and splicing efficiency was improved by the maturase activity of the IEP. However, spliced transcripts were not expressed. Although intron splicing was not detected in human cells, and homing of Pl.LSU/2 in *E. coli* and *S. cerevisiae* could not be demonstrated, these data provide the first functional characterization of the PI.LSU/2 IEP and the first evidence that the Pl.LSU/2 group II intron splicing occurs *in vivo* in eukaryotes in an IEP-dependent manner.

KEYWORDS

Gene therapy; integrating vector; group II intron; Pl.LSU/2; Intron-encoded protein; splicing; homing; *Pylaiella littoralis*

TABLE OF CONTENTS

ABBREVIATIONS

(c)PPT	(central) polypurine tract
AAV	Adeno-associated virus
ADA-SCID	Adenosine deaminase-Severe combined immunodeficiency
bp	Base pair
CA	Capsid protein
CALML3	Calmodulin-like 3 protein
CAT	Chloramphenicol acetyltransferase
cDNA	Complementary DNA
CFC	Colony-forming cells
CGD	Chronic granulomatous disease
CIS	Common integration site
DBD	DNA-binding domains
DMD	Duchenne muscular dystrophy
DNA	Deoxyribonucleic acid
DSB	Double-strand break
EBS	Exon-binding site
eGFP	Enhanced green fluorescent protein
GFP	Green fluorescent protein
GST	Glutathione-S-transferase
Gu-HCl	Guanidine hydrochloride
HD-Ad	Helper-dependent Adenoviral vector
HE	Homing endonuclease
His	Histidine
HIV	Human immunodeficiency virus
HR	Homologous recombination
HRP	Horseradish peroxydase
HSC	Hematopoietic stem cells
HSV	Herpes simplex virus
IBS	Intron-binding site
IDLV	Integrase-deficient lentiviral vector
IEP	Intron-encoded protein
IL2RG	Interleukin-2 gamma c common chain receptor
IMAC	Immobilized metal ion affinity chromatography

INDELs	Insertions and/or deletions
IPTG	Isopropyl β-D-1 thiogalactopyranoside
IR	Inverted repeat
kb	Kilobase
kDa	Kilodaltons
LEDGF	Lens epithelium-derived growth factor
LTR	Long-terminal repeat
LV	Lentiviral vector
miRNA	MicroRNA
MOI	Multipicity of infection
MoMLV	Moloney murine leukemia virus
mRNA	Messenger RNA
NHEJ	Non-homologous end joining
NLS	Nuclear localization signal
nt	Nucleotide
ORF	Open-reading frame
PB	Piggy Bac
PBS	Primer binding site / Phosphate buffered saline
PCR	Polymerase chain reaction
PGK	Phosphoglycerate kinase
pI	Isoelectric point
pre-mRNA	Precursor mRNA
Q-PCR	Quantitative PCR
RBS	Ribosome binding site
rHIV	Recombinant HIV
RNA	Ribonucleic acid
RNase	Ribonuclease
RNP	Ribonucleoprotein
RRE	Rev responsible element
rRNA	Ribosomal RNA
RT	Reverse transcriptase
SB	Sleeping Beauty
SCID-X1	X-linked severe combined immunodeficiency
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis

SIN	Self-inactivating
SV40	Simian virus 40
TALE(N)	Transcription activator-like effector (nuclease)
T-ALL	T-cell acute lymphoblastic leukemia
TEV	Tobacco Etch virus
tRNA	Transfer RNA
VCN	Vector copy number
VSV-G	Vesicular stomatitis virus glycoprotein
WAS	Wiskott-Aldrich Sydrome
WASp	Wiskott-Aldrich Sydrome protein
WPRE	Wookchuck hepatis virus posttranscriptional regulatory element
WT	Wild-type
X-ALD	X-linked Adenoleukodystrophy
ZF(N)	Zinc-finger (nuclease)

LIST OF FIGURES AND TABLES

(Sorted in order of appearance)

RÉSUMÉ

(Section in French)

En thérapie génique, le transgène d'intérêt est généralement amené au sein des cellules cibles par l'utilisation de transporteurs, appelés vecteurs. Les différents types de vecteurs existant peuvent être classés selon leur nature (viraux, non-viraux), ou encore en fonction de leur capacité à induire ou non une intégration du transgène dans l'ADN génomique des cellules cibles. Cette intégration présente un avantage notable pour la correction de cellules ayant un pouvoir prolifératif important, comme c'est le cas des cellules souches. En effet, l'insertion du transgène dans ces cellules permet alors d'assurer son expression stable et à long terme, et ce au sein de l'ensemble de la population cellulaire descendante, même si l'expression du transgène peut parfois être diminuée ou éteinte par des modifications épigénétiques (phénomène de « silencing »). Cette relative stabilité de l'expression ne s'observe pas en absence d'intégration, ceci étant dû principalement à la dilution du matériel génétique épisomal au cours des divisions cellulaires.

Les différentes approches intégratives développées pour des applications cliniques diffèrent les unes des autres par le profil d'intégration du transgène au sein du génome. En effet, cette intégration peut s'effectuer soit de façon aléatoire ou semi-aléatoire, soit de façon « orientée », ou encore de façon site-spécifique.

La première catégorie regroupe les vecteurs basés sur les transposons à ADN (profil d'intégration aléatoire vis-à-vis de certaines structures génomiques comme les îlots CpG ou les unités de transcription), les vecteurs rétroviraux (vecteurs γ-rétroviraux et lentiviraux), et les recombinases. Les vecteurs rétroviraux sont historiquement les premiers vecteurs à avoir été utilisés en clinique chez l'homme, et leur utilisation est très largement répandue aujourd'hui. Ils offrent l'avantage de transduire un grand nombre de type cellulaire et d'induire une intégration efficace du transgène dans le génome de la cellule hôte. Néanmoins, les vecteurs γ-rétroviraux présentent un profil d'intégration biaisé vers les sites d'initiation de la transcription et les îlots CpG, tandis que les vecteurs lentiviraux présentent un biais vers les unités transcriptionnelles. De façon générale, l'insertion d'un transgène au sein du génome peut induire une mutagénèse insertionnelle, liée à la dérégulation de l'expression de certains gènes. Dans le cas où une copie du vecteur est insérée au sein d'un oncogène, cette dérégulation peut entrainer une oncogenèse. Les profils d'intégration caractéristiques des vecteurs rétroviraux favorisant l'insertion de l'ADN proviral autour des unités transcriptionnelles et/ou des promoteurs augmente ainsi le risque de mutagénèse. L'insertion aléatoire ou semi-aléatoire d'un vecteur représente donc un risque génotoxique non négligeable qu'il convient d'évaluer dans le cadre de protocoles cliniques.

Suite à l'avènement de ce type d'effet indésirable lors d'études cliniques utilisant des vecteurs dérivés de rétrovirus, des stratégies alternatives ont été développées afin de surmonter ce problème. Certaines approches, basées sur l'utilisation de domaines spécifiques de liaisons à l'ADN, sont développées pour permettre une insertion « orientée » du transgène au sein du génome hôte. Ces domaines de liaison à l'ADN, reconnaissant une séquence spécifique du génome, peuvent être fusionnés directement à la protéine impliquée dans le mécanisme d'intégration du vecteur (intégrase pour les vecteurs lentiviraux ; transposase pour les transposons), ou à une protéine partenaire interagissant avec un des composants du complexe d'intégration, ceci afin de diriger la machinerie d'intégration vers une région choisie du génome. Ces stratégies permettent en théorie de favoriser l'insertion du vecteur aux alentours de la région ciblée, mais n'empêchent toutefois pas une intégration hors de la région d'intérêt.

L'élaboration d'approches permettant une intégration du transgène au niveau d'un site précis et unique du génome est actuellement un champ d'action largement étudié en thérapie génique. En effet, ce type de stratégie permettrait, outre de s'affranchir du risque de mutagénèse insertionnelle, d'effectuer de la réparation génique, ouvrant ainsi la voie à des traitements de maladies génétiques à transmission dominante. Actuellement, ces approches sont basées sur l'utilisation de nucléases site-spécifiques, comme les zinc-finger nucléases (ZFN), les méganucléases, ou encore plus récemment les TALEN (transcription activator-like effector nuclease). Ces nucléases sont capables d'induire une coupure double-brin au niveau d'un site spécifique de l'ADN. Lorsqu'une molécule d'ADN donneur comportant des régions homologues au site d'intégration est présente au sein du noyau des cellules cibles, la cassure double-brin peut être réparée par recombinaison homologue. Si cet ADN donneur contient le transgène d'intérêt, il sera alors intégré précisément au niveau de la coupure double-brin. Ces stratégies, bien qu'étant prometteuses, présentent certains risques qu'il convient d'évaluer. En effet, la coupure double-brin du génome des cellules hôtes peut potentiellement s'effectuer ailleurs qu'au niveau du site choisi si la spécificité de la nucléase utilisée n'est pas assez grande. De plus, la réparation de la coupure double-brin peut être effectuée par un mécanisme alternatif à la recombinaison homologue, le NHEJ (non-homologous end joining), qui est souvent source d'erreurs. Il convient également de noter que l'efficacité de ces approches reste à l'heure actuelle trop faible pour être généralisée dans les approches de thérapie génique.

Enfin, des approches mixtes ont été mises en place et consistent à combiner l'activité de transfert d'acides nucléiques au sein des cellules d'un vecteur à la machinerie d'intégration d'un autre vecteur. Ces vecteurs « hybrides » sont particulièrement intéressants lorsque le vecteur possédant l'activité de transfert d'acides nucléiques ne permet pas une intégration du matériel génétique au sein du génome hôte (comme c'est le cas pour les vecteurs dérivés des Adénovirus, ou des virus AAV associés à l'adénovirus), ou bien lorsque l'efficacité d'acheminement de la machinerie d'intégration au noyau cellulaire est faible (comme c'est le cas des vecteurs non-viraux, ou plasmides, utilisés pour exprimer

les transposons/transposases, ou nucléases site-spécifique). Dans ce dernier cas, si l'efficacité globale d'intégration du transgène dans le génome est bien améliorée, les questions de sécurité liées au système intégratif utilisé demeurent.

L'ensemble de ces données montre l'intérêt d'évaluer la sécurité des vecteurs intégratifs actuellement utilisés en clinique, mais aussi de développer des stratégies alternatives permettant un ciblage de l'insertion du transgène au niveau d'un site précis du génome des cellules hôtes.

Au cours de ce travail de thèse, effectué sous la direction du Pr. Javier Perea au sein du laboratoire « Immunologie moléculaire et Biothérapies » à Généthon dirigé par le Dr. Anne Galy, j'ai principalement évalué la possibilité d'utiliser l'intron de groupe II Pl.LSU/2 de l'algue brune *Pylaiella littoralis* comme vecteur site-spécifique en ciblage génomique. Pour cela, la majeure partie de mon travail a porté sur la caractérisation des activités catalytiques de cet intron *in vivo*, ainsi que l'étude des activités biochimiques de la protéine qu'il code.

J'ai également eu l'opportunité de collaborer avec l'équipe d'Anne Galy sur une étude visant à évaluer l'efficacité et la sureté de vecteurs lentiviraux en déterminant le niveau de transduction de progéniteurs hématopoïétiques par une mesure du nombre de copies de vecteur au sein de clones cellulaires dérivés de progéniteurs isolés (CFC, colony-forming cells). La correction génique de progéniteurs hématopoïétiques est une stratégie thérapeutique efficace pour le traitement de plusieurs types maladies génétiques, comme le syndrome de Wiskott-Aldrich (WAS). WAS est un déficit immunitaire héréditaire rare de transmission récessive liée au chromosome X, dû à des mutations dans le gène codant la protéine WASp. Un essai clinique initié en Allemagne et basé sur la transduction de progéniteurs hématopoïétiques de patients par des vecteurs rétroviraux comportant le transgène codant WASp, suivie de la réimplantation de ces cellules corrigées au sein du patient, a montré l'efficacité thérapeutique de cette stratégie. Un essai clinique international et multi-centrique de phase I/II, auquel l'équipe d'Anne Galy est associée, est actuellement en cours. Cet essai est basé sur l'utilisation d'un vecteur lentiviral auto-inactivé dérivé du VIH-1 (virus de l'immunodéficience humaine de type 1). En fonction des conditions expérimentales utilisées, les vecteurs lentiviraux peuvent transduire un nombre variable de cellules et intégrer un nombre variable de copies d'ADN proviral au sein du génome des cellules hôtes. Il est couramment accepté que l'efficacité d'un vecteur est corrélée positivement au pourcentage de cellules transduites ainsi qu'au nombre de copies de vecteur intégrées. Néanmoins, il a également est montré que le nombre de copies de vecteur intégrées est corrélé à la génotoxicité. Il apparait donc nécessaire de contrôler la distribution des copies de vecteur au sein des cellules transduites de façon à pouvoir évaluer l'efficacité et la sécurité des protocoles expérimentaux utilisés, et ainsi définir la fenêtre thérapeutique du vecteur. S. Charrier et collaborateurs ont donc développé et validé une méthode simple de quantification du nombre de copies de vecteurs intégrées au niveau des CFC basée sur la PCR quantitative. Des clones de lignées cellulaires transduits par le vecteur lentiviral

ont été générés et utilisés comme contrôles pour démontrer la faisabilité, la sensibilité, et la reproductibilité de cette méthode de quantification. Mon implication dans ce travail s'est effectuée lors de cette étape. Cette étude, publiée en 2011 dans le journal Gene Therapy, a montré que la fréquence de progéniteurs hématopoïétiques transduits et la distribution du nombre de copies de vecteur intégrées dans les CFC dépendent des conditions expérimentales utilisées, telles que la dose de vecteur, le nombre de tour de transduction, ou bien la méthode de purification du vecteur. L'ensemble des résultats obtenus démontrent l'importance d'une telle évaluation afin d'optimiser les protocoles de transduction pré-cliniques et cliniques utilisant des vecteurs lentiviraux de façon à définir pour chaque protocole les conditions expérimentales permettant d'assurer une efficacité de transduction suffisante tout en maximisant la sécurité du protocole.

Si l'évaluation des risques liés à l'insertion non ciblée de vecteurs intégratifs est une étape nécessaire à l'élaboration de protocoles cliniques, le ciblage de l'intégration à un site spécifique du génome pourrait représenter une solution majeure au problème de mutagénèse insertionnelle. Cependant, certains obstacles liés aux systèmes actuellement développés (efficacité et spécificité) restent encore à surmonter avant d'élargir leur utilisation en clinique. C'est pourquoi il existe un intérêt à évaluer et développer de nouveaux systèmes permettant un ciblage de l'intégration au niveau du génome. La majeure partie de ce travail de thèse a donc consisté à évaluer la possibilité d'utiliser un intron de groupe II en ciblage génomique.

Les introns de groupe II sont des éléments mobiles naturels présents dans les génomes procaryotes ou d'organelles d'eucaryotes. Une fois transcrit, ils ont la caractéristique de pouvoir s'auto-épisser *in vitro* dans des conditions ioniques spécifiques sans l'intervention d'aucune protéine. Cette fonction catalytique d'auto-épissage est directement liée à la molécule d'ARN elle-même ; ainsi ces ARN catalytiques ont été nommés « ribozymes ». Cet épissage s'effectue classiquement *via* deux étapes de transesterification et est dépendant du repliement de l'ARN précurseur en une structure catalytiquement active. Dans certains cas, les introns de groupe II possèdent un cadre ouvert de lecture codant une protéine, appelée IEP (Intron-encoded protein), qui participe au repliement de l'intron ARN *in vivo* et favorise ainsi l'épissage de l'intron. Cette protéine présente d'autres activités biochimiques impliquées dans le mécanisme de mobilité des introns. Les introns de groupe II sont en effet capables de se propager au sein des génomes au niveau d'un site précis par un mécanisme nommé « homing », détaillé ci-dessous. Le site naturel d'insertion des introns correspond, dans un génome sans intron, à la jonction des deux exons desquels il s'est excisé. La reconnaissance du site d'insertion s'effectue principalement par appariement de séquences entre certaines régions de l'intron et le site cible. Cette caractéristique a permis de développer des introns « re-ciblés » pouvant s'intégrer à un site spécifique choisi dans le génome de procaryotes simplement en modifiant certaines séquences de l'intron impliquées dans la reconnaissance du site d'intégration. Il existe d'ailleurs un système de knock-out commercialisé (TargeTron gene knock-out system, Sigma Aldrich) développé

pour *E. coli* et basé sur l'utilisation de l'intron de groupe II Ll.LtrB de *Lactococcus lactis*, pouvant être « re-ciblé ».

Les mécanismes d'épissage et de homing des introns de groupe II reposent en grande partie sur la structure de l'ARN intronique. La structure secondaire des introns de groupe II est hautement conservée et consiste en six domaines (DI à DVI) arrangés autour d'une boucle centrale. Chaque domaine possède des rôles spécifiques impliqués dans le repliement, les réarrangements conformationels, et/ou l'activité catalytique des introns. Il existe plusieurs intéractions entre les différents domaines de l'intron, permettant la formation d'une structure tertiaire conservée. Cette structure est à l'origine de la formation du site actif, essentiel à l'activité catalytique de l'intron. Les activités catalytiques d'épissage et « d'épissage inverse » de l'intron impliquent, en plus des intéractions tertiaires existants au sein de la molécule, un appariement de séquence entre l'intron et les exons flanquants. Notons que ces intéractions sont identiques lors de l'épissage de l'intron, de l'épissage inverse de l'intron au sein de l'ARN messager mature, ou encore de l'« épissage inverse » de l'intron au sein d'une cible ADN.

Lorsque l'intron a adopté sa structure catalytique et que les appariements avec les exons flanquants se sont effectués, l'intron peut alors s'épisser. Cet épissage requiert la liaison d'ions Mg^{2+} au niveau du site actif de l'intron, et s'effectue généralement *via* deux étapes de transesterification, conduisant à la formation d'un intron ARN épissé en lariat (via une liaison de l'extrémité 5' de l'intron au 2'-OH du A de branchement). L'épissage peut également conduire à la formation d'un ARN intronique linéaire si le nucléophile impliqué lors de la première étape n'est pas le A de branchement, mais est externe.

Une grande partie des introns de groupe II code une protéine multifonctionnelle, l'IEP. *In vivo*, le repliement d'un intron de groupe II est favorisé par l'activité maturase de son IEP. En général, cette protéine possède quatre domaines conservés : le domaine X, responsable de l'activité maturase, le domaine RT, responsable de l'activité de transcriptase inverse (reverse transcription), le domaine D de liaison à l'ADN, et le domaine En, responsable de l'activité endonucléase. Ces trois dernières activités biochimiques sont impliquées dans le mécanisme de homing de l'intron. Ce mécanisme, comprenant différentes étapes, est basé sur la formation d'une particule ribonucléoprotéique (RNP) formée de l'IEP et de l'intron épissé en lariat, obtenu suite à l'épissage de l'intron *in vivo* favorisé par l'IEP. C'est ce complexe qui est impliqué dans la reconnaissance du site cible d'intégration : l'IEP peut en effet effectuer des intéractions avec 2 à 6 nucléotides du site cible et induire un déroulement local de la structure en double hélice, permettant ainsi à l'intron de s'apparier avec le site cible sur une longueur de 13 à 15 nucléotides. L'intron effectue alors un épissage inverse sur le brin sens à la jonction des deux exons. L'IEP, *via* son activité endonucléase, clive alors le brin anti-sens 9 ou 10 nucléotides en aval de cette jonction, générant ainsi une extrémité 3'-OH utilisée par l'IEP pour synthétiser un ADN complémentaire à l'intron *via* son activité de transcriptase inverse. Une copie double brin d'ADN

complémentaire correspondant à l'intron est enfin synthétisée et intégrée par la machinerie de réparation cellulaire.

Certains introns de groupe II ont été utilisés en ciblage génomique chez les procaryotes. En effet, il a été montré que des modifications dans les séquences introniques responsable de l'appariement aux exons inhibent l'épissage de l'intron, et que des mutations complémentaires des exons restaurent l'épissage. De même, le site d'intégration de certains introns peut être modifié en mutant ces mêmes séquences introniques impliquées dans la reconnaissance du site cible. L'intron de *Lactococcus lactis* Ll.LtrB a été largement étudié en ce sens. Le site d'intégration est choisi en fonction des nucléotides reconnus par l'IEP, qui constituent les seules positions fixes. Chaque gène du génome d'*E. coli* contient plusieurs sites potentiels d'intégration, étant donné le faible nombre de ces positions fixes. Les séquences de l'intron impliquées dans la reconnaissance du site d'intégration sont alors modifiées pour s'apparier au site choisi. L'intron est en général délété de la séquence codant l'IEP, qui est exprimée en *trans*, et un transgène peut alors y être cloné. Ces introns « re-ciblés » peuvent induire des modifications génomiques ciblées de façon très efficace dans les cellules procaryotes et induire l'insertion de transgène de façon site-spécifique. Néanmoins, les tentatives d'utilisation des introns de groupe II dans les cellules eucaryotes se sont révélées infructueuses. Il semblerait qu'au moins un des obstacles empêchant l'activité catalytique de l'intron (splicing et/ou homing) dans les cellules eucaryotes soit lié à un environnement cellulaire ionique défavorable. En effet, l'efficacité de homing de l'intron Ll.LtrB dans des oocytes de *Xénopus Laevis*, ou des embryons de *Drosophilia Melanogaster* ou de Zebrafish (*Danio renio*) peut être nettement améliorée en injectant les RNPs en présence de $MgCl_2$.

Dans ce contexte, nous avons choisi d'étudier l'intron de *Pylaiella littoralis* Pl.LSU/2. En effet, il a été montré que cet intron est capable de s'auto-épisser *in vitro* à des concentrations particulièrement faibles de Mg^{2+}. A ce jour, cet intron est le seul possédant cette caractéristique. Cette faible dépendance de l'intron Pl.LSU/2 vis-à-vis du Mg^{2+} pourrait faire de lui un bon candidat pour du ciblage génomique dans les cellules eucaryotes. Cet intron possède un cadre ouvert de lecture codant théoriquement une protéine présentant tous les domaines conservés des IEP. Néanmoins, ni la capacité d'épissage *in vivo* ou de homing de cet intron, ni les activités biochimiques de la protéine qu'il code n'avaient été étudiées lors de l'initiation de ce projet de thèse.

J'ai donc dans un premier temps caractérisé les activités biochimiques de l'IEP codée par l'intron Pl.LSU/2. Cette étude a nécessité plusieurs étapes d'optimisations afin de purifier la protéine en vue de sa caractérisation biochimique. L'IEP a été exprimée en utilisant différents systèmes (expression dans *E. coli*, expression *in vitro*, ou expression dans les cellules d'insectes Sf9 par baculovirus), et purifiée. Etant donné l'absence d'anticorps disponibles dirigés contre l'IEP, nous avons choisi d'exprimer l'IEP fusionnée à différentes étiquettes (Histidine ou Glutathione-S-transférase) en vue de

sa purification par chromatographie d'affinité. Nous avons tout d'abord optimisé l'expression chez *E. coli* de l'IEP fusionnée en N-terminal à la GST (GST-IEP) sous forme soluble, puis nous avons purifié cette protéine en utilisant une résine chargée en glutathion. La GST-IEP a ainsi pu être partiellement purifiée. Des protéines mutantes, la GST-IEPmtDD-, dont le motif catalytique YADD a été changé en YAAA, et la GST-IEPΔRT5, qui porte une déletion d'une partie du domaine conservé RT5, censées être défectives pour l'activité de transcriptase inverse (RT), ont également été exprimées et purifiées. Enfin, en contrôle, nous avons exprimé et purifié la GST seule, dans les mêmes conditions. L'activité RT de ces protéines a par la suite été testée en utilisant les fractions de protéines partiellement purifiées. Cette expérience a montré une activité RT pour la GST-IEP, mais également pour les mutants. La fraction de protéines GST partiellement purifiée n'a, elle, pas montré d'activité RT. Nous avons donc émis l'hypothèse qu'une protéine présentant une activité RT contaminait les fractions GST-IEP et mutants, empêchant ainsi d'évaluer l'activité biochimique de l'IEP.

Nous avons donc testé d'autres systèmes d'expression, en espérant s'affranchir de cette contamination. L'IEP, fusionnée en N-terminal à une étiquette de six résidus histidine et à un épitope V5 (HisV5-IEP) a été exprimée *in vitro* (cell-free) en utilisant un kit commercial. Dans les conditions testées, l'HisV5-IEP n'a pas pu être détectée par électrophorèse et coloration au bleu de Coomassie, indiquant que la protéine n'est pas exprimée ou trop faiblement pour pouvoir étudier pas la suite ses activités biochimiques. Ce système a donc été abandonné, car les éventuelles optimisations sont assez limitées.

Le système d'expression dans les cellules d'insectes Sf9 par baculovirus a ensuite été évalué. Pour cela, la séquence codant l'IEP fusionnée en N-terminal à une étiquette de six résidus histidine (His-IEP) a été insérée au sein du génome du baculovirus, puis un stock de baculovirus recombinant a été produit. Les cellules Sf9 infectées par le baculovirus codant l'His-IEP expriment la protéine lors du cycle viral du baculovirus. Cette expérience a montré que l'His-IEP était majoritairement exprimée sous forme insoluble dans les cellules Sf9, limitant ainsi sa purification. Les optimisations étant également assez limitées avec ce système, nous avons choisi de réévaluer l'expression et la purification de l'IEP chez *E. coli*.

L'étiquette fusionnée à l'IEP a donc été changée afin d'éviter la contamination mise en évidence lors de la purification de la GST-IEP. Nous avons donc exprimé l'HisV5-IEP chez *E. coli* et purifié la protéine par chromatographie d'affinité en utilisant une résine chargée en Ni^{2+}. Les mutants HisV5-IEPmtDD- et HisV5-IEPΔRT5 ont également été exprimé et purifié. Nous avons testé plusieurs conditions de purifications : dénaturantes, non-dénaturantes en présence du détergent zwitterionique CHAPS, et natives. Toutes ces purifications ont conduit à l'obtention de fractions de protéines partiellement purifiées, mais dont la pureté était tout de même satisfaisante. Néanmoins, aucune de ces fractions n'a montré d'activité RT.

Nous avons alors émis l'hypothèse que l'activité RT de l'IEP pouvait être instable en absence d'intron ARN, comme c'est le cas pour l'IEP de l'intron Ll.LtrB. Nous avons donc co-exprimé l'HisV5-IEP et l'intron Pl.LSU/2 chez *E. coli* et purifié l'HisV5-IEP théoriquement complexée à l'intron ARN en RNP, ceci par deux méthodes : chromatographie d'affinité en utilisant une résine chargée en Ni^{2+} ou ultracentrifugation sur une solution de sucrose. Le mutant HisV5-IEPmtDD- a également été co-exprimé avec l'intron ARN et purifié dans les mêmes conditions. Ces expériences ont permis de montrer que l'HisV5-IEP contenue dans les RNP présente une activité RT uniquement lorsque les RNP sont purifiées par ultracentrifugation en sucrose. Cette activité est dépendante de la dose de RNP utilisée et du temps d'incubation de la réaction. Le mutant HisV5-IEPmtDD- ne présente, lui, aucune activité RT, comme attendu. Ces résultats indiquent que les conditions de purification de l'IEP par chromatographie d'affinité sont délétères pour l'activité biochimique de la protéine. Par la suite, nous avons testé l'activité RT de l'HisV5-IEP exprimée seule chez *E. coli* et purifiée par ultracentrifugation en sucrose. Les résultats obtenus montrent une activité RT de la protéine significativement différente de celle obtenue avec la protéine mutante HisV5-IEPmtDD-. L'HisV5-IEP présente donc une activité RT soit seule, soit complexée en RNP à l'ARN intronique.

Dans un second temps, nous avons voulu évaluer la capacité de l'intron Pl.LSU/2 à s'épisser *in vivo* dans une cellule eucaryote. Pour cela, nous avons développé une stratégie pouvant permettre la mise en évidence directe de l'épissage de l'intron chez *Saccharomyces cerevisiae*. L'intron Pl.LSU/2, flanqué d'une partie de ses exons naturels, a été cloné dans un plasmide d'expression de levure en amont du gène URA3 codant l'orotidine 5-phosphate decarboxylase (Ura3p) et en aval d'un épitope HA. L'expression de l'ensemble de ces séquences est placée sous le contrôle d'un même promoteur. L'épissage de l'intron Pl.LSU/2 permettrait d'établir un cadre ouvert de lecture comprenant l'épitope HA, les deux exons (E2 et E3), et le gène URA3, pouvant donc conduire à l'expression d'une protéine de fusion HA-E2-E3-URA3. La stratégie est donc la suivante : le plasmide contenant l'intron Pl.LSU/2 est transformé dans une souche mutante de levure dont le gène URA3 est défectif, et les levures sont ensuite étalées sur un milieu minimum contenant ou non de l'uracile. Les clones poussant sur le milieu sans uracile témoigneraient alors d'une expression de la protéine de fusion HA-E2-E3-URA3 et donc d'un épissage de l'intron Pl.LSU/2. Afin d'effectuer cette expérience, nous avons construit un plasmide contrôle contenant les séquences HA, E2, E3, et URA3 en phase, afin de s'assurer que l'expression de cette protéine de fusion permette bien une croissance des levures sur un milieu sans uracile. De plus, afin d'évaluer l'activité maturase de l'IEP, nous avons également construit un plasmide permettant l'expression conditionnelle de l'IEP fusionnée en C-terminal à un épitope c-Myc et possédant des signaux de localisation nucléaire (NLS). La séquence de l'IEP utilisée ici est codon-optimisée pour la traduction dans les cellules humaines. L'expression de l'IEP, placée sous le contrôle du promoteur Gal10, est réprimée lorsque les levures se trouvent en présence de glucose et induite en présence de galactose. Lors de cette expérience, aucun clone de levure n'a été

obtenu sur milieu sans uracile, même chez les levures exprimant l'IEP. Les levures contenant le plasmide contrôle peuvent, elles, pousser sur le milieu sans uracile. A la suite de ces résultats, nous avons voulu analyser les ARN exprimés chez la levure par RT-PCR et RT-PCR quantitative. Ces résultats ont montré que l'intron Pl.LSU/2 s'épisse bien chez la levure et que l'expression de l'IEP permet d'augmenter l'efficacité de l'épissage de l'intron. L'analyse des protéines exprimées par western blot a montré que la protéine de fusion HA-E2-E3-URA3 n'est pas détectée chez les levures contenant le plasmide avec intron. Ce dernier résultat explique donc l'absence de clones sur milieu sans uracile. Deux hypothèses peuvent être émises à la suite de cette étude : 1) l'épissage de l'intron Pl.LSU/2 n'étant pas assez efficace, la quantité d'ARN messager mature ne permet pas une expression suffisante de protéines HA-E2-E3-URA3 pour être détectée par western blot ; ou 2) il existe un blocage de la traduction de l'ARN messager mature dans les cellules empêchant l'expression de la protéine HA-E2-E3-URA3. Cette dernière hypothèse a été formulée par une autre équipe ayant obtenu des résultats comparables lors d'une étude similaire réalisée en utilisant l'intron Ll.LtrB.

Aux vues de l'ensemble des résultats obtenus chez la levure, nous avons voulu évaluer la capacité d'épissage de l'intron Pl.LSU/2 dans les cellules humaines. Cette étude a été réalisée par une collaboratrice de l'équipe. Pour assurer une expression suffisante de l'intron et de l'IEP dans les cellules, nous avons décidé d'utiliser des vecteurs lentiviraux dérivés du VIH-1 afin, d'une part d'établir des lignées stables exprimant l'intron, et d'autre part d'exprimer l'IEP dans ces lignées stables. Lors de ces expériences, différentes formes de l'intron ont été utilisées : le domaine IV, contenant naturellement la séquence codant l'IEP, a été plus ou moins délété. En effet, il avait été montré pour l'intron Ll.LtrB qu'une structure du domaine IV correspondait à une région de liaison à l'IEP de Ll.LtrB. L'utilisation de différentes formes d'intron Pl.LSU/2 présentant un domaine IV plus ou moins délété pourrait donc permettre de déterminer la ou les régions impliquées dans la liaison de l'IEP à l'intron, pré-requise à l'activité maturase de l'IEP. Nous avons tout d'abord vérifié la transcription des différentes formes d'intron dans les lignées stables de cellules humaines HCT 116, ainsi que l'expression de l'IEP, fusionnée en C-terminal à un épitope c-Myc, possédant des signaux de localisation nucléaire, et codon-optimisée pour la traduction dans les cellules humaines. L'analyse par RT-PCR et RT-PCR quantitative des ARN exprimés dans les lignées stables contenant les différentes formes d'intron et exprimant ou non l'IEP n'a pas permis de mettre en évidence un épissage de l'intron Pl.LSU/2. Il semblerait donc qu'au contraire des résultats obtenus chez la levure, l'intron Pl.LSU/2 ne s'épisse pas ou très peu efficacement dans les cellules humaines.

L'ensemble de ces résultats portant sur la caractérisation des activités biochimiques de l'IEP de Pl.LSU/2 et de la capacité d'épissage de l'intron dans la levure et dans les cellules humaines a été décrit dans un article soumis au journal PLoS ONE, dont la publication a été acceptée sous réserve de modifications.

Enfin, une étude sur la capacité de homing de l'intron Pl.LSU/2 chez *E. coli* et chez la levure a été menée. Nous avons tout d'abord évalué le homing de l'intron Pl.LSU/2 chez *E. coli* en utilisant une stratégie adaptée d'une méthode utilisée pour mettre en évidence le homing de Ll.LtrB chez *E. coli*. L'intron Pl.LSU/2, flanqué d'une partie de ses deux exons (E2 et E3) est cloné dans un plasmide d'expression procaryote (plasmide donneur) en aval de la séquence codant l'HisV5-IEP. Une partie de la séquence codant l'IEP (au niveau du domaine IV de l'intron) est délétée et remplacée par le gène de résistance à la kanamycine (Kan^R) positionné en sens inverse par rapport à l'intron. Ce gène Kan^R est interrompu par l'intron de groupe I *td*, positionné dans le même sens que l'intron Pl.LSU/2 et pouvant s'auto-épisser très efficacement chez *E. coli*. L'expression de l'ensemble de ces séquences s'effectue à partir d'un même promoteur et de façon conditionnelle. Un second plasmide (plasmide accepteur), possédant le gène de résistance au chloramphénicol, est également utilisé et contient les exons E2 et E3 juxtaposés : cette séquence correspond au site d'intégration naturel de l'intron Pl.LSU/2. Lors de la transcription effectuée à partir du plasmide donneur, l'intron *td* devrait s'épisser, permettant ainsi une restauration du gène Kan^R. L'intégration de l'intron Pl.LSU/2 portant le gène Kan^R restauré permettrait alors l'expression de ce gène et l'apparition de clones bactériens sur un milieu contenant de la kanamycine. La stratégie est donc la suivante : les deux plasmides sont transformés chez *E. coli* puis l'expression de l'intron Pl.LSU/2 et de l'HisV5-IEP est induite. Les bactéries sont ensuite étalées sur un milieu nutritif contenant soit du chloramphénicol seul, soit du chloramphénicol et de la kanamycine (sélection des clones portant au moins le plasmide accepteur dans lequel l'intron s'est intégré). Malheureusement, durant les expériences effectuées, aucun clone n'a été obtenu sur le milieu contenant de la kanamycine. Afin d'étudier de façon plus approfondie l'état des plasmides contenus dans les bactéries, une analyse par restriction enzymatique a été effectuée sur l'ADN plasmidique extrait des bactéries à différents temps après induction de la transcription de l'intron Pl.LSU/2 et de l'HisV5-IEP. Cette analyse n'a pas permis de détecter le plasmide accepteur dans lequel se serait intégré l'intron Pl.LSU/2. Le homing de Pl.LSU/2 chez *E. coli* n'a donc pas pu être déterminé.

Le homing peut s'effectuer si au préalable l'intron Pl.LSU/2 est épissé dans la cellule. Or, nous n'avons pas mis en évidence la capacité d'épissage de l'intron chez *E. coli*. En revanche, comme mentionné ci-dessus, nous avons montré un épissage de l'intron Pl.LSU/2 chez la levure, dont l'efficacité est augmentée *via* l'activité maturase de l'IEP. Nous avons donc cherché à évaluer le homing de Pl.LSU/2 chez la levure. Pour cela, nous avons utilisé les mêmes plasmides que lors de l'étude de l'épissage de Pl.LSU/2 chez la levure (plasmide contenant l'intron et plasmide exprimant l'IEP), et inclus un troisième plasmide, le plasmide accepteur de homing, contenant les exons E2 et E3 juxtaposés et constituant le site d'intégration naturel de l'intron. Les levures ont tout d'abord été transformées par le plasmide contenant l'intron et le plasmide accepteur, puis les cellules doublement recombinantes ont été transformées ou non par le plasmide exprimant l'IEP. Dans cette stratégie, le homing de l'intron Pl.LSU/2 n'est pas mis en évidence par la restauration d'un gène rapporteur ; il est

donc nécessaire d'analyser l'état des plasmides contenus dans les cellules afin d'évaluer le homing. Après une période de culture, l'ADN plasmidique a donc été extrait des levures et analysé par PCR afin de mettre en évidence la présence de plasmide accepteur dans lequel l'intron serait intégré. Cette analyse a révélé la présence de ce plasmide dans toutes les levures, même celles non transformées par le plasmide codant l'IEP. En absence d'IEP, le mécanisme de homing ne peut pas avoir lieu, car l'IEP est directement impliquée dans ce mécanisme *via* ses activités RT et endonucléase. Les résultats obtenus pourraient donc être expliqués par la mise en place d'un processus de recombinaison homologue entre le plasmide contenant l'intron et le plasmide accepteur, puisque ces deux plasmides présentent des régions d'homologie (E2 et E3). Dans ce contexte, le homing de Pl.LSU/2 n'a donc pas pu être déterminé. Le choix d'une stratégie alternative, similaire à celle utilisée chez *E. coli*, s'avère donc nécessaire.

Ce travail de thèse constitue donc un apport dans la caractérisation de l'intron de groupe II Pl.LSU/2 et a permis de démontrer l'épissage de l'intron et l'activité maturase de l'IEP *in vivo* chez la levure. L'activité de transcriptase inverse de l'IEP a également été démontrée *in vitro*. La caractérisation du mécanisme de homing de l'intron reste à être effectuer, et de futures optimisations sont requises en vue de l'utilisation des introns de groupe II en ciblage génomique dans les cellules humaines.

MOTS-CLÉS

Thérapie génique; vecteurs intégratifs; intron de group II; Pl.LSU/2 ; Intron-encoded protein; épissage; homing; *Pylaiella littoralis*

INTRODUCTION

This chapter consists in a literature review and describes intellectual background to the work presented in this thesis. A short history of gene therapy is presented and the different integrating approaches are described. Targeting strategies are depicted followed by a focus on group II introns. Finally, the aim of this work, which principally concerns the characterization of a promising group II intron, is outlined.

1 - GENE THERAPY

Gene therapy consists in transferring nucleic acids into target cells of a patient in order to obtain a therapeutic effect. Classical gene therapy is the technology by which genes or small DNA molecules are delivered to human cells, tissues, or organs to correct a genetic defect, or provide new therapeutic functions for the ultimate purpose of preventing or treating diseases. These approaches have been diversified from the treatment of monogenic disorders to cancer treatments or prevention of a disease by transferring genes encoding therapeutic proteins as a vaccine. More recently, nucleic acids transfer has been used to perform gene repair rather than supplementation, which opens the path of dominant genetic disorders treatment. Gene therapy has also been developed to modulate the expression of an endogenous protein by targeting the messenger RNA, either to restore its expression or to degrade it (Watts JK and Corey DR 2012). After decades of research, gene therapy has now been successfully used to treat a number of disorders in humans.

Nucleic acid transfer into cells requires a transporter, also called vector. Indeed, transfection of naked DNA without vectorization is poorly efficient due to several cellular hurdles such as the crossing of the cellular membrane, the routing to the nucleus, and the crossing of the nuclear envelope. Two different strategies have been exploited in gene therapy: either a direct *in vivo* delivery of the vector, or an *ex vivo* strategy, which consists in the delivery of the vector to patient target cells before re-engraftment of these modified cells into the patient. The *in vivo* delivery of the vector could be achieved by different administration routes (intravenous systemic administration, or local administration such as intraocular, intramuscular, etc.), which efficiency can be impeded by several extracellular barriers such as immunity, natural filters, vascular walls, extracellular matrix, etc. The feasibility of *ex vivo* gene therapy is usually less complex than those of *in vivo* gene therapy due to the absence of extra-cellular barriers and as the required quantity of vector is usually lower (at least compared to *in vivo* systemic administration).

The idea of gene therapy emerged in the early 60s, which was at this time only theoretical. In 1964, Edward Tatum declared in a perspective paper: "*Within the next hundred years great advances can be expected in the control of mutational processes, in the design and synthesis of genetic determinants, and in the development of techniques for the introduction of such new genetic determinants into the genome of living organisms.*" (Tatum EL 1964). He detailed his idea during a symposium in 1966 with the description of what will define the basis of *ex vivo* gene therapy: "*The first successful genetic engineering will be done with the patient's own cells, for example, liver cells, grown in culture. The desired new gene will be introduced by directed mutation, or from normal cells of another donor by transduction, or by direct DNA transfer. The rare cell with the desired change will then be selected, grown into a mass culture, and reimplanted in the patient's liver.*" (Campbell TL 1966).

The first two clinical trials were initiated in the early 70s and 80s and were both unapproved. In the first trial, the wild-type Shope papilloma virus was delivered directly into patients presenting an arginase deficiency syndrome with the hope that the viral arginase would replace the missing enzyme in these patients (Friedmann T 1992). The second gene therapy trial consisted in a transfection of bone marrow cells collected from beta-thalassemia patients with plasmids encoding the human beta-globin gene, before re-infusion of the cells into the patients. Although no adverse effect was observed in both cases, no therapeutic benefit resulted from those trials.

The major significant bottleneck for realization of gene therapy at this time was thus to efficiently insert foreign DNA into human cells. The need of efficient vectors for the introduction of transgenic DNA into mammalian cells stimulated researches in this field. A number of improvements in non-viral vector strategies were performed using liposomes (Schaefer-Ridder M et al. 1982), electroporation (Neumann E et al. 1982), and later the polycation polyethyleneimine (Boussif O et al. 1995). It was not until the late 70s that the potential use of viruses as gene therapy vectors became evident. Viruses are indeed naturally-occuring vehicles for the introduction of foreign DNA into cells. The first viral vector was based on Simian Virus 40 (SV40) in 1979 (Hamer DH and Leder P 1979) followed by the development of γ-retroviral vectors (Shimotohno K and Temin HM 1981; Wei CM et al. 1981; Tabin CJ et al. 1982), adenoviral vectors (Van Doren K et al. 1984) and adeno-associated virus (AAV) vectors (Hermonat PL and Muzyczka N 1984). These vector systems were later refined, for example by splitting viral genomes so that viral protein sequences could be removed (Mann R et al. 1983), or the development of retroviral vectors based on the lentivirus HIV-1 to transduce non-dividing cells (Naldini L et al. 1996). Viral vectors are now deficient for replication and the maximum of viral sequences are removed. A comparison of the major gene therapy vectors is depicted in Table I-1.

	γ-retrovirus	Lentivirus	Foamy virus	Herpes virus	Adenovirus	AAV	Non-viral
Nucleic acid in vector	RNA	RNA	RNA	DNA	DNA	DNA	DNA
Packaging capacity	9 kb	10 kb	12 kb	30 kb	30 kb	4.6 kb	Unlimited
Tropism	Broad	Broad	Broad	Neurotropic	Broad	Broad	Broad
Integration into host genome	Yes	Yes	Yes	No	No	Rarely	Rarely
Transgene expression	Long-term	Long-term	Long-term	Transient	Transient	Long-term in post mitotic cells	Transient
Transduction of post-mitotic cells	-	+	+	+++	+++	++	++
Pre-existing immunity	None	None	None	Yes	Yes	Yes	None
Safety concerns	Insertional mutagenesis	Insertional mutagenesis	Insertional mutagenesis	Inflammatory responses	Inflammatory responses	Low risk of insertional mutagenesis	-

Table I-1: Properties of major gene therapy vectors.
(-): zero; (+): low; (++): moderate; (+++): high. *Adapted from (Nathwani AC et al. 2005).*

The development of viral vectors improved the efficiency of gene transfer so that clinical benefit from gene therapy could be expected. This enables the first fully regulated clinical trial in 1990 (Anderson WF et al. 1990). In this trial, two children with severe combined immunodeficiency caused by a lack of adenosine deaminase activity (ADA-SCID) were treated by *ex vivo* gene therapy using autologous T-lymphocytes, transduced with a Moloney murine leukaemia virus (MoMLV)-derived vector encoding the ADA gene, and subsequently retransplanted in the patient. No adverse effects were observed in this ADA-SCID trial, and a significant expression of ADA was detected in transduced cells recovered from the patients (Blaese RM et al. 1993; Blaese RM et al. 1995). However, transient transgene expression required regular infusions of transduced cells and enzyme replacement therapy.

This was due to the fact that the corrected cells were differentiated T-lymphocytes with a limited proliferative potential. Therefore, in 1992, another gene therapy clinical trial using a MoMLV-derived vector for the treatment of ADA-SCID was initiated, but this time targeted both autologous peripheral blood lymphocytes and hematopoietic stem cells (HSC) (Bordignon C et al. 1995). This trial led to both short-term and long-term reconstitution of the patient immune system and correction of growth failure. However, enzyme replacement therapy was still required. Although the clinical benefit of these two studies were limited, they open the way of a successful clinical trial by Aiuti and coworkers, where nonmyeloablative bone marrow conditioning facilitated the engraftment of gene modified HSC (Aiuti A et al. 2002; Aiuti A et al. 2009).

Subsequently, the number of clinical trials initiated from the 1990s increased significantly (Fig. I-2).

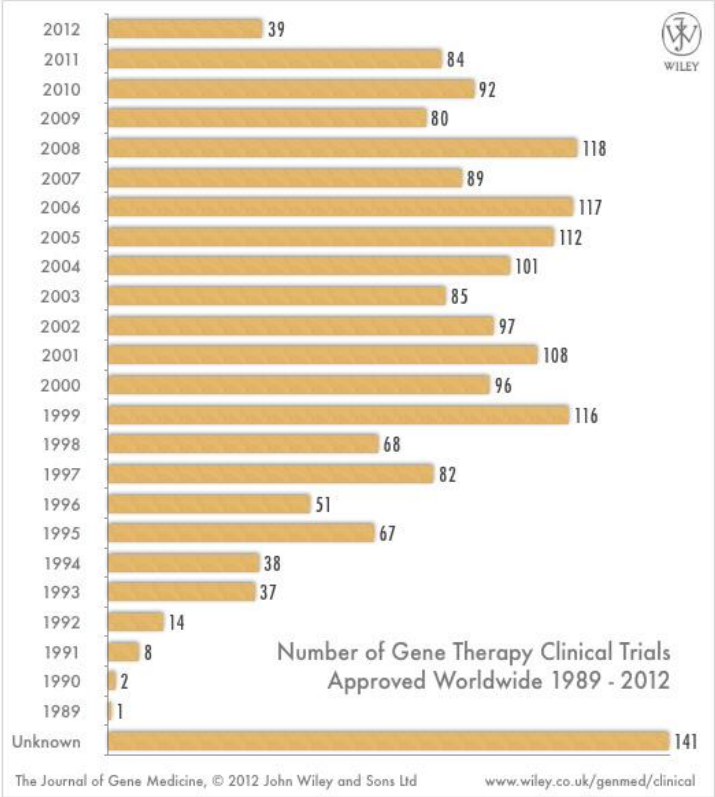


Figure I-2: Number of gene therapy clinical trials from 1989.

Source: The Journal of gene Medicine; Clinical Trials database (www.wiley.co.uk/genmed/clinical ; update June 2012). Unknown section corresponds to clinical trials for which information of approved/initiated year is missing.

The first serious adverse event due to a gene therapy protocol occurred in 1999 during a trial using an adenoviral vector to treat the liver metabolic disorder ornithine transcarbamylase deficiency. The patient developed a massive systemic inflammatory response to the adenoviral vector, leading to multiple organ failure and death within few days of vector administration (Raper SE et al. 2003). Investigations on the design of this clinical protocol outlined several protocol violations and unreporting of previous adverse events on animal model. This marked the need of correct clinical trial designs and also of studying immune responses to vectors used in gene therapy.

The first successful gene therapy clinical trial was reported in 2000 in a study initiated in Paris for X-linked severe combined immunodeficiency (SCID-X1) (Cavazzana-Calvo M et al. 2000). In this trial, HSC were extracted from the bone marrow of 10 children and transduced with a MoMLV retroviral vector encoding the interleukin-2 gamma c common chain receptor (*IL2RG*). Cells were then infused into patients intravenously. In all but one patient, a polyclonal T-cell repertoire was found within years of treatment and an antigen-specific response to immunization was detected (Schmidt M et al. 2005; Hacein-Bey-Abina S et al. 2010). This immune system reconstitution enables the withdrawal of immunoglobulin therapy in the majority of the patients and they now can live outside of a sterile confinement. An associated trial was conducted in London with 10 patients incorporated (Gaspar HB et al. 2004) and also resulted in effective immune reconstitution (Gaspar HB et al. 2011). However, both trials experienced serious adverse events (unexpected at this time) with the formation of T-cell leukemia-like expansions (Hacein-Bey-Abina S et al. 2003a). These events have occurred in 5 of 20 treated patients of the two studies combined, leading to one death. The other patients were successfully treated by chemotherapy and entered in remission. The initiated event in those expansions is an insertion of the retroviral vector near oncogenes causing a dysregulated expression of these genes through the action of an enhancer contained in the vector (Hacein-Bey-Abina S et al. 2003b; Hacein-Bey-Abina S et al. 2008; Howe SJ et al. 2008). The subject of insertional mutagenesis is discussed in more details in section 2.2 - of the introduction.

Other notable gene therapy clinical trials have demonstrated therapeutic benefits. They include lentiviral transduction of HSC for treatment of the β-thalassemia anemia (Cavazzana-Calvo M et al. 2010), lentiviral vector transduction of HSC for treatment of X-linked adrenoleukodystrophy (X-ALD) (Cartier N et al. 2009), lentiviral vector transduction of HSC for treatment of the Wiskott-Aldrich syndrome (WAS) (Boztug K et al. 2010; Galy A and Thrasher AJ 2011), retroviral transduction of epidermal stem cells for treatment the junctional epidermolysis bullosa (Mavilio F et al. 2006), retroviral anti-melanoma T-cell receptor immunotherapy (Johnson LA et al. 2009), retroviral T-cell suicide gene therapy to control proliferation following T-cell and bone marrow transplant for leukemia (Bonini C et al. 1997), AAV-mediated neurotransmitter production for treatment of Parkinson's disease (Kaplitt MG et al. 2007), and AAV-mediated expression of RPE65 (Retinal pigment epithelium-specific 65-kDa protein) in retina for treatment of Leber's congenital amaurosis (Bainbridge JW et al. 2008; Maguire AM et al. 2009; Simonelli F et al. 2010; Jacobson SG et al. 2012).

However, gene therapy trials are currently hampered by a number of technical hurdles. The first bottleneck can be represented by the inability of the vector to efficiently transduce the target cell population, as appear to be the case in a clinical trial for cystic fibrosis (Griesenbach U and Alton EW 2009). Moreover, expression of the transgene may be lost following promoter silencing, as may have occurred during a clinical trial for treatment of the chronic granulomatous disease (CGD) (Ott MG et al. 2006; Grez M et al. 2011). Efficient engraftment and expansion of cell transplants modified *ex vivo* may be limited in absence of significant survival advantage for transduced cells, as was observed in a clinical trial for anti-HIV gene therapy (Mitsuyasu RT et al. 2009). Patients may also develop immune response to the transgene product or the vector itself (Manno CS et al. 2006; Mingozzi F et al. 2009; Mendell JR et al. 2010).

Although clinical application of gene therapy remains experimental due to technical challenges, sustained research efforts are made to overcome those hurdles. To date, gene therapy protocols have improved the health (sometimes dramatically) for dozens of patients.

A number of gene therapy approaches currently use tools that can enable active integration of the therapeutic transgene into target cell chromosomes. Indeed, even though episomal nuclear DNA can integrate into chromosomes under the action of host DNA repair proteins (Stephen SL et al. 2008), this mechanism is likely too inefficient to be useful for most applications. Strategies for which the therapeutic transgene is integrated into the host chromosome are better candidates for targeting cells that have rapid turnover, as hematopoietic cells. Gene therapy that target mitotic cells and/or tissues thus required the integration of the transgene for long-term transgene expression. The following chapter describes the major classes of integrative approaches in gene therapy.

2 - INTEGRATIVE APPROACHES

Several strategies, either viral or non-viral, are used or developed to allow transgene integration into the host cell genome. The transgene integration can be random, semi-random, oriented, or site-specific, depending on the system used.

2.1 - RANDOM AND SEMI-RANDOM INTEGRATION

2.1.1 - DNA transposon vectors

DNA transposons are naturally mobile genetic elements residing in the genome as repetitive sequences that can move through a direct cut-and-paste mechanism called transposition, in which the transposon gets excised from the donor locus and is subsequently integrated into another location by the transposase protein. DNA transposons were first extensively used as genetic tools in invertebrates and in plants for transgenesis or insertional mutagenesis (Osborne BI and Baker B 1995; Plasterk RH 1996; Hayes F 2003). The most studied DNA transposon for applications in mammals is the Tc1/*mariner*-type Sleeping beauty (SB) transposon.

The SB transposon was reconstructed by consensus alignment of inactive transposon "fossils" from a number of salmonid fish (Ivics Z et al. 1997; Ivics Z et al. 2009). The complete wild-type transposon consists of a transposase protein coding sequence flanked by two nonidentical 230 bp inverted repeats (IR) (Fig. I-3A). The region between the left IR and the transposase coding sequence is able to promote weak transcription of transposase (Moldt B et al. 2007). The 340 amino acid transposase consists of an N-terminal domain presenting DNA binding domains, a nuclear localization signal (NLS), and a C-terminal DDE-type catalytic domain, which is common to integrases and transposases from many mobile elements, like retroviruses (Cui Z et al. 2002; Ikeda R et al. 2007), and contains three carboxylate residues believed to be responsible for coordinating metal ions needed for DNA cleavage activity. The transposase is able to cleave the DNA transposon from flanking DNA and reintegrate it elsewhere by a "cut-and-paste" mechanism (Fig. I-3B). A 32-34 nt transposase binding site is present at the inner and outer end of each IR (Izsvak Z et al. 2002). After binding of the transposase to these sites, a tetramer of transposase is subsequently formed, bringing the two IR into close proximity, and the transposase then cleaves the DNA at the IR. This cleavage is performed only when the transposon is flanked by TA dinucleotides, and is enhanced when the flanking sequence is a TATA motif. The excised transposon subsequently integrates into a new target DNA at a TA-dinucleotide. The non-homologous end joining (NHEJ) DNA repair pathway allows the integration of the transposon into the target DNA, resulting in a 3 bp footprint on the donor DNA and duplication of the TA-dinucleotide at the target site (Yant SR and Kay MA 2003; Izsvak Z et al. 2004). Potential TA-dinucleotides targets differ in their attractiveness for SB integration due to local DNA characteristics, and structural prediction of integration sites preferences has been described (Geurts AM et al. 2006). However, the integration profile of the SB transposon is random with respect to genomic features such as transcription units and CpG islands (Vigdal TJ et al. 2002), with about 35 % of SB integrations in genes.

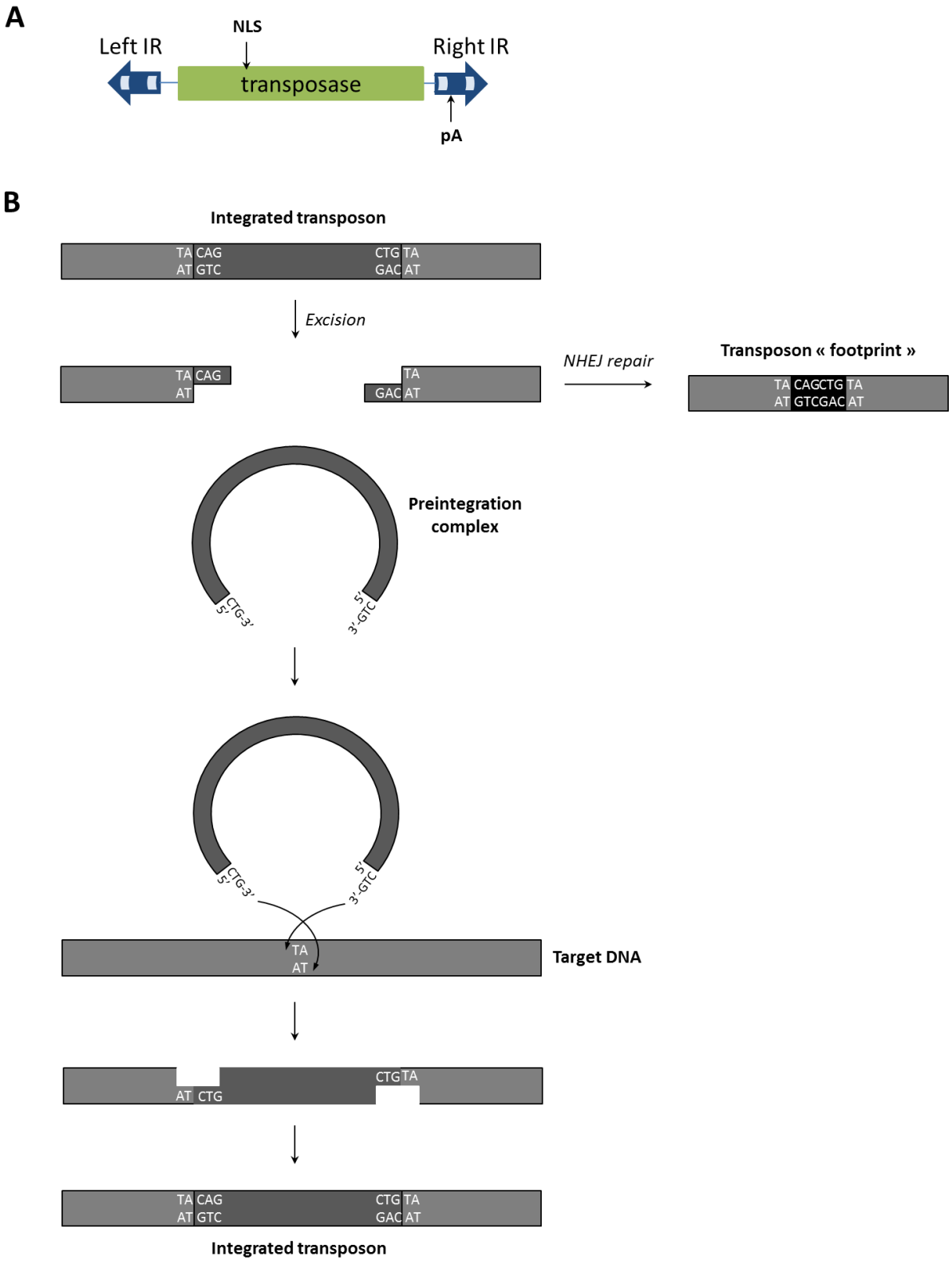


Figure I-3: Sleeping beauty transposon and its transposition mechanism.

(***A***) Schematic representation of wild-type SB transposon consisting of flanking inverted repeats (Left and right IR) and a transposase coding sequence. Light gray boxes: transposase binding sites; NLS: nuclear localization signal; pA: polyadenylation signal. (***B***) Mechanism of SB transposition. The integrated SB transposon is excised by the transposase protein, leaving the donor DNA with 3 bp transposon-mediated overhangs. The donor DNA is repaired by the NHEJ pathway, resulting in the formation of a transposon "footprint". The excised SB

transposon can insert a DNA site containing a TA dinucleotide, which is duplicated at each end of the integration site during the transposon integration. *Adapted from (Plasterk RH et al. 1999).*

Subsequently to the reconstruction of active SB transposon, several other transposons, such as the hAT-like transposon Tol2 from the medaka fish *Oryzias latipes* (Kawakami K and Noda T 2004; Kawakami K 2007), the PiggyBac (PB) transposon from the cabage looper moth *Trichoplusia ni* (Ding S et al. 2005), or the Tc1-like Frog Prince transposon from the frog *Rana pipens* (Miskey C et al. 2003) were shown to effectively transpose in mammalians. Among them, Tol2, SB and PB transposons have been extensively evaluated for gene therapy (Kang Y et al. 2009; Hackett PB et al. 2010; Swierczek M et al. 2012). DNA transposons have thus emerged as particularly attractive vectors for gene therapy. As non-viral vector, they can be delivered into primary cells by conventional transfection techniques, like electroporation (Mates L et al. 2009). Moreover, as there are no viral antigen contained in the vector, they are potentially less immunogenic than viral vectors. DNA transposons combine the desired features possessed by naked DNA and the ability to insert transgenes into host chromosomes, thus enable long-term transgene expression. Finally, the use of DNA transposons may overcome some of manufacturing hurdles intrinsic to the production of high-titered batches of vectors in viral vectors gene therapy protocols.

To turn the SB DNA transposon into a gene delivery tool, three systems were developed (Fig. I-4). The first is composed of a unique plasmid containing the gene of interest with its own promoter and flanked by the transposon terminal repeat sequences in their natural inverted orientation (IR) required for transposition, and the transposase ORF with its promoter (Mikkelsen JG et al. 2003) (Fig. I-4A). It is also possible to provide the transposase on a separate plasmid or by the direct delivery of transposase mRNA (Wilber A et al. 2006) (Fig. I-4B). The wild-type size of SB transposon is 1.7 kb, and each 1 kb increase in transposon length leads to a decrease of 30% in the transposition efficiency (Izsvak Z et al. 2000). However, up to 8 kb inserts can be efficiently transferred with the SB transposon, and the generation of "sandwich" transposons with two complete SB transposons elements flanking the transgene (Fig. I-4B; two ended arrows) enhances transposition of large sequences (> 10 kb) (Zayed H et al. 2004). In each system described, the transposase binds the inverted repeats and catalyzes the excision of the gene of interest from the plasmid, as well as its integration into the host cell genome.

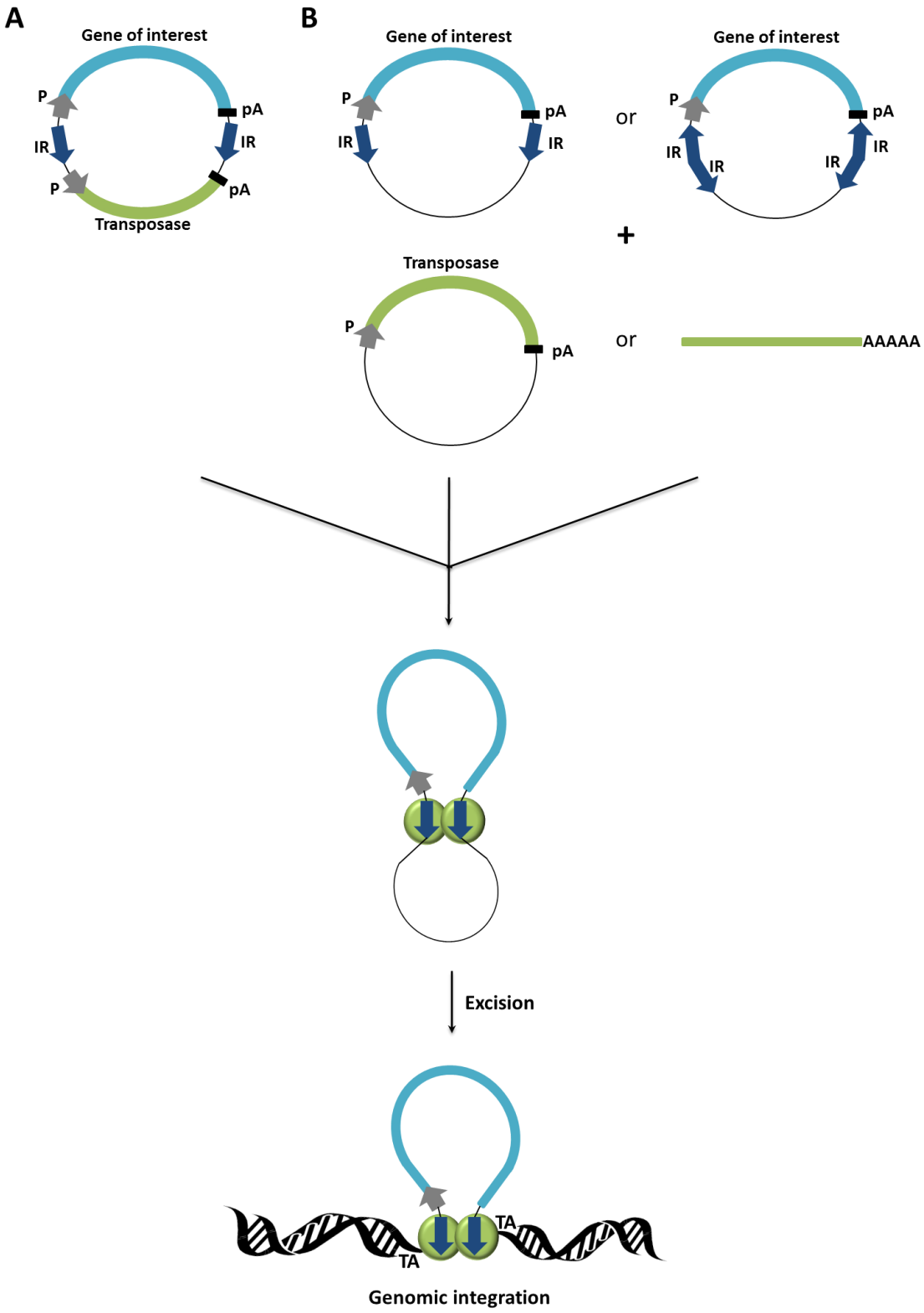


Figure I-4: Schematic representation of plasmid-based SB vector systems.

(A) One plasmid system: the transposase coding sequence and the SB transposon containing the gene of interest are on the same plasmid. (***B***) The transposase is provide either on a separate plasmid or by direct delivery of transposase mRNA (green solid line with polyA queue). The gene of interest can be flanked by the SB inverted repeats (IR) in their natural orientation, or by two IR in a "sandwich" configuration (two-ended blue arrows). pA: polyadenylation signal; P: promoter. The expressed transposase (green circles) binds the inverted repeats

and catalyzes the excision of gene of interest and its subsequent genomic integration at TA dinucleotides, which are duplicated and subsequently flank the transposon at the integration site. *Adapted from (Izsvak Z et al. 2010).*

The level of transposase expression in target cells using SB and PB transposon vectors must be tightly controlled, as overexpression of transposase actually reduces the level of transposition by a phenomenon called overproduction inhibition, while Tol2 transposition is directly proportional to the level of transposase and thus does not exhibit overproduction inhibition (Geurts AM et al. 2003; Wu SC et al. 2006). The mechanism of overproduction inhibition, which has been reported in other *Tc1/mariner* transposons, is not known, but may represent an autoregulatory system (Lohe AR and Hartl DL 1996).

As mentioned above, the efficiency of transposition from transposon vector plasmid declines with the transgene size, limiting its useful capacity. The PB transposon-mediated system can still efficiently transpose inserts of up to 14 kb and the Tol2 transposon vector can transfer genes of up to 11 kb without loss of transposition activity (Izsvak Z et al. 2000; Ding S et al. 2005; Balciunas D et al. 2006).

Constant improvements of transposases led to the development of a set of hyperactive transposases. To date, the most active SB transposase recently developed is called *SB100X* and displays a 100-fold higher transposition efficiency than the originally reconstructed protein (Mates L et al. 2009). The PB transposase, for which the coding sequence was codon-optimized for translation in target cells (*mPB*), exhibits enhanced transposition activity in mouse embryonic stem cells (Bjork BC et al. 2010), and a recent study allows the development of an hyperactive PB transposase, which yields to a 17-fold enhanced excision and 9-fold enhanced integration activity compared to the mPB (Yusa K et al. 2011; Doherty JE et al. 2012).

The introduction of a transgene into the host genome can be targeted for silencing. This phenomenon is often observed when the transgene is inserted by viral vectors. Some studies have reported post-integrative silencing of insertions by epigenetic modifications of the locus, and showed that the promoter within the cargo transgene construct has a major influence in triggering epigenetic modifications of the integrated transposon (Garrison BS et al. 2007). However, a recent study using individual clones containing transposon integrations demonstrated that transposon vector systems using *SB100X*, *Tol2* and *mPB* all showed low levels (1.7 – 3.8 %) (Grabundzija I et al. 2010).

One of the first demonstrations that transposons could be used in gene therapy was obtained by transfecting keratinocytes from patients suffering of junctional epidermolysis bullosa with the sleeping Beauty vector system encoding laminin (Ortiz-Urda S et al. 2003). The gene-modified cells were able to regenerate human skin on immunodeficient mice. Other demonstrations were recently made with the Sleeping Beauty vector system using *SB100X* to integrate SB transposon containing GFP or DsRed marker genes in $CD34^+$ hematopoietic progenitors (Mates L et al. 2009; Xue X et al. 2009). The use of *SB100X* in these studies allowed efficient and stable gene transfer in up to 50% of the hematopoietic colonies and gene-marked cells were able to reconstitute multilineage hematopoietic system after transplantation into immunodeficient mice. SB-based vector system has also been used in animal models for the correction of hemophilia A (Ohlfest JR et al. 2005; Liu L et al. 2006; Kren BT et al. 2009) and hemophilia B (Yant SR et al. 2002), tyrosinemia type I (Montini E et al. 2002), mucopolysaccharidosis I and VII (Aronovich EL et al. 2007) and diabetes (Heggestad AD et al. 2004). The SB transposon-mediated system has recently received the NIH Recombinant DNA Advisory

Committee agreement to be use in a human clinical trial. SB transposon carrying a chimeric antigen receptor is used to genetically modify human T-cells, rendering them specifically cytotoxic for $CD19^+$ B-lineage tumors (Xue X et al. 2009; Hackett PB et al. 2010).

The PB transposon has been explored for both *in vivo* (Ding S et al. 2005; Nakanishi H et al. 2010) and *ex vivo* (Grabundzija I et al. 2010; Di Matteo M et al. 2012) gene therapy. Also, several studies have shown that the PB transposon-mediated system could be an efficient and safe approach to achieve the genetic reprogramming of mouse or human fibroblasts into induced pluripotent stem (iPS) cells (Kaji K et al. 2009; Woltjen K et al. 2009; Yusa K et al. 2009).

As mentioned above, SB transposons do not appear to display integration bias toward genes or CpG islands, and exhibit a random pattern of integration (Vigdal TJ et al. 2002). It was also shown that the distribution of SB transposon integrations inside intergenic sequences is non-random, with a strong bias toward microsatellite repeats (Yant SR et al. 2005). In contrast, the PB transposon exhibit a non-random integration profile with a bias to transcription units (Wilson MH et al. 2007). It has been reported that the frequency of PB integrations in a window of 50 kb around transcription start sites of almost 900 known proto-oncogenes (from the Sanger Cancer Gene Census) was around 2.5% in human primary T cells (Galvan DL et al. 2009). Another study of transposon integrations to sites within a window of either 400 kb or 1000 kb around almost 2100 cancer-related gene (from CancerGene database) shows a higher frequency of Tol2 integrations (up to 9.4%) than PB integrations (up to 6.6%) (Meir YJ et al. 2011). It was also shown that Tol2 and PB transposons are more prone to induce abnormal clonal expansion than SB transposon (Huang X et al. 2010), which can evolve into tumor developments. Although the SB transposon shows a global random integration pattern and displays only weak promoter/enhancer activity, it can still potentially induce adverse genotoxic event. The use of insulator sequences within the transposon was thus evaluated and showed a reduction of the risk of *cis* activation of neighboring genes around the integration site (Walisko O et al. 2008).

Even if DNA transposon-mediated vector system represents an attractive approach in gene therapy, some advances and analyses are still needed. Transfection technologies need to be developed to enable efficient uptake of the transposon and transposase plasmids by the target cells. Although a major breakthrough has been made with the development of the *SB100X* transposase, which can support up to 50% of stable gene transfer in $CD34^+$ cells (Mates L et al. 2009), transposon-mediated vector system remains less efficient than viral vector, with up to 80% of $CD34^+$ transduction using a lentiviral vector in optimal conditions. In addition, detailed analyses of the potential genotoxicity induced by the use of DNA transposon vectors have to be conducted as it was done previously for γ-retroviral and lentiviral vectors, with the use of tumor-prone mouse models or *in vitro* genotoxicity assays.

2.1.2 - Retroviral vectors

The family of retroviruses, known as *Retroviridae*, consists of a number of enveloped positive-sense single-stranded RNA viruses for whom reverse transcription and chromosomal integration of the viral genome are essential stages of the viral cycle. Within this family, the γ-retrovirus, lentivirus and spumavirus (foamy virus) genera have been developed as vectors for gene therapy.

The retroviral vector analyzed in this thesis was an HIV-1-based lentiviral vector, so the following summary is focused principally on this virus. However, the important differences existing between lentiviruses and γ-retroviruses or foamy viruses are also described.

Over the course of their viral cycle, retroviruses alternate between two major forms: the provirus and the virion (reviewed in (Coffin JM et al. 1997)). The provirus consists of double-stranded DNA integrated into a host cell chromosome. Viral RNA and proteins are expressed from the provirus using the host's transcription and translation machinery. These are subsequently packaged at the host plasma membrane into virion particles which are then enveloped by a host-derived lipid membrane. The resulting virion can bind to and enter a new host cell, reverse transcribe its genome to regenerate the double-stranded DNA form and finally integrate it into the host chromosome as a new provirus.

HIV-1 RNA genome is approximately 9 kb in length and encodes several viral proteins (Fig. I-5). The viral genome contains structural (*gag*, *pol* and *env*), regulatory (*rev* and *tat*) and accessory (*vif*, *vpr*, *vpu* and *nef*) genes, which are flanked by long terminal repeats (LTR). The gag gene encodes viral core proteins, which are the matrix protein, the capsid p24 protein (CA), the nucleocapsid, and peptides p2, p1 and p6. The Gag-Pol precursor protein encodes essential replication enzymes, which are the reverse transcriptase (RT), the integrase, and the protease. The *env* gene encodes the glycopolyprotein envelope precursor (gp160), which can be cleaved by a protease to generate the surface glycoprotein gp120 and the transmembrane glycoprotein gp41.

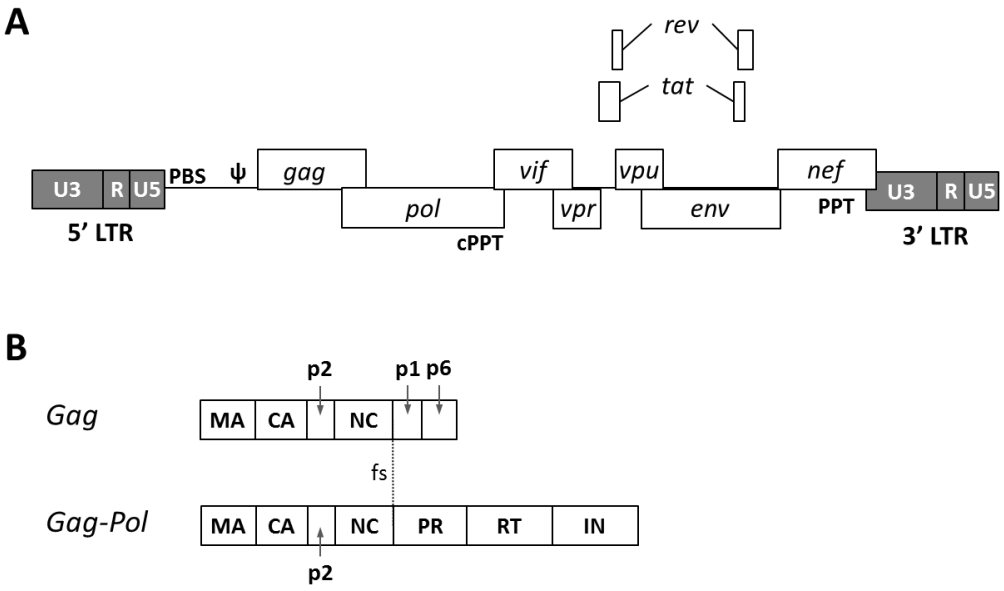


Figure I-5: Schematic representation of HIV-1 provirus and polyprotein structure.

(A) Structure of HIV-1 provirus. LTR: long terminal repeat (subdivided into U3, R and U5 regions); PBS: tRNA primer binding site; ψ: RNA packaging signal; gag: polyprotein encoding virion structural proteins; pol: polyprotein encoding viral enzymes; vif/vpr/vpu/nef: genes encoding accessory proteins; rev and tat: genes encoding regulatory proteins; PPT: polypurine tract. (***B***) HIV-1 Gag and Gag-pol polyproteins. MA: matrix protein; CA: capsid p24 protein; p1, p2 and P6: spacer peptides; NC: nucleocapsid protein; fs: ribosomal frameshift site; PR : protease; RT: reverse transcriptase; IN: integrase.

The viral cycle of HIV-1 is composed by several steps, including transcription of the provirus, translation of viral proteins, virion assembly and budding, virion maturation, cell entry, uncoating, reverse transcription, nuclear import and integration.

Transcription of proviral DNA by the host RNA polymerase II (PolII) enzyme is driven by the U3 promoter region within the 5' LTR. The regulation of HIV-1 transcription occurs through both host transcription factors (Pereira LA et al. 2000) and tat protein. In absence of tat, viral transcription from 5' LTR yields short and non-polyadenylated RNA (Ratnasabapathy R et al. 1990; Zhou Q and Sharp PA 1995). In contrast to HIV-1, γ-retroviruses do not regulate their transcription in this way, but foamy viruses express a transactivator protein with a similar function (Lochelt M et al. 1994). HIV-1 produces at least 30 different mRNAs through alternative splicing, though about half of these mRNA remains unspliced (Purcell DF and Martin MA 1993). The splicing of γ-retroviruses is less complex, resulting in only two mRNA species. In all retroviruses, the unspliced mRNA, which is the molecule uptake into virion, is required for the production of the viral proteins it encodes. For HIV-1, it is exported from the nucleus via the cellular protein Crm1 and the rev viral protein, which contains a nuclear export signal (Meyer BE et al. 1996) and binds the unspliced mRNA to the Rev response element (RRE) (Malim MH et al. 1989; Bogerd HP et al. 1998). γ-retroviruses such as the Moloney murine leukemia virus (MoMLV) do not express a Rev protein and use a different mode of export of their unspliced mRNA, which may be mediated by the ψ packaging signal (Smagulova F et al. 2005). In the Mason Pfizer Monkey virus, the "constitutive transport element" sequence contained in the unspliced mRNA binds the cellular RNA export factor Tap to allow the unspliced mRNA export (Gruter P et al. 1998).

The translation of HIV-1 mRNA, which is carried out by host ribosomes, is promoted by rev protein through an increase of the ribosomal association (D'Agostino DM et al. 1992). The initiation of translation does not occur at the 5' cap of viral mRNA, but is probably initiated at the internal ribosome entry site (IRES) located upstream of *gag* (Buck CB et al. 2001). The translation of unspliced mRNA gives rise to Gag and Gag-Pol polyproteins (Fig. I-5B), which are translated in a ratio of approximately 20:1. Gag-Pol translation is achieved by a bypass of the gag termination codon through a ribosome frameshift (Fig. I-5B; fs) of one nucleotide back into the Pol reading frame (Jacks T et al. 1988). Gag polyproteins is myristoylated during translation and this modification together with a N-terminal membrane binding domain allow the protein to interact with the membrane (Henderson LE et al. 1983; Zhou W et al. 1994).

The Env polyprotein is translated from a spliced viral mRNA and carries an N-terminal signal peptide which allows its targeting to the rough endoplasmic reticulum. The C-terminal region of Env contains hydrophobic amino acids, which are inserted into the membrane and thus acting as transmembrane anchors (Perez LG et al. 1987). The env protein is glycosylated in the endoplasmic reticulum, which is a critical step for correct Env folding and cleavage (Pal R et al. 1989; Li Y et al. 1993). Env polyprotein, in their oligomeric state, are cleaved by host proteases, resulting in heterodimers of the surface and transmembrane proteins. This cleavage exposes the fusogenic peptide of the transmembrane protein (Gallaher WR 1987), making the Env oligomers competent for cell fusion.

In addition of the basic genes (*gag*, *pol* and *env*) encoded by all retroviruses, HIV-1 expresses the regulatory genes *tat* and *rev* described above and the four accessory genes *vif*, *vpr*, *vpu* and *nef*. These accessory proteins are often multifunctional and their diverse roles in the HIV-1 life cycle are

extensively studied (reviewed in (Malim MH and Emerman M 2008)). It appears that HIV-1 accessory proteins frequently act to protect the virus from restriction factors such as host proteins which serves as antiviral defenses. Vif is thought to be a viral countermeasure against antiviral activity of APOBEC3G (Mariani R et al. 2003). The Vpr protein has several proposed functions such as assisting nuclear import of the preintegration complex (Heinzinger NK et al. 1994), causing cell cycle arrest at the G2/M state (He J et al. 1995; Re F et al. 1995), or activating transcription from the HIV-1 LTR promoter (Felzien LK et al. 1998). Among other functions, Vpu and Nef act to protect from superinfection by reducing the CD4 presentation at the cell surface (Garcia JV and Miller AD 1991; Bour S et al. 1995).

Virion assembly requires the subcellular co-localization of the viral Gag, Gag-Pol and Env proteins together with the viral genome and a number of essential host factors. The Gag protein is the major structural component of the immature virion and can induce assembly and budding of virus-like particles in the absence of other viral components (Shioda T and Shibuta H 1990). Gag multimerization is essential in the formation of virions and there are estimated to be 2000-5000 molecules of Gag in each HIV-1 virion (Briggs JA et al. 2004). Unspliced viral genome is recruited to nascent virions as a dimer through an interaction between the nucleocapsid region of Gag and the packaging signal (ψ) (Clever JL et al. 2002). Among the host factor recruited to virions (reviewed in (Ott DE 2002)), the tRNA primer that is used for first strand DNA synthesis during reverse transcription is packaged via an interaction with the Gag-Pol polyprotein (Khorchid A et al. 2000). The recruitment of Env appears to be a non-specific mechanism, as nascent virions are able to incorporate envelope proteins from other viruses and this phenomenon is called pseudotyping (Zavada J 1982). Rather than a specific recognition of Gag and Env proteins, it is suggested that co-localization of these factors occurs within a specific subcellular structure, possibly lipid rafts (reviewed in (Briggs JA et al. 2003).

Budding is the process by which rafts of multimerized Gag and other components form into spherical bodies surrounded by host-derived lipid membrane. Several observations lead to suggest that retroviruses make use of a pre-existing host exosomal pathway in both virion budding and cell entry (Gould SJ et al. 2003). Indeed, the Tsg101 multivesicular budding protein plays an essential role in HIV-1 budding (Garrus JE et al. 2001). In addition, exosomes can form directly at the plasma membrane or via budding into endosomes within the cell, which are mechanisms that can be related to the observation of HIV-1 budding into endosomes in macrophages and direct budding at the cell surface in T-cells (Raposo G et al. 2002).

During and after budding, virions undergo a morphological transition known as maturation, which induces a shift in virion electron density from the envelope to the core. This process is dependent upon Gag and Gag-Pol cleavages by the viral protease. In the mature virion, (Fig. I-6), the matrix protein is proposed to be bound to the envelope, CA forms the outer shell of the virion core, and the nucleocapsid is associated with the viral genome.

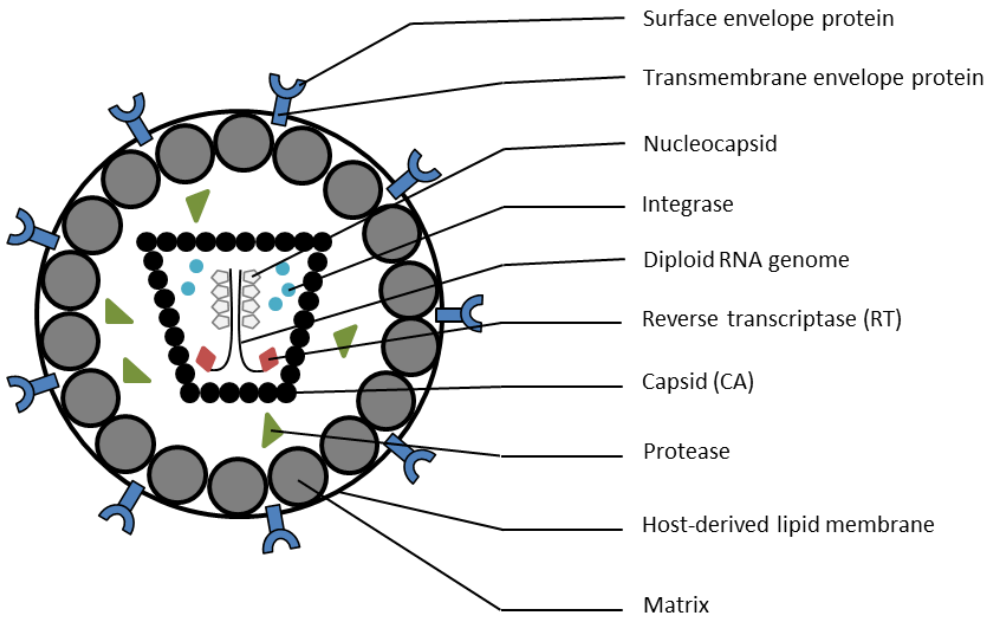


Figure I-6: Schematic representation of a mature HIV-1 virion.

The major determinant of viral tropism is recognition of a host cell surface receptor by the virion Env protein. In addition, an number of non-receptor molecules have been implicated in initial binding if HIV-1 to the cell surface (Fortin JF et al. 1998; Mondor I et al. 1998; Nisole S et al. 1999; Lin CL et al. 2000). Following initial contact, a specific interaction between Env and the host receptor is required for successful viral entry. The primary receptor for HIV-1 is CD4 (Dalgleish AG et al. 1984), but successful infection also require the coreceptor CCR5 (Alkhatib G et al. 1996; Choe H et al. 1996; Deng H et al. 1996; Doranz BJ et al. 1996; Dragic T et al. 1996) and/or CXCR4 (Feng Y et al. 1996), which are both involved in chemokine signaling. It is thought that the initial contact between the surface protein and CD4 induces a conformational change in the former protein, revealing a strongly conserved, high-affinity coreceptor binding domain. Coreceptor binding leads to exposure of the transmembrane fusion peptide (Bosch ML et al. 1989), which may allow an interaction of the fusion peptide to the cell membrane. For the fusion between the host and virus membranes, retroviruses employ the type I fusion pathway (reviewed in (Colman PM and Lawrence MC 2003)). The viral transmembrane protein is believed to insert its fusion peptide into the host cell membrane, subsequently collapsed into a six helical bundle conformation. The membranes are then into close proximity for fusion, resulting in the pore formation (Markosyan RM et al. 2003). Enveloped viruses are also able to fuse host membranes at endosomal membranes following endocytosis, which can take place via clathrin-coated (Daecke J et al. 2005) or calveolae (Hovanessian AG et al. 2004). Reduction of host membrane glycosphingolipid (Hug P et al. 2000) and cholesterol (Manes S et al. 2000) content is known to reduce HIV-1 infection, suggesting a role of lipid rafts.

Uncoating refers to post-entry changes in the protein composition of the intact viral core as it becomes first a reverse transcription complex in which viral DNA synthesis occurs, and then a preintegration complex which is competent for integration into the host genome. Uncoating, reverse transcription and nuclear transport appear to be related processes. RT, IN, the nucleocapsid, the phosphorylated matrix protein (Kaushik R and Ratner L 2004), and Vpr have been detected in the HIV-1 reverse transcription and preintegration complexes, while CA appears to rapidly dissociate (Bukrinsky MI et al. 1993;

Miller MD et al. 1997; Fassati A and Goff SP 2001). Reverse transcription and preintegration complexes are thought to migrate toward the nucleus via interactions with the host cytoskeleton, and more precisely via actin microfilaments (Wilk T et al. 1999; Towers GJ 2007) and microtubules (de Soultrait VR et al. 2002; McDonald D et al. 2002).

The requirement for reverse transcription during the replicative cycle is the defining feature of retroviruses. The essential activities for this process are provided by the viral reverse transcriptase, which has two major activities. Its N-terminal portion carries out RNA- or DNA-dependent polymerization, while its C-terminal portion has RNase H activity. DNA synthesis is thought to be relatively error-prone, with estimates mutation rates of 10^{-4} to 10^{-5} mutations per base-pair per cycle (reviewed in (Laakso MM and Sutton RE 2006)). The primer for minus-strand DNA synthesis is a host-derived $tRNA^{lysine-3}$ (Ratner L et al. 1985), which is annealed to the primer binding site (PBS) (Lanchy JM et al. 1998). Minus-strand DNA synthesis proceeds to the 5'-end of the viral genomic RNA and is followed by transfer of the minus-strand cDNA to the RNA 3'-end (Varmus HE et al. 1978), which occurs through interactions between the cDNA and the 3' RNA 5 sequences. The transferred minus-strand DNA acts as a primer for continued DNA synthesis up to the 5'-end of the remaining RNA (5'-end of the PBS). This process creates an RNA:DNA duplex, which is cleaved by the RT via its RNase H activity. The PPT, located upstream the 3' LTR, is then specifically cleaved, resulting in the formation of primer used for plus-strand DNA synthesis by RT (reviewed in (Rausch JW and Le Grice SF 2004)). The RNA-DNA junction at the 3'-end of the PPT is then cleaved by a second RT enzyme (Gotte M et al. 1999), and Plus-strand synthesis continues until the end of the tRNA PBS. The tRNA is then removed by RT RNase H activity (Pullen KA et al. 1992). The completion of the plus-strand DNA is followed by its transfer, which occurs through base-pairing of the PBS on the two DNA strands (Wakefield JK and Morrow CD 1996). The DNA synthesis is then continued to the end of both LTR. The presence of a second central PPT (cPPT) is an original feature of lentiviruses, including HIV-1 (Charneau P and Clavel F 1991), and it acts as a second primer for plus-strand synthesis.

Lentiviruses such as HIV-1 can efficiently infect non-dividing cells (Lewis PF and Emerman M 1994), while γ-retroviruses such as MoMLV cannot (Roe T et al. 1993). This observation led to postulate that γ-retroviruses are unable to cross the nuclear envelope and so can only access to chromosomes for integration when the nuclear envelope breaks during mitosis. Yamashita and Emerman showed that an HIV-1-based virus in which Vpr was removed, the cPPT was inactivated by mutations, and the HIV-1 In and matrix protein were replaced with their MoMLV counterparts has no difference in infectivity between dividing and nondividing cells (Yamashita M and Emerman M 2005), implying that the ability of lentiviruses to infect nondividing cells is not induced by these nuclear-localizing factors. The authors argued instead that CA is the key determinant, as HIV-1 particles carrying MoMLV CA protein are unable to infect nondividing cells (Yamashita M and Emerman M 2004). The suggested process involves the nondissociation of CA from the MoMLV cores that could impede it to enter the nucleus may be by its too large size.

Integration of viral DNA into a host chromosome (reviewed in (Hindmarsh P and Leis J 1999)) is catalyzed by the integrase (IN), which contains three domains: an N-terminal domain with an HHCC zinc finger motif (Engelman A and Craigie R 1992); the catalytic core domain in which a DDE motif coordinates the Mg^{2+} ions that catalyze integration (Gao K et al. 2004); and a C-terminal domain

which contains an SH3 fold involved in the sequence-nonspecific DNA binding (Engelman A et al. 1994). Linear viral DNA in the nucleus may be converted to 1-LTR or 2-LTR circles by homologous recombination (HR) or NHEJ in the nucleus, but unlike the linear form, they are not efficient substrates for integration (Brown PO et al. 1987; Lobel LI et al. 1989). Two nucleophilic attacks are involved in the integration mechanism and each takes place at each end of the linear DNA molecule. This process is initiated by the Mg^{2+} ion of IN and is known as 3'-end processing (Van Maele B et al. 2006). The 3' hydroxyl group formed on processed viral LTR is sufficiently nucleophilic to attack the target DNA once the IN has bound a host cell chromosome, and both LTRs attack the target DNA, resulting in a joining intermediate (Craigie R et al. 1990), which is subsequently resolved by unpairing the target strand bases between the two positions at which the strands are joined. The two sections of complementary, unpaired sequence at either end of the proviral insertion are most likely repaired by host DNA repair factors (Yoder KE and Bushman FD 2000).

A number of host proteins are thought to participate in HIV-1 integration such as barrier to autointegration factor, High Mobility Group A1, EED, p300 (reviewed in (Turlure F et al. 2004) and (Van Maele B et al. 2006)), and the chromatin-tethering lens epithelium-derived growth factor LEDGF/p75, which has an major role in HIV-1 integration (reviewed in (Engelman A and Cherepanov P 2008)). In addition to its tight association with chromatin throughout the cell cycle, LEDGF binds to HIV-1 integrase via a C-terminal (Cherepanov P et al. 2004). Knockdown of LEDGF expression or overexpression of dominant negative LEDGF containing an integrase binding domain but no chromatin binding domain severely inhibits HIV-1 integration (Llano M et al. 2006). It has been suggested that chromatin-associated LEDGF acts as a chromosomal receptor for the HIV-1 preintegration complex, tethering it to the chromosome and thereby promoting the strand transfer reaction.

The pattern of chromosomal integration by HIV-1 and other retroviruses was for many years believed to be virtually random. More recently, the occurrence of severe adverse events related to retroviral integration during gene therapy, combined with the availability of methods to clone, sequence, and map integration sites, has led to a greater understanding of retroviral integration site preferences. Retroviral integration is not thought to be strongly dependent upon primary sequence, but weak palindromic consensus sequences have been detected at HIV-1 integration sites (Holman AG and Coffin JM 2005. The weakness of the consensus sequences suggests that the contribution of primary sequence to HIV-1 integration site selection is not one of specific recognition of base pairs by integrase, but rather that particular sequences result in local physical properties of DNA such as bendability and protein deformability that are conducive to the mechanism of integration {Bushman, 1994 #564; Wu X et al. 2005). On a larger scale, HIV-1 displays a preference for integration into actively transcribing genes (Wang GP et al. 2007). The mechanism by this preference comes about is not well understood (reviewed in (Bushman F et al. 2005)). Knockdown of LEDGF/p75 in target cells biases HIV-1 integration away from active genes and towards CpG islands and transcription start sites, suggesting a role for this factor in HIV-1 integration site selection (Ciuffi A et al. 2005). Although there is a high degree of sequence homology between retroviral integrases, γ-retrovirus and foamy virus integrases do not interact with LEDGF/p75 and integrate preferentially near CpG islands and transcription start sites. Following integration, the presence of a provirus may disrupt nearby host genes or regulatory elements with potentially deleterious effects on host cell function, as described thereafter in Introduction section 2.2 -.

The abilities of retroviruses to efficiently infect target cells via receptor-mediated uptake and to integrate their viral genome into a host chromosome have made them attractive tools for gene transfer and gene therapy. To date, retroviral vectors remain the most commonly used vectors in gene therapy trials with adenoviral vectors (Fig. I-7).

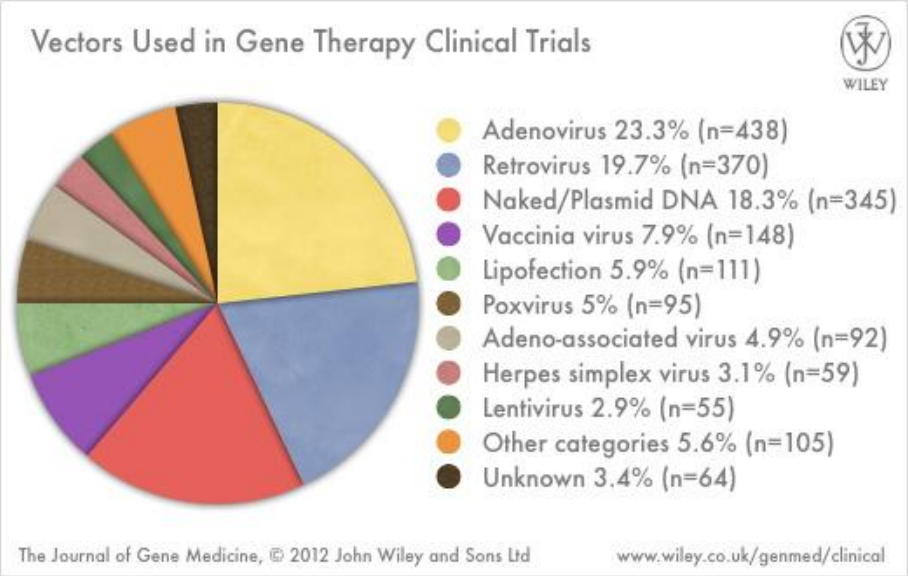


Figure I-7: Vectors used in gene therapy clinical trials.

Source: The Journal of Gene Medicine; Clinical Trials database (www.wiley.co.uk/genmed/clinical ; update June 2012). Unknown corresponds to clinical trials for which information of the vector type used is missing.

Retroviral vector based on the γ-retroviruses MoMLV were one the earliest viral vector developed for gene therapy (See Introduction section 1 -) and lentiviral vectors based on HIV-1 were developed later using many of the principles of the original γ-retroviral system. The major structural change in moving virus to vector was to split the genome into the non-coding sequences required in *cis* for gene transfer and the viral coding sequences required only in *trans* in the producer cell. This separation renders the vector capable of only one round of infection, since no viral proteins will be produced in the target cell. Three generations of HIV-1-based lentiviral vectors have been developed successively to increase their level of safety. Only the last generation of lentiviral vector is thereafter described.

The system is composed by four plasmids: a transfer vector containing the essential *cis* elements and the transgene; a packaging plasmid expressing Gag and Gag-Pol; a plasmid expressing Rev; and a plasmid expressing the envelope pseudotype (Fig. I-8). Separation of the genome onto multiple plasmids reduces the risk of recombination resulting in the production of replication-competent retroviruses. This was historically a significant safety concern for early retroviral vector production systems which routinely gave rise to replication competent viruses through recombination (Donahue RE et al. 1992) but has not been observed with modern split-genome systems.

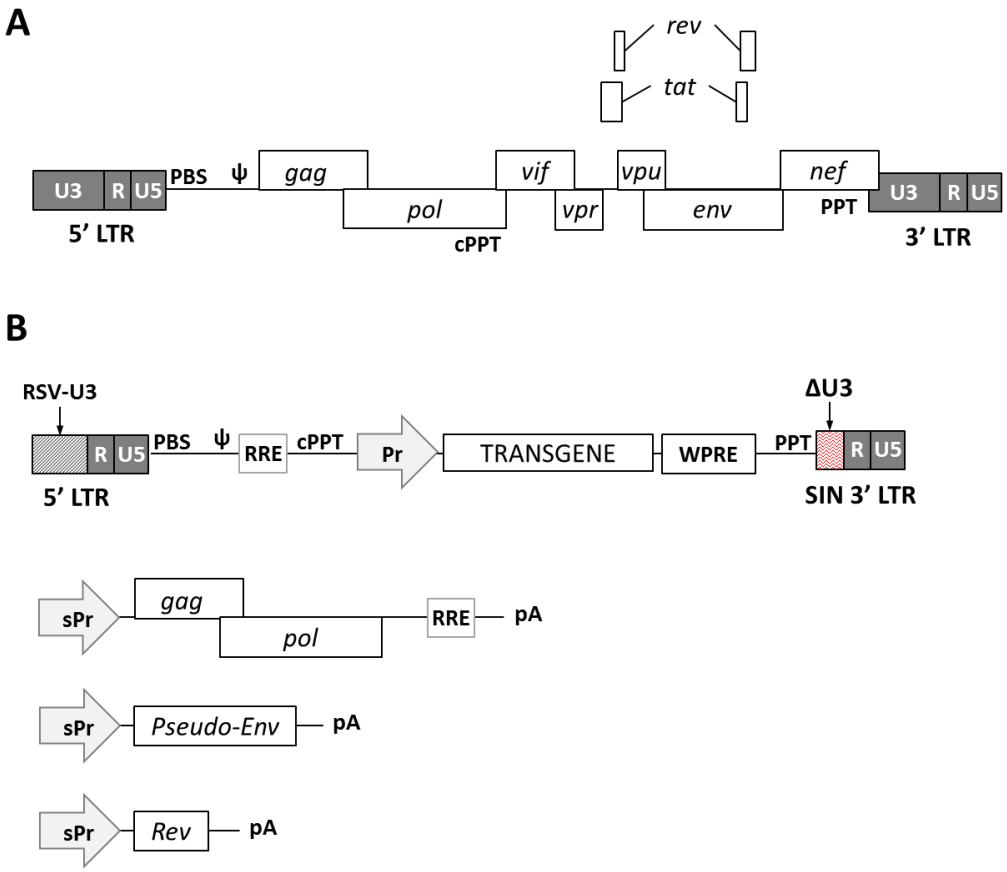


Figure I-8: HIV-1 provirus and third generation lentiviral vectors.

(***A***) Map of a wild-type HIV-1 provirus. LTR: long terminal repeat (subdivided into U3, R and U5 regions); PBS: tRNA primer binding site; ψ: RNA packaging signal; gag: polyprotein encoding virion structural proteins; pol: polyprotein encoding viral enzymes; vif/vpr/vpu/nef: genes encoding accessory proteins; rev and tat: genes encoding regulatory proteins; PPT: polypurine tract. (***B***) Map of a four plasmid third generation lentiviral vector system. RSV-U3: U3 region of the Rous sarcoma virus (RSV); RRE: Rev response element; cPPT: central polypurin tract; Pr: promoter; WPRE: woodchuck hepatis virus posttranscriptional regulatory element; SIN 3' LTR: Self-inactivating LTR; ΔU3: deleted U3 region lacking promoter activity; sPr: strong promoter; pA: polyadenylation signal; Pseudo-Env: envelope pseudotype. The function of all components of the lentiviral vector system is described in the text.

The transfer vector consists of the viral LTRs, the RNA packaging signal (ψ), and the transgene expression cassette. The viral LTRs, containing essential sequences for transcription, reverse transcription and integration, has been modified in order to enhance the safety of lentiviral vector system. The 5' LTR U3 region can be replaced by strong viral promoters such as the Rous sarcoma virus (RSV) U3 region, thus enabling this third generation vectors to be Tat-independent (Dull T et al. 1998; Kim VN et al. 1998). The 3' LTR of the transfer vector is also mutated, so that almost all of U3 region is removed in order to eliminate its promoter/enhancer activity (Zufferey R et al. 1998). This mutation is duplicated in the 5' LTR during reverse transcription, and thus is present on both LTRs of the proviral DNA. This Self-inactivating mutation (SIN) thus reduces drastically the propagation of spontaneously produced replication-competent recombinant HIV-like viruses, due to the inability of transcribing the full viral genome. It also reduces insertional activation of cellular oncogenes by residual promoter activities of integrated LTRs (See Introduction section 2.2 -) and mobilization of integrated vectors by wild-type virus. Other improvements include the insertion of the cPPT element

to improve the transduction efficiency of nondividing cells (Follenzi A et al. 2000; Demaison C et al. 2002; Van Maele B et al. 2003), and a woodchuck posttranscriptional regulatory element (WPRE) which increases the amount of unspliced RNA in both nucleus and cytoplasm (Zufferey R et al. 1999; Ramezani A et al. 2000; Brun S et al. 2003). It was reported that a short "X-protein" encoded within WPRE might be oncogenic (Kingsman SM et al. 2005), so inactivation of the "X-protein" by mutation in WPRE was performed and was shown to unaffect WPRE function (Zanta-Boussif MA et al. 2009). Though strong constitutive promoter are commonly used to drive transgene expression, regulated expression from transfer vectors can be achieved to some extend with the use of endogenous promoters {Martin, 2005 #669}(Dupre L et al. 2004), tissue-specific promoters (Hioki H et al. 2007; Richard E et al. 2008), inducible promoters (Benabdellah K et al. 2011) or even with the insertion of microRNA (miRNA) target sequences to impede transgene expression in unwanted cells (Brown BD et al. 2006; Brown BD et al. 2007; Papapetrou EP et al. 2009).

In this third generation vector system, Rev is encoded in a separate plasmid. Rev/RRE is required for optimal lentivector production, as it overcomes the nuclear retention of the lentiviral genomic RNA, which is mediated by inhibitory sequences involving the splice donor and *gag*. Alternative Rev/RRE-independent systems have been developed, such as the use of codon-optimized gag-pol genes (Kotsopoulou E et al. 2000), or the use of the constitutive transport element of Mason-Pfizer monkey virus (Wagner R et al. 2000; Wodrich H et al. 2001).

Retroviral vectors are able to incorporate envelope proteins from a wide range of enveloped viruses if they are co-expressed in producer cells, a phenomenon known as pseudotyping. The choice of envelope pseudotype alters the target cell specificity and physical properties of the virion (Cronin J et al. 2005). The most commonly used envelope protein is the vesicular stomatitis virus glycoprotein (VSV-G) which confers stability and broad tropism to viral particles (Burns JC et al. 1993). However, VSV-G-pseudotyped vectors at high concentrations can represent immunostimulatory elements at high concentrations (Pichlmair A et al. 2007), whereas the non-specific tropism may pose some safety concerns due to gene transfer into undesired cell types. In this context, envelope glycoproteins from other retroviruses having broadly but less ubiquitously tropism are exploited, such as the feline endogenous retrovirus RD114 Env or the gibbon ape leukemia virus (GALV) Env, which allow efficient transduction of $CD34^+$ hematopoietic progenitors (Hanawa H et al. 2002; Relander T et al. 2005). In addition, some efforts are made to generate targeting lentiviral vectors using a ligand protein or antibody fused to viral glycoproteins to retarget the lentiviral particles to specific cell-surface molecules (Yang L et al. 2006; Frecha C et al. 2008; Funke S et al. 2008; Gennari F et al. 2009).

The level of expression from integrated retroviral vectors is subject to positional effects whereby adjacent chromosomal elements modulate the level of transgene expression (Lewinski MK et al. 2005). This positional effect can induce a transcriptional silencing of the integrated provirus, which may involve methylation and histone deacetylation (Ellis J and Yao S 2005), thus limiting the potential therapeutic benefit. The use of chromatin insulators in the vector backbone can reduce these effects (Emery DW et al. 2000; Knight S et al. 2012).

The safety concerning the use of retroviruses-derived vectors in gene therapy is discussed in section 2.2 -.

2.1.3 - Recombinases

Recombinases are specialized proteins which catalyze site-specific recombination between short recognition sites present on two DNA molecules. Many recombinases have been identified in a large number of prokaryotes and eukaryotes, but almost all fall into two families, namely the tyrosine and serine recombinases.

The tyrosine recombinase family (also known as λ integrases) include the Cre recombinase from bacteriophage P1 (Sternberg N et al. 1986) and Flp recombinase from the yeast *Saccharomyces cerevisiae* (Hartley JL and Donelson JE 1980), while the serine recombinase family includes ΦC31 integrase from the bacteriophage of *Streptomyces* species (Kuhstoss S and Rao RN 1991; Rausch H and Lehmann M 1991). The mechanism of recombination used by each family has been well-studied (reviewed in (Smith MC and Thorpe HM 2002; Grindley ND et al. 2006)). Although the two are mechanistically quite distinct, there are similarities between the two recombination processes. Both tyrosine and serine recombination sites contain two inverted recombinase binding sites. The size of the total recombination site varies between recombinases but is greater than 30bp for the most commonly used systems. A number of recombinases have been investigated for use in gene therapy, and the most studied is the ΦC31 integrase. This section describes in more details the mode of action, the applications and the safety of ΦC31.

In its natural host bacterial cell, ΦC31 integrase catalyzes the integration of the ΦC31 phage genome into the bacterial genome in a precise and undirectionaly way. Integration is mediated by a binding of ΦC31 integrase to its attachment sites, named *attB* (bacterial attachment site) and *attP* (phage attachment site), which are present on both DNA molecule and are approximately 50% identical. This process is only phage-dependent and does not require any host cell cofactors. The integration leads to the formation of two hybrid sites, *attL* and *attR*, which are not substrates of ΦC31 (Thorpe HM and Smith MC 1998). During recombination, serine recombinases introduce one nick into each of the four DNA strands, which are subsequently resolved with integration of the DNA. When at least one parental molecule is circular (e.g. a plasmid) and the other is linear (e.g. a chromosome), the net result of recombination is the insertion of the circular parent into the recombination site of the linear parent. The different properties of serine recombinase such as unidirectional integration, large recognition sites and the absence of cofactors have made of ΦC31 integrase an attractive tool for gene engineering.

A significant limitation is the requirement for target recombination sites in genome. Although human cells contain no perfectly matched recombination sites for ΦC31, divergent pseudo-*attP* sites are present, and can be used by ΦC31 integrase with reduced efficiency (Groth AC et al. 2000). A survey of the integration sites used by ΦC31 integrase in human cell lines found that several hundred potential sites exist, but the majority of integrations take place at a small subset of these (Chalberg TW et al. 2006). Of these hotspots, 19 integrations sites accounted for approximately 56% of the integration events. These hotspots are located across intergenic regions, introns, and exons in approximately the same proportion, with a slight preference for transcribed regions. An enhanced ΦC31 integrase has recently been developed which is 2-fold more efficient than the wild-type ΦC31 (Keravala A et al. 2009).

Numerous proofs of concept have been reported for the use of ΦC31 as non-viral gene therapy vector system. In an hemophilia B animal model, a hydrodynamic tail injection of an *attB*-containing plasmid flanking the human factor IX and a ΦC31 integrase expressing plasmid was performed in knockout mice for factor IX, and approximately 10% of normal factor IX activity level was shown, well above the therapeutic level (Keravala A et al. 2011). The Duchenne muscular dystrophy (DMD) mouse model mdx was transplanted with muscle precursor cells (MPCs), which had been previously transfected with a ΦC31 vector system integrating a mini-dystrophin gene (Quenneville SP et al. 2007). The modified MPCs transplantation leads to the expression of the mini-dystrophin in muscle fibers and to the reconstitution of the dystrophin complex.

The genotoxicity of ΦC31 vector system was recently evaluated in cord-line epithelial cells (CECLs) isolated from the outer membrane of human umbilical cords (Sivalingam J et al. 2010). The cells were found to have a ΦC31 integrase-mediated integration efficiency of 3.0%. The analysis of 44 independent integration events from polyclonal population revealed 18 distinct loci with 8p22 pseudo-attP site most frequently recovered. The analyses of individual clones showed that most integrations events are found in endogenous retrovirus element (Weiss RA 2006), and it was also shown by analysis of genome copy number on a polyclonal population that loci copy gain and loss could occur through the use of ΦC31, as well as translocations.

Even if there are no reports of oncogenic transformation due to ΦC31 integrase expression *in vivo* (Sivalingam J et al. 2010), safety concerns have been raised following reports that the enzyme can cause chromosomal rearrangements in mammalian cells (Liu J et al. 2006).

While the use of ΦC31 integrase-mediated non-viral vector system seems to be promising, extensive further studies are needed to enhance the integration efficiency and evaluate more precisely the safety of these vectors. It is noteworthy that even under optimal specific conditions, the ΦC31 system could not allow integration at a unique and precise site of the genome.

2.2 - INSERTIONAL MUTAGENESIS

Vector integration into host chromosomes is necessarily a mutagenic event in that it alters the primary DNA sequence of the host. Insertion of DNA may affect functional elements already present at the integration site in a number of ways. Firstly, vectors may contain promoters and/or enhancers able to transactivate neighbouring host genes or dysregulate host promoters at long distances in either direction from the integration site. Secondly, vectors may insert into and disrupt coding sequences, resulting in abnormal or prematurely terminated transcripts. Thirdly, vector insertion may disrupt other regulatory elements such as miRNA cistrons.

Transactivation of neighboring genes by integrated vector

The first insertional mutagenesis event following a gene therapy protocol was obtained during the two principal trials for gene therapy of SCID-X1, which have been performed on a total of 10 patients (Cavazzana-Calvo M et al. 2000; Gaspar HB et al. 2004). The conduct of these two trials was very similar. Autologous bone marrow was extracted from patients, selected for $CD34^+$ to enrich for HSCs and hematopoietic progenitors, and transduced *ex vivo* with a MoMLV-derived γ-retroviral vector carrying the IL2RG cDNA. Cells were then infused back into patients. Both trials were highly successful, resulting in engraftment and expansion of modified cells, correction of γc signaling, and

significant immune reconstitution. However, four patients in the French trial and one in the English trial experienced a serious adverse event in the form of a dysregulated expansion similar to T-cell acute lymphoblastic leukemia (T-ALL) 2-6 years after treatment (Hacein-Bey-Abina S et al. 2003a; Howe SJ et al. 2008). One of these patients subsequently died, but the others responded well to standard anti-leukemia chemotherapy and retained a functioning adaptive immune system after treatment. The initiating event in these leukemic events appears to have been integration of the γ-retroviral vector into host chromosomes nearby known T-ALL proto-oncogenes leading to dysregulation of their expression (Hacein-Bey-Abina S et al. 2003b).

In all cases, a latent period on the order of years was observed between transplantation of gene modified cells and subsequent leukemic expansions. At the time of expansion, the polyclonal T-cell population displays a clonal dominance of one or few T-cell clones. In four of the five leukemic patients, dominant clones were found to contain retroviral insertions within or near the known T-ALL proto-oncogene LMO2 (Pike-Overzet K et al. 2007; Hacein-Bey-Abina S et al. 2008). This gene was overexpressed in mature T lymphocytes, probably as a result of the enhancer activity of the vector promoter. Insertions near other T-ALL proto-oncogenes such as SPAG6, CCND2, and LYL1 have also been identified. A general model has been proposed in which insertional mutagenesis leads to continued expression of developmental genes which are normally expressed in HSCs but normally downregulated during immature T-cell development, and this continued expression disrupts the normal T-cell expansion and maturation processes in the thymus (Rabbitts TH 1998). However, the vector-mediated transactivation on proto-oncogenes is still in debate. Several reports postulate that the IL2RG expression from the vector could cooperate with oncogenic transformation (Dave UP et al. 2004; Pike-Overzet K et al. 2007).

Other insertional mutagenesis events have been reported. A recent report of a leukemia apparition in a WAS gene therapy trial using a retroviral-derived vector showed an association to a LMO2 insertional activation (Trobridge GD 2011). Retroviral vector-mediated clonal proliferation of gene-corrected myeloid cells in patients treated for CGD was also reported (Ott MG et al. 2006). In the dominant clones, clusters of vector integrations were found in MDS1/EVI1, PRDM16 and SETBP1 loci. Paradoxically, this clonal dominance is thought to have contributed to the success of the therapy, increasing the proportion of corrected cells. However, silencing of the transgene has occurred and three patients developed myelodysplasia with monosomy 7. It was postulated that the insertional activation-mediated overexpression of EVI1 gene was the cause of a genomic instability, aberrant expansion, and myelodysplasia (Stein S et al. 2010). In the case of the CGD insertional mutagenesis, the strong enhancer activity of the spleen focus-forming virus (SFFV) LTR is thought to have induce the transactivation of EVI1, as the use of a γ-retroviral vector that do not contain SFFV LTR in another clinical trial did not induce clonal expansions while therapeutic benefit in all treated patients was obtained (Kang EM et al. 2010).

All these clonal expansions resulted from a transactivation of proto-oncogenes by enhancer/promoter activities of the integrated vector. This can be induced by the vector LTRs in the case of retroviral vectors or by the internal promoter/enhancer used to express the transgene.

Several approaches are currently developed to overcome this enhancer activity of retroviral vector LTRs. The use of SIN vectors with internal promoter expressing the transgene has been shown to reduce genotoxicity, as this type of vector design was less prone to cause tumors in a tumor-prone mouse model (Montini E et al. 2009; Montini E and Cesana D 2012). In hematopoietic cells, SIN-

lentiviral vectors integrate near oncogenes at least twice less than LTR-driven γ-retroviruses (Cattoglio C et al. 2007). The transactivation of neighboring genes with SIN-lentiviral or SIN-γ-retroviral vectors is largely dependent on the type of vector's internal promoter (Hargrove PW et al. 2008) and the use of physiological promoters reduces genotoxicity (Zychlinski D et al. 2008; Modlich U et al. 2009). SIN-vectors with internal physiological promoters have been evaluated in numerous clinically relevant models such as WAS, CGD, X1-SCID, recessive dystrophic epidermolysis bullosa, ADA-SCID, or junctional epidermolysis bullosa (Di Nunzio F et al. 2008; Thornhill SI et al. 2008; Trinh AT et al. 2009; Titeux M et al. 2010; Avedillo Diez I et al. 2011; Papanikolaou E et al. 2012; van der Loo JC et al. 2012a; van der Loo JC et al. 2012b).

In the case of DNA transposon vectors, it was shown that the terminal inverted repeats of the SB transposon have very low intrinsic promoter/enhancer activity and thus cannot activate endogenous genes near its integration site (Walisko O et al. 2008). However, transactivation of neighboring genes can still occur via internal promoter/enhancer activity.

A promising strategy to reduce vector-mediated genotoxicity is the incorporation of enhancer-blocking insulators into the vector. Insulators are DNA sequences that block the activity of enhancers on promoters when located between them (Bushey AM et al. 2008). Insulators have several advantageous functions. Firstly, they can protect the neighboring genes of integrated vector from its potential enhancer activities. They can also homogenize the expression of the transgene irrespective of the chromosomal insertion site and even reduce silencing of the transgene. The protective function of the chicken hypersensitive site-4 insulator has been evaluated in several studies using retroviral vectors (Malik P et al. 2005; Aker M et al. 2007; Arumugam PI et al. 2007; Evans-Galea MV et al. 2007; Li CL and Emery DW 2008; Arumugam PI et al. 2009; Li CL et al. 2009; Gaussin A et al. 2012) and DNA transposons (Walisko O et al. 2008), with promising results.

Alteration of host gene transcripts

The integration of the vector in gene therapy can also alter the nature of nearby genes transcripts. An example is observed in a gene therapy trial for β-thalassemia using a lentiviral vector (Cavazzana-Calvo M et al. 2010). A clonal expansion was caused by alteration of the HMGA2 gene expression. In this case, the integrated vector leads to expression of a novel HMGA2 transcript resistant to degradation. As in a CGD trial, the clonal expansion in treated patients is thought to have contributed to the therapeutic benefit. To date, neither clinical evidences supporting the existence of a preleukemic state nor significant alteration of hematopoietic imbalance were found.

Genotoxicity of integrated vectors can thus be induced by posttranscriptional deregulation of host gene expression. This includes the generation of chimeric, read-through viral and cellular transcripts, expressed when the transcription from a provirus inserted into a transcription unit bypass the polyadenylation signal and continues into the cellular gene (Zaiss AK et al. 2002; Schambach A et al. 2007; Almarza D et al. 2011), aberrant splicing (Moiani A and Mavilio F 2012; Moiani A et al. 2012), and premature transcript termination. These chimeric viral-cellular transcripts are subsequently processed by splicing involving both viral and cellular acceptor and donor sites (Uren AG et al. 2005). Read-through fusion transcripts have been shown to cause tumors in experimental models using non-SIN-retroviral vectors (Li Z et al. 2002; Montini E et al. 2009). SIN-lentiviral vectors was also found to induce read-through transcripts (Almarza D et al. 2011; Cesana D et al. 2012), although with less frequency than non-SIN-lentiviral vectors. The recoding of vector splice sites together with the use of

strong polyadenylation signal can drastically reduce read-through transcripts and improve the safety of the vectors.

This mechanism of read-trough is also observed with the use of DNA transposons (Kool J and Berns A 2009) and could be potentially found for any integrative vectors. Careful attention must thus be provided on this issue.

Disruption or alteration of the expression of regulatory elements

Another possibility of insertional mutagenesis-mediated oncogenesis is a dysegulation of the expression of regulatory elements. A high proportion of the genome is transcribed and generates non-coding elements such as miRNAs and long non-coding RNAs. MiRNAs are known to regulate gene expression either by translational repression or mRNA cleavage (Du T and Zamore PD 2005). Several studies have shown that miRNAs may have either proto-oncogenesis properties (Hayashita Y et al. 2005; He L et al. 2005) or tumor suppressor activity (O'Donnell KA et al. 2005). The dysregulation of the expression of oncogenic miRNAs either by transactivation of by read-through may induce tumorigenesis mechanism, as well as disruption of tumor suppressor miRNAs. In fact, oncogenic miRNA cistron activation has already been observed in tumors generated in mice by MoMLV provirus integration (Wang CL et al. 2006). This highlights that vectors integration sites must be analyzed extensively and that insertions in non-coding regions are not necessarily synonym of safety.

Evaluation of the safety of vectors by analyzing vector integrome

It is now recognized that events of clonal dominance *in vivo* is associated to vector insertion sites (Montini E et al. 2006; Kustikova OS et al. 2007; Montini E et al. 2009). Indeed, early integration site profiling from clinical trials showed clusters of insertions *in vivo* from treated patients correlating with the occurrence of adverse events (Deichmann A et al. 2007; Schwarzwaelder K et al. 2007). The presence of specific target regions (CIS: common integration sites) in dominant myeloid clones was also observed in the two patients treated for CGD (Ott MG et al. 2006) and similar finding was recently shown in the WAS clinical trial (Boztug K et al. 2010), supporting the idea of a potential correlation between the presence of CIS and an increased risk of aberrant clonal expansion *in vivo*. A recent study exploiting integration site profiles from experimental models and ALD lentiviral gene therapy trial showed that the distribution of integration sites along the CIS could predict their genotoxic potential (Biffi A et al. 2011). Another recent study analyzed more than 7000 insertion sites retrieved from multiple clinical trials and showed the presence of shared CIS among the trials as well as restricted number of specific loci as preferential targets for retroviral integrations *in vitro* and *in vivo* (Deichmann A et al. 2011).

However, it remains undefined to what extend the presence of CIS is the result of positive clonal selection *in vivo* after cell infusion in the patient, or if it derives from preferential targeting sites at the time of transduction (Cattoglio C et al. 2007). Integration sites selection during *in vitro* transduction seems to be influenced by cellular determinants. The presence of CIS in patient's sample could not be *per se* responsible for abnormal expansion. Indeed, CIS involved in the genotoxic integrations in the SCID-X1 trials were also found at the same frequency in a ADA-SCID trial without leading to leukemic clonal expansions in patients (Aiuti A et al. 2007). Other factors including disease background (Kustikova O et al. 2010), the nature of the transgene, and the occurrence of additional mutations unrelated to vector insertions are also involved in aberrant expansions of transduced clones.

In ADA-SCID, a comparison of vector insertion sites retrieved from pre- and post-transplantation in two gene therapy approaches using either HSC or T-cells target cells showed that vector integration preferences is cell-specific and is related to both epigenomic state and expression profile of the cell type at the time of transduction (Biasco L et al. 2011). These types of bioinformatics strategies with a comprehensive analysis of insertion sites profiles coupled with analyses of epigenome and transcriptome (Cattoglio C et al. 2010a; Cattoglio C et al. 2010b) will be helpful to determine the nature of CIS detection in gene therapy applications and predict genotoxic risk in gene therapy.

All integrating approaches using non-targeting vector integration displays inherent mutagenesis risk. The development of targeting integrative strategies could represent an attractive alternative in gene therapy protocols. Numerous strategies and tools have been evaluated since the past two decades and some of them were found to be promising. Three parameters have to be carefully evaluated during the development of gene targeting tools: the efficiency of integration, the specificity of integration and the genotoxicity induced by the strategy adopted. Several strategies developed for oriented or targeted insertion are described thereafter.

2.3 - ORIENTED INTEGRATION

An approach to target DNA insertion into a chosen location of the genome is to retarget an existing semi-random integration mechanism. Several approaches have been evaluated in this attempt. The first involve the direct fusion of the transposase/integrase to sequence specific DNA-binding domains (DBDs) for retarget the preintegration complex (Fig. I-9A). Another approach consists in tethering the preintegration complex using a DNA-binding proteins that interact with either the DNA to be integrated (Fig. I-9B) or with the recombinase/transposase/integrase (Fig. I-9C).

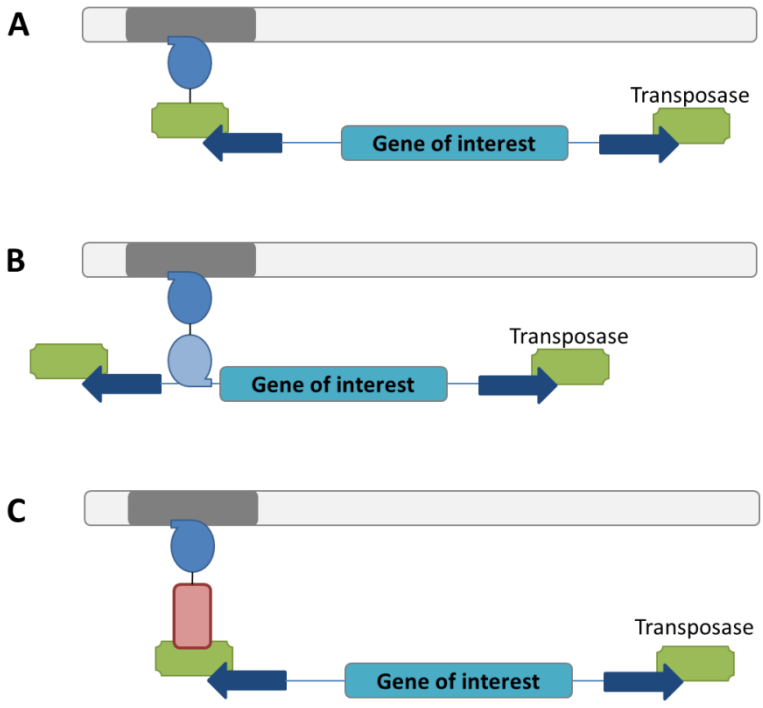


Figure I-9: DBDs-mediated strategies to target gene insertion illustrated for DNA transposon system.
The DNA transposon system is composed of the transposable element containing the gene of interest and flanked by inverted repeats (IRs, blue arrows). The transposase (green rectangle), which binds the IRs, catalyzes the

transposition. (***A***) The targeting of a specific DNA site (dark gray rectangle) is achieved by using a DBD (dark blue circle) recognizing the target site and fused to the transposase. (***B***) Targeting is achieved by using two DBDs fused, one recognizing the DNA target site (dark blue circle) and the other binding the transposable element (light blue circle). (***C***) Targeting is achieved by using a protein interacting with the transposase (red rectangle) fused to a DBD recognizing the DNA target site (Dark blue circle).

2.3.1 - Direct fusion of a DNA-binding domain to transposase/integrase

The feasibility of a system involving the fusion of a DBD to transposase/integrase was firstly evaluated *in vitro*. The integrase of avian sarcoma virus fused to the DBD of the *E. coli* LexA protein showed an alteration of the integration profile with hot spots of integrated vector near a region containing LexA operators without altering the processivity of the integrase (Katz RA et al. 1996). The HIV-1 IN fused to DBD of the phage λ repressor was also able to target integrations near the λ repressor binding sites *in vitro* with no changes in the activity of HIV-1 IN (Bushman FD 1994). The three ZF DNA binding domain of the transcription factor Zif268 fused to HIV-1 IN showed a bias in the integration patterns near its specific binding sites *in vitro* (Bushman FD and Miller MD 1997). However, the infectivity of HIV-1 vector encoding this fusion IN was totally abolished. This strategy was applied to synthetic E2C six-finger ZF domain recognizing a 18 bp DNA site within the 5'-untranslated region of the erbB-2 human gene. *In vitro* study showed that fusion HIV-1 IN-E2C is able to target integrations near the 18 bp E2C binding site and the fusion protein with an efficiency of up to 60% and retains its processivity (Tan W et al. 2004).

The fusion HIV-1 IN-E2C was subsequently tested for retargeting provirus integration to the human chromosomal E2C binding site *in vivo* in HeLa cell line, and it was shown an 10-fold increase of insertions near E2C binding site (Tan W et al. 2006), with however a drastically lower infectivity of viruses carrying fusion HIV-1 IN-E2C of 1 to 24% compared to wild-type IN viruses.

Retargeting using Fusion transposases was also evaluated. The PB transposase was fused to the DNA-binding domain of the Gal4 transcription factor, recognizing a 17 bp DNA site called upstream activating sequence (UAS). Retargeting of PB-Gal4 mediated transposition was evaluated in mosquito embryos on a plasmid target (Maragathavally KJ et al. 2006). The authors demonstrated that 67% of transpositions have occurred into a TTAA site (natural PB target site) located 1 kb upstream of the UAS plasmidic sequence. The SB transposase was fused to either Gal4 DBD or E2C DBD and retargeting of transposition was evaluated on plasmid DNA target sites in HeLa cell line (Yant SR et al. 2007). The authors showed that both SB-Gal4 and SB-E2C induce respectively an 11-fold increase and 8-fold increase of transpositions in the region surrounding the corresponding DNA target site, while fusion transposases activities were decreased to 26% and 20% of the wild-type, respectively. The targeting activity of fusion SB to the artificial three-finger ZF binding domain of the Jazz protein to the 9 bp Jazz binding site located in the utrophin gene (Corbi N et al. 2000) was evaluated in human HeLa cells (Ivics Z et al. 2007). While the fusion SB-Jazz transposase retained only 15% of activity compared to the wild-type SB transposase, no targeted events near the Jazz binding site in HeLa genome could be identified.

Direct fusions of DBDs to transposase/integrase have shown to be challenging. It appears that this type of direct fusion is deleterious to the structure of the protein, generally leading to the decrease of the fusion protein activity. In the case of integrase-DBD fusions, the infectivity of virions is also altered.

Further improvements in the fusion protein design are still needed to impede a too drastic alteration of the structure/activity of the transposase/integrase.

2.3.2 - Use of a DNA-binding domain fusion of a partner protein

The strategy of using a DBD fusion to a partner protein that interact either with transposase/integrase or directly with the DNA to be integrated should be less deleterious regarding the transposase/integrase activity as those proteins are not modified.

Ciuffi et al engineered an artificial tethering factor based on LEDGF/p75, which naturally directs HIV-1 intergation into transcription units. The strategy consists on the construct of the integrase binding domain of LEDGF/p75 fused to the DNA-binding domain of the phage λ repressor. An *in vitro* integration reaction was performed using purified LEDGF- λ repressor tethering factor and HIV-1 integrase to catalyze integration into DNA containing the λ repressor target site. The presence of the fusion tethering factor was found to increase the rate of integration at sites surrounding the target site (Ciuffi A et al. 2006). This alteration of HIV-1 integration preferences using LEDGF/P75 fusion proteins was recently confirmed *in vivo* (Gijsbers R et al. 2010; Silvers RM et al. 2010). The transient overexpression of a fusion protein composed by the integrase binding domain of LEDGF/p75 to heterochromatin protein 1α in HEK 293 cells modify the integration profile of an HIV-1-based lentiviral vector from transcription units to heterochromatin (Silvers RM et al. 2010). The same results were found with the use of fusion protein composed of the integrase domain of LEDGF/p75 and the heterochromatin protein 1β (Gijsbers R et al. 2010), which induce a shift in HIV-1 lentiviral vector integration to heterochromatin regions in HeLaP4-CCR5 cells.

The modification of a partner of the SB transposase has also been tested to evaluate the potential of tethering-mediated targeting (Ivics Z et al. 2007). The authors used the protein-protein interaction domain of the SB transposase called N57 (Izsvak Z et al. 2002) to construct an artificial fusion protein composed by N57 and the tetracycline repressor, which binds the tetracycline response element DNA sequence. A transgenic HeLa cell line containing a TRE-driven eGFP gene was created. It was shown the targeting efficiency was about 10% with integrated SB transposons spanning in a 2.5 kb region around the TRE-eGFP locus (Ivics Z et al. 2007).

The retargeting of SB transposition was also achieved using a fusion protein composed of two DBDs: the tetracycline repressor, which recognizing the tetracycline response element of the transgenic HeLa cell line, and LexA, which recognize the LexA operator sequence inserted in the SB transposon (Ivics Z et al. 2007). In this experiment, 2 integrations out of 400 in total were identified around the TRE-eGFP locus. The authors also used another tethering protein consisting in a fusion of LexA and a SAF-box domain, which binds specifically to scaffold/matrix attachments regions (Kipp M et al. 2000). These sequences are involved in the modulation of chromatin structure and localization within the nucleus, allowing a regulation of gene expression by the formation of chromatin loops that become accessible to the transcription machinery. The use of a LexA-SAF-box fusion protein was shown to increase by 4-fold the frequency of SB transposon insertions near scaffold/matrix attachments regions (Ivics Z et al. 2007).

The use of DBDs fusion proteins that tether the integration complex is actually not enough efficient for therapeutic applications and non-targeted integration still occur at a high frequency. In one hand, the binding of the endogenous LEDGF/p75 to HIV-1 IN will compete to those of the DBD-

LEDGF/p75; and in the other hand, the natural DNA-binding of SB transposon will compete to the tethering activity of DBDs fusion protein. Several studies are again needed to overcome these off-target issues before considering the use of these strategies on gene therapy.

2.4 - TARGETED INTEGRATION

The development of targeting vectors that can integrate the DNA sequence of interest at a precise and unique location in the human genome is the ultimate goal in gene therapy. In gene targeting, a DNA fragment introduced into cells is able to replace a portion of endogenous chromosomal DNA through homologous recombination (HR). For example, mouse ES cells can be transfected with partially homologous template DNA in order to produce specific genomic alterations (Doetschman T et al. 1987). This has contributed greatly to basic biological research as it enables the production of adult mice carrying specific genomic alterations (Capecchi MR 2001). Gene targeting is a highly attractive approach to gene therapy as it offers the potential to insert therapeutic DNA at a known, "safe" location, or even to correct disease-causing mutations *in situ*. However, gene targeting in mammalian cells is extremely inefficient, with just 1 in 10^6 mouse ES cells carrying the desired insertion and 100 to 1000-fold more carrying background integrations elsewhere in the genome.

It was shown first in yeast (Kostriken R et al. 1983) and later in mammalian cells (Rouet P et al. 1994) that the efficiency of gene targeting could be enhanced by the introduction of a double strand DNA break (DSB) at the target site using a site-specific endonuclease. The free ends at DSBs mark them as substrates for host DNA repair pathways, stimulating HR between the target and template DNA (reviewed in (Sung P and Klein H 2006)). Rouet et al reported that the expression of a targeting endonuclease increased the rate of gene targeting by 100- fold.

Several nucleases have been evaluated since and display different features.

2.4.1 - Meganucleases

Meganucleases are naturally-occurring site-specific nucleases. The first two meganucleases, HO (Kostriken R et al. 1983) and I-SceI (Jacquier A and Dujon B 1985) have been identified in yeast mobile genetic elements. HO is known to naturally induce mating type switching in yeast by cleavage at the nuclear MAT locus followed by recombination. I-SceI is encoded by a mitochondrial group I intron and participate of the spreading, also called homing, of this intron to intronless copy of the gene (Jacquier A and Dujon B 1985; Chevalier BS and Stoddard BL 2001). For both proteins, their recognition sequences are known to be 18 bp long (Nickoloff JA et al. 1986; Colleaux L et al. 1988). Since then, hundreds of meganucleases have been identified in eukaryotes, bacteria and archae (Chevalier BS and Stoddard BL 2001). A large number of meganucleases are encoded by mobile genetic element such as group I introns or inteins and several have been shown to participate in the homing of mobile elements, so that they have been named Homing endonucleases (HEs).

All HEs recognize long DNA cleavage sequence (> 12 bp). This length of recognition site is usually sufficient to be unique in mammalian cells. However, as with recombinases, none of these enzymes have useful target sites in human cells. Retargeting their substrate specificity by protein engineering was needed for their use in human gene therapy. The crystal structure of the I-CreI HE (Heath PJ et al. 1997), and the structure of the protein bound to its target (Jurica MS et al. 1998) were resolved in the late nineties. HEs are classified into four families of proteins: the His-Cys box family, the GIY-YIG

family, the HNH family and the LAGLIDADG family. This latter is the well characterized, with nine members which have been crystallized. These studies allow the identification of a conserved core structure, characterized by a αββαββα fold. Generally, two αββαββα folds are facing each other and contribute to the active center. LAGLIDADG HEs can be homodimeric such as I-CreI, targeting a palindromic or pseudopalindromic DNA sequence, or monomeric such a I-SceI, targeting non-palindromic sequences. Within each αββαββα fold, two relatively independent subdomains were identified.

The analysis of HEs crystal structures allowed the engineering by combinatorial approaches of several mutants with a define cleavage specificity (Steuer S et al. 2004; Chen Z and Zhao H 2005; Arnould S et al. 2006; Ashworth J et al. 2006; Rosen LE et al. 2006; Silva GH et al. 2006; Smith J et al. 2006; Eastberg JH et al. 2007; Eklund JL et al. 2007; Niu Y et al. 2008). The coexpression of two I-CreI variants yields a heterodimeric species with a specific cleavage target (Arnould S et al. 2006; Smith J et al. 2006). However, the co-production of two I-CreI variants may result in the formation of homodimers rather than expected heterodimers. A possible strategy for overcoming this issue involves the engineering of the protein interface between the two monomers to impair the formation of functional homodimers and thus favor heterodimer formation (Fajardo-Sanchez E et al. 2008). Another strategy involves the creation of single-chain molecules adapted from natural monomeric HEs (Epinat JC et al. 2003; Li H et al. 2009). The use of either I-CreI variants (Smith J et al. 2006) or I-CreI variants with improved specificity (Grizot S et al. 2009; Grizot S et al. 2010) has been recently shown to induce targeted recombination in human HEK 293 cells at the clinically relevant RAG1 loci, which is mutated in SCID-X1. The improved I-CreI variants could induce 3-6% of targeting RAG1 recombination in transfected cells (Grizot S et al. 2009; Grizot S et al. 2010). Other clinically relevant loci have been successfully targeted by engineered HEs, such as XPC gene, involved in the Xenodrema Pigmentosum disease (Arnould S et al. 2007; Grizot S et al. 2009; Grizot S et al. 2010). These frequencies should be sufficient for gene repair strategies when the corrected cells subsequently display a selective advantage. However, many diseases are not on this particular case and require much higher levels of efficacy.

Recent studies analyzing the potential impact of chromosomal context and epigenetics on the efficiency of meganucleases-mediated genome editing has shown that the chromatin accessibility modulates the efficacy of meganucleases (Daboussi F et al. 2012; Valton J et al. 2012). The efficiency of targeting mutagenesis (0.1% to 6%) was strongly correlated with the subsequent homologous gene targeting (<0.1% to 15%) (Daboussi F et al. 2012). The chromatin state of the cell thus appears to play a major role in the targeting efficiency. This was further confirmed by the identification of genes regulating gene targeting in a high-throughput approach (Delacote F et al. 2011). The authors showed that siRNAs directed against the ATF7IP gene, encoding a protein involved in chromatin remodeling, stimulated by 3- to 8-fold homologous gene targeting in various loci and cell types. These findings could explain the variation of homologous gene targeting efficiency observed in different cell types and cell cycle phases and open the way to the development of strategies improving homologous gene targeting.

Meganucleases-induced DSBs can also be repaired by NHEJ, an error-prone process that frequently results in micro-insertions or micro-deletions (INDELs) at the cleavage site (Liang F et al. 1998). The propensity of a cell to use either NHEJ of HR varies with the cell type (Paques F and Duchateau P

2007) and also with the cell cycle (Kadyk LC and Hartwell LH 1992; Takata M et al. 1998; Gasior SL et al. 2001; Rothkamm K et al. 2003). Although NHEJ repairing of meganuclease-induced DSB can be an issue to gene therapy protocols; it can be used *per se* to correct a mutated gene. This was evaluated in human myoblasts nucleofected with several engineered meganucleases targeting different sites of the dystrophin gene carrying an out-of-frame deletion (Rousseau J et al. 2011). Mutations in the dystrophin gene are involved in DMD. It was shown that the use of a meganuclease targeting the exon 50 of the dystrophin resulted in INDELs for which 44% of them would have permit the restoration of the dystrophin reading frame in DMD patients with deletions of exons 51, 51-53 or 51-60. Moreover, approximately 36% of the INDELs produced would have permit restoration of normal dystrophin reading frame in DMD patients with a deletion of exons 51-56.

The parameters that remain to be improved when using meganucleases in therapeutic applications are the level of specificity and the level of induced homologous recombination while minimizing the frequency of NHEJ DSB-mediated repair. Meganucleases are promising tools as they exhibit high sequence specificity, cleaving as few as 1 in 10^8-10^9 random DNA sequences (Gimble FS et al. 2003; Scalley-Kim M et al. 2007). However, several meganucleases have been shown to cause some genomic instability as a result of NHEJ-mediated DSB repair (Rouet P et al. 1994; Monnat RJ, Jr. et al. 1999; Allen C et al. 2003; Guirouilh-Barbat J et al. 2004; Weinstock DM et al. 2006). Further studies on the potential genotoxicity of meganucleases inducing DSBs have to be conducted in clinically relevant animal models. The use of homing endonucleases that induce nicks in target DNA instead of DSBs is currently under development (Niu Y et al. 2008; McConnell Smith A et al. 2009) and could reduce genomic instability associated with DSBs.

2.4.2 - Zinc-Finger nucleases

Zinc-finger nucleases (ZFNs) are versatile and effective targeting reagent, which have separate DNA-binding and DNA cleavage domains. The development of ZFNs was initiated by the observation that the natural restriction enzyme FokI has physically separable binding and cleavage domains (Li L et al. 1992), with no apparent sequence specificity for the cleavage domain. It was also shown that the cleavage site could be redefine by changing the site recognition domain (Kim YG and Chandrasegaran S 1994; Kim YG et al. 1996; Kim YG et al. 1998).

The crystal structure of the DNA binding domain from a zinc finger transcription factor (Pavletich NP and Pabo CO 1991) showed a relatively simple recognition motif in which a series of looped polypeptides, or “fingers”, contact three DNA bases per finger using three amino acids within each finger. The structure suggested that zinc finger transcription factors were in fact modular, and that the fingers could be interchanged to alter the DNA binding site specificity. It has been shown that endonuclease domains such as that of FokI can be fused to zinc finger DNA binding domains to construct ZFNs artificial endonucleases which can be targeted against an extremely large number of chromosomal sites (Kim YG et al. 1996).The FokI cleavage domain must dimerize to achieve DNA cleavage (Bitinaite J et al. 1998; Smith J et al. 2000), and the best way to induce this dimerization is the construction of two sets of fingers recognizing neighboring sequences, each fused to a monomeric FokI cleavage domain (Fig. I-10A). The optimum configuration involve the introduction of a short linker between the FokI cleavage domain and the zinc fingers with a spacer of 5 or 6 bp between the two ZF recognition sites that are in inverted orientation (Fig. I-10A). As meganucleases, ZFNs induce DSBs, which are then repaired either by NHEJ or HR pathways (Fig. I-10B).

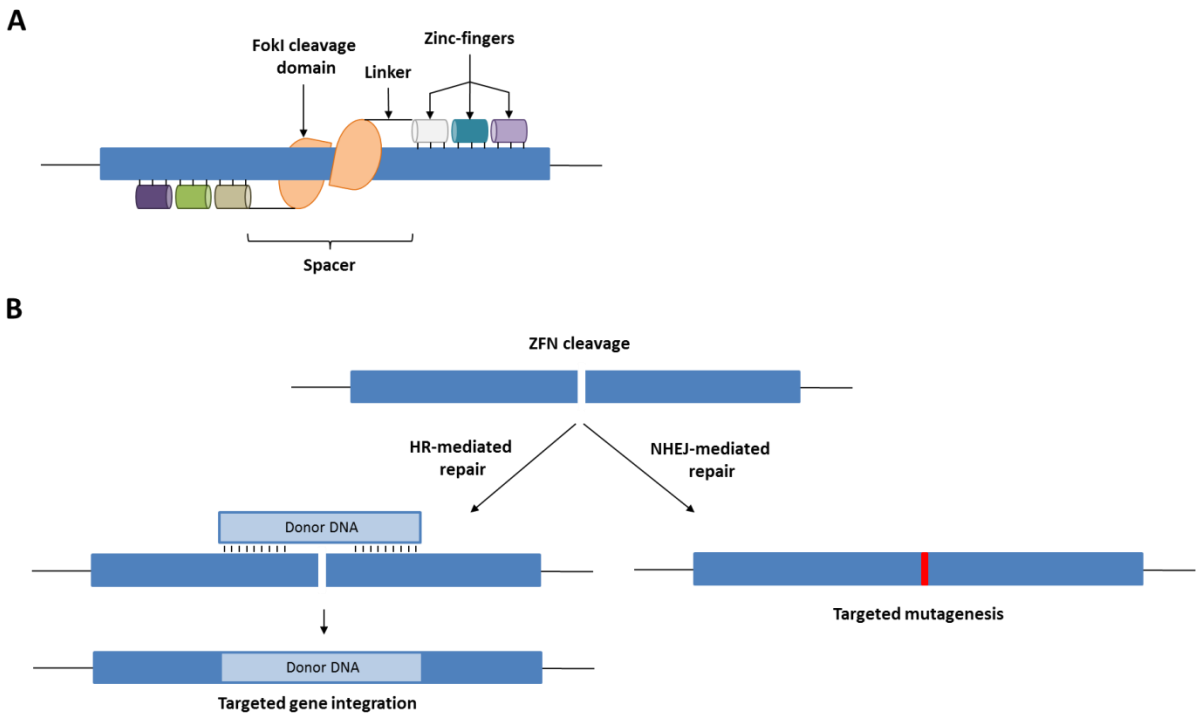


Figure I-10: Schematic representation of ZFNs bound to DNA and ZFNs cleavage repair pathways.

(***A***) Schematic representation of ZFNs DNA recognition. Zinc-fingers contacts with the DNA sequence are indicated thin lines. The spacer between the two zinc-finger binding sites are usually 5-6 bp. (***B***) After ZFNs DNA cleavage inducing a DSB, two repairing pathways can be used by the cell: Non-homologous end-joining (NHEJ) pathway, which usually incorporate mutations such as micro-deletions or micro-insertions and result to a targeted mutagenesis, and homologous recombination (HR) if a donor DNA containing sequences homologous to the DNA targeted site is added, resulting in an homologous targeted gene intergration.

The "modular assembly" approach stimulated the collection of libraries of fingers, each finger recognizing a different DNA triplet (Segal DJ et al. 1999; Dreier B et al. 2005). Construction of artificial zinc finger domains proceeds by assembly of fingers to give the desired binding specificity (18-24 bp target site) followed by *in vivo* selection of functional domains (Maeder ML et al. 2008). ZFNs are not thought to be more efficient than natural endonucleases at stimulating gene targeting, but their flexibility of specificity has raised the prospect of useful applications.

The first successful gene targeting using ZFNs in human cells pair was performed in HEK 293 cell line co-transfected with a ZFN expressing plasmid and a donor plasmid with an efficiency of approximately 1.8 gene targeting events per 10,000 cells (Porteus MH and Baltimore D 2003). Using ZFNs and template DNA delivered by plasmid transfection, Urnov and coworkers were also able to introduce small sequence changes into the endogenous IL2RG locus in 20% of K562 cells and 5% of primary human $CD4^+$ T lymphocytes (Urnov FD et al. 2005). It was later shown that up to 8 kb of heterologous sequence can be inserted into this locus in cell lines by flanking it with 750 bp sequences homologous to the sequence surrounding the DSB (Moehle EA et al. 2007). Gene addition of this type was shown at the CCR5 locus in 35% of human hematopoietic K-562 cells, 39% of Jurkat cells, and up to 0.11% of primary $CD34^+$ hematopoietic progenitor cells transduced with a ZFN-expressing integrase-defective lentiviral vector (Lombardo A et al. 2007). Human embryonic stem cells (hESCs) and human induced pluripotent stem cells (hiPSCs) were also genetically modified using ZFN

targeting different loci, such as OCT4 gene or PITX3 gene (Hockemeyer D et al. 2009). Although, it appear that HR-mediated gene targeting using ZFNs in hESCs and hiPSCs is poorly efficient (Zou J et al. 2009).

ZFNs can also be used to targeted mutagenesis of a gene of interest, as the DSB induced by ZFNs can be repaired by the error-prone NHEJ pathway (Fig. I-10B) (Bibikova M et al. 2002). This strategy has been used to knock out expression of CCR5, a major co-receptor for HIV-1 infection of T lymphocytes, in order to protect these cells from HIV infection (Perez EE et al. 2008). The authors used an adenoviral vector to transiently express the ZFN targeting the CCR5 locus in order to enhance the ZFN delivery in targeted cells. Gene disruption of CCR5 in human primary $CD4^+$ T cells was achieved with an efficiency of up to 50%. In February 2009, Sangamo Biosciences has begun a clinical trial using this system (http://www.sangamo.com/pipeline/index.html). The ZFN-mediated CCR5 gene disruption system is also in a pre-clinical phase for *ex vivo* anti-HIV therapy, using $C34^+$ hematopoietic progenitors as the target cells (Holt N et al. 2010), and the evaluation of ZFN-mediated CXCR4 gene disruption system is under study (Yuan J et al. 2012). Functional correction of the factor IX gene was also demonstrated *in vivo* in a humanized mouse model of hemophilia B using an intra-peritoneal (I.P.) injection of two adeno-associated virus vectors, serotype 8 (AAV8) expressing a ZFN targeting the factor IX gene and the DNA sequence to be integrated (Li H et al. 2011), resulting in the restoration of a plasma factor IX level of approximately 2-3% of normal.

One concern with the application of nucleases for gene addition or knockout is the cytotoxicity observed with intracellular endonuclease expression (Alwin S et al. 2005; Porteus MH 2006). In general, endonuclease toxicity is thought to be due to off-target DNA cleavage. Recent improvements have been made by redesigning the dimer interface to prevent unwanted homodimerization (Miller JC et al. 2007; Szczepek M et al. 2007; Sollu C et al. 2010). ZNF architecture modification is also being addressed to further enhance the efficiency and specificity of ZFNs (Doyon Y et al. 2011). Even though off-target cleavage has been much reduced by these latter techniques, it is difficult to anticipate the occurrence of adverse events, as very low frequency of deleterious events can be amplified by selection based on growth advantage. Development of highly sensitive methods for the detection of DSBs in ZFN expressing cells is currently under study ((Gabriel R et al. 2011; Pattanayak V et al. 2011) and reviewed in (Mussolino C and Cathomen T 2011)).

Another inherent problem with the use of nucleases inducing DSB is the repairing by the error-prone NHEJ pathway, at least in the case where an HR-mediated gene integration is wanted. Inducing new mutations at the target site in a high proportion of the cell (e.g. NHEJ is more efficient than HR) could be problematic. However, some cases may tolerate new mutations, as long as a sufficient number of cells are well corrected, as it could be the case of IL2RG targeting. The use of ZFNs technology for therapeutic applications has thus to involve extensive analyses of the frequency of off-target and NHEJ-mediated mutagenesis to avoid any potential deleterious genome alteration events.

2.4.3 - TALENs

The major hurdle of using either meganucleases or ZFNs in genomic targeting is the challenge of engineer new nucleases with wanted DNA binding specificities. Although some improvements have been made to facilitate the design, the techniques used require labor-intense selection and still empirical. An alternative class of engineered nuclease, which has several advantages over meganucleases and ZFNs, have been developed recently and is based on the identification of novel

DNA-binding proteins called Transcription activator-like effectors (TALEs) (Boch J et al. 2009; Moscou MJ and Bogdanove AJ 2009).

TALEs are proteins produced by *Xanthomonas* pathogens during host plant infection and delivered to the nucleus of the plant cells. The TALEs proteins act as transcription factors by binding to specific DNA sequences and activating gene expression. The central region of the protein is composed by tandem repeats of 34 amino acids sequences, named monomers, which are involve in the DNA recognition and binding (Kay S et al. 2007; Romer P et al. 2009) (Fig. I-11A). The sequence of each monomer is highly conserved, except in two positions at amino acids 12 and 13, named repeat variable diresidues (RVDs). The identity of these RVDs determines the nucleotide-binding specificity of each monomer, with one monomer binding one nucleotide of the DNA target (Boch J et al. 2009; Moscou MJ and Bogdanove AJ 2009). It was shown that a simple cipher specify the target base of each RVDs (NI target nucleotide A, HD target nucleotide C, NG target nucleotide T and NN target nucleotides G or A). The linear sequence of monomers in TALEs thus determines the target DNA sequence bound by TALEs. The only fixed position is a thymine at the beginning of the DNA target site, which is probably bound by a region of the protein within the nonrepetitive N-terminus domain. Each natural TALEs is thus able to specifically bind a DNA sequence of 35 bp long beginning with a T. This modular architecture of TALEs has been used to design site-specific nucleases by fusing the cleavage domain of FokI to the TALEs monomers, creating TALE nucleases (TALENs) (Christian M et al. 2010) (Fig. I-11B).

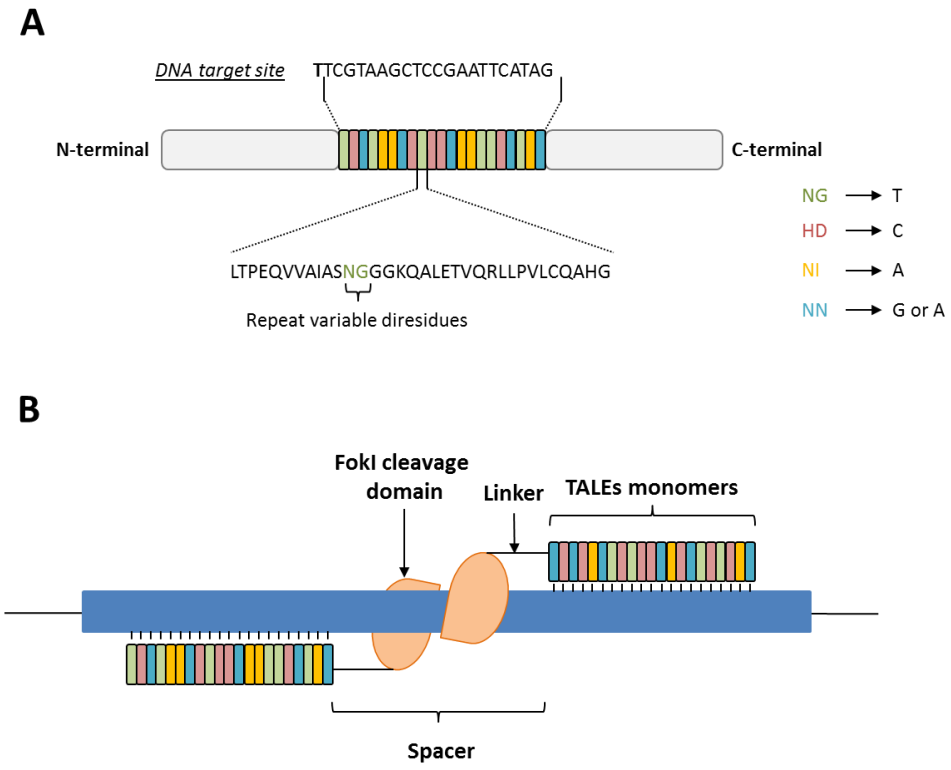


Figure I-11: Schematic representation of TALEs and TALENs

(***A***) Structure of natural TALEs from *Xanthomonas* pathogen. Each DNA binding monomer (colored rectangles) consists of 34 amino acids highly conserved except amino acid in position 12 and 13 which varies among TALEs monomers (repeat variable diresidus, RVDs). Each RVDs determines the bounded nucleotide (NG = T; HD = C; NI = A; NN = G or A). The amino acid sequence of the adjacent monomers specifies the sequence of

targeted DNA, which always begin with a T (in bold letter). (***B***) Schematic representation of engineered TALE nucleases (TALENs). TALEs monomers are fused to FokI cleavage domain using a linker. Each TALENs is engineered to recognize specific sequences separated by a spacer for which the length (from 12 to 40 bp) depends directly on the length of the linker (from 17 to >200 amino acids). The left TALEN binds the top strand of its DNA target site and the right TALEN binds the bottom strand of its DNA target site. The binding of each TALENs to their DNA target site induce a dimerization of FokI domain and a DSB at the DNA target site.

The fact that TALEs monomers target only one nucleotide makes them the most modular site-specific nucleases. This way, TALENs can theoretically target any DNA sequence. To date, several locus of the human genome have been successfully targeted by TALENs.

TALENs were used in human K-562 cell line to induce gene modification in CCR5 locus by HR-mediated repair pathway with a frequency of up to 16% (Miller JC et al. 2011). Another group has demonstrated TALENs mediated site specific genome modifications in hESCs and hiPSCs at OCT4, PITX3, and AAVS1 loci by HR-mediated repair (Hockemeyer D et al. 2011), but targeting efficiency in non-selected population was not determined. The DNA target sites chosen by this group have already been targeted by ZFNs in a previous study (Hockemeyer D et al. 2009), and the TALENs targeting efficiency was approximately the same than those of ZFNs, with also low frequency of off-targets. Another comparison between well-characterized ZFNs targeting CCR5 locus (used in the clinical trial initiated by Sangamo Biosciences; see Introduction section 2.4.2 -), ZFN targeting IL2RG locus, and engineered TALENs targeting the same DNA sites was also performed in HEK 293 cells (Mussolino C et al. 2011). In this study, the targeted efficiency of nucleases was assayed by analyzing the frequency of DSB repaired by error-prone NHEJ pathway. They showed that TALEN was about half as effective that ZFN on IL2RG locus, with frequencies of targeted genomic modifications of up to 14% for TALEN and 37% for ZFN, and slightly more efficient than ZFN on CCR5 locus, with frequencies of targeted genomic modifications of up to 17% for TALEN and 14% for ZFN. Moreover, off-target activity to the CCR2 locus was as low as 1% for TALEN while it was about 11% for ZFN. The cytotoxicity of TALEN was also shown to be lower than those of ZFN (~ 80% of cell survival) than with ZFN (~40% cell survival).

The construction of TALENs expressing constructs has very recently been improved by the development of new methods (Cermak T et al. 2011; Li T et al. 2011; Morbitzer R et al. 2011; Zhang F et al. 2011; Briggs AW et al. 2012; Sanjana NE et al. 2012), and guidelines for design of TALENs are also available (Doyle EL et al. 2012). TALENs have been showed to display less off-target activity than ZFNs (Mussolino C et al. 2011). A possible explanation to this difference is the existence of the conserved 5' T nucleotide in all natural TALEs DNA binding sites. It has been postulated that this 5' T nucleotide could pair with an unknown region of the N-terminal domain of TALEs and it was shown that this T position is critical for TALE-DNA interaction. Thus, this fixed and obligate 5' T nucleotide binding could represent a great impede to off-target TALENs binding and cleavage. It is also likely that the high length of TALENs target site (usually 34-38 bp) compared to that of ZFNs (18-24 bp) will also contribute to the higher specificity and thus lower toxicity of these new site-specific nucleases. However, further studies are still needed to fully assess the specificity and efficiency of TALENs, in particular in human primary cells and in animal models.

2.5 - COMBINATORIAL APPROACHES

An alternative approach to retargeting integration is to provide a given vector with an entirely new integration mechanism. A number of such hybrid vectors have been investigated which combine the gene transfer activity of one vector with the integration activity of another (Table I-12). This combination is particularly useful when either the gene transfer vector is normally non-integrating (Adenoviral vectors or AAV vectors) or when the delivery of the gene transfer vector system to target cells is poorly efficient (non-viral based delivery of transposases, recombinases, ZFNs or meganucleases). In this way, it is possible to generate vectors which combine desirable cell tropism and integration properties for particular gene therapy applications.

Gene transfer vector	**Integration mechanism**	**Integration profile**	**References**
Ad	AAV-Rep	Site-specific	(Recchia A et al. 1999; Recchia A et al. 2004; Goncalves MA et al. 2005; Goncalves MA et al. 2006; Wang H and Lieber A 2006; Goncalves MA et al. 2008)
Ad	SB transposase	random	(Yant SR et al. 2002; Hausl MA et al. 2010)
	ΦC31 integrase	Pseudo-random (into pseudo *attP* sites)	(Ehrhardt A et al. 2007)
	MoMLV intergrase	Semi-random (into active genes)	(Zheng C et al. 2000; Murphy SJ et al. 2002)
	Foamy virus integrase	Random	(Picard-Maureau M et al. 2004)
HSV-1	AAV-Rep	Site-specific	(Heister T et al. 2002; Wang Y et al. 2002; Liu Q et al. 2006)
HSV-1	SB transposase	Random	(Bowers WJ et al. 2006; de Silva S and Bowers WJ 2011)
	MoMLV integrase	Semi-random (into active genes)	(de Felipe P et al. 2001)
IDLV	ZFN	Site-specific	(Lombardo A et al. 2007; Cornu TI et al. 2008)

Gene transfer vector	Integration mechanism	Integration profile	References
	SB transposase	Random	(Staunstrup NH et al. 2009; Vink CA et al. 2009; Moldt B et al. 2011)
IDLV	Meganuclease	Site-specific	(Cornu TI and Cathomen T 2007)
AAV	Meganuclease	Site-specific	(Miller DG et al. 2003; Porteus MH et al. 2003)

Table I-12: Hybrid vector systems.
Ad: adenoviral vector; HSV-1: herpes simplex virus type 1 vector; IDLV: integrase-deficient lentiviral vector; AAV: adeno-associated viral vector; AAV-Rep: AAV Rep protein involved in the AAV site-specific integration into the AAVS1 locus; SB: Sleeping Beauty; MoMLV: Moloney murine leukemia virus.

AAV-Rep mediated integration

AAV is a single-stranded, non-enveloped DNA virus belonging to the family *Parvoviridae*. Unlike retroviruses, AAV does not require chromosomal integration for progression through the virus replication cycle. Perhaps as a consequence, the efficiency of AAV integration is considerably lower than that of retroviruses. Nonetheless, AAV can integrate into the host genome by a Rep-mediated mechanism. In the presence of AAV large replication protein Rep78 or Rep68, wild-type AAV integrates preferentially into a site on human chromosome 19 known as AAVS1 (Kotin RM et al. 1990).

The large Rep proteins are essential non-structural proteins with helicase and strand and sequence-specific endonuclease activities (Im DS and Muzyczka N 1990). During replication of the AAV DNA genome, Rep binds to a $(GAGC)_3$ Rep binding element present in the virus inverted terminal repeats and nicks the double-stranded DNA genome at a terminal resolution site. A Rep binding element is also present at the AAVS1 site on chromosome 19, and Rep is able to tether the AAV genome to the AAVS1 site (Weitzman MD et al. 1994). However, most intracellular wild-type AAV genomes do not integrate, and the efficiency of Rep-mediated integration is about 10-15%. In recombinant AAV vectors lacking Rep, integration occurs at a very low frequency and randomly throughout the genome (Miller DG et al. 2005).

The construction and production of recombinant AAV vectors containing Rep expressing cassettes and allowing transgene site-specific integration is impractical and uneasy because of the very limited packaging size of AAVs and also because of the deleterious effect that Rep is known to have on viral replication. To overcome these issues, hybrid vectors were designed using either Adenoviral vector or Herpes simplex virus (HSV)-1 vector as gene transfer vector.

Adenoviruses are double-stranded DNA non enveloped viruses with a virus genome from 26 to 45 kb. Recombinant adenoviral vectors are currently the most efficient vectors for gene transfer into a high number of cells. However, Ad viruses induce acute toxicity, inflammatory and cytotoxic immune response against viral proteins, and the Ad genome predominantly persists as an episomal state and is not integrated into host cell chromosome, thus impeding a long-term expression of the transgene in

mitotic cells. The last generation of recombinant Adenoviral vector is the Helper-dependent adenoviral vector (HD-Ad), which lacks all viral genes and only retain the adenoviral inverted repeat sequences and the packaging signal (Andrews JL et al. 2001). The production of HD-Ad requires a helper virus vector containing all the viral genes for replication and production and in which the packaging signal is excisable (Parks RJ et al. 1996). The absence of all viral genes in HD-Ad induces a high cloning capacity of up to 37 kb and significantly reduces the limitations associated with immune responses and limited transgene expression (Morral N et al. 1998; Morsy MA et al. 1998; Schiedner G et al. 1998; Thomas CE et al. 2000). HD-Ad was thus used to create hybrid vectors using the AAV-rep mediated integration system (Recchia A et al. 2004).

In 1999, Recchia *et al*. developed a two-vector HD-Ad system. The first HD-Ad vector carries the Rep78 gene under the control of either T7 (with the phage λ DNA for expression of the T7 RNA polymerase) or α-1-antitrypsin liver specific promoter, which both allows overcoming the negative effects of Rep expression during vector production. The other HD-Ad vector carries the transgene flanked by AAV inverted terminal repeats. They showed that up to 35% of the insertions occurred into the AAVS1 site in human hepatoma HepG2 cell line. The tropism of HD-Ad hybrid vectors was later broaden with the generation of fiber-modified Ad capsids (Goncalves MA et al. 2006; Wang H and Lieber A 2006; Goncalves MA et al. 2008). Indeed, several clinically relevant cell types, such as human hematopoietic stem and progenitor cells or mesenchymal stem cells and myoblasts are refractory to transduction by conventional HD-Ad vectors, due to the absence of the Ad (and Coxsackie B virus) CAR receptor on the cell surface of these cells (Shayakhmetov DM et al. 2000; Knaan-Shanzer S et al. 2001; Knaan-Shanzer S et al. 2005; Goncalves MA et al. 2006). Using fiber-modified Ad capsid, human muscle cells (Goncalves MA et al. 2006; Goncalves MA et al. 2008) and human hematopoietic cells (Wang H and Lieber A 2006) were successfully transduced and Rep-mediated site-specific integration of large transgenes (14 or 27 kb) into AAVS1 was demonstrated.

The HSV is also used to develop hybrid vectors. HSV is an enveloped virus which belongs to the family *herpesviridae*. Its genome is approximately 152 kb in size. Replication-defective HSV amplicon vectors based on HSV-1 serotype are usually used for gene delivery (Cuchet D et al. 2007). HSV-1 amplicons are helper-dependent vectors carrying a DNA plasmid containing the HSV-1 origins of replication, the packaging signal and the gene of interest (Spaete and Frenkel 1982). They have a high packaging capacity of up to 100 kb (Wade-Martins et al. 2001) and are able to transduce a large number of dividing and non-dividing cells, with a high tropism for neuronal cells (Sena-Esteves et al. 2000). As adenoviral vectors, they persist as episomes into transduced cells. They are also able to establish latency while maintaining some transcriptional activity. However, some hurdles has to be overcome, such as the silencing of most viral and non-viral promoters after injection of the vector in the brain (Suzuki M et al. 2006), which limit the duration of the transgene expression. The strategy consisting in using genetic elements from AAV that confer genetic stability by integrating the transgene into host chromosome is one of the approaches being developed.

Gene integration in AAVS1 site was also showed using HSV-1/AAV hybrid two-vector system (Heister T et al. 2002; Wang Y et al. 2002). The system was further improved by using a tetracycline-regulated Rep expression system to tightly control Rep expression and allowed the design of a one-vector system (Recchia A et al. 2004). Site-specific AAVS1 integrations were shown into human primary cells (0.1 to 16%) and transgenic mice (0.2% to 2% in mice liver hepatocytes). To further improve the production of Rep-containing hybrid vectors, Liu *et al*. placed the rep promoter upstream

of the transgene and the Rep coding sequence downstream of the transgene and the whole integrating cassette was flanked first with AAV inverted terminal repeats and then with LoxP sequences (Liu Q et al. 2006). Rep is thus not expressed during hybrid vector production, yielding titers as high as standard HSV-1 amplicons vectors. However, the expression of the Cre recombinase in transduced target cell induces a circularization to the LoxP sequences allowing the Rep coding sequence to become close to its promoter and thus expressed. The integration of the AAV inverted terminal repeat-cassette induces a termination of the Rep expression. The integration efficiency was about 20% in HEK 293 cells expressing Cre with 70% of correctly targeted integrations. However, this system can only be used in cells expressing the Cre recombinase, so that adaptations are needed to its use in common gene therapy applications.

Although the Rep-mediated hybrid vector systems have been shown to be very promising, several hurdles to their use in therapeutic applications still exist. The genotoxic risk induced by non-targeted integrations has to be carefully evaluated. Indeed, the mechanism of Rep-mediated integration frequently results in multiple insertions as well as significant and unpredictable rearrangements to both vector and host DNA (reviewed in (McCarty DM et al. 2004)). As with other proteins that induce double-strand beaks into the host genome, the risk of chromosomal rearrangements and genomic mutations is the major issue to be addressed. In addition, the poor efficiency of Rep-mediated integration, in particular in human primary cells, limits its use in clinical applications. Further improvements are thus required before considering the use in clinic of Rep-mediated site-specific integrations with hybrid vectors.

Hybrid vectors for efficient delivery of transposase/ recombinase/nucleases

Efficient delivery of transposase, recombinase or site-specific nucleases systems into target cells using naked DNA remains a tough challenge. In this context, the integration machineries based on transposase, recombinase, and site-specific nucleases have been used together with non-integrating virus vectors in order to obtain hybrid vectors combining the efficiency of virus vectors to transduce target cells as well as deliver genes into the host nucleus and the integration machinery of transposase/recombinase /nucleases.

The SB transposase was used in hybrid vector either to allow transgene integration in the case of hybrid vectors using non-integrating viral vector such as adenoviral vectors (Yant SR et al. 2002; Hausl MA et al. 2010) or HSV-1 viral vectors (Bowers WJ et al. 2006; de Silva S and Bowers WJ 2011), or to overcome the genotoxic effect of semi-random integration pattern into active genes of HIV-1-based lentiviral vectors (Staunstrup NH et al. 2009; Vink CA et al. 2009; Moldt B et al. 2011).

Yant *et al*. developed a two-vector system using the HD-ad gene delivery system (Yant SR et al. 2002). The first HD-Ad vector contains the SB transposon carrying the human factor IX gene flanked with Flp recombinase recognition target. These sequences are used to circularize the DNA because the SB transposase is more efficient on circular DNA than on linear DNA and the HD-Ad packaged DNA is predominantly linear. The second HD-Ad carries Flp and SB transposase coding sequences. In order to evaluate the possible long-term expression of the human factor IX, which is the missing protein in hemophilia B patients, systemic *in vivo* delivery of these two hybrid vectors was performed in mice. It showed efficient transduction of mouse hepatocytes (up to 45%) and stable integration of the transgene leading to the expression of human factor IX at therapeutic levels in mice undergoing rapid

liver cell cycling (Yant SR et al. 2002). The same system was used in a recent study and those results were confirmed with therapeutic level of factor IX of approximately 2-4% of normal in mice (Hausl MA et al. 2010). This last study also evaluated the system in a dog model form hemophilia B with the canine factor IX gene used as the transgene. The authors observed high levels of circulating canine factor IX in treated dogs one week post-injection followed by a drop to low but therapeutic levels after two weeks (Hausl MA et al. 2010). However, an adaptive immune response against the adenoviral vector capsid epitopes was observed by an increase level of anti-adenoviral neutralizing antibodies.

Bowers et al. used the SB transposon in a hybrid vector based on HSV-1 gene delivery to target neuronal cells and achieve transgene integration and long-term expression (Bowers WJ et al. 2006). The HSV-1 viral vector was also used because it induces only a mild inflammatory response (Olschowka JA et al. 2003). The system was based on two hybrid vectors, one carrying the SB transposase gene under the control of HSV immediate early 4/5 promoter, and the other carrying the SB transposon containing a β-galactosidase-neomycin resistant fusion gene under the control of the Rous sarcoma virus LTR promoter element. In contrast to the HD-Ad system, in which the packaged DNA is linear, the HSV-1 system allows the packaging of circular DNA. The efficiency of transposition in baby hamster kidney cultured cells was approximately 10 to 15% (Bowers WJ et al. 2006). The intracranially injection of the two hybrid vectors in mice embryos *in utero* resulted in the neuronal expression of the fusion β-galactosidase-neomycin protein in all brain sections of treated mice sacrificed 97 days post-transduction (Bowers WJ et al. 2006), suggesting that neuronal precursors cells have underwent SB transposition (reviewed in (de Silva S and Bowers WJ 2011)).

The hybrid lentiviral vector/SB transposon was used to both efficiently deliver the SB transposon system into the cells and overcome the biased integration pattern of lentiviral vectors to active transcription units. The lentiviral vector used is a HIV-1-based integrase-defective lentiviral vector (IDLV). The integrase activity of IDLVs is abolished through the introduction of the D64V mutation in the integrase (Vargas J, Jr. et al. 2004). After transduction with these IDLVs/SB hybrid vectors, SB transposition occurs from 1-LTR or 2-LTR circles, naturally formed. Staunstrup et al. designed a two hybrid vectors system, one carrying the SB100X transposase under the control of a phosphoglycerate kinase (PGK) promoter, and the other carrying the two SB transposon inverted repeats together with a puromycin reporter gene (Staunstrup NH et al. 2009). They showed that SB transposition from LTR circles was efficient in HEK 293 cell line; although the number of puromycin resistant clones was approximately 12-fold lower than with a conventional integrase-proficient LV. They also demonstrated that the integration profile using IDLV/SB hybrid vector was identical to those of the SB transposon, with a random integration pattern, as previously shown (Yant SR et al. 2005). This latter results was also confirmed using the less active SB11 transposase (Vink CA et al. 2009). Stable and random integrations were also observed in human primary fibroblasts and keratinocytes (Moldt B et al. 2011), but transposition efficiency was significantly lower than in human cells lines, with up to 0.03% of transposition in primary cells compared to up to 10% of transposition in cell lines.

The use of hybrid vectors using SB transposons is promising but the efficiency of transposition in human primary cells still needs to be improved.

The integration machinery of the ΦC31 integrase has been used to design an HD-Ad/ΦC31 hybrid vector, evaluated in mice (Ehrhardt A et al. 2007). The ΦC31-mediated integration of human factor IX gene in mice liver was demonstrated by injection of a two HD-Ad vector system followed by rapid cell cycling of mouse hepatocytes and resulted in the expression of human factor IX at therapeutic

level (12% of normal) until 122 days post-injection (Ehrhardt A et al. 2007). The authors also shown that the integration pattern of the transgene was not site specific, with only one insertion into a previously characterized ΦC31 hot spot (Olivares EC et al. 2002) on 40 analyzed integration sites. This suggests that the ΦC31-mediated integration mechanism is not as specific as it was thought to be.

Hybrid vector are also used to efficiently deliver site-specific nucleases such as ZFNs and meganucleases. Two examples of IDLV-based hybrid vectors are described.

Firstly, Lombardo et al. generated IDLVs able to integrate by homologous recombination a donor DNA consisting of the GFP gene flanked by IL2RG or CCR5 homologous regions. In this study, homologous recombination was stimulated by ZFN. Three IDLV were used, two for expression of each half of the ZFN dimer and one for the delivery of the donor DNA template and cells were analyzed by FACS about 15 days post-infection. When using a ZFN targeting the IL2RG locus, up to 6.3% of K562 cells infected by the three IDLV were GFP positive, while 1.1% of cells became GFP positive in absence of the ZFN (Lombardo A et al. 2007). When the authors used a ZFN targeting the CCR5 locus, gene addition occurred in up to 44% of K562 cells with a 2% background integration rate and in up to 0.06% of $CD34^+$ hematopoietic progenitor cells with a 0.005% background integration rate (Lombardo A et al. 2007).

Secondly, Cornu et al. reported a lentiviral vector able to undergo gene targeting by homologous recombination stimulated by the meganuclease I-SceI (Cornu TI and Cathomen T 2007). In this study, the I-SceI expression cassette and the homologous repair template were cloned into two separate IDLVs and co-transduced into target cells previously transduced with either retroviral vector or lentiviral vector encoding the eGFP gene. Gene conversion at a chromosomal eGFP target was observed in approximately 1% of HEK 293 cells transduced with both IDLVs and 0.03% of cells transduced with the template IDLV only.

The major limitation in these approaches is thus the efficiency of the HR-mediated repair of DSBs induced by the nucleases, as discussed in Introduction section 2.4 -.

2.6 - SUMMARY AND ALTERNATIVE STRATEGY

The different approaches actually used and developed to integrate a transgene into a host chromosome can be split into two major groups. The first group is composed of integration systems that cannot target the integration at a unique site of the genome, such as retroviral vectors, DNA transposons and recombinases. The major limitation of their use is the risk of insertional mutagenesis. The DNA transposon system is thought to be less genotoxic than retroviral vectors because of their random integration pattern. Still, the insertional mutagenesis risk is not overcome, as approximately 35% of DNA transposons-mediated integrations occur in gene (Vigdal TJ et al. 2002). This limitation leads researchers to develop other integration systems, which compose the second group of approaches. Engineered site-specific nucleases which can induce homologous recombination by creating a double-strand break at a specific location of the genome are thus evaluated, as well as AAV-Rep-mediated site-specific integration. Although being attractive and promising, these two strategies also imply some adverse events that have to be faced out. The occurrence of off-target integrations is obviously a major limitation, even though much effort has been made to increase the specificity of the nucleases to their target site. Moreover, all site-specific integration systems currently developed involve the creation of a double-strand break (DSB) into the genome, which can be repaired by the error-prone

non-homologous end joining (NHEJ) pathway. DSB are thus mutagenic and can also induce chromosomal recombination.

The safety concerns related to the use of both of these approaches lead to the development of new tool that could target integration with a radically different mechanism. In this context, the use of group II introns could represent an attractive alternative.

3 - GROUP II INTRONS

3.1 - GENERAL INTRODUCTION

Introns constitute the DNA regions in a gene that are excised from precursor mRNA (pre-mRNA) by the process of splicing. Introns can be divided into four major classes: spliceosomal introns, tRNA introns, group I introns, and group II introns. Additional classes have also been reported in the literature (i.e. group III introns in *Euglena gracillis*, tRNA-like introns in archae) (reviewed in (Michel F and Ferat JL 1995) and (Abelson J et al. 1998)). Each class of intron is characterized by a unique biochemical mechanism for splicing. It appears that the actual splicing reaction for most introns may be catalyzed by an RNA component rather than a protein enzyme.

Some classes of introns such as group I, group II and tRNA introns can be folded into characteristic structures. Group I and group II introns were initially found in yeast mitochondrial genome and were classified into two separate groups, based on sequence and/or secondary structural characteristics (Michel F et al. 1982). Several introns in other organisms were subsequently identified using RNA structural conservation pattern. These two groups appear to be unrelated, even though some similarities exist between them.

Some members of group I and group II introns have been shown to self-splice *in vitro* under specific conditions in the absence of any proteins or RNAs. The catalytic function required for splicing resides in the RNA molecule itself and thus these catalytic RNA molecules have been called "ribozymes". Group I introns were among the first ribozymes to be identified and Thomas Cech was awarded the Nobel Prize in chemistry for the discovery in 1989 (Kruger K et al. 1982). Other naturally occurring ribozymes such as group II introns were subsequently identified. The self-splicing observed for members of both of them occurs via two transesterification steps. However, the splicing of group I introns differ from those of spliceosomal and group II introns. The initiation of the first splicing step is carried out by an external guanidine nucleotide (exoG) that acts as the nucleophile in the first transesterification step (reviewed in (Woodson SA 2005) and (Stahley MR and Strobel SA 2006)). The reaction of splicing liberate a linear form of the group I intron, which sometimes can be circularized in a secondary reaction. In contrast, the splicing of spliceosomal and group II introns involves an internal nucleotide (bulged adenosine) that acts as the nucleophile in the first transesterification step.

In some cases, these group I and group II introns contain an open-reading frame (ORF). The splicing of group I and group II introns *in vivo* have been shown to involve these intron-encoded proteins (IEP). The IEPs are capable of assisting in the splicing of the intron in which they are found and have also additional functions in the intron mobility.

Indeed, group I (Jacquier A and Dujon B 1985; Dujon B 1989) and group II introns have spread among different species by a mobility mechanism called homing. Group I intron homing is initiated by the recognition of the target sequence (the junction of the two exons in an intronless genome) by its IEP, called homing endonuclease (HE) or meganuclease (See Introduction section 2.4.1 -). HE then creates a double-strand break, subsequently repaired by homologous recombination using the intron-containing copy of the gene as a template. The mechanism of group II intron homing is quite different: the DNA target site, which naturally corresponds to the junction of the two exons in an intronless

genome, is recognized mainly by base-pairing with the intron RNA. The intron RNA can subsequently reverse splice into the sense strand at the junction of the two exons. The IEP then cleaves the antisense strand downstream of the junction and reverse transcribe the intron RNA. A double strand cDNA copy of the intron is then integrated by DNA repair mechanism. These site-specific homing mechanisms have been adapted to practical applications, as described with the engineering of meganucleases for targeted gene integration (See Introduction section 2.4.1 -), and more recently with the development of retargeted group II introns as gene targeting vectors.

In the next chapters, the different characteristics of group II introns such as their structure, folding, splicing, and mobility mechanisms are described, followed by a description of the *Pylaiella littoralis* Pl.LSU/2 group II intron studied in this work.

3.2 - STRUCTURE AND FOLDING OF GROUP II INTRONS

3.2.1 - Intron RNA structure

The secondary structure of group II intron was first determined based on two intron sequences from yeast mitochondria (Michel F et al. 1982). This model was later supported by comparative analysis when more genome sequences became available and also by biochemical studies, with however some minor modifications (Michel F et al. 1989; Kwakman JH et al. 1990; Chanfreau G and Jacquier A 1994; Michel F and Ferat JL 1995). Group II intron secondary structure consists of six helical domains (I to VI) radiating from a central wheel (Fig. I-13) and bringing the 5' and 3' splice sites in close proximity (Michel F and Ferat JL 1995; Qin PZ and Pyle AM 1998). These domains of group II introns have specific roles in folding, conformational rearrangements, and/or catalysis.

It exist several interactions between the different group II intron domains (Fig. I-13: indicated by Greek letters) that allow the formation of a conserved tertiary structure, juxtaposing distant sequences to form an active site. Strikingly, even though these RNA elements show a high conservation in structure features and organization, they have only few conserved primary sequences. The few strictly conserved sequences are the consensus at the 5' (…↓GUGYG…) and 3' (…AY↓…) splice site, some nucleotide on the linker region (between introns domains), a large part of the domain V, some regions of domain I and the bulged adenosine of the domain VI.

Even though all group II introns fold into a similar overall secondary structure, they can be divided into three major subclasses, IIA, IIB and IIC by correlating specific secondary structural features (Toor N et al. 2001; Toro N 2003; Simon DM et al. 2008). Elements of the IIA and IIB classes are almost twice the size (~ 800 nt, excluding the IEP ORF) of those from the IIC class (~ 450 nt), which are presumed to be more ancient (Toor N et al. 2001; Toro N 2003). The subgroups IIA and IIB introns were later subdivided in subfamilies (IIA1 and IIA2; IIB1 and IIB2). The characteristics of all group II intron domains are presented thereafter, with the description of structural specificities associated with the different subclasses.

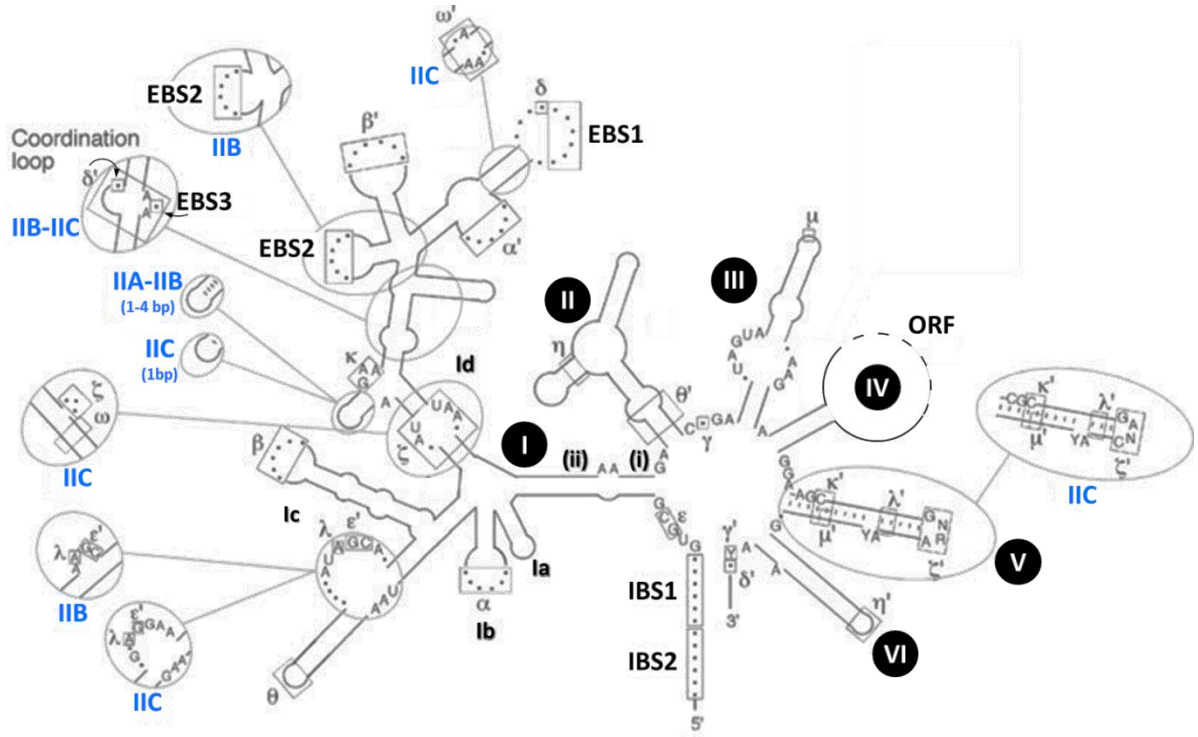


Figure I-13: Representation of a group IIA intron RNA secondary structure.

Intron domains I to VI are indicated, as well as DI subdomains I(i), I(ii), Ia, Ib, Ic, and Id. Notable variations in IIB and IIC introns (noted in blue) are indicated in circles. Boxes indicate sequences involved in tertiary interactions (Greek letters, EBS, IBS). The structure of DIV is not represented here. The open reading frame encoding the IEP is indicated by dotted line in DIV. *Adapted from (Lambowitz AM and Zimmerly S 2011).*

- Domain I is the largest domain of the RNA structure divided into four subdomains (Ia-Id) and serves as a scaffold for the assembly of other domains into a catalytically active tertiary structure. Indeed, it was shown that the domain I is the first to fold, followed by sequential folding of the other domains (reviewed in (Pyle AM et al. 2007)). Thus, the folding of DI appears to be the rate-limiting step of the overall intron folding (Su LJ et al. 2005). DI is essential for catalysis (Michel F and Ferat JL 1995) and exon recognition, which explains its necessity for both splicing and mobility.

DI is held together through the α-α' Watson-Crick base-pairing interaction, identified by phylogenetic analyses (Jacquier A and Michel F 1987; Michel F et al. 1989), and shown to be functionally important for self-splicing *in vitro* (Harris-Kerr CL et al. 1993). An additional β-β' pairing takes part in the preorganization of the intron structure to the active form (Toor N et al. 2001; Simon DM et al. 2008). In group IIC introns, a ω-ω' interaction is also involved in the folding of DI. DI folds independently of the other domains (Fedorova O and Zingler N 2007). It has been demonstrated that a small substructure/region in domain Id is crucial for compaction and folding. This region, designated as "folding control element", is composed of κ and ζ elements involved in the κ-κ' and ζ-ζ' interactions important for the docking of the domains V (Costa M and Michel F 1995; Boudvillain M and Pyle AM 1998) (Keating KS et al. 2008). Indeed, as mentioned above, DI is the scaffold for all other domains, with DV in the middle and the other domain structures stacking upon each other (Dai L et al. 2008; Toor N et al. 2008a; Toor N et al. 2008b; Michel F et al. 2009; Toor N et al. 2009). Other motifs that are essential in the formation of the active site are the ε-ε' and λ-λ' interactions that place the 5' splice site near the DV and are involved in catalysis either directly or in

positioning the conserved first intron nucleotide (G) to promote the nucleophilic attack (Jacquier A and Michel F 1990; Boudvillain M et al. 2000; de Lencastre A et al. 2005; de Lencastre A and Pyle AM 2008). The X-ray crystal structure of the *Oceanobacillus iheyensis* IIC intron has revealed that the ε' and λ are components of a functional substructure in subdomain Ic, called the z-anchor, that makes multiple contacts with the subdomain domain I(i) forming a binding interface for domain V and nucleotides at the 5'-end of the intron, thereby mediating the structural integrity of the core (Toor N et al. 2008a). They are also a strong binding site for divalent metal ion such as Mg^{2+}, and thus are suggested to contribute to the intron structure stabilization. In addition, the θ-θ' tetraloop-receptor interaction with DII is thought to be involved in the stabilization of the intron native structure (Costa M et al. 1997a), and also in the recruiting of DIII and the linker J2/3 into the active site (Podar M et al. 1998b).

In class IIB and IIC introns, DI also contains an internal asymmetric loop in subdomain Id, which is referred to as the coordination loop. It has been proposed that this loop supports the docking of the branch-point of DVI and all other components essential for splicing (de Lencastre A et al. 2005; Hamill S and Pyle AM 2006). Indeed, the coordination loop contains EBS3 involved in the EBS3/IBS3 interaction, and the δ' nucleotide involved in the δ-δ' base-pairing with the δ base located upstream of EBS1 (sequences involved in the intron/exons interactions; see following section). These studies have thus postulated that all reaction components are aligned in close proximity in a single active site prior to splicing and that the configuration of the core is maintained throughout the whole splicing process. The docking of DVI and its static nature is however questioned (Michel F et al. 2009). It was suggested that DVI could be positioned between the Ic subdomain and the coordination loops (Pyle AM 2010). The lack of information on DVI in the crystal structure of *O. iheyensis* group II intron also supports the dynamic nature of DVI (Toor N et al. 2010). In addition, a recent study proposed a different receptor for the docking of DVI than the one previously suggested (coordination loop; (Hamill S and Pyle AM 2006)), which is located in subdomain Ic (Li CF et al. 2011), also supporting the idea of a major translocation of DVI between the two splicing steps of the branching pathway.

- Domain II is mainly involve in tertiary interactions within the intron and in conformational changes of the intron during splicing (Fedorova O et al. 2003; Fedorova O and Pyle AM 2005). In addition to the θ-θ' interaction with DI, DII harbors the η-η' tetraloop-receptor interaction with DVI (Costa M et al. 1997a). The η-η' interaction is suggested to induce a conformational change of the intron between the two transesterification steps that moves the DVI out of the overall intron structure (Chanfreau G and Jacquier A 1996; Costa M et al. 1997a).

- Domain III has been shown to enhance the catalytic efficiency, but is not absolutely essential for catalysis (Koch JL et al. 1992; Qin PZ and Pyle AM 1998). DIII has been shown to interact with DV via the μ-μ' interaction (Fedorova O and Pyle AM 2005).

- The intron domain IV is the most varying region in secondary RNA structure, and can contain the ORF encoding a multifunctional protein called IEP (intron-encoded protein). This structure does not directly contribute to catalysis, but when present it influences both splicing and mobility (Fedorova O and Zingler N 2007).

- Domain V is one of the smallest domains (usually 34 nt) and represents the main catalytic center of group II introns. Almost every nucleotide in DV has a major role in the intron's function and this domain is the most phylogenetically conserved primary sequence of the entire intron (Michel F and Ferat JL 1995; Fedorova O and Zingler N 2007). At the 5'-end of DV is located the catalytic triad AGC (or CGC) which forms together with a highly conserved dinucleotide bulge at the 3'-end of DV a negatively charged pocket that binds two coordinating metal ions (usually Mg^{2+}) required for catalysis (Sigel RK et al. 2000; Zhang L and Doudna JA 2002; Sigel RK et al. 2004; Toor N et al. 2008a; Toor N et al. 2008b). The dinucleotide bulge forms with the J2/3 linker and the DV catalytic triad a triple helix, bringing together the catalytic essential residues of the intron (Toor N et al. 2008a). DV and DI form together the catalytic core and are the only elements that are absolutely required for minimal catalytic activity of the intron (Koch JL et al. 1992; Michels WJ, Jr. and Pyle AM 1995). The aforementioned κ-κ', μ-μ', λ-λ', and ζ-ζ' interactions involve sequences of DV.

- Domain VI is necessary for the splicing via the branching pathway, as it contains the bulged adenosine that serves as the branch point. It was shown that the flipping of the bulged adenosine is the crucial feature that enables the intron to splice via branching rather than hydrolysis (Chu VT et al. 1998). Indeed, the first step of group II introns splicing can be initiated by either the bulged adenosine, leading to the formation of a branched intron lariat (See Introduction section 3.3.1 -), or by external H_2O molecule or hydroxyl (OH^-), leading to linear spliced intron (See Introduction section 3.3.2 -). Mutations in either the bulged adenosine or in surrounding G/U wooble pairs impede the branching pathway and the intron splicing occurs via the hydrolytic pathway.

- Lastly, the linker regions between the different domains have been shown to have an important role in several aspects of intron folding and catalysis (de Lencastre A and Pyle AM 2008).

During evolution, group II introns have developed different modes of target site recognition, especially IIC intron group.

3.2.2 - Intron/exons boundaries

Two types of tertiary interactions exist for group II introns: the ones between intron and exon sequences and the ones between two intron sequences. These interactions are involved in both forward and reverse splicing reactions and in the overall tertiary structure of the intron.

Group II introns RNAs indeed recognize their targets sites, either RNA or DNA, via specific base-pairing with the exon sequences (Fig. I-14).

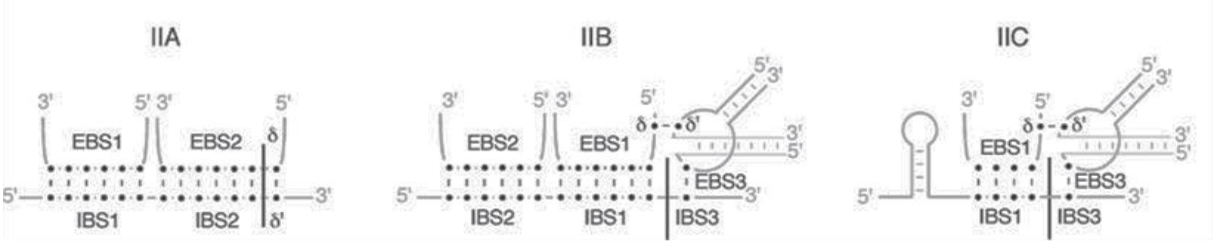


Figure I-14: Base-pairing interactions used by IIA, IIB and IIC introns with the exons at the target site.
EBS: exon-binding site; IBS: intron-binding site. The junction between 5' and 3' exons is represented by the black line. *Taken from (Lambowitz AM and Zimmerly S 2011).*

For group IIA and IIB introns, the 5' exon is defined through two interactions with the intron domain I. The exon-binding sites (EBS) 1 and 2 of the intron (See Fig. I-13) pair to their corresponding intron-binding sites (IBS) 1 and 2 located in the 3'-end of the 5' exon, thus forming a 12-15 bp interaction with the 5' exon (Fig. I-14) (Qin PZ and Pyle AM 1998; Boudvillain M et al. 2000; Costa M et al. 2000). The resulting two recognition duplexes mediate the high-interaction specificity and cleavage-site fidelity, giving the proper conformation of the 5' splice site for transesterification or hydrolytic cleavage (Jacquier A and Michel F 1987). The affinity of binding between the 5' exon and intron DI via EBS/IBS interactions was shown to be strongest when the intron is completely folded and the catalytic core correctly folded (Costa M and Michel F 1999). It has been suggested that a premature exon binding could prevent or delay the correct folding of the intron into its active state. Class IIC introns differ from IIA and IIB introns with respect to 5' exon definition, as they preferentially insert downstream of a transcriptional terminator stem-loop structure which is thought to substitute in part for the missing IBS2/EBS2 interaction (Fig. I-14) (Toor N et al. 2006; Fedorova O and Zingler N 2007; Robart AR et al. 2007).

The 3' exon is defined by two single-base-pair interactions, which also vary between the RNA structural classes. The first is the γ-γ' interaction, which involve a nucleotide located in the J2/3 linker region between domains II and III (γ) and the last intron base (γ') (See Fig. I-13). The second interaction is either δ-δ' for class IIA introns or EBS3/IBS3 for IIB and IIC introns where δ /EBS3 are in different locations in domain I and δ'/IBS3 is the first base of the 3' exon (See Fig. I-13) (Jacquier A and Michel F 1990; Costa M et al. 2000). Class IIB and IIC introns also have δ-δ' interactions, but with a different δ' nucleotide positioned in the coordination loop of domain I and involved in a different aspect of exon recognition. The γ-γ' and EBS-IBS3 interactions seem to play a minor role in the splice site recognition for IIB introns, as disruption of these interactions affect mainly the efficiency of the second splicing step but not the fidelity of 3' splice site selection (Costa M et al. 2000). In contrast, IIA introns appear to be somewhat more sensitive to substitutions in γ-γ' nucleotides, which can lead to the use of cryptic 3' splice sites. Furthermore, domain VI is thought to guide the 3' intron-exon junction into the catalytic active site in a passive way, ensuring an efficient second splicing step with high fidelity (Jacquier A and Jacquesson-Breuleux N 1991).

Class IIA and IIB introns form a continuous binding interface to 5' and 3' exons with EBS1 and the δ nucleotide that respectively binds IBS1 and the first 3' exon nucleotide δ' (Fig. I-14) (Jacquier A and Jacquesson-Breuleux N 1991). The main difference is that the exons are largely internalized for the class IIB introns as opposed to class IIA introns, for which they are mostly bound to the surface of the ribozyme (Dai L et al. 2008). In 2008, the X-ray crystal structure of the *O. iheyensis* class IIC intron in a post-catalytic state with ligated exon substrate has revealed that the exon junction is presented as a continuous strand over the important active sites in domain V (Toor N et al. 2008a; Toor N et al. 2008b; Toor N et al. 2009; Toor N et al. 2010). This study also confirms that EBS1 and EBS3 motifs are linked together in a common exon binding interface by the δ- δ' interaction (Fig. I-14). The crystal structure of the *O. iheyensis* group IIC intron in a pre-catalytic state has been published very recently (Chan RT et al. 2012). It was observed that the overall structure of this pre-catalytic intron was quite similar to those of the post-catalytic intron previously determined (Toor N et al. 2008a), suggesting that no drastic conformational changes of DI to DV occur during the splicing. This structure together with the previous one allows the authors to build a theoretical model of the splicing pathway (Chan RT et al. 2012). The next section describes the different splicing pathways used by group II introns.

3.3 - SPLICING MECHANISM

Group II intron splicing is catalyzed by the intron RNA molecule itself. This reaction requires the correct folding of the RNA molecule into its catalytically active secondary and tertiary structures, forming the active site described above which binds catalytically essential Mg^{2+} ions.

Some group II introns have been naturally split into segments by genomic rearrangements and can be separated in two or more distinct transcribed segments (Glanz and Kuck 2009). However, these segments can reassociate via tertiary interactions between group II intron domains and *trans*-splice to produce a functional mRNA. This *trans*-splicing mechanism is observed in several species, and particularly in plant mitochondria and chloroplasts (Goldschmidt-Clermont et al. 1991; Bonen 1993; Knoop et al. 1997; Qiu and Palmer 2004; Bonen 2008)

Three different mechanisms of splicing have been described so far, leading to different forms of spliced intron molecules (Fig. I-15).

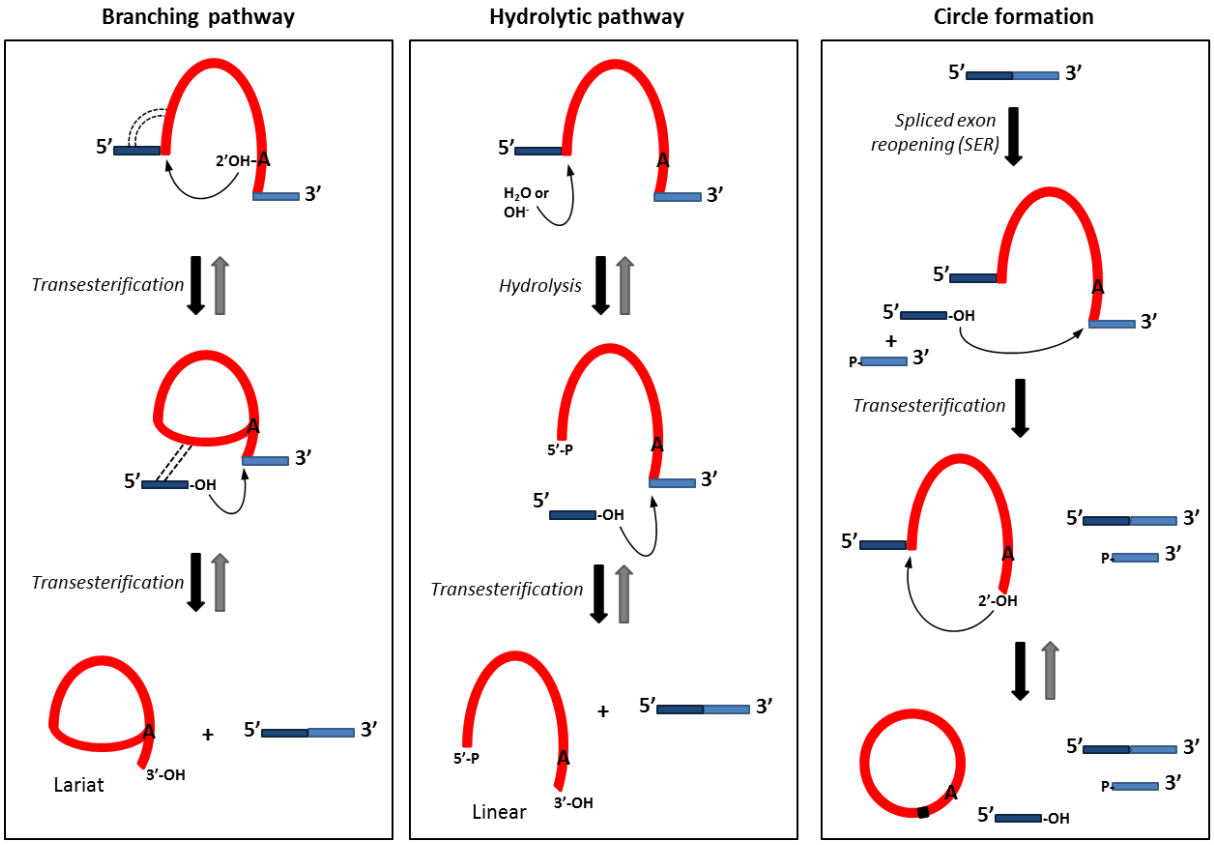


Figure I-15: Schematic representation of group II introns splicing reactions.

Red line: group II intron; Dark blue line: 5' exon; light blue line: 3' exon. The bulged adenosine residue and the water molecule acting as nucleophiles are represented with their 2' hydroxyl groups that initiate the first step of branching and hydrolysis splicing, respectively. Base-pairing between the 5' exon and the intron are indicated by dotted lines. The minor reaction leading to the formation of intron circle is also represented. This reaction is thought to result from the spliced exons reopening (SER). Nucleophilic attacks are indicated by thin black arrows. Thick solid black and gray arrows indicate forward and reverse direction of the corresponding reaction, respectively. Details of each reaction are in the text. *Adapted from (Lehmann K and Schmidt U 2003).*

3.3.1 - Branching pathway

Ribozyme activity was the first property assigned to group II introns (van der Veen R et al. 1986; Peebles CL et al. 1987). A major pathway by which group II introns excise themselves from the pre-mRNA is the branching pathway. Branching splicing occurs by two transesterification reactions as a two-step process (Fig. I-15; left panel) (reviewed in (Lehmann K and Schmidt U 2003)). The 5' splice site is put by several interactions in close proximity to the 2'-OH group of a specific bulged adenosine in domain VI (Fig. I-15; left panel, 2'OH-A), which performs a nucleophilic attack and breaks the phosphodiester bond of the 5' junction in a classical SN2 displacement mechanism (Padgett RA et al. 1994; Podar M et al. 1995). It results the formation of a 2'-5' bond between the first nucleotide of the intron and the bulged adenosine, and the release of the 5' exon. The 3'-end of the intron remains covalently bound to the 3' exon, forming an intron-exon splicing intermediate (van der Veen R et al. 1986; Lehmann K and Schmidt U 2003). After the first cleavage reaction, the 5' exon is still tightly linked to the intron via base-pairing interactions (Fig. I-15; dotted lines) (Jacquier A and Michel F 1987; Jacquier A and Jacquesson-Breuleux N 1991). This way, its 3'-OH group is correctly positioned to attack the 3' splice site in the second transesterification reaction. This leads to the release of a free intron lariat and ligated exons. This second step also proceeds via a SN2 displacement mechanism, but a phosphate substitution at the two splice sites has revealed inverted stereoisomeric preferences (Padgett RA et al. 1994; Podar M et al. 1995). Group II introns are dependent on divalent metal ions for folding and catalysis and have a two-metal ion coordination for the leaving groups at the catalytic center (Piccirilli JA 2008; Toor N et al. 2008a).

Both of the two transesterification reactions are reversible (Fig. I-15). The first step is the splicing rate-limiting for most self-splicing group II introns (Daniels DL et al. 1996). The rate constant of this first step is equal in the forward and reverse direction (Chin K and Pyle AM 1995). The intermediates are usually not detected as the second forward reaction is much faster than the reverse and thus drives the forward reaction to completion. The reverse splicing reaction is considerably slower, although under suitable reaction conditions, reverse splicing can be quite efficient (Muller MW et al. 1991; Aizawa Y et al. 2003). Reverse splicing of group II intron is not limited to RNA substrates as they can also reverse splice into DNA molecules and thus provides the basis for intron mobility ((Zimmerly S et al. 1995a; Yang J et al. 1996; Cousineau B et al. 1998); reviewed in (Lambowitz AM and Zimmerly S 2004)).

3.3.2 - Hydrolytic pathway

In addition to the lariat splicing pathway, where the nucleophile is internal, group II introns splicing can be performed via a hydrolytic pathway where the nucleophile attacking the 5'exon-intron junction in the first step is water or a hydroxyl ion (Lehmann K and Schmidt U 2003). This hydrolysis step releases the 5' exon and a linear intron attached to the 3' exon (Fig. I-15; middle panel). The second step is identical to that of the branching pathway and the final products are ligated exons and a linear intron. *In vitro*, the balance between the branching and hydrolysis reactions is strongly influenced by the choice of monovalent cation used (Daniels DL et al. 1996). This balance may also differ depending on the subclass of group II intron. Some introns have been shown to splice *in vitro* only through the hydrolytic pathway (Granlund M et al. 2001), and *in vivo* the hydrolytic reaction is an active pathway for introns lacking the branch-point nucleotide (Podar M et al. 1998a; Vogel J and Borner T 2002). It

has recently been shown that a linear intron can catalyze efficient reverse splicing, suggesting an alternative pathway for mobility (Roitzsch M and Pyle AM 2009).

3.3.3 - Circle formation

Both lariat and linear introns have been shown to be involved in the spliced exons reopening *in vitro* reaction (Jarrell KA et al. 1988; Daniels DL et al. 1996). This alternative reaction, shown for some group II introns, consists in the hydrolysis of the 5'-3' exon junction after recognition by excised intron molecule (Lehmann K and Schmidt U 2003; Fedorova O and Zingler N 2007). This reaction, while leaving the intron unchanged, occurs with same stereo chemistry as splicing and the cleavage is performed at the exact junction position (Lehmann K and Schmidt U 2003; Michel F et al. 2009). This spliced exon reopening (SER) reaction is thought to be involved in the generation of intron circles. A fully circular intron form, first discovered as a by-product of *in vitro* splicing, has been shown *in vivo* in bacteria and plant mitochondria (Murray HL et al. 2001; Li-Pook-Than J and Bonen L 2006; Molina-Sanchez MD et al. 2006). In the circularization pathway, a free 5' exon is suggested to attack the 3' splice site of an unspliced precursor mRNA, leaving a 5' exon still covalently linked to the intron. The 2'-OH group of the intron 3'-end subsequently attacks the 5' splice site, releasing the 5' exon and a circular intron with a 2'-5' linkage at the circle junction (Murray HL et al. 2001).

3.4 - INTRON-ENCODED PROTEINS

3.4.1 - Description of IEPs

Almost all bacterial group II introns and about half in chloroplasts and mitochondria contain an open-reading frame usually located in the loop of domain IV. These intron-encoded proteins, called IEP, have generally several conserved domains, and are involved in the splicing of their corresponding intron *in vivo* (See Introduction section 3.4.3 -) as well as in their mobility (See Introduction section 3.5 -). To date, the best characterized IEP, named LtrA, is encoded by the *Lactococcus lactis* Ll.LtrB group II intron (Matsuura M et al. 1997).

In general, group II intron-encoded proteins contain four major conserved domains: RT (reverse transcriptase), X (maturase), D (DNA-binding), and En (DNA endonuclease) (Fig. I-16). Each domain shows different level of phylogenetical and functional conservation.

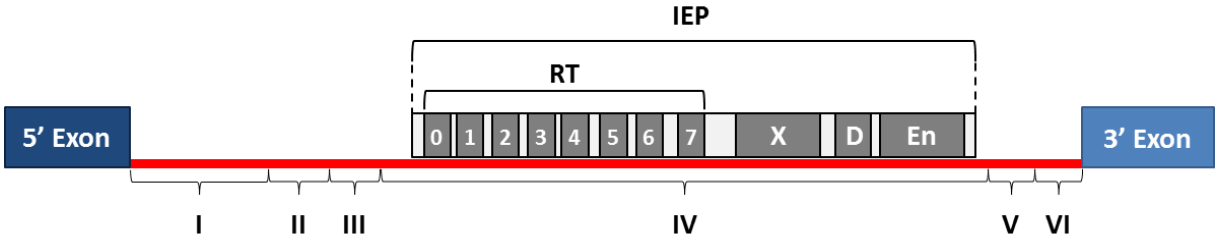


Figure I-16: Schematic representation of intron-encoded protein conserved domains.

The 5' and 3' exons are indicated in dark and light blue rectangles, respectively. The intron is represented by a red line, with intron domains I to VI. The intron ORF, located in the domain IV, is characterized by four different conserved domains: reverse transcriptase (RT), containing the RT blocks 0-7; maturase (X), DNA-binding (D) and endonuclease (En) domains.

At the N-terminal of IEP is located the typical RT domain, divided into eight blocks (RT0 and RT1-7). RT blocks 1-7 are common segments to all retroelements and correspond to the palm and finger structure of HIV-1 RT (Kohlstaedt LA et al. 1992). The RT5 block contains the highly conserved YNDD motif (where N can be any amino acid), identified as part of the RT active site (Xiong Y and Eickbush TH 1990; Steitz TA et al. 1993), where X is more frequently an alanine, but can also be another amino acid. The RT0 block can be considered as an N-terminal extension of the RT domain and this subdomain is conserved among RTs of non-LTR retrotransposons (Malik HS et al. 1999). In contrast to retroviral RTs, group II introns and non-LTR retrotransposons RTs contain some sequences between the different RT blocks which can present structural conserved features potentially important for the RT function (Malik HS et al. 1999; Blocker FJ et al. 2005). The first biochemical evidence of an RT activity of group II intron IEP has been demonstrated for the mitochondrial aI1 and aI2 of *Saccharomyces cerevisiae* (Kennell JC et al. 1993). Subsequently, RT activity of IEPs from bacterial group II introns such as Ll.LtrB (Matsuura M et al. 1997), RmInt1 from *Sinorhizobium meliloti* (Martínez-Abarca F et al. 1999), or G.st.I1 from *Geobacillus stearothermophilus* (Vellore J et al. 2004; Moretz SE and Lampson BC 2010) has been demonstrated.

The domain X, usually referred as maturase domain, is involved in the splicing activity of group II introns. Indeed, mutations in this domain affect the splicing activity of its corresponding intron *in vivo* (Moran JV et al. 1994). Although this domain is poorly conserved in sequence, it is present in all known IEPs. It is characterized by three predicted α-helices, which are structures found in the thumb domain of retroviral RTs (Blocker FJ et al. 2005). Together with the RT domain, the X domain participates in the binding of the intron RNA and promotes the folding of the intron into its catalytically active structure (Saldanha R et al. 1999; Wank H et al. 1999; Cui X et al. 2004).

Domain D was functionally defined for Ll.LtrB group II intron. Although this domain is not conserved in sequence among introns ORF, it is characterized by two pairs of cysteines residues that fit the consensus of a class of Zinc finger DNA binding motif. Several studies have shown that these conserved cysteine pairs seem to maintain the structure of the DNA endonuclease region (San Filippo J and Lambowitz AM 2002). Domain D and domain En participate in the binding of the IEP to the target DNA during the mobility of the intron.

The En domain is a DNA-dependent endonuclease domain of the HNH family (Shub DA et al. 1994) which functions with Mg^{2+} to cleave the target DNA antisense strand during intron mobility to prime the reverse transcription (Zimmerly S et al. 1995a; Guo H et al. 1997; Singh NN and Lambowitz AM 2001; San Filippo J and Lambowitz AM 2002).

While all IEPs characterized are shown to contain the conserved X domain, several IEP sequences have derived and lost some conserved sequences in the RT and En domains. Some group II intron IEPs do not contain the required sequences for RT activity, but conserved their function in RNA splicing. This can be represented by the loss of some RT blocks such as the chloroplast MatK IEP of *Nicotiana tabacum* (presence of RT5-7 blocks) (Mohr G et al. 1993) or the mitochondrial MatR IEP of *Arabidopsis Thaliana* (presence of RT6 block). This type of proteins with degenerated RT domains is rarely found in bacterial group II introns, suggesting a higher mobility than organellar group II introns. The En domain is frequently absent in group II introns. It was shown that RmInt1, which encodes an IEP lacking the En domain, uses a different mechanism to prime the reverse transcription (See Introduction section 3.5.1 -).

A small group of fungal mitochondrial group II introns are shown to encode IEP belonging to the LAGLIDADG DNA endonucleases, often encoded by group I introns (See Introduction section 2.4.1 -). It was recently shown that the LAGLIDADG IEP from a functional group II intron located in the rns gene of *Leptographium truncatum* was able to create a DSB at the splice site, but does not enhance the intron splicing *in vitro* and does not appear to bind the intron RNA precursor transcript (Mullineux ST et al. 2010). The evolutionary relationships between this IEP and the group II intron remain an open question.

The IEPs can be either translated from their start codon located within the intron domain IV such as Ll.LtrB IEP, or they can be translated in frame with the 5' exon, as IEP of aI1 and aI2 from *S. cerevisiae COX1* gene. These latter chimeric IEPs are only found among mitochondrial introns (Zimmerly S et al. 2001; Dai L et al. 2003). The analyses of the proteins have however revealed that these chimeric proteins are rapidly processed by proteolytic cleavage (Moran JV et al. 1994; Zimmerly S et al. 1999), probably by a protease encoded in the nucleus and addressed to mitochondria (Van Dyck E et al. 1995; Arlt H et al. 1998).

3.4.2 - IEP Lineages

Phylogenetic analysis of IEP ORFs identified from the growing number of intron sequences has shown that group II introns can be further subdivided into nine classes: mitochondrial-like (ML), chloroplast-like 1 (CL1), chloroplast-like 2 (CL2), and bacterial A, B, C, D, E1/E2, F classes (Fig. I-17). The CL1 and CL2 classes are not monophyletic, but split into four clades (CL1A, CL1B, CL2A and CL2B). Introns from all phylogenetic classes can be found in bacteria, while only classes ML and CL are found in organelles. Each IEP lineage is associated with a distinct RNA secondary structure: ML with IIA, CL1 with IIB1, CL2 with IIB2, bacterial class C with IIC, while bacterial class A, B, D, E and F and associated with different IIB structures (Fig. I-17) (Michel F et al. 1989; Simon DM et al. 2008; Simon DM et al. 2009) (Toor N et al. 2001; Zimmerly S et al. 2001; Toro N 2003).

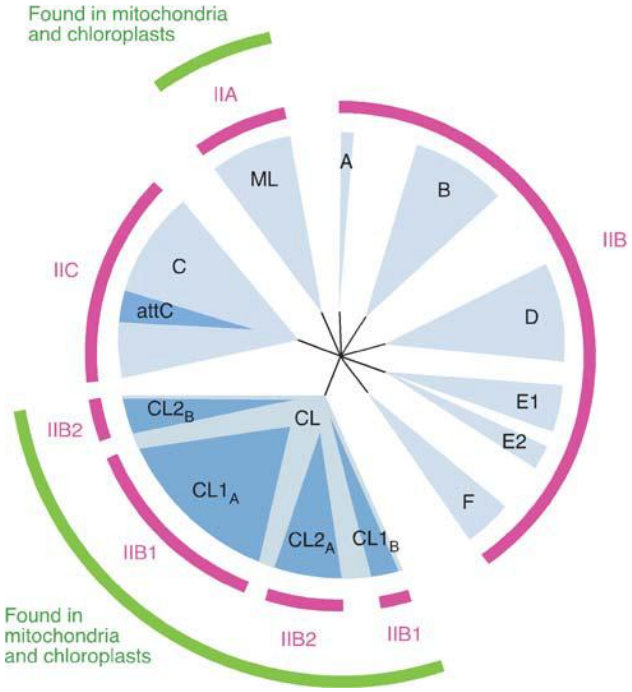


Figure I-17: Group II intron IEP ORF lineages.

The lineages of group II intron IEP are indicated by blue sectors (mitochondrial-like [ML], chloroplast-like 1 [CL1], chloroplast-like 2 [CL2], and bacterial A, B, C, D, E1/E2, F) (Simon DM et al. 2009). Sublineages (CL1A, CL1B, CL2A, CL2C, and bacterial C for which the intron inserts after *attC* sites) are shown in dark blue sectors. Corresponding RNA structural subgroup are indicated in purple. IEP lineage found in mitochondria and chloroplast are indicated by green arcs. *Figure taken from (Lambowitz AM and Zimmerly S 2011).*

The different subclasses can be found in a mix of host organisms, although some subclasses seem to be somewhat restricted to particular bacterial phylogenetic groups. Single species can harbor introns of several subclasses clearly showing that the introns are mobile elements or that there has been a lot of horizontal transfers of introns (Dai L and Zimmerly S 2002b; Robart AR and Zimmerly S 2005; Simon DM et al. 2009). Interestingly, comparison of RNA secondary structures between the different subclasses indicates that the catalytic RNA has specific features that are unique to each group and strongly suggests that RNA structure has coevolved with the sequences of the IEP (Toor N et al. 2001; Simon DM et al. 2009). The coevolution is suggested to be due to the strong biochemical interactions that exist between the IEP and the catalytic RNA. This relies on the fact that both the protein and ribozyme RNA are required for the splicing reaction and the mobility event of group II introns *in vivo*.

3.4.3 - IEP-mediated splicing

Most studies of the mechanism of group II intron folding and catalysis *in vitro* are conducted in relatively extreme reaction settings, with high salt and Mg^{2+} concentrations (> 50 mM) and elevated temperature (> 40°C) to reach optimal conditions, compared to the physiological conditions in the cell (Peebles CL et al. 1986; van der Veen R et al. 1986; Jarrell KA et al. 1988; Matsuura M et al. 1997). This is necessary to ensure high enough splicing reactivity. Under near-physiological conditions, intron folding is very slow and the structure is unstable (Fedorova O et al. 2007; Fedorova O and Zingler N 2007). Therefore, most or all group II introns probably require protein factors to stabilize active structure and/or resolve misfolded intermediates to allow efficient splicing *in vivo* (Lehmann K and Schmidt U 2003; Fedorova O et al. 2007). The best-characterized protein factors that participate in

intron splicing *in vivo* are the intron-encoded proteins. The first proof of maturase function of the IEP was determined by the genetic analyses of intron mutants aI1 and aI2 of *COX1* intron from *S. cerevisiae* (Carignani G et al. 1983; Moran JV et al. 1994). This dependence of the *in vivo* splicing reaction from the IEP was also demonstrated for the bacterial Ll.LtrB intron, where deletions or missense mutations into the LtrA ORF lead to a complete block of the intron splicing *in vivo*. In addition, it was shown that the LtrA-mediated *in vitro* splicing was ATP-independent and can be achieved at low Mg^{2+} concentrations (5 mM) (Matsuura M et al. 1997; Saldanha R et al. 1999). At those ionic conditions, the Ll.LtrB intron is unable to fold alone into its catalytically active structure. The IEP has thus a "chaperone" activity on the intron RNA and promote its correct folding into its catalytically active structure *in vivo* (Fig. I-18).

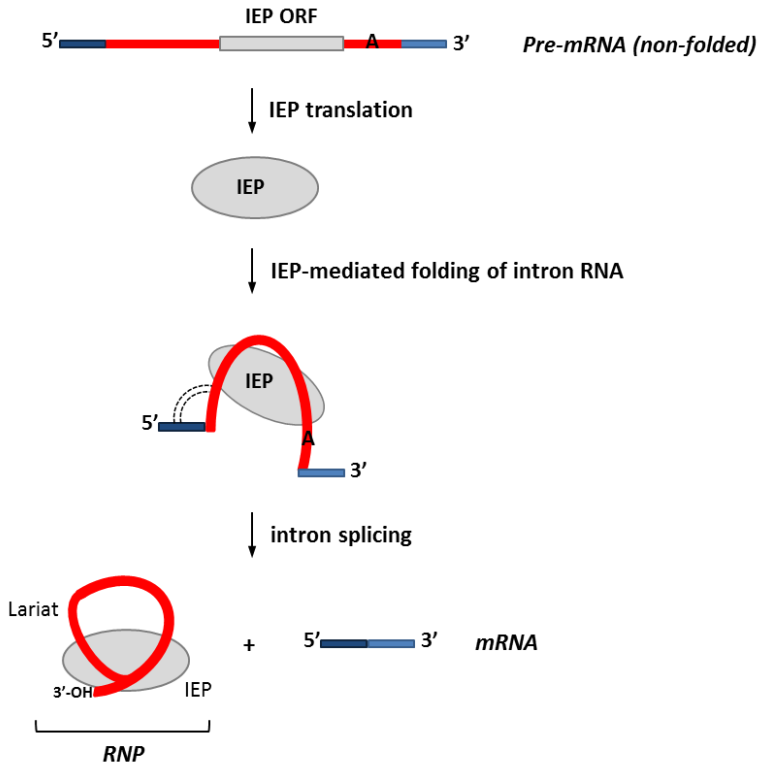


Figure I-18: IEP-dependent intron splicing *in vivo*.

The ionic strength *in vivo* induces a very slow folding of group II introns (red line), which could accumulate as misfolded intermediates. The major splicing factor *in vivo* is the IEP (gray ellipse), which binds the intron and promotes its folding into the catalytically active structure, leading to the intron splicing and the ligation of the exons (dark and light blue rectangles). This raises to the formation of a ribonucleoparticle (RNP) composed of the intron lariat and the IEP.

The LtrA protein was shown to bind tightly and specifically to several part of intron RNA, leading to a stabilization of its active structure (Matsuura M et al. 2001; Noah JW and Lambowitz AM 2003). The main contact between LtrA and the intron RNA is done at an idiosyncratic stem-loop structure located at the beginning of Ll.LtrB intron domain IV and was called DIVa (Wank H et al. 1999) (Fig. I-19).

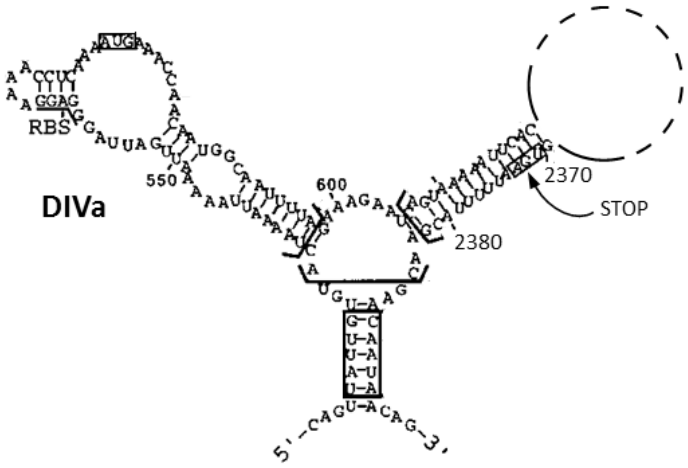


Figure I-19: Representation of the Ll.LtrB DIV secondary structure.

Initiation (AUG) and termination (STOP) codons of LtrA translation are boxed. RBS: putative Shine Dalgarno sequence. The stem-loop structure of DIVa is represented. The rest of the DIV sequence is depicted by dotted line. *Adapted from (Wank H et al. 1999).*

This binding involves the recognition of specific bases in the terminal loop and helical bulges (Watanabe K and Lambowitz AM 2004). Additional contacts are made with the conserved core regions that further stabilize the active RNA structure (Wank H et al. 1999; Matsuura M et al. 2001; Dai L et al. 2008). The deletion of DIVa does not impede the maturase activity of the protein, suggesting that LtrA can bind directly to the core regions. It was postulated that the binding of LtrA to DIVa represents an auto-regulation system, as Ll.LtrB DIVa contains the Shine-Dalgarno sequence and the start codon of the LtrA. However, it was also shown that the aI2 IEP, which is translated as a chimeric protein with exons 1 and 2 of the *COX1* gene in *S. cerevisiae*, also binds both DIVa and catalytic core regions (Huang HR et al. 2003), suggesting a conserved interaction mode rather than a specific auto-regulation. The aI2 intron is also able to splice *in vivo* in an IEP-dependent manner in absence of DIVa and without a drastic loss of efficiency (Huang HR et al. 2003). Other studies show that introns deleted of the binding site in DIV retain residual maturase-dependent splicing *in vitro* and *in vivo*, suggesting that the contacts to other regions are sufficient to promote the splicing even in absence of the DIVa primary binding site (Wank H et al. 1999; Matsuura M et al. 2001). The mapping of the binding sites between LtrA and its intron reveals several positions spanning from DI to DVI (Matsuura M et al. 2001; Dai L et al. 2008). The mapping of these sites onto a three dimensional model of the Ll.LtrB structure indicates that they form a large binding surface, extending from DIVa to contiguous DI, DII and DVI regions (Dai L et al. 2008). The mapping of the regions of LtrA required to the intron RNA binding onto a three dimensional model of the protein indicates that an RNA-binding surface extends from the RT to the X domain and also includes the N-terminal extension of the protein, which is thought to bind the intron DIVa region (Cui X et al. 2004; Blocker FJ et al. 2005; Gu SQ et al. 2010).

The IEP-dependent splicing of group II intron appears to be specific for some group II introns. For example, the LtrA protein is able to promote the splicing of its corresponding intron, Ll.LtrB, but not of other self-splicing introns such as the yeast aI2 and aI5γ as well as the *E. coli* IntB (Saldanha R et al. 1999). In contrast, other maturases in different plastid systems have evolved to become general group II intron splicing factors. The MatK protein of the *tnrKI* group II intron that represents the only

known putative maturase in chloroplasts of higher plants is thus able to bind and promote the splicing of multiple ORF-less chloroplastic introns IIA (Ems SC et al. 1995; Vogel J et al. 1999; Zoschke R et al. 2010). Another interesting example is provided by the nuclear encoded n-Mat proteins of flowering plants. Evolutionary progressions have brought these mitochondrial IEPs to loss mobility functions and splice multiple introns (Nakagawa N and Sakurai N 2006; Keren I et al. 2009).

3.4.4 - Nuclear-encoded accessory factors

Accessory factors involved in group II splicing are also encoded by nuclear genes. These factors may assist in splicing directly by binding to the group II intron and either stabilize the active structure or resolve misfolded intermediate structures (Huang HR et al. 2005; Kohler D et al. 2010). They also may have indirect effects. For example, the MW3 and ME4 genes encode mitochondrial carrier proteins which are suggested to suppress group II splicing defects by altering the ionic balance within the mitochondrion (Grivell LA 1995). In contrast, the yeast Mss116p protein, which is a member of the DEAD-box family of ATP-dependent RNA helicases, is thought to promote the splicing of mitochondrial introns (Seraphin B et al. 1989; Huang HR et al. 2005). Indeed, MS116 null mutants are defective in splicing of all four mitochondrial group II introns, as well as in the splicing of all mitochondrial group I introns and other RNA process. Interestingly, other DEAD-box proteins are shown to compensate for the Mss116p loss of function (Huang HR et al. 2005). It is now know that the DEAD-box proteins are able to unwind the RNA (Mohr S et al. 2006; Del Campo M et al. 2007; Halls C et al. 2007; Del Campo M et al. 2009), which is also a system used by proteins of the spliceosome to improve specific structuration during splicing.

3.5 - MOBILITY OF GROUP II INTRONS

Group II introns are mobile genetic elements that can insert into a specific DNA sequence, which corresponds to the junction of the two exons (from which they splice) in an intronless genome. This mobility occur by a mechanism, called retrohoming (or target-primed reverse transcription), and is mediated by the ribonucleoparticle (RNP) formed after the IEP-dependent intron splicing *in vivo* and consisting of the IEP bound to the intron lariat RNA (See Fig. I-18). Group II intron mobility occurs either at a specific target site with a high efficiency (retrohoming), but can also occur at ectopic sites with a low frequency (retrotransposition).

3.5.1 - Retrohoming

The first demonstration of the mobile nature of group II introns was obtained by genetic analyses of *S. cerevisiae* (Meunier B et al. 1990). During crossing between haploid strains, the authors observed that aI1 and aI2 mitochondrial group II introns have homed into intronless alleles at a frequency of about 90%, and that this homing was abolished by mutations of either the IEP or the intron RNA inhibiting intron splicing. Retrohoming is a multistep process, involving the reverse splicing of the intron RNA into the sense strand of the DNA target site, the cleavage of the antisense strand by the endonuclease activity of the IEP, the reverse transcription of the intron RNA by the reverse transcriptase activity of the IEP and the integration of double-stranded cDNA copy of the intron by DNA repair mechanisms (Fig. I-20). Retrohoming rely on the formation of the RNP after IEP-mediated intron splicing. The DNA target site for retrohoming, usually extending to 30-35 bp, is recognized by both the IEP and the intron RNA (See following section). The IEP promotes local unwinding of the DNA target site, so that the intron RNA can base pair with the target site (Singh NN and Lambowitz AM 2001; Aizawa Y et

al. 2003). The first step of retrohoming consists of the reverse splicing of the intron RNA into the sense strand of the DNA target site at the exact junction of the two exons (Zimmerly S et al. 1995a; Yang J et al. 1996). This mechanism of reverse splicing into double-stranded DNA involves the same EBS-IBS and δ-δ' interactions required for splicing and reverse splicing into RNA (Mohr G et al. 2000; Singh NN and Lambowitz AM 2001). Then, the endonuclease (En) domain cleaves the antisense strand 9-10 bases downstream the exons junction, generating a primer used for reverse transcription (RT) of the inserted intron RNA (Zimmerly S et al. 1995b). The resulting cDNA is finally integrated by cellular DNA repair mechanisms.

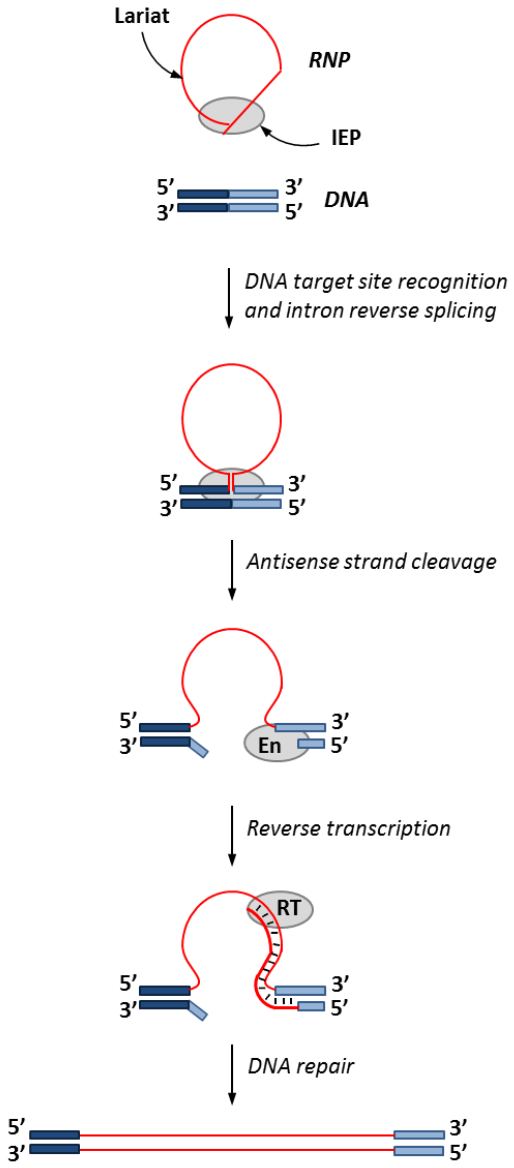


Figure I-20: General group II intron retrohoming mechanism.

The spliced intron RNA lariat (thin red line) forming with the IEP a RNP particle, which recognizes the DNA target site (Exons 5' in dark blue, and exon 3' in light blue).. The intron RNA is reverse spliced into the sense strand at the junction of the two exons. After cleavage of the antisense strand downstream of the exon junction (En activity), the IEP (gray ellipse) reverse transcribe the intron RNA into cDNA (thick red line) via its reverse transcription activity (RT) and a double-stranded cDNA copy of the intron is integrated by cellular DNA repair mechanisms. *Adapted from (Lambowitz AM and Zimmerly S 2004).*

Many IEPs in bacteria lack the En domain, and the corresponding introns use another mobility pathway that requires a primer provided by the DNA replication fork (Ichiyanagi K et al. 2003; Zhong J and Lambowitz AM 2003). The *S. meliloti* RmInt1, a bacterial IIB intron whose IEP lacks the En domain uses two endonuclease-independent retrohoming pathways. A major pathway occurs by reverse-splicing of the intron RNA into single-stranded DNA at the replication fork and the nascent lagging strand is used as a primer for reverse-transcription (Fig. I-21A) (Martinez-Abarca F et al. 2004). Another process involves retrohoming before the replication fork and reverse transcription is primed using the nascent leading strand (Fig. I-21B) (Martinez-Abarca F et al. 2004). For both endonuclease-independent mobility pathways, the intron recruits several host factors to complete the integration into the new genomic location (Beauregard A et al. 2008).

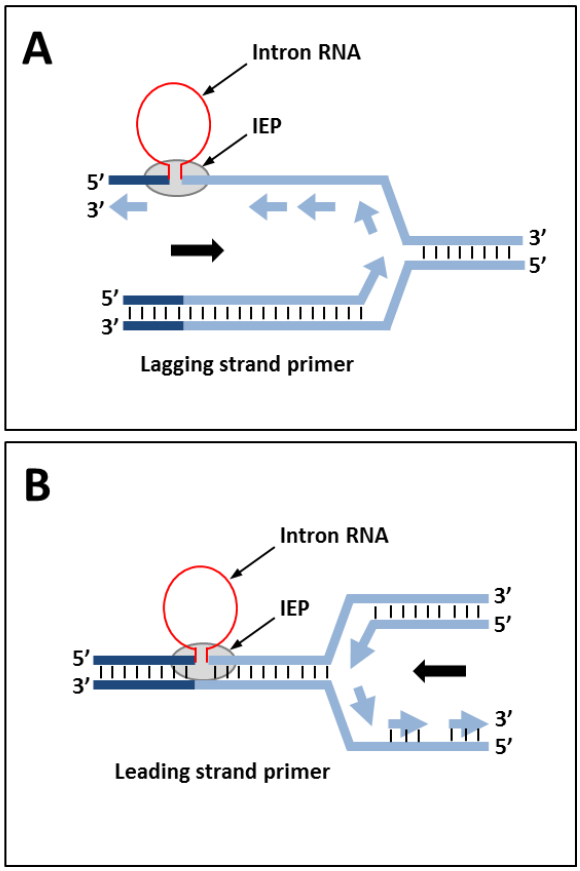


Figure I-21: Endonuclease-independent homing pathways.
The nascent strand in the DNA replication fork is used as primer for reverse transcriptase. In pathway (***A***) the intron reverse splices into single-stranded DNA at the replication using the lagging strand as a primer, while in pathway (***B***) the intron inserts into double-stranded DNA before passage of the replication fork. In the pathway (***B***) represented, the primer is leading strand, but the lagging strand can be used if the replication fork is in the opposite direction. These pathways are used for retrohoming by introns whose IEP lacks the En domain, or for retrotransposition into ectopic sites (See Introduction section 3.5.3 -). The black arrow indicates the direction of replication. The 5' exon and 3' exon are indicated in dark and light blue lines, respectively. *Adapted from (Lambowitz AM and Zimmerly S 2004).*

3.5.2 - DNA target site recognition

The DNA target site recognition during homing has been studied for the *L. lactis* Ll.LtrB (Guo H et al. 2000; Mohr G et al. 2000; Perutka J et al. 2004), the *S. cerevisiae* aI1 (Yang J et al. 1998) and aI2

(Guo H et al. 1997), the *S. meliloti* RmInt1 (Jimenez-Zurdo JI et al. 2003), and bacterial class C (Granlund M et al. 2001; Dai L and Zimmerly S 2002a) introns (Fig. I-22), and more recently for the *E. coli* EcI5 intron (Garcia-Rodriguez FM et al. 2011).

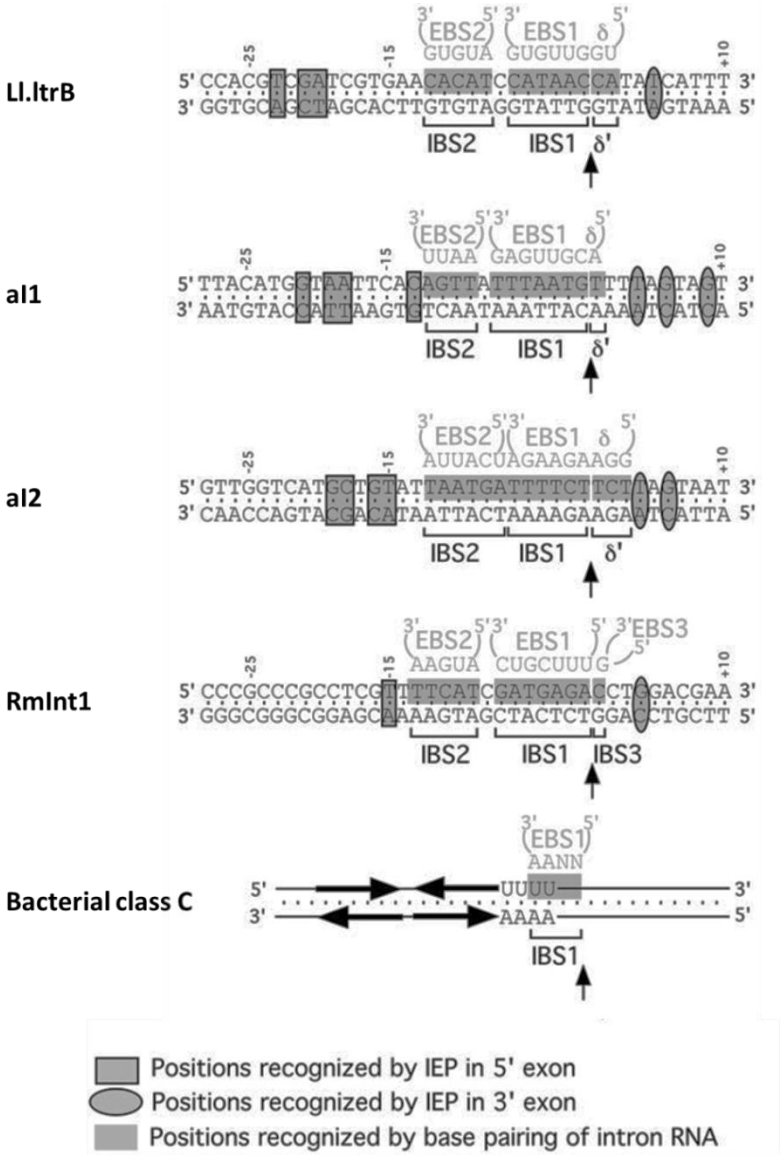


Figure I-22: DNA target site recognition.

IBS and δ' regions of the DNA target site recognized by base pairing with EBS and δ regions of the intron RNA are indicated for the *L. lactis* Ll.LtrB, the *S. cerevisiae* aI1 and aI2, the *S. meliloti* RmInt, and bacterial class C introns. In addition, the inverted repeat of Rho-independent transcription terminator, after which bacterial class C introns insert if followed by an IBS1, is also indicated. *Figure taken from (Lambowitz AM and Zimmerly S 2004).*

In each case, both components of group II intron RNPs recognize the DNA target site. The intron RNA pairs to the 5' and 3' exons via EBS/IBS and δ-δ' pairing previously described, and the IEP interacts with several downstream and upstream nucleotides (Guo H et al. 1997; Mohr G et al. 2000; Singh NN and Lambowitz AM 2001; Jimenez-Zurdo JI et al. 2003). The pairing between the intron RNA and the DNA target site appears to be crucial for the reverse splicing and homing as mutations in the DNA target impede those mechanism, which are rescued with compensatory mutations of the

intron RNA (Guo H et al. 1997). We can observe that a relatively limited number of positions are recognized by the IEP, and most of the target specificity rely on base pairing with the intron RNA. However, for Ll.LtrB, aI1 and aI2 introns, mutations of the nucleotides recognized by the IEP in the 5' exon inhibit both reverse splicing and antisense strand cleavage, while 3' exon mutation impede only the antisense strand cleavage. Detailed analyses of the mechanism of DNA recognition by Ll.LtrB have revealed that RNPs first bind nonspecifically to the DNA and scan the sequence until DNA target specific binding (Aizawa Y et al. 2003). Initial contacts are thus made by the IEP through interactions with nucleotides of the 5' exon at the sense strand, and these contacts promote a local unwinding of the DNA (Singh NN and Lambowitz AM 2001). This unwinding allows an efficient pairing of the intron RNA to the target site via EBS-IBS and δ-δ' and reverse splice (Singh NN and Lambowitz AM 2001). It has been showed that these primary contacts are crucial for unwinding, as mutations of critical bases in the 5' exon do not inhibit reverse splicing into single-stranded DNA target (Zhong J and Lambowitz AM 2003). Second strand cleavage requires additional contacts of the IEP to the 3' exon (Singh NN and Lambowitz AM 2001), with one critical base for Ll.LtrB target site.

3.5.3 - Retrotransposition

While retrohoming is the predominant mobility pathway, at a much lower frequency (typically 10^{-4}, 10^{-5}) group II introns are also able to invade noncognate (ectopic) sites through retrotransposition (Dickson L et al. 2001; Ichiyanagi K et al. 2003). The retrotransposition mobility events of Ll.LtrB in *L. lactis* follow the same mechanism as the main retrohoming pathway described above for RmInt1 with insert into single-stranded DNA (Ichiyanagi K et al. 2002; Ichiyanagi K et al. 2003). The target sequences usually have good match for IBS1, but not for IBS2 neither for the sequences recognized by the IEP. Different host organism may also influence which mobility pathways introns use, as the Ll.LtrB intron in *E. coli* retrotransposition by inserting into double-stranded DNA with varying priming mechanism (Coros CJ et al. 2005). The retrotransposition, with its lower sequence specificity, is evolutionary important and has allowed the spread of group II introns to new and different genomic locations. The reverse splicing step during the mobility event ensures that the intron will be excised from the mRNA transcript, thereby minimizing the damage on the host.

3.5.4 - Applications in targeted genome editing

The mode of DNA target site recognition by group II intron, involving largely base-pairing between the intron RNA and the DNA target site, and the finding that mutations in the DNA target site could be compensate by mutations in the intron (Guo H et al. 1997) have led to the development of engineered group II introns for gene targeting. It is indeed possible to retarget the insertion of group II intron at a chosen specific DNA site by simply modifying the base pairing sequences EBS of the intron (Eskes R et al. 1997; Guo H et al. 2000; Karberg M et al. 2001; Zhuang F et al. 2009a; Garcia-Rodriguez FM et al. 2011).

The development of these so called "targetrons" has first been achieved with the *L. lactis* Ll.LtrB group IIA intron for gene targeted disruption or insertion in bacteria (Karberg M et al. 2001; Frazier CL et al. 2003; Zhong J et al. 2003; Perutka J et al. 2004). The system used is now commercially available (TargeTron® Gene Knockout System, Sigma-Aldrich). Retargeted Ll.LtrB intron is usually deleted for the LtrA (IEP) ORF and is expressed from a donor plasmid with short flanking exons (Guo H et al. 2000; Zhong J et al. 2003). The IEP is expressed from a position just downstream of the 3'

exon. It was indeed shown that the homing efficiency could be significantly increased by deleting the LtrA ORF from the intron (Guo H et al. 2000). It was suggested that this Ll.LtrB intron-ΔORF is more resistant to the nucleolytic cleavage (Matsuura M et al. 1997). This way, the Ll.LtrB intron-ΔORF cannot splice once integrated without the expression of LtrA *in trans*. Retargeted introns are designed by computer program that search within the selected gene for best matches to the fixed positions recognized by the IEP and to the δ nucleotide recognized by the δ' nucleotide of the intron (Perutka J et al. 2004). The small number of fixed position (< 5) enables the targeting of several sites into any gene of *E. coli*. Once potential DNA target sites are found, the EBS1 and EBS2 of the intron are modified to match with the selected DNA target site. The IBS1 and IBS2 sequences of the 5' exon are also modified to pair with EBS1 and EBS2 intron sequences. Indeed, EBS/IBS pairing is required for the intron splicing in *E. coli*, which induces the formation of the homing catalytic RNP molecules. The retargeting of the Ll.LtrB intron is made by PCR amplification of an intron PCR template, which contains a part of the 5' exon with IBS1 and IBS2, and the beginning of the Ll.LtrB intron with EBS1 and EBS2 (Fig. I-23). Primers allowing IBS and EBS sequences modification are designed using an associated computer program and the PCR product amplified is subsequently ligated in the plasmid encoding the Ll.LtrB intron and LtrA intron-encoded protein (Fig. I-23; pACD4K-C).

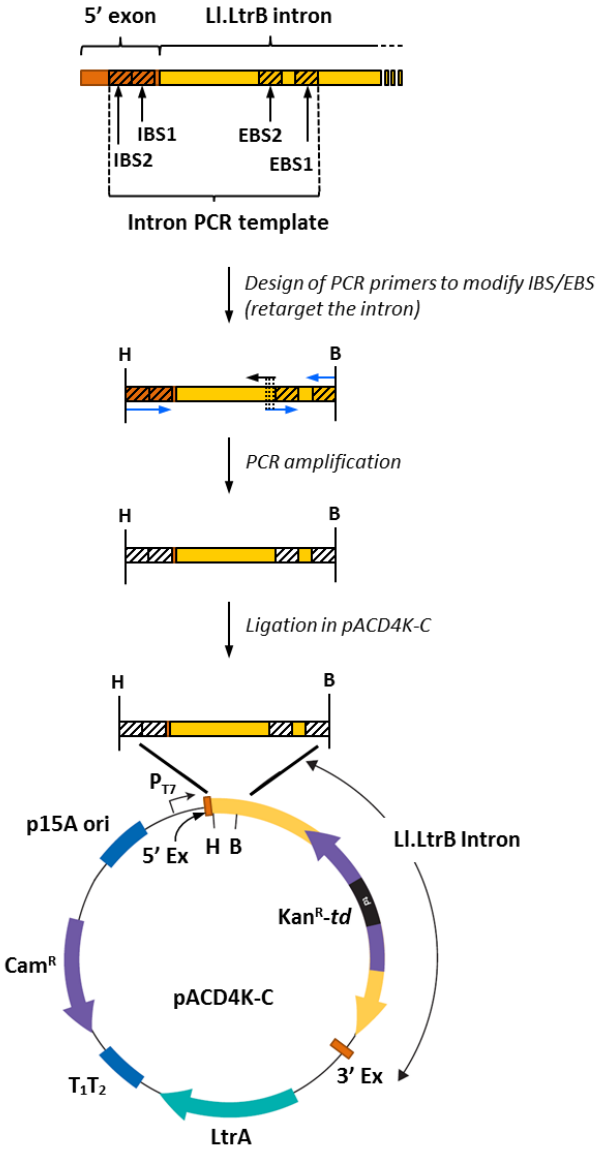


Figure I-23: Retargeting of Ll.LtrB intron in the TargeTron® Gene Knockout System.

A computer program is used to identify target sites in the gene of interest, and indicates the primers sequences that will be used to retarget the intron by PCR. The retargeted fragment also contains a part of the 5' exon sequence with retargeted IBS1 and IBS2. Retargeted intron segment is ligated into a linearized vector that contains the remaining intron components (Ll.LtrB intron sequence in yellow; 5' and 3' exon sequences in orange; LtrA protein in blue). IBS: intron-binding site; EBS: exon-binding site; H: Hind III restriction site; B: BsrG I restriction site; designed primers for retargeting: blue arrow; supplied primer for retargeting: black arrow; PT7: T7 promoter; 5' ex: 5' exon sequence; Kan^R-*td*: Kanamycin resistance gene interrupted by the *td* group I intron; 3' ex: 3' exon sequence; LtrA: Ll.LtrB intron-encoded protein; T1T2: T1-T2 transcription terminators; Cam^R: chloramphenicol resistance gene; p15A ori: p15A low-copy replication origin. *Adapted from "TargeTron® Gene Knockout System" manual.*

The plasmid pACD4K-C encoding the retargeted Ll.LtrB group II intron is then transformed into *E. coli*, the retargeted intron and LtrA are expressed, leading to the formation of the RNP and the integration of the intron into the DNA target site. *E. coli* cells in which the intron has been integrated can then be selected by kanamycin resistance.

Using this retargeted group II introns system, insertion at the chromosomal target site occurs at frequencies of about 1% in *E. coli* without selection. However, the method using selection of the homing event with the Kan^R-*td* retrotransposition-activated selectable marker (RAM marker) has been developed (Zhong J et al. 2003) and is based on a previously retrotransposition-indicator gene marker system (Ichiyanagi K et al. 2002). The strategy uses a Kan^R gene marker inserted in the DIV loop of Ll.LtrB, previously deleted for LtrA ORF, in the reverse orientation with respect to the Ll.LtrB group II intron orientation. The Kan^R gene is interrupted by the efficient self-splicing *td* group I intron, inserted in the forward orientation. During retrohoming (or retrotransposition), the *td* group I intron is excised from the RNA intermediate, enabling selection of the Kan^R marker after integration of the Ll.LtrB group II intron into a DNA target site. Nearly 100% of selected colonies using this RAM marker method have the desired targeting insertion.

Several applications of this gene knockout system in various Gram-negative or –positive bacteria species have been published (Chen Y et al. 2005; Yao J et al. 2006; Heap JT et al. 2007; Pearson MM and Mobley HL 2007; Yao J and Lambowitz AM 2007) (Malhotra M and Srivastava S 2008; Rodriguez SA et al. 2008; Sayeed S et al. 2008). In addition, Jones *et al*. have shown that retargeted group II introns could be used to insert a functional copy of a gene into the defective gene, or insert functional exons preceded by a splice acceptor site to overcome defective downstream exons (Jones JP, 3rd et al. 2005) in *E. coli*. However, the target DNA sites used in this study were inserted into plasmids rather than be chromosomal targets. Indeed, the efficiency of group II intron homing in plasmid target is much higher than in chromosomal DNA sites.

Recently, other group II introns such as the EcI5 group IIB intron from *E. coli* (Zhuang F et al. 2009a) and the RmInt1 group IIB intron from *S. meliloti* (whose IEP lacks the D and En domains) (Garcia-Rodriguez FM et al. 2011) have been used for gene targeting in *E. coli*. EcI5 was shown to be significantly more active than Ll.LtrB with specific chromosomal integration efficiencies of up to 98% without any selection (Zhuang F et al. 2009a). These new retargeted group II introns extend the range of accessible target sites for gene targeting in *E. coli* and other bacteria.

These successful gene targeting applications in bacteria have led to evaluate the use of group II introns as gene targeting vectors in eukaryotes (Guo H et al. 2000; Mastroianni M et al. 2008; Zhuang F et al. 2009b). The first attempt was performed in human HEK 293 and CEM T cell lines with the Ll.LtrB group II intron, retargeted to insert into CCR5 human gene and HIV-1 provirus (Guo H et al. 2000). The targets, inserted in plasmids, were co-transfected with RNP particles, previously reconstituted with *in vitro* self-spliced Ll.LtrB intron RNA and purified LtrA (Saldanha R et al. 1999) and packaged into liposomes. Although no information on the efficiency of targeted integration has been provided, the authors demonstrated intron insertion into the plasmid target DNA sites by PCR analyses. It is possible that the authors have to transfect RNP particles rather than express the intron and IEP into the human cells because of inefficient or inexistent splicing of the intron into human cells, even in presence of the IEP. In addition, in these initial experiments, group II intron integration into plasmid targets sites in human cells appear to be much less efficient than in bacteria, as detection of integration events required nested PCR analysis. To date, this work is the only attempt of a direct use of group II intron for gene targeting in human cells. To further increase the homing efficiency in an eukaryotic environment, Mastroianni *et al*. used additional $MgCl_2$ during microinjection of retargeted RNP particles into *Xenopus laevis* oocytes, and *Drosophilia melanogaster* and zebrafish (*Danio rerio*) embryos (Mastroianni M et al. 2008). Indeed, as mentioned previously, the splicing and homing of

group II intron are dependent on Mg^{2+}. It was shown that mutations in mitochondrial Mg^{2+} transport proteins in yeast strongly inhibited the splicing of all mitochondrial group II introns (Wiesenberger G et al. 1992; Gregan J et al. 2001a; Gregan J et al. 2001b). The authors postulated that the free Mg^{2+} concentration available in eukaryotic cells could be insufficient to permit the Ll.LtrB intron homing, as 10 mM of Mg^{2+} are required for its reverse splicing into DNA target sites (Saldanha R et al. 1999). The authors showed that the use of $MgCl_2$ during microinjection of retargeted RNP could enhance Ll.LtrB intron homing into its natural DNA site located in a plasmid, with integration efficiency of up to 27% in *X. laevis* oocytes (Mastroianni M et al. 2008; Zhuang F et al. 2009b). Specific insertion of a retargeted Ll.LtrB intron into a chromosomal target in *D. melanogaster* was also demonstrated, but again detection required nested PCR, implicating a very low homing efficiency.

The use of group II introns in human genome engineering could thus represent an attractive alternative to currently used strategies. Such group II intron mediated gene targeting is likely to be very specific, as group II intron mobility requires a base pairing with the target site of usually 12-16 bp, with additional contacts between the IEP and the DNA target site, extending the DNA target site selection to 30-35 bp. The design and production of engineered group II introns would be much easier and faster than those of site-specific nucleases. However, some hurdles are to be overcome. One of the most important is the problem of the very low efficiency of homing (and probably splicing) of the group II introns used in eukaryotic and human cells.

3.6 - DISTRIBUTION, CLASSIFICATION AND EVOLUTIONARY HYPOTHESES

Since the identification of group II introns as independent structural intron class (Michel F et al. 1982), the number of known group II introns has grown to hundreds of members. They are in low frequency in mitochondria of fungi, sporadically found in organellar genomes of algae, while numerous group II intron are found in organellar genomes of higher plants (Michel F et al. 1989). An exceptional example of a specie in which group II intron are overrepresented is *Euglena gracilis* with as much as 91 group II introns identified (Doetsch NA et al. 1998). Group II introns were also identified in proteobacteria and blue algae and are thought to be “ancestors” of mitochondrial and chloroplasts group II introns (Ferat JL and Michel F 1993; Ferat JL et al. 1994; Mills DA et al. 1996; Shearman C et al. 1996). Sequencing projects of bacterial genomes have revealed that group II introns are also widespread in eubacteria and a few members are found in archeabacteria (Martinez-Abarca F and Toro N 2000; Dai L et al. 2003; Toro N et al. 2007; Valles Y et al. 2008).

The close relationship between the IEP and intron RNA structure and the presence of the IEP in all intron subclasses have also led to the hypothesis that the group II intron ancestor was essentially a retroelement (Toor N et al. 2001; Dai L and Zimmerly S 2002b). The presence of the IEP in a quite conserved location in domain IV suggests that the IEP was acquired once by insertion of a retroelements into an already catalytic ribozyme. Alternatively, the self-splicing ability might have been developed later by a retroelement in order to prevent host damage. The “retroelement ancestor hypothesis” predicts that the various structural lineages of group II introns arose by coevolution with the IEP from an ancestor intron in bacteria, which had an RNA structure characterizing by a mix of the different known structural features and a compact reverse transcriptase ORF (Toor N et al. 2001; Robart AR and Zimmerly S 2005). Bacterial introns are usually not found in important housekeeping genes, but rather in intergenic regions or other mobile elements (Dai L and Zimmerly S 2002b; Ichiyanagi K et al. 2003; Robart AR and Zimmerly S 2005). These properties also support their

"selfish" retroelement character, as they insert in genomic locations that minimize their impact on the host and/or will favorer their spread. Two of the group II intron subclasses are predicted to have migrated to the organelles of eukaryotes (the mitochondrial and chloroplast-like lineages), which was followed by loss of the IEP and degeneration in several RNA features, especially in plants where almost all group II introns are ORF-less (Toor N et al. 2001). Organellar group II introns are inserted in many highly conserved genes essential for respiration and photosynthesis and therefore must retain efficient splicing properties. As opposed to the bacterial introns, organellar elements behave more like splicing-only elements and rely on host-encoded splicing factors (Toor N et al. 2001; Lehmann K and Schmidt U 2003; Robart AR and Zimmerly S 2005).

There are also several similarities in RNA structure and splicing mechanism between group II introns and nuclear spliceosomal introns (Valadkhan S 2007). An evolutionary hypothesis is that group II introns invaded the eukaryotic nucleus and then have been successively fragmented, while retaining the fundamental catalytic mechanism of self-splicing with the evolution from *cis*-acting elements to *trans*-acting such as small RNAs that became dependent on host protein factors for the splicing reaction (Sharp PA 1991). This transition may have involved the split of ancestral group II introns in catalytically inactive spliceosomal introns and the catalytically active RNA moiety of the spliceosome (Martin W and Koonin EV 2006). This theory is based on the endosymbiosis between an ancestral α-proteobacterium and an archeal host. Several observations support this idea, as the fact that fragmented bacterial and organellar group II introns can perform efficient splicing *in trans* (Knoop V et al. 1997; Belhocine K et al. 2008). Some domains have been shown to act *in trans* on fragmented introns to promote the splicing, thus strongly suggesting a direct evolution connexion between group II introns and spliceosomal snRNA. Other clues such as the the presence of metal-ion binding sites in DV and U6 snRNA, the branch site motifs in DVI, similar to the U2-intron pairing, the structural similarities between terminal regions of spliceosomal introns and those of group II introns, and the structural and functional similarities between DV and U6 snRNA strongly support this hypothesis (Sharp PA 1985; Jacquier A 1990; Madhani HD and Guthrie C 1992; Shukla GC and Padgett RA 2002; Villa T et al. 2002; Seetharaman M et al. 2006; Keating KS et al. 2010; Rogozin IB et al. 2012).

There is also a relationship between group II introns and the non-LTR retroelements found in higher eukaryotes (Malik HS et al. 1999; Robart AR and Zimmerly S 2005). The RT segments of the protein of the two types of elements are phylogenetically and structurally related and both elements are mobile through a similar mechanism (Lambowitz AM and Zimmerly S 2004; Beauregard A et al. 2008).

Altogether, this has put up the scenario that mobile group II introns may be the ancestors of spliceosomal introns and non-LTR retroelements, and therefore may have played a substantial role in the evolution of the eukaryotic genome as predecessors of the spliceosome and retrotransposons.

3.7 - *PYLAIELLA LITTORALIS* PL.LSU/2 GROUP II INTRON

The analysis of the mitochondrial genomic region encoding the large subunit ribosomal RNA (LSU rRNA) of the brown algae *Pylaiella littoralis* has revealed the presence of four group IIB introns in the gene (Fontaine JM et al. 1995). Interestingly, introns Pl.LSU/1 to Pl.LSU/3 were found to exhibit high canonical secondary and tertiary structure related to those of RNA structural class IIB1, while Pl.LSU/4 belongs to the RNA structural class IIB2. The first three introns are also closely related with regards to their primary sequences, suggesting that they have transposed in *cis* along the gene (Ferat JL et al. 1994). Moreover, when a LSU rRNA DNA probe was incubating with total RNA from the algae, a strong signal was observed corresponding to mature LSU rRNA, suggesting that those group

II introns are spliced correctly in their natural host (Fontaine JM et al. 1995). Several introns of the LSU rRNA gene are also not present in some other strains of *P. littoralis* originating from different geographic location (Fontaine JM et al. 1995; Ikuta K et al. 2008), suggesting that some of them are the result of recent invasion. The first two introns were found to contain an ORF encoding a RT-like protein located in the DIV loop and related to non-LTR retrotransposons RT (Michel F et al. 1982; Michel F and Lang BF 1985). An alignment of the amino acid sequence of those RT-like proteins revealed the presence of the conserved domains found in group II intron-encoded proteins, which are RT (reverse transcriptase), X (maturase), D (zinc-finger-like DNA binding), and En (endonuclease) domains (Fontaine JM et al. 1995). A phylogenetic analysis of these RT-like proteins revealed that they belong to the lineage of RT from plastid, cyanobacteria, and γ-proteobacterial (class CL1, See Fig. I-17), which is consistent with a co-evolution of the group II intron and their inserted RT, as the corresponding group II introns belong to the RNA structural class IIB1.

The self-splicing ability *in vitro* of Pl.LSU/1 to Pl.LSU/3 group II introns was assayed under different conditions (Costa M et al. 1997b), and the results revealed that Pl.LSU/2 is a remarkably efficient ribozyme catalyst. Pl.LSU/2 was found to be highly reactive under most of the conditions tested. It was observed that no improvement of the catalytic activity was detected at high ionic strength (1 M of NH_4Cl, $(NH_4)_2SO_4$, or KCl) when magnesium concentration was raised from 10 mM to 100 mM (Costa M et al. 1997b), while most group II introns require high magnesium concentrations for efficient self-splicing *in vitro* (usually > 50 mM) (Peebles CL et al. 1986; van der Veen R et al. 1986; Jarrell KA et al. 1988; Matsuura M et al. 1997). More interestingly, the Pl.LSU/2 intron was shown to still function at the very low Mg^{2+} concentration of 0.1 mM in presence of 1 M NH_4Cl, with a splicing of about 10% of precursor molecules in 1 hr (Costa M et al. 1997b). Pl.LSU/2 can also undergo self-splicing by the branching pathway and the hydrolysis pathway *in vitro*, depending on the conditions used. It was also shown that a homogenous and correctly folded population of Pl.LSU/2 RNA could be obtained at low Mg^{2+} concentration and that a high proportion of intron lariat RNA could be obtained under optimal conditions.

The secondary and tertiary structures of Pl.LSU/2 intron RNA were subsequently investigated by biochemical analyses (Costa M et al. 1998; Costa M and Michel F 1999; Costa M et al. 2000; Li CF et al. 2011). These studies allowed the refinement of the secondary structure of intron domain V and provided experimental evidence for a number of structural characteristics such as α-α', θ-θ', ζ-ζ' tertiary interactions. In addition, requirements for efficient 5' exon-intron binding were determined and correspond to EBS/IBS pairing as well as the presence of the intron domain V and the completion of the intron active site formation, and EBS3/IBS3 as well as δ-δ' interactions were identified and their role in 5' and 3' exons binding to the intron were outlined. More recently, nuclear magnetic resonance (NMR) spectrometry studies were performed to further describe the overall structure of the catalytic domain V (Dayie KT 2005; Seetharaman M et al. 2006) and the dynamic profile of this domain (Eldho NV and Dayie KT 2007). The structure of the DV domain of Pl.LSU/2 group II intron was very recently refined *in silico* (Henriksen NM et al. 2012). The secondary structure of Pl.LSU/2 group II intron with its known tertiary interactions is represented in Fig. I-24.

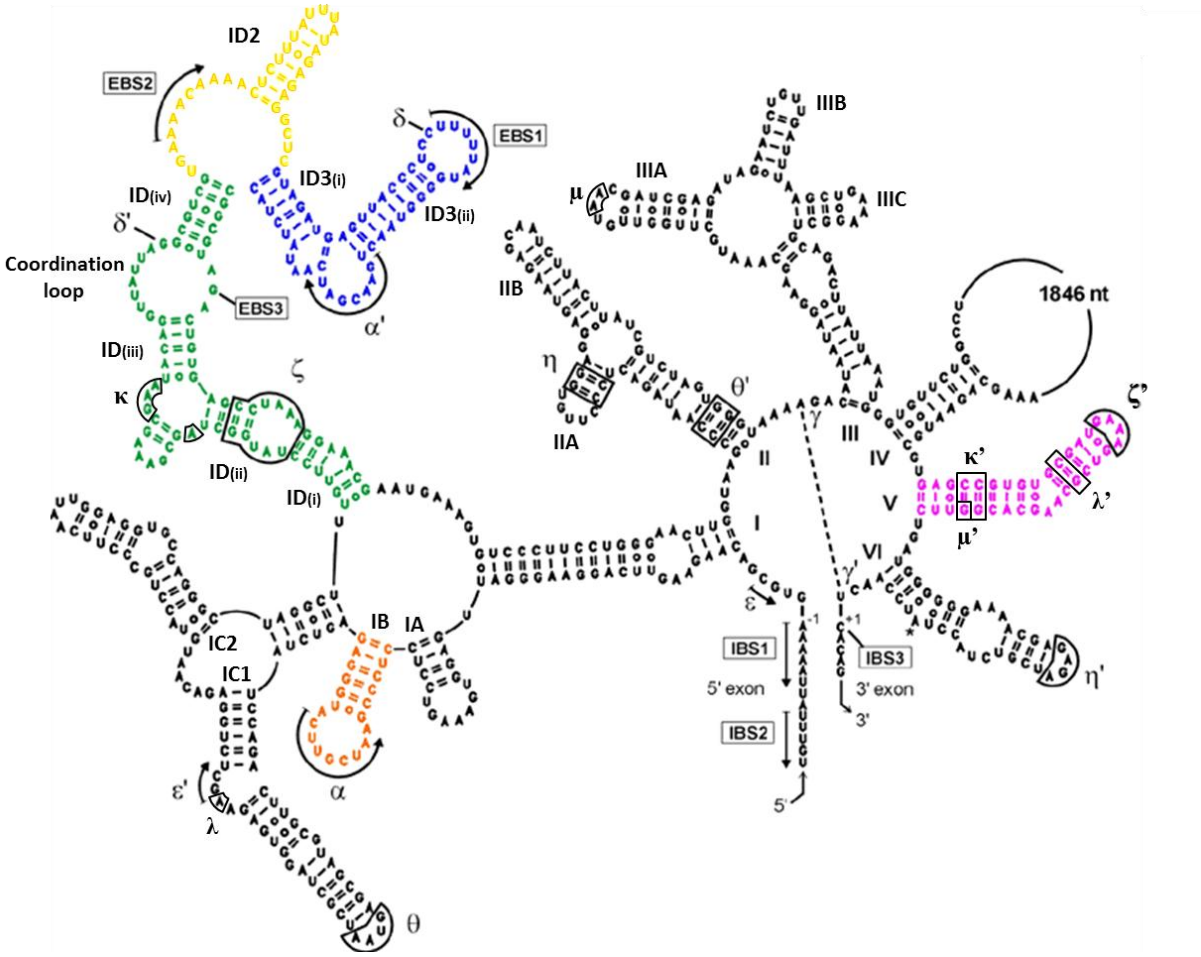


Figure I-24: Secondary structure of Pl.LSU/2 group II intron with tertiary interactions.

EBS and IBS sequences are indicated and tertiary interactions involving intron sequences are noted by Greek letters. The remaining 1846 nt of DIV is represented by the black line and contains the IEP ORF. *The figure was taken from the F. Michel lab website (http://www.cgm.cnrs-gif.fr/michel/) and contains minor modifications.*

In 2001, the complete sequence of the brown algae mitochondrial genome of *P. littoralis* (strain L. Kjellm) becomes available (Oudot-Le Secq MP et al. 2001). The amino acid sequence of the Pl.LSU/2 IEP (gi|15150713) was thus annotated using previous results of Pl.LSU/2 IEP alignment with other intron-encoded protein amino acid sequences (Fontaine JM et al. 1995) (Fig. I-25).

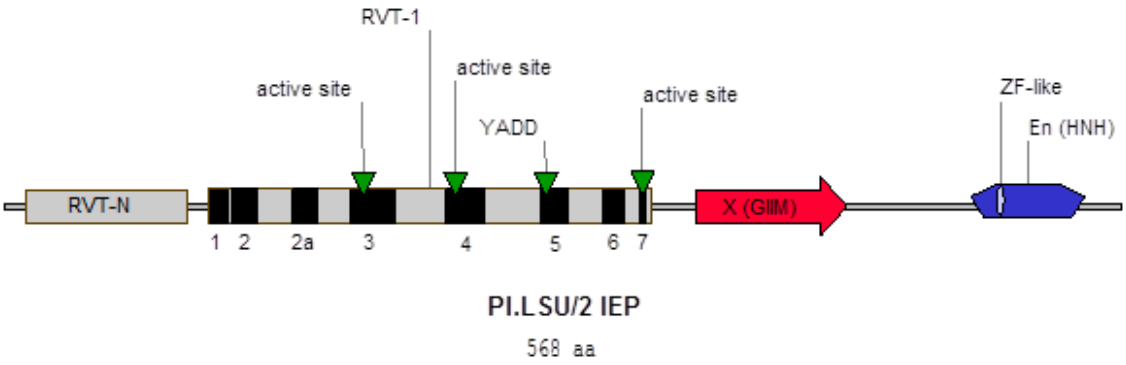


Figure I-25: Schematic representation of the Pl.LSU/2 IEP with its conserved domains.

RVT-N: N-terminal domain of reverse transcriptase (Pfam 13655); RVT-1: RNA-dependent DNA polymerase (Interpro IPR000477, Pfam 00078); black rectangles: RT blocks 1 to 7; active site: putative active sites of the RT domain (See NCBI annotation); YADD: conserved RT catalytic motif; X (GIIM): Group II intron maturase-specific domain (Interpro 013597, Pfam 08388); En (HNH): Endonuclease domain presenting the HNH endonuclease signature (Interpro IPR002711/IPR003615); ZF-like: DNA-binding domain with zinc-finger characteristic cysteines signature $C_{(2-3)}C$.

As mentioned above, the Pl.LSU/2 IEP sequence presents all the characteristic domains of group II intron-encoded proteins. The RT domain is here represented by the N-terminal RVT-N domain present in some reverse transcriptase proteins, such as bacterial reverse transcriptases, and the typical RVT-1 domain (RNA-dependent DNA polymerase) found in all reverse transcriptase proteins. This latter domain contains the conserved RT blocks 1 to 7 and presents four putative active sites (putative NTP binding sites or nucleic acid binding site), including the highly conserved YNDD catalytic motif (YADD in Pl.LSU/2 IEP). The Pl.LSU/2 sequence also contains the conserved X domain, corresponding to the group II intron maturase-specific domain, and the endonuclease domain with the HNH endonuclease signature. The DNA-binding domain corresponding to the ZF-finger characteristic cysteines signature $C_{(2-3)}C$ has been defined by alignment of the Pl.LSU/2 IEP sequence with those of other group II intron-encoded proteins (Fontaine JM et al. 1995).

Several observations show that the Pl.LSU/2 group II intron is a highly active ribozyme *in vitro*, and together with the observation of mature LSU rRNA in algae (Fontaine JM et al. 1995), the absence of degeneration of its IEP encoding sequence suggest that this protein could also be active *in vivo*. However, there is no experimental validation of the biochemical activities of the IEP and no direct proof of catalytic activity of the Pl.LSU/2 intron *in vivo*.

4 - AIM OF THE THESIS

The literature review presented in introduction shows that several strategies are currently developed to achieve stable gene transfer in gene therapy. The most widely used integrative gene therapy vectors are those based on retroviruses. However, safety concerns with regards to potential insertional mutagenesis led to intensive research for the development of alternative tools. The ultimate goal in the gene therapy field would be the development of techniques enabling precise and highly specific gene repair (or gene disruption in the case of anti-viral therapies). To date, the use of site-specific nucleases is the only way to achieve site-specific gene repair. Although these approaches seem to be very promising, several hurdles still have to be overcome. The main concern in all the different gene targeting approaches developed so far is the safety of these new tools. Indeed, in most cases, these strategies involve a double-strand break of the DNA followed by DNA repair by homologous recombination. Such an approach is potentially mutagenic and could induce genomic rearrangements if off-targets are not avoided. The presence of off-target integrations is itself a concern, as the goal of these site-specific gene targeting approaches is precisely to induce DNA integration at a chosen and unique location of the genome. In addition, these targeting strategies are based on engineered proteins, whose design and production are often labor-intensive.

Mobile group II introns are self-splicing ribozymes already used in gene targeting approaches in bacteria (Karberg M et al. 2001; Frazier CL et al. 2003; Zhong J et al. 2003; Perutka J et al. 2004). They can be retargeted to a chosen specific site of the genome by simply modifying the intron sequences involved in the recognition of their DNA target site. Indeed, as described in introduction, group II intron can integrate into double-stranded DNA at a specific and unique site, mainly recognized by base-pairing between the spliced intron RNA and the DNA target site. The mechanism of group II intron integration is highly specific, as it involve a base-pairing of about 12-15 nt, and additional contacts are made between the intron-encoded protein and the DNA target site, increasing the number of connected nucleotides to 15-21 nt (reviewed in (Lambowitz AM and Zimmerly S 2004)). The design and production of these targeting tools are very easy and fast, as it can be performed by PCR amplification. However, the attempt of using group II introns for gene targeting in human cells has faced some important bottlenecks (Guo H et al. 2000). Indeed, the mostly used Ll.LtrB group II intron appears to be relatively inefficient for both splicing and homing in human cells, and more generally in eukaryotic cells . The main barrier seems to be the high requirement in Mg^{2+} ions of Ll.LtrB for catalysis (Mastroianni M et al. 2008; Zhuang F et al. 2009b). A general solution would thus be the use of a group II intron with a lower dependence on Mg^{2+} for catalysis.

The mitochondrial Pl.LSU/2 intron from the brown algae *Pylaiella littoralis* has been shown to be a highly active catalyst *in vitro* and can retain catalytic activity even in presence of remarkably low Mg^{2+} concentration (Costa M et al. 1997b). The secondary and tertiary structures of this group II intron are also defined (Costa M et al. 1998; Costa M and Michel F 1999; Costa M et al. 2000; Li CF et al. 2011). In addition, this intron contains in its DIV loop an open-reading frame putatively encoding an IEP, whose sequence analysis reveals the presence of all conserved domains of group II intron-encoded proteins (Fontaine JM et al. 1995). We postulated that the unusually low requirement of Mg^{2+} for the Pl.LSU/2 intron catalytic activity, which is a unique feature, could make this intron a good candidate for specific genetic engineering in human cells. To use this intron as a site-specific

vector, two molecules have to be functionally active *in vivo*: the Pl.LSU/2 intron RNA and the Pl.LSU/2 IEP. Although the catalytic activity of the intron was well defined *in vitro*, its catalytic activity *in vivo* was not reported yet, and the biochemical activities of the IEP were not determined. In this context, the aim of this study was to further characterize this intron in order to evaluate its potential use for gene targeting in human cells. The Pl.LSU/2 intron-encoded protein was expressed and purified in order to assay its potential biochemical activities. Several expression hosts and purification systems were used to achieve the purification of catalytically active IEP. The reverse transcriptase activity of the IEP was subsequently assayed. In addition, the Pl.LSU/2 intron splicing, mediated or not by the IEP, was evaluated *in vivo* first in *S. cerevisiae*, and then in a human cell line. Ultimately, the homing capacity of Pl.LSU/2 intron into its natural DNA target site was evaluated in *E. coli* and *S. cerevisiae*.

Even though site-specific gene therapy vectors could represent safer strategies and open the way of dominant gene disorders treatment, their development is still under study. To date, successful clinical trials using retroviral vectors have highlighted the usefulness of these integrating vectors for gene therapy (Mavilio F et al. 2006; Aiuti A et al. 2009; Cartier N et al. 2009; Boztug K et al. 2010; Cavazzana-Calvo M et al. 2010; Hacein-Bey-Abina S et al. 2010; Gaspar HB et al. 2011). Although much effort has been made to increase the safety of retroviral vectors such as the development of SIN-vectors, it seems necessary to assay their safety in gene therapy protocols. The risk of insertional mutagenesis is directly correlated with the number of integrated vector copies into transduced cells, and a too high number of insertions per cell has more probability of inducing abnormal clonal expansions, as usually more than one transforming event is required for tumorigenesis (Moolten FL and Cupples LA 1992; Du Y et al. 2005a; Du Y et al. 2005b; Modlich U et al. 2005).

During my thesis, I had the opportunity to participate to a work developing and validating a simple method to quantify the vector copy number at the single-cell level. This study was conducted by the Anne Galy's team at Genethon with the aim of determining vector safety and optimizing protocols of gene therapy using HIV-1-based lentiviral vectors, and in particular for the treatment of the Wiskott-Aldrich syndrome. This work has been published in 2011 in Gene therapy (Charrier S et al. 2011) and is summarized in the following section.

PART I: ANALYSIS OF LENTIVIRAL VECTOR COPY NUMBER

1 - SUMMARY OF THE WORK

The Wiskott-Aldrich syndrome (WAS) is a rare (1/200,000 male births) X-linked immunodeficiency due to mutation in the WAS gene encoding a protein named WASp (for Wiskott-Aldrich syndrome protein) (Derry JM et al. 1994). WASp is a major regulator of the cytoskeleton expressed in hematopoietic cells (reviewed in (Bouma G et al. 2009)). This disease is characterized by several clinical manifestations such as micro-thrombocytopenia (diminution of the number of platelets), recurrent infections caused by the immunodeficiency, eczema, etc. (Sullivan KE et al. 1994). When HLA-compatible donors exist, WAS patients can be treated by allogenic hematopoietic stem cells transplantation (reviewed in (Pai SY and Notarangelo LD 2010)). However, patients without HLA-compatible donors and presenting a mutated form of the WAS gene associated with severe clinical manifestations have no other therapeutic option than gene therapy (reviewed in (Galy A and Thrasher AJ 2011)).

As mentioned in Introduction section 1 -, a first gene therapy clinical trial has been opened recently in Germany for the treatment of WAS patients using (non-SIN) γ-retroviral vectors (Boztug K et al. 2010). Although clear success was demonstrated, one patient developed a vector-induced T-cell leukemia, due to vector insertion near the proto-oncogene LMO2. In addition to this adverse event, several hurdles inherent to the use of γ-retroviral vectors for gene therapy have led to the development of lentiviral vectors (LV) with improved safety for gene therapy of WAS (reviewed in (Galy A et al. 2008)). Indeed, encouraging results were obtained from preclinical studies using recombinant HIV-1-derived (rHIV) SIN lentiviral vectors (LV) encoding WASp (Dupre L et al. 2006; Marangoni F et al. 2009; Zanta-Boussif MA et al. 2009; Scaramuzza S et al. 2012). The assessment of the vector safety was done by analyzing the insertional pattern (Mantovani J et al. 2009), which showed usual rHIV insertion profile, as well as evaluating the vector-mediated transformation ability *in vitro* (Modlich U et al. 2009) and *in vivo* in an animal model (Marangoni F et al. 2009), demonstrating a higher safety of rHIV LV compared to γ-retroviral vectors with regards to insertional mutagenesis. These results led to the opening of a international multicenter phase I/II clinical trial (Galy A and Thrasher AJ 2011).

To set up a gene therapy protocol, two parameters have to be equilibrated: a sufficient transduction level needs to be achieved to obtain the required therapeutic level of protein expression, and on the other hand, the number of vector insertions per cell has to be kept as low as possible to reduce the risk of insertional mutagenesis. The insertional mutagenesis risk increases as the integrated virus copy number rises (Moolten FL and Cupples LA 1992; Du Y et al. 2005a; Du Y et al. 2005b; Modlich U et al. 2005). The toxicity of multiple vector insertions may also result in the apoptosis of transduced cells at high vector dose (Arai T et al. 1999). Increasing the vector dose or the number of transduction round to improve the transduction efficiency could thus not be appropriate in terms of safety (Kustikova et al. 2003). Hence, for each clinical application, retroviral transduction methods require optimizations not only to maximize the transduction efficiencies but also to deliver a narrow range of integrated copies per cell. A quantitative analysis of both transduction efficiency and vector copy number (VCN) per cell should provide insights to assist the optimization of gene therapy protocols.

In the case of hematopoietic stem/progenitors cells transduction, the initial frequency of transduced cells and the number of vector integrations in individual cells can be assessed on colony-forming cells (CFC), which are formed by hematopoietic progenitor cells during a low cell density culture. Thus, a

comparison of the number of transgene-positive CFC to the total can estimate the transduction efficiency of the hematopoietic progenitor cells population. The quantification of VCN in CFC can also provide information on multiple vector copies, in contrast to the quantification of an average VCN in a heterogenous cell population. A VCN quantification method on CFC by Q-PCR was already proposed, although not experimentally validated (Schuesler T et al. 2009).

In this work, Charrier et al. developed and validated a simple method to quantify VCN in individual CFC relying on a single-step genomic DNA extraction and a duplex Q-PCR analysis. Three control cell clones with known VCN and integration sites (Mantovani J et al. 2009) were generated. In this study, I confirmed by Southern blot the VCN of these control cells clones. These clones were then used to assess the sensitivity, accuracy and reproducibility of the developed method. It was shown that the Q-PCR method was sufficiently sensitive to determine VCN per cell of as low as 100 cells. Although the range of VCN determined was similar to expected values, VCN of these control clones was underestimated by 30-40%. However, this method still provides dose-dependent results and can thus evaluate the distribution of VCN at the single-cell level. This Q-PCR technique was then used to evaluate the transduction efficiency of $CD34^+$ hematopoietic progenitor cells by a LV encoding the GFP (GFP-LV). Positive correlation was found between CFC positives for vector and CFC expressing the GFP. In addition, the number of CFC positive for vector obtained by Q-PCR was very similar to theoretical data estimated from the average mean transduction rate using the Poisson distribution and corresponding to the probability of transduction of single hematopoietic cells (Fehse B et al. 2004). Altogether, these data contributed to validate the method consisting of the determination of VCN on CFC by Q-PCR to quantify the frequency of transduction of hematopoietic progenitor cells and determine the distribution of VCN per cell on the transduced cell population.

This method was subsequently used to evaluate various conditions of hematopoietic progenitor cells transduction by a GFP-LV and a clinically relevant LV encoding WASp (WASP-LV). The effect of the vector concentration and the number of round of transduction were evaluated. It was shown that for the GFP-LV, the frequency of transduced cells significantly increased with higher vector dose and that the distribution of VCN was modified, with a higher median values and range of VCN. On the other hand, the use of two hits of transduction with the GFP-LV had only a little effect on the improvement of the frequency of transduced cells, but resulted in a clear modification of VCN distribution, with an increased number of cells with high VCN. The results were not similar when using the WASP-LV. The frequency of transduced cells was also increased when using higher vector doses but the distribution of VCN was not significantly impacted.

Interestingly, this Q-PCR method was applied to evaluate the activity of a WASP-LV purified by chromatography with clinical grade manufacturing techniques (Merten OW et al. 2011) and used in current WAS gene therapy clinical trials (Galy A and Thrasher AJ 2011). Indeed, high concentrated rHIV particles are usually required for efficient transduction of primary cells (Haas DL et al. 2000) and a common concentration technique rely on ultracentrifugation of vector-containing cell extracts. However, such materials contain contaminating elements as cellular debris, membrane fragments and proteins, which can be toxic to target cells (Selvaggi TA et al. 1997; Reiser J 2000; Tuschong L et al. 2002; Baekelandt V et al. 2003). Thus, alternative methods to concentrate and extensively purify gene therapy vector, allowing its clinical use, were developed in the context of the WAS gene therapy clinical trial. Comparable levels of transduction were determined between WASP-LV purified by

chromatography and the previously evaluated WASP-LV purified by ultracentrifugation. As for WASP-LV purified by ultracentrifugation, it was shown that increasing the vector concentration improved the transduction efficiency without modifying the VCN distribution. The effect of two hits of transduction was also evaluated and showed an increase of the frequency of transduced cells, with a better improvement than those observed with the use of higher vector dose. The modification of the VCN distribution described with the GFP-LV showing an increase in the number of cells with high VCN does not occurred with two hits of transduction by WASP-LV. Altogether, these data indicate that different preparations of LV could have different behavior impacting on the transduction efficiency and on the level of integrated vector copies per cell.

This study also reveals a bias in the VCN distribution among different types of cells, with higher VCN values in erythroid colonies than in myeloid colonies, probably due to the experimental conditions that favored preferential transduction of erythroid progenitor cells.

To conclude, a simple method based on Q-PCR for analyzing vector copy number in individual CFC has been developed and validated, and has determined the frequency of transduction and the distribution of vector copies in hematopoietic progenitor cell population. This work demonstrated the influence of experimental conditions as well as the vector type on these two parameters. Such a method would provide important data for optimizing gene therapy protocols and can be used together with vector insertion site analyses to assess vector safety.

2 - ARTICLE 1

Open

Gene Therapy (2011) 18, 479–487

www.nature.com/gt

ORIGINAL ARTICLE

Quantification of lentiviral vector copy numbers in individual hematopoietic colony-forming cells shows vector dose-dependent effects on the frequency and level of transduction

S Charrier[1,2,3], M Ferrand[1,2,3], M Zerbato[1,2,3], G Précigout[1,2,3], A Viornery[1,2,3], S Bucher-Laurent[1], S Benkhelifa-Ziyyat[1,2,3], OW Merten[1], J Perea[1,2,3] and A Galy[1,2,3]

Lentiviral vectors are effective tools for gene transfer and integrate variable numbers of proviral DNA copies in variable proportions of cells. The levels of transduction of a cellular population may therefore depend upon experimental parameters affecting the frequency and/or the distribution of vector integration events in this population. Such analysis would require measuring vector copy numbers (VCN) in individual cells. To evaluate the transduction of hematopoietic progenitor cells at the single-cell level, we measured VCN in individual colony-forming cell (CFC) units, using an adapted quantitative PCR (Q-PCR) method. The feasibility, reproducibility and sensitivity of this approach were tested with characterized cell lines carrying known numbers of vector integration. The method was validated by correlating data in CFC with gene expression or with calculated values, and was found to slightly underestimate VCN. In spite of this, such Q-PCR on CFC was useful to compare transduction levels with different infection protocols and different vectors. Increasing the vector concentration and re-iterating the infection were two different strategies that improved transduction by increasing the frequency of transduced progenitor cells. Repeated infection also augmented the number of integrated copies and the magnitude of this effect seemed to depend on the vector preparation. Thus, the distribution of VCN in hematopoietic colonies may depend upon experimental conditions including features of vectors. This should be carefully evaluated in the context of *ex vivo* hematopoietic gene therapy studies.
Gene Therapy (2011) **18,** 479–487; doi:10.1038/gt.2010.163; published online 16 December 2010

Keywords: Lentivirus; CD34+ cell; Q-PCR; method; hematopoietic cells; CFC

INTRODUCTION

Genetically corrected hematopoietic stem cells (HSC) can be used as an alternative to allogenic HSC transplantation for the correction of several types of inherited diseases.[1] Different vectors have been used for the gene modification of HSC, but recombinant HIV-1-derived (rHIV) lentiviral vectors (LV) appear to be promising for gene-modification of the long-term repopulating HSC population, as shown in several preclinical models[2–4] and in clinical trials in man.[5] Thus, novel clinical applications of rHIV vectors are actively developed, including for instance a treatment of Wiskott–Aldrich syndrome (WAS). A LV encoding the WAS protein (WASP) is being developed to treat this life-threatening X-linked primary deficiency.[6]

Depending on the experimental conditions, rHIV vectors can transduce variable percentages of cells and integrate variable numbers of copies of proviral DNA into the genome of target cells. Indeed, it has been shown that human HSC are permissive to the integration of multiple copies of rHIV. As high as six integrations per cell have been detected on average in populations of human hematopoietic cells engrafted in the bone marrow of immuno-incompetent mice.[7] The stable insertion of genes in hematopoietic progenitor cells has a significant impact as these cells will transmit their genomic heritage to a considerable number of cells given their proliferation and differentiation potential. The biological potency of the vector is expected to correlate positively with the frequency of transduced cells and also with the number of integration per cell unless transgene silencing is observed, as suggested in some studies.[8] At the same time, genotoxicity related to the number of vector insertions per cell can result from inappropriate transgene expression in cells[9] or from effects of elements contained in the integrated cassette,[10] thus requiring a strict control of the number of gene insertion per cell. It is therefore important to determine the distribution of vector copies in the infected cell population at the single cell level to assess the efficacy and safety, that is, the therapeutic window of integrative vectors.

Several methods have been used to measure the transduction of human hematopoietic cells with LV but there are few reports of validated techniques to determine vector copy distribution in a population from single-cell measures. In earlier studies, average number of vector copies have been measured in cell populations using Southern blot detection of vector sequences, calibrated on dilutions of genomic DNA and a housekeeping gene .[11] More recently, quantitative PCR (Q-PCR) has been used to provide more precise measures and the technique can be calibrated on dilutions of a plasmid bearing both vector and genomic sequences to determine average vector copy number (VCN) in a human cell population

[1]Genethon, Evry, France; [2]Inserm, U951, Evry, France and [3]University of Evry Val d'Essonne, UMR_S951, Evry, France
Correspondence: Dr A Galy, Genethon, 1 bis rue de l'Internationale, 91002 Evry, France.
E-mail: galy@genethon.fr
Received 6 July 2010; revised 18 October 2010; accepted 20 October 2010; published online 16 December 2010

transduced with a rHIV LV.[12] Such average values does not provide any indication on the initial frequency of transduced cells in the target population nor does it estimate the numbers of integrations in individual transduced cells. Hematopoietic progenitor cells have the ability to form colonies arising from a single cell when cultured at low cell density in semi-solid medium. After a culture period of about 2 weeks in methylcellulose, individual colonies can be picked and the genomic DNA of the cells can be extracted from these colony-forming cells (CFC) with proteinase K and phenol/chloroform to determine the presence or absence of vector by PCR and agarose gels, thus, determining the frequency of transduced hematopoietic progenitor cells in the initially-infected population of cells.[13] Protocols for the quantification of VCN in CFC by Q-PCR have been reported recently but not validated with experimental data.[14] Here, we have developed a simplified method relying on a single-step genomic DNA extraction and a duplex Q-PCR method to quantify VCN in individual CFC. We have validated this approach experimentally. Transduced and cloned human cell lines were generated as controls and used to demonstrate the feasibility, sensitivity and reproducibility of this protocol. Measures of VCN in CFC have been used to evaluate various conditions for the transduction of hematopoietic progenitor cells with LV encoding the green fluorescence protein (GFP) reporter transgene (GFP-LV) or WASP (WASP-LV). Results show that the frequency of transduced CFC and the distribution of VCN in these cells could be augmented by repeated infection. This approach should be useful to optimize rHIV transduction protocols and to verify vector safety.

RESULTS

Generation of controlled human cell lines to measure rHIV vector integration

Transduced human cells containing a known amount of rHIV integration can serve as control materials to analyze VCN in cells. A panel of such control cells was generated by infecting the human fibrosarcoma HT1080 cell line with a GFP-LV, then selecting and characterizing stably-transduced single-cell clones with variable numbers of VCN. HT1080 cells were used because of their rapid growth and few chromosomal changes from the diploid karyotype (http://www.lgcstandards-atcc.org/ and Ref. 15). Following transduction with the vector and two rounds of single-cell cloning, a single HT1080 cell clone HT4-A was first selected and analyzed. Vector insertion sites were determined with the vector integration tag analysis (VITA) technique described elsewhere[16] showing a single vector insertion in chromosome 19 (Table 1). Subsequently, this clone was re-infected with the GFP-LV and two daughter clones displaying different levels of expression of GFP were selected by single-cell cloning. These HT4-A2 and HT4-A6 cells showed, respectively, three and eight vector insertions in various genomic locations, including the initial insertion in chromosome 19. The number of vector insertion sites in the HT4-A, HT4A-2 and HT4-A6 was confirmed by Southern-blot analysis using a single-restriction enzyme (*Xba*I) to reveal the LV integration banding pattern. As shown in Figure 1a, we observed one band in clone HT4-A, three bands in HT4-A2 and eight bands in clone HT4-A6. With the slight over-exposition of the blot, a very weak band appears at about 8 kb in all HT4 clones and corresponds to a nonspecific hybridation in the stringency conditions used in the experiment.

In previous studies, we have used a duplex Q-PCR for comparative amplification of vector-specific sequences (human immunodeficiency virus or woodchuck hepatitis post-transcriptional regulatory element) versus a human cellular gene (the human albumin gene (*ALB*) for which two copies are present per cell) to determine the number of rHIV vector insertions in human cells.[12] The reaction is calibrated by a standard curve made by serial dilutions of a plasmid carrying a single copy of vector bearing WPRE and of the ALB sequence (see supplementary Table S1). Analysis of the genomic DNA from HT4-A, HT4-A2 and HT4-A6 cell clones by such Q-PCR protocol determined 1.1 ± 0.2, 2.8 ± 0.7 and 7.1 ± 1.3 VCN in repeated experiments (Table 1). These results were consistent with the determination of vector insertions by VITA and by Southern blot in each of the clones. The numbers of vector copy also correlated strongly with the expression of the integrated GFP transgene in the different cell clones as determined by the mean fluorescence intensity of the cells in flow cytometry (Figure 1b).

Such concurring results from VITA, Southern blot, Q-PCR and flow cytometry characterize these three cell lines, which can be used for analytical purposes as a panel of control cells carrying a known range of rHIV vector integrations.

Validation of the Q-PCR method to determine VCN in hematopoietic progenitor cells

Hematopoietic progenitor cells are endowed with sufficient proliferation potential so that single progenitor cells can form visible colonies when cultured in semi-solid medium in the presence of cytokines.

Table 1 Characterization of vector copies and vector insertion sites in HT1080 clones

Clones	*VCN by Q-PCR (Taqman)*			*Insertion site analysis by VITA (n=2)*		
	n	*VCN Ave ± s.d. (extremes)*	*Nb IS*	*Size(bp)_Chrom(no.)_Position*	*Gene*	*Description function*
HT4-A	37	1.1 ± 0.2 (0.74–1.5)	1	**107_19_13093976**	**BTBD14B**	**Transcriptional regulator**
				285_8_81041615	MRPS28	Mitochondrial ribosomal protein
HT4-A2	16	2.79 ± 0.65 (1.9–3.6)	3	183_14_71097682	SIPA1L1	Signal-induced proliferation-associated protein
				107_19_13093976	**BTBD14B**	**Transcriptional regulator**
				164_3_178970571	No gene	
				136_17_10245972	MYH8	Myosin heavy chain 8, skeletal muscle
				114_17_44346886	UBE2Z	Ubiquitin-conjugating enzyme E2Z
HT4-A6	17	7.13 ± 1.32 (5.3–8.7)	8	330_17_77173163	NPLOC 4	Nuclear protein localisation 4
				63_12_61459846	PPM 1H	Protein phosphatase 1H
				68_7_19590033	No gene	
				62_7_101494262	CUX-1	Cut-like homeobox 1a
				107_19_13093976	**BTBD14B**	**Transcriptional regulator**

Abbreviations: Q-PCR, quantitative-PCR; s.d., standard deviation; VCN, vector copy numbers; n, number of Q-PCR experiments.
The bold data represent the identical insertion that is expected to be found in the three different cell lines.

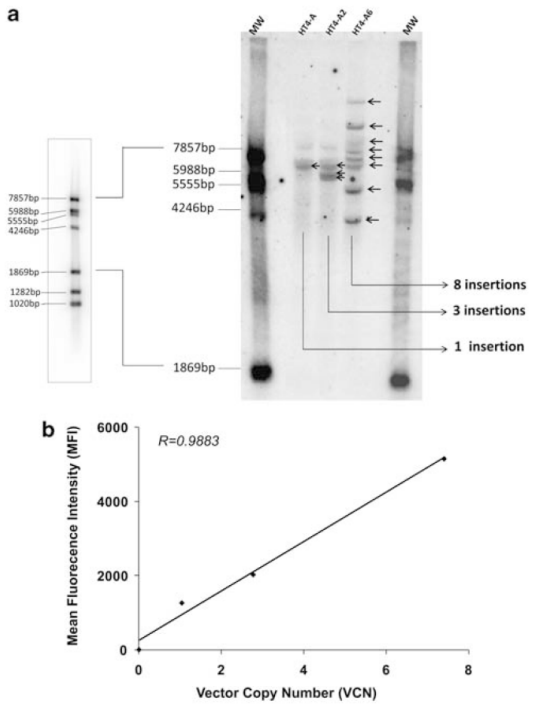


Figure 1 Characterization of the HT4-A, HT4-A2 and HT4-A6 clones. (**a**) Southern Blot on *Xba*I-digested genomic DNA from the HT1080 clones. The probes bands are the *Xba*I fragments released from an internal region to the vector and the flanking genomic DNA. (**b**) Correlation between VCN obtained by Q-PCR and MFI obtained by flow cytometry in the three clones.

The transduction of hematopoietic progenitor cells can therefore be assessed at the clonal level by measuring VCN in a single unit of CFC in which all cells originate from a single progenitor cell. This prompted us to assess if the duplex Q-PCR technique that we employed previously could be used to quantify VCN in CFC. One technical challenge is that CFC contain small numbers of cells (between 1000–10 000 cells) requiring specific DNA extraction method, therefore requiring an evaluation of the sensitivity, feasibility and robustness of the approach.

First, we determined if the sensitivity of our Q-PCR protocol would be adequate in this desired cell number range. Serial dilutions of the standard curve plasmid showed the expected amplification of HIV versus ALB or WPRE versus ALB sequences with as little as 10^2 copies of plasmid corresponding to 8.26×10^{-7} ng of DNA per reaction (see supplementary Table S1). The amplification of HIV or WPRE sequences was considered to be equivalent. Considering the size of the plasmid in relation to the size of the human genome and the amount of DNA per cell, this level would be compatible with the amplification of the integrated proviral vector DNA in approximately 100 cells. The sensitivity of the Q-PCR is therefore in principle adequate for the analysis of CFC.

Second, we determined whether we could amplify small amounts of cellular genomic DNA material under conditions compatible with the study of CFC. Extraction of genomic DNA from CFC was performed by a single-step proteinase K lysis, which is reportedly successful with small numbers of cells.[17] The effects of cell number and of the extraction conditions were tested. Genomic DNA was obtained from decreasing numbers of HT4-A, HT4-A2 and HT4-A6 cells after extraction with proteinase K in the presence or not of methylcellulose. As shown in Figure 2a, the cycle threshold (CT) for the ALB sequence increased correspondingly to the reduction in cell sample size. The lowest amount of genomic DNA that could provide an interpretable signal with a CT of 31.2 ± 0.5 (n=94) corresponded to 100 cells, thus confirming the range of predicted sensitivity of the Q-PCR. The

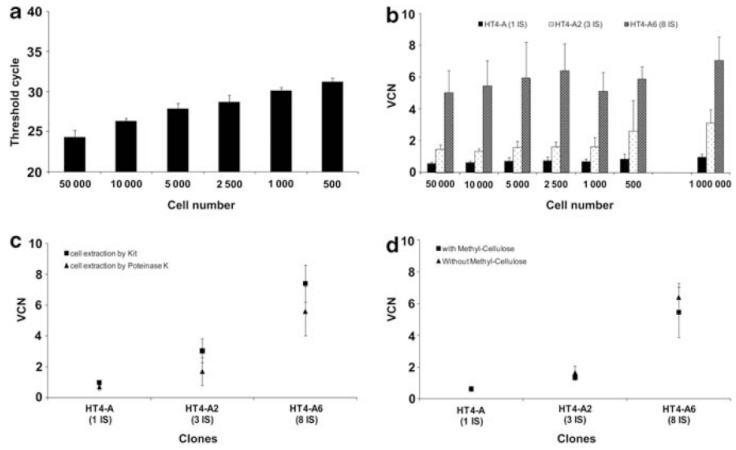


Figure 2 Determination of VCN by Q-PCR. (**a**) The sensitivity of the Q-PCR was evaluated by measuring CT values (average of duplicate measures) for the amplification of the albumin gene sequences in decreasing numbers of control cells. Q-PCR was carried out from 1/6 of the total genomic DNA extracted from 500–50 000 cells. (**b**) VCN in the HT4-A, HT4-A2 and HT4-A6 clones were measured after proteinase K lysis extractions of separate preparations of cells ranging from 500–50 000 cells per condition. The Q-PCR was performed on 1/6 of the extracted genomic DNA as in Figure 2a. The control indicated on the right side of the graph corresponds to Q-PCR on genomic DNA obtained from 1×10^6 cells extracted with a commercial Promega kit. (**c**) Comparison of VCN values obtained on the three clones using genomic DNA extracted either by proteinase K lysis from a number of cells inferior to 5×10^4 or using a commercial kit and 1×10^6 cells per extraction. (**d**) Comparison of VCN values obtained by Q-PCR from the three clones using genomic DNA extracted by proteinase K lysis in the presence (5 µl per condition) or absence of methylcellulose.

determination of VCN from variable amounts of genomic DNA obtained from each of the clones showed similar values over the range of cell-equivalent tested (Figure 2b). However, lower VCN values than expected were obtained from the Q-PCR amplification of small amounts of genomic DNA extracted with proteinase K from HT4-A, HT4-A2 and HT4-A6 cells (Figure 2b). Under these conditions, the copy number was inferior by about 30–40% from the expected value. With the six cell dilutions shown in Figure 2b, clone HT4-A gave an average of 0.7 ± 0.1 VCN (range 0.6–0.8), which is 30% less than expected value of 1; HT4-A2 gave an average of 1.7 ± 0.4 VCN (range 1.3–2.5), which is 40% less than expected value of 3; and HT4-A6 gave an average of 5.7 ± 0.5 VCN (range 5–6.4), which is 30% less than expected value of 8. To understand the origin of this suboptimal accuracy, we investigated several parameters. Copy numbers measured from genomic DNA obtained from large numbers of cells with a commercial DNA extraction kit were close to the expected values (VCN 0.97, 3.12 and 7.1 for each of the three clones) (Figure 2b controls on the right side of graph), demonstrating the accuracy of the Q-PCR amplification step in itself when it is performed on optimal material. The quality of the genomic DNA was therefore examined. The proteinase K extraction method was found to give lower VCN values than a commercial DNA extraction kit even with large numbers of cells, although this lower trend was not found to be statistically significant (Figure 2c). The presence or absence of methylcellulose with the cells did not significantly impact the values of VCN in the clones (Figure 2d). Thus, the genomic DNA extraction step in itself appears to be a critical parameter for the accuracy of VCN determination in CFC by Q-PCR. The advantage of the rapid single-step proteinase K procedure must therefore be mitigated by an underestimation of VCN values by 30–40%. However, with this limitation defined, the approach remains useful and provides dose-dependent results in order to evaluate the distribution of VCN at the clonal level.

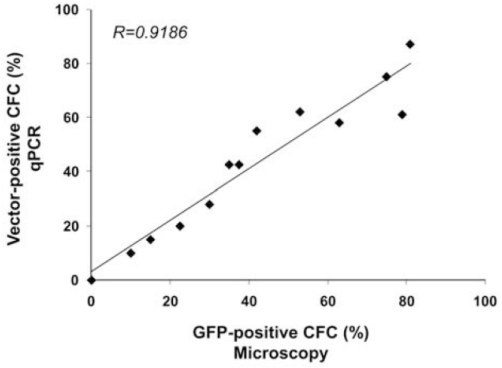


Figure 3 Correlation between GFP-positive CFC observed by microscopy and vector-positive CFC evaluated by Q-PCR (n=13 experiments).

Third, we used this Q-PCR technique to evaluate the transduction of cord blood CD34+ hematopoietic progenitor cells with various concentrations of the GFP-LV by comparing the number of vector copies in relation to transgene expression in individual CFC. Infected CD34+ cells were seeded in methylcellulose and after 2 weeks, colonies were scored under epifluorescence microscopy to determine GFP expression. Genomic DNA was extracted from each colony with proteinase K to measure vector copies by duplex Q-PCR. The colonies were scored positive for vector when they displayed a specific signal even though VCN values could be lower than one copy per cell (cut-off value 0.1). Under these conditions, a positive correlation (r^2=0.92; n=13) was found between the frequency of CFC expressing the GFP protein and scoring positive for vector by Q-PCR (Figure 3). There were only 4 ± 7% of CFC scored positive by microscopy that had no detectable vector integration by qPCR and inversely, only 6 ± 5% of CFC scored negative by microscopy that gave a positive signal by qPCR.

It has been proposed that the probability of transduction of single hematopoietic cells in a preparation can be estimated from the average mean transduction rate using the Poisson distribution analysis.[18] Assuming that each cell has equivalent probability of being transduced and that the distribution of events is not modified by cell culture, then the transduction rate can predict the percentage of cells receiving at least one copy of vector. The transduction rate can be estimated from the mean average copy number per cell in a whole population of cells. In a series of experiments, CD34+ cells were infected with the GFP-LV at 2×10^8 IG per ml giving mean average VCN of 1.4 ± 0.4 in the whole population (n=3 experiments) and in this set of experiments, the presence of vector was measured by Q-PCR on CFC (Table 2 and Figure 5a). This showed 70% vector-positive individual CFC (Table 2 and Figure 5a), which is close to the expected transduction efficiency of 65% calculated from the Poisson distribution as per Ref. 18. Similar findings were made with another vector. In six independent transductions of CD34+ cells with a WASP-LV vector, we measured a mean average value of 0.4 ± 0.1 vector copies per cell in the cultured whole cell population (data not shown). In one of these six experiments, the transduction frequency of CFC was found to be 40% by Q-PCR (Table 2), which is also very close to an expected 35% transduction frequency on the basis of Poisson distribution.

Altogether, these results experimentally validate the feasibility of using Q-PCR to quantify the frequency of transduction in single CFC.

Table 2 Effects of vector concentration on the transduction of CFC

LV vectors	*LV concentration (IG ml^{-1})*	*Number of CFC tested*	*Vector-positive CFC (%)*	*Median VCN within transduced CFC (range)*
GFP-LV (ultracentrifuged) n=3 experiments	2×10^7	120	48	0.9 (0.1–4.9)
	20×10^7	120	70	1.2 (0.1–12.7)
WASP-LV (ultracentrifuged) n=2 experiments	5×10^7	78	27	1.1 (0.2–4.4)
	10×10^7	73	32	0.9 (0.1–7.1)
WASP-LV (chromatography-purified) n=2 experiments	5×10^7	143	40	0.5 (0.1–11.9)
	10×10^7	40	50	0.7 (0.2–4.5)

Abbreviations: CFC, colony-forming cell; GFP-LV, green fluorescence protein-lentiviral vector; VCN, vector copy numbers.
These experiments are also related in Figure 5 (see Figure legend).

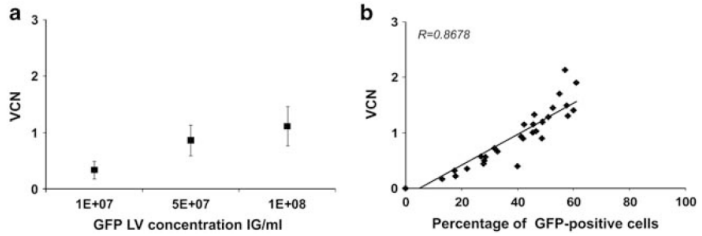


Figure 4 (**a**) Transduction of CD34+ cells with increasing concentrations of a GFP-LV tested from 0.1 to 10×10^7 IG ml^{-1} in eight independent experiments. Results show the mean average VCN determined on the cells expanded in liquid culture in the presence of cytokines for 2 weeks after genomic DNA kit extraction (mean ± s.d.). (**b**) Correlation between average VCN and the frequency of GFP expression measured by FACS in 40 experiments in which various GFP-LV vector concentrations were tested.

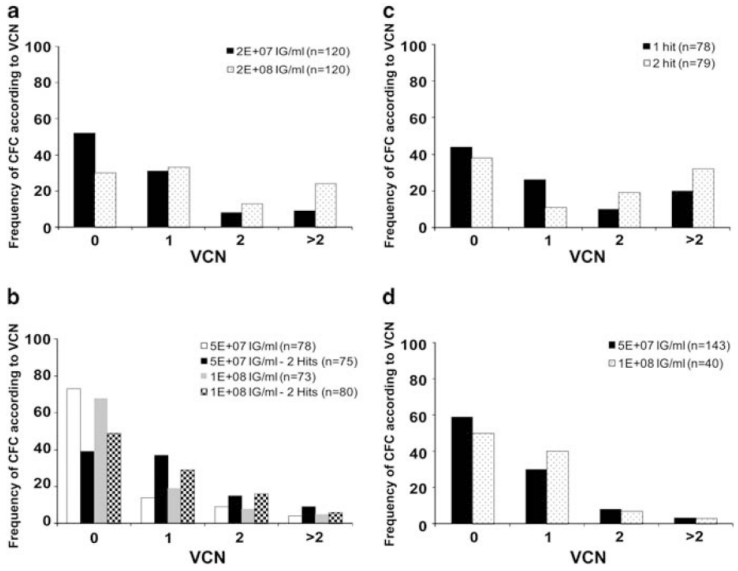


Figure 5 Effects of experimental conditions on the frequency of transduction and on the distribution of vector copies in the population. CFC with VCN values comprised between 0 and 0.1 were categorized as 0; those comprised between 0.1 and 1.1 were categorized as 1; those comprised between 1.2 and 2.1 were categorized as 2 and those with VCN superior to 2.1 were categorized as >2. Bars represent average percentage of CFC in each category over the total number of CFC analyzed. The number of CFC analyzed is indicated between brackets for each graph. (**a**) Transduction with an ultracentrifuged GFP-LV using various concentrations of vector. Results represent data pooled from three separate transduction experiments. (**b**) Transduction with an ultracentrifuged WASP-LV using several concentrations of vectors and either one or two consecutive infections (hits). Results represent data pooled from two separate transduction experiments. (**c**) Transduction with the GFP-LV at concentration of 2×10^8 IG ml^{-1} given once or twice. Results represent data pooled from two separate transduction experiments. (**d**) Transduction with two batches of chromatography purified WASP-LV using two concentrations. Data from one experiment.

Application of the Q-PCR on CFC to evaluate the effects of vector concentration and number of hits on the transduction of hematopoietic progenitor cells

Various parameters, such as the concentration of vector and the number of hits of vector, can be adapted to optimize the infection of hematopoietic progenitor cells with LV. Increasing concentrations of a GFP-LV augmented the percentage of GFP-expressing cells in the bulk population of cultured CD34+ cells as expected[19] and augmented the mean average VCN in the population (Figure 4a), thus, resulting in a good correlation between transgene expression and cell transduction (Figure 4b). At the individual progenitor cell level, increasing the concentration of LV increased the frequency of transduced CFC (Table2 and Figure 5a). In addition, this also modified the distribution of VCN in the progenitor cell population as shown by increased median values or range of VCN when using higher vector concentrations (Table 2). The frequency of CFC containing greater than two vector copies per cell was also augmented as represented by categories of frequency according to VCN in Figure 5a. Yet, in the conditions tested, the majority of transduced CFC only integrated one copy of vector per cell.

The effects of multiple rounds of infection on CFC transduction were analyzed. CD34+ cells were infected once for 6 h or twice in a consecutive manner, by adding the same concentration of vector for a second time. A second hit of infection with the GFP-LV had little impact on the frequency of vector-positive CFC as seen in Table 3, but instead, it shifted the median of VCN values from 1.4 to 2.2. This

Table 3 Effects of repeated infection on the transduction of CFC

LV vectors	*Number of hits*	*Number of CFC tested*	*Vector-positive CFC (%)*	*Median VCN within transduced CFC (range)*
GFP-LV	1	78	56	1.4 (0.1–14.5)
	2	79	62	2.2 (0.1–20.9)
WASP-LV	1	151	29	0.9 (0.1–7.1)
	2	155	53	1 (0.1–6.5)

Abbreviations: CFC, colony-forming cell; GFP-LV, green fluorescence protein-lentiviral vector; VCN, vector copy numbers.
These experiments are also related in Figure 5 (see Figure legend). Vectors were tested at the concentration of 2×108 IG ml^{-1} (GFP-LV) or $0.5\text{-}1\times108$ IG ml^{-1} (WASP-LV) and data were pooled from two separate transductions with each vector.

effect is also illustrated in Figure 5c where higher percentages of CFC are found in categories of cells with VCN equal or superior to two. The range of VCN obtained after infection with the GPF-LV is very broad, even with one hit. In contrast, performing a second hit with the WASP-LV, increased the frequency of transduced CFC but did not modify the median or range of VCN values (Table 3). However, it increased the percentage of CFC with two copies of vector (Figure 5b) without significant impact on higher categories. Thus, these results show that different types of LV have the potential to behave differently according to the experimental conditions. With the WAS-LV tested here, the transduction efficiency could be more effectively augmented by two consecutive hits of vector rather than doubling the vector concentration.

Application of the Q-PCR on CFC to evaluate the activity of a purified WASP-LV

In the perspective of clinical gene therapy studies for WAS, we have developed a large-scale purification process for the WASP-LV using chromatography and membrane steps.[20] Preparations of chromatography-purified WASP-LV were generated during this development phase and were previously reported.[21] Two batches of chromatography-purifed WASP-LV were tested here to determine the ability to transduce individual CFCs in comparison with the same vector concentrated by ultracentrifugation. Comparable levels of transduction were found between these two types of preparations as shown in Table 2. Approximately 40–50% of the CFC were transduced with the chromatography-purified vector compared with about 30% with the ultracentrifuged vector, but the median of VCN were slightly higher in the latter (0.9–1.1) than the former (0.5–0.7). Approximately 90% of the CFC that were transduced with the chromatography-purified WASP-LV had less than two vector copies per cell as shown in Figure 5d (with both concentrations tested, 3% of CFC have >2 VCN and 7–8% have two VCN). Thus, contrary to what observed with the GFP-LV, increasing the concentration of the WASP-LV improved the transduction frequency, yet, had little effect on VCN augmentation per cell. This suggests that a plateau of transduction was probably reached with the highest value of 10^8 IG per ml of this preparation of LV. In a non-mutually exclusive manner this also suggests that different preparations of LV may have different infectivity properties for hematopoietic progenitor cells.

Transduction of hematopoietic cell subsets

To assess the transduction of the different types of hematopoietic progenitor cells, we examined the effects of the GFP-LV and WASP-LV on the VCN values in each type of hematopoietic colony, combining all experimental conditions tested for each vector. This showed that the VCN were within a similar range in colony-forming units (CFU)-granulocyte, monocyte (GM) and CFU-granulocyte, erythrocyte, monocyte, megakaryocyte (GEMM) colonies and in majority inferior to four copies per cell as 83% of the CFC were transduced with the GFP-LV (Figure 6a) and 96% of the CFC were transduced with the WASP-LV (Figure 6b). However, it is also remarked that both GFP and WASP LV generate higher VCN values in erythroid colonies (colony-forming unit, erythroid (CFU-E)/burst-forming unit, erythoid (BFU-E)) than in myeloid and/or mixed colonies (CFU-GEMM). The difference between VCN values in erythroid colonies versus other categories alone or combined, is statistically significant ($P<0.05$, Student's t test).

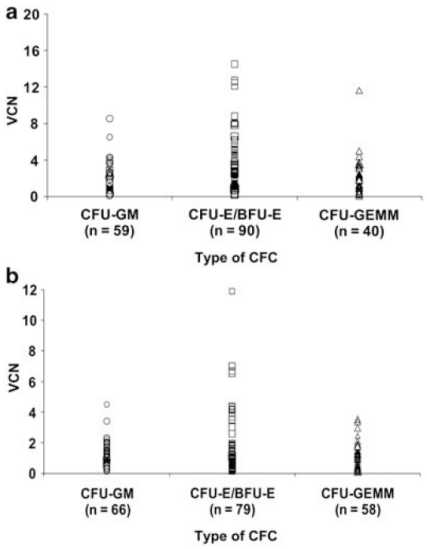


Figure 6 Distribution of the VCN in the different type of the CFC (CFU-GM, CFU-mix and BFU-E) from CD34+ cells transduced by (**a**) GFP-LV or (**b**) by WASP-LV.

DISCUSSION

We herein describe and validate an analytical method to measure rHIV VCN in human CFC, providing experimental data on the transduction of hematopoietic progenitor cells, in particular with a relevant WASP vector.

The method described in this paper is simple and rapid, comprising a single-step extraction of genomic DNA followed by a duplex Q-PCR to amplify the vector and cellular sequences simultaneously. This simplicity presents an advantage over previously-published protocols that analyze the presence of gene transfer vectors in hematopoietic colonies with protocols combining cell lysis, DNA extraction with phenol chloroform or isopropanol, PCR amplification and agarose gene analysis.[4,13] More recent protocols combine these DNA extraction methods with Q-PCR analysis,[14] but to our knowledge, without being validated experimentally. Thus, we herein show that a simple protocol can be used and is sufficiently sensitive to reliably determine the frequency of transduced CFC according to expected values. Following transduction of CD34+ cells with a GFP-LV, there is a good correlation between the frequency of PCR-positive CFC and the expression of the transgene. In addition, the transduction frequency of CFC is coherent with values calculated from the average copy number in the bulk population of CD34 cells using Poisson's distribution of

single events.[18] The precision of the method is comparable with that of Q-PCR performed in standard conditions. Indeed, comparable standard deviations are found on the three control cell lines whether using genomic DNA extraction kits and large amounts of cell material or in conditions mimicking those used for CFC (see Figures 2b and c). The method is therefore simple, sensitive and precise. However, the simplicity of the method must be mitigated by a suboptimal accuracy. Testing the three control characterized cell lines in the same conditions as CFC, we find a 30–40% underestimation of VCN values compared with the expected values. This underestimation is not caused by insufficient quantities of genomic DNA as the Q-PCR is sensitive within the range of cells analyzed. The limitation is probably caused by insufficient quality of the genomic DNA, which prevents the optimal amplification of vector-specific sequences, but the presence of methylcellulose can be excluded as a factor. In spite of this suboptimal accuracy, VCN results obtained with this simple method provide coherent results. Indeed, the VCN obtained on the three control cell lines reflect the expected range of rHIV copies inserted in these cells. The distribution of VCN in individual CFC is consistent with expected values from an idealized Poisson's distribution.[18] Theoretical calculations predict that when a mean vector copy number is inferior to two in a cell population, then among transduced cells the majority of individual cells should contain one copy of vector. This distribution was obtained with the GFP-LV or with the WASP-LV and indeed we measured that the majority (about 60 %) of the CFC contained one copy. Also, with mean transduction rates inferior to one, an idealized distribution would predict that less than 10% of individual cells should contain more than two copies per cell and this is also what we measured in experiments using the chromatography-purified WASP-LV. Thus, in spite of a slight underestimation of the VCN values in CFC, the distribution of cells according to VCN categories appears to be consistent with expected data from theoretical calculations. To take into account this possible underestimation, one could apply a corrective factor on the basis of the 30–40% underestimation that was observed with the three control cell lines using this technique. Altogether, our data show that at this point, the Q-PCR method is sufficiently sensitive, precise and acceptably accurate to provide a meaningful measure of the frequency of vector-positive CFC and the distribution of vector copies at the clonal level within a population of hematopoietic progenitor cells.

Further effort outside the scope of this article are needed to improve the accuracy of this technique by optimizing the DNA extraction step or the performance of amplification of rHIV sequences in proteinase K-extracted DNA. In addition, this simple technique could be adapted to high-throughput analysis to obtain data on large numbers of clones. Although the use of CFC is a very practical manner to obtain clones of hematopoietic cells, it is strongly biased for cells of the myelo-erythroid lineage. Further development could be envisioned with the analysis of single cell sorted cultures, which could enable the evaluation of transduction at the single cell level in various hematopoietic cell lineages.

Measuring the number of vector copies in individual target cells is important to assess the therapeutic window of integrative vectors. Their biological potency but also their inherent risk of genotoxicity is related to the number of vector copies integrated per cell. Determining the VCN in hematopoietic progenitor cells in relation to gene expression in these cells or their progeny, could be used to gauge the biological activity of the integrated cassette and may reveal potential occurrence of gene silencing as suggested by some studies.[8] In order to evaluate the genotoxic potential of vectors, the Q-PCR on CFC method could be used in combination with an analysis of vector insertion sites. Three characterized human cell lines, which we have been derived from HT1080 cells, could be relevant controls for both types of analytical measures. These cells contain known numbers of rHIV integration, which are in the range of events expected to occur following transduction of HSC. In addition, these cells contain known sites of rHIV integration, which could also serve as references for quality control and validation of protocols aiming to identify vector insertion sites.

Following transduction of CD34+ cells with the two different LVs tested in our study, revealed a slight bias in the distribution of VCN in the different types of progenitor cells. BFU-E seemed to integrate more copies than CFU-GM or CFU-GEMM. As erythroid cells arise from primitive cells such as CFU-GEMM, this would indicate that the experimental conditions favoured the preferential transduction of erythroid-restricted progenitor cells. This is not unexpected as the transduction medium contained stem cell factor (SCF) and thrombopoietin (TPO), which may provide a strong erythroid progenitor cell stimulus. Alternatively, we cannot exclude that the measure of VCN is not as accurate or precise in erythroid colonies as in other cells. Erythroid-restricted colonies are composed of cells that start to compact their chromatin in the course of their development. Further studies are needed to address this point more thoroughly, in particular, in the context of gene therapy studies targeting erythroid progenitor cells.

There is a strong interest in optimizing the conditions for *ex vivo* hematopoietic gene transfer in the perspective of experimental or clinical applications. Several previous studies have optimized parameters such as cell concentration, cytokines, medium, timing, concentration of vector, multiplicity of infection or number of transduction hits with rHIV vectors. In many cases, such studies were performed with GFP-encoding LV, relying on transgene expression levels[8,19,22] but providing little to no information on the amount of vector integrated in target cells or the frequency of the targeted population. In this paper, we show that the Q-PCR analysis of CFC is applicable to evaluate the infectivity of CD34+ cells with LV in a transgene-independent manner. Increasing the concentration of infectious vector or repeating the infection were two different but effective ways to augment the level of transduction as measured by the percentage of vector-positive cells and by the distribution of VCN within transduced cells. In some experiments, repeating the infection seemed more effective than increasing the vector concentration to augment the percentage of vector-positive colonies as seen in Figure 5b. This indicates that clonogenic cells could become more permissive to rHIV transduction during the *ex vivo* culture. Further optimization could be undertaken to increase this effect and improve the levels of transduction. On the other hand, these results reveal that two consecutive hits can augment the frequency of progenitor cells with more than two copies per cell. Therefore a repeat transduction strategy may become toxic to a fraction of the cells and this should be carefully tested. The effect of repeated hits on high VCN was clearly seen with the ultracentrifuged vectors but not with the chromatography-purified batches. This information cannot be obtained from gene expression analyses based on average numbers. To our knowledge, this is the first time that transduction efficiency is documented from quantitative measures of VCN at the single-cell level. Thus, our results strongly suggest that different preparations could behave differently in pharmacological terms, for infection of hematopoietic progenitor cells resulting in variable levels of vector copies per cell. The process of transduction is complex, but the mechanisms behind the observed differences in VCN may involve viral binding or entry as well as proviral integration capacity.

In the perspective of clinical gene therapy studies for WAS, we have developed a large-scale purification process for rHIV LV as described in Ref. 21, and more extensively in Ref. 20. The chromatography-purified WASP-LV has showed an acceptable safety profile in pre-clinical tests, notably with respect to hematopoietic progenitor cell survival and differentiation, as confirmed here. The CFC transduction levels obtained with this batch of vector were similar to those obtained with the same WASP-LV purified by ultracentrifugation. Transduction is effective (about 40–50% of the CFC can be transduced), whereas providing only a low number of copies (60–85% of CFC contain 1 copy per cell and >90% of the CFC contain equal or less than two copies per cell), even after repeat infection. If we take into account that there is possibly an underestimation factor for VCN in CFC and apply a 30–40% correction, then >90% of the cells would have no more than three copies per cell. Thus, although a probability exists that some cells will integrate high VCN, our data suggest that the chromatography-purified WASP-LV provides a relatively safe level of transduction of hematopoietic cells.

In conclusion, the analysis of vector copy number in individual CFC with Q-PCR determines the frequency and the distribution of vector copies in the population of hematopoietic progenitor cells that were initially targeted by the vector. It is clear that these values are influenced by the experimental transduction conditions and by the type of vector tested. Such data are important to optimize preclinical and clinical transduction protocols with LVs for *ex vivo* hematopoietic gene therapy applications.

MATERIALS AND METHODS

Generation and titration of lentiviruses

The PGK-GFP/VSVg rHIV vector (GFP-LV) or w1.6hWASP/VSVg rHIV vector (WASP-LV) were produced by transient quadri-transfection of 293T cells and were purified by ultracentrifugation as previously reported.[12] In some experiments a WASP-encoding LV produced similarly by transient transfection was concentrated and purified through a series of chromatography and membrane steps as reported.[20] Vectors were titered as infectious genomes (IG) per ml on HCT116 cells using duplex Q-PCR as previously described[12] and were also titered for p24 levels using an ELISA (Perkin Elmer, Waltham, MA, USA). In some experiments, vectors encoding the GFP were titered by flow cytometry instead, and results were multiplied by a factor 2 to match with IG values, as determined by repeated comparisons between the two titration methods (data not shown). Vector batches used in the study are listed in supplementary Table S2.

Clones derived from HT1080 cells

The fibrosarcoma HT1080 cells originated from ATCC (CCL-121, American Type Culture Collection, Manassas, VA, USA) were grown in DMEM supplemented with glutamine and antibiotics, and containing 10% fetal calf serum. HT1080 cells were transduced with a rHIV encoding the GFP under control of the human phosphor-glycerate kinase promoter kindly provided by Dr L Naldini (Tiget, Milan, Italy). Expression of GFP was measured by flow cytometry. Using the same culture medium, the cells were first cloned by fluorescence-activated cell sorting and then a second cloning was performed by limit dilution culture. One clone (HT4-A) containing one copy was identified and further transduced with the same vector to generate subclones. Such re-infected HT4-A cells were then cloned by limit dilution and wells were screened by fluorescent microscopy and flow cytometry to identify cells presenting different mean fluorescence intensity for GFP expression. Two daughter clones (HTA-A2 and HT4-A6) expressing different levels of fluorescence were selected and characterized by Q-PCR, Southern blot and sequencing.

Southern Blot

The number of rHIV insertions in HT4, HT4-A2 and HT4-A6 cells was determined by Southern blot using genomic DNA extracted by chloroform/phenol extraction with AutoGen extractor, NA2000 (Geneworx, Blénod lès Pont à Mousson, France) after proteinase K (Invitrogen, Cergy Pontoise, France) lysis. A measure of 40 µg of genomic DNA of each cell sample was digested with 70 units of *Xba*I (Gibco BRL/Invitrogen, Cergy Pontoise, France) at 37 °C over night and the digested DNA samples were subjected to an 0.8% agarose gel electrophoresis. The DNA fragments in the gel were transferred onto a Nylon membrane (Amersham, Piscataway, NJ, USA). The membrane was hybridized with 50% deionized formamide over night at 42 °C with a 2185 bp fragment of the integrative lentiviral DNA obtained by AflII-BamHI digestion of the plasmid pCCLsincppthPGK-GFP-WPRE and labeled with [α-^{32}P]dATP using the Prime-it Random Primer Labeling kit (Stratagene, Lyon, France). The final wash was preformed with 0.1× SSC-1% SDS at 68 °C for 1 h. After 10 days of exposure on a phosphor screen (Molecular Dynamics PhosphorImager System; GE Healthcare Bio-Sciences, Piscataway, NJ, USA), radioactive bands were revealed by the Storm system (GE-Healthcare Bio-Sciences).

Analysis of vector genomic insertion sites

rHIV insertion sites were determined following NlaIII digestion of the genomic DNA of HT4-A, HT4-A2 and HT4-A6 cell clones as described previously.[16] Sequencing reactions were performed using Big Dye terminator sequencing chemistry (Applied Biosystems, Carlsbad, CA, USA) from the M13 forward or M13 reverse primers and run on a 377-XL Applied Biosystems automated sequencer. Junctions obtained were matched on the human genome using the BLAT program (UCSC human Genome Working draft (March 2006 (NCBI36/hg18). The description of the human transcripts was obtained from ftp://hgdownload.cse.ucsc.edu.

Human CD34+ cells source, transduction and culture

Umbilical cord blood progenitor CD34+ cells were obtained by immunomagnetic selection (Miltenyi Biotec, Paris, France) from mononuclear cell fractions of cord blood samples obtained from uncomplicated births at Hopital Louise Michel, Evry, France, in compliance with French national bioethics law. Cells were first pre-activated by culturing overnight 5×10^4 cells in 0.2 ml of X-vivo20 medium (Lonza, Levallois Perret, France) supplemented with 50 U ml^{-1} penicillin, 50 µg ml^{-1} streptomycin and 2 mM L-glutamine (Gibco BRL/Invitrogen), SCF (25 ng ml^{-1}), Flt-3 ligand (50 ng ml^{-1}), TPO (25 ng ml^{-1}), IL-3 (10 ng ml^{-1}) (R&D Systems, Lille, France). Pre-activated cells were then infected with LV using different concentration of vectors ranging from 2×10^7 IG per ml to 2×10^8 IG per ml for 6 h in the presence of polybrene (6 µg ml^{-1}). When cells were re-infected, the second hit of vector was added overnight. At the end of the transduction step, cells were washed and either plated in semi-solid medium for CFC assays or cultured in the presence of cytokines for *in vitro* expansion. For *in vitro* expansion, the transduced CD34+ cells were seeded at 5×10^4 cells per ml in 24-well flat-bottom plates in X-vivo20 medium supplemented with 10% fetal calf serum (Gibco), penicillin/streptomycin, L-glutamine and the same recombinant cytokines as used for transduction. Fresh medium was added every 3 days. Cultures were incubated at 37 °C 5% CO_2 for 14 days.

CFC assays

CFC assays were performed in duplicate by plating 1000 transduced or untransduced cells per ml of Methocult (H4434), a complete methylcellulose medium supplemented with human cytokines (Stem Cell Technologies, Vancouver, CA, USA) according to the manufacturer's instructions. After 2 weeks of culture, 5% CO_2, 37 °C in humid atmosphere, CFU-E/BFU-E, CFU-GM and CFU-GEMM were counted by inverse microscopy using standard visual criteria.

Q-PCR and analysis of CFC

Well-isolated colonies were aspirated carefully with a pipette tip under the miscroscope and the cells were suspended into 100 µl of phosphate-buffered saline in a 96-well U-bottom plates. The plates were centrifuged at 1500 r.p.m. for 10 min. Green fluorescent (GFP+) colonies were identified by fluorescence microscopy. Medium was aspirated and the cells pelleted from each CFC unit were suspended in 10 µl of phosphate-buffered saline. Genomic DNA was extracted from these pellets using proteinase K lysis consisting in adding 20 µl of lysis buffer (0.3 mM Tris HCl, pH 7.5; 0.6 mM $CaCl_2$; 1.5 % Glycerol; 0.675% Tween-20; and 0.3 mg ml^{-1} Proteinase K) to each well and incubating the plate at 65 °C for 30 min, 90 °C for 10 min and at 4 °C to end the reaction for a minimum of 10 min. After lysis, 30 µl of water was added to each well to obtain

60 µl of genomic DNA preparation. In all experiments using this technique, the Q-PCR was always carried out from 10 µl of this genomic DNA preparation (that is, amplifying 1/6 of the extracted material), and in duplicate.

The Q-PCR consisted of a duplex detection of WPRE or HIV sequences normalized to ALB, as described for titrations. The probes were conjugated to FAM for HIV or WPRE sequences and to VIC for Albumin. Amplification reactions (25 µl) contained 10 µl of genomic DNA and 15 µl of TaqMan buffer (Absolute Q-PCR Rox Mix, ABgene AB-1139/B), 0.1 µM primers (forward and reverse), 0.1 µM TaqMan probe and consisted of 40 cycles at 95 °C (15 s) then 60 °C (1 min) on an ABI PRISM 7700 sequence detector (Applied Biosystems). Standard amplification curves were obtained by serial dilutions of the pRRLcpptPGKGFP-WPRE-Alb plasmid containing the appropriate sequences in cis from the vectors and *ALB* gene. All PCR measures were performed at least in duplicate. All Q-PCR experiments on CFC include samples from untransduced CFC as negative controls.

Data were edited using the Primer Express software (Applied Biosystems). Data were interpreted in the linear portion of the standard curve. Linear regression coefficient of the standard curve should be >0.99. In this portion of the curve the ratio between HIV/ALB of the plasmid standard is equivalent to 1 ± 0.2 (n=28). The detection's threshold is determined with this standard. When the CT of Albumin is superior to 32 cycles, the ratio HIV/ALB is different from the value of 1 ± 0.2 defined in the linear portion of the curve. The duplicate CT should vary by less than 0.5 CT. The ALB and HIV or WPRE CT of H_2O should locate between 35 and 40. HIV CT of untransduced cells should be like H_2O. ALB CT of untransduced cells and transduced samples should be around 23–28. The amount of WPRE/ALB or HIV/ALB is equimolar to the amount of pRRLcpptPGKGFP-WPRE-Alb plasmid. However, because in diploid cells, there are two molecules of ALB per cell, the number of integrated VCN per cell is determined by multiplying the ratio WPRE/ALB or HIV/ALB by two. Results of VCN inferior to 0.1 copy per cell were considered to be negative.

Preparation of genomic DNA from cells other than CFC for Q-PCR

Genomic DNA from cell lines or from bulk CD34+ were extracted with the 'Wizard genomic DNA purification kit' (Promega Corporation, Madison, WI, USA) or with a single-step proteinase K lysis as described above.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are very grateful for technical help from Khalil Seye and Gregory Cedrone, and thank Nicolas Laroudie and the vector production group for providing vector batches and in particular the chromatography-purifed R&D batches. We also thank Guillaume Sirantoine for sequencing the vector insertion sites, Daniel Stockholm and William Vainchenker for their input with methods and Thierry Larmonnier and Lucia Braga-Vacherie from the Genethon Biological Resources Center for help with cell line banking. We are also very grateful to the staff of the Maternité de l'hôpital Louise Michel, Evry, France for providing umbilical cord samples and to AFM (Association Francaise contre les myopathies) and the EC HEALTH-FP6 integrated project CONSERT for sponsoring this project.

1 Fischer A, Cavazzana-Calvo M. Gene therapy of inherited diseases. *Lancet* 2008; **371**: 2044–2047.
2 Piacibello W, Bruno S, Sanavio F, Droetto S, Gunetti M, Ailles L *et al.* Lentiviral gene transfer and *ex vivo* expansion of human primitive stem cells capable of primary, secondary, and tertiary multilineage repopulation in NOD/SCID mice. Non-obese diabetic/severe combined immunodeficient. *Blood* 2002; **100**: 4391–4400.
3 Enssle J, Trobridge GD, Keyser KA, Ironside C, Beard BC, Kiem HP. Stable marking and transgene expression without progression to monoclonality in canine long-term hematopoietic repopulating cells transduced with lentiviral vectors. *Hum Gene Ther* 2010; **21**: 397–403.
4 Woods NB, Fahlman C, Mikkola H, Hamaguchi I, Olsson K, Zufferey R *et al.* Lentiviral gene transfer into primary and secondary NOD/SCID repopulating cells. *Blood* 2000; **96**: 3725–3733.
5 Cartier N, Hacein-Bey-Abina S, Bartholomae CC, Veres G, Schmidt M, Kutschera I *et al.* Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* 2009; **326**: 818–823.
6 Galy A, Roncarolo MG, Thrasher AJ. Development of lentiviral gene therapy for Wiskott Aldrich syndrome. *Expert Opin Biol Ther* 2008; **8**: 181–190.
7 Woods NB, Muessig A, Schmidt M, Flygare J, Olsson K, Salmon P *et al.* Lentiviral vector transduction of NOD/SCID repopulating cells results in multiple vector integrations per transduced cell: risk of insertional mutagenesis. *Blood* 2003; **101**: 1284–1289.
8 Liu Y, Hangoc G, Campbell TB, Goodman M, Tao W, Pollok K *et al.* Identification of parameters required for efficient lentiviral vector transduction and engraftment of human cord blood CD34(+) NOD/SCID-repopulating cells. *Exp Hematol* 2008; **36**: 947–956.
9 Chang AH, Sadelain M. The genetic engineering of hematopoietic stem cells: the rise of lentiviral vectors, the conundrum of the ltr, and the promise of lineage-restricted vectors. *Mol Ther* 2007; **15**: 445–456.
10 Modlich U, Bohne J, Schmidt M, von Kalle C, Knoss S, Schambach A *et al.* Cell-culture assays reveal the importance of retroviral vector design for insertional genotoxicity. *Blood* 2006; **108**: 2545–2553.
11 Mohamedali A, Moreau-Gaudry F, Richard E, Xia P, Nolta J, Malik P. Self-inactivating lentiviral vectors resist proviral methylation but do not confer position-independent expression in hematopoietic stem cells. *Mol Ther* 2004; **10**: 249–259.
12 Charrier S, Dupre L, Scaramuzza S, Jeanson-Leh L, Blundell MP, Danos O *et al.* Lentiviral vectors targeting WASp expression to hematopoietic cells, efficiently transduce and correct cells from WAS patients. *Gene Ther* 2007; **14**: 415–428.
13 Ailles L, Schmidt M, Santoni de Sio FR, Glimm H, Cavalieri S, Bruno S *et al.* Molecular evidence of lentiviral vector-mediated gene transfer into human self-renewing, multipotent, long-term NOD/SCID repopulating hematopoietic cells. *Mol Ther* 2002; **6**: 615–626.
14 Schuesler T, Reeves L, Von Kalle C, Grassman E. Copy number determination of genetically-modified hematopoietic stem cells. *Methods Mol Biol* 2009; **506**: 281–298.
15 Chen TR, Hay RJ, Macy ML. Intercellular karyotypic similarity in near-diploid cell lines of human tumor origins. *Cancer Genet Cytogenet* 1983; **10**: 351–362.
16 Mantovani J, Charrier S, Eckenberg R, Saurin W, Danos O, Perea J *et al.* Diverse genomic integration of a lentiviral vector developed for the treatment of Wiskott-Aldrich syndrome. *J Gene Med* 2009; **11**: 645–654.
17 Rook MS, Delach SM, Deyneko G, Worlock A, Wolfe JL. Whole genome amplification of DNA from laser capture-microdissected tissue for high-throughput single nucleotide polymorphism and short tandem repeat genotyping. *Am J Pathol* 2004; **164**: 23–33.
18 Fehse B, Kustikova OS, Bubenheim M, Baum C. Pois(s)on–it's a question of dose. *Gene Ther* 2004; **11**: 879–881.
19 Haas DL, Case SS, Crooks GM, Kohn DB. Critical factors influencing stable transduction of human CD34(+) cells with HIV-1-derived lentiviral vectors. *Mol Ther* 2000; **2**: 71–80.
20 Merten OW, Charrier S, Laroudie N, Fauchille S, Dugué C, Jenny C *et al.* Large-scale manufacture and characterisation of a lentiviral vector produced for *ex vivo* gene therapy application. *Hum Gene Ther* 2010 Nov 2. [Epub ahead of print] PMID: 21043787.
21 Denard J, Rundwasser S, Laroudie N, Gonnet F, Naldini L, Radrizzani M *et al.* Quantitative proteomic analysis of lentiviral vectors using 2-DE. *Proteomics* 2009; **9**: 3666–3676.
22 Millington M, Arndt A, Boyd M, Applegate T, Shen S. Towards a clinically relevant lentiviral transduction protocol for primary human CD34 hematopoietic stem/progenitor cells. *PLoS One* 2009; **4**: e6461.

Supplementary Information accompanies the paper on Gene Therapy website (http://www.nature.com/gt)

Supplemental Table S1: Sensitivity of the Q-PCR method

		HIV and Albumin detection by duplex Q-PCR			WPRE and Albumin detection by duplex Q-PCR		
Number of copies of the standard curve plasmid	**Quantity of DNA (ng) per reaction**	**ALB CT**	**HIV CT**	**HIV/ALB Ratio**	**ALB CT**	**WPRE CT**	**WPRE/ALB Ratio**
10^7	$8.26x10^{-2}$	14.9 ± 0.6	13.9 ± 0.6	0.9 ± 0.1	14.9 ± 0.6	14.6 ± 0.6	1.2 ± 0.2
10^6	$8.26x10^{-3}$	18.5 ± 0.9	17.3 ± 0.8	1 ± 0.1	18.3 ± 0.6	18.1 ± 0.6	1.1 ± 0.1
10^5	$8.26x10^{-4}$	22 ± 0.6	20.6 ± 0.6	1.1 ± 0.1	21.7 ± 0.8	21.8 ± 0.7	0.9 ± 0.1
10^4	$8.26x10^{-5}$	25.2 ± 0.7	23.8 ± 0.7	1.1 ± 0.1	24.9 ± 0.9	25.2 ± 0.9	0.7 ± 0.1
10^3	$8.26x10^{-6}$	28.8 ± 0.7	27.4 ± 0.9	1 ± 0.2	28.4 ± 0.8	28.5 ± 0.6	0.9 ± 0.2
10^2	$8.26x10^{-7}$	32 ± 1	30.6 ± 0.8	1 ± 0.2	31.5 ± 0.8	31.3 ± 0.6	1 ± 0.4
H_2O	0	39.5 ± 1.5	35.2 ± 2.4	NA	38.5 ± 2.2	35.8 ± 1.9	NA

Supplemental Table S2: Characteristics of the lentiviral vector batches used in the study.

Batch	Transgene	Type of purification	Transgene	Infectious titer	p24
1	GFP	Ultracentrifugation	GFP	6.5×10^9 TU/ml	4.4×10^4 ng/ml
2	GFP	Ultracentrifugation	GFP	7.5×10^9 TU/ml	1.4×10^5 ng/ml
3	GFP	Ultracentrifugation	GFP	3.2×10^9 IG/ml	3×10^4 ng/ml
4	WASP	Ultracentrifugation	WASP	8.2×10^9 IG/ml	Not done
5	WASP	Ultracentrifugation	WASP	2.3×10^{10} IG/ml	4.7×10^4 ng/ml
6	WASP	Chromatography	WASP	2.1×10^8 IG/ml	4.9×10^4 ng/ml
7	WASP	Chromatography	WASP	2.1×10^8 IG/ml	7.6×10^3 ng/ml

Titer determined as TU/ml by FACS or as IG/ml by qPCR; 1 IG = 2 TU

PART II: Pl.LSU/2 GROUP II INTRON CHARACTERIZATION

1 - INTRODUCTION

As mentioned in introduction, the development of new tools for site-specific genomic engineering is an active research area in the field of gene therapy. Mobile group II introns genetic elements represent an attractive strategy, as they can integrate into a specific DNA site by the homing mechanism, which is not based on a double-strand break in contrast to all currently used tools.

To date, group II introns as gene targeting vectors are only used in prokaryotes, due to their inefficacy in eukaryotes. In the attempt of developing an efficient group II intron as a gene targeting tool in human cells, we decided to evaluate and characterize the *Pylaiella littoralis* Pl.LSU/2 group II intron. This intron was indeed shown to efficiently self-splice *in vitro* at low Mg^{2+} concentrations (Costa M et al. 1997b). The major hurdle of the use of group II introns in eukaryotes, and in particular in human cells, appears to be due to at least unfavorable ionic environment that impede intron catalytic activity (Mastroianni M et al. 2008; Zhuang F et al. 2009b). We thus postulated that Pl.LSU/2 group II intron, which is active *in vitro* even under remarkably stringent ionic conditions, could be more efficient in eukaryotic cells than other group II introns.

The Pl.LSU/2 intron contains in its domain IV an ORF encoding a protein that presents all group II intron-encoded protein (IEP) conserved domains (Fontaine JM et al. 1995; Fontaine JM et al. 1997). However, no report on the biochemical activities of this protein was available at the beginning of the work. The first aim of the project was thus to characterize the biochemical activities of the Pl.LSU/2 IEP. Group II intron IEPs are required to the intron homing into target DNA. It was thus crucial to determine if the Pl.LSU/2 could be active. To perform biochemical analyses on Pl.LSU/2, it was necessary to produce and purify this protein. Several strategies were thus attempted, such a chromatography purifications and centrifugation in a sucrose cushion. Much effort was made to purify the protein by chromatography, because this system is easily scalable so that various studies could be considered, such as biochemical and structural studies. The results are described in the following section. The purification of Pl.LSU/2 IEP by sucrose centrifugation was successful and allowed the characterization of the reverse transcriptase activity of the protein. These results are included in article 2.

To further characterize both the Pl.LSU/2 intron and IEP, we evaluated the splicing capacity of the Pl.LSU/2 intron in eukaryotes. Indeed, in an ideal gene targeting system based on group II intron, the intron would be expressed directly into the cells and would splice in the nucleus. The expression of the IEP in *trans*, addressed to the nucleus, would then allow the RNP formation and the homing mechanism. Both components of this system could be delivered with the use of non-integrating virus-derived vectors. In this context, the characterization of the *in vivo* Pl.LSU/2 splicing ability was a prerequisite to the development of Pl.LSU/2 intron as gene targeting tool. Because of the facility to perform genetic and mechanistic studies in *Saccharomyces cerevisiae*, we first evaluated both the splicing capacity of intron Pl.LSU/2 and the maturase activity of the IEP in that host. We then assayed the intron catalytic activity in a human cell line. The results are described in article 2.

Finally, we attempted to evaluate the homing capacity of Pl.LSU/2 in its natural target site first in *E. coli* and then in *S. cerevisiae*. The results of these preliminary assays are described in section 4 -.

2 - DEVELOPMENT AND OPTIMIZATION OF PURIFICATION STRATEGIES FOR PL.LSU/2 INTRON-ENCODED PROTEIN

The Pl.LSU/2 intron-encoded protein (IEP) sequence analysis reveals the presence of conserved catalytic domains shared by other group II intron-encoded proteins: reverse transcriptase (RT), maturase (X) and endonuclease (En). The first putative activity of the Pl.LSU/2 IEP that we evaluated was the reverse transcriptase. Indeed, several published methods that allow the testing of an RT activity are available and easily realizable.

Production of proteins, whether for biochemical analysis, therapeutics or structural studies, requires the success of three individual factors: expression, solubility and purification. Although a number of expression hosts are available for protein production, the standard still remains *E. coli* (Baneyx F 1999; Goulding CW and Perry LJ 2003). However, the percentage of soluble heterologous proteins expressed in *E. coli* is usually less than 23% (Chambers SP et al. 2004; Marblestone JG et al. 2006). There is a general perception that solubility problems can often be solved by using eukaryotic hosts, such as insect cells (with the baculovirus expression system), yeast or mammalian. Another promising alternative is the cell-free protein synthesis, which has been improved dramatically in recent years. In this work, I have used *E. coli*, insect cells and cell-free expression systems in order to express the Pl.LSU/2 IEP in a properly folded and soluble form.

Unlike many other enzymes, the RT activity of a protein could not be assayed directly from cell extracts. This is due to the fact that unfractionated extracts are likely to contain contaminating RNases and DNases naturally produced by the host in which the expression is performed. The RNases could degrade the RNA template required for the cDNA synthesis during the reverse transcriptase activity assay and the DNases could degrade the cDNA produced, thus biasing the reaction analysis. Moreover, cells extracts may also contain contaminating RNA-dependent DNA polymerases which could also lead to misinterpretation of the results. Purification of the IEP from these contaminants is therefore necessary before assaying its RT activity.

Different purification approaches can be considered to purify a protein. The most used purification method is the chromatography. However, almost all proteins lose their activity during manipulation. It is thus important to purify the protein as quick as possible. Highly specific methods, such as those based on bioaffinity (antibody-antigen interaction) or those based on the use of fusion tags such as 6xHis or glutathione-S-transferase (GST), allow in some cases the purification of a highly pure protein in a single step. As no antibody directed against the Pl.LSU/2 IEP is yet available, we decided to use both GST and 6xHis (metal binding) as tags for IEP purification.

2.1 - GST-TAGGED IEP IN *E. COLI*

The GST protein is a 26-kDa eukaryotic protein, which is well expressed in *E. coli* and was shown to improve the solubility and enhance the expression of some target proteins (Smith DB and Johnson KS 1988; Kim S and Lee SB 2008). GST has a biospecific affinity for glutahione ligand and can thus bind to resin-immobilized glutathione. The GST tag has to be properly folded to bind glutathione, thus the fusion protein needs to be soluble and in non-denaturing conditions for efficient purification. GST-tagged proteins can be eluted under mild conditions using free reduced-glutathione at neutral pH.

The plasmid pGST-IEP (See plasmid map), derived from pGEX-4T1 (See Appendix plasmid map), was constructed to express the GST-IEP fusion protein. The GST protein is less efficient at improving protein solubility when positioned at the C-terminal end of the target protein even if it still functions as an affinity tag (Smith DB and Johnson KS 1988). We have thus tagged the IEP in N-terminal to ensure maximal improvement of the solubility. The IEP sequence was inserted in frame with the GST just downstream of a thrombin recognition site (Fig. R-1; T). The Thrombin protease can be used to release the IEP from the GST. GST-IEP expression is driven by an IPTG (Isopropyl β-D-1 thiogalactopyranoside)-inducible *tac* promoter (Fig. R-1; *tac*). The translation initiation codon is located upstream the GST ORF (Fig. R-1; ATG), and a ribosome binding site (Fig. R-1; RBS) is present to allow efficient translation of the fusion protein.

Figure R-1: Schematic representation of GST-IEP expression cassette.

tac: *tac* promoter; RBS: Ribosome Binding Site; ATG: translation initiation codon; GST: Glutathione-S-transferase coding sequence ; T: Thrombin protease recognition site; Term: transcription terminator.

2.1.1 - Expression in BL21 Star (DE3) and purification

The GST-IEP protein fusion protein was first expressed in the *E. coli* strain BL21 Star (DE3), available in the laboratory. BL21 is a protease-deficient strain engineered to maximize expression of full-length protein. Its derivative BL21 Star strain contains a mutation in the gene encoding RNase E (*rne*131), which is one of the major sources of mRNA degradation (Kido M et al. 1996; Lopez PJ et al. 1999). *E. coli* BL21 Star (DE3) was transformed with pGST-IEP, grown at 37°C until OD_{600nm} reached 0.5 and GST-IEP expression was induced for 3 hrs with 0.1 mM of IPTG. This first purification experiment was performed on 100 ml of *E. coli* culture using the batch purification method (use of the resin without column). The culture was pelleted and lysed to obtain the total protein fraction (T). The soluble protein fraction (S) was then loaded onto the Glutathione-Sepharose resin. The flow-through (Ft) was collected and the resin was washed thrice (W_1, W_2 and W_3). Purified protein fraction (P) was finally obtained by an elution at room temperature for 10 min with 10 mM of reduced glutathione. The Glutathione-Sepharose resin (R) was also analyzed to evaluate the amount of proteins that remain bound to the resin after elution. The GST-IEP molecular mass is expected to be around 91-kDa. All protein fractions were analyzed by SDS-PAGE (Sodium dodecyl sulfate polyacrylamide gel) with Coomassie blue staining (Fig. R-2).

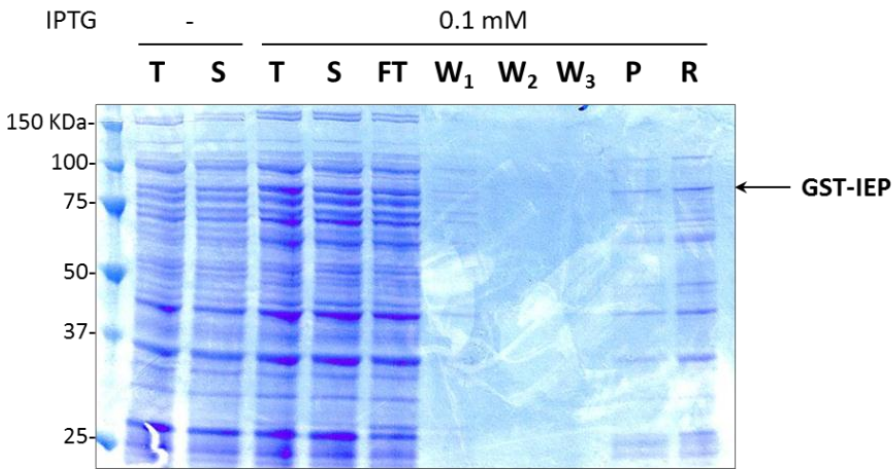


Figure R-2: Expression of GST-IEP in BL21 Star (DE3) and purification.

GST-IEP expression was induced from 100 ml *E. coli* culture at OD_{600nm} 0.5 with 0.1 mM of IPTG at 37°C for 3 hrs. A negative control was also performed from 10 ml of *E. coli* culture without IPTG induction (-). T: total protein fraction (1/47 of T(-) fraction and 1/470 of T(0.1 mM) fraction); S: soluble protein fraction (1/47 of S(-) fraction and 1/470 of S (0.1 mM) fraction); Ft: Flow-through from the Glutathione-Sepharose resin after binding of GST-IEP (1/470 of the fraction); W_1 to W_3: wash protein fractions 1 to 3 (1/190 of the fractions); P: purified protein fraction eluted from the resin (1/19 of the fraction); R: proteins which remain bound to the resin after elution (1/7 of the fraction). Numbers at *left* indicate molecular mass marker.

SDS-PAGE shows no detectable over-expression of the GST-IEP (Fig. R-2; fraction T). However, a protein that ran between the 75-kDa and the 100-kDa markers at the approximate size of the GST-IEP was found in the purified and resin fractions (Fig. R-2; fractions P and R). Those fractions were highly contaminated with a lot of *E. coli* proteins and/or degradation products. It became apparent that the GST-IEP expression and purification were not optimal.

2.1.2 - Expression in BL21 Star (DE3) pRARE, purification, and RT activity assay

When we analyzed the protein sequence of the GST-IEP, we found that it contains several codons underrepresented in *E. coli*. Indeed, most amino acids are encoded by more than one codon, and each organism carries its own bias in the usage of the 61 available amino acid codons. In each cell, the tRNA population closely reflects the codon bias of the mRNA population (Ikemura T 1981; Dong H et al. 1996). When the mRNA of an heterologous target gene is overexpressed in *E. coli*, differences in codon usage can impede translation due to the demand for one or more tRNAs, which may be rare or lacking in the tRNAs population of the host (Goldman E et al. 1995; Kane JF 1995; Kurland C and Gallant J 1996). Examination of codon usage in all 4,290 *E. coli* genes reveals a number of codons that are underrepresented in *E. coli* (Nakamura Y et al. 2000) (Table R-3; indicated in red).

amino acid	codon	fraction in all genes	fraction in highly expressed genes
Arg	AGG	0.022	**0.003**
	AGA	0.039	**0.006**
	CGG	0.098	**0.008**
	CGA	0.065	**0.011**
	CGU	0.378	0.643
	CGC	0.398	0.330
Gly	GGG	0.151	0.044
	GGA	0.109	**0.020**
	GGU	0.337	0.508
	GGC	0.403	0.428
Ile	AUA	0.073	**0.006**
	AUU	0.507	0.335
	AUC	0.420	0.659
Leu	UUG	0.129	0.034
	UUA	0.131	0.055
	CUG	0.496	0.767
	CUA	0.037	**0.008**
	CUU	0.104	0.056
	CUC	0.104	0.080
Pro	CCG	0.525	0.719
	CCA	0.191	0.153
	CCU	0.159	0.112
	CCC	0.124	**0.016**

Table R-3: Codon usage in *E. coli* of five amino acids.

Arg: arginine; Gly: glycine; Ile: isoleucine; Leu: leucine; Pro: proline. Codon usage is indicated as the fraction of all possible codons for a given amino acid. "All genes" is the fraction represented in all 4,290 coding sequences of *E. coli* (Nakamura Y et al. 2000). "Highly expressed" genes is the fraction represented in 195 genes highly and continuously expressed during exponential growth (Henaut A and Danchin A 1996). In red are indicated codons that are underrepresented in *E. coli*.

In particular, arginine codon AGA, AGG and CGA, isoleucine codon AUA and leucine codon CUA represent less than 8% of their corresponding population of codons (Table R-3; in red and underlined). The codon usage of highly expressed genes in *E. coli* demonstrates a more extreme bias. Indeed, in addition of the codons mentioned before, codon GGA for glycine, CGG for arginine and CCC for proline fall to less than 2% of their respective populations (Table R-3; in red and bold letters). Under growth conditions used to overexpress target genes in *E. coli*, it is likely in many cases that the resident tRNA population available for target protein synthesis would resemble to that of highly expressed genes in Table R-3. We then analyzed the IEP sequence to determine the content of "problematic" codons (Fig. R-4).

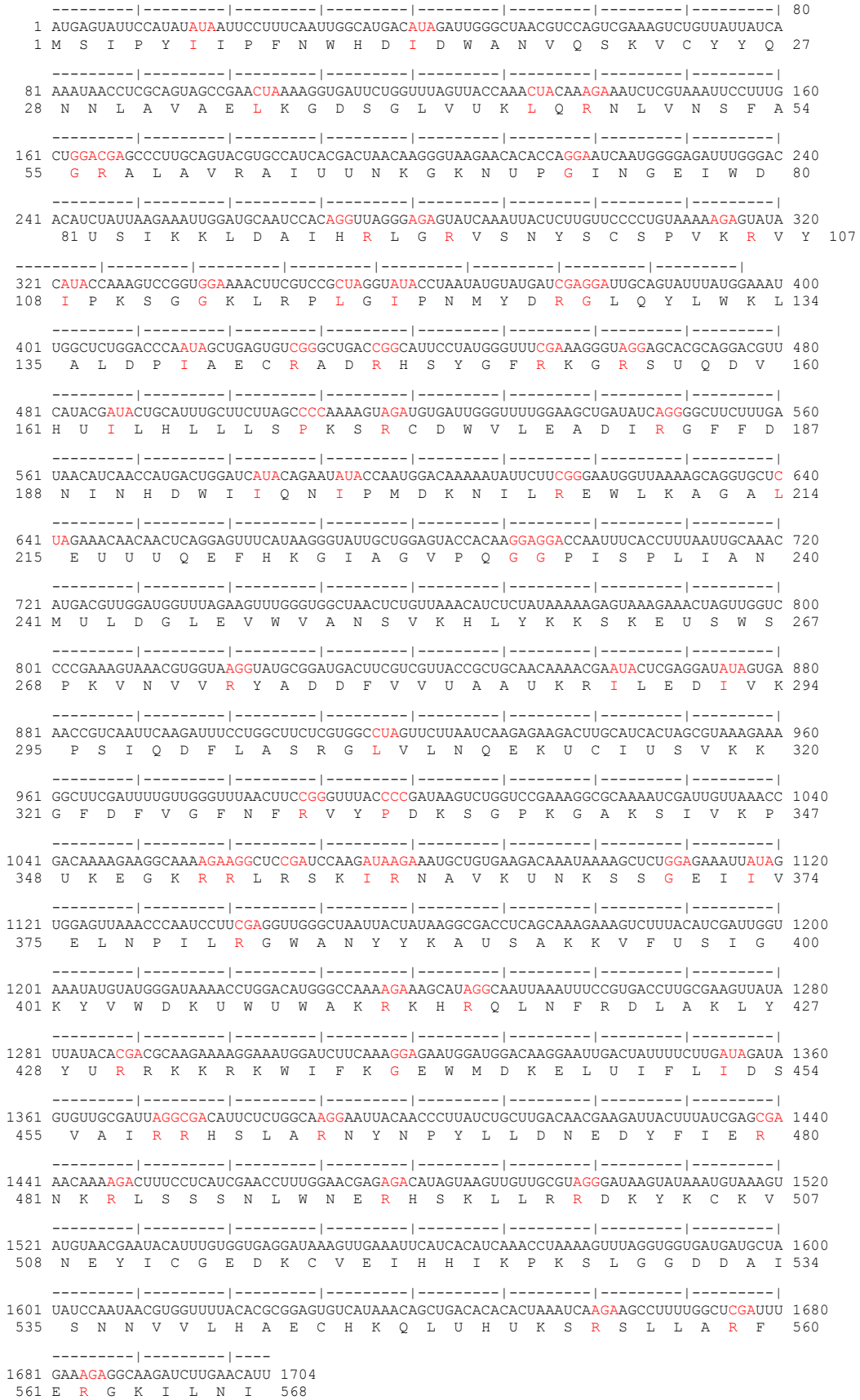


Figure R-4: Underrepresented codons in *E. coli* in the IEP sequence.

The IEP nucleotide sequence contains 64 of the "problematic" Arg, Gly, Ile, Leu and Pro codons (in red).

The IEP sequence contains 64 of these codons for which the corresponding tRNA is rare or lacking in the *E. coli* tRNA population (Fig. R-4; in red). It is likely that the low GST-IEP expression in *E. coli* is due to the lack of tRNA corresponding to these codons.

To circumvent this problem, we decided to use the plasmid pRARE (See Appendix plasmid map). This plasmid carries tRNA genes for all of the "problematic" rarely used codons; Arg, Ile, Gly, Leu and Pro, except for Arg CGA and CGG. The expression of tRNAs in pRARE is driven by their native promoters. The pRARE plasmid, containing a p15a origin of replication, is compatible with the pGST-IEP plasmid, which contain a ColE1 origin of replication. Numerous reports confirm the efficacy of plasmid-mediated tRNA supplementation (Baca AM and Hol WG 2000; Sorensen HP et al. 2003).

*(a) **Expression and purification***

The pRARE plasmid was transformed in BL21 Star (DE3) and chemically competent BL21 Star (DE3) pRARE cells were then prepared. To evaluate the efficiency of GST-IEP expression in this novel strain, small-scale cultures were performed. A lower induction temperature (32°C) was also tested, as it was shown that it could improve the expression of soluble proteins in *E. coli* (Hammarstrom M et al. 2002). Cell pellets from 2 ml of expression samples induced at 37°C and 32°C for 3 hrs with or without 1 mM of IPTG were prepared. Total (T), insoluble (I) and soluble (S) protein fractions were loaded on a SDS-PAGE gel (Fig. R-5).

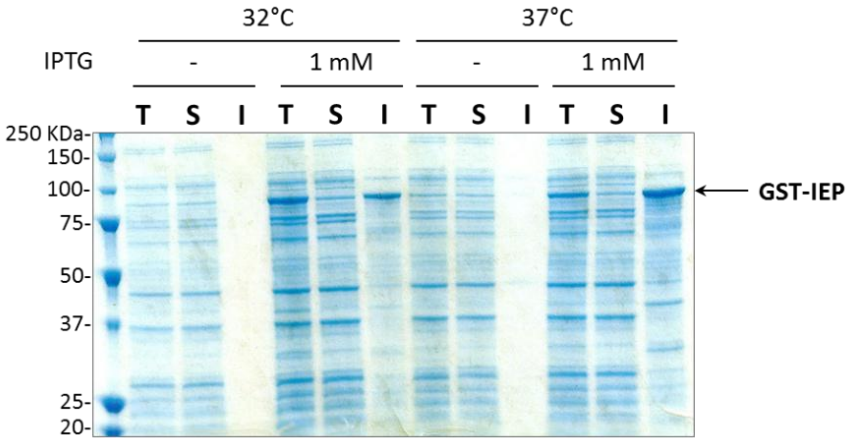


Figure R-5: GST-IEP expression in BL21 Star (DE3) pRARE strain.

A 10 ml culture of BL21 Star (DE3) pRARE transformed with pGST-IEP expression plasmid was grown at 37°C until OD_{600nm} reached 0.5. The culture was then split into four 2.5 ml sample. GST-IEP expression was induced at 32°C or 37°C with 1 mM of IPTG for 3 hrs. Negative controls without IPTG induction (-) were also performed for both induction temperatures. Total (T, 1/20 of the fractions), insoluble (I, 1/8 of the fraction), and soluble (S, 1/20 of the fractions) protein fractions were analyzed on a 10% SDS-PAGE gel and stained with Coomassie blue. Numbers at *left* indicate molecular mass marker.

SDS-PAGE gel shows that GST-IEP is over-expressed in *E. coli* BL21 Star (DE3) pRARE strain at 32°C and 37°C (Fig. R-5; 1 mM IPTG, fractions T). The protein is at the expected size of 91-kDa. The efficiency of GST-IEP expression appears to be quite similar at 32°c and 37°C. Notably, the fraction of insoluble GST-IEP is lower when the culture is induced at 32°C rather than 37°C (Fig. R-5; 32°C, 1 mM IPTG, fraction I). This result indicates that GST-IEP becomes less insoluble when the induction temperature is decreased. In this experiment, the low expression problem of GST-IEP in *E. coli* was

solved by the use of plasmid-mediated tRNA complementation. However, the GST-IEP is mainly found in the insoluble protein fraction.

The purification of GST-IEP from the insoluble fraction is possible but would require a solubilization process under denaturing conditions followed by a refolding step. Those conditions do not ensure a correct protein structure *in fine* and this could affect both GST binding to the resin and biochemical activities of the fusion protein. The optimization of the solubility of GST-IEP is thus crucial. Factors such as drastically reduced temperature or induction conditions (lowering IPTG concentration and induction time, greater culture aeration) have in some specific cases lead to enhance soluble protein production in *E. coli* (Shirano Y and Shibata D 1990). Indeed, growth at a temperature range of 15-23°C could lead to significant reduction in degradation of the expressed protein (Spiess C et al. 1999; Hunke S and Betton JM 2003). It has also been shown that heat shock proteases induced during over-expression of proteins in *E. coli* are poorly active at lower temperature conditions. Moreover, macromolecular crowding of proteins at concentrations of 200-300 mg/ml in the cytoplasm of *E. coli* (inclusion bodies) suggests a highly unfavorable protein-folding environment, especially during recombinant high-level expression (van den Berg B et al. 1999). Lowering the expression rate by reducing the induction temperature would increase the available time for protein folding, thus minimize the formation of inclusion bodies containing unfolded/misfolded aggregated proteins. In this context, we have induced the expression of GST-IEP at 18°C with various IPTG concentrations. Total (T), soluble (S), and insoluble (I) protein fractions were analyzed by Coomassie blue staining and western blot using an HRP-conjugated anti-GST antibody (Fig. R-6).

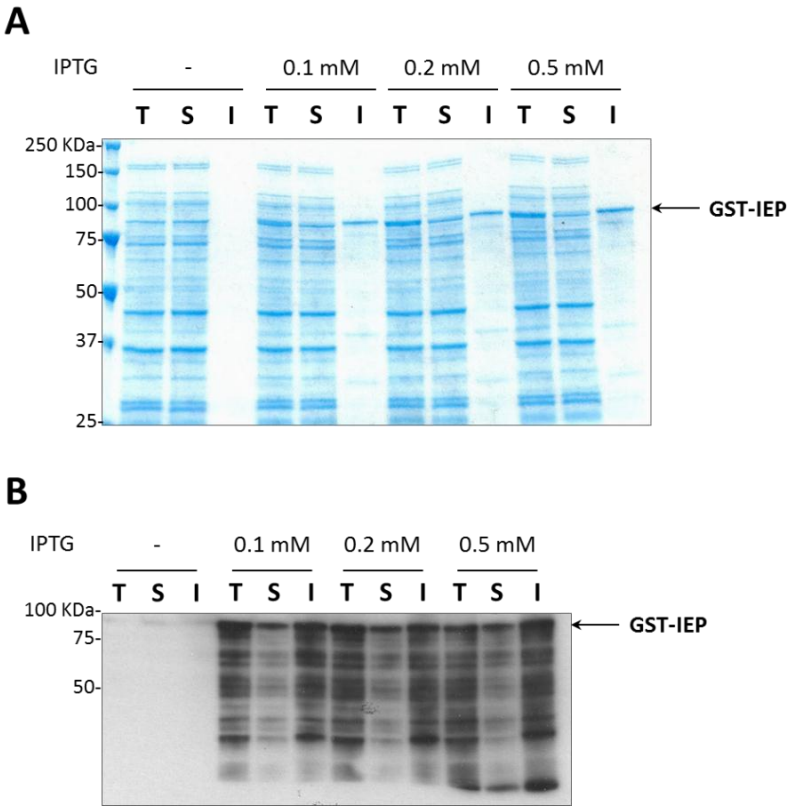
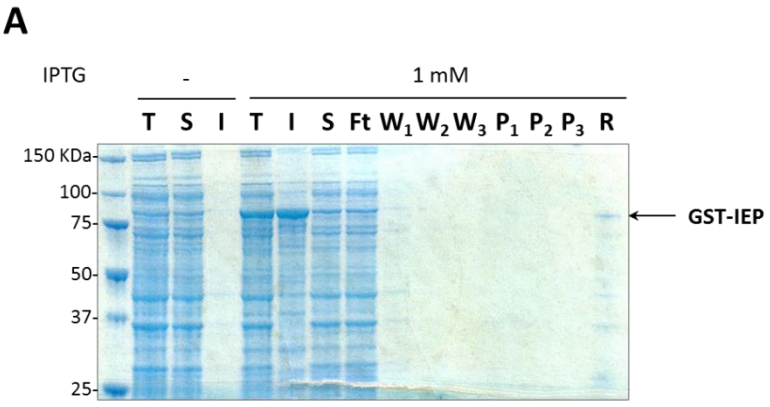


Figure R-6: GST-IEP expression in BL21 Star (DE3) pRARE at 18°C with various IPTG concentrations.
BL21 Star (DE3) pRARE was transformed with pGST-IEP expression plasmid. 2 ml cultures were grown at 37°C until OD_{600nm} reached 0.5. Cultures were then transferred at 18°C and incubated for 20 min. GST-IEP

expression was then induced with 0.1, 0.2 or 0.5 mM of IPTG for 3 hrs at 18°C. Negative control was also performed from 2 ml of culture without IPTG induction (-). Total (T, 1/20 of the fractions), insoluble (I, 1/8 of the fractions) and soluble (S, 1/20 of the fractions) protein fractions were analyzed. (***A***) Coomassie blue SDS-PAGE gel of proteins fractions. (***B***) Western blot analysis using an HRP-conjugated mouse anti-GST antibody. Numbers at *left* indicate molecular mass marker.

The Coomassie blue stained gel shows that the GST-IEP is over-expressed in all conditions (Fig. R-6A; 0.1 to 0.5 mM IPTG, fractions T). A significant amount of GST-IEP is found in insoluble fractions, whatever the IPTG concentration used (Fig. R-6A; 0.1 to 0.5 mM IPTG, fractions I). The soluble form of GST-IEP is difficult to detect by Coomassie staining, as an *E. coli* protein ran just below the GST-IEP (Fig. R-6A; fractions T - and S -). To verify the presence of the GST-IEP in the soluble fraction, a western blot analysis was performed (Fig. R-6B). It shows that the protein is expressed as a soluble form at all IPTG concentrations tested (Fig. R-6B; fractions S). The fraction of soluble GST-IEP is not increased when the IPTG concentration is decreased from 0.5 to 0.1 mM. Notably, GST-IEP is here subjected to degradation, as shown by the presence of multiple degradation products in all induced fractions (Fig. R-6B). The low induction temperature seems to enhance the solubility of GST-IEP but the variation of IPTG concentration does not appear to influence the solubilization of the protein.

To ensure a good purification efficiency of GST-IEP, it is required to equilibrate the expression rate and the solubility of the protein in order to obtain a fair amount of soluble protein at the end of the purification process. A small-scale purification was performed at 4°C on a 20 ml *E. coli* culture induced with 1 mM of IPTG at 18°C for 3 hrs to ensure a good expression rate and maximize the solubility of the protein. All purification steps including the elution step were here performed at 4°C to minimize protein degradation. Protein samples were analyzed by SDS-PAGE with Coomassie blue staining and by western blot (Fig. R-7).



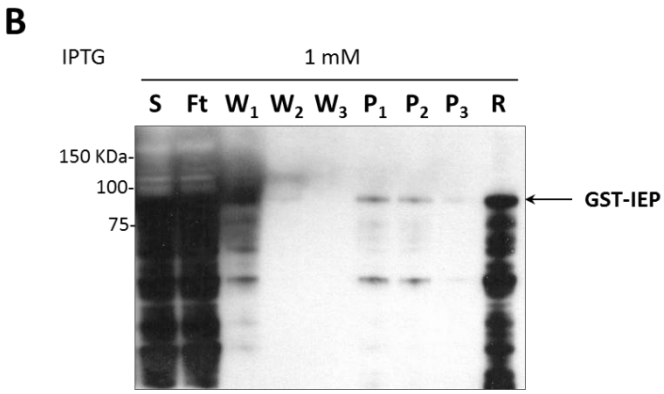


Figure R-7: Purification of GST-IEP expressed in BL21 Star (DE3) pRARE.

A 20 ml culture of BL21 Star (DE3) pRARE transformed with pGST-IEP expression plasmid was grown at 37°C until OD_{600nm} reached 0.7. The culture was transferred at 18°C and continued for 20 min. GST-IEP expression was then induced with 1 mM of IPTG for 3 hrs. A negative control was also performed from 20 ml of *E. coli* culture without IPTG induction (-). T: total protein fraction (1/100 of the fraction); S: soluble protein fraction (1/100 of the fraction); I: insoluble protein fraction (1/20 of the fraction); Ft: Flow-through from the Glutathione-Sepharose resin after binding of GST-IEP (1/100 of the fraction); W_1 to W_3: wash protein fractions 1 to 3 (1/50 of the fraction); P_1 to P_3: purified protein fractions successively eluted from the resin (1/5 of the fraction); R: proteins which remain bound to the resin after elutions (1/5 of the fraction). Numbers at *left* indicate molecular mass marker. (***A***) Coomassie blue stained 10% SDS-PAGE gel. (***B***) Western blot analysis using an HRP-conjugated mouse anti-GST antibody.

Coomassie blue stained SDS-PAGE shows that GST-IEP is over-expressed in this experiment (Fig. R-7A; 1 mM IPTG, fraction T), but is mainly found as an insoluble form (Fig. R-7A; 1 mM IPTG, fraction I). Nevertheless, a small fraction of soluble GST-IEP has bound to the resin (Fig. R-7A; 1 mM IPTG, fraction R). The western blot reveals also that the binding to the resin does not seem to be very strong as some non-negligible amount of the soluble GST-IEP is found in the flow-through and in the first wash fraction (Fig. R-7B; 1 mM IPTG, fractions Ft and W_1). Eluted GST-IEP is not detectable on the Coomassie blue stained gel (Fig. R-7A; 1 mM IPTG, fractions P_1 to P_3) but is highlighted by western blot (Fig. R-7B; fraction P_1 and P_2). It becomes apparent that binding and elution steps require some optimizations.

The loss of a high amount of soluble GST-IEP in the flow-through could be caused by a saturation of the resin. Therefore, the amount of resin was increased 3-fold in regards to the manufacturer's instructions. To improve the elution efficiency, the concentration of reduced glutathione was increased from 10 mM to 50 mM (Fig. R-8).

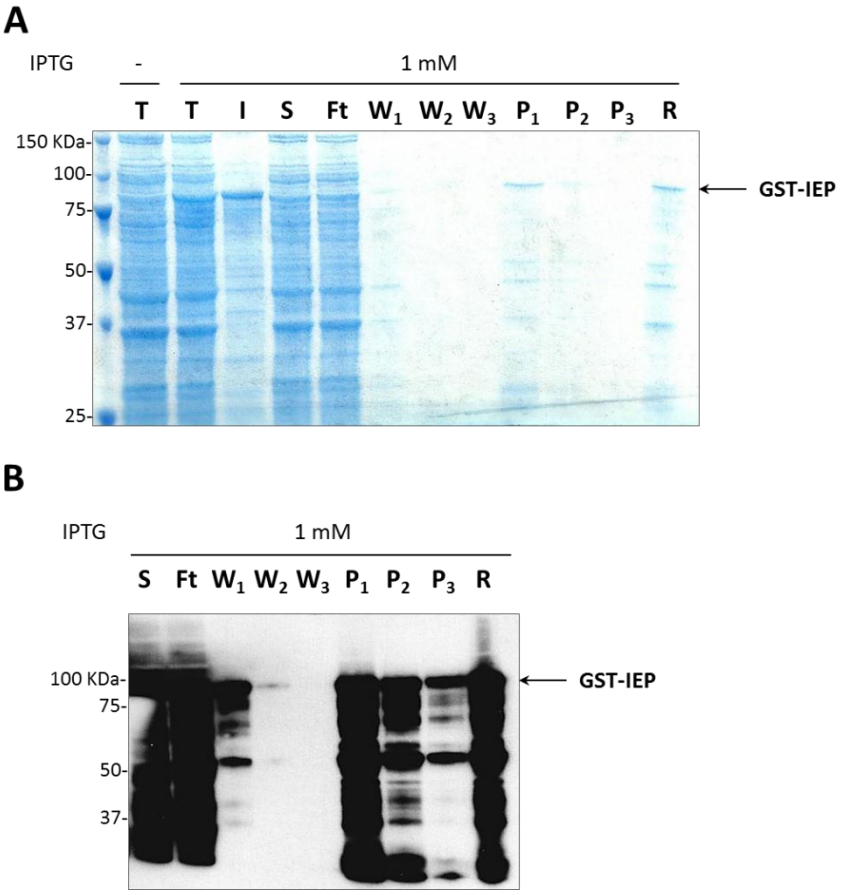


Figure R-8: Purification of GST-IEP using 50 mM of reduced glutathione for the elution and 3-fold amount of resin.

A 30 ml culture of BL21 Star (DE3) pRARE transformed with pGST-IEP expression plasmid was grown at 37°C until OD_{600nm} reached 0.6. The culture was transferred at 18°C and continued for 20 min. GST-IEP expression was then induced with 1 mM of IPTG for 3 hrs. A negative control was also performed from 30 ml of *E. coli* culture without IPTG induction (-). T: total protein fraction (1/100 of the fraction); S: soluble protein fraction (1/100 of the fraction); I: insoluble protein fraction (1/20 of the fraction); Ft: Flow-through from the Glutathione-Sepharose resin after binding of GST-IEP (1/100 of the fraction); W_1 to W_3: wash protein fractions 1 to 3 (1/50 of the fraction); P_1 to P_3: purified protein fractions successively eluted from the resin (1/10 of the fraction); R: proteins which remain bound to the resin after elutions (1/3 of the fraction). Number at *left* indicate molecular mass marker. (***A***) Coomassie blue stained 10% SDS-PAGE gel. (***B***) Western blot analysis using an HRP-conjugated mouse anti-GST antibody.

Coomassie stained gel shows that the GST-IEP is detected in the first purified protein fraction (Fig. R-8A; 1 mM IPTG, fraction P_1). The increase of reduced glutathione concentration has thus improved the GST-IEP elution efficiency, even if a non-negligible amount of GST-IEP still remains bound to the resin after elutions (Fig. R-8A; 1 mM IPTG, fraction R). Three successive elution steps were performed successively and showed that most of GST-IEP is eluted from the first elution step (Fig. R-8A; 1 mM IPTG, fractions $P_{1)}$. Purified proteins fractions (Fig. R-8A; 1 mM IPTG, P_1 and P_2) and resin fraction (Fig. R-8A; 1 mM IPTG, R) are contaminated by *E. coli* proteins, which could have bound non-specifically to the resin, and/or GST-IEP degradation products. The presence of degradation products is confirmed by western blot (Fig. R-8B). Western blot also shows that the use of a greater amount of resin does not seem to circumvent the problem of the loss of GST-IEP during

purification. Indeed, the protein is still found in the flow-through and the first wash (Fig. R-8B; 1 mM of IPTG, fractions Ft and W_1). The elution of GST-IEP was improved by the use of a greater concentration of reduced glutathione, but the low binding of GST-IEP to the resin was not solved.

This low affinity binding of GST-IEP to the resin may be due to an altered conformation of the GST. To improve the binding of GST-fusion protein, it is usually recommended to perform this step at 4°C, which was already done in the protocol used before. Thus, it seems difficult to further optimize this step. The elution of the GST-IEP could also be optimized by using a greater glutathione concentration but this could lead to decrease the purity of the purified protein fraction and a relatively large amount of GST-IEP is already eluted. As the GST is a relatively large tag, it may interfere with the proper folding of the IEP, impeding its biochemical activities to be assayed. It could thus be necessary to remove it at the end of the purification process using thrombin cleavage. However, some groups have shown the possibility of generating proper crystal structures of fusion proteins (Smyth DR et al. 2003). In most cases, functional tests can be performed using intact GST-tagged fusion protein. GST removal is thus not always necessary. Reverse transcriptase activity assays can thus be considered on GST-IEP. For these assays, GST-IEP protein has to be formulated in a neutral storage buffer. GST-IEP purification was performed as previously on a 150 ml *E. coli* induced culture and a dialysis step was added at the end of the process. The ionic strength of the elution buffer was increased with the addition of 120 mM of NaCl to prevent the binding of the protein to the dialysis membrane. Each GST-IEP purified fractions were thus dialyzed against the elution buffer lacking reduced glutathione (Fig. R-9).

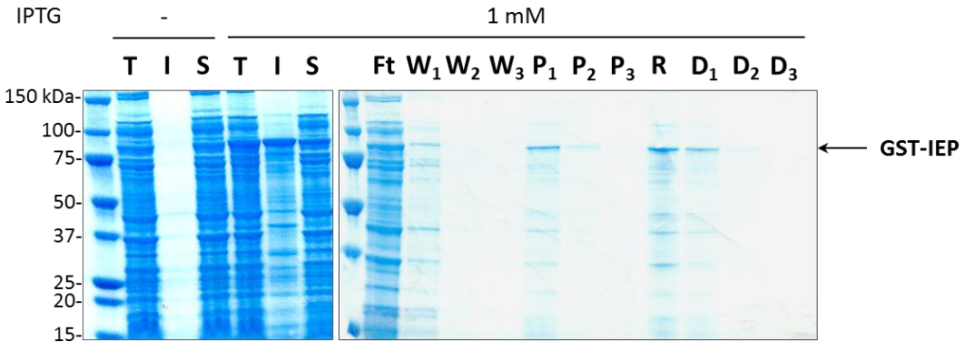


Figure R-9: Purification of GST-IEP and dialysis.

Coomassie blue stained SDS-PAGE gel containing 20 µl of each purification fractions. A 150 ml culture of BL21 Star (DE3) pRARE transformed with pGST-IEP expression plasmid was grown at 37°C until OD_{600nm} reached 0.6. The culture was transferred at 18°C and continued for 20 min. GST-IEP expression was induced with 1 mM of IPTG for 3 hrs. A negative control was also performed from 10 ml of *E. coli* culture without IPTG induction (-). T: total protein fraction (1/35 of fraction T(-) and 1/535 of fraction T(1 mM)); I: insoluble protein fraction (1/22 of fraction I- and 1/350 of fraction I+); S: soluble protein fraction (1/35 of fraction S(-) and 1/535 of fraction S(1 mM)); Ft: Flow-through from the Glutathione-Sepharose resin after binding of GST-IEP (1/535 of the fraction); W_1 to W_3: wash protein fractions 1 to 3 (1/215 of the fraction); P_1 to P_3: purified protein fractions successively eluted from the resin (1/32 of the fractions); R: proteins which remain bound to the resin after elutions (1/16 of the fraction); D_1 to D_3: purified protein fractions after dialysis against elution buffer lacking reduced glutathione (1/30 of the fractions). Numbers at *left* indicate molecular mass marker.

The SDS-PAGE gel shows that only a few amount of GST-IEP is lost during the dialysis step (Fig. R-9; 1 mM IPTG, fractions D_1 to D_3). The dialyzed fraction (Fig. R-9; 1 mM IPTG, fraction D_1) obtained

using the first purified fraction (Fig. R-9; 1 mM IPTG, fraction P_1) contains a sufficient amount of GST-IEP, which is the predominant protein in the fraction. This purified and dialyzed fraction can be thus used in a reverse transcriptase (RT) assay.

A negative control is required in the RT assay. Indeed, GST-IEP is only partially purified as some contaminating *E. coli* proteins (and/or GST-IEP degradation products) are still present and could bias the RT assay. A mutant form of the GST-IEP (GST-IEP mtDD-) was constructed by site-directed mutagenesis of the pGST-IEP plasmid (pGST-IEPmtDD-). The catalytic YADD motif contained in the RT5 domain of the IEP (See Fig. I-23) is mutated in YAAA. Mutation of these asparagine residues is commonly used to abolish the RT activity of other group II intron-encoded proteins (Matsuura M et al. 1997) and also abolish RT activity of HIV-1 reverse transcriptase (Larder BA et al. 1989). This GST-IEP mtDD- protein should thus be RT-defective. Several clones obtained by transformation of BL21 Star (DE3) pRARE with the pGST-IEPmtDD- were verified by sequencing the GST-IEP mutated sequence. Small-scale expression experiments were then performed on positive clones. One of these clones, which showed a good expression rate, was finally chosen to purify the GST-IEP mtDD- (Fig. R-10).

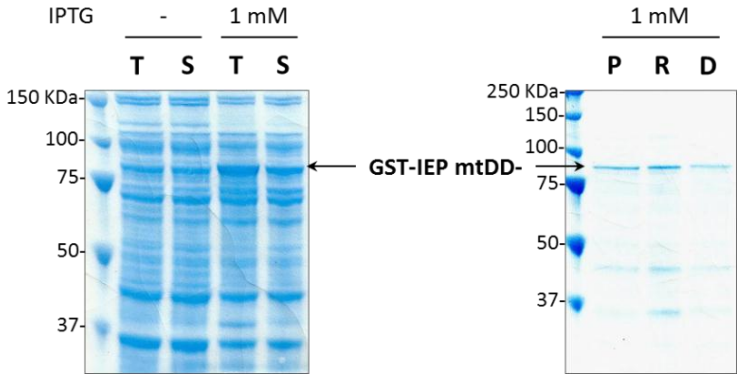


Figure R-10: Expression and purification of the mutant GST-IEP mtDD-.

Coomassie blue stained SDS-PAGE gel. A 100 ml culture of BL21 Star (DE3) pRARE transformed with pGST-IEP mtDD- expression plasmid was grown at 37°C until OD_{600nm} reached 0.6. The culture was transferred at 18°C and continued for 20 min. GST-IEP mtDD- expression was then induced with 1 mM of IPTG for 3 hrs. A negative control was also performed from 50 ml of *E. coli* culture without IPTG induction (-). T: total protein fraction (1/175 of the fraction T(-) and 1/350 of the fraction T(1 mM)); S: soluble protein fraction (1/175 of the fraction S(-) and 1/350 of the fraction S(1 mM)); P: purified protein fraction eluted from the resin (1/21 of the fraction); R: proteins which remain bound to the resin after elution (1/11 of the fraction); D: purified protein fraction after dialysis against elution buffer lacking reduced glutathione (1/19 of the fraction). Numbers at *left* indicate molecular mass marker.

In this experiment, only one elution step was performed, as most of the wild-type GST-IEP was shown to be recovered from the first elution. SDS-PAGE shows that GST-IEPmtDD- is overexpressed in BL21 Star (DE3) pRARE (Fig. R-10; 1 mM IPTG, fraction T) and about 50% is soluble (Fig. R-10; 1 mM IPTG, fraction S). A relatively good amount of the mutant GST-IEP mtDD- is purified (Fig. R-10; 1 mM IPTG, fraction P), even if a non-negligible amount is remains bound to the resin (Fig. R-10; 1 mM IPTG, fraction R). In this experiment, about 50% of the protein is lost during dialysis (Fig. R-10; 1 mM IPTG, fraction D). However, as for the wild-type GST-IEP, the mutant GST-IEP mtDD- is

the predominant protein found in the purified and dialyzed fraction, although some *E. coli* proteins and/or degradation products still contaminate the fraction.

The wild-type and mutant mtDD- GST-IEP purified and dialyzed fractions obtained previously were subsequently used to assay the IEP reverse transcriptase activity.

(b) Reverse transcriptase activity

RT activity is assayed using the artificial template-primer substrate poly(rA)-oligo(dT)$_{12\text{-}18}$. The RT activity is indicated by incorporation of [α-^{32}P]dTTP during the reaction (See Material and methods section 3.5.4 -). The first experiment consisted in a time course of potential RT activity of the wild-type GST-IEP using a fixed volume of dialyzed protein fraction, while the same volume of GST-IEP mtDD- dialyzed fraction was used at the maximal time point (Fig. R-11).

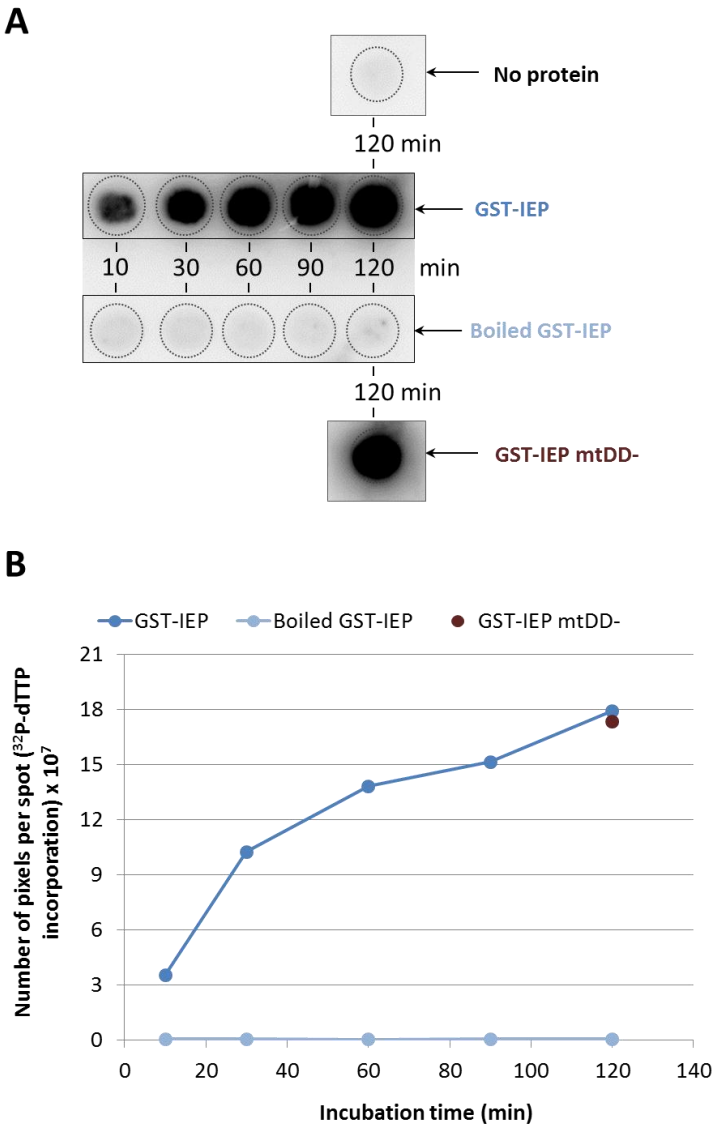


Figure R-11: RT assay with GST-IEP and GST-IEP mtDD-.

(***A***) Poly(rA)-oligo(dT)$_{12\text{-}18}$ and 8 µl of dialyzed protein fractions were used. RT reactions with GST-IEP were performed at 37°C for 10, 30, 60, 90 and 120 min. Negative controls consisting of GST-IEP protein fractions incubated 10 min at 90°C before RT reactions were also subjected to the time course (boiled GST-IEP). RT

reactions without protein (No protein; background) and with GST-IEP mtDD- were performed at 37°C for 120 min. (***B***) Data representing the number of pixels per spot were quantified with ImageQuant™ software and corrected to the background. Dark blue line: GST-IEP; light blue line: boiled GST-IEP; dark red spot: GST-IEP mtDD-.

Figure R-11A shows the membrane image acquisition. The background, consisting in an assay using the dialysis buffer (Fig. R-11A; No protein), shows no signal. The data were then quantified and corrected to the background. Results are represented in figure R-11B. The time course performed using GST-IEP fraction shows an RT activity positively correlated to the reaction time (Fig. R-11B; GST-IEP, 10 to 120 min). This RT activity is abolished by denaturation of proteins contained in the dialyzed fraction (Fig. R-11B; boiled GST-IEP), indicating that these results are not artifacts from the experiment. Surprisingly, the GST-IEP mtDD- dialyzed fraction also shows an RT activity similar to that found for the wild-type GST-IEP dialyzed fraction (Fig. R-11B; GST-IEP mtDD-). This result was not expected, as this mutant should be RT-defective. It suggests that the RT activity found with these purifications does not depend on GST-tagged IEP, but probably reflects a bias induced by a contaminating *E. coli* protein presenting a reverse transcriptase activity and contained in dialyzed fractions.

To further confirm or infirm these results, another strain of *E. coli* was used as the expression host and the RT activity of two additional control proteins was assayed.

2.1.3 - Expression in ArcticExpress (DE3)RIL, purification, and RT activity assay

The use of another *E. coli* strain could possibly circumvent the contamination problem highlighted in the previous experiment. We found at this time that the ArcticExpress (DE3)RIL strain, derived from BL21 (DE3) strain, was engineered to enhance protein solubility at low temperatures. Indeed, this strain co-expresses the cold-adapted chaperonins Cpn10 and Cpn60 from *Oleispira Antarctica*, which show high protein refolding activities at temperatures of 4-12°C (Ferrer M et al. 2003). The use of this strain could potentially increase the yield of active soluble recombinant protein, allowing better purification efficiency, and thus minimizing the amount of contaminating *E. coli* proteins. This strain also expresses tRNA for arginine codons AGG/AGA, isoleucine codon AUA and leucine codon CUA, overcoming codon usage bias.

In addition, a second mutant form of GST-IEP was here introduced. This new mutant is deleted of the RT5 domain (GST-IEP ΔRT5). The corresponding protein is expected to have a molecular mass of 88-kDa. The use of this mutant together with the GST-IEPmtDD- mutant, purified at different times, should allow us to conclude or not to a contamination of purified protein fractions by an *E. coli* reverse transcriptase protein. Indeed, as for the YAAA mutation, the deletion of the RT5 domain should abolish any RT activity of the IEP.

We also included in our experiment another control consisting in the expression, purification, and RT activity assay of GST protein alone. The GST protein is expressed from the pGEX-4T1 plasmid. The use of this control should allow us to determine if the contaminant *E. coli* protein presenting the RT activity is co-eluted only when the IEP is present or not. Indeed, detection of RT activity with GST purified sample would indicate a direct or indirect binding of the contaminant to the resin or to the GST protein.

GST, wild-type GST-IEP and mutants GST-IEP mtDD- and GST-IEP ΔRT5 proteins were thus expressed in ArcticExpress (DE3)RIL strain. A 400 ml *E. coli* culture was induced at 15°C for 18 hrs with 0.1 mM of IPTG. Proteins were then purified as shown previously. Coomassie blue staining and western blot analyses of purified and dialyzed proteins fractions were performed (Fig. R-12A) and RT activity was subsequently assayed (Fig. R-12B).

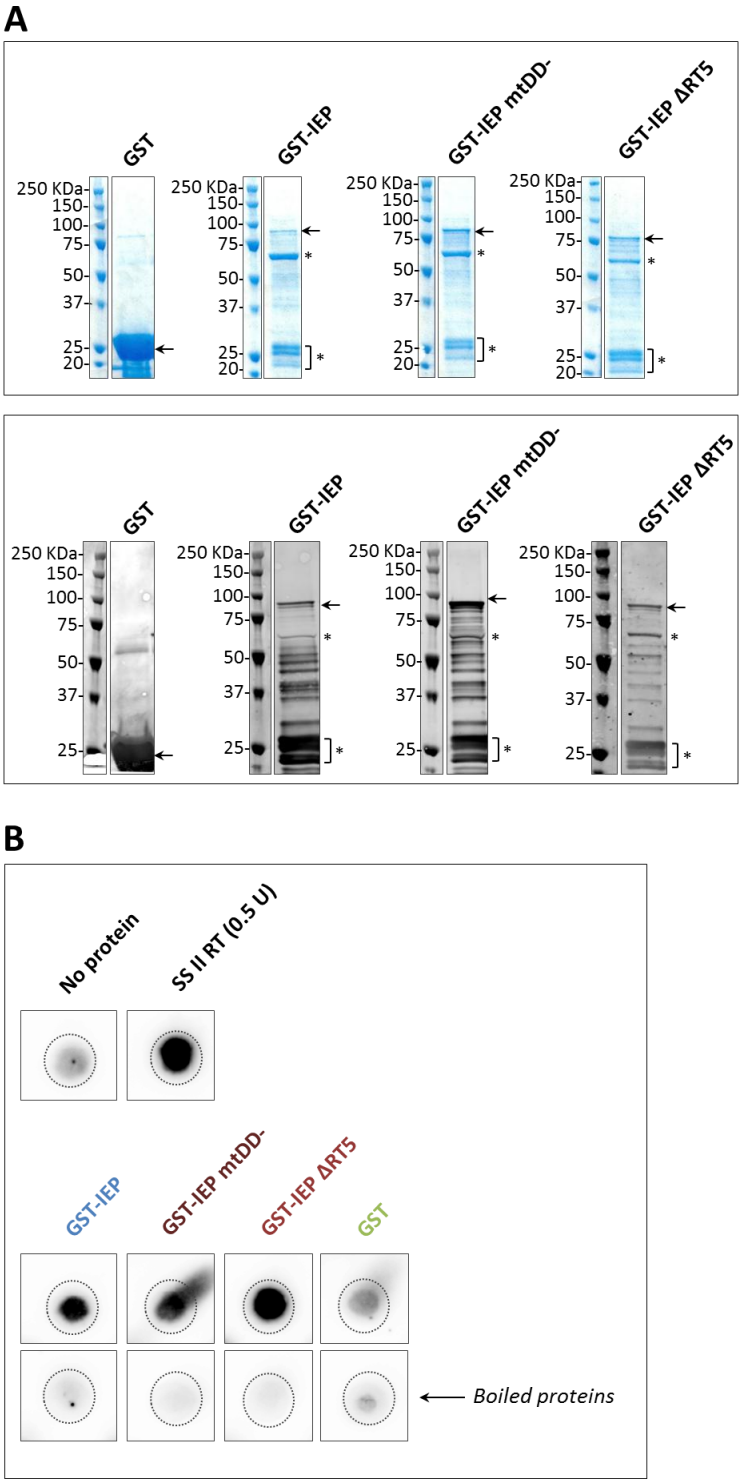


Figure R-12: RT assay with GST-IEP, GST-IEP mtDD-, GST-IEP ΔRT5 and GST.

(***A***) 400 ml culture of ArcticExpress (DE3)RIL transformed with the appropriate expressing plasmid was grown at 32°C until OD_{600nm} reached 0.6. The culture was transferred at 15°C and continued for 25 min. GST and GST-

tagged proteins expression was then induced with 0.1 mM of IPTG for 18 hrs. *Upper panel*: Coomassie blue stained SDS-PAGE gel of purified protein fraction after dialysis against elution buffer lacking reduced glutathione (1/20 of the fraction loaded). *Lower panel*: Western blot analysis of protein fractions using an HRP-conjugated mouse anti-GST antibody (1/20 of the fraction loaded). GST-tagged proteins degradation products are indicated by asterisks. Numbers at *left* indicate molecular mass marker. (***B***) RT assay with poly(rA)-oligo$(dT)_{12\text{-}18}$ and 5 µl of dialyzed protein fractions. RT reactions were performed at 37°C for 1 hr with GST-IEP, GST-IEP mtDD-, GST-IEP ΔRT5 and GST. Negative controls consisting of dialysis buffer (No protein) and protein fractions incubated 10 min at 90°C (boiled proteins) were also subjected to RT reactions, as well as positive control consisting of the SuperScript II reverse transcriptase (0.5 units).

Figure R-12A shows that GST and GST-tagged proteins (IEP and mutants), expressed in *E. coli* strain ArcticExpress (DE3)RIL, were purified with a relatively good efficiency (Fig R-12A; upper panel). GST protein is highly expressed in *E. coli* and highly soluble under conditions used, allowing the purification of a great amount of GST protein (Fig. R-12A; upper panel, GST). All GST-tagged purified and dialyzed protein fractions show a similar profile: full-length proteins are co-purified with contaminant proteins (Fig. R-12A; upper panel, GST-IEP, GST-IEP mtDD- and GST-IEP ΔRT5). Western blot analysis shows that a high amount of those proteins are degradation products (Fig. R-12A; lower panel, indicated by asterisks). However, some proteins detected by Coomassie blue staining are not detected by western blot (Fig. R-12A), indicating the presence of some *E. coli* contaminant proteins. Even so, those purified and dialyzed protein fractions were used to assay the RT activity of IEP. As previously, the background showed no signal (Fig. R-12B; no protein) and wild-type and mtDD- fractions have high RT activity (Fig. R-12B; GST-IEP and GST-IEP mtDD-). We also observed that the mutant GST-IEP ΔRT5 fraction presents an RT activity similar to those of wild-type and mtDD- fractions (Fig. R-12B; GST-IEP ΔRT5). RT activity was abolished when proteins were denatured before the reaction (Fig R-12B; boiled proteins). Again, it indicates that signals found are not artifacts of the experiment. These results confirm those obtained before and indicate that a protein purified in all GST-tagged protein fractions is responsible for the RT activity in those assays. Interestingly, the assay using GST purified and dialyzed protein fraction shows no RT activity (Fig. R-12B; GST). This suggests that the contaminating *E. coli* protein, responsible for the RT activity, is co-eluted specifically with the IEP.

To determine if the *E. coli* BL21 (DE3) strain contains an ORF encoding a reverse transcriptase protein, we used BLASTP (Altschul SF et al. 1997; Altschul SF et al. 2005) with the Pl.LSU/2 IEP protein sequence (gi|15150713) as a query. The *E. coli* BL21 (DE3) complete genomic sequence (Jeong H et al. 2009) was used in this research. The protein whom sequence produces the most significant alignment is a reverse transcriptase protein (gi|254287748) encoded by the retron EC86 (Fig. R-13).

```
>gi|251784373|ref|YP_002998677.1|  reverse transcriptase [Escherichia coli BL21(DE3)]
 gi|254287749|ref|YP_003053497.1|  retron EC86 RNA-directed DNA polymerase-like protein [Escherichia
coli BL21(DE3)]
Length=320

 GENE ID: 8115105 ybl35 | reverse transcriptase [Escherichia coli BL21(DE3)]

 Score = 45.1 bits (105),  Expect = 4e-08, Method: Compositional matrix adjust.
 Identities = 31/93 (33%), Positives = 49/93 (53%), Gaps = 20/93 (22%)

Query  218  VPQGGPISPLIANMTLDGLEVWVANSVKHLYKKSKETSWSPKVNVVRYADDFVVTAATKR  277
            +PQG P SP +AN+     L+  +        Y  S+   ++       RYADD  ++A +
Sbjct  159  LPQGAPSSPKLANLICSKLDYRIQG-----YAGSRGLIYT------RYADDLTLSAQS--  205

Query  278  ILEDIVKPSIQDFL----ASRGLVLNQEKTCIT  306
             ++ +VK   +DFL     S GLV+N +KTCI+
Sbjct  206  -MKKVVKA--RDFLFSIIPSEGLVINSKKTCIS  235
```

Figure R-13: BLASTP alignment result using the Pl.LSU/2 IEP sequence against the *E. coli* BL21 (DE3) complete genomic sequence.

The BLASTP result represented here is the most significant alignment found (E-value 4 x 10^{-8}).

This *E. coli* reverse transcriptase presents a sequence similar to those of Pl.LSU/2 IEP on a portion of 93 amino acids, with 53% of positive matches (Fig. R-13; positives 49/93). This protein corresponds to the reverse transcriptase encoded by the retron EC86 (Lampson BC et al. 2005). It contains two related reverse transcriptase domains: the Interpro IPR000477 domain, which is also present in Pl.LSU/2 IEP, and the IPR000123 domain, which contains a signature specific for the RNA-dependent DNA polymerase of bacterial retrons (Lim D and Maas WK 1989; Poch O et al. 1989; Inouye M and Inouye S 1991; Inouye S and Inouye M 1995).

The alignment of amino acid sequences of Pl.LSU/2 IEP and the EC86 reverse transcriptase (RT) shows that the latter presents moderate to high conservation in seven RT blocks (Fig. R-14; RT blocks 1 to 6).

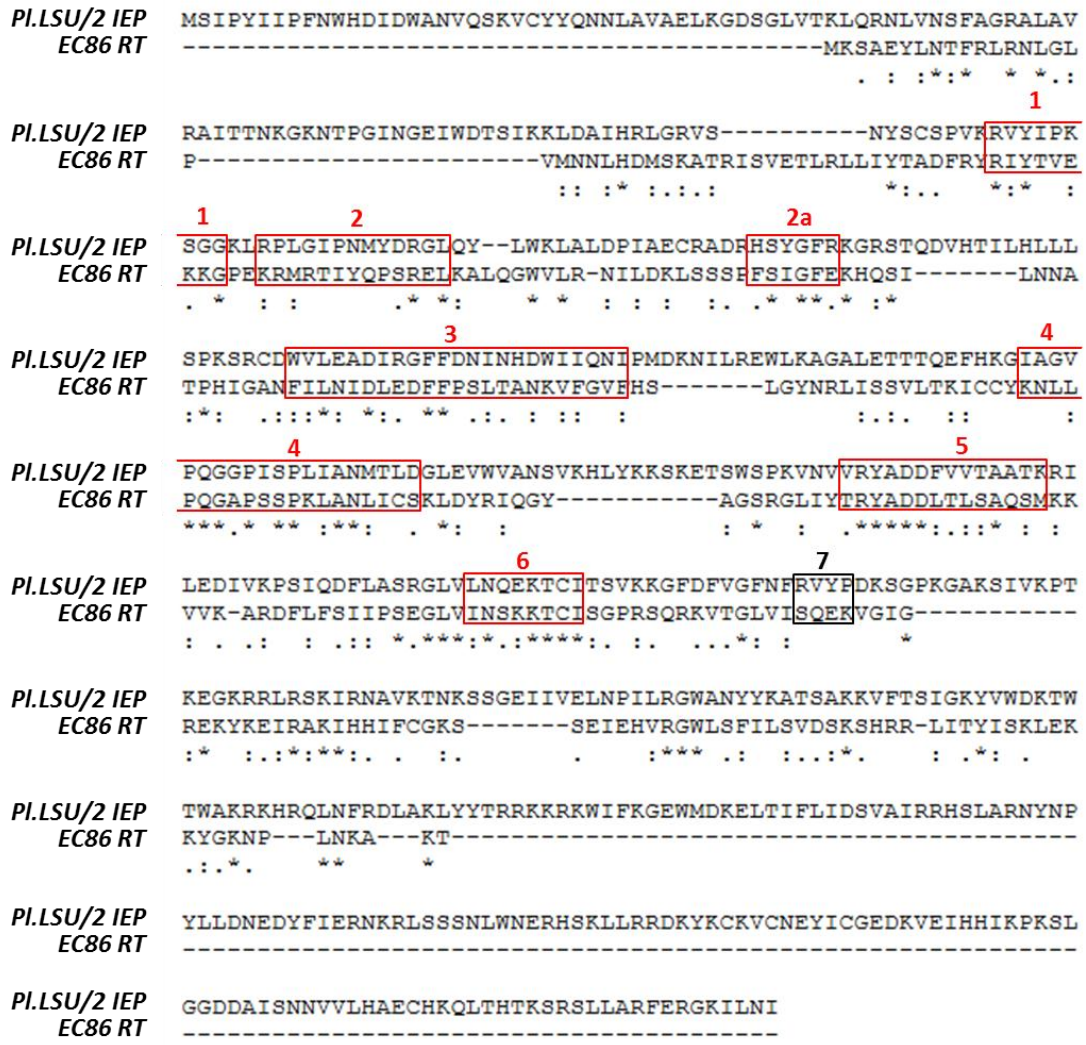


Figure R-14: Alignment of Pl.LSU/2 IEP and EC86 RT amino acid sequences.

The alignment was performed using ClustalW. Asterisks (*) indicate fully conserved residues, colons (:) indicate strong group (STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY, FYW), points (.) indicate weak groups (CSA, ATV, SAG, STNK, STPA, SGND, SNDEQK, NDEQHK, NEQHRK, FVLIM, HYF). Pl.LSU/2 IEP RT blocks 1 to 7 are indicated by rectangles. The RT7 block presenting no conservation between the two sequences is indicated in black.

E. coli BL21 (DE3) strain, from which are derived BL21 Star (DE3) and ArcticExpress (DE3)RIL strains, thus contains an ORF encoding an RNA-dependent DNA polymerase, which could correspond to the protein that bias the GST-tagged IEP RT activity assays. In these conditions, it is not possible to determine if the IEP presents an RT activity. Even though the contaminant *E. coli* protein shows an RT activity in our experiments, it does not necessarily imply that IEP is inactive. Moreover, the cleavage of GST from GST-tagged IEP should probably not overcome this problem, as the contaminating protein seems to co-eluate with IEP.

To circumvent this contamination bias, we decided to express the IEP in other expression systems. We chose to evaluate the expression of the Pl.LSU/2 IEP with the cell-free and the Sf9/baculovirus expression systems.

2.2 - CELL-FREE EXPRESSION SYSTEM

Cell-free protein synthesis is an attractive alternative to *E. coli*, since it offers a simple approach to rapid synthesis of properly folded proteins. Several improvements have been made in this field, allowing the production of large amount of functional proteins in a modified *E. coli* cell-free lysate (Klammt C et al. 2006). In addition, *E. coli* cell-free lysate allows a higher productivity than eukaryotes lysates, as it has been shown that the rate of peptide growth is five to ten times slower in eukaryotes (Netzer WJ and Hartl FU 1997; Hartl FU and Hayer-Hartl M 2002). We thus decided to evaluate the IEP expression with a cell-free system using an *E. coli* lysate. We used the p151-IEP expression plasmid (See Appendix plasmid map) initially constructed to express the IEP tagged in N-terminal with an histidine tag in *E. coli*. This plasmid is compatible with the cell-free expression system as described in the manufacturer's instructions (Expressway™ Cell-free *E. coli* expression system; Life Technologies, Invitrogen).

The histidine tag (also called His-tag) is one of the most widely used purification tags. It generally consists in six (6xHis) histidine residues. The small size of the His-tag usually minimizes interference with the folding and structure of the target protein (Carson M et al. 2007). Histidine residues can readily coordinate with metal ions such as Ni^{2+} immobilized on a resin for purification. If exposed on the surface of the protein, it should bind to a Sepharose resin that has been charged with Ni^{2+} allowing non-tagged proteins to pass straight through. Elution can then be carried out by imidazole or low pH, allowing pure or nearly pure protein to be prepared. This purification method, called Immobilized Metal ion Affinity Chromatography (IMAC), can be performed under native and denaturing conditions, since the His-tag does not require a specific conformation for metal binding. The binding of His-tagged proteins to Ni^{2+}-charged resins is usually more efficient under denaturing conditions as the His-tag becomes more exposed. However, purify a protein under denaturing conditions implies to refold the protein and this does not ensure a recovery of the catalytic protein conformation.

In p151-IEP plasmid, the IEP sequence is placed downstream of a stretch of six histidine residues (Fig. R-15A; 6xHis) and a V5 epitope tag, allowing the expression of the IEP fused in N-terminus with His and V5 tags (HisV5-IEP). His and V5 tags can be removed from the IEP by the use of the Tobacco Etch Virus (TEV) protease that cleave at the TEV recognition site located between the HisV5 tag and the IEP sequence (R-15A, TEV). Expression of the fusion HisV5-IEP is driven by a T7 promoter (R-15A; T7) recognized by the T7 RNA polymerase of the cell-free expression kit. The template was prepared according to the manufacturer's instructions. The control plasmid pEXP5-NT/CALML3 (See Appendix plasmid map) supplied with the kit and expressing was used to express the control His-tagged human calmodulin-like 3 protein (His-CALML3) (R-15A). The first aim was to determine if a sufficient yield of HisV5-IEP could be produced by this *in vitro* translation system. The expression of HisV5-IEP and His-CALML3 was performed according to the manufacturer's instructions during 6 hrs at 30°C (Fig. R-15B).

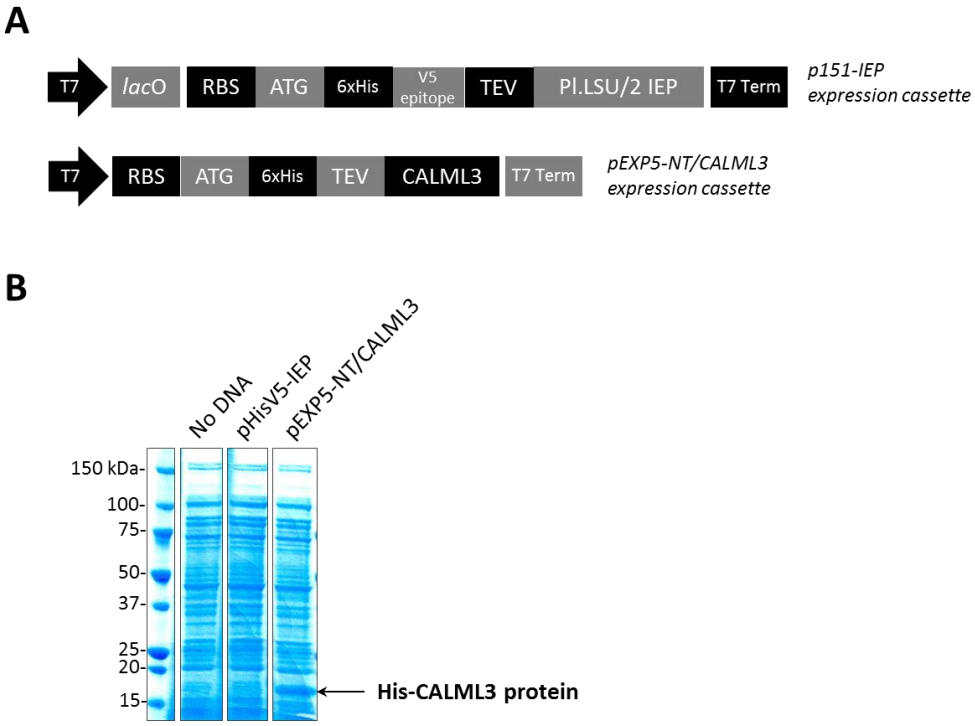


Figure R-15: Cell-free expression of HisV5-IEP.

(***A***) Schematic representation of expression cassettes used. T7: T7 promoter; *lac*O: *lac* Operator; RBS: Ribosome Binding Site; ATG: translation initiation codon; 6xHis: Stretch of six histidine residues; V5: V5 epitope; TEV: Tobacco Etch Virus protease recognition site; T7 Term: T7 terminator; CALML3: human calmodulin-like 3 protein. (***B***) Coomassie blue stained SDS-PAGE gel of total protein fractions. Cell-free expression of HisV5-IEP and His-CALML3 was performed as described by the manufacturer's instructions during 6 hrs at 30°C. A condition without expression plasmid (No DNA) was also performed as a control.

The HisV5-IEP fusion protein is expected to have a molecular mass of 69-kDa. Unfortunately, figure R-15B shows that no supplemental expressed protein is detected between the HisV5-IEP condition (Fig. R-15B; HisV5-IEP) and the negative control condition (Fig. R-15B; No DNA), in which no plasmid DNA was added. The HisV5-IEP fusion protein is not overexpressed by the cell-free system in the conditions used. In contrast, a protein that ran between the 15-kDa and the 20-kDa markers and probably corresponding to the control 19.5-kDa His-CALML3 protein is overexpressed with the control pEXP5-NT/CALML3 plasmid, as shown on the SDS-PAGE (Fig. R-15B; His-CALML3 protein).

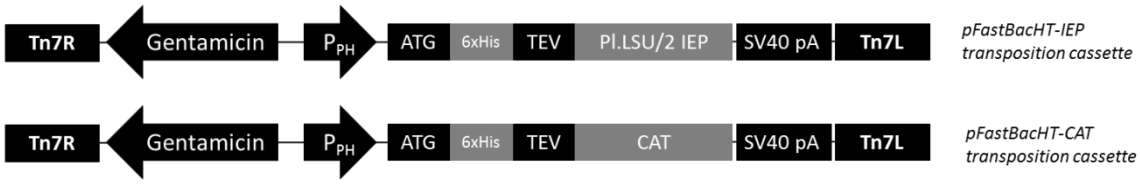
We used here all conditions required to obtain the highest protein yield such as the incubation temperature and time used, as recommended by the manufacturer. Nevertheless, the cell-free expression system did not allow the expression of the HisV5-IEP in a sufficient yield that could be detectable by a Coomassie staining. Although a western blot analysis could have determined if the HisV5-IEP was expressed at a low level, it was decided to drop out these experiments, as high yield of protein is required for biochemical analyses, and further optimizations with this expression system are quite limited. We thus decided to evaluate the IEP expression in insect Sf9 cells using the baculovirus expression system.

2.3 - BACULOVIRUS EXPRESSION SYSTEM

The baculovirus has been commonly used for the production of recombinant proteins in insect cells and baculovirus/insect cell expression system has been widely reviewed (Patterson RM et al. 1995; Kost TA et al. 2005). Recombinant baculoviruses allowing the expression of IEP in Sf9 insect cells were generated using a method based on site-specific transposition of an expression cassette into a baculovirus shuttle vector propagated in *E. coli* (Luckow VA 1993). We used the Bac-to-Bac® Baculovirus expression system (Life Technologies, Invitrogen): the Pl.LSU/2 IEP sequence was cloned into a baculovirus donor plasmid downstream of a stretch of six histidine residues (Fig. R-16A; 6xHis) (pFastBacHT-IEP, See Appendix plasmid map), allowing the expression of the IEP fused in N-terminal with a His tag (His-IEP). His-IEP molecular mass is expected to be around 69-kDa. The expression of His-IEP is driven by the strong polyhedrin promoter P_{PH} (Fig. R-16A; P_{PH}), which is activated in the very late phase of baculovirus infection, starting from 18-24 hrs postinfection. A TEV recognition site (Fig. R-16A; TEV) is located between the His-tag and the IEP sequence and could be used for the cleavage of the His-tag from the fusion protein. A donor plasmid supplied with the kit (pFastBac™HT-CAT, See Appendix plasmid map) containing the fusion protein His-CAT encoding sequence (Chloramphenicol Acetyltransferase) is also used as a control (Fig. R-16A). The protocol used to generate recombinant baculoviruses and express His-tagged proteins is described in Materials and methods.

Sf9 cells were infected by recombinant baculoviruses at a multiplicity of infection (MOI) of 5. Seventy-two hours after infection, Sf9 cells were pelleted and lysed to obtain both soluble (S) and insoluble (I) protein fractions. Soluble fractions were used for His-IEP and His-CAT purification by IMAC. The soluble protein fraction was loaded onto a Ni^{2+}-charged resin. The flow-through (Ft) was collected and the resin was washed four times using three wash buffers containing increasing concentrations of imidazole (W_1, W_2, W_3 and W_4). Eight fractions of purified proteins (P_1 to P_8) were collected from one elution with 1 M of imidazole. The Ni^{2+}-charged resin (R) was also analyzed to evaluate the amount of proteins that remain bound to the resin after elution (Fig. R-16B).

A



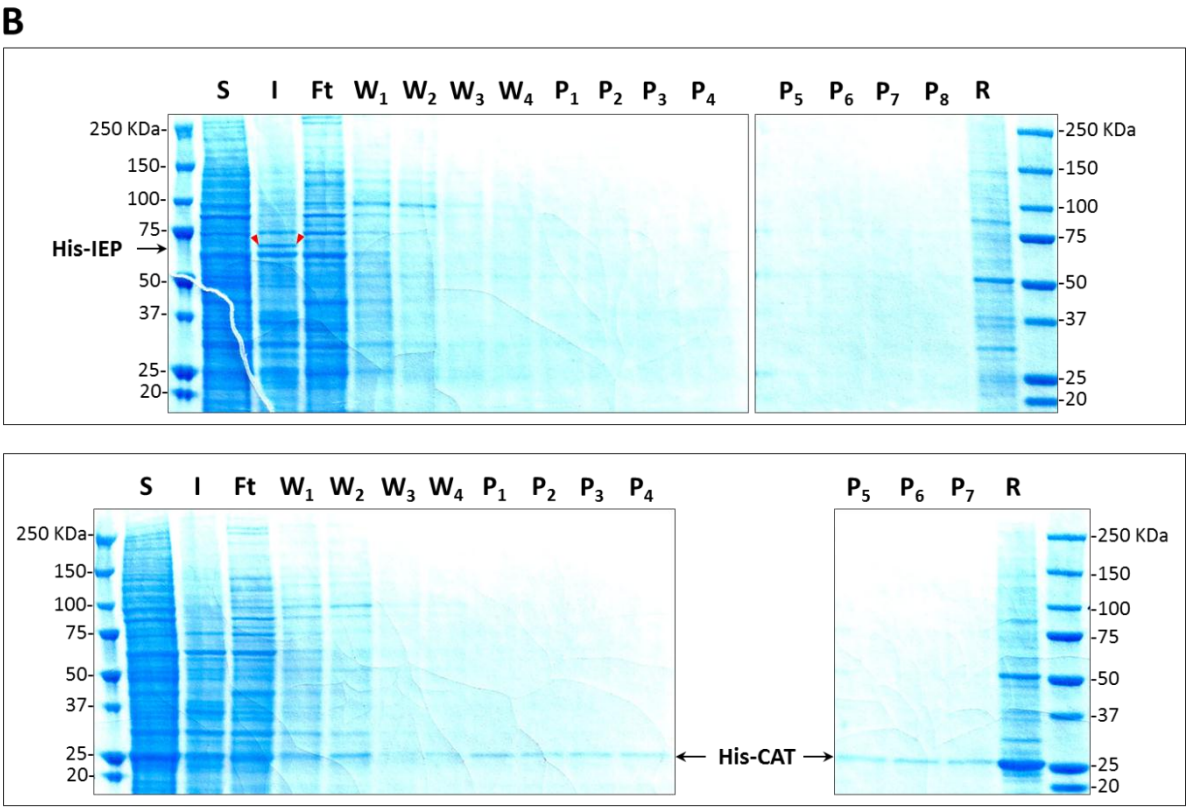


Figure R-16: His-IEP expression by the baculovirus/Sf9 system.

(***A***) Schematic representation of expression cassettes of pFastBacHT-IEP and –CAT vectors. Tn7R and Tn7L: mini Tn7 elements permitting site-specific transposition into baculovirus genome (bacmid DNA); Gentamicin: Gentamicin resistance gene used for selection recombinant bacmid DNA in *E. coli*; P_{PH}: Polyhedrin promoter; ATG: translation initiation codon; 6xHis: Stretch of six histidine residues; TEV: Tobacco Etch Virus protease recognition site; Pl.LSU/2 IEP: Pl.LSU/2 IEP coding sequence; CAT: Chloramphenicol Acetyltransferase coding sequence; SV40 pA: SV40 polyadenylation signal. (***B***) Coomassie blue stained SDS-PAGE gel of purification fractions. *Upper panel*: His-IEP purification. *Lower panel*: His-CAT purification. S: soluble protein fraction (1/500 of the fraction); I: insoluble protein fraction (1/150 of the fraction); Ft: Flow-through from the Ni^{2+}-charged resin after binding of His-tagged protein (1/500 of the fraction); W_1 to W_4: wash protein fractions 1 to 4 1/500 of the fractions); P_1 to P_8: purified protein fractions eluted from the resin (1/35 of the fractions); R: proteins which remain bound to the resin after elutions (1/70 of the fraction). His-IEP is indicated by red arrowheads. Numbers at *left* indicate molecular mass marker.

Figure R-16B shows that His-IEP, expressed from Sf9 cells, is mainly found in the insoluble protein fraction (Fig. R-16B; upper panel, fraction I, indicated by the red rectangle). The protein is at the expected size of 69-kDa. The soluble protein fraction, used for His-IEP purification, contains almost no His-IEP: the protein is thus not detectable in any purification fractions, from the flow-through to the resin fraction (Fig. R-16B; upper panel). In contrast, the control His-CAT protein, which is at the expected size of 28-kDa, is mostly expressed in its soluble form (Fig. R-16B; lower panel, fraction S). The IMAC purification process was efficient to purify this protein, as observed by the SDS-PAGE analysis (Fig. R-16B; lower panel).

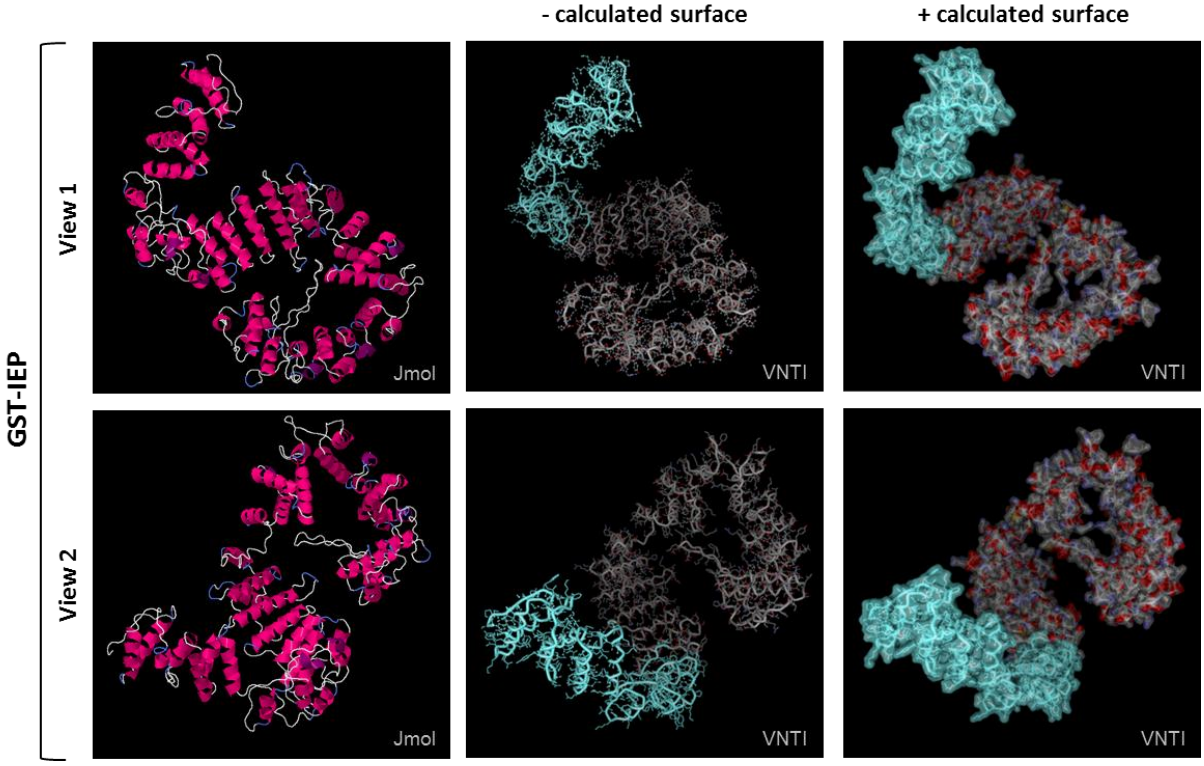
The expression and solubility of His-IEP in Sf9 cells should be optimized in order to allow its purification under native conditions. Western blot analysis could also have determined if a small

amount of His-IEP is soluble and purified. However, I did not perform these experiments, because optimizations are again limited with the baculovirus/Sf9 expression system. Three major parameters can be adjusted, as the MOI, the time of cell harvest and the cell line, but these adjustments should mainly impact on the protein yield, not on its solubility. For this reason and also because of the great flexibility and facility to work with *Escherichia coli*, we have decided to evaluate the expression of a His-tagged IEP in *E. coli*.

2.4 - HIS-TAGGED IEP IN *E. COLI*

The results obtained for RT assays with GST-tagged IEP purified fractions showed a RT activity using both wild-type and mutants GST-IEP fractions. It was concluded that a contaminant *E. coli* reverse transcriptase protein, which has been co-purified with the GST-tagged proteins, has bias the RT reactions. The use of a different tag could induce a different conformation of the fusion protein and impede the purification of this contaminant protein and thus circumvent this problem. We chose the HisV5 tag, used in the cell-free expression system (See Results section 2.2 -), because of its small size which should not alter the IEP conformation.

To evaluate the influence of GST and HisV5 tags on the IEP conformation, GST-IEP and HisV5-IEP tridimensional conformations were predicted using the I-TASSER server (Zhang Y 2008; Roy A et al. 2010) (http://zhanglab.ccmb.med.umich.edu/I-TASSER/). The tridimensional model presenting the higher confidence level was retrieved for each protein. The modeling accuracy is indeed estimated by calculation of the C-score (-1.5 for GST-IEP and -1.00 for HisV5-IEP models). Models with C-score ≥ -1.5 are expected to have a correct fold (Roy A et al. 2010). The predicted 3D structure models were then analyzed using Jmol (Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/) and Vector NTI® (Life Technologies) softwares (Fig. R-17).



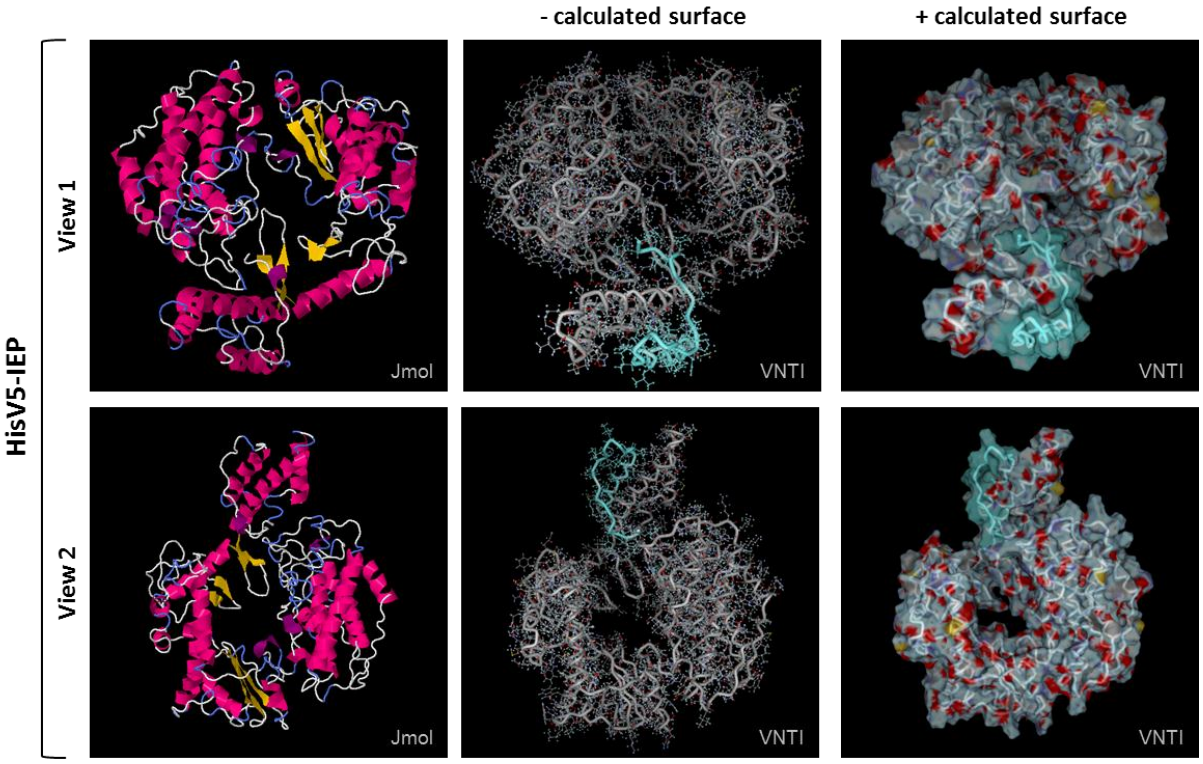


Figure R-17: Predicted 3D structures of GST-IEP and HisV5-IEP.

The 3D structure models of GST-IEP and HisV5-IEP with the highest confidence level predicted using the I-TASSER server were analyzed with the Jmol software (Jmol) and the 3D molecule viewer component of the Vector NTI® software (VNTI). Two views of each protein predicted structure are shown (views 1 and 2). In Jmol views, α-helixes are in magenta, 3_{10}-helixes are in purple, β-strands are in yellow, and turns are in blue. Atoms are not represented in Jmol views. In VNTI, the surface of each predicted molecule was calculated (+ calculated surface) or not (- calculated surface). Atoms are represented in the "ball and stick" style (carbon in gray, hydrogen in pale blue, nitrogen in pale purple, oxygen in red, and sulfur in yellow). The tag sequences (GST and HisV5) are colored in cyan.

Figure R-17 shows that GST-IEP and HisV5-IEP predicted 3D structures are different. Indeed, GST-IEP is predicted to contain only α- helixes, 3_{10}- helixes, and turns, while HisV5-IEP predicted structure shows three antiparallel β-sheets (Fig. R-17; Jmol views). The overall structures (Fig. R-17; - calculated surface, VNTI views) and the predicted protein surfaces (Fig. R-17; + calculated surface, VNTI views) seem quite different: GST-IEP presents a relaxed structure, while HisV5-IEP structure appears to be more compacted, forming a hole in the center of the structure. Although these results are only theoretical, the use of the small HisV5 tag could allow the IEP to adopt a different conformational structure and this feature may possibly overcome the co-purification of the *E. coli* contaminant protein showed in GST-IEP (WT and mutants) purified fractions.

2.4.1 - Expression in BL21 Star (DE3) pRARE and purification under native conditions

The plasmid p151-IEP was initially designed to express the Pl.LSU/2 IEP in *E. coli*. This plasmid places the IEP ORF, fused in its N-terminal to a 6xHis stretch and a V5 epitope, downstream of a T7/*lac* promoter (Fig. R-18A). The T7/*lac* promoter is specifically recognized by the T7 RNA polymerase, which is expressed by the DE3 bacteriophage lambda lysogenic. This bacteriophage must

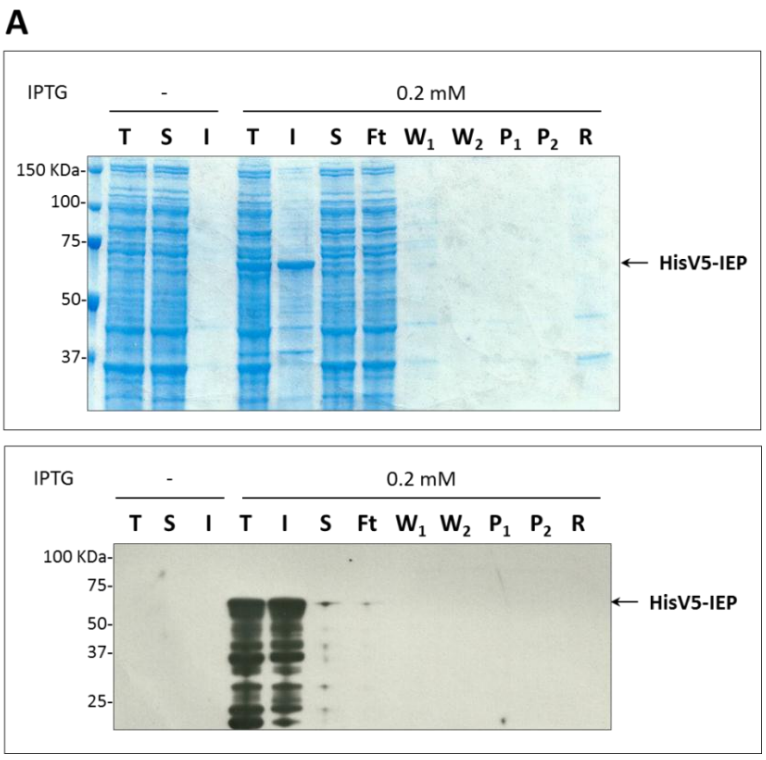
be integrated in the used *E. coli* strain to express His-tagged protein. The expression of T7 RNA polymerase is driven by the *lacUV5* promoter, inducible by IPTG. The T7/*lac* promoter contains also a *lac* operator sequence immediately downstream of the T7 promoter used to regulate basal expression of the protein (Fig. R-18A). A TEV recognition site (Fig. R-18A; TEV) is also located between the IEP sequence and the tags to allow their removal with the TEV protease. The expression and purification were first optimized with the wild-type version of HisV5-IEP before purification of mutants. The p151-IEP plasmid was transformed in the BL21 Star (DE3) pRARE *E. coli* strain. This strain was chosen according to the results obtained with the GST-IEP purification, which showed that a good yield of soluble GST-IEP could be expressed in this strain. HisV5-IEP expression was induced using several IPTG concentrations in order to determine the best induction conditions to obtain high amount of soluble proteins. Induction was performed, as for GST-IEP expression, at 18°C for 3 hrs (Fig. R-18B).

Figure R-18: HisV5-IEP expression in BL21 Star (DE3) pRARE using different IPTG concentrations.
(***A***) Schematic representation of HisV5-IEP expression cassette. T7: T7 promoter; *lac*O: *lac* Operator; RBS: Ribosome Binding Site; ATG: translation initiation codon; 6xHis: Stretch of six histidine residues; V5: V5 epitope; TEV: Tobacco Etch Virus protease recognition site; T7 Term: T7 terminator. (***B***) Protein fractions were loaded onto a SDS-PAGE gel. HisV5-IEP expression was induced from 2 ml *E. coli* culture at OD_{600nm} 0.5 with

0.1 mM, 0.2 mM or 0.5 mM of IPTG. A negative control was also performed from 2 ml of *E. coli* culture without IPTG induction (-). T: total protein fraction (1/20 of the fractions); S: soluble protein fraction (1/20 of the fractions); I: insoluble protein fraction (1/8 of the fractions). *Upper panel*: Coomassie blue stained SDS-PAGE gel. *Lower panel*: Western blot analysis. A monoclonal mouse HRP-conjugated anti-V5 antibody was used to detect the IEP. Numbers at *left* indicate molecular mass marker.

SDS-PAGE shows that a protein that ran below the 75-kDa marker is expressed in *E. coli* whatever the IPTG concentration used and is at the expected size of HisV5-IEP (69-kDa) (Fig. R-18B; upper panel, 0.1 to 0.5 mM IPTG, fractions T). Western blot analysis confirms that this protein is HisV5-IEP (Fig. R-18B; lower panel, 0.1 to 0.5 mM IPTG). The amount of HisV5-IEP increases with the concentration of IPTG used (Fig. R-18B; upper panel, 0.1 to 0.5 mM IPTG, fractions T). But in the meantime, the protein becomes more insoluble (Fig. R-18B; 0.1 to 0.5 mM IPTG, fractions I). The soluble protein is detected by Coomassie blue staining at 0.1 mM of IPTG (Fig. R-18B, upper panel, 0.1 mM IPTG, fraction S) and its yield also increases with IPTG concentration. These results are confirmed by western blot analysis (Fig. R-18B; lower panel). HisV5-IEP is subjected to very little degradation (Fig R-18B; lower panel). Soluble HisV5-IEP can thus be expressed in *E. coli* under these conditions.

To obtain high yield of soluble HisV5-IEP purified in native conditions, the expression rate has to be equilibrated. The solubility of the protein is improved by low expression rate; however it is important to maintain a sufficient *E. coli* growth to maximize the yield of expressed proteins. In this context, a small scale IMAC purification assay in native conditions was performed as described in Results section 2.2 -. Ten ml *E. coli* cultures transformed with p151-IEP were induced at 18°C for 3 hrs with 0.2 mM or 1 mM of IPTG (Fig. R-19).



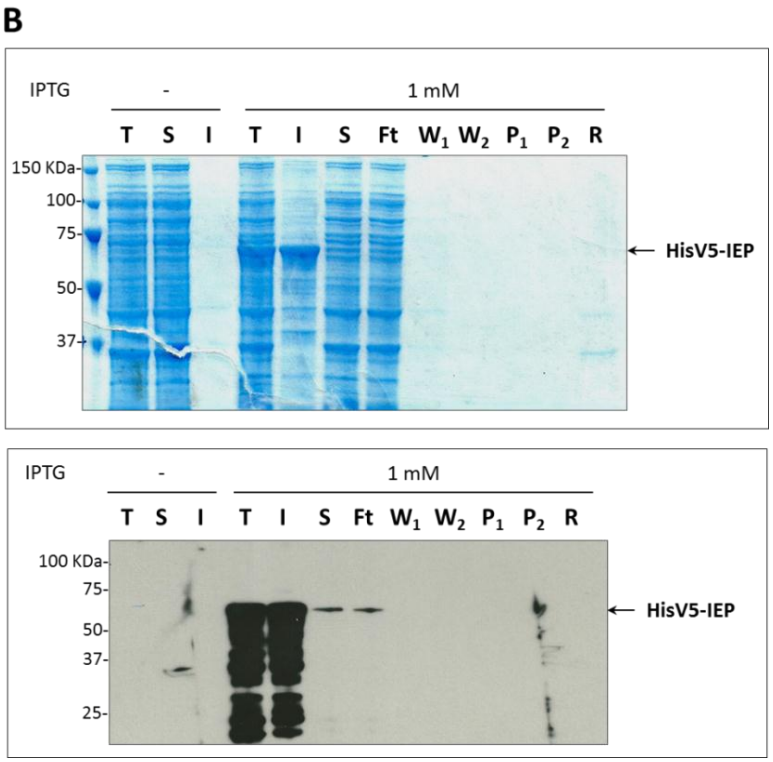


Figure R-19: Purification in native conditions of HisV5-IEP expressed from *E. coli* BL21 Star (DE3) pRARE.

HisV5-IEP expression was induced from 10 ml *E. coli* culture at OD_{600} 0.6 with 0.2 mM (***A***) or 1 mM (***B***) of IPTG. A negative control was also performed from 5 ml of *E. coli* culture without IPTG induction (-). Protein fractions were loaded onto SDS-PAGE gels. T: total protein fraction (1/35 for the T(-) fraction, 1/70 for the T(0.2 and 1 mM) fraction); S: soluble protein fraction (1/35 for the S(-) fraction, 1/70 for the S(0.2 and 1 mM) fraction); I: insoluble protein fraction (1/32 for the I- fraction, 1/64 for the I+ fraction); Ft: Flow-through from the Ni^{2+}-charged resin after binding of HisV5-IEP (1/70 of the fraction); W_1 and W_2: wash protein fractions 1 and 2 (1/35 of the fractions); P_1 and P_2: purified protein fractions successively eluted from the resin (1/3.5 of the fraction); R: proteins which remain bound to the resin after elutions (1/2 of the fraction). Upper panels: Coomassie blue stained SDS-PAGE gels. Lower panels: Western blots. A monoclonal mouse HRP-conjugated anti-V5 antibody was used to detect the IEP. Numbers at *left* indicate molecular mass marker.

Coomassie blue stained gels show that HisV5-IEP is mainly insoluble, whatever the IPTG concentration used (Fig. R-19A and B; upper panels). As expected, the protein yield increases with IPTG concentration. It seems that all soluble HisV5-IEP comes out in the flow-through during the purification (Fig. R-19A and B; upper panels, fractions Ft). This result is confirmed by western blot analysis (Fig. R-19A and B; lower panels, fractions Ft). It is apparent that HisV5-IEP does not bind to the Ni^{2+}-charged resin under conditions used. It suggests that the His-tag could be hidden inside the tertiary structure of the protein so that the binding to the resin cannot occur. We can also notice that HisV5-IEP is subjected to degradation in this experiment (Fig. R-19A and B; lower panels, fractions T and I).

A purification of the protein under denaturing conditions could circumvent the non-binding issue. Indeed, the HisV5-tag would then be more exposed allowing the fusion protein to bind the resin. A refolding step must follow the purification in these conditions in order to characterize the biochemical

activities of the HisV5-IEP, but it is noteworthy that recovering of the protein in its proper catalytic conformation is not ensured.

2.4.2 - Expression in BL21 Star (DE3) pRARE, purification in denaturing condition and RT activity assay

The HisV5-IEP expressed in *E. coli* BL21 Star (DE3) pRARE strain was thus purified under denaturing conditions using either 6 M of guanidine hydrochloride (Gu-HCl) or 8 M of urea. Purifications were followed by a refolding step consisting in successive dialyses against buffers containing less and less denaturants (See Materials and methods section 3.4.2 -; Fig. R-20). This multi-step refolding strategy should allow a relatively slow refolding process, which would prevent the aggregation of improperly folded proteins.

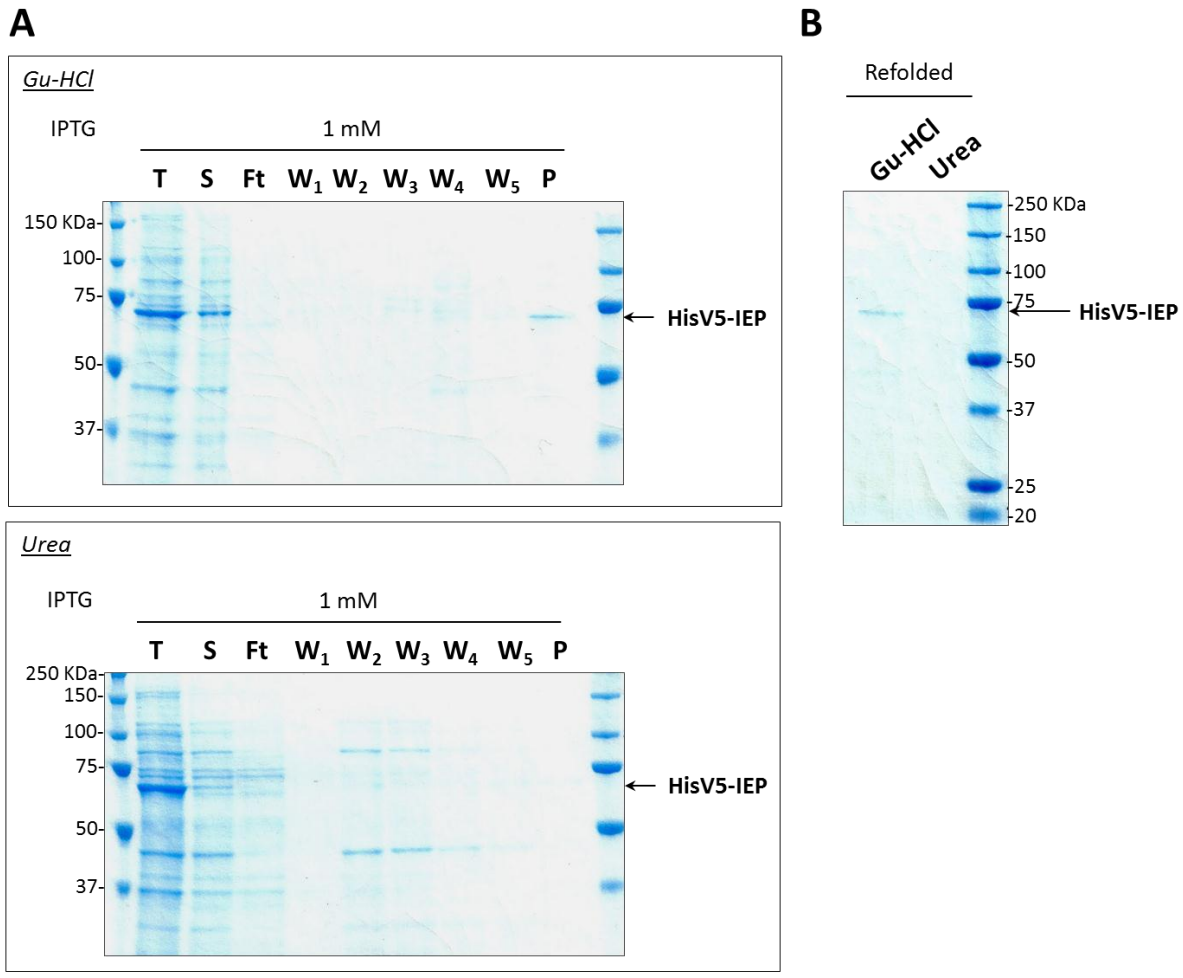


Figure R-20: Purification of HisV5-IEP under denaturing conditions.

(***A***) HisV5-IEP expression was induced from 200 ml *E. coli* culture at OD_{600nm} 0.6 with 1 mM of IPTG for 3 hrs at 18°C. A negative control was also performed from 10 ml of *E. coli* culture without IPTG induction (not shown). T: total protein fraction (1/300 of the fractions); S: soluble protein fraction (1/300 of the fractions); Ft: Flow-through from the Ni^{2+}-charged resin after binding of HisV5-IEP (1/300 of the fractions); W_1 to W_5: wash protein fractions 1 to 5 (1/1000 of the fractions); P: purified protein fraction eluted from the resin (1/20 of the fractions). *Upper panel*: Coomassie blue stained SDS-PAGE gel of protein fractions collected during purifications of HisV5-IEP performed using guanidine hydrochloride (Gu-HCl) from 100 ml of induced *E. coli*

culture. *Lower panel*: Coomassie blue stained SDS-PAGE gel of protein fractions collected during purifications of HisV5-IEP performed using urea. (***B***) Coomassie blue stained SDS-PAGE gel of refolded protein fractions (1/20 of the fractions). Proteins purified under denaturing conditions, using either guanidine hydrochloride (Gu-HCl) or urea, were refolded by a multi-step dialysis process. Numbers at *left* indicate molecular mass marker.

Figure R-20A shows first that the HisV5-IEP solubilization efficiency depends on the denaturant used. Guanidine hydrochloride allows a high solubilization of HisV5-IEP (Fig. R-20A; upper panel, fractions T and S), while only a small amount of HisV5-IEP is solubilized with urea (Fig. R-20A; lower panel, fractions T and S). The HisV5-IEP denaturation induces an improvement of the binding of the protein to the Ni^{2+}-charged resin, as the HisV5-IEP is not detected in the flow-through (Fig. R-20A; fractions Ft). HisV5-IEP is found in the purified protein fractions when using GuHCl (Fig. R-20A; upper panel, fractions P). The purity of this purified fraction is very high (Fig. R-20A; upper panel, fraction P). The fact that HisV5-IEP denatured by urea is not detectable in the purified fraction (Fig. R-20A; lower panel, fraction P) is due to the very low amount of HisV5-IEP in the soluble fraction used for purification. Those purified fractions were dialyzed by a multi-step dialysis process in order to refold the proteins (Fig. R-20B). SDS-PAGE of refolded proteins shows that almost no HisV5-IEP is lost during the process. Again, HisV5-IEP previously purified with Gu-HCl is well detected by Coomassie blue staining in contrast to HisV5-IEP purified by urea (Fig. R-20B) due to the difference of HisV5-IEP concentration in both purified fractions (Fig. R-20A; fractions P). These results showed that high yield of nearly pure HisV5-IEP can be purified under denaturing conditions using Gu-HCl. This highly pure fraction could thus be used to assay the RT activity of the IEP.

The RT assay requires the use of a negative control protein. Thus, the control plasmid p151-IEPmtDD- was constructed. This plasmid should express a mutant RT-defective HisV5-IEP (HisV5-IEP mtDD-) in which the catalytic motif YADD is replaced by YAAA. HisV5-IEP mtDD- was subsequently purified under Gu-HCl denaturing conditions and refolded by the multi-step dialysis (Fig. R-21).

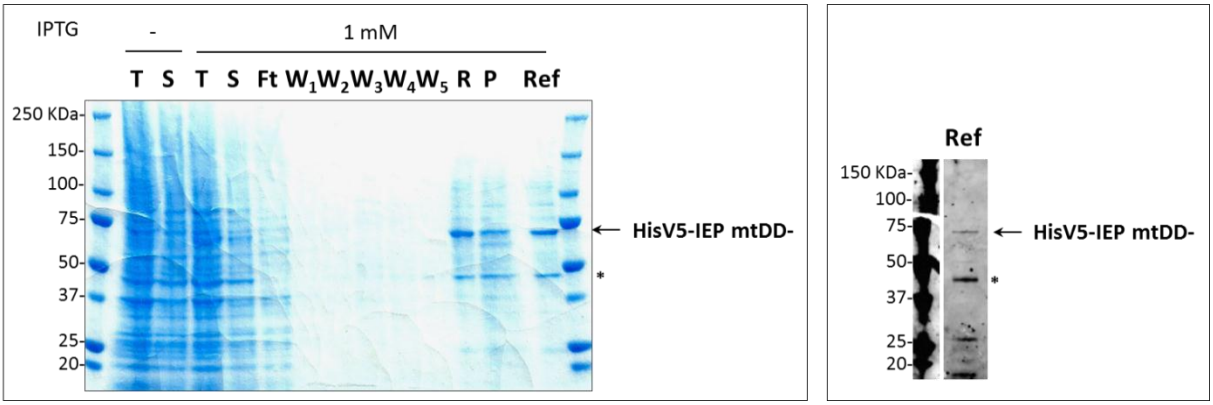


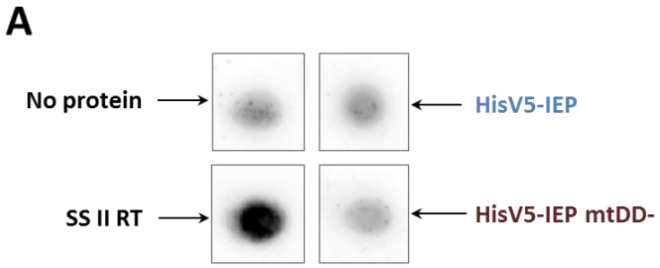
Figure R-21: Purification of HisV5-IEP mtDD- under denaturing conditions using Gu-HCl.

HisV5-IEP mtDD- expression was induced from 100 ml *E. coli* culture at OD_{600nm} 0.6 with 1 mM of IPTG at 18°C for 3 hrs. A negative control was also performed from 5 ml of *E. coli* culture without IPTG induction. *Left panel*: Coomassie blue stained SDS-PAGE gel of protein fractions collected during purifications of HisV5-IEP mtDD- performed using Gu-HCl. T: total protein fraction (1/20 of the T(-) fraction and 1/300 of the T(1 mM) fraction); S: soluble protein fraction (1/20 of the S(-) fraction and 1/300 of the S(1 mM) fraction); Ft: Flow-through from the Ni^{2+}-charged resin after binding of HisV5-IEP (1/300 of the fraction); W_1 to W_5: wash protein fractions 1 to 5 (1/1000 of the fractions); R: proteins which remain bound to the resin after elution (1/10 of the

fraction); P: purified protein fraction eluted from the resin (1/20 of the fraction); Ref: refolded protein fraction using multi-step dialysis process (1/20 of the fraction). *Right panel*: Western blot analysis of refolded protein fraction ((1/20 of the fraction) using a mouse anti-V5 antibody. Asterisk indicate HisV5-IEP mtDD- degradation product. Numbers at *left* indicate molecular mass marker.

Figure R-21 shows that HisV5-IEP mtDD- is over-expressed in *E. coli* (Fig. R-21; left panel, 1 mM IPTG, fraction T) and well solubilized by GuHCl (Fig. R-21; left panel, 1 mM IPTG, fraction S). The YAAA mutation has no adverse impact on the protein expression. High yield of nearly pure HisV5-IEP mtDD- can also be obtained in those denaturing conditions (Fig. R-21; left panel, 1 mM IPTG, fraction P) and the refolding step allow the recovery of all HisV5-IEP mtDD- (Fig. R-21; left panel, 1 mM IPTG, fraction Ref). Notably, a protein that ran between the 37-kDa and the 50-kDa markers is co-purified in this experiment (Fig. R-21; left panel, indicated by asterisk). This protein appears to be a degradation product of the mutant IEP, as shown by western blot (Fig. R-21; right panel). Purified and refolded IEP (WT and mtDD-) protein fractions can thus be assayed for RT activity.

The RT activity of HisV5-IEP and mutant HisV5-IEP mtDD-, purified under denaturing conditions with Gu-HCl and refolded by multi-step dialysis, was subsequently assayed *in vitro* using poly(rA)-oligo$(dT)_{12\text{-}18}$ template. RT activity is determined by incorporation of [α-^{32}P]dTTP nucleotides. RT reactions were performed using either wild-type (WT) or mutant (mtDD-) HisV5-IEP as the source of reverse transcriptase. Control experiments were also performed using either the dialysis buffer corresponding to the background (No protein), or the SuperScript® II reverse transcriptase (SS II RT) as positive control. Radioactive products were spotted on DE81 filter and exposed on a phosphor screen. The Image obtained highlights the presence of radioactive products (Fig. R-22A). Quantification of reactions was then performed using image analysis software by determining the number of pixels per spot, which is correlated to the RT activity of the protein used (Fig. R-22B).



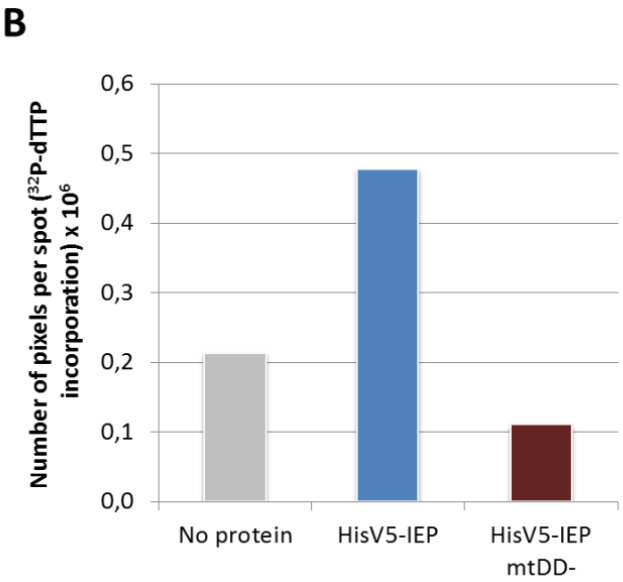


Figure R-22: RT assay with HisV5-IEP and HisV5-IEP mtDD- purified under denaturing conditions.
(***A***) 8 µl of refolded protein fractions were used. RT reactions without proteins (No protein) and with HisV5-IEP and HisV5-IEP mtDD- mutant were performed at 37°C for 45 min. Positive control consists of 0.05 U of SuperScript® II reverse transcriptase (SS II RT). (***B***) Data, representing the number of pixels per spot, were quantified with ImageQuant™ software. Light gray bar: No protein; blue bar: HisV5-IEP; dark red bar: HisV5-IEP mtDD-.

Figure R-22A shows that the background condition (No protein) does not present any signal. Data were subsequently quantified (Fig. R-22B). No statistically significant differences could be observed between the wild-type IEP (Fig. R-22B; HisV5-IEP), the mutant IEP (Fig. R-22B; HisV5-IEP mtDD-) and the background (Fig. R-22B; No protein). This experiment was repeated and no RT activity could be demonstrated using purified protein fractions obtained under denaturing conditions. This result suggests that the conditions used to purify and/or refold HisV5-IEP did not allow the protein to recover its active conformation. In this context, we chose to test the purification of HisV5-IEP under non-denaturing conditions.

2.4.3 - Expression in Rosetta-gami B (DE3), purification in non-denaturing conditions, and RT activity assay

In the attempt to further optimize the expression of soluble HisV5-IEP in *E. coli*, we analyzed the predicted disulfide bonds formation in HisV5-IEP. Indeed, the proper folding of proteins containing cysteine residues may involve the formation of disulfide bonds. Notably, the prediction of disulfide bridges in HisV5-IEP using the DIpro software (http://scratch.proteomics.ics.uci.edu/) (Cheng J et al. 2006) reveals that three disulfide bonds can occur (Table R-23).

Bond number	Cys1 Position	Cys2 Position
1	207	175
2	133	57
3	538	546

Table R-23: Predicted disulfide bonds in HisV5-IEP sequence.

The prediction was performed using DIpro 2.0 software (http://scratch.proteomics.ics.uci.edu/). The predicted disulfide bonds (Bond number 1 to 3) are ordered by probability in descending order. Cysteines predicted to be involved in these bonds (Cys1 and Cys2 position) are at position 57, 133, 175, 207, 538 and 546. The total number of cysteines in HisV5-IEP sequence is 9.

Among the 9 cysteines contained in HisV5-IEP, 6 are predicted to form disulfide bonds (Table R-23). In this context, we decided to use a more adapted *E. coli* strain: Rosetta-gami B (DE3). This strain combines the key features of *E. coli* Tuner, Origami and Rosetta strains. *E. coli* Tuner strain, which is a derivative of BL21 strain, enables adjustable levels of protein expression throughout all cells in a culture. Indeed, this *lacZY* deletion mutant allows uniform entry of IPTG into all cells, which produces a homogeneous level of induction. The *E. coli* Rosetta strain carries the pRARE plasmid, thus overcoming the codon bias. Finally, *E. coli* Origami strain contains mutations in thioredoxine reductase and glutathione reductase that enhance disulfide bonds formation. Protein expression in Rosetta-gami B (DE3) strain is expected to yield about 10-fold more active proteins than in another host.

HisV5-IEP and mutants HisV5-IEP mtDD- and HisV5-IEP ΔRT5 (plasmid p151-IEPΔRT5) were expressed in *E. coli* Rosetta-gami B (DE3). The induction was performed at 30°C for 4 hrs, which was shown to be the optimal condition to obtain large amount of soluble proteins in these cells. Proteins were subsequently purified under non-denaturing conditions with CHAPS. Indeed, the CHAPS zwitterionic detergent was used here to further improve the solubilization of proteins. CHAPS detergent is known to solubilize proteins in their native state and without altering their native charge. In order to formulate purified proteins in a neutral buffer, a final step consisting in a dialysis of purified protein fractions against elution buffer lacking CHAPS was performed (Fig. R-24).

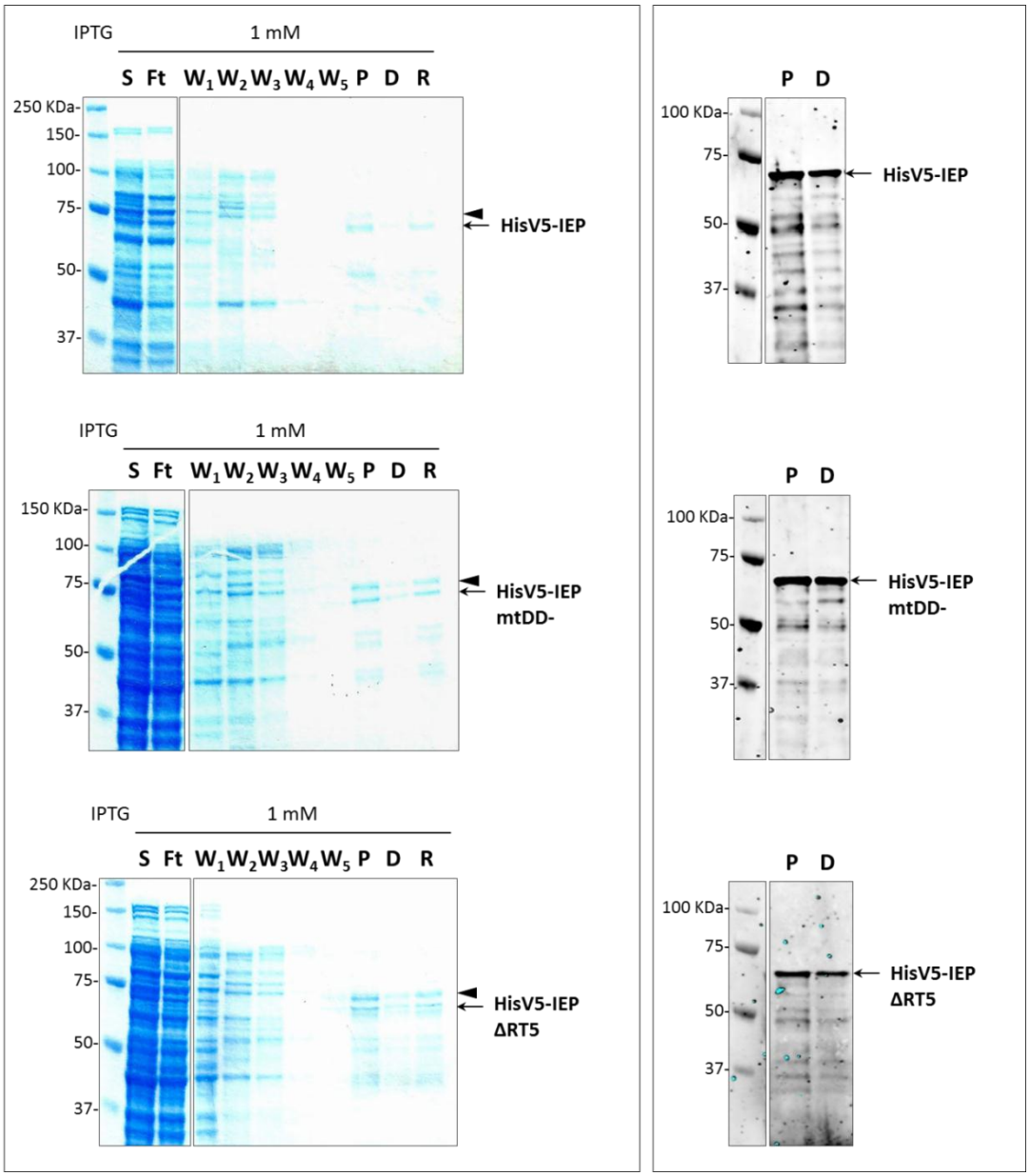


Figure R-24: Purification of HisV5-IEP and mutants under non-denaturing conditions with CHAPS.

HisV5-IEP and mutants (mtDD- and ΔRT5) expression was induced from 100 ml *E. coli* culture at OD_{600nm} 0.8 with 1 mM of IPTG for 4 hrs at 30°C. A negative control was also performed from 5 ml of *E. coli* culture without IPTG induction (not shown). *Left panel*: Coomassie blue stained SDS-PAGE gel of protein fractions collected during purifications performed with CHAPS. S: soluble protein fraction (1/450 of the fractions); Ft: Flow-through from the Ni^{2+}-charged resin (1/450 of the fractions); W_1 to W_5: wash protein fractions 1 to 5 (1/2500 of the fractions); P: purified protein fraction eluted from the resin (1/30 of the fractions); D: purified protein fraction after dialysis (1/30 of the fractions); R: proteins which remain bound to the resin after elution (1/15 of the fractions). Black arrowhead indicates *E. coli* co-purified protein contaminant. *Right panel*: Western blot analysis, using a mouse anti-V5 antibody, of purified protein fraction before (P; 1/30 of the fractions) and after (D; 1/30 of the fractions) dialysis. Numbers at *left* indicate molecular mass marker.

Figure R-24 shows that soluble HisV5-IEP and mutants are expressed in *E. coli* Rosetta-gami B (DE3) (Fig R-24; left panel, fractions S). These proteins are also purified with a relatively good yield and purity (Fig. R-24; left panel, fractions P). Nonetheless, a contaminant *E. coli* protein that ran just above HisV5-tagged proteins is co-purified in this experiment (Fig. R-24; left panel, indicated by

black arrowhead). Moreover, HisV5-IEP and mutants seem to undergo degradation, as showed by Coomassie blue staining (Fig. R-24; left panel) and western blot (Fig R-24; right panel). A non-negligible amount of HisV5-tagged proteins are also lost during dialysis (Fig. R-24; left panel, fractions D), probably because of a sticking of proteins to the dialysis membrane. These dialyzed protein fractions can still be used in an RT assay.

We therefore assayed the RT activity of HisV5-IEP and mutants (mtDD- and ΔRT5) purified under non-denaturing conditions with CHAPS, as described previously (Fig. R-25).

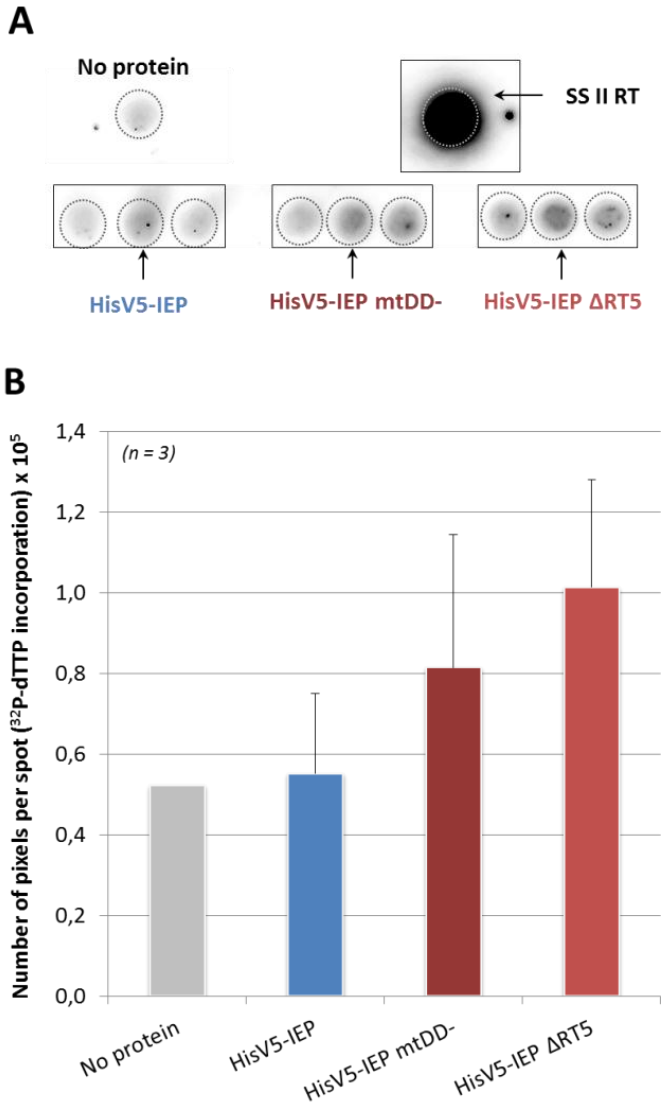


Figure R-25: RT assay with HisV5-IEP and mutants purified under non-denaturing conditions with CHAPS.

(***A***) 8 µl of dialyzed protein fractions were used. RT reactions without proteins (No protein) and with HisV5-IEP, HisV5-IEP mtDD- mutant and HisV5-IEP ΔRT5 mutant were performed at 37°C for 45 min. Positive control consists of 0.5 U of SuperScript® II reverse transcriptase (SS II RT). Reactions using HisV5-IEP and mutants were performed in triplicates. (***B***) Data, representing the number of pixels per spot, were quantified with ImageQuant™ software. Light gray bar: No protein; blue bar: HisV5-IEP; dark red bar: HisV5-IEP mtDD-; Light red bar: HisV5-IEP ΔRT5. Data are the mean of experimental triplicates and standard deviation is represented by thin lines.

Membrane image analysis shows that the background condition does not present any signal (Fig. R-25A; No protein). Data were subsequently quantified (Fig. R-25B). The RT assay using HisV5-tagged proteins purified under non-denaturing conditions with CHAPS (Fig. R-25B; HisV5-IEP) shows no significant differences of activity compared to the background or mutants (Fig. R-25B; HisV5-IEP mtDD- and HisV5-IEP ΔRT5). Again, this could be due to an inability of the IEP to achieve a correct folding and/or to an instability of its catalytic conformation in presence of CHAPS, even if CHAPS should theoretically not alter the native state of proteins.

We thus have decided to perform a purification of HisV5-IEP and its mutants under completely native conditions.

2.4.4 - Expression in Rosetta-gami B (DE3), purification in native conditions, and RT activity assay

We used here the Rosetta-gami B (DE3) *E. coli* strain as the expression host. A final dialysis step was performed to remove imidazole used during the elution and formulate purified proteins in a neutral buffer (Fig. R-26; See Materials and methods section 3.4.2 -).

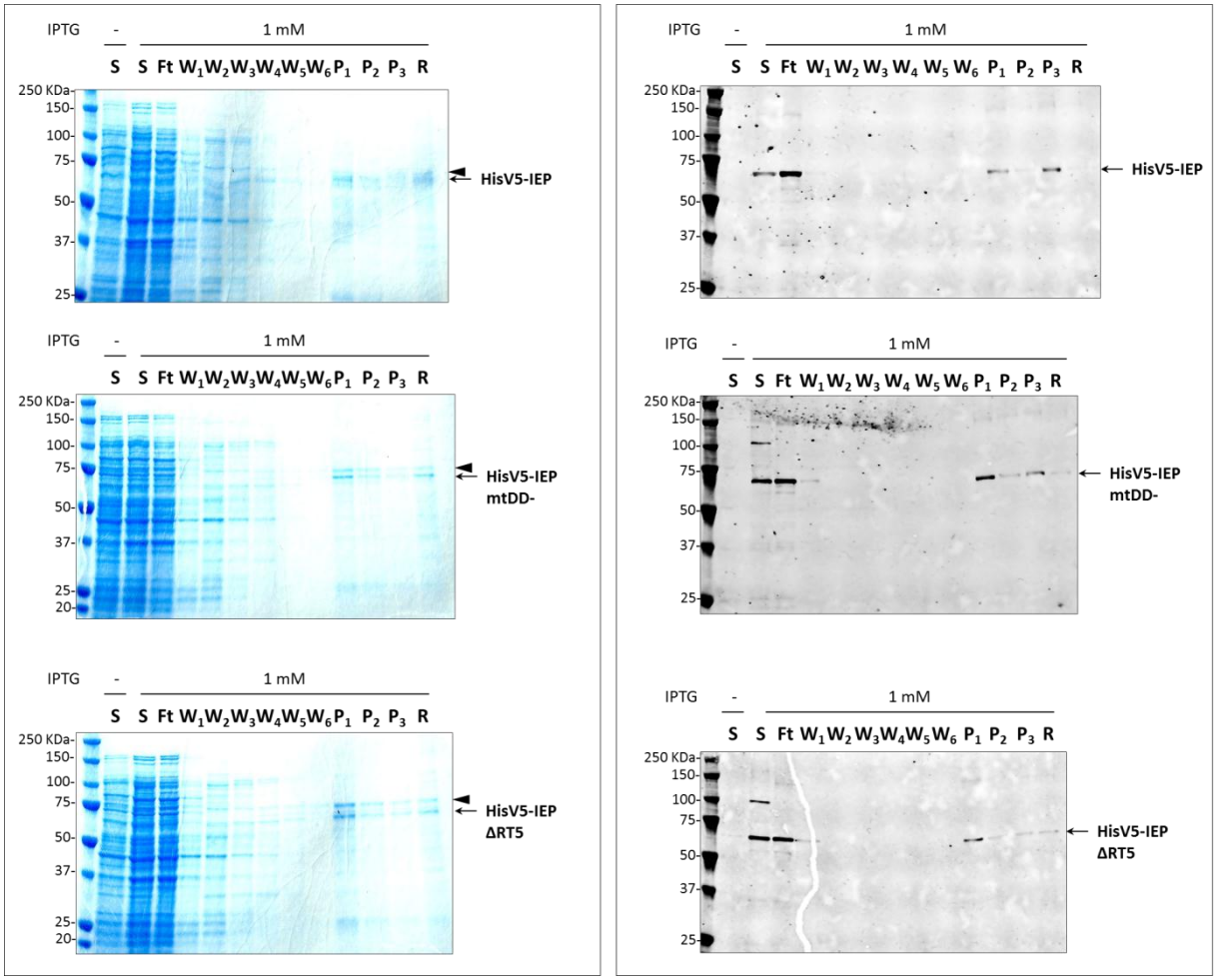


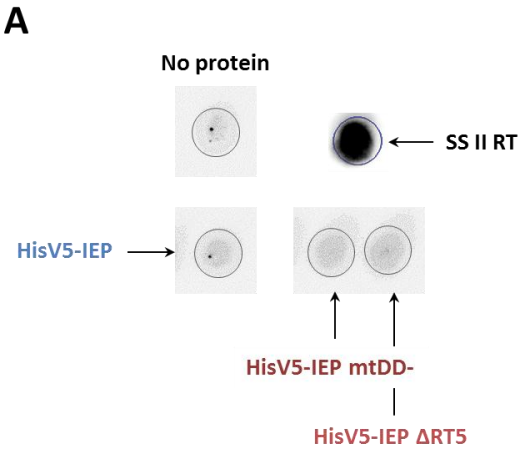
Figure R-26: Purification under native conditions of HisV5-IEP and mutants, expressed in Rosetta-gami B (DE3).

HisV5-IEP and mutants (mtDD- and ΔRT5) expression was induced from 100 ml *E. coli* culture at OD_{600nm} 0.8 with 1 mM of IPTG for 4 hrs at 30°C. A negative control was also performed from 5 ml of *E. coli* culture without IPTG induction. *Left panel*: Coomassie blue stained SDS-PAGE gel of protein fractions collected during

the purification. S: soluble protein fraction (1/25 of the S(-) fraction and 1/450 of the S(1 mM) fraction); Ft: Flow-through from the Ni^{2+}-charged resin (1/450 of the fraction); W_1 to W_6: wash protein fractions 1 to 6 (1/2500 of the fraction); P_1 to P_3: purified protein fractions successively eluted from the resin (1/30 of the fraction); R: proteins which remain bound to the resin after elution (1/15 of the fraction). Black arrowhead indicates *E. coli* co-purified protein contaminant. *Right panel*: Western blot analysis using a mouse anti-V5 antibody. Numbers at *left* indicate molecular mass marker.

Coomassie blue stained SDS-PAGE shows that wild-type HisV5-IEP and mutants (mtDD- and ΔRT5), expressed in Rosetta-gami B (DE3) *E. coli* strain, are successfully purified under native conditions (Fig. R-26; left panel, fractions P). The 75-kDa contaminant *E. coli* protein is also co-purified with HisV5-tagged proteins (Fig. R-26; left panel, indicated by black arrowhead). Almost all HisV5-tagged proteins are eluted during the first elution step (Fig. R-26; left panel, fractions P_1). The western blot analysis demonstrates the presence of HisV5-IEP and mutants in purified fractions (Fig. R-26; right panel, fractions P). The identity of the two major proteins in those purified protein fractions was also determined by mass spectrometry analysis (See Results section 3.3.1 -). Notably, His-tagged proteins are not subjected to degradation under those conditions, as shown by western blot (Fig. R-26; right panel).

These purified protein fractions were then used to assay the reverse transcriptase activity *in vitro* of HisV5-IEP and its derivative mutant proteins (Fig. R-27).



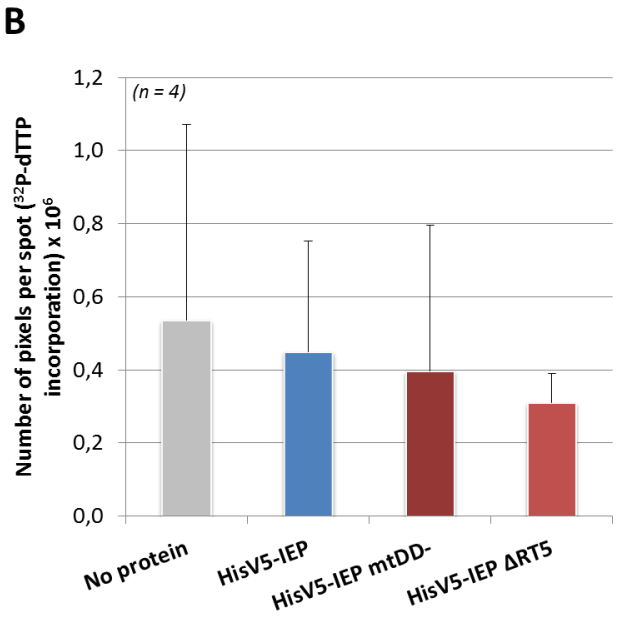


Figure R-27: RT assay with HisV5-IEP and mutants purified under native conditions.

(***A***) 8 µl of purified protein fractions were used. RT reactions without proteins (No protein) and with HisV5-IEP, HisV5-IEP mtDD- and HisV5-IEP ΔRT5 mutants were performed at 37°C for 45 min. Positive control consists of 0.03 U of SuperScript® II reverse transcriptase (SS II RT). (***B***) Data, representing the number of pixels per spot, were quantified with ImageQuant™ software. Data are the mean of four different experiments and standard deviation is represented by thin lines. Light gray bar: No protein; blue bar: HisV5-IEP; dark red bar: HisV5-IEP mtDD-; Light red bar: HisV5-IEP ΔRT5.

Figure R-27A shows the resulting membrane image of one experiment. Quantification of data for a total of four independent experiments was performed (Fig. R-27B). We observe that, even when purified under native conditions, the HisV5-IEP does not display any RT activity. Indeed, no statistically significant difference could be determined between all conditions (Fig. R-27B). This suggests that either the protein folding and/or stability are altered during the IMAC purification process or that the HisV5 tag impedes the proper folding of the IEP. It also indicates that the contaminant *E. coli* protein, which was co-purified with GST-tagged proteins and responsible for the RT activity in previous RT assays, is obviously not present in purified HisV5-tagged protein fractions. The purification process is very similar for His-tagged proteins and GST-tagged proteins. The conformation and activity of this contaminant *E. coli* protein should be not be altered by the IMAC purification. Thus, the lack of RT activity in IMAC purification fractions likely reflects the absence of this contaminant and in addition the absence and/or instability of the IEP active conformation.

2.5 - CONCLUSION

This chapter outlined the attempts to express and purify previously uncharacterized IEP from *Pylaiella littoralis* Pl.LSU/2 group II intron in order to assay its potential RT activity. The Pl.LSU/2 intron indeed contains an open-reading frame theoretically encoding a putative protein which presents reverse transcriptase, maturase and endonuclease domains shared by other group II intron-encoded proteins.

We choose to express the IEP as a fusion protein tagged with either GST or HisV5 tags to purify the IEP by affinity chromatography.

The GST soluble protein, used as a tag for further purification, was first selected for the possibility of enhancing the overall solubility of the tagged IEP. GST-tagged IEP was expressed in BL21 Star (DE3) pRARE *E. coli* strain, after identifying problematic codons rarely used in *E. coli*. Protein expression studies revealed an overexpressed protein, running at the expected size of 91-kDa and mainly insoluble. Attempts were made to reduce the level of insoluble protein, including varying the temperature of induction, the concentration of IPTG and also use the ArcticExpress (DE3)RIL strain to express the protein. However, only little improvements were seen in the ratio of soluble protein to insoluble. Although the protein was at approximately 20% solubility, the overall expression yield of the protein was high, so it was deemed unnecessary to pursue any further work to enhance solubility as enough protein was soluble and further purified using Glutathione-Sepharose resin.

Purified protein fractions were shown to be partially contaminated by *E. coli* proteins. Nonetheless, reverse transcriptase activity was still tested, as GST-tagged proteins (IEP and mutants) were always predominant in those fractions. *In vitro* RT assays revealed an RT activity using fractions containing GST-IEP but also with mutants GST-IEP mtDD- and GST-IEP ΔRT5 fractions. In contrast, no RT activity was found with the GST negative control. Mutants GST-IEP mtDD- and ΔRT5 were expected to be RT-deficient. Therefore, the activity showed in those assays could not be attributed to GST-tagged proteins. We speculated that a contaminating *E. coli* reverse transcriptase protein, co-eluted along with the IEP, was responsible for this RT activity.

To overcome this bias, we decided to evaluate other expression systems and host.

We evaluated the expression of HisV5-IEP and His-IEP in the cell-free expression system and the baculovirus/Sf9 expression system, respectively. But these expression systems did not allow either a sufficient expression rate or solubility of the fusion protein and optimizations are limited in these systems, so that we did not pursue these experiments.

Finally, the expression in *E. coli* and purification of HisV5-tagged IEP was tested. The HisV5-tag was chosen because of its small length so that the structure of the IEP should not be affected. The structure of the fusion HisV5-IEP can potentially be different than those of GST-IEP and this could possibly circumvent the problem of contamination with the *E. coli* protein co-eluted along with GST-IEP that showed an RT activity.

The protein was first expressed in BL21 Star (DE3) pRARE and purified by IMAC under native conditions but results showed that HisV5-IEP could not bind to the Ni^{2+}-charged resin. An internal position of the tag into the tertiary structure of the fusion protein could be the cause of this inconvenient. To overcome this problem, HisV5-IEP was purified under denaturing conditions. High yield of pure HisV5-IEP was obtained with those conditions. A refolding step consisting in a multi-step dialysis process was performed in order to slowly remove the denaturant. Those conditions should theoretically avoid any protein precipitation, so that a correct refolding of proteins could be achieved. RT assays using refolded protein fractions were performed and showed no RT activity of wild-type HisV5-IEP. It suggests that proteins were not allowed to recover their correct catalytic conformation during the refolding process.

We thus decided to use less stringent purification conditions while promoting the solubility of HisV5-IEP. HisV5-IEP was purified in non-denaturing conditions in presence of CHAPS. Another *E. coli* strain was also used in the following experiments. Rosetta-gami B (DE3) indeed contains mutations that enhance the disulfide bond formation. HisV5-IEP was predicted *in silico* to contain three disulfide bonds. The formation of these probable disulfide bonds could enhance the correct

tertiary conformation and/or stability of the protein. Using this strain could thus promote the correct folding of HisV5-IEP. High yield of nearly pure HisV5-IEP was obtained with those non-denaturing purification conditions. The protein was subjected to little degradation and a 75-kDa contaminant *E. coli* protein was also co-purified. Unfortunately, no RT activity of the wild-type HisV5-IEP could be detected. It suggests that the use of CHAPS detergent has impeded the correct HisV5-IEP conformation or has unstabilized the protein conformation.

Therefore, HisV5-IEP expressed in Rosetta-gami B (DE3) was purified under native conditions. High yield of nearly pure HisV5-IEP was obtained, although the 75-kDa *E. coli* contaminant was also co-eluted. Unfortunately, those HisV5-IEP purified protein fractions showed no reverse transcriptase activity.

The conducted experiments did not allow the characterization of the IEP RT activity. In GST-IEP RT assays, an *E. coli* contaminant protein seemed to bias the reactions, and HisV5-IEP RT activity could not be demonstrated. We decided to co-express HisV5-IEP with the Pl.LSU/2 intron RNA in *E. coli* in order to assay the RT activity of HisV5-IEP contained in RNP particles. RNPs were purified by two methods: IMAC and sucrose centrifugation. All the results obtained are briefly summarized in the next section and detailed in the submitted article 2.

3 - ARTICLE 2

3.1 - SUMMARY OF THE WORK

This article presents the results obtained on the characterization of Pl.LSU/2 IEP biochemical activities and on the *in vivo* splicing activity of the Pl.LSU/2 intron in yeast. We showed that Pl.LSU/2 IEP presents an RT activity *in vitro* when expressed in presence or absence of the Pl.LSU/2 intron RNA in *E. coli*. We also demonstrated the maturase activity of Pl.LSU/2 IEP *in vivo* in *S. cerevisiae*, promoting the splicing of Pl.LSU/2 intron. Nevertheless, we failed to detect the splicing of Pl.LSU/2 in a human cell line.

In their natural environments, the transcription of group II introns is concomitant with the expression of the IEP that they carried. It has been shown that for the *Lactococcus lactis* Ll.LtrB group II intron-encoded protein, the co-expression of the intron RNA is required for the proper folding and stability of the IEP in its active conformation (Matsuura M et al. 1997). The attempts to purify biochemically active IEP protein, tagged in N-terminal with either the GST protein or the HisV5 tag, and purify by IMAC were not conclusive, as described in the previous chapter. Thus, the purification of RNP particles containing the IEP and the spliced intron RNA could be required to allow the folding of the Pl.LSU/2 IEP in its active form. RNP particles can be purified by IMAC if the containing-IEP is fused to the His-tag. Another method consists in separate this ribonucleoprotein complex from host soluble proteins by a sucrose centrifugation process, which allows concentration of morphologically intact particles (Kennell JC et al. 1993; Matsuura M et al. 1997).

. In this context, we used a plasmid which allows the co-expression of HisV5-IEP and Pl.LSU/2 intron in *E. coli*. Soluble protein extracts were then used to purify RNP particles potentially formed in *E. coli*. We tested both RNPs purification methods: IMAC (See Results section 3.3.2 -) and sucrose cushion centrifugation (See article 2). We showed that RNP particles containing HisV5-IEP were expressed and purified by both methods. The RT assay using both RNP particles preparations showed that HisV5-IEP contained in RNPs has an RT activity when RNPs are purified by sucrose cushion centrifugation (See article 2). In those experiments, mutations in the RT catalytic YADD motif of the IEP abolish its RT activity. In contrast, no RT activity of HisV5-IEP contained in RNPs purified by IMAC was found (See Results section 3.3.2 -). It suggests that the IMAC purification process alters the correct folding and/or stability of the IEP. This finding potentially explains the results obtained with HisV5-IEP purified by IMAC, showing an absence of activity. Altogether, these results led us to assay the RT activity of HisV5-IEP alone purified by sucrose cushion centrifugation (See article 2). We showed that HisV5-IEP can be purified by sucrose cushion centrifugation with a high yield of purity, and that, even in absence of co-expression of Pl.LSU/2 intron RNA, HisV5-IEP purified by sucrose cushion centrifugation has an RT activity *in vitro*. This indicates that the protein is able to achieve its correct folding by itself and does not necessarily require the intron RNA to be stabilized, in contrast to the Ll.LtrB IEP (Matsuura M et al. 1997).

The determination of the Pl.LSU/2 IEP functionality *in vitro* marked another step of the Pl.LSU/2 intron characterization. To further evaluate the Pl.LSU/2 intron catalytic activity, an experimental strategy was designed to evaluate its splicing *in vivo* in *S. cerevisiae*. We constructed a splicing reporter plasmid in which the intron, flanked by its two exons, is placed just downstream of the URA3

gene lacking a transcription start codon. This construction should permit the transcription of a precursor mRNA from which the Pl.LSU/2 can splice. This way, the URA3 gene should be translated only upon precise Pl.LSU/2 intron splicing, leading to the production of the fusion E2-E3-Ura3p protein. This system can theoretically allow the rapid determination of the Pl.LSU/2 intron splicing using the URA3 selection. In addition, one of the aims of this study was to determine the maturase activity of the Pl.LSU/2 IEP. To do so, the IEP coding sequence of the intron, located in its domain IV, was here partially deleted, and the IEP was conditionally expressed from a separate plasmid. The ura3- *S. cerevisiae* strain harboring the splicing reporter plasmid was then transformed or not with the inducible IEP-expressing plasmid. The cells were then plated onto a minimum media lacking uracil; the growth of colonies on this medium indicating a Pl.LSU/2 intron splicing. In all experiments conducted, we repeatedly failed to detect any colonies on this selective medium. However, the analysis of RNAs showed the presence of spliced mRNA in all conditions and demonstrated that the Pl.LSU/2 splicing efficiency was significantly increased upon IEP expression in yeast cells. Nevertheless, in spite of the detection of the spliced mRNA, the E2-E3-Ura3p protein expression was not detected by western blot, explaining the absence of clones on –ura medium. The Pl.LSU/2 intron was thus shown to splice in yeast cells, and this splicing is promoted by the maturase activity of Pl.LSU/2 IEP. However, the system that we used is not convenient for the intron splicing detection, due to the failure of (sufficient) Ura3p expression.

We finally evaluated the ability of the Pl.LSU/2 intron to splice in a human cell line. Four stable HCT 116 human cell lines expressing different forms of the intron, whose domain IV was more or less deleted, were established by transducing HCT 166 with intron-expressing lentiviral vectors (LVs). LVs encoding the IEP, used or not in N-terminal with the GFP were also used. We demonstrated that the IEP was expressed in HCT 116 cell line and verified its nuclear localization. The four stable cell lines were then transduced or not with IEP-LVs and GFP-IEP-LVs. The analysis of RNAs extracted from the cells showed that the four forms of the intron used were transcribed, but we unfortunately failed to detect any trace of spliced mRNA in these human cell lines, even upon IEP expression. The Pl.LSU/2 intron was thus shown to be unable to (efficiently) splice in human cells.

To conclude, we have characterized biochemical activities of the Pl.LSU/2 IEP and demonstrated the Pl.LSU/2 intron splicing in *S. cerevisiae*. The Pl.LSU/2 IEP was shown to have an RT activity both alone or when complexed in RNP particle. The splicing of Pl.LSU/2 intron was demonstrated in yeast and its efficiency was shown to increase upon IEP expression, indicating a maturase activity of the IEP *in vivo*. However, no Pl.LSU/2 splicing could be evidenced in human cells. Thus, further optimizations are required for the use of group II introns in human gene targeting.

3.2 - ARTICLE 2

The brown algae Pl.LSU/2 group II intron-encoded protein has functional reverse transcriptase and maturase activities.

Madeleine Zerbato, Nathalie Holic, Sophie Moniot-Frin, Dina Ingrao, Anne Galy, and Javier Perea*

Inserm, U951, Evry F91002; University of Evry Val d'Essonne, UMR S_951, Evry, F91002; Genethon, Evry, F91002, France.

* To whom correspondence should be addressed. Tel: +33 169472833; Fax: +33 169472838; Email: perea@genethon.fr

Present Address: Pr. Javier Perea, Genethon, 1 bis rue de l'Internationale, BP60, 91002 Evry Cedex, France.

ABSTRACT

Group II introns are self-splicing mobile elements found in prokaryotes and eukaryotic organelles. These introns propagate by homing into precise genomic locations, following assembly of a ribonucleoprotein complex containing the intron-encoded protein (IEP) and the spliced intron RNA. Engineered group II introns are now commonly used tools for targeted genomic modifications in prokaryotes but not in eukaryotes. We speculate that the catalytic activation of currently known group II introns is limited in eukaryotic cells. The brown algae *Pylaiella littoralis* Pl.LSU/2 group II intron is uniquely capable of *in vitro* ribozyme activity at physiological level of magnesium but this intron remains poorly characterized. We purified and characterized recombinant Pl.LSU/2 IEP. Unlike other IEPs, Pl.LSU/2 IEP displayed high level of reverse transcriptase activity without intronic RNA. The Pl.LSU/2 intron could be engineered to splice accurately in *Saccharomyces cerevisiae* and splicing efficiency was increased by the maturase activity of the IEP. However, spliced transcripts were not expressed. Furthermore, intron splicing was not detected in human cells. While further tool development is needed, these data provide the first functional characterization of the PI.LSU/2 IEP and the first evidence that the Pl.LSU/2 group II intron splicing occurs *in vivo* in eukaryotes in an IEP-dependent manner.

INTRODUCTION

Prokaryotic and eukaryotic organelle introns are mobile elements able to integrate specifically in the exon junction of an intronless genome [1-3]. This property, called "homing", contributes to the spreading of introns and has been used for precise *in vivo* genome engineering [4-13]. Two general homing mechanisms have been described depending on the type of intron: group I introns encode a very specific nuclease (meganuclease) to produce a double-strand break (DSB) in the intronless genome at the junction of exons. The DSB is then repaired by homologous recombination (HR) with a template DNA coming from the "invader" genome [14]. Group II introns are ribozymes that self-splice from precursor RNA yielding excised intron lariat RNAs. The lariat splicing intermediate recognizes the DNA junction between the exons of the intronless genome and integrates the genome forming a DNA-RNA hybrid. After reverse transcription, the template intronic RNA is degraded and the gap is repaired by a DNA polymerase. In spite of very divergent mechanisms used for homing, both group I or group II introns rely on the expression of a protein coded by the intron itself (IEP, Intron-Encoded Protein) for the homing process [2]. These IEPs often carry different activities: maturase (to help the proper splicing of the intron) [15-22], double strand endonuclease (in group I introns) [17,18,20,22], single strand endonuclease and reverse transcriptase (in group II introns) [23-28]. Homing always results in the incorporation of a copy of the intron into the intronless genome.

The ability of group I and group II introns to recognize and to integrate into a specific genomic site has been exploited to generate various knock-out or knock-in models in mammalian cells [4,12,13,29,30], plants [31-33], and bacteria [5,34-38]. Specific genomic targeting is obtained by changing the native target recognition sequences using rational engineering or directed molecular evolution [13,39]. At present, group II intron-derived genomic targeting strategies are only used in bacteria. However, using group II introns in mammalian cells could represent some advantages over the currently existing technologies. In mammalian cells, meganucleases and strategies based on FokI restriction endonuclease coupled to engineered Zinc fingers [32] or Transcription Activator-Like Effectors [40,41] are used for specific genomic insertion. Both of these approaches utilize DSB repair processes which occur mainly by non-homologous end joining (NHEJ) or by HR in the presence of a DNA template [42-44]. These approaches are limited by low efficiency and safety concerns. The radically different homing mechanism of group II introns could be an alternative to DSB mediated gene engineering.

There are only a few examples of the use of group II introns in eukaryotes. An initial proof of concept in mammalian cells has been reported by Guo *et al.* [4] using the *Lactococcus lactis* Ll.LtrB intron to target two genes carried into plasmids in HEK 293 human cells. These initial experiments showed homing into the plasmid as detected by PCR with specific oligonucleotides but efficacy was not measured. The splicing of the Ll.LtrB intron inside the HEK 293 cells was also not demonstrated since the authors introduced directly the purified LtrA-lariat ribonucleoparticles into the cell to obtain homing. However, it is known that the Ll.LtrB intron can acquire its correct tertiary structure in bacteria with the help of LtrA as used in commercial gene knock-out systems [5,45-47].

Several elements may contribute to the limitations in the use of group II introns in eukaryotes. Ribozyme activity of group II introns depends on the correct folding of RNA into a specific tertiary structure [48]. This activity is necessary for both the intron splicing and its insertion into target DNA.

Most of group II introns are able to self-splice *in vitro* in the absence of proteins but have to be "chaperoned" by proteins *in vivo* [48,49]. These proteins may not function optimally in eukaryotic cells. Self-splicing of group II introns depends on the correct tertiary structure folding of the intronic RNA, and Mg^{2+} cations contribute to this folding by stabilizing the RNA tertiary structure [48,50]. Mg^{2+} is also required for catalysis of group II introns [51] and its concentration is critical for proper self-splicing *in vitro*. It is therefore possible that most group II intron cannot function optimally in eukaryotic cells where the free Mg^{2+} concentration is estimated to be around 1-2 mM [52]. For example, the optimal *in vitro* reaction conditions for the well-known bacterial *Lactococcus lactis* Ll.LtrB intron IEP-assisted RNA splicing use 5 mM Mg^{2+} and 10 mM of Mg^{2+} are required for an optimal reverse splicing reaction of RNPs into DNA target site [53]. The importance of Mg^{2+} concentration is emphasized by a recent study showing that Ll.LtrB RNP particles were able to insert efficiently into a plasmid target in eukaryotic nuclei by providing $MgCl_2$ during the microinjection of RNPs in *Xenopus laevis* oocytes, and *Drosophilia melanogaster* and zebrafish (*Danio renio*) embryos [29]. In this work, we have chosen to study the Pl.LSU/2 intron from the mitochondrial large subunit rRNA gene of the brown algae *Pylaiella littoralis* because of its ability to self-splice *in vitro* at unusually low Mg^{2+} concentrations (0.1 mM) [54]. We reasoned that this characteristic could make the Pl.LSU/2 intron a good candidate to be used as a tool for genome manipulation in eukaryotic cells. The Pl.LSU/2 intron structure presents a highly canonical secondary structure model consisting in six helical domains (I to VI) radiating from a central wheel [55,56]. Although the Pl.LSU/2 intron tertiary structure has been characterized [57-59], there is no report on the homing property of this intron nor on its potential biochemical activities *in vivo*. Moreover, the putative intron-encoded protein of PI.LSU/2 has not been described. The fact that the Pl.LSU/2 intron sequence presents in its domain IV an open reading frame containing putative domains of most group II intron IEPs [56] suggests that at least one of the Pl.LSU/2 IEP putative activities has been preserved by evolution.

We have therefore characterized the biochemical activities of the Pl.LSU/2 IEP as well as the Pl.LSU/2 intron activity *in vivo* in eukaryotic cells in order to evaluate the possibility of using this intron in genome engineering. Here, we show that active recombinant Pl.LSU/2 IEP can be produced and purified in *Escherichia coli*. The purified IEP shows a reverse transcriptase activity *in vitro* both alone or when expressed together with the intronic RNA. The Pl.LSU/2 intron is able to splice properly in yeast cells helped by the maturase activity of its IEP.

MATERIALS AND METHODS

Strains and human cell lines

Pl.LSU/2 intron and Pl.LSU/2 intron-encoded protein were expressed in *E. coli* strain Rosetta-gami B(DE3) F^- *ompT hsdS*$_B$ (r_B^- m_B^-) *gal dcm lacY1 ahpC* (DE3) *gor522*::Tn*10 trxB* pRARE (Cam^R, Kan^R, Tet^R) (EMD4Biosciences, Novagen). This strain was grown in LB (Luria-Bertoni) medium with 50 µg/ml of carbenicilin and/or 25 µg/ml of chloramphenicol. *In vivo* splicing of Pl.LSU/2 intron was evaluated in *S. cerevisiae* BY4742 *MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0* (S288C) [60]. This strain was grown in YPD (Yeast extract Peptone Dextrose) and/or in SD (Synthetic Dextose) medium containing dropout supplement mix (according to Molecular Cloning, A Laboratory Manual, by Sambrook J. & Russell D.). HEK 293T and HCT 116 human cells (obtained from ATCC (CRL-11268 and CCL-247 respectively, American Type Culture Collection, Manassas, VA, USA) were grown in DMEM supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin (PS) (Life technologies, Invitrogen). Cells were passaged with TrypLE express 1X (Life technologies, Invitrogen).

Plasmids and oligonucleotides

Plasmids and oligonucleotides used in this study are listed in Supplemental information (Supplemental Tables S1 A and S1 B, respectively).

Structure analysis

Pl.LSU/2 intron secondary structure has already been described [55,56]. The secondary structure of the intron RNA domain IV was predicted with the sFold 2.2 software (http://sfold.wadsworth.org) [61,62] (Supplemental Figure S1).

Expression of the Pl.LSU/2 intron-encoded protein (IEP) in *E. coli*

E. coli strain Rosetta-gami B (DE3) was transformed with the appropriate expression plasmid and single colonies were inoculated into 4 ml of LB containing appropriate antibiotics. Precultures were shaken at 170 rpm at 32°C overnight, inoculated into 100 ml of LB medium without antibiotics and grown at 32°C for 3-6 hr, until OD_{600nm} reached 0.4-0.8. Induction was started by addition of IPTG (1 mM final), and the incubation was continued for 3 hr at 30°C. Cells were then collected by centrifugation (1,900*g* for 10 min at 4°C), and washed once with 20 ml of 150 mM NaCl. The washed cell pellet was stored at -80°C overnight.

Purification of the Pl.LSU/2 IEP and RNP particles by sucrose cushion centrifugation

The IEP, tagged in N-terminus with an histidine stretch (6xHis) and a V5 epitope (GKPIPNPLLGLDST) was purified by sucrose cushion centrifugation, as described [23]. The washed cell pellet was resuspended in 4 ml of ice-cold buffer A (50 mM Tris-HCl at pH 7.4, 1 mM EDTA, 1 mM DTT, 10% (v/v) glycerol), and lysozyme was added to a final concentration of 4 mg/ml. After 45 min of incubation on ice, cells were lysed by three cycles of freeze/thawing between -70°C and 37°C, followed by addition of 2.5 volumes of HKCTD buffer (25 mM Tris-HCl at pH 7.4, 500 mM KCl, 50 mM $CaCl_2$, 5 mM DTT). Lysate was then centrifuged (14,000*g* for 15 min at 4°C), and supernatant was layered over 5ml of 1.85 M sucrose containing HKCTD and centrifuged in a Beckman Type 70 Ti rotor (50,000*g* for 17h at 4°C). The resulting pellet was gently washed with 1 ml of ice-cold Milli-Q water and then dissolved in 25 µl of ice cold 10 mM Tris-HCl at pH 8.0, and 1 mM DTT. IEP mtDD- and RNP particles containing IEP WT or IEP mtDD- were purified according to the same procedure.

Purified proteins and RNPs preparations were stored at -80°C. The yield of RNP particles was 25-90 OD_{260nm} units per 100 ml of culture, with 1 OD_{260nm} unit of RNP containing 0.84 ± 0.12 µg of IEP. The yield of purified proteins was 55-150 µg per 100 ml of culture. Quantification of proteins was performed by using Bio-Rad DC Protein Assay Kit 1 (BioRad Laboratories) according to the manufacturer's instructions. All protein fractions were analyzed by Coomassie-blue staining of SDS-PAGE and western blotting.

RNA extraction

Total cellular RNA from yeast cells was extracted from 5 ml of mid-log phase cultures in SD minimal medium (minimal SD base or minimal SD base Gal/Raf, Clontech). After centrifugation (3,800*g* for 5 min at 4°C), the cell pellet was washed with 1 ml of Milli-Q water and resuspended in 2 ml of buffer Y1 (1 M sorbitol, 0.1 M EDTA) supplemented with 0.1% β-mercaptoethanol and 300 U of lyticase (from *Arthrobacter luteus*, >2,000 units/mg protein, Sigma). After 30 min of incubation at 30°C on a rotary shaker, total cellular RNA was extracted from the resulting spheroplasts using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. Total cellular RNA from human cells was isolated with RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. In both cases, a DNase I treatment (RNase-free DNase set, Qiagen) was performed on-column during RNA purification according to the manufacturer's instructions.

Protein extraction

To obtain yeast total proteins, 20 ml of the culture were centrifuged (3,800*g* for 5 min at 4°C). The cell pellet was washed with 5 ml of Milli-Q water, resuspended in CelLytic Y reagent (2.5 ml/g cell pellet, Sigma-Aldrich) supplemented with 7 mM DTT and incubated at 25°C for 20 min on a rotary shaker. The resulting lysate was centrifuged at 18,000*g* for 10 min at 4°C. Proteins were then precipitated with 2 volumes of acetone for 1 hr at 4°C and recovered by centrifugation at 18,000*g* for 15 min at 4°C. Protein pellet was then washed with ethanol and resuspended in 0.5 volume of 50 mM Tris-HCl at pH 8.5, 8 M Urea, and 10 mM DTT.

Yeast nuclear proteins were extracted from 200 ml culture after centrifugation (4,000*g* for 5 min at 4°C). Cell pellets were washed first with 25 ml of Milli-Q water and then with 3 ml of spheroplasting buffer (1 M Sorbitol, 50 mM K_2HPO_4 at pH 6.5, 0.018% β-mercaptoethanol). Cells were resuspended in 3 ml of spheroplasting buffer containing 500 U of lyticase and incubated 30 min at 30°C on a rotary shaker. The spheroplasts were collected by centrifugation (4,500*g* for 5 min at 4°C), washed with 3 ml of ice-cold spheroplasting buffer, resuspended in 8 ml of ice-cold buffer L (18% Ficoll 400, 20 mM K_2HPO_4 at pH 6.8, 1 mM $MgCl_2$, 0.5 mM EDTA, 2 mM PMSF, 1 µg/ml aprotinin) and lysed on ice by 20 strokes of a dounce homogeneizer. The resulting lysate was then centrifuged (3,500*g* for 10 min at 4°C). The resulting supernatant was centrifuged in a Beckman SW 55Ti rotor (58,000*g* for 35 min at 4°C) and pelleted nuclei were resuspended in 200 µl of ice-cold buffer NP (0.34 M sucrose, 20 mM Tris-HCl at pH 7.4, 50 mM KCl, 5 mM $MgCl_2$, 2 mM PMSF, 1 mg/ml aprotinin). Nuclear protein extracts were stored at -80°C.

Human total protein extracts were prepared following washing of cells and lysis of the pellets in buffer containing 50 mM Tris-HCl pH7.5, 200 mM NaCl, 1 mM EDTA, 1 mM PMSF, 1% Triton X-100, 0.1% SDS, 0.5% sodium deoxycholate, and 10% glycerol supplemented with protease inhibitors cocktail.

Protein concentrations were determined with the Bio-Rad DC Protein Assay kit 1.

Western blot analyses

Proteins were resolved on a denaturing 10% polyacrylamide-SDS gel (Criterion XT Bis-Tris gels, Biorad), transferred onto a nitrocellulose membrane (Hybond ECL, Amersham) and probed with the appropriate antibodies.

Immunoblots of the 6xHis/V5-tagged Pl.LSU/2 IEP expressed in *E. coli* were probed with mouse anti-V5 antibody (1:5,000 dilution, Life technologies, Invitrogen) followed by IRDye 680-conjugated goat anti-mouse antibody (1:8,000 dilution, LI-COR Biosciences) and immunoreactive bands were detected with the Odyssey infrared scanner (Li-Cor).

Yeast total proteins (30 µg/lane) were immunoblotted with mouse anti-HA antibody (1:200 dilution, Santa Cruz Biotechnologies) and rat anti-Tub1p antibody (1:5,000 dilution, Abcam) to detect respectively the expressed IEP and Tubulin 1 protein (Tub1p), then with Horseradish peroxidase (HRP)-conjugated goat anti-mouse antibody (1:25,000 dilution, Jackson ImmunoResearch) and HRP-conjugated rabbit anti-rat antibody (1:1,000 dilution, Dako). Bands were revealed by chemiluminescence (Supersignal West Dura Extended Duration Substrate, Thermoscientific).

Yeast nuclear proteins (80 µg/lane) and human total proteins (30 µg/lane) were probed with an HRP-conjugated mouse anti-c-Myc antibody (1:3,000 dilution, Life technologies, Invitrogen) to detect the myc-tagged NLS-IEP. For yeast nuclear proteins, a mouse anti-TATA binding protein (TBP) antibody (final concentration of 2 µg/ml, Abcam) was added and HRP-conjugated goat anti-mouse antibody (1:20,000 dilution) was used as secondary antibody. Bands were revealed by chemiluminescence.

RT assays

RT activity was assayed for 45 min at 37°C in 14 µl of reaction medium containing 10 mM KCl, 10 mM $MgCl_2$, 50 mM Tris-HCl at pH 8.0, 5 mM DTT, 0.05% NP40, 1 µg of poly(rA)-oligo$(dT)_{12-18}$ (Amersham), 10 µCi of [α-^{32}P]dTTP (3,000 Ci/mmole, Perkin-Elmer) and 1 µg of RNase A (Sigma Aldrich), as described [23]. Reactions were started by the addition of either IEP or RNPs preparations and stopped by addition of EDTA (50 mM final). Radioactive products were spotted on a DE81 filter (Whatman) which was washed twice in 2X SSC. After an overnight exposure on a phosphor screen (Molecular Dynamics PhosphorImager System; GE Healthcare Bio-Sciences), radioactive spots were detected by the Storm system (GE-Healthcare Bio-Sciences) and data were analyzed with ImageQuant software (GE Healthcare Life Sciences).

Reverse transcription

Reverse transcription of yeast or human cellular RNA was carried out in 20 µl of reaction medium using the Verso cDNA Kit (Thermo Scientific). One microgram of total RNA was incubated for 5 min at 70°C and then mixed with 1X reverse transcription buffer, 0.5 mM of each dNTPs, 2 µM of RM-R oligonucleotide (yeast RNA; Supplemental Table S1 B) or 300 ng of random hexamers and 125 ng of anchored oligo dT (human RNA), 1 µl of RT Enhancer enzyme, and 1 µl of Verso reverse transcriptase. After incubation at 42°C for 1 hr, reaction was stopped by heating 2 min at 95°C. A control reaction without Verso reverse transcriptase was performed to ensure the absence of DNA contamination in the RNA samples (minus-Verso RT samples).

PCR

Two µl of yeast cDNA were used as a template for PCR in 50 µl of reaction medium containing 1 unit of Phusion high-fidelity DNA polymerase (Thermoscientific), 62.5 µM of each dNTPs, PCR buffer, and 0.2 µM of each p1 and p2 or p3 and p4 primers. The amplification consisted of 30 cycles at 98°C

for 30 sec, 70°C for 30 sec and 72°C for 30 sec. PCR products were analyzed on 12% polyacrylamide gels.
Two µl of human cDNA were used as a template for PCR in 40 µl of reaction medium containing 0.8 unit of Phusion high-fidelity DNA polymerase, 200 µM of each dNTPs, PCR buffer, and 0.25 µM of each p5 and p6 primers. The amplification consisted of 30 cycles at 98°C for 10 sec, 58°C for 20 sec and 72°C for 1 min 30 sec. PCR products were analyzed on 1.2% agarose gels.

Quantitative PCR
The qPCR consisted of a SYBR Green-based detection of cDNA with specific primers hybridizing E2 and the E2-E3 junction or primers hybridizing E2 and the E2-Intron junction. Amplification reactions (25 µl) contained 5 µl of a 1/10 dilution of cDNA and 12.5 µl of qPCR buffer (Power SYBR Green PCR master mix, Applied Biosystems), 0.3 µM of each primers and consisted of 40 cycles at 95°C for 15 sec then at 65°C for 1 min on a 7900HT Real-Time PCR System (Applied Biosystems). To ensure the absence of non-specific amplifications, a dissociation step was added consisting in a final cycle at 95°C (15 sec) then 65°C (15 sec) and finally 95°C (15 sec). To quantify cDNA copy number obtained respectively from spliced or precursor mRNA, standard amplification curves were made by serial dilutions of appropriate linearized plasmids. All PCR measures were performed at least in duplicate. All qPCR experiments include samples from mock vector-transformed cells and minus-Verso RT samples as negative controls.
Data were edited using the Sequence Detection Systems 2.3 software (Applied Biosystems) and interpreted in the linear portion of the standard curve. The following test acceptability criteria were used: linear regression coefficient of the standard curve >0.98; less than 0.5 CT variation for duplicate samples; CT comprised between 35 and 40 for H_2O; mock vector-transformed cells and minus-Verso RT samples.

Generation and titration of lentiviral vectors
Recombinant lentiviral vectors (LVs) using the pRRL backbone plasmid [63] (Supplemental Table S1 A) were constructed to express the indicated transgenes under the control of the human phosphoglycerate kinase (PGK) promoter. VSV-G-pseudotyped particles were produced by quadritransfection of HEK 293T cells (Supplemental Table S1 A) as previously described [63,64]. Harvested virus particles were concentrated by ultracentrifugation and titered in infectious genome particles (IG) as previously described [64].

Lentiviral vector transduction of human cell lines
HCT 116 cells were seeded at 10^5 cells per well in 12-well plates 16-24 hrs prior transduction. Cells were transduced with LVs using concentrations ranging from 10^6 to 10^8 IG/ml in the presence of 6 µg/ml of polybrene. Transduction medium was replaced with fresh medium 6 hrs after transduction. A second hit of transduction was performed the following morning in the same conditions using LVs encoding proteins. Cells were removed and analyzed 48 hrs later. For LVs stably expressing different forms of Pl.LSU/2 intron, cells were expanded for about 10 days. Vector copy numbers per cell were calculated by quantitative PCR on cell lines genomic DNA as previously described [64] (0.9 for intron-ΔDIV; 0.3 for intron-DIVa; 4.7 for intron-DIVab, and 1.2 for full length intron).

SUPPORTING INFORMATION

Supplemental Figure S1 and Supplemental informations (Supplemental Table S1 and Supplemental references).

RESULTS

Pl.LSU/2 IEP contained in RNP particles presents an RT activity *in vitro*

The *Pylaiella littoralis* Pl.LSU/2 group II intron is located in the mitochondrial gene encoding the large ribosomal RNA (Fig. 1A; LSU rRNA gene). This intron contains in its domain IV an open-reading frame presenting the predicted conserved domains of group II intron-encoded proteins which are the reverse transcriptase (RT), DNA-binding domain (D), maturase (X) and endonuclease (En) [15,23,25,28,53] (Fig. 1A). Subsequently, recombinant Pl.LSU/2 intron-encoded protein (IEP) was expressed and purified in order to measure its predicted biochemical activities.

Expression of IEP in RNP particles. Group II intron-encoded proteins are known to be fully active when intron RNA is coexpressed with the IEP, as a result of stabilization of the protein in its active conformation [23]. To express the Pl.LSU/2 IEP in RNP particles, we designed the plasmid p151-E+I+IEP for the induction of both the Pl.LSU/2 intron and IEP expression in *E. coli* (Fig. 1A). This plasmid contains the Pl.LSU/2 intron and its flanking exons (50 last nucleotides of exon 2 and 71 first nucleotides of exon 3) [54] cloned downstream of the phage T7 promoter in the expression vector pET151/D-TOPO (Supplemental Table S1 A). The IEP ORF is fused to a 6xHis and a V5 epitope tags in its N-terminus used to the detection of the protein by western blot. This vector could potentially allow the expression of both the intron RNA and the wild-type IEP ORF (IEP WT) with its own Shine-Dalgarno sequence for translation. We also constructed a derivative negative control plasmid (p151-E+I+IEPmtDD-) in which the catalytic YADD motif of the IEP RT domain is mutated to YAAA (Fig. 1A, position indicated by a diamond; Supplemental Table S1 A) [65]. This plasmid is expected to express a mutant RT-defective IEP (IEP mtDD-).

The expression plasmids p151-E+I+IEP and p151-E+I+IEPmtDD- were transformed in *E. coli* Rosetta-gami B (DE3). After induction, soluble protein fractions were used to purify RNP particles by sucrose cushion centrifugation [23]. The SDS-PAGE analysis shows a major band at the expected size of IEP WT and mtDD- (Fig. 2A; Coomassie, 69 kDa, black arrow). Other bands are also detected by Coomassie staining and correspond to *E. coli* contaminant proteins and/or IEP (WT and mtDD-) degradation products. The western blot analysis confirms the presence of IEP WT and mtDD- at the expected size (Fig. 2A; WB) and shows the presence of some degradation products. A non-specific band is also detected by western blot at 150-kDa in the RNPs mtDD- preparation (Fig. 2A; WB). RNP particles containing either IEP WT or IEP mtDD- can thus be purified using sucrose cushion centrifugation.

RT activity. The open reading frame of the intron Pl.LSU/2 contains a conserved RT domain. The RT activity of Pl.LSU/2 IEP (WT and mtDD-) in RNP particle preparations was assayed with the artificial template-primer substrate poly(rA)-oligo$(dT)_{12\text{-}18}$. We show that RNP particles from cells expressing p151-E+I+IEP and purified by sucrose cushion centrifugation have an RT activity (Fig. 2B; RNP WT). Quantitatively, the RT activity of 0.1 OD_{260nm} units of RNP particles is similar to that of 0.03 units of the commercial SuperScript® II reverse transcriptase (SSII RT; Life technologies, Invitrogen). The time course of RT reactions shows that the RT activity of WT IEP-containing RNPs increases over incubation time (Fig. 2C) and is positively correlated with the amount of RNP particles (Fig. 2D). As expected, the RT activity of the Pl.LSU/2 IEP is abolished by point mutations in the conserved YADD motif (Fig. 2B, 2C and 2D; RNP mtDD- conditions) and is similar to that of the background

condition (Fig. 2C; No protein). The 6xHis tag was also used to purify RNP particles by immobilized metal-ion affinity chromatography (IMAC) on a Ni^{2+}-charged column but no RT activity was ever detected in those RNP particle preparations (data not shown), suggesting that this purification process destabilizes the complex. These results indicate that the Pl.LSU/2 IEP in RNPs particles can be expressed in *E. coli* and purified by sucrose centrifugation thereby preserving its RT activity and thus its active conformation.

Pl.LSU/2 IEP is active *in vitro* without the help of the intron RNA

Isolated Pl.LSU/2 IEP has in vitro *RT activity.* To analyze the intrinsic property of the Pl.LSU/2 IEP, we expressed the IEP in *E. coli* without coexpressing the intron RNA. We used the plasmid p151-IEP (Fig. 1A) which places the IEP ORF fused to the 6xHis and V5 epitope tags immediately downstream of the phage T7 promoter and Shine-Dalgarno sequence of the vector. This plasmid version was also constructed with a mutation of the YADD motif (Fig. 1A; indicated by a diamond). Both WT and mutant proteins were expressed in *E. coli* and then purified by sucrose cushion centrifugation. SDS-PAGE shows that the WT and the RT-defective IEPs (mtDD-) are successfully purified from *E. coli* lysates (Fig. 3A; Coomassie). No other proteins are detected by Coomassie staining. The identity of the purified proteins is verified by western blot (Fig. 3A; WB) and by mass spectrometry (data not shown).

RT assays with Pl.LSU/2 IEP purified by sucrose cushion centrifugation show that the protein displays an intrinsic *in vitro* RT activity (Fig. 3B; IEP WT). This activity is abolished by mutations in the RT domain as expected (Fig. 3B; IEP mtDD-). The positive control SuperScript® II Reverse transcriptase (0.03 U) shows a number of pixels per spot of $18.1 \times 10^6 \pm 5.78$ (not shown). It is noteworthy that no RT activity is detected following IMAC purification of IEP (WT and mtDD-, data not shown), which is consistent with previous findings obtained with RNPs. These data suggest that purification in sucrose cushion preserves the activity of the IEP and thereby its correct folding. It also demonstrates that recombinant IEP has intrinsic RT activity and does not necessarily require the help of intron RNA to become active.

Splicing assay of Pl.LSU/2 intron in *S. cerevisiae*

In vitro self-splicing of Pl.LSU/2 has been demonstrated [54] using a deleted form of the intron in the domain IV (See Fig. 1B; intron-ΔDIV). However, most group II introns need the maturase activity of their IEP to achieve *in vivo* splicing. In *Lactococcus lactis* Ll.LtrB group II intron, the LtrA intron-encoded protein interacts with a substructure located in the beginning of the domain IV, and other contacts are made with the conserved core regions of the intron [27,66]. In order to evaluate the influence of the IEP on the splicing of the Pl.LSU/2 intron in yeast, we included a part of the domain IV in our construction. This domain corresponds to DIVab indicated in the secondary structure of the domain IV (1870 nucleotides) predicted by the S-fold software (See Fig. 1B; Supplemental Fig. S1). This DIVab domain retains putatively most of the predicted secondary structure of domain IV (S-fold prediction, not shown).

To study the expression and splicing of the Pl.LSU/2 intron in yeast, we developed a *URA3*-based intron-splicing reporter expressed from a 2-µ plasmid (Fig. 4A; pEgpIIE-URA3; Supplemental Table S1 A). The Pl.LSU/2 (DIVab) intron flanked by its natural exons (50 last nucleotides of exon 2 and 71 first nucleotides of exon 3) [54] was fused to the coding sequence of URA3, which encodes a orotidine 5-phosphate decarboxylase (Ura3p). The *URA3* gene is read in-frame only upon precise Pl.LSU/2

splicing, which should lead to Ura3p expression and growth of an ura3- (*ura3Δ0)* strain of *S. cerevisiae* on minimal media lacking uracil (Fig. 4A). Expression of the Ura3p can be detected with an HA tag added upstream of the exon 2, generating the fusion HA-E2-E3-Ura3p. A control plasmid (pEE-URA3) was also used in this assay to ensure that the hybrid HA-E2-E3-URA3 protein is able to complement the ura3- mutation.

Previous studies have shown that optimal splicing of the bacterial Ll.LtrB group II intron in bacteria and yeast requires the maturase activity of its encoded protein, LtrA, by promoting the folding of the intron RNA into its catalytically active structure. The open reading frame of Pl.LSU/2 intron contains a conserved X domain similar to that of LtrA. Therefore, to determine if Pl.LSU/2 IEP has maturase activity facilitating the splicing of its intron, we first attempted to demonstrate intron splicing through functional restoration of yeast growth in the *URA3*-based intron-splicing reporter assay. In this system, the IEP protein was conditionally expressed using an inducible GAL10 promoter plasmid (Fig. 4B; pNLS-IEP^{co}) and intron sequences were delivered *in trans*. For these experiments, we used an IEP sequence that was codon-optimized for translation in human cells (IEP^{co}) tagged in C-terminus with c-Myc epitopes, and containing nuclear localization signals (NLS) of the SV-40 T-antigen (Fig. 4B) to address the protein to the nucleus. The inducible/repressible expression of the NLS-IEP^{co} was confirmed respectively in galactose (Gal) and glucose (Glc) media and nuclear localization of the protein was verified by western blot analysis on yeast nuclear protein extracts (data not shown).

Yeasts were thus transformed with either the control pEE-URA3 plasmid or the pEgpIIE-URA3 plasmid. Subsequently, the NLS-IEP^{co} expressing plasmid was transformed or not in yeast carrying pEgpIIE-URA3 and the number of colonies on medium containing or not uracil was determined for each condition (Fig. 4C). Functional restoration of growth could not be demonstrated in the *URA3*-based intron splicing reporter assay. Yeasts expressing the EgpIIE-URA3 cassette repeatedly fail to grow on the –ura selective medium, even when EgpIIE-URA3 is coexpressed with NLS-IEP^{co} (Fig. 4C). However, yeasts grow on minimal medium lacking uracil when the cells are transformed with the pEE-URA3 plasmid encoding the hybrid E2-E3-URA3 protein proving the orotidine 5-phosphate decarboxylase activity of the hybrid protein, as it confers the ability to complement the ura- mutation (Fig. 4C). Thus, the splicing of the Pl.LSU/2 intron could not be demonstrated through this assay into yeast cells.

Pl.LSU/2 group II intron can splice in yeast in an IEP-dependent manner

Since the assay reads-out both RNA splicing and translation of the spliced mRNA, further investigation was conducted to determine if RNAs were transcribed, spliced and translated. To determine if the Pl.LSU/2 intron was spliced from the precursor mRNA, an RT-PCR analysis was performed on RNA extracted from yeast cells harboring either pEE-URA3 or pEgpIIE-URA3 expressed in the presence or absence of NLS-IEP^{co}. Two different pairs of primers were used to specifically amplify cDNA obtained from precursor or spliced mRNA (Fig. 5A; Supplemental Table S1 B). We first show that the pEgpIIE-URA3 splicing reporter allows the expression of precursor mRNA in all conditions (Fig. 5B; precursor cDNA, pEgpIIE-URA3). As expected, the control plasmid pEE-URA3 allows the expression only of a mRNA corresponding to the spliced mRNA (Fig. 5B; pEE-URA3). In absence of NLS-IEP^{co}, a poorly efficient splicing of Pl.LSU/2 can be detected from yeast harbouring the pEgpIIE-URA3 reporter (Fig. 5B; spliced cDNA, pEgpIIE-URA3). It is worth noting that splicing efficiency is significantly increased when NLS-IEP^{co} is expressed into yeast cells (Fig. 5B; spliced cDNA, Glucose- / Galactose+). Accurate Pl.LSU/2 splicing was verified by

sequencing across the splice junction of the RT-PCR spliced products (data not shown). These results were quantified using an RT-qPCR analysis performed on four independent experiments (Fig. 5C). Precursor and spliced cDNA copy numbers were calculated and ratios of spliced/precursor cDNA were determined to measure splicing efficiency. We confirm that in absence of NLS-IEP^{co} expression (Glc + / Gal -), the splicing of Pl.LSU/2 is poorly efficient while the Pl.LSU/2 intron splicing efficiency in yeast cells is significantly increased upon NLS-IEP^{co} expression (Fig. 5C; Glc- / Gal+; between 2.1 and 7.9 fold change compared to background; $p < 0,032$, *t*-test using a unilateral pair-wise comparison). These results indicate first that the Pl.LSU/2 intron can be expressed in yeast and also that the IEP presents a maturase activity *in vivo*.

Spliced Pl.LSU/2 group II transcripts are not translated

The results of the phenotypic analysis showing an absence of clones on medium lacking uracil (Fig. 4C) and the level of splicing observed in yeast harboring the pEgpIIE-URA3 reporter and expressing the NLS-IEP^{co} (Fig. 5C) suggested a blockade in expression of the spliced messenger. Indeed, we had not been able to detect by western blot the Ura3p hybrid protein resulting theoretically from the Pl.LSU/2 spliced mRNA (Fig. 5D; pEgpIIE-URA3, Ura3p), even upon robust NLS-IEP^{co} expression (Fig. 5D). In contrast, high levels of Ura3p hybrid protein are expressed in yeast harboring pEE-URA3 (Fig. 5D). The absence of detectable level of Ura3p in yeast carrying pEgpIIE-URA3 thus explains the failure of functional growth restoration in the *URA3*-based intron splicing reporter assay.

Splicing assay of Pl.LSU/2 intron in human cells

To study if the splicing of the Pl.LSU/2 intron could also occur in human cells, we first aimed to determine if the Pl.LSU/2 intron could be expressed in the colon carcinoma HCT 116 cell line. We tested various forms of the intron in which the domain IV was more or less extensively deleted (See Fig. 1B; ΔDIV, DIVa, DIVab and full length intron). To stably express the various intron sequences in human cells, we transduced HCT 116 cells with lentiviral gene transfer vectors (LVs) expressing the different forms of the PI.LSU/2 intron flanked by the last 50 nucleotides of exon 2 and the first 71 nucleotides of exon 3 (Fig. 6A; upper panel; pRRL-intron +/- DIV). Transduced HCT 116 cells were expanded for at least ten days to establish stable cell lines expressing the different forms of the intron. RT-PCR using primers p5 and p6 hybridizing E2 and E3, respectively (Fig. 6A; upper panel; Supplemental Table S1 B) shows expression of the appropriate precursor transcripts in each stable cell lines (Fig. 6A; lower panel; amplicons of 736 bp for ΔDIV intron, 1020 bp for DIVa intron, 1472 bp for DIVab intron and 2534 bp for full length intron). Nevertheless, we did not find any trace of spliced mRNA in any of the stable cell lines analyzed (data not shown). These results suggest that Pl.LSU/2 intron RNA can be expressed but does not splice in human cells.

To determine if the intron-encoded protein could promote splicing of Pl.LSU/2 intron in human cells, we induced the expression of Pl.LSU/2 IEP^{co} or GFP-IEP^{co} fusion protein in human HCT 116 cells by transduction with protein-expressing LVs (Fig. 6B; upper panel). Forty eight hours following transduction of HCT 116 cells with the LV, western blot analysis demonstrates expression of the IEP^{co} or GFP-IEP^{co} in the cells (Fig. 6B; lower panel). GFP expression was also confirmed by western blot using an anti-GFP antibody and by evidence of nuclear localization following microscopy analysis (data not shown).

The previously described Pl.LSU/2 intron expressing stable lines were then transduced by these LVs. Western blot analysis showed the expression of IEP^{co} and GFP-IEP^{co} in these cells (data not

shown). To determine the Pl.LSU/2 intron splicing capacity in the presence of IEP, RNAs were extracted and analyzed by RT-qPCR. Primers p7 and p4 (Supplemental Table S1 B), hybridizing E2 and E2-E3 junction respectively, were used to detect the spliced mRNA. Primers p1 and p2 (Supplemental Table S1 B), hybridizing E2 and E2-Intron junction respectively, were used to detect the precursor mRNA. Ratios of spliced/precursor cDNA copy number, which determine the splicing efficiency, are less than 2.10^{-4} in every condition tested and for every form of the Pl.LSU/2 intron tested (data not shown). This result shows that, in contrast to yeast, the Pl.LSU/2 intron is unable to splice efficiently in human cells in this context, even in presence of the intron-encoded protein that was codon-optimized for translation in human cells.

DISCUSSION

This study provides the first functional description of the ribozyme activity of the Pl.LSU/2 group II intron *in vivo*. A catalytically-active recombinant PI.LSU/2 intron-encoded protein can be produced in *E. coli* and purified. This protein presents a reverse transcriptase activity *in vitro* both alone or when coexpressed with the intronic RNA. It also displays a maturase activity which facilitates the splicing of the Pl.LSU/2 intron *in vivo* in yeast in a IEP dose-dependent manner.

These results contribute to characterize the poorly studied Pl.LSU/2 intron. Prior to our study, no information was yet available on the *in vivo* properties of this intron, neither in its natural environment nor in experimental systems. The description of the RT and maturase activities of the IEP confirms some of the predicted properties encoded by the domain IV of the intron which was known to contain an open reading frame theoretically encoding a protein presenting the four characteristic domains (RT, X/maturase, D and En) of other group II intron-encoded proteins [55,56].

One limitation of the use of group II introns in eukaryotes could be the requirement for a correct folding of both the IEP and the intron RNA, therefore it is important to assess the cooperation between the intron and IEP. Here, we show for the first time that the Pl.LSU/2 IEP displays a RT activity when complexed as RNP. In addition, unlike other group II introns IEP, the PI.LSU/2 IEP has an intrinsic RT activity *in vitro* in the absence of the intron RNA. In a previous study of the Lambowitz group [23], the *Lactococcus lactis* Ll.LtrB IEP (LtrA) correct folding was shown to be facilitated by the unspliced precursor or intron RNA: LtrA protein was less active and had unstable RT activity when the exon 1 and the 5' end of the intron were missing from the expression plasmid. The intron RNA appeared to have a chaperone-like activity promoting the proper folding of the LtrA protein. In the case of Pl.LSU/2, the IEP reverse transcriptase activity could be clearly demonstrated even in the absence of co-expression with the intronic RNA. Similarly, Vellore et al. expressed in *E. coli* the G.st.I1 intron-encoded protein (called *trt*) from *Geobacillus stearothermophilus* fused to a 6xHis tag without co-expressing the intron RNA [67]. The authors showed an RT activity using partially purified protein fraction. However, it is noteworthy that the authors did not used the negative control consisting of the *trt* IEP mutated in the YADD catalytic motif. The data presented herein clearly demonstrate the PI.LSU2 IEP self-activation properties which may present an advantage for its use in heterologous systems.

The particular biochemical properties of this IEP may also facilitate its use *in vivo*. The PI.LSU/2 intron can be engineered to splice *in vivo* in yeast and this is facilitated in a dose-dependent manner by the Pl.LSU/2 IEP. Although the Pl.LSU/2 IEP encoding gene used here is a codon-optimized sequence for translation in human cells, IEP translation in yeast is enough efficient to improve splicing of Pl.LSU/2. Presumably the maturase activity of the PI.LSU/2 IEP promotes the folding of the intron RNA into its catalytic tertiary structure *in vivo.* Interestingly, a residual splicing could be detected by RT-PCR and RT-qPCR using RNA extracts from yeasts that do not express the IEP. This IEP-independent splicing could have occur *in vivo* (which would be consistent with the ability of this intron to splice *in vitro* even under not optimal Mg^{2+} conditions [54]) or *in vitro* during the RT-PCR reactions precisely because this ribozyme is highly active *in vitro*.

In spite of evidence of functional properties of the PI.LSU/2 intron, we failed to develop a productive system for genomic targeting at this stage. While we demonstrated the presence of a spliced mRNA in yeast cells, there was no translation of the spliced mRNA since restoration of the URA3 ORF did not lead to the expected growth on a minimal medium without uracil. It is possible

that the yield of spliced mRNA in yeast cells is not sufficient to the expression of a detectable amount of Ura3p proteins by western blot and that the yield of expressed Ura3p does not reach the required threshold for yeast growth on medium lacking uracil. However, the possibility of a translation blockade is reminiscent of results already reported by Chalamcharla et al. with the Ll.LtrB intron using different reporter genes [68]. The authors demonstrated that Ll.LtrB intron splicing occured predominantly in the cytoplasm of yeast cells and that precursor mRNA was subjected to nonsense-mediated mRNA decay (NMD). They also showed that the spliced mRNA was subjected to an NMD-independent translation blockade. However, the mechanism involved remains unknown. The authors speculated that the pairing mechanism between the intron lariat and the spliced mRNA via EBS (exon binding sites)/IBS (intron binding sites) interactions [6] could impede the spliced mRNA translation [68]. One would also postulate that the translation blockade could result from RNA sequestration to cytoplasmic microdomains such as P-bodies or stress granules that play important role in mRNA processing, including repression of translation and mRNA decay [69-71]. Anyway, the fact that this translation blockade also occurs through the use of Pl.LSU/2 group II intron supports the Chalamcharla et al. hypothesis that the mechanism involved is transposable to other group II introns and may has impeded their spread in eukaryotic nuclear genes.

We then tested the ability of the intron to splice in human cells. Here, we used different forms of the Pl.LSU/2 intron with deletions of various sizes of the domain IV in order to determine if one of these parts are required for intron high-affinity binding to the IEP. The domain IV of the Ll.LtrB intron is known to bind its IEP [72], but is not needed for *in vitro* splicing, as in absence of it, a residual splicing occurs [27,66]. In the same way, the domain IV of the yeast *coxI*-I2 intron is required for stable binding of its IEP and additional contacts with the catalytic core of the intron promote the splicing [73]. The domain IV of group II introns appears to guide the interactions or anchor the IEP to catalytic core regions of the intron. In spite of the use of various domain IV coding regions, in spite of using a humanized codon-optimized IEP, in spite of evidence of expression of the IEP and intron in human cells and in spite of the expected nuclear localization of the recombinant IEP in human cells, we failed to detect any splicing of Pl.LSU/2 intron in human cells. Misfolding of IEP and/or intron RNA could explain this splicing defect, but alternatively, defective nuclear/cytoplasmic compartimentalization or inadapted environnement could also be implicated.

The adaptation of group II introns homing mechanism could be considered in the context of a gene repair strategy approach in gene therapy. In current gene therapy assays, the transgene integration with retroviral vectors is not site-specific. Expression of therapeutic transgenes is sometimes unregulated due to *cis*-acting elements present in the neighbouring of the insertion site. Insertions near oncogenes also present a risk of activation by promoters or enhancers carried by the vector [74-79]. All of these issues could be circumvented by targeting the insertion into the original site (gene repairing). The repaired wild type copy will be under the control of original cis-regulating sequences. Strategies based on genome double strand breaks (DSB) in order to achieved reparation by homologous recombination are today thouroughly investigated. The use of group II intron could be an alternative approache avoiving the putative genotoxicity due to off-target DSB.

We present encouraging data that suggest that PI.LSU/2 group II intron could have advantageaous qualities for engineering genomic targeting strategies. This intron and its IEP function *in vivo* in yeast. However, the use of Pl.LSU/2 and other group II introns in human genomic engineering will require further optimizations.

ACKNOWLEDGMENTS

We are very grateful for technical help from Damien Mansard, Cécile K. Lopez, Alexandrine Garnier, Khalil Seye, Stéphanie Matrat and Marie-Noëlle Monier in our group at Genethon. We also thank François Michel and Maria Costa (Centre de Génétique Moléculaire, CNRS, Gif-sur-Yvette) for providing plasmids containing the Pl.LSU/2 intron and for helpful discussions, Guillaume Sirantoine for sequencing, Stéphanie Bucher and the molecular biology development group for titrating vectors batches and Thierry Larmonier and Lucia Braga-Vacherie from the Genethon Biological Resources Center for help with human cell line banking. We also are very grateful to Fedor Svinartchouk and the Biomarkers group for their significant input with mass spectrometry analyses of purified proteins.

FUNDING

This work was supported by the Association Française contre les myopathies (AFM); the Institut national de la santé et de la recherche médicale (INSERM); the University of Evry Val d'Essonne; and by the 7th European Commission Framework Program (FP7) "PERSIST" [agreement 222878].

REFERENCES

1. Saldanha R, Mohr G, Belfort M, Lambowitz AM (1993) Group I and group II introns. FASEB J 7: 15-24.

2. Lambowitz AM, Zimmerly S (2004) Mobile group II introns. Annu Rev Genet 38: 1-35.

3. Toor N, Keating KS, Pyle AM (2009) Structural insights into RNA splicing. Curr Opin Struct Biol 19: 260-266.

4. Guo H, Karberg M, Long M, Jones JP, 3rd, Sullenger B, et al. (2000) Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. Science 289: 452-457.

5. Karberg M, Guo H, Zhong J, Coon R, Perutka J, et al. (2001) Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. Nat Biotechnol 19: 1162-1167.

6. Mohr G, Smith D, Belfort M, Lambowitz AM (2000) Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. Genes Dev 14: 559-573.

7. Garcia-Rodriguez FM, Barrientos-Duran A, Diaz-Prado V, Fernandez-Lopez M, Toro N (2011) Use of RmInt1, a group IIB intron lacking the intron-encoded protein endonuclease domain, in gene targeting. Appl Environ Microbiol 77: 854-861.

8. Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, et al. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature 435: 646-651.

9. Alwin S, Gere MB, Guhl E, Effertz K, Barbas CF, 3rd, et al. (2005) Custom zinc-finger nucleases for use in human cells. Mol Ther 12: 610-617.

10. Mussolino C, Morbitzer R, Lutge F, Dannemann N, Lahaye T, et al. (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. Nucleic Acids Res 39: 9283-9293.

11. Hockemeyer D, Wang H, Kiani S, Lai CS, Gao Q, et al. (2011) Genetic engineering of human pluripotent cells using TALE nucleases. Nat Biotechnol 29: 731-734.

12. Grizot S, Smith J, Daboussi F, Prieto J, Redondo P, et al. (2009) Efficient targeting of a SCID gene by an engineered single-chain homing endonuclease. Nucleic Acids Res 37: 5405-5419.

13. Arnould S, Perez C, Cabaniols JP, Smith J, Gouble A, et al. (2007) Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. J Mol Biol 371: 49-65.

14. Haugen P, Simon DM, Bhattacharya D (2005) The natural history of group I introns. Trends Genet 21: 111-119.

15. Mohr G, Perlman PS, Lambowitz AM (1993) Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucleic Acids Res 21: 4991-4997.

16. Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM (2004) A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to promote RNA splicing. J Mol Biol 340: 211-231.

17. Delahodde A, Goguel V, Becam AM, Creusot F, Perea J, et al. (1989) Site-specific DNA endonuclease and RNA maturase activities of two homologous intron-encoded proteins from yeast mitochondria. Cell 56: 431-441.

18. Wenzlau JM, Saldanha RJ, Butow RA, Perlman PS (1989) A latent intron-encoded maturase is also an endonuclease needed for intron mobility. Cell 56: 421-430.

19. Schafer B, Wilde B, Massardo DR, Manna F, Del Giudice L, et al. (1994) A mitochondrial group-I intron in fission yeast encodes a maturase and is mobile in crosses. Curr Genet 25: 336-341.

20. Ho Y, Kim SJ, Waring RB (1997) A protein encoded by a group I intron in Aspergillus nidulans directly assists RNA splicing and is a DNA endonuclease. Proc Natl Acad Sci U S A 94: 8994-8999.

21. Szczepanek T, Jamoussi K, Lazowska J (2000) Critical base substitutions that affect the splicing and/or homing activities of the group I intron bi2 of yeast mitochondria. Mol Gen Genet 264: 137-144.

22. Belfort M (2003) Two for the price of one: a bifunctional intron-encoded DNA endonuclease-RNA maturase. Genes Dev 17: 2860-2863.

23. Matsuura M, Saldanha R, Ma H, Wank H, Yang J, et al. (1997) A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical

demonstration of maturase activity and insertion of new genetic information within the intron. Genes Dev 11: 2910-2924.

24. Guo H, Zimmerly S, Perlman PS, Lambowitz AM (1997) Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. EMBO J 16: 6835-6848.

25. Kennell JC, Moran JV, Perlman PS, Butow RA, Lambowitz AM (1993) Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. Cell 73: 133-146.

26. Singh NN, Lambowitz AM (2001) Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. J Mol Biol 309: 361-386.

27. Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM (1999) A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. Mol Cell 4: 239-250.

28. San Filippo J, Lambowitz AM (2002) Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. J Mol Biol 324: 933-951.

29. Mastroianni M, Watanabe K, White TB, Zhuang F, Vernon J, et al. (2008) Group II intron-based gene targeting reactions in eukaryotes. PLoS One 3: e3121.

30. Barzel A, Privman E, Peeri M, Naor A, Shachar E, et al. (2011) Native homing endonucleases can target conserved genes in humans and in animal models. Nucleic Acids Res 39: 6646-6659.

31. Shukla VK, Doyon Y, Miller JC, DeKelver RC, Moehle EA, et al. (2009) Precise genome modification in the crop species Zea mays using zinc-finger nucleases. Nature 459: 437-441.

32. Durai S, Mani M, Kandavelou K, Wu J, Porteus MH, et al. (2005) Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. Nucleic Acids Res 33: 5978-5990.

33. Porteus MH (2009) Plant biotechnology: Zinc fingers on target. Nature 459: 337-338.

34. Yao J, Lambowitz AM (2007) Gene targeting in gram-negative bacteria by use of a mobile group II intron ("Targetron") expressed from a broad-host-range vector. Appl Environ Microbiol 73: 2735-2743.

35. Chen Y, McClane BA, Fisher DJ, Rood JI, Gupta P (2005) Construction of an alpha toxin gene knockout mutant of Clostridium perfringens type A by use of a mobile group II intron. Appl Environ Microbiol 71: 7542-7547.

36. Yao J, Zhong J, Fang Y, Geisinger E, Novick RP, et al. (2006) Use of targetrons to disrupt essential and nonessential genes in Staphylococcus aureus reveals temperature sensitivity of Ll.LtrB group II intron splicing. RNA 12: 1271-1281.

37. Zarschler K, Janesch B, Zayni S, Schaffer C, Messner P (2009) Construction of a gene knockout system for application in Paenibacillus alvei CCM 2051T, exemplified by the S-layer glycan biosynthesis initiation enzyme WsfP. Appl Environ Microbiol 75: 3077-3085.

38. Zhuang F, Karberg M, Perutka J, Lambowitz AM (2009) EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. RNA 15: 432-449.

39. Smith J, Grizot S, Arnould S, Duclert A, Epinat JC, et al. (2006) A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. Nucleic Acids Res 34: e149.

40. Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, et al. (2010) Targeting DNA double-strand breaks with TAL effector nucleases. Genetics 186: 757-761.

41. Li T, Huang S, Jiang WZ, Wright D, Spalding MH, et al. (2011) TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. Nucleic Acids Res 39: 359-372.

42. Bibikova M, Golic M, Golic KG, Carroll D (2002) Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. Genetics 161: 1169-1175.

43. Jeggo PA (1998) DNA breakage and repair. Adv Genet 38: 185-218.

44. van Gent DC, Hoeijmakers JH, Kanaar R (2001) Chromosomal stability and the DNA double-stranded break connection. Nat Rev Genet 2: 196-206.

45. Frazier CL, San Filippo J, Lambowitz AM, Mills DA (2003) Genetic manipulation of Lactococcus lactis by using targeted group II introns: generation of stable insertions without selection. Appl Environ Microbiol 69: 1121-1128.

46. Zhong J, Karberg M, Lambowitz AM (2003) Targeted and random bacterial gene disruption using a group II intron (targetron) vector containing a retrotransposition-activated selectable marker. Nucleic Acids Res 31: 1656-1664.

47. Perutka J, Wang W, Goerlitz D, Lambowitz AM (2004) Use of computer-designed group II introns to disrupt Escherichia coli DExH/D-box protein and DNA helicase genes. J Mol Biol 336: 421-439.

48. Michel F, Ferat JL (1995) Structure and activities of group II introns. Annu Rev Biochem 64: 435-461.

49. Huang HR, Rowe CE, Mohr S, Jiang Y, Lambowitz AM, et al. (2005) The splicing of yeast mitochondrial group I and group II introns requires a DEAD-box protein with RNA chaperone function. Proc Natl Acad Sci U S A 102: 163-168.

50. Qin PZ, Pyle AM (1998) The architectural organization and mechanistic function of group II intron structural elements. Curr Opin Struct Biol 8: 301-308.

51. Toor N, Keating KS, Taylor SD, Pyle AM (2008) Crystal structure of a self-spliced group II intron. Science 320: 77-82.

52. Romani AM, Maguire ME (2002) Hormonal regulation of Mg2+ transport and homeostasis in eukaryotic cells. Biometals 15: 271-283.

53. Saldanha R, Chen B, Wank H, Matsuura M, Edwards J, et al. (1999) RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. Biochemistry 38: 9069-9083.

54. Costa M, Fontaine JM, Loiseaux-de Goer S, Michel F (1997) A group II self-splicing intron from the brown alga Pylaiella littoralis is active at unusually low magnesium concentrations and forms populations of molecules with a uniform conformation. J Mol Biol 274: 353-364.

55. Fontaine JM, Goux D, Kloareg B, Loiseaux-de Goer S (1997) The reverse-transcriptase-like proteins encoded by group II introns in the mitochondrial genome of the brown alga Pylaiella littoralis belong to two different lineages which apparently coevolved with the group II ribosyme lineages. J Mol Evol 44: 33-42.

56. Fontaine JM, Rousvoal S, Leblanc C, Kloareg B, Loiseaux-de Goer S (1995) The mitochondrial LSU rDNA of the brown alga Pylaiella littoralis reveals alpha-proteobacterial features and is split by four group IIB introns with an atypical phylogeny. J Mol Biol 251: 378-389.

57. Costa M, Christian EL, Michel F (1998) Differential chemical probing of a group II self-splicing intron identifies bases involved in tertiary interactions and supports an alternative secondary structure model of domain V. RNA 4: 1055-1068.

58. Costa M, Michel F (1999) Tight binding of the 5' exon to domain I of a group II self-splicing intron requires completion of the intron active site. EMBO J 18: 1025-1037.

59. Costa M, Michel F, Westhof E (2000) A three-dimensional perspective on exon binding by a group II self-splicing intron. EMBO J 19: 5007-5018.

60. Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, et al. (1998) Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. Yeast 14: 115-132.

61. Ding Y, Chan CY, Lawrence CE (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA 11: 1157-1166.

62. Ding Y, Lawrence CE (2003) A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res 31: 7280-7301.

63. Dull T, Zufferey R, Kelly M, Mandel RJ, Nguyen M, et al. (1998) A third-generation lentivirus vector with a conditional packaging system. J Virol 72: 8463-8471.

64. Charrier S, Dupre L, Scaramuzza S, Jeanson-Leh L, Blundell MP, et al. (2007) Lentiviral vectors targeting WASp expression to hematopoietic cells, efficiently transduce and correct cells from WAS patients. Gene Ther 14: 415-428.

65. Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9: 3353-3362.

66. Matsuura M, Noah JW, Lambowitz AM (2001) Mechanism of maturase-promoted group II intron splicing. EMBO J 20: 7259-7270.

67. Vellore J, Moretz SE, Lampson BC (2004) A group II intron-type open reading frame from the thermophile Bacillus (Geobacillus) stearothermophilus encodes a heat-stable reverse transcriptase. Appl Environ Microbiol 70: 7140-7147.

68. Chalamcharla VR, Curcio MJ, Belfort M (2010) Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. Genes Dev 24: 827-836.

69. Olszewska M, Bujarski JJ, Kurpisz M (2012) P-bodies and their functions during mRNA cell cycle: mini-review. Cell Biochem Funct 30: 177-182.

70. Thomas MG, Loschi M, Desbats MA, Boccaccio GL (2011) RNA granules: the good, the bad and the ugly. Cell Signal 23: 324-334.

71. Buchan JR, Parker R (2009) Eukaryotic stress granules: the ins and outs of translation. Mol Cell 36: 932-941.

72. Rambo RP, Doudna JA (2004) Assembly of an active group II intron-maturase complex by protein dimerization. Biochemistry 43: 6486-6497.

73. Huang HR, Chao MY, Armstrong B, Wang Y, Lambowitz AM, et al. (2003) The DIVa maturase binding site in the yeast group II intron aI2 is essential for intron homing but not for in vivo splicing. Mol Cell Biol 23: 8809-8819.

74. Hacein-Bey-Abina S, von Kalle C, Schmidt M, Le Deist F, Wulffraat N, et al. (2003) A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. N Engl J Med 348: 255-256.

75. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, McCormack MP, Wulffraat N, et al. (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science 302: 415-419.

76. Hacein-Bey-Abina S, Garrigue A, Wang GP, Soulier J, Lim A, et al. (2008) Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J Clin Invest 118: 3132-3142.

77. Hacein-Bey-Abina S, Hauer J, Lim A, Picard C, Wang GP, et al. (2010) Efficacy of gene therapy for X-linked severe combined immunodeficiency. N Engl J Med 363: 355-364.

78. Ott MG, Schmidt M, Schwarzwaelder K, Stein S, Siler U, et al. (2006) Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. Nat Med 12: 401-409.

79. Stein S, Ott MG, Schultze-Strasser S, Jauch A, Burwinkel B, et al. (2010) Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. Nat Med 16: 198-204.

FIGURES LEGENDS

Figure 1. Schematic representation of plasmids and Pl.LSU/2 intron domain IV used. (*A*) Gray rectangles: Predicted protein domains shared by other group II intron ORFs; E2 and E3: exons flanking the Pl.LSU/2 group II intron [56]; bold line: Pl.LSU/2 group II intron; broken line: vector sequences; PT7: Specific promoter of the T7 bacteriophage RNA polymerase; TT7: T7 bacteriophage RNA polymerase transcription terminator; 6xHis: histidine tag; V5: V5 epitope; hatched rectangles: 50 last nt of E2 and 71 first nt of E3; diamond: position in which the RT catalytic motif YADD is mutated in YAAA on negative control plasmids (p151-E+I+IEPmtDD- and p151-IEPmtDD-). (*B*) Predicted RNA secondary structure of Pl.LSU/2 intron domain IV (DIV). Start and Stop codons of the IEP ORF are indicated. DIVa: section conserved in the intron-DIVa form (deletion from 244 to 1772 nt of DIV). DIVab: section conserved in the intron-DIVab form (deletion from 369 to 1446 nt of DIV). ΔDIV : section conserved in the intron-ΔDIV form (section highlighted in gray); the section from 8 to 1818 nt of DIV is replaced by the sequence CCTAGGATCT [54]. The detailed putative secondary structure is available in Supplemental Figure S1.

Figure 2. RT activity of Pl.LSU/2 IEP in RNP particles purified from *E. coli*. (*A*) SDS-PAGE analysis of IEP-containing RNPs preparations by Coomassie-blue staining (Coomassie) and western blot (WB). Quantities of RNPs used were 9 OD_{260nm} units for RNPs with IEP WT (RNP WT) and 18 OD_{260nm} units for RNPs with IEP mtDD- (RNP mtDD-). A monoclonal mouse anti-V5 antibody was used to detect the IEP by western blot. The IEP is indicated by the black arrow. Numbers at *left* indicate molecular mass markers in kilodaltons (KDa). (*B*) RT assays with 0.1 OD_{260nm} units of RNPs. Dark gray bar: RNPs containing IEP WT; Light gray bar: RNPs containing IEP mtDD-; Hatched bar: control SuperScript® II reverse transcriptase (SS II RT). Data represent the number of pixels per spot indicating the [α-^{32}P]dTTP incorporation for each reaction. Data are the means of at least three independent experiments with the standard deviation indicated by thin lines. Data were subjected to a *t*-test using a unilateral pair-wise comparison procedure. A highly significant difference is indicated by asterisks ($p < 6 \times 10^{-4}$). (*C*) Time course of RT reaction. RT activity of RNPs containing IEP (RNPs WT, dark gray dots) was assayed with 0.1 OD_{260nm} units of RNPs. Reactions were performed using various incubation times. Experiments were also performed without RNP particles (No RNPs, light gray dots) or with RNP particles containing the mutant Pl.LSU/2 IEP mt DD- (RNPs mtDD-, gray dots). (*D*) Dose-dependent effect on RT activity. RT activity of RNPs containing IEP (RNPs WT, dark gray dots) or mutant IEP (RNPs mtDD-, light gray dots) was assayed with different quantities of RNPs for 45 min.

Figure 3. RT activity of Pl.LSU/2 IEP purified from *E. coli*. (*A*) Analysis of IEP purification by SDS-PAGE with Coomassie-blue staining (Coomassie) and western blot (WB). Volumes loaded contain 5-10 µg of purified protein fraction. A monoclonal mouse anti-V5 antibody was used to detect the IEP by western blot. The IEP is indicated by the black arrow. Numbers at *left* indicate molecular mass markers in kilodaltons (KDa). (*B*) RT assays with 100 ng of IEP. Dark gray bar: IEP WT; Light gray bar: IEP mtDD-. Data are the mean of two independent experiments with the standard deviation indicated by thin lines. Data were subjected to a *t*-test using a unilateral pair-wise comparison procedure. A significant difference is indicated by asterisk ($p < 0.036$).

Figure 4. *In vivo* splicing assay of Pl.LSU/2 intron in *Saccharomyces cerevisiae*. (*A*) Schematic representation of the group II intron splicing reporter assay. PPGK: Phosphoglycerate kinase gene promoter; TPGK: Phosphoglycerate kinase gene transcription terminator; E2 and E3: 50 last nt of exon 2 and 71 first nt of exon 3; bold line: Pl.LSU/2 intron-DIVab; light gray rectangle: URA3 ORF; HA: HA epitope; AUG: Translation start codon. See text for detailed description of the assay. (*B*) Schematic representation of the NLS-IEP^{co} expressing plasmid. PGAL10: UDP-Galactose epimerase gene promoter; TGAL10: UDP-Galactose epimerase gene transcription terminator; dark gray rectangle: human codon-optimized Pl.LSU/2 IEP ORF (IEP^{co}); 3xNLS: nuclear localization signals; Myc: stretch of 3 c-Myc epitopes. (*C*) Numbers of yeast colonies growing on appropriate minimal medium containing (+ura) or not (-ura) uracil. Strain carrying pEE-URA3 was used as a control. Strain carrying pEgpIIE-URA3 splicing reporter was transformed or not with the NLS-IEP^{co} expressing plasmid (pNLS-IEP^{co}). Three resulting transformants were cultured until OD_{600nm} reached 2.5 and applied on the appropriate minimal medium using a 10^{-4} dilution of the culture. Data indicate the numbers of yeast colonies on the plate per ml of the dilution.

Figure 5. IEP-mediated Pl.LSU/2 splicing *in vivo*. (*A*) Schematic representation of amplification products from precursor and spliced cDNA obtained by RT-PCR. Black rectangle: Pl.LSU/2 intron-DIVab; E2 and E3: 50 last nt of exon 2 and 71 first nt of exon 3; gray rectangle: URA3 ORF; black arrow: primer; p1 and p2: forward and reverse primers amplifying cDNA derived from the precursor mRNA (111 bp product); p3 and p4: forward and reverse primers amplifying cDNA derived from the spliced mRNA (77 bp product). (*B*) Acrylamide electrophoresis of RT-PCR products.Total RNA were extracted from yeast carrying pEE-URA3 or pEgpIIE-URA3 and transformed (+) or not (-) with the NLS-IEP^{co} expressing plasmid (pNLS-IEP^{co}). Cells were grown in presence (+) or absence (-) of glucose or galactose (inducer of the NLS-IEP^{co} expression). P: amplification product from precursor cDNA; S: amplification product from spliced cDNA. Numbers at *left* indicate molecular mass marker in base pair (bp). (*C*) Quantification of the Pl.LSU/2 *in vivo* splicing by RT-qPCR. cDNA copy number obtained from spliced and precursor mRNA were calculated using standard amplification curves made by serial dilutions of SphI-linearized pEE-URA3 and pEgpIIE-URA3 plasmids, respectively. Ratios of spliced/precursor cDNA copy number were determined and data were normalized on condition 1 (pNLS-IEPco (-); Glc+ ; Gal-). Four independent experiments are represented (Exp 1 to Exp 4). (*D*) Western blot analysis of Ura3p made from pEE-URA3 or pEgpIIE-URA3. Yeast strains were cultivated in glucose (Glc+ ; Gal-) or in galactose (Glc- ; Gal+) and in absence (-) or presence (+) of pNLS-IEP^{co}. Tubulin 1 protein (Tub1p) expression was also determined, as well as nuclear expression of NLS-IEP^{co} and TATA-Binding protein (TBP).

Figure 6. *In vivo* splicing assay of Pl.LSU/2 intron in human cell lines. (*A*) *Upper panel.* Schematic representation of Pl.LSU/2 intron transfer cassettes (pRRL-Intron +/- DIV) used in this study. Several sizes of the intron domain IV are used (See Fig. 1B; ΔDIV, DIVa, DIVab and full length DIV; Supplemental Table S1 A). HCT 116 cells were transduced by the corresponding VSV-G-pseudotyped lentiviral vectors (LVs) to establish stable cell lines expressing the four different forms of Pl.LSU/2 intron. Light gray rectangle : chimeric 5' LTR (RSV-R-U5); P_{PGK}: phosphoglycerate kinase gene promoter; E2 and E3: 50 last nt of exon 2 and 71 first nt of exon 3; black rectangle : Pl.LSU/2 intron with various size of the DIV (See Fig. 1B); WPRE: Woodchuck hepatitis post-transcriptional regulation element; p5 and p6: forward and reverse primers amplifying cDNA derived from precursor and spliced RNA; dark gray rectangle : 3' LTR (ΔU3-R-U5). *Lower panel.* Agarose electrophoresis of

RT-PCR products. HCT 116 cells were transduced with Pl.LSU/2 intron-expressing LVs (ΔDIV-LV, DIVa-LV, DIVab-LV or Full-LV) and stable cell lines were established. Total RNAs were extracted from the four stable cell lines expressing the different forms of the intron and precursor mRNAs were detected by RT-PCR. (*B*) *Upper panel.* Schematic representation of Pl.LSU/2 IEP gene transfer cassettes used in this study. Black rectangle: codon-optimized sequence of Pl.LSU/2 IEP for translation in human cells, in frame with GFP encoding sequence (GFP-IEP^{co}) or not (IEP^{co}); 3xNLS: nuclear localization signals; Myc: stretch of 3 c-Myc epitopes. *Lower panel.* Western blot analysis. HCT 116 cell line was either untransduced (NT) or transduced with either IEP^{co}-LV or GFP-IEP^{co}-LV. Total proteins were extracted from lysates 48h after transduction and a monoclonal mouse anti-c-myc antibody was used to detect the IEP^{co} and the GFP-IEP^{co}.

Figure S1. Secondary structure of the Pl.LSU/2 intron domain IV predicted by sFold. Pl.LSU/2 intron domain IV (DIV; 1870 nts) predicted RNA secondary structure obtained with the sFold software (http://sfold.wadsworth.org). The domain IV used is from nucleotide 494 to nucleotide 2363 of the Pl.LSU/2 intron.

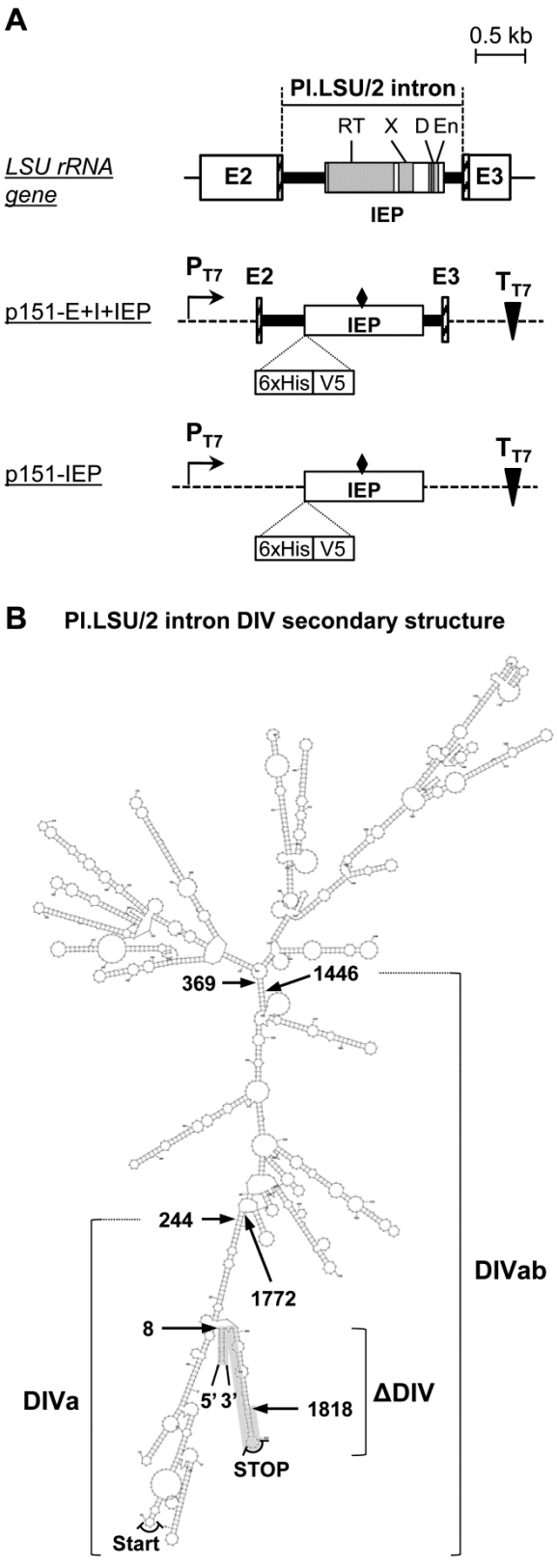

A
0.5 kb
Pl.LSU/2 intron
RT X D En
LSU rRNA gene
E2
E3
IEP
P T7
E2
E3
T T7
p151-E+I+IEP
IEP
6xHis V5
P T7
T T7
p151-IEP
IEP
6xHis V5
B
Pl.LSU/2 intron DIV secondary structure
369
1446
244
1772
8
DIVa
5' 3'
1818
ΔDIV
STOP
Start
DIVab

Figure 1

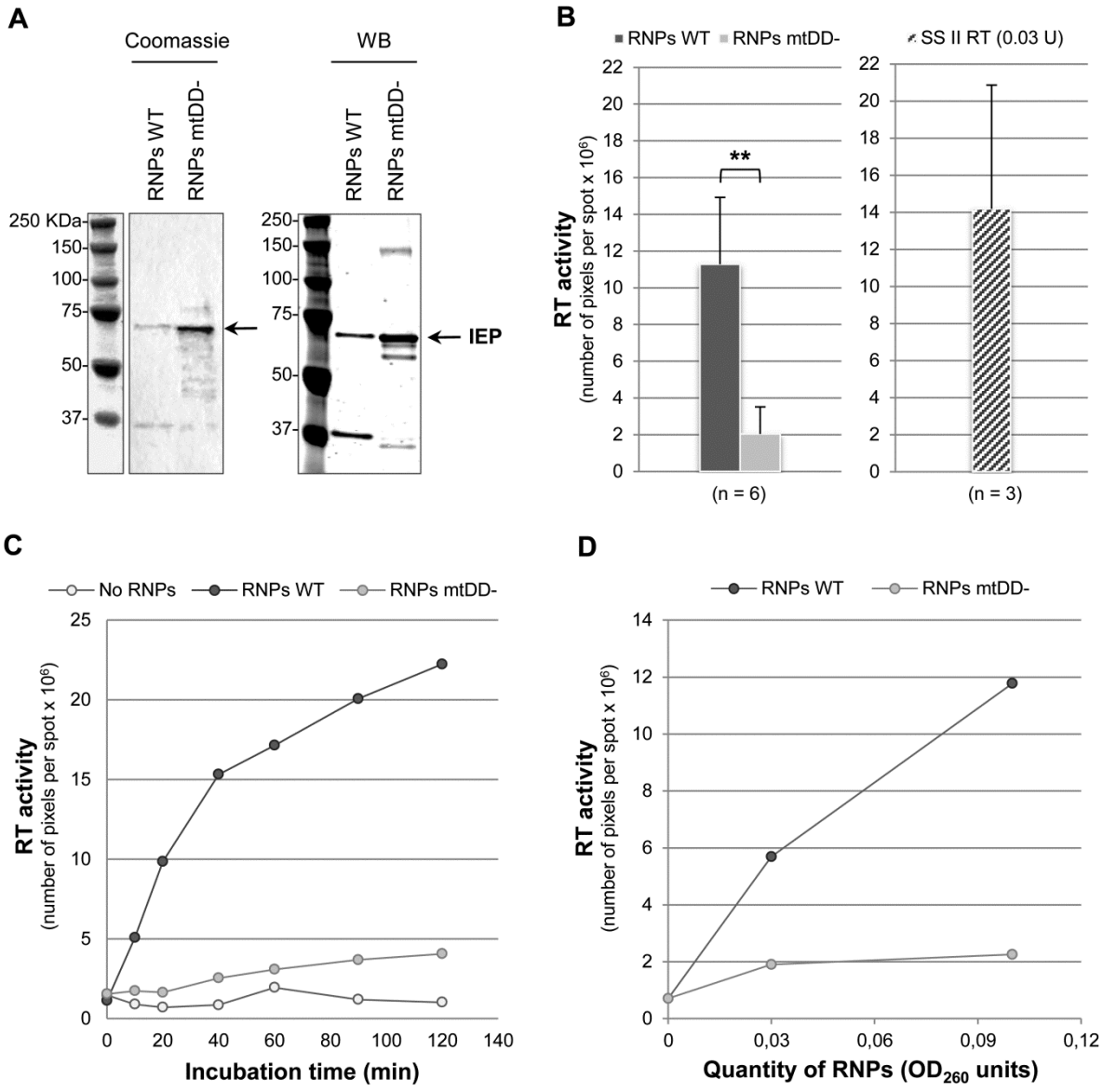

A
Coomassie
WB
RNPs WT
RNPs mtDD-
250 KDa
150
100
75
50
37
IEP
B
RNPs WT
RNPs mtDD-
SS II RT (0.03 U)
RT activity
(number of pixels per spot x 10⁶)
**
(n = 6)
(n = 3)
C
No RNPs
RNPs WT
RNPs mtDD-
RT activity
(number of pixels per spot x 10⁶)
Incubation time (min)
D
RNPs WT
RNPs mtDD-
RT activity
(number of pixels per spot x 10⁶)
Quantity of RNPs (OD260 units)

Figure 2

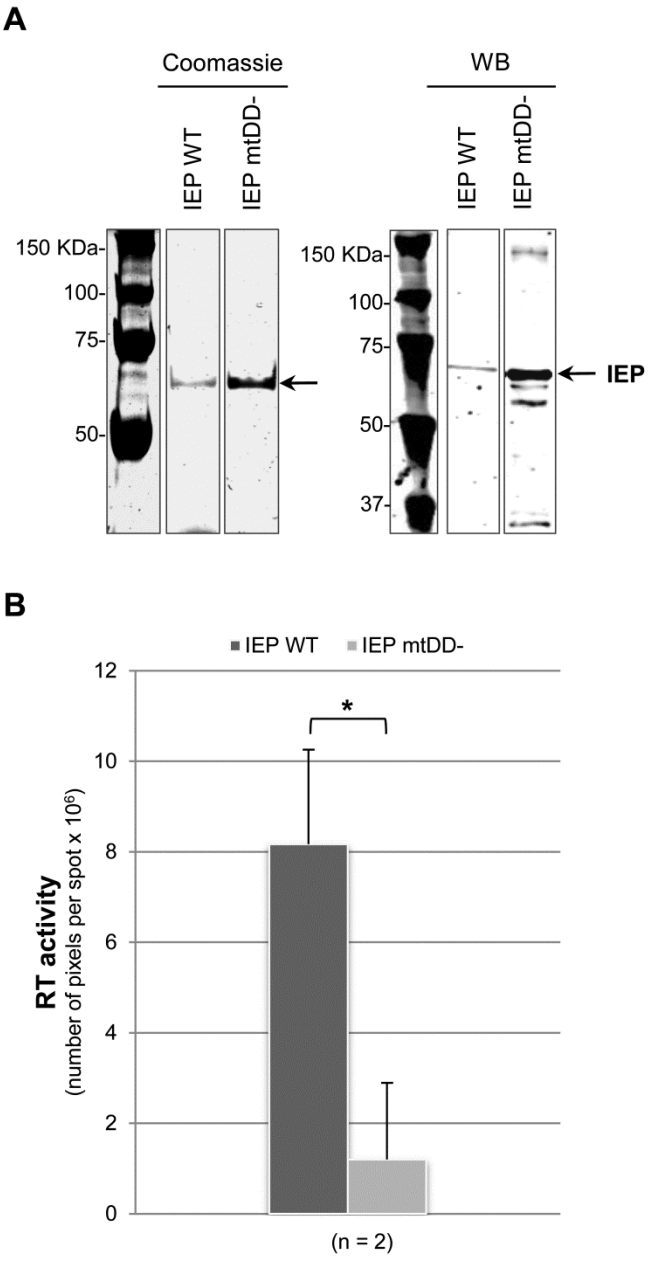
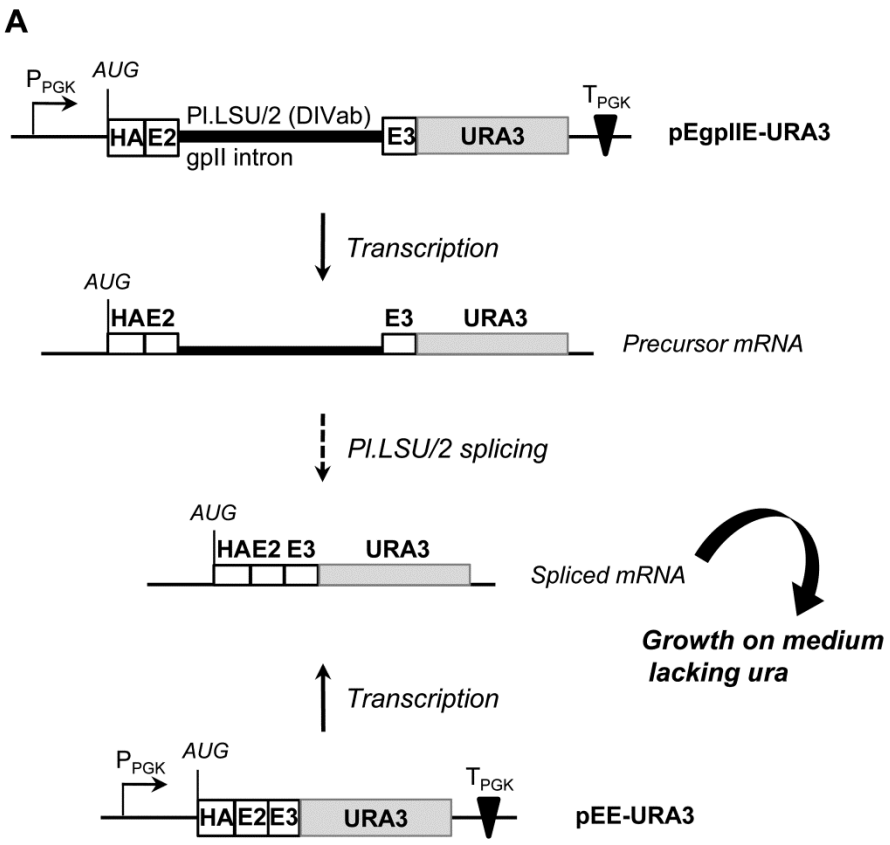

A
Coomassie
WB
IEP WT
IEP mtDD-
IEP WT
IEP mtDD-
150 KDa-
100-
75-
50-
150 KDa-
100-
75-
50-
37-
IEP
B
IEP WT
IEP mtDD-
12
10
8
6
4
2
0
*
RT activity
(number of pixels per spot x 10⁶)
(n = 2)

Figure 3



C

	Number of colonies / ml	
	+ ura	*- ura*
pEE-URA3	> 3.10^3	> 3.10^3
pEgpIIE-URA3	> 3.10^3	0
pEgpIIE-URA3 +pNLS-IEPco	> 3.10^3	0

Figure 4

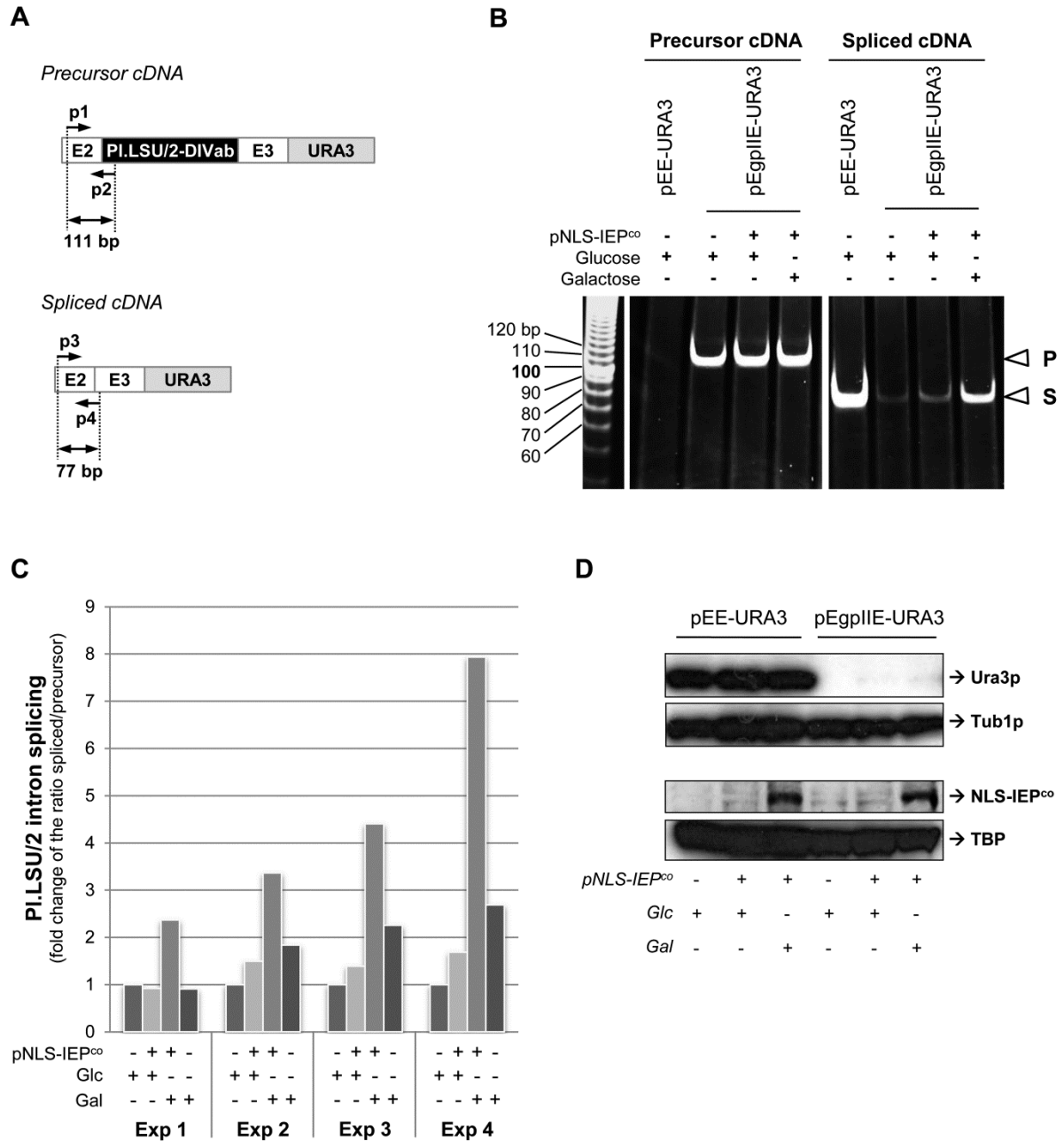

A
Precursor cDNA
p1
E2
Pl.LSU/2-DIVab
E3
URA3
p2
111 bp
Spliced cDNA
p3
E2
E3
URA3
p4
77 bp
B
Precursor cDNA
Spliced cDNA
pEE-URA3
pEgpIIE-URA3
pEE-URA3
pEgpIIE-URA3
pNLS-IEPco
Glucose
Galactose
120 bp
110
100
90
80
70
60
P
S
C
Pl.LSU/2 intron splicing
(fold change of the ratio spliced/precursor)
9
8
7
6
5
4
3
2
1
0
pNLS-IEPco
Glc
Gal
Exp 1
Exp 2
Exp 3
Exp 4
D
pEE-URA3
pEgpIIE-URA3
Ura3p
Tub1p
NLS-IEPco
TBP
pNLS-IEPco
Glc
Gal

Figure 5

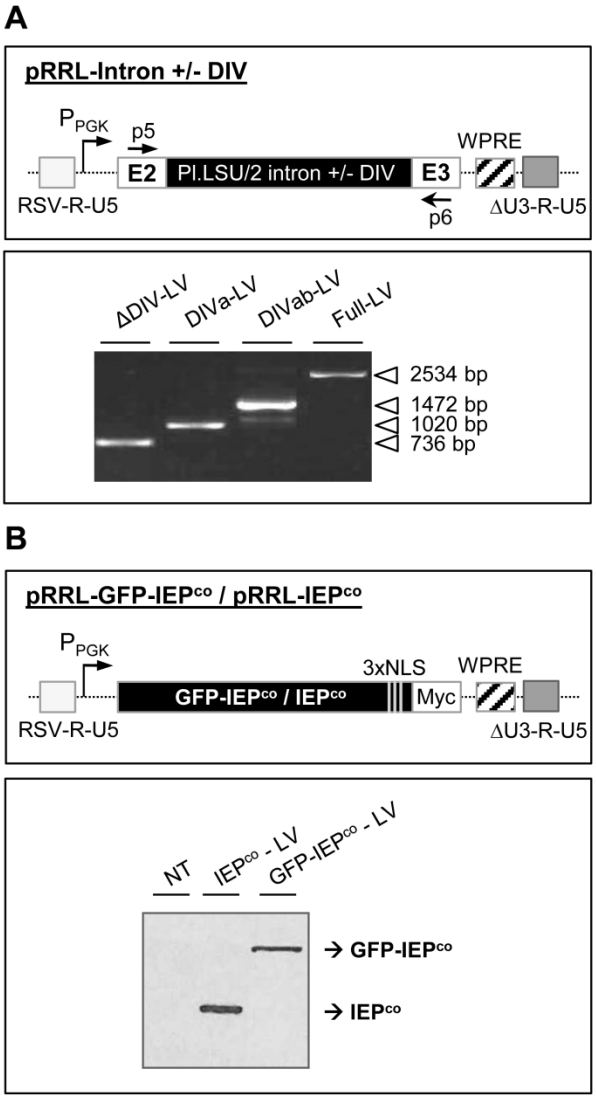
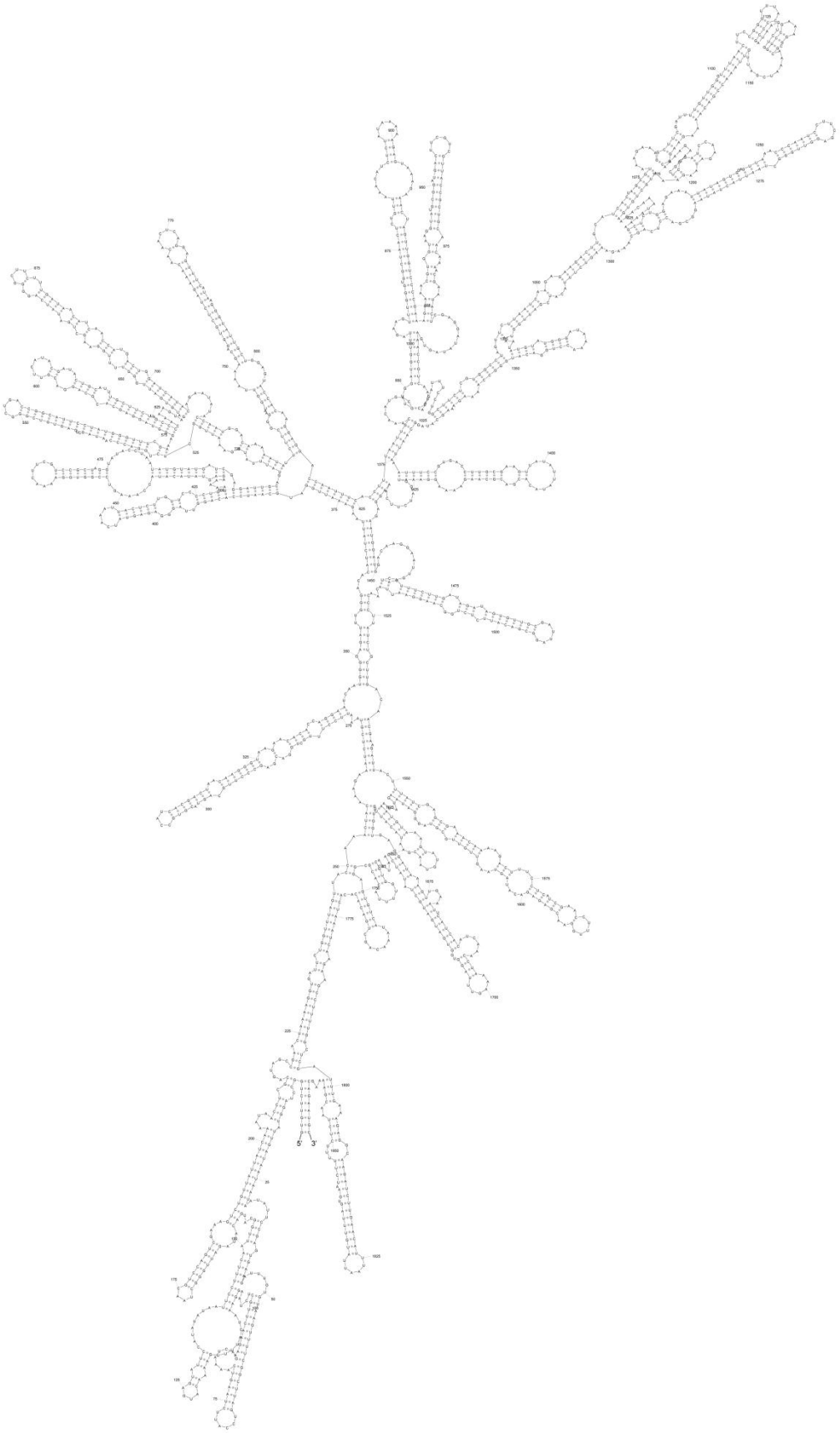

A
pRRL-Intron +/- DIV
P PGK
p5
E2
Pl.LSU/2 intron +/- DIV
E3
WPRE
RSV-R-U5
p6
ΔU3-R-U5
ΔDIV-LV
DIVa-LV
DIVab-LV
Full-LV
2534 bp
1472 bp
1020 bp
736 bp
B
pRRL-GFP-IEPco / pRRL-IEPco
P PGK
3xNLS
WPRE
GFP-IEPco / IEPco
Myc
RSV-R-U5
ΔU3-R-U5
NT
IEPco - LV
GFP-IEPco - LV
→ GFP-IEPco
→ IEPco

Figure 6

Supplemental Figure S1

DIV domain



Supplemental Table S1. Plasmids and oligonucleotides used in this work

A. Plasmids

Plasmid	Relevant characteristics	Reference
pET151/D-TOPO	*E. coli* T7 expression vector allowing a directional TOPO cloning in frame of a 6xHis/V5 tag.	Invitrogen
p151-IEP	Pl.LSU/2 IEP fused to a 6xHis/V5 tag in its N-terminus and expressed from a T7 promoter on the pET151 plasmid.	This work
p151-IEPmtDD-	p151-IEP derivative vector in which the conserved YADD motif of the IEP RT domain is changed in YAAA. The resulting IEP mtDD- protein should be RT-defective.	This work
p151-E+I+IEP	Pl.LSU/2 full length group II intron expressed from a T7 promoter on the pET151 plasmid and containing the IEP fused to a 6xHis/V5 tag in its N-terminus.	This work
p151-E+I+IEPmtDD-	p151-E+I+IEP derivative in which the conserved YADD motif of the IEP RT domain is changed in YAAA.	This work
pBFG1	2µ plasmid containing the LEU2 gene selection marker, 3 HA epitopes and a PGK promoter.	(Yelin et al. 1999)
pCI-neo	Mammalian expression vector containing a neomycin phosphotransferase gene.	Promega
pCIneo-E2E3mGFP	In frame Pl.LSU/2 flanking exons E2 (the last 50 nt) and E3 (the first 71 nt) containing a mutation that converts a STOP codon in E3 in tyrosine (E3m) and fused to the 5'-end of the GFP ORF on the pCI-neo plasmid.	This work
pCIneo-E2IntronDIVabE3mGFP	pCIneo derived plasmid containing a Pl.LSU/2 intron deleted form in which section from nucleotide 369 to nucleotide 1446 of DIV domain was removed (intron DIVab), flanked by the last 50 nt of E2 and the first 71 nt of E3 with a mutation that converts a STOP codon in tyrosine (E3m) and fused to the 5'-end of the GFP ORF.	This work
pEE-URA3	Pl.LSU/2 flanking exons (the last 50 nt of E2 and the first 71 nt of E3m) fused to 3 HA epitopes in its N-terminus and to a URA3 gene lacking the start codon in its C-terminus expressed from a PGK promoter on the pBFG1 plasmid.	This work
pEgpIIE-URA3	URA3-based PL.LSU/2 intron splicing reporter. The last 50 nt of E2, the Pl.LSU/2 intron DIVab form, see above) and the first 71 nt of E3 with a mutation that converts a STOP codon in tyrosine (E3m) are cloned just downstream of a URA3 gene lacking the start codon. The whole cassette is expressed from a PGK promoter on the pBFG1 plasmid.	This work

pNLS-IEP^{co}	Codon-optimized Pl.LSU/2 NLS-IEP fused to a c-myc epitope in its C-terminus and expressed from the GAL10 promoter of the pYEF1 plasmid on the pRS413 plasmid.	This work
pPl.LSU/2	Cloning plasmid containing the Pl.LSU/2 group II intron flanked by the last 50 nt of exon 2 and the first 71 nt of exon 3.	(Costa et al. 1997b)
pPl.LSU/2-ΔDIV	pPl.LSU/2 derivative construct in which most of the intron DIV domain was removed : section from nucleotide 8 to nucleotide 1818 of the DIV domain was replaced by CCTAGGATCT.	(Costa et al. 1997b)
pRRL-backbone	Advanced generation SIN Tat-independent HIV lentiviral vector system.	(Charrier et al. 2007)
pRRL-intron-ΔDIV	pRRL-backbone derived plasmid containing containing the Pl.LSU/2 group II intron with ΔDIV deletion (see above) flanked by the last 50 nt of exon 2 and the first 71 nt of exon 3.	This work
pRRL-intron-DIVa	pRRL-backbone derived plasmid containing the last 50 nt of exon 2, a Pl.LSU/2 intron deleted form in which section from nucleotide 244 to nucleotide 1772 of DIV domain was removed, and the first 71 nt of exon 3.	This work
pRRL-intron-DIVab	pRRL-backbone derived plasmid containing the last 50 nt of exon 2, the Pl.LSU/2 intron DIVab form (See above) and the first 71 nt of exon 3.	This work
pRRL-intron-Full	pRRL-backbone derived plasmid containing Pl.LSU/2 group II intron flanked by the last 50 nt of exon 2 and the first 71 nt of exon 3.	This work
pRRL-GFP	pRRL-backbone derived plasmid containing GFP ORF	This work
pRRL-GFP-IEP^{co}	pRRL-backbone derived plasmid containing codon-optimized Pl.LSU/2 IEP ORF fused to the GFP ORF in its N-terminus and to3 NLS and a c-myc epitope in its C-terminus.	This work
pRRL-IEP^{co}	pRRL-backbone derived plasmid containing codon-optimized Pl.LSU/2 IEP fused to 3 NLS and a c-myc epitope in its C-terminus.	This work
pRS413	ARS-CEN plasmid with the HIS3 gene selection marker.	(Sikorski and Hieter 1989; Christianson et al. 1992)
pRS426	2µ plasmid containing the URA3 gene selection marker.	(Christianson et al. 1992)
pUC57	*E. coli* cloning plasmid.	Genescript
pUC57-NLS-IEP^{co}	Codon-optimized Pl.LSU/2 IEP with 3 NLS and a c-myc epitope in its C-terminus.	This work

pYEF1	2µ plasmid containing the GAL10 promoter and terminator.	(Cullin and Minvielle-Sebastia 1994)

B. Oligonucleotides

Oligonucleotide	**Sequence (5' to 3')**	**Use**
BamHI-K-IEP	GGGATCCACCATGAGTATTCCTTACATAATTCCG	Amplification of IEP^{co} ORF
DM2	GTAGCTTTCGAAGCTTTACCTGCCGGCACC	Amplification of E2E3m and E2-IntronDIVab-E3m
DM3	GGTAAAGCTTCGAAAGCTACATATAAGGAA	Amplification of URA3 ORF
DM4	GAATTCAGTTTTTTAGTTTTGCTGG	Amplification of URA3 ORF
EcoRI-Stop-Myc	AGAATTCCTAGGCAGCGCCGTTCAG	Amplification of IEP^{co} ORF
p1	CTTTTATCTTTGACACAAAATCGGGGG	Amplification of the precursor cDNA (qPCR)
p2	TCCTGAACTTCTTGTCGCACTTTTTA	Amplification of the precursor cDNA (qPCR)
p3	AGGATCCCAGCTTTTATCTTTGACACA	Amplification of E2E3m, E2-IntronDIVab-E3m, and the spliced cDNA (qPCR)
p4	CGAGTTAGCAGAGACCTGTGTTTTTA	Amplification of the spliced cDNA (qPCR)
p5	CTTTTATCTTTGACACAAAATCG	Amplification of the precursor cDNA (PCR)
p6	GCAGGTGTCAGTCCCTATACA	Amplification of the precursor cDNA (PCR)
p7	ATTCACGCGTGGTACCTCTAGAA	Amplification of the spliced cDNA (qPCR)
PlLSU2-AS2	TTAAATGTTCAAGATCTTGC	Amplification of IEP ORF
PlLSU2-S2	CACCATGAGTATTCCATATATA	Amplification of IEP ORF
PlLSU2-SacI-AS	CGAGCTCTCGATAAGCTTTACCTGCCG	Amplification of Intron (domains IVb, V and VI)-E3
PlLSU2-XbaI-AS	GCTCTAGAGTTTTCAAAATGATTTCCTTAGAGCAAG	Amplification of E2-Intron (domains I, II, III and IVa)
PlLSU2-XbaI-S	GCTCTAGAACTAGTGGATCCCCCGGGCTGCA	Amplification of E2-Intron (domains I, II, III and IVa)

PlLSU2-XhoI-S	AAACGAATACTCGAGGATATAGTGAAACCG	Amplification of Intron (domains IVb, V and VI)-E3
RM-R	GTGTGCATTCGTAATGTCTGCCCATTCT	Reverse transcription of total yeast RNA
Sal-GAL-F	GTCGACCTAAACTCACAAATTAGAGCTTC	Amplification of the GAL10 promoter and terminator
Xba-GAL-R	TCTAGATGTGAGTTAGCTCACTCATTAG	Amplification of the GAL10 promoter and terminator
YAAA_F	TGGTAAGGTATGCGGCTGCCTTCGTCGTTACCGC	Site-directed mutagenesis of the YADD motif in the RT domain of IEP
YAAA_R	GCGGTAACGACGAAGGCAGCCGCATACCTTACCA	Site-directed mutagenesis of the YADD motif in the RT domain of IEP

Supplemental references

Christianson TW, Sikorski RS, Dante M, Shero JH, Hieter P. 1992. Multifunctional yeast high-copy-number shuttle vectors. *Gene* **110**: 119-122.

Cullin C, Minvielle-Sebastia L. 1994. Multipurpose vectors designed for the fast generation of N- or C-terminal epitope-tagged proteins. *Yeast* **10**: 105-112.

Sikorski RS, Hieter P. 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics* **122**: 19-27.

Yelin R, Rotem D, Schuldiner S. 1999. EmrE, a small Escherichia coli multidrug transporter, protects Saccharomyces cerevisiae from toxins by sequestration in the vacuole. *Journal of bacteriology* **181**: 949-956.

3.3 - ADDITIONNAL RESULTS

3.3.1 - Mass spectrometry analysis of HisV5-IEP purified fractions

The fusion HisV5-IEP and HisV5-IEP mtDD-, expressed in Rosetta-gami B (DE3), were purified using two methods: IMAC and sucrose centrifugation. In the case of IMAC purification, we showed that a 75-kDa *E. coli* contaminant protein was also co-purified with the 69-kDa HisV5-tagged proteins (See Fig. R-23). In contrast, the sucrose purifications of HisV5-tagged proteins were free of any contaminant proteins (See Fig. 3A of article 2; Coomassie). To confirm the identity of HisV5-IEP and HisV5-IEP mtDD- in both purification fractions and also identify the 75-kDa protein in IMAC purification fraction, we analyzed the purified proteins by mass spectrometry (MS). The 69-kDa and the 75-kDa bands were excised from SDS-PAGE gel stained with Coomassie blue, trypsin-digested in order to generate peptides and analyzed by MALDI-TOF/TOF.

(a) MS spectra

The first analysis consisted in peptide fingerprint mass mapping, as an initial identification of proteins in each sample. We used the search program MS-Fit (http://prospector2.ucsf.edu). This program allows to compare the experimental mass values of digested peptides with theoretical values from proteins databases, calculated using the cleavage specificity of the enzyme used (trypsin). We used here the UniProtKB/SwissProt annotated database in order to identify the 75-kDa *E. coli* contaminant protein and also determine if some *E. coli* proteins would be found in the 69-kDa bands, which are expected to contain HisV5-IEP or HisV5-IEP mtDD-. The results obtained for HisV5-IEP IMAC purification are described (Fig. R-28).

A

69-kDa band

Protein Hit Number	# mat % mat	% Cov	Protein MW (Da)/pI	Accession #	Species	Gene	Protein Name
1	**16/10**	**35.3**	**66851/5.6**	**Q5PKV9**	SALPA	glmS	Glucosamine-fructose-6-phosphate aminotransferase
2	**16/10**	**35.3**	**66878/5.6**	**Q8ZKX1**	SALTY	glmS	Glucosamine-fructose-6-phosphate aminotransferase
3	**15/10**	**31.0**	**66779/5.6**	**Q8FBT4**	ECOL6	glmS	Glucosamine-fructose-6-phosphate aminotransferase
4	**15/10**	**31.0**	**66867/5.6**	**Q83IY4**	SHIFL	glmS	Glucosamine-fructose-6-phosphate aminotransferase
5	**15/10**	**31.0**	**66895/5.6**	**P17169**	ECOLI	glmS	Glucosamine-fructose-6-phosphate aminotransferase

1. SALPA: *Salmonella paratyphi A.*
2. SALTY: Salmonella typhimurium.
3. ECOL6: *Escherichia coli* O6.

4. SHIFL: *Shigella flexneri.*
5. ECOLI: *Escherichia coli* (strain K12).

B

75-kDa band

Protein Hit Number	# mat % mat	% Cov	Protein MW (Da)/pI	Accession #	Species	Gene	Protein Name
1	**49/15**	**42.7**	**123725/9.5**	**Q44363**	RHIRD	traA	Conjugal transfer protein traA
2	**32/10**	**52.4**	**74290/6.4**	**C4ZU97**	ECOBW	arnA	Bifunctional polymyxin Resistance protein ArnA
3	**32/10**	**52.4**	**74290/6.4**	**B1X8W8**	ECODH	arnA	Bifunctional polymyxin Resistance protein ArnA
4	**32/10**	**52.4**	**74290/6.4**	**B1IXT2**	ECOLC	arnA	Bifunctional polymyxin Resistance protein ArnA
5	**32/10**	**52.4**	**74290/6.4**	**P77398**	ECOLI	arnA	Bifunctional polymyxin Resistance protein ArnA

1. RHIRD: *Rhizobium radiobacter*.
2. ECOBW: *Escherichia coli* (strain K12 / MC4100 / BW2952).
3. ECODH: *Escherichia coli* (strain K12 / DH10B).
4. ECOLC: *Escherichia coli* (strain ATCC 8739 / DSM 1576 / Crooks).
5. ECOLI: *Escherichia coli* (strain K12).

Figure R-28: MS-Fit searches.
The 75-kDa and 69-kDa bands detected in HisV5-IEP IMAC purification fractions were excised from SDS-PAGE gel, trypsin-digested and analyzed by MADI-TOF/TOF. Resulting MS-data were subjected to MS-Fit searches. #mat: number of peptides of the MS-data query that matched with theoretical data of the protein hit; % mat: percentage of matches; % Cov: percentage of the protein hit covering by the matched peptides; MW (Da): protein molecular weight in Daltons; pI: protein isoelectric point; Accession#: UniProtKB/SwissProt accession number. (***A***) MS-Fit closest matches found with the 69-kDa sample as a query. (***B***) MS-Fit closest matches found with the 75-kDa sample as a query. Proteins with relevant molecular weight are highlighted in yellow.

MS-Fit results show that relevant matches are found with MS data obtained with the 69-kDa sample of IMAC purification fraction (Fig. R-28A). The 5 closest matches correspond to the glucosamine-fructose-6-phosphate aminotransferase protein, whose molecular weight is around 67-kDa and the third and fifth closest matches correspond to *E. coli* species (Fig. R-28A). The matched peptides cover 31% of the protein hit, which is sufficient to be relevant and indicates that the 69-kDa band, expected to contain either HisV5-IEP is contaminated by the 67-kDa Glucosamine-fructose-6-phosphate aminotransferase *E. coli* protein. Same results were obtained with the 69-kDa sample of HisV5-IEP mtDD- IMAC purification fraction and HisV5-IEP sucrose cushion centrifugation fraction (data not shown).

MS-Fit results obtained with MS-data of the 75-kDa sample show that the first closest match correspond to a 124-kDa *Rhizobium radiobacter* protein (Fig. R-28B; protein hit #1). The protein molecular weight is not relevant as the band excised from the gel was just below the 75-kDa marker. In contrast, the other closest matches correspond to a 74-kDa *E. coli* protein (Fig. R-28B; Bifunctional

polymyxin Resistance protein ArnA, highlight in yellow). In each case, the matched peptides cover 52.4% of the protein hit. These results indicate that the contaminant protein in the HisV5-IEP IMAC fraction corresponds to the 74-kDa *E. coli* Bifunctional polymyxin Resistance protein ArnA. Same results were obtained with the HisV5-IEP mtDD- IMAC fraction.

The MS-Fit search cannot identify the Pl.LSU/2 IEP as a match. Indeed, the protein database used is UniProtKB/Swiss-Prot and the Pl.LSU/2 IEP sequence is not an entry in this database. The Pl.LSU/2 IEP sequence is so far unreviewed and thus appears as an entry only in the UniProtKB/TrEMBL database.

In order to determine the presence and identity of HisV5-IEP and HisV5-IEP mtDD- in the 69-kDa bands, we have compared experimental MS-data obtained using MALDI-TOF/TOF to theoretical Pl.LSU/2 IEP MS-data. Therefore, we have performed an *in silico* MS-digest (http://prospector2.ucsf.edu) of HisV5-IEP. MS-digest is a program that determinates the theoretical fingerprint mass pattern of a protein. The program performs an *in silico* digestion of the query protein to obtain theoretical mass values of expected digested peptides (Fig. R-29). MS-digests of HisV5-IEP and HisV5-IEP mtDD- are identical because the YAAA mutation in HisV5-IEP mtDD- does not change the trypsin cleavage profile.

HisV5-IEP *in silico* MS-Digest results
Protease used: Trypsin
pI of Protein: **9.8**
Protein MW: **69168**

1	MHHHHHHGKP	IPNPLLGLDS	TENLYFQGID	PFTMSIPYII	PFNWHDIDWA
51	NVQSKVCYYQ	NNLAVAELKG	DSGLVTKLQR	NLVNSFAGRA	LAVRAITTNK
101	GKNTPGINGE	IWDTSIKKLD	AIHRLGRVSN	YSCSPVKRVY	IPKSGGKLRP
151	LGIPNMYDRG	LQYLWKLALD	PIAECRADRH	SYGFRKGRST	QDVHTILHLL
201	LSPKSRCDWV	LEADIRGFFD	NINHDWIIQN	IPMDKNILRE	WLKAGALETT
251	TQEFHKGIAG	VPQGGPISPL	IANMTLDGLE	VWVANSVKHL	YKKSKETSWS
301	PKVNVVRYAD	DFVVTAATKR	ILEDIVKPSI	QDFLASRGLV	LNQEKTCITS
351	VKKGFDFVGF	NFRVYPDKSG	PKGAKSIVKP	TKEGKRRLRS	KIRNAVKTNK
401	SSGEIIVELN	PILRGWANYY	KATSAKKVFT	SIGKYVWDKT	WTWAKRKHRQ
451	LNFRDLAKLY	YTRRKKRKWI	FKGEWMDKEL	TIFLIDSVAI	RRHSLARNYN
501	PYLLDNEDYF	IERNKRLSSS	NLWNERHSKL	LRRDKYKCKV	CNEYICGEDK
551	VEIHHIKPKS	LGGDDAISNN	VVLHAECHKQ	LTHTKSRSLL	ARFERGKILN
601	I				

m/z (mi)	Start	End	Missed Cleavages	Sequence
802.4894	586	592	1	(K) SRSLLAR(F)
832.4887	95	102	1	(R) AITTNKGK(N)
834.3992	296	302	0	(K) ETSWSPK(V)
852.5050	118	124	1	(K) KLDAIHR(L)
871.4785	459	464	1	(K) LYYTRR(K)
879.4968	346	353	1	(K) TCITSVKK(G)
879.5298	427	434	1	(K) KVFTSIGK(Y)
894.4581	180	186	1	(R) HSYGFRK(G)
900.5149	338	345	0	(R) GLVLNQEK(T)
901.4203	415	421	0	(R) GWANYYK(A)
907.5036	160	166	0	(R) GLQYLWK(L)
948.5050	440	446	1	(K) TWTWAKR(K)
948.5513	139	147	1	(R) VYIPKSGGK(L)
970.5330	448	454	1	(K) HRQLNFR(D)
970.5429	580	587	1	(K) QLTHTKSR(S)
977.5163	81	89	0	(R) NLVNSFAGR(A)
990.5255	364	372	1	(R) VYPDKSGPK(G)
991.5683	588	595	1	(R) SLLARFER(G)
1028.6463	373	382	1	(K) GAKSIVKPTK(E)
1049.5262	294	302	1	(K) SKETSWSPK(V)
1050.6167	119	127	1	(K) LDAIHRLGR(V)
1071.6309	236	243	1	(K) NILREWLK(A)
1083.5139	128	137	0	(R) VSNYSCSPVK(R)
1086.6517	376	385	1	(K) SIVKPTKEGK(R)
1100.5769	167	176	0	(K) LALDPIAECR(A)
1100.6575	551	559	0	(K) VEIHHIKPK(S)
1104.6160	450	458	1	(R) QLNFRDLAK(L)
1108.5283	177	185	1	(R) ADRHSYGFR(K)
1142.6204	455	463	1	(R) DLAKLYYTR(R)
1157.7001	90	100	1	(R) ALAVRAITTNK(G)
1173.6586	70	80	1	(K) GDSGLVTKLQR(N)
1205.5738	354	363	0	(K) GFDFVGFNFR(V)
1219.5776	207	216	0	(R) CDWVLEADIR(G)
1239.6150	128	138	1	(R) VSNYSCSPVKR(V)
1272.5235	540	550	0	(K) VCNEYICGEDK(V)
1300.6420	308	319	0	(R) YADDFVVTAATK(R)
1333.6688	353	363	1	(K) KGFDFVGFNFR(V)
1339.6504	469	478	1	(K) WIFKGEWMDK(E)
1359.6692	415	426	1	(R) GWANYYKATSAK(K)
1361.6920	516	526	1	(K) RLSSSNLWNER(H)
1374.7601	78	89	1	(K) LQRNLVNSFAGR(A)
1401.7485	296	307	1	(K) ETSWSPKVNVVR(Y)
1409.7206	125	137	1	(R) LGRVSNYSCSPVK(R)
1432.7067	244	256	0	(K) AGALETTTQEFHK(G)

m/z (mi)	Start	End	Missed Cleavages	Sequence
1442.7420	167	179	1	(K) LALDPIAECRADR(H)
1442.7678	428	439	1	(K) VFTSIGKYVWDK(T)
1444.7729	148	159	0	(K) LRPLGIPNMYDR(G)
1456.7431	308	320	1	(R) YADDFVVTAATKR(I)
1462.7107	205	216	1	(K) SRCDWVLEADIR(G)
1483.7369	435	445	1	(K) YVWDKTWTWAK(R)
1487.8441	81	94	1	(R) NLVNSFAGRALAVR(A)
1489.8625	479	491	0	(K) ELTIFLIDSVAIR(R)
1503.6276	538	550	1	(K) CKVCNEYICGEDK(V)
1539.8741	401	414	0	(K) SSGEIIVELNPILR(G)
1557.7768	517	529	1	(R) LSSSNLWNERHSK(L)
1627.8149	56	69	0	(K) VCYYQNNLAVAELK(G)
1632.8989	338	352	1	(R) GLVLNQEKTCITSVK(K)
1644.8228	103	117	0	(K) NTPGINGEIWDTSIK(K)
1645.9636	479	492	1	(K) ELTIFLIDSVAIRR(H)
1772.9177	103	118	1	(K) NTPGINGEIWDTSIKK(L)
1773.9428	144	159	1	(K) SGGKLRPLGIPNMYDR(G)
1802.0171	189	204	0	(R) STQDVHTILHLLLSPK(S)
1807.8802	354	368	1	(K) GFDFVGFNFRVYPDK(S)
1829.9392	101	117	1	(K) GKNTPGINGEIWDTSIK(K)
1867.9912	303	319	1	(K) VNVVRYADDFVVTAATK(R)
1883.0597	398	414	1	(K) TNKSSGEIIVELNPILR(G)
1944.0801	321	337	0	(R) ILEDIVKPSIQDFLASR(G)
1989.0076	240	256	1	(R) EWLKAGALETTTQEFHK(G)
1989.0626	160	176	1	(R) GLQYLWKLALDPIAECR(A)
2015.1396	187	204	1	(K) GRSTQDVHTILHLLLSPK(S)
2045.1502	189	206	1	(R) STQDVHTILHLLLSPKSR(C)
2077.9502	498	513	0	(R) NYNPYLLDNEDYFIER(N)
2100.1812	320	337	1	(K) RILEDIVKPSIQDFLASR(G)
2236.1682	473	491	1	(K) GEWMDKELTIFLIDSVAIR(R)
2317.1070	217	235	0	(R) GFFDNINHDWIIQNIPMDK(N)
2320.0881	498	515	1	(R) NYNPYLLDNEDYFIERNK(R)
2333.2587	148	166	1	(K) LRPLGIPNMYDRGLQYLWK(L)
2354.1631	540	559	1	(K) VCNEYICGEDKVEIHHIKPK(S)
2385.2119	56	77	1	(K) VCYYQNNLAVAELKGDSGLVTK(L)
2422.2765	401	421	1	(K) SSGEIIVELNPILRGWANYYK(A)
2642.2634	493	513	1	(R) HSLARNYNPYLLDNEDYFIER(N)
2787.3843	560	585	1	(K) SLGGDDAISNNVVLHAECHKQLTHTK(S)
2813.4192	217	239	1	(R) GFFDNINHDWIIQNIPMDKNILR(E)
2825.5771	321	345	1	(R) ILEDIVKPSIQDFLASRGLVLNQEK(T)
3160.6320	551	579	1	(K) VEIHHIKPKSLGGDDAISNNVVLHAECHK(Q)
3203.7133	257	288	0	(K) GIAGVPQGGPISPLIANMTLDGLEVWVANSVK(H)
3517.6668	207	235	1	(R) CDWVLEADIRGFFDNINHDWIIQNIPMDK(N)
3745.0146	257	292	1	(K) GIAGVPQGGPISPLIANMTLDGLEVWVANSVKHLYK(K)

Figure R-29: HisV5-IEP *in silico* MS-digest.

The protein sequence of HisV5-IEP was subjected to an *in silico* tryptic MS-digest to obtain theoretical mass values (m/z; mass/charge) of expected digested peptides. The trypsine cleaves predominantly after arginine (R) and lysine (K) residues (underlined in the sequence). mi: monoisotopic peptide mass calculated with the use of the lowest common isotope for each element (^{12}C, ^{1}H, ^{14}N, ^{16}O, ^{32}S, ^{31}P); Start: first amino acid of the expected peptide; End: last amino acid of the expected peptide. The expected peptide ions are listed by m/z values.

We then compared theoretical peptide mass values of the *in silico* HisV5-IEP MS-digest (Fig. R-29) with the experimental mass values of digested peptides obtained using MALDI-TOF/TOF with the 69-kDa samples. The results obtained with HisV5-IEP WT and mtDD- IMAC fractions are described below (Fig. R-30).

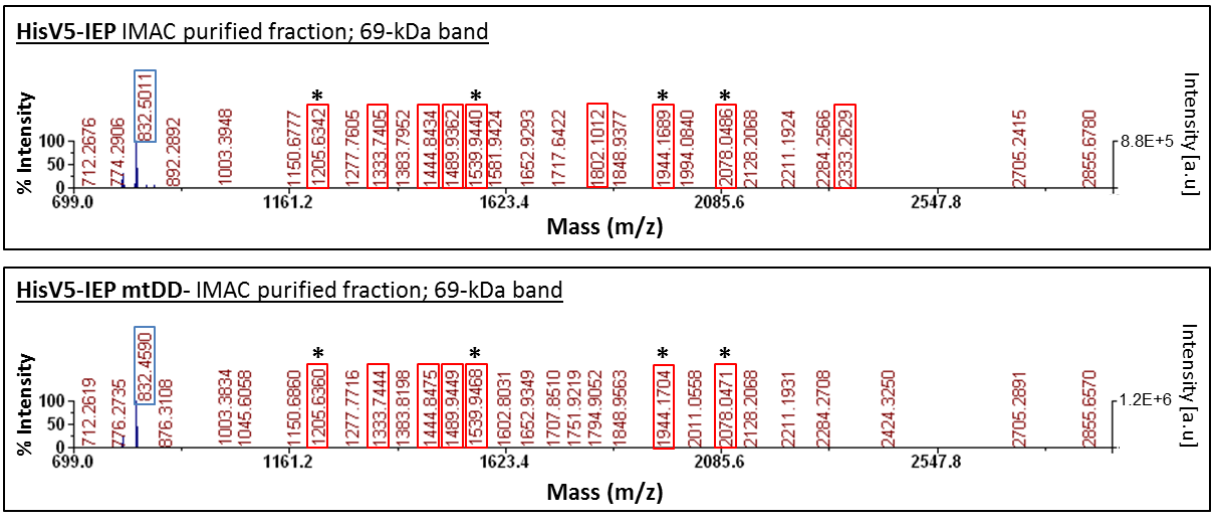


Figure R-30: Identification of HisV5-IEP and HisV5-IEP mtDD- purified by IMAC using MALDI-TOF/TOF.

After IMAC purification and separation by SDS-PAGE, bands expected to contain HisV5-IEP and HisV5-IEP mtDD- were excised from the 69-kDa mass range, digested with trypsin and analyzed using MALDI-TOF/TOF. % Intensity: percentage of peaks intensity; a.u: arbitrary unit. HisV5-IEP (upper panel) and HisV5-IEP mtDD- (lower panel) lysis products identified by the *in silico* MS-digest (See Fig. R-29), are indicated by red rectangles. Peptides for which an MS/MS fragmentation will be performed are indicated by asterisks. The HCCA matrix peak is indicated by blue rectangle.

MS spectra of the 69-kDa samples reveal the presence of 7 peptides whose mass values match with theoretical mass values of HisV5-IEP *in silico* digested peptides without missed cleavages (Fig. R-30; indicated by red rectangles). These results indicate the presence of HisV5-IEP and HisV5-IEP mtDD- in the 69-kDa band on IMAC purification fractions. The identity of HisV5-IEP WT and mtDD- purified by sucrose cushion centrifugation was also confirmed by MS spectra analysis (not shown).

(b) MS/MS spectra

To further confirm the identification of HisV5-IEP, four of the matched peptides (Fig. R-30; mass values of 1205.6342; 1539.9440; 1944.1689 and 2078.0486; indicated by asterisks), found using MALDI-TOF/TOF with HisV5-IEP IMAC purified fraction and theoretically corresponding to HisV5-IEP peptides without missed cleavages, were analyzed by MS/MS fragmentation. MS/MS fragmentation is used here to produce sequence information of the 4 peptides, also called parental

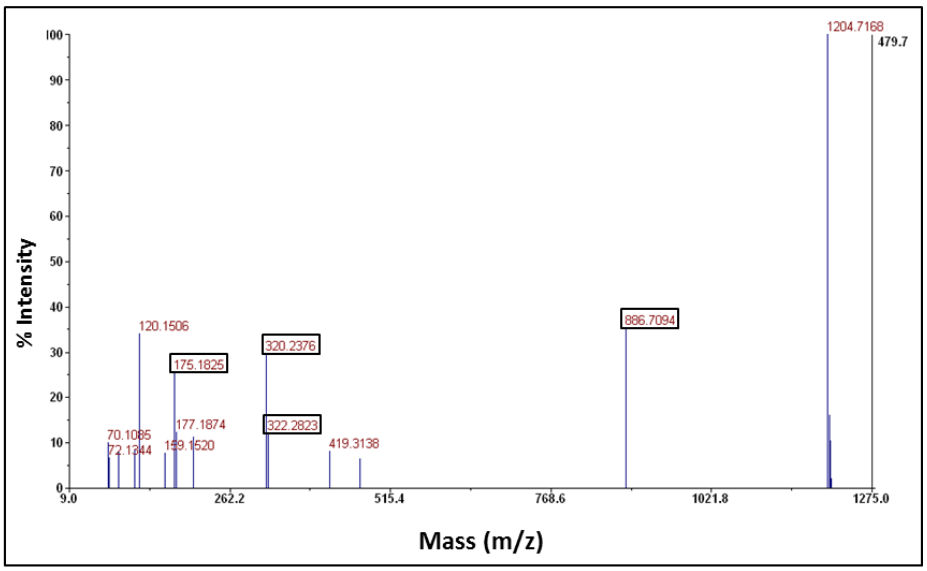
ions, by fragmentation inside the mass spectrometer and analysis of the fingerprint pattern of resulting fragment ions. In parallel, theoretical fingerprint patterns of the expected GFDFVGFNFR peptide 1 of m/z 1205.5738, SSGEIIVELNPILR peptide 2 of m/z 1539.8741, ILEDIVKPSIQDFLASR peptide 3 of m/z 1944.0801 and NYNPYLLDNEDYFIER peptide 4 of m/z 2077.9502, which were identified by the *in silico* MS-digest of HisV5-IEP (See Fig. R-29), are calculated. The comparison between theoretical and experimental patterns can then allow the identification of these four parental ions (Fig. R-31).

A

GFDFVGFNFR

MH^{+1}(mono)
1205.5738

b				y
---	1	G	10	---
205.0972	2	F	9	1148.5524
320.1241	3	D	8	1001.4839
467.1925	4	F	7	886.4570
566.2609	5	V	6	739.3886
623.2824	6	G	5	640.3202
770.3508	7	F	4	583.2987
884.3937	8	N	3	436.2303
1031.4621	9	F	2	322.1874
---	10	R	1	175.1190

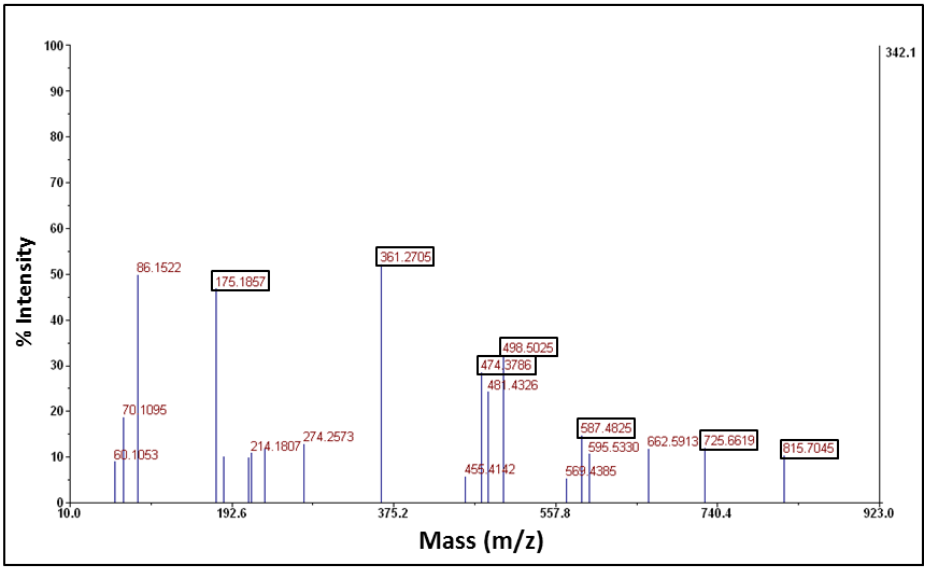


B

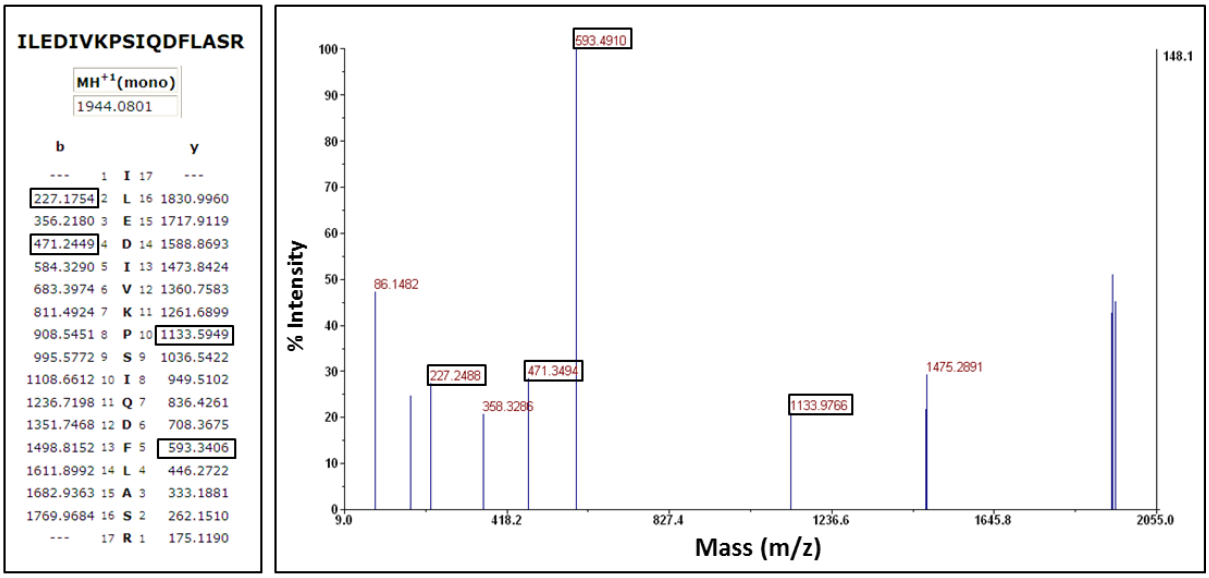
SSGEIIVELNPILR

MH^{+1}(mono)
1539.8741

b				y
---	1	S	14	---
175.0713	2	S	13	1452.8421
232.0928	3	G	12	1365.8100
361.1354	4	E	11	1308.7886
474.2195	5	I	10	1179.7460
587.3035	6	I	9	1066.6619
686.3719	7	V	8	953.5778
815.4145	8	E	7	854.5094
928.4986	9	L	6	725.4668
1042.5415	10	N	5	612.3828
1139.5943	11	P	4	498.3398
1252.6783	12	I	3	401.2871
1365.7624	13	L	2	288.2030
---	14	R	1	175.1190



C



D

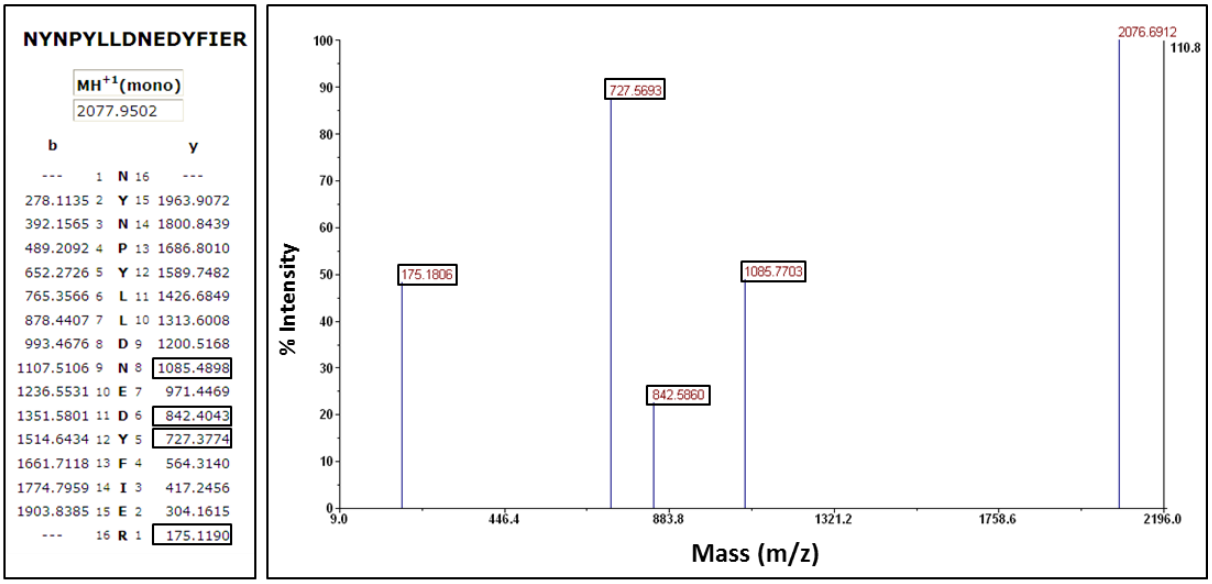


Figure R-31: MS/MS fragmentation of four peptides found by MALDI-TOF/TOF analysis of the 69-kDa sample in HisV5-IEP IMAC purified fraction.

The peptides whose mass values match with theoretical data are indicated by black rectangles. MH^{+1} (mono): monoisotopic peptide mass; b: peptide fragment ion when charge is retained at the N-terminus; y: peptide fragment ion when charge is retained at the C-terminus. Theoretical fingerprint pattern of the GFDFVGFNFR peptide 1 of m/z 1205.5738 (***A***; left panel), SSGEIIVELNPILR peptide 2 of m/z 1539.8741 (***B***; left panel), ILEDIVKPSIQDFLASR peptide 3 of m/z 1944.0801 (***C***; left panel), and NYNPYLLDNEDYFIER peptide 4 of m/z 2077.9502 (***D***; left panel) were manually compared with MS/MS fragmentations performed by MALDI-TOF/TOF on the parental ion of m/z 1205.6342 (***A***; right panel), 1539.9440 (***B***; right panel), 1944.1689 (***C***; right panel), and 2078.0486 (***D***; right panel) respectively.

MS/MS fragmentation of the parental ions of m/z 1205.5738, 1539.9440, 1944.1689 and 2078.0486 found by MALDI-TOF/TOF reveals the presence of fragment ions whose mass values match with

theoretical data of expected peptides (Fig. R-31; indicated by black rectangles). This confirms the identity of the parental ions analyzed and further demonstrates that HisV5-IEP and HisV5-IEP mtDD- are well contained in the 69-kDa band obtained by IMAC purification.

In conclusion, the mass spectrometry analysis of the HisV5-IEP and HisV5-IEP mtDD- IMAC purification fractions allowed the identification of the 74-kDa *E. coli* Bifunctional polymyxin Resistance protein ArnA in the 75-kDa band. Moreover, we determined that the 69-kDa band observed in the IMAC and sucrose cushion purification fractions contains HisV5-IEP or HisV5-IEP mtDD- as well as the 67-kDa glucosamine-fructose-6-phosphate aminotransferase *E. coli* protein. Notably, the matched peptides found using MALDI-TOF/TOF analysis of the 69-kDa sample of IMAC purification fraction cover 18% of HisV5-IEP and are located all along the protein sequence from the RVT-1 domain to the HNH domain of HisV5-IEP (Fig. R-32). It is accepted that a covering of about 20% is required to an identification of a protein by MALDI-TOF/TOF, and the correct identification of a minimum of 2 peptides by MS/MS further confirm the identification. We can thus admit that the HisV5-IEP is correctly identified under these criteria.

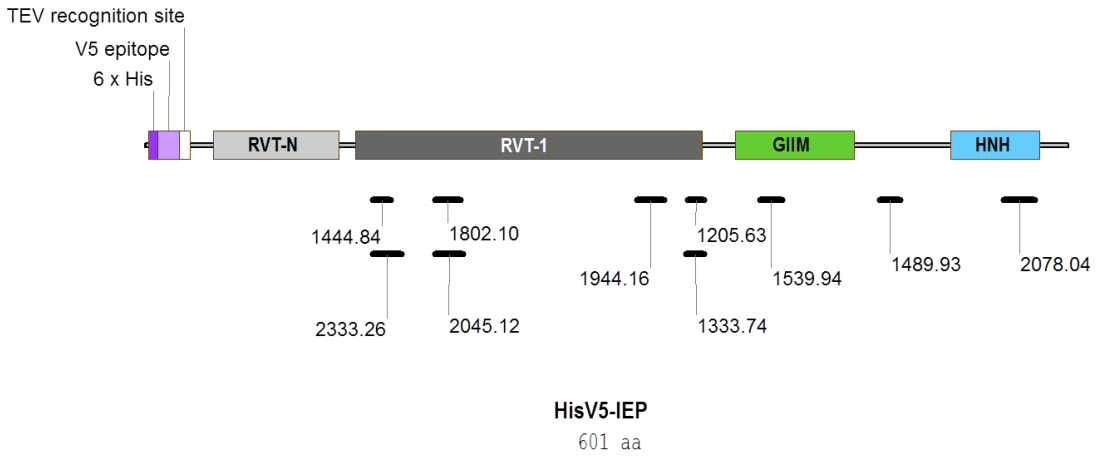


Figure R-32: Schematic representation of the HisV5-IEP covering.

Matched peptides found by MADLI-TOF/TOF analysis with the 69-kDa sample of HisV5-IEP IMAC purification fraction are represented by black lines and located on the HisV5-IEP sequence. The peptides m/z values are indicated for each peptide. 6xHis: Histidine tag; RVT-N: Pfam N-terminal domain of reverse transcriptase; RVT-1: Pfam reverse transcriptase domain (RNA-dependent DNA polymerase); GIIM: Pfam Group II intron, maturase-specific domain; HNH: Pfam HNH endonuclease domain; aa: amino acids.

3.3.2 - RNP particles purification by IMAC

As detailed previously (See Results section 2.4 -), the expression of HisV5-IEP in *E. coli* followed by purification using IMAC did not allow to demonstrate the RT activity of the IEP. We postulated that the Pl.LSU/2 intron RNA was required for the proper catalytic conformation of the IEP. Thus, we expressed both Pl.LSU/2 IEP and the Pl.LSU/2 intron in Rosetta-gami B (DE3) (See article 2). Before testing the centrifugation in sucrose cushion (See article 2), we first decided to use the IMAC purification process under native conditions to purify RNP particles. The same protocol than those used to purify HisV5-tagged proteins under native conditions (See Results section 2.4.4 -) was used to purify RNPs containing wild-type or mtDD- HisV5-IEP (Fig R-33A). The RT activity of purified RNP particles was then assayed (Fig. R-33B and R-33C) as described previously.

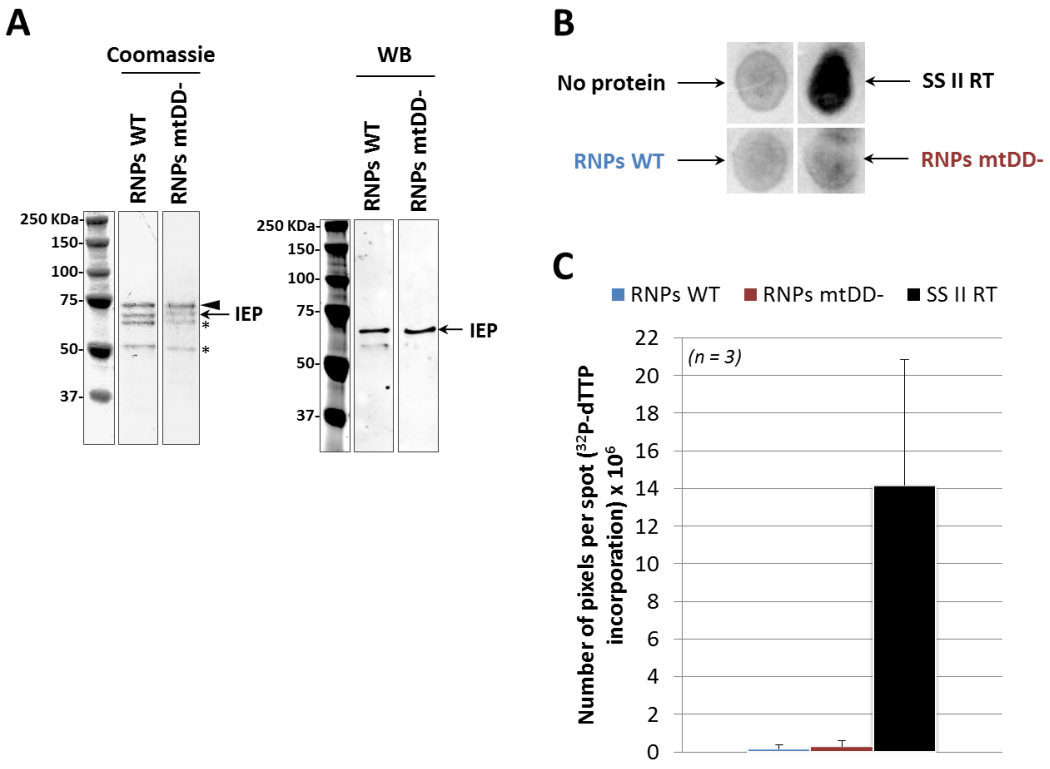


Figure R-33: RT activity of HisV5-IEP in RNP particles purified from *E. coli* by IMAC.

(***A***) SDS-PAGE analysis by Coomassie blue staining (Coomassie) and western blot (WB) of HisV5-IEP-containing RNPs purified by IMAC. Quantities of RNPs used were 9 and 18 OD_{260nm} units for RNP containing wild-type HisV5-IEP (RNPs WT) and RNP containing mutant HisV5-IEP mtDD- (RNPs mtDD-), respectively. A monoclonal mouse anti-V5 antibody was used to detect the HisV5-IEP by western blot. The IEP is indicated by black arrow. The 75-kDa *E. coli* contaminant protein is indicated by black arrowhead. Asterisks indicate *E. coli* contaminant proteins and/or degradation products. (***B***) RT reactions without proteins (No protein) and with 0.1 OD_{260nm} unit of RNPs WT or RNPs mtDD- were performed at 37°C for 45 min. Positive control consists of 0.03 U of SuperScript® II reverse transcriptase (SS II RT) diluted in the dialysis buffer used during RNPs purification. (***C***) Data, representing the number of pixels per spot, were quantified with ImageQuant™ software. Blue bar: RNPs containing HisV5-IEP; dark red bar: RNPs containing HisV5-IEP mtDD-; black bar: SuperScript® II reverse transcriptase. Data are the mean of three independent experiments and standard deviation is indicated by thin lines.

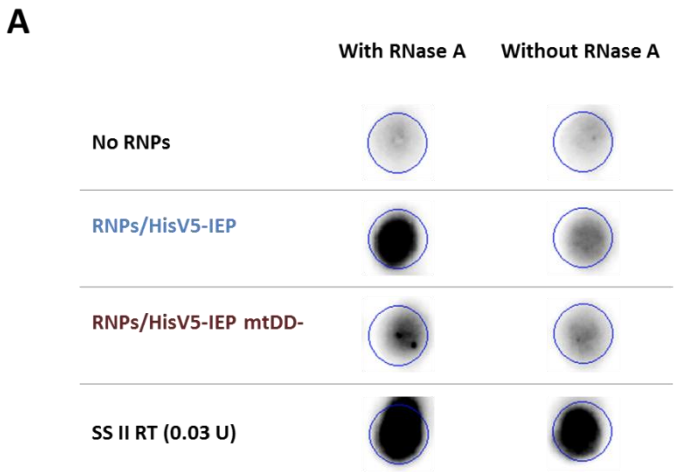
Coomassie blue stained SDS-PAGE shows that four proteins are present in the RNPs purified fractions with one protein at the expected size of HisV5-IEP (WT and mtDD-) (Fig. R-33A, Coomassie, indicated by black arrow). The western blot analysis confirms the identity of this protein as HisV5-IEP (WT and mtDD-) (Fig. R-33A; WB). The 75-kDa *E. coli* contaminant protein, which was co-purified with HisV5-IEP during IMAC purification under native conditions (See Fig. R-24), is also co-purified with HisV5-IEP contained in RNPs (Fig. R-33A; indicated by black arrowhead). The other proteins (Fig. R-33A; indicated by asterisks) can be *E. coli* contaminant proteins and/or HisV5-IEP (WT and DD-) degradations products. Even if RNPs particles could only be partially purified, RT assays were performed using these RNPs preparations. Figure R-33B shows the resulting membrane image of one experiment. We observe that the HisV5-IEP contained in RNPs and purified by IMAC does not display any RT activity. Quantification of data of three independent experiments was performed and

shows that both RNPs containing wild-type and mutant HisV5-IEP have no RT activity (Fig R-33C). These results confirmed that the IMAC purification process alters the protein folding and/or stability.

3.3.3 - Influence of RNase A on RT activity of Pl.LSU/2 IEP contained in RNPs

In early RT assays using artificial exogenous RNA template, we have evaluated the influence of an RNase A treatment on the RT activity of Pl.LSU/2 IEP contained in RNPs. RNase A is a ribonuclease specific to single-stranded RNA. Previous studies have shown that an RNase A digestion of the endogenous RNA contained in RNPs could have different implications on the RT activity of the intron-encoded protein depending on the group II intron used. Unlike the *Saccharomyces cerevisiae* aI2 group II intron, for which an RNase A digestion of RNP particles is necessary to release the RT from endogenous RNA (Moran JV et al. 1995; Zimmerly S et al. 1999), the *Lactococcus lactis* Ll.LtrB group II intron-encoded protein has essentially the same RT activity in presence or absence of RNase A (Matsuura M et al. 1997). In the case of the aI2 intron, the endogenous RNA has to be digested just prior to *in vitro* RT reactions using exogenous templates.

We thus evaluated the influence of an RNase A treatment on the RT activity of Pl.LSU/2 IEP contained in RNPs. *In vitro* RT assays were performed with HisV5-IEP-containing RNPs purified by sucrose centrifugation, as described in article 2, except that RNase A was added or not to the RT reaction medium (Fig. R-34).



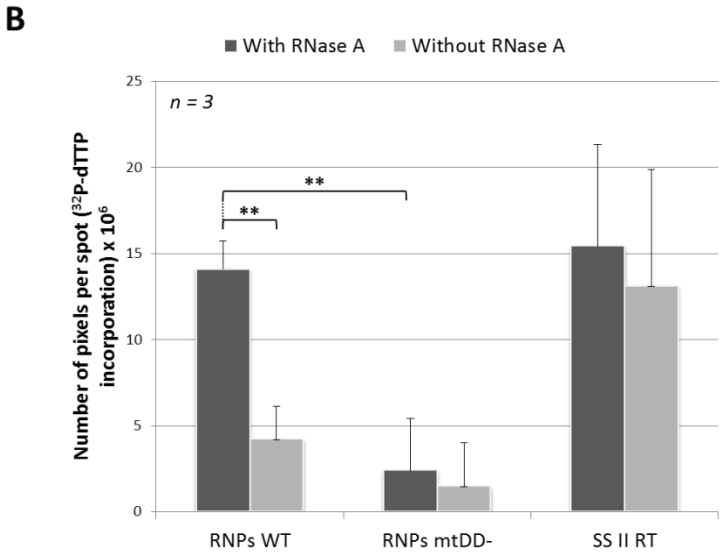


Figure R-34: Influence of RNase A treatment on RT activity of Pl.LSU/2 IEP contained in RNPs purified by sucrose centrifugation.

(***A***) RT reactions were performed without RNPs (No RNPs) or with 0.1 OD_{260nm} units of RNPs purified by sucrose centrifugation and containing either wild-type HisV5-IEP (RNPs/HisV5-IEP) or mutant HisV5-IEP (RNPs/HisV5-IEP mtDD-). Positive control consists of RT assay using 0.03 U of SuperScript® II reverse transcriptase (SS II RT). RT activity was assayed in presence (with RNase A) or absence (without RNase A) of RNase A in the reaction medium. (***B***) Data, representing the number of pixels per spot, were quantified with ImageQuant™ software and corrected to the background. Data are the mean of three independent experiments with the standard deviation indicated by thin lines. Data were subjected to a t-test using a unilateral pair-wise comparison procedure. A highly significant difference is indicated by asterisks ($p < 3.10^{-3}$).

Figure R-34 shows that HisV5-IEP contained in RNPs particles has an RT activity in presence of RNase A in the reaction medium (Fig. R-34; RNPs/HisV5-IEP, dark gray bar). In contrast this activity is highly significantly impede in absence of RNase A (Fig. R-34; RNPs/HisV5-IEP, light gray bar, $p < 3.10^{3}$). As expected, the RT activity of the protein is abolished by point mutations in the catalytic YADD motif of the RT domain (Fig. R-34; RNPs/HisV5-IEP mtDD-). The SuperScript II reverse transcriptase has an RT activity both in presence and absence of RNase A (Fig. R-34; SSII RT), as the activity of the SuperScript II is not expected to be influenced by an RNase A treatment. These results indicate that, as for the *S. cerevisiae* aI2 intron-encoded protein, the RT activity of Pl.LSU/2 IEP using an exogenous RNA template requires the release of the IEP from the endogenous RNA. This could suggest that a high binding affinity occurs between Pl.LSU/2 IEP and intron RNA so that the binding of IEP to the artificial RNA template in RT assays would be too rare without digestion of endogenous RNA.

4 - HOMING OF PL.LSU/2 GROUP II INTRON

4.1 - INTRODUCTION

The major aim of this work was to characterize the Pl.LSU/2 group II intron in order to evaluate its potential use in targeted genome engineering. This strategy relies on the ability of group II intron to transpose to a specific DNA target site by the homing mechanism. The homing of group II intron is achieved by RNP particles, formed after the IEP-mediated intron RNA splicing and composed by the IEP and the intron lariat. Both components of the RNP recognizes the DNA target site and the intron is integrated after several steps involving a "reverse splicing" of the intron into the sense strand of the DNA target site, the antisense-strand DNA cleavage and reverse transcription of the intron by the IEP, and finally the integration of a newly synthetized double-stranded cDNA copy of the intron into the DNA target site by cellular repair mechanisms. The homing property of several group II introns such as the bacterial Ll.LtrB and the yeast aI2 group II introns has been demonstrated. In contrast, the homing capacity of Pl.LSU/2 group II intron has never been characterized.

The Pl.LSU/2 IEP catalytic activities and the ability of the Pl.LSU/2 intron to splice *in vivo* are required for the homing process. We have shown in article 2 that Pl.LSU/2 IEP has an RT activity *in vitro* either alone or contained in RNP particles. We also demonstrated that the Pl.LSU/2 IEP has also a maturase activity, promoting the Pl.LSU/2 intron splicing in yeast cells. Two of the three biochemical activities of the IEP were thus demonstrated and the *in vivo* splicing ability of Pl.LSU/2 intron was also showed.

It was then desirable to determine whether Pl.LSU/2 intron homing could be observed. We first evaluated the homing of Pl.LSU/2 intron in *E. coli*. In parallel, the demonstration of IEP-promoted Pl.LSU/2 intron splicing in *S. cerevisiae* led us to design an homing assay in yeast.

4.2 - HOMING OF PL.LSU/2 IN *E. COLI*

Because of the facility of performing mechanistic studies in *E. coli*, the evaluation of Pl.LSU/2 intron homing was first attempted in that host, even though the splicing of Pl.LSU/2 intron was not demonstrated in *E. coli*. The transcription and translation take place in the same cellular compartment in bacteria. We expected that this feature would be favorable to the homing mechanism. The *E. coli* homing assay was performed using the natural intron DNA target site inserted into a plasmid in order to increase the chance of homing detection. We used a strategy adapted from a commercially available gene knockout system (TargeTron™, Sigma-Aldricht) based on the homing property of Ll.LtrB group II intron. A retrohoming-activated marker (RAM)-twintron selective approach was thus designed to detect retrohoming of Pl.LSU/2 in *E. coli* (Fig. R-35).

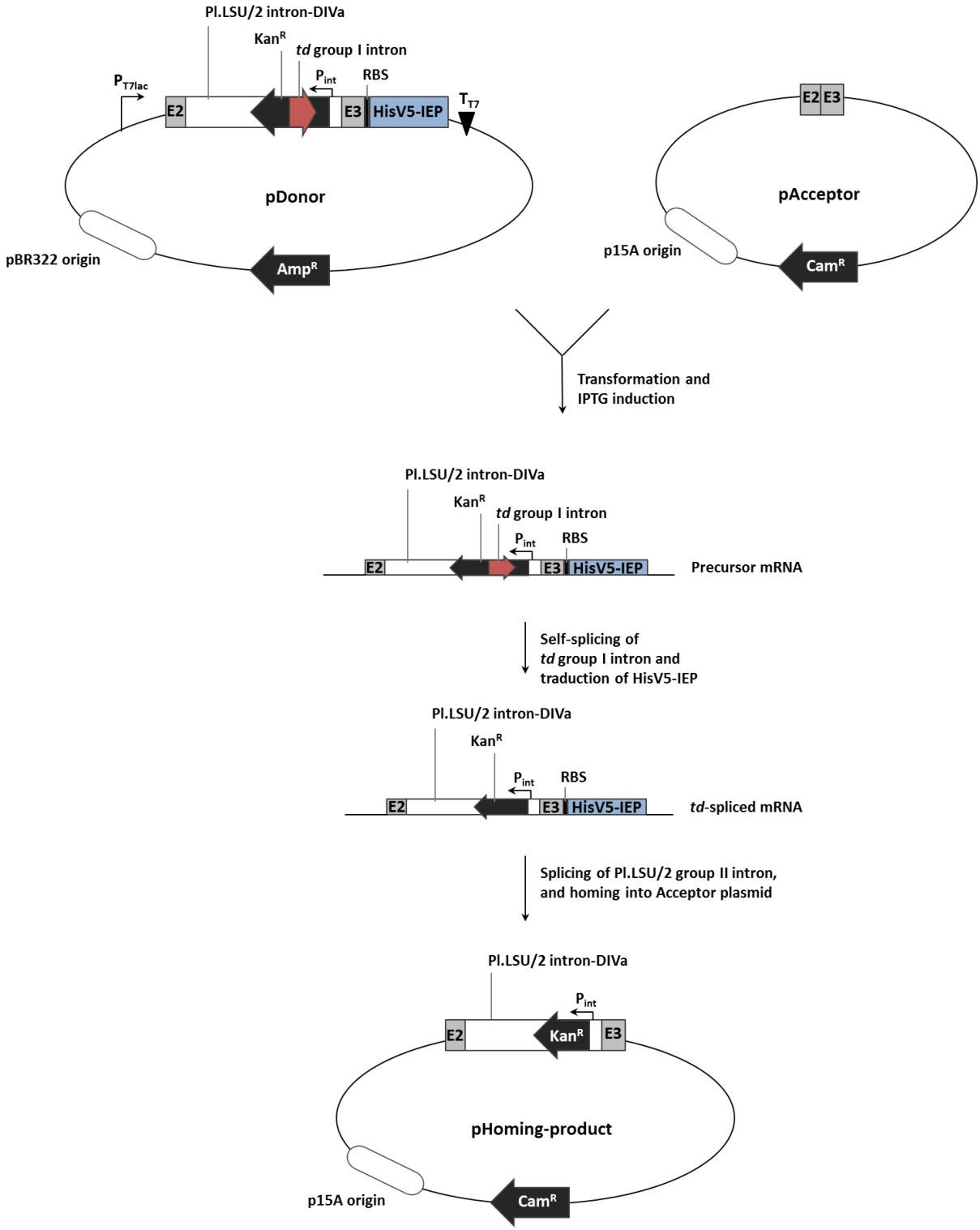


Figure R-35: RAM-targetron strategy in *E. coli*.

Pl.LSU/2 intron DIVa: deleted form of Pl.LSU/2 intron (See article 2; Fig. 1B); E2: 50 last nt of exon 2; E3: 71 first nt of exon 3; Kan^R: kanamycin resistance gene; P_{int}: internal kanamycin promoter; RBS: ribosome binding site; HisV5-IEP: Pl.LSU/2 IEP fused in N-terminal with 6xHis tag and V5 epitope; T_{T7}: T7 transcription terminator; AmpR: ampicillin resistance gene; pBR322 origin: pBR322 high-copy replication origin; CamR:

Chloramphenicol resistance gene; p15A origin: p15A low-copy replication origin. The strategy is described in the text.

This RAM-twintron strategy consists in the transformation of *E. coli* by two plasmids: a donor and an acceptor plasmid. The donor plasmid is a high-copy plasmid containing the Pl.LSU/2 group II intron with its flanking exons (E2 and E3) cloned downstream a T7lac promoter and upstream the HisV5-IEP encoding sequence. A ribosome binding site (RBS) is located just upstream of the HisV5-tag to allow its translation. The Pl.LSU/2 intron used in this assay contains a deletion of the domain IV (Pl.LSU/2 intron-DIVa; See article 2) which removes a part of the IEP ORF. The intron contains in its domain IV the kanamycin resistance gene used as retrohoming-activated marker (Kan^R) disrupted by the efficient self-splicing *td* group I intron. The Kan^R selectable marker is inserted in the reverse orientation into Pl.LSU/2 intron domain IV, and the self-splicing *td* group I intron is inserted in the forward orientation. The self-splicing *td* group I intron inserted into the Pl.LSU/2 intron forms a twintron.

The acceptor plasmid is a low-copy plasmid containing the potential natural DNA target site of Pl.LSU/2, corresponding to the last 50 nt of exon 2 (E2) and the 71 first nt of exon 3 (E3).

The donor and acceptor plasmids are co-transformed in *E. coli* BL21 Star (DE3) and the transcription of the donor cassette containing the RAM-twintron and the HisV5-IEP ORF is induced with IPTG. The self-splicing of intron *td* from this RNA intermediate restores the Kan^R ORF. The reverse orientation of the rescued Kan^R ORF on the *td*-spliced mRNA implies that the marker could be activated only upon Pl.LSU/2 intron integration into the acceptor plasmid via the homing mechanism (Fig R-35). Pl.LSU/2 homing events can thus be detected by the growth of kanamycin resistant *E. coli* colonies. The use of the twintron strategy allows to ensure that the homing mechanism rely on an RNA intermediate.

Five independent homing assays were conducted as described thereafter. *E. coli* BL21 Star (DE3) was cotransformed with pDonor and pAcceptor plasmids. A 130 ml culture was then incubated at 32°C to OD_{600nm} 0.4. The culture was then split in four equal parts which were either non-induced or induced for 30 min, 1 hr or 2 hrs with 2 mM of IPTG at 32°C. Two ml of each culture were then plated on LB-agar containing either chloramphenicol (Cam) or chloramphenicol and kanamycin (Cam-Kan). *E. coli* cells that grow under Cam selection contain the pAcceptor plasmid with or without integrated Pl.LSU/2 intron, while cells that grow under Cam-Kan selection necessarily contains the pAcceptor plasmid in which Pl.LSU/2 intron has been integrated. The homing frequency can thus be calculated by the ratio of the number of Cam^R-Kan^R colonies to the number of Cam^R colonies.

Unfortunately, no Cam^R-Kan^R colonies were found on 2.9×10^7 ($\pm$ 1.65×10^5) Cam^R colonies (data representing the mean of five independent experiments). Several hypotheses have been formulated to explain these results: the HisV5-IEP could not been expressed from the pDonor plasmid, the Kan^R marker could display an expression defect, or the Pl.LSU/2 intron could have not been integrated.

To ensure that these results were not the consequence of an absence of HisV5-IEP expression, proteins were extracted from 20 ml of the uninduced and induced cultures and HisV5-IEP expression was verified by western blot (Fig. R-36).

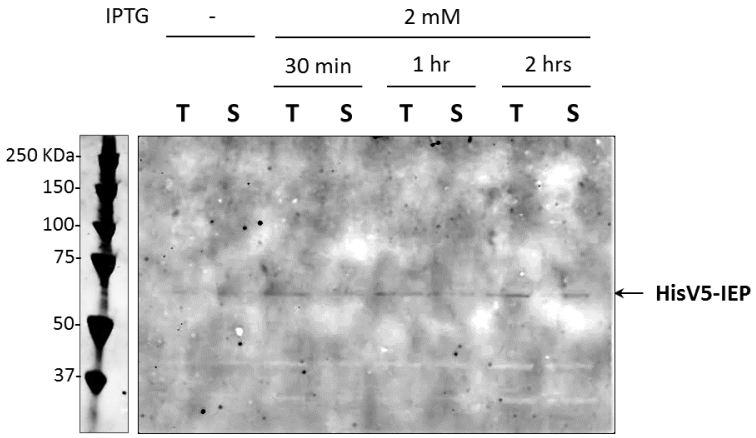


Figure R-36: Western blot analysis of protein expressed in *E. coli* during RAM-targetron homing assay.
1/70 of total (T) and soluble (S) protein fractions extracted from 20 ml cultures of *E. coli* transformed by both pDonor and pAcceptor plasmids and induced or not (-) at OD_{600nm} of 0.4 with 2 mM of IPTG for 30 min, 1 hr and 2 hrs. Protein fractions are loaded onto SDS-PAGE gel and HisV5-IEP is detected using an anti-V5 antibody.

The molecular mass of HisV5-IEP is expected to be around 69-kDa. Figure R-36 shows that HisV5-IEP is expressed at low levels in induced cultures and whatever the time of induction used (Fig. R-36; 2 mM IPTG, 30 min, 1 hr and 2 hrs, fractions T). The protein is also detectable in the soluble fractions (Fig. R-36; 2 mM IPTG, fractions S). These results indicate that the expression cassette from pDonor is transcribed and translated. However, the yield of HisV5-IEP expression remains low, even though the mRNA is transcribed from a high-copy plasmid. Nevertheless, the absence of Cam^R-Kan^R clones is probably not the consequence of a failure in HisV5-IEP expression.

To determine if the absence of Cam^R-Kan^R clones is due to a failure in the Kan^R marker expression in *E. coli* cells even upon Pl.LSU/2 homing, plasmid DNA were extracted from 10 ml of the uninduced and induced cultures and analyzed by restriction digestion using enzymes that give differential restriction profiles for pDonor, pAcceptor and pHoming product. SspI restriction enzyme was used to digest 1 µg of plasmid DNA. The detection of a 676 bp restriction fragment should conclude to the retrohoming of Pl.LSU/2 intron in the DNA target site of pAcceptor plasmid (Fig. R-37A; pHoming product, indicated in bold). Restriction fragments of each plasmid DNA sample were thus analyzed by electrophoresis on agarose gels (Fig. R-37B).

A

	pDonor	pAcceptor	pHoming product
SspI	5703 bp 2316 bp 1633 bp 252 bp	4838 bp 1727 bp	5814 bp 1727 bp **676 bp** 252 bp

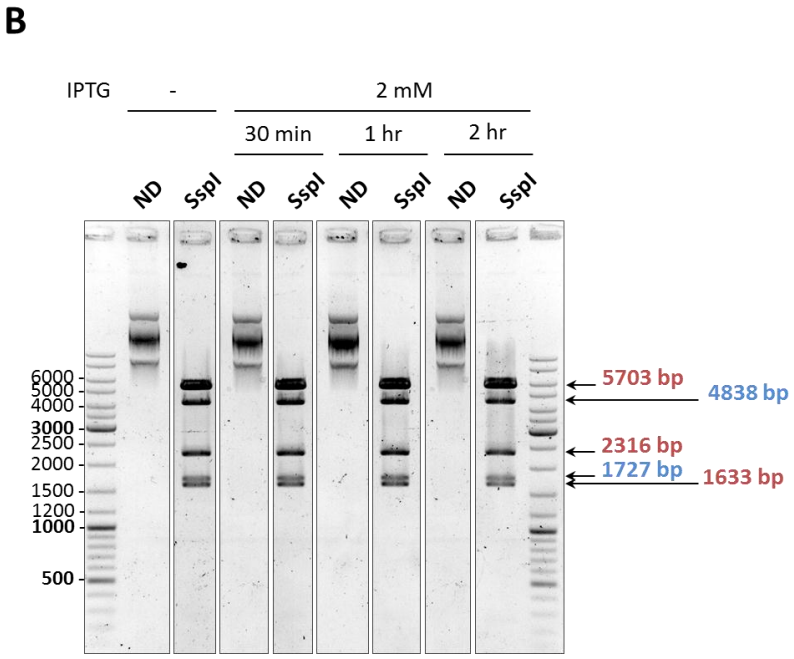


Figure R-37: Restriction digestions of plasmid DNA extracted during *E. coli* homing assay.

(***A***) Expected restriction fragments for pAcceptor, pDonor and pHoming after SspI digestion. Restriction fragment that could conclude to the presence of a plasmid with integrated Pl.LSU/2 intron (pHoming product) is indicated in bold. (***B***) SspI restriction profiles. The length of restriction products observed are indicated at *right*: pDonor corresponding restriction products are in red and pAcceptor corresponding restriction products are in blue. Molecular mass marker is indicated at *left*.

The analysis of plasmid DNA restriction profiles shows that restriction products of only pDonor and pAcceptor are detected (Fig. R-37B; indicated in red and blue, respectively). Restriction profiles are identical for each plasmid DNA extracted sample (Fig. R-37; - IPTG and 2 mM IPTG 30 min, 1 hr and 2 hrs). The 676 bp fragment from the pHoming-product plasmid was not detected by restriction digest analysis. We subsequently confirm this result by the use of several other enzymes (Data not shown). This suggests that the homing of Pl.LSU/2 intron into its DNA plasmid target has not occurred. It could also indicate that the homing frequency is too low to be detected by restriction digestion.

The homing of the Pl.LSU/2 intron could not been demonstrated in our *E. coli* assay. We did not perform further experiments principally because of a lack of time. In addition, another work that we conducted in parallel led us to evaluate the homing of Pl.LSU/2 intron in yeast.

4.3 - HOMING OF PL.LSU/2 IN *S. CEREVISIAE*

The study of the Pl.LSU/2 intron splicing in *S. cerevisiae* showed that Pl.LSU/2 intron can splice in yeast and that the splicing efficiency is increased when the IEP is expressed (See article 2). It could suggest that in yeast, the Pl.LSU/2 IEP and the intron lariat can form RNP particles, which are the catalytic molecules involved in the homing mechanism. Thus, we postulated that the Pl.LSU/2 intron homing could occur in yeast cells. The strategy used to evaluate the homing capacity of the Pl.LSU/2 intron in yeast was directly adapted from the yeast splicing assay. The two plasmids constructed for the yeast splicing assay (Fig. R-38, pEgpIIE-URA3 and pNLS-IEP^{co}) and an acceptor plasmid containing the potential natural DNA target site of Pl.LSU/2 (Fig. R-38; pE2E3) were used here.

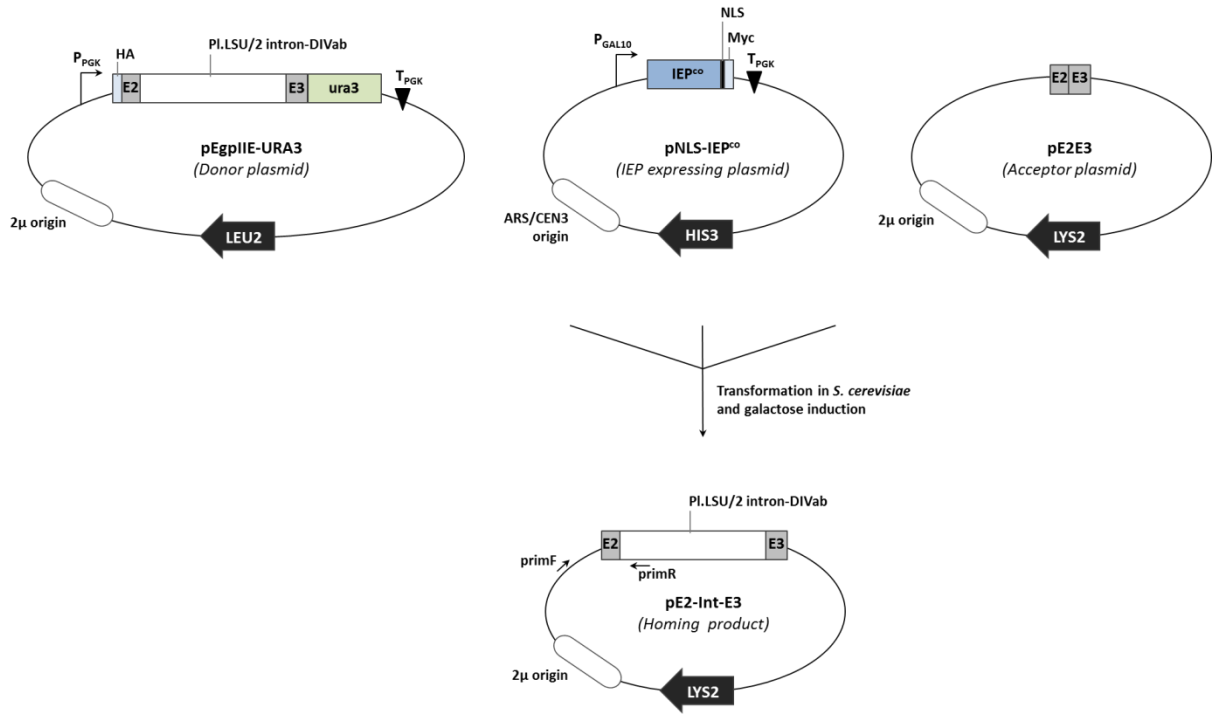


Figure R-38: Yeast homing assay strategy.

P_{PGK}: PGK promoter; HA: stretch of 2 HA tags; Pl.LSU/2 intron-DIVab form: deleted form of Pl.LSU/2 intron (See article 2, Fig 1B); E2: 50 last nt of exon 2; E3: 71 first nt of exon 3; URA3: Ura3p encoding sequence; T_{PGK}: PGK transcription terminator; LEU2: Leucine encoding selection gene; 2µ origin: 2µ high-copy replication origin; P_{GAL10}: galactose-inducible GAL10 promoter; IEP^{co}: IEP encoding sequence codon-optimized for translation in human cells; NLS: stretch of 3 nuclear localization signals; Myc: c-Myc epitope; T_{PGK}: PGK transcription terminator; HIS3: Histidine encoding selection gene; ARS/CEN3 origin: Autonomously Replicating Sequence/yeast centromere low-copy replication origin; LYS2: Lysine encoding selection gene; PrimF/R: oligonucleotides. The strategy is described in the text.

The strategy adopted here is not based on a retrohoming-activated marker as in the *E. coli* homing assay, so that the homing can only be detected by plasmid DNA analysis. The plasmid pEgpIIE-URA3 is used as the Pl.LSU/2 intron donor plasmid, and pNLS-IEP^{co}, allows the inducible Pl.LSU/2 IEP expression. The plasmid pE2E3 corresponds to the acceptor plasmid and contains the last 50 nt of exon 2 (Fig. R-38; E2) and the 71 first nt of exon 3 (Fig. R-38; E3), used here as the Pl.LSU/2 DNA target site. The expression of the Pl.LSU/2 IEP in yeast has previously been shown to promote the Pl.LSU/2 intron splicing from the precursor mRNA transcribed from the pEgpIIE-URA3 plasmid (See article 2), potentially leading to the formation of a ribonucleoparticle. This theoretically formed RNP could thus recognize its DNA target site located in the pE2E3 plasmid. The homing of Pl.LSU/2 intron should subsequently lead to the formation of the plasmid pE2-Int-E3 (Fig. R-38; homing product).

S. cerevisiae BY4742 strain was first co-transformed by the donor (pEgpIIE-URA3) and acceptor (pE2E3) plasmids. Double transformants were then selected and transformed or not by the IEP expressing plasmid (pNLS-IEP^{co}). As the Pl.LSU/2 IEP expression in *S. cerevisiae* BY4742 strain has already been demonstrated during the yeast splicing assay (See article 2), we have not verified the Pl.LSU/2 IEP expression in this homing assay. Yeasts were then cultured in either glucose- or galactose-containing medium in order to repress or induce NLS-IEP^{co} expression, respectively. After 22 hrs of culture, plasmid DNA was extracted from yeast cells and analyzed by PCR amplifications.

PCR amplification analysis was chosen instead of restriction digestion, as used previously in the *E. coli* homing assay, because of the higher sensitivity of the PCR technique. Amplification primers PrimF and PrimR (See Fig. R-38) were used to specifically amplify the acceptor plasmid in which the Pl.LSU/2 intron would have been inserted, leading to the synthesis of a 485 bp amplification product (Fig. R-39).

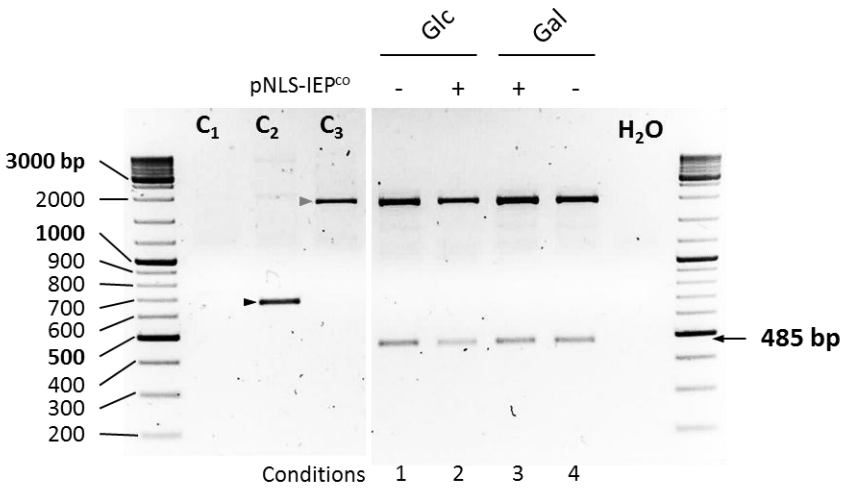


Figure R-39: Aragose electrophoresis of PCR amplifications of plasmid DNA extracted during yeast homing assay.

S. cerevisiae harboring the pEgpIIE-URA3 donor plasmid and the pE2E3 acceptor plasmid were either transformed (+) or not (-) by pNLS-IEP^{co} and cultivated for 22 hrs at 28°C in glucose- (Glc) or galactose- (Gal) containing medium. Plasmid DNA was extracted from 4 ml of these cultures. 60 ng of plasmid DNA was used to specifically amplify the homing product (pE2-Int-E3) by PCR with primF and primR primers (See Fig. R-38). Amplification reactions without plasmid DNA (H_2O), or with 20 ng of each pE2E3 (C_1), pNLS-IEP^{co} (C_2) and pEgpIIE-URA3 (C_3) plasmid were also performed as controls. Non-specific amplification products obtained with C_2 and C_3 controls are indicated by black and gray arrowheads, respectively. Molecular mass marker is indicated at *left*.

Figure R-39 shows that no DNA contamination during PCR amplification has occurred, as shown by the absence of amplifications product without plasmid DNA (Fig. R-39; H_2O). We can also observe that non-specific amplification products are detected in the controls: a product of about 650 bp is found using the pNLS-IEP^{co} plasmid (Fig. R-39; C_2, black arrowhead) and an approximately 2000 bp product is found using the donor pEgpIIE-URA3 plasmid (Fig. R-39; C_3, gray arrowhead). The 2000 bp band is also detected in every condition of the homing experiment. Surprisingly, we observe the 485 bp amplification product expected in presence of the homing product (pE2-Int-E3) in every conditions of the homing experiment (Fig. R-39; conditions 1 to 4, 485 bp product). This finding was not expected. Indeed, this indicates that the pE2-Int-E3 plasmid is present in yeast cells in which either the pNLS-IEP^{co} plasmid is absent (Fig. R-39; conditions 1 and 4) or the NLS-IEP^{co} expression is repressed (Fig. R-39; condition 2). In these cells, the homing process should not be permitted in absence of the IEP, as this mechanism theoretically involves the endonuclease and reverse transcriptase activities of the IEP.

To confirm that these 485 bp products correspond to an amplification of the pE2-Int-E3 plasmid, the 485 bp fragments were extracted from the gel and sequenced using the four oligonucleotides

represented in Figure R-40A. Sequences obtained were subsequently aligned with the expected pE2-Int-E3 plasmid sequence (Fig. R-40B).

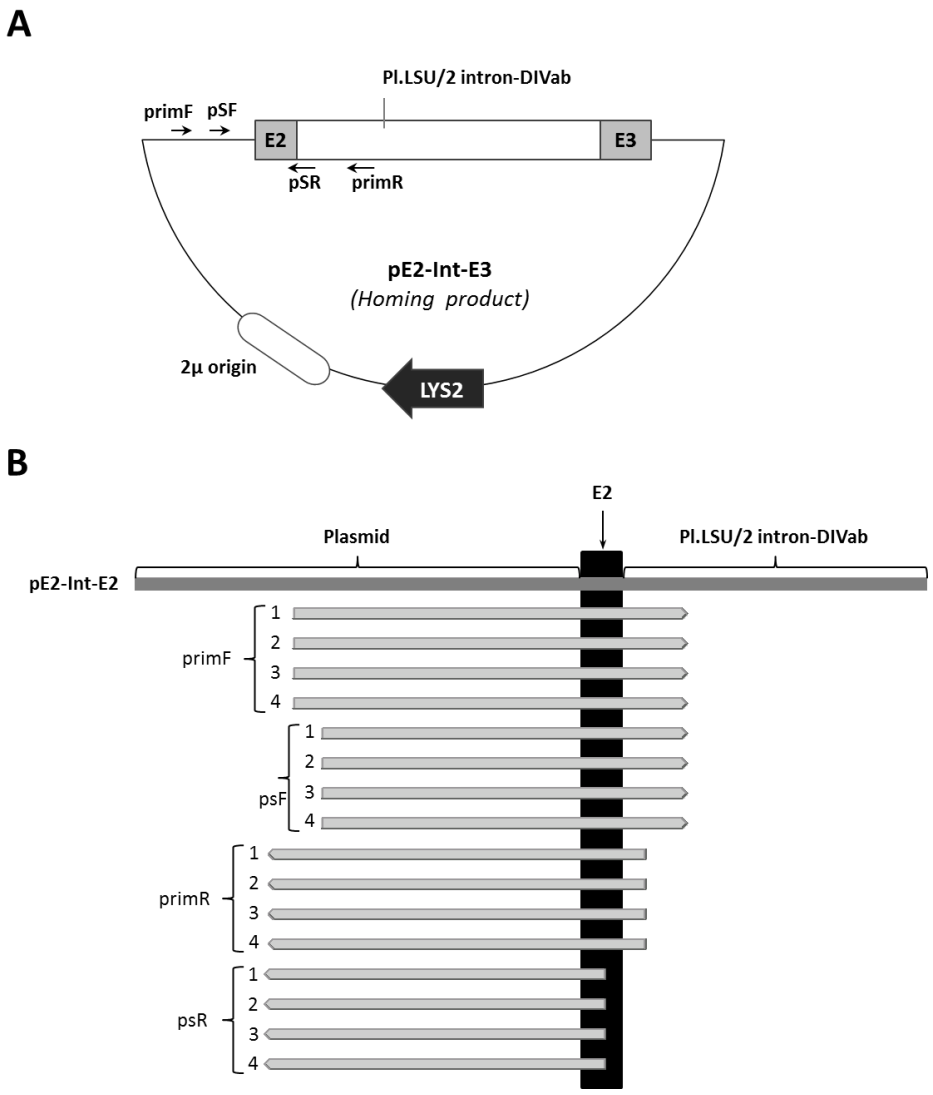


Figure R-40: Sequencing analysis of the 485 bp amplification products obtained in yeast homing assay.
(***A***) Schematic representation of the pE2-Int-E3 plasmid, expected upon the Pl.LSU/2 intron integration into the E2E3 target site in pE2E3 acceptor plasmid, with the sequencing primers used (primF, pSF, pSR and primR). (***B***) Representation of the alignment of the expected pE2-Int-E3 plasmid sequence (dark gray rectangle) and the 485 bp products sequences obtained with primers primF, pSF, pSR and primR (light gray rectangle). The exon 2 sequence (E2) is indicated by a black rectangle. Acceptor plasmid and Pl.LSU/2 intron-DIVab sequences are indicated with brackets.

Figure R-40B shows that sequences obtained for all 485 bp amplification products are correctly aligned with the expected pE2-Int-E3 sequence. These results confirm that the plasmid pE2-Int-E3, resulting from the Pl.LSU/2 intron integration into the E2-E3 DNA target site at the junction of the two exons, is present in all conditions tested in the yeast homing assay.

The detection of Pl.LSU/2 intron integration into the acceptor plasmid in yeast that either does not harbor the pNLS-IEP^{co} expressing plasmid or in which the NLS-IEP^{co} expression is repressed could have occur by homologous recombination, which is an efficient process in *S. cerevisiae*. Indeed, the

acceptor pE2E3 and the donor pEgpIIE-URA3 plasmids shared homologous DNA sequences. They both contain the last 50 nt of E2 and the 71 nt of E3. Only one nucleotide mispairing can occur, as the 16th nucleotide of E3 (adenine) is replaced by a thymine in pEgpIIE-URA3. The Pl.LSU/2 intron located in the pEgpIIE-URA3 donor plasmid is thus flanked by homology sequences of at least 50 bp, and this length of homology is sufficient to achieve efficient homologous recombination in *S. cerevisiae* (Manivasakam P et al. 1995). The strategy adopted here does not permit the distinction between Pl.LSU/2 intron retrohoming and homologous recombination events involving the donor and acceptor plasmids. Another strategy seems to be required to evaluate the Pl.LSU/2 homing in yeast.

GENERAL DISCUSSION AND FUTURE PERSPECTIVES

The main goal of this thesis was to characterize the group II intron Pl.LSU/2 from the perspective of its use as a site-specific gene therapy vector.

Integrative vectors currently used in clinical trials can induce insertional mutagenesis. The evaluation and optimization of the safety of integrative vectors such as retroviral vectors has been highlighted by adverse events encountered in different clinical protocols. During this thesis, I had the opportunity to collaborate on such a study conducted with the aim of evaluating the safety of a lentiviral vector used in a gene therapy clinical trial for the treatment of the hematopoietic WAS disorder. The results of this study are described in article 1. As discussed in this article, it seems required to combine analyses of vector insertion sites and vector copy number per corrected cell in order to assess the biological potency and the risk of genotoxicity of integrating vectors.

A solution to further enhance the safety of clinical protocols would be the use of gene targeting systems. Several strategies based on the use of site-specific nucleases are currently developed to enable gene targeting. All these strategies start by generating a double-stand break of the host DNA at a specific site, which has to be repaired by homologous recombination with a provided template. Those approaches present an inherent mutagenesis risk due to off-target DSB or to the possibility of introducing some errors during the repair mechanism. The development of alternative strategies that would not be based on DSB would thus be of a great interest. In this context, we were interested in natural mobile group II introns that can transpose to a specific DNA target site by the homing mecnanism, and evaluated the possibility of the use of Pl.LSU/2 group II intron as gene targeting vector. For this purpose, we have characterized the biochemical activities of its encoded protein and evaluated the intron catalytic activities *in vivo*.

A major part of this work consisted in the development of expression and purification strategies for the Pl.LSU/2 IEP biochemical characterization. The expression of soluble IEP was optimized and the use of classical methods relying on affinity chromatography allowed the protein to be partially purified. However, these purification methods were not conclusive as active IEP for reverse transcription could not be obtained. All these results are discussed in the following section, and further optimizations are proposed.

The use of the Pl.LSU/2 intron as targeting vector assumes that the intron homing should occur. As the *in vivo* intron splicing is a prerequisite to its homing into DNA targets, the first study that we conducted was the characterization of the Pl.LSU/2 *in vivo* splicing. I designed a strategy based on a Pl.LSU/2 splicing-dependent Ura3p complementation in *S. cerevisiae* that should allow a direct determination of the Pl.LSU/2 splicing in yeast. I showed that the intron could splice in yeast and that the splicing efficiency was improved by the maturase activity of the IEP. However, translated proteins from spliced transcripts were not detected, and splicing of Pl.LSU/2 could not been demonstrated in human cells. These results are discussed in article 2, and further perspectives are proposed thereafter.

Finally, we have attempted to evaluate the homing of Pl.LSU/2 in *E. coli* and *S. cerevisiae*. Two different strategies were used for this purpose. However, the Pl.LSU/2 homing could not been demonstrated in any of both hosts. These results and future perspectives are discussed hereafter.

1 - EXPRESSION AND PURIFICATION OF THE Pl.LSU/2 IEP

The first step in the characterization of the Pl.LSU/2 group II intron was to assess the biochemical activities of its IEP. At the beginning of this work, no information was available on the potential activities of the putative Pl.LSU/2 IEP. The fact that the open-reading frame contained in Pl.LSU/2 intron presented all conserved domains of intron-encoded proteins (See Fig. I-24) suggested that the functionality of this protein could have been preserved by evolution. In order to study the activities of the Pl.LSU/2 IEP, the protein must be first expressed and purified. We chose to first test for its reverse transcriptase activity, as a simple and fast method was already published and used to demonstrate the reverse transcriptase activity of the Ll.LtrB IEP.

1.1 - CHOICE OF THE EXPRESSION HOST

The Pl.LSU/2 IEP was expressed as a fusion protein using different tags to allow its subsequent purification. I tested several expression systems such as *E. coli*, Sf9/baculovirus, and cell-free, and we showed that the expression in *E. coli* has to be optimized to obtain sufficient amount of soluble IEP. In this purpose, I used tRNA complementation, several *E. coli* strains, I evaluated the influence of the growth temperature and IPTG concentration for the induction of IEP expression, and I tested different methods of lysis (not shown). Although it was observed that a non-negligible amount of IEP was always found to remain insoluble, expression conditions could be optimized enough to obtain detectable amount of soluble protein. However, it would be interesting to test the expression of the IEP in other hosts such as *S. cerevisiae* or mammalian cells. Indeed, *E. coli* does not support post-translational modifications, and the Pl.LSU/2 IEP could be subjected to such modifications potentially affecting its overall folding and/or stability. If so, the expression of IEP in a host enable to perform post-translational modifications could improve the purification of active IEP. However, the IEP expression in insect cells showed that the IEP was mainly expressed in an insoluble form (See Fig. R-16). It is worth noting that the promoter driving the transcription of the IEP in this experiment is considered as a strong promoter, so that fine regulation of the transcription rate, and more generally the protein expression rate, could not been achieved with this system. In contrast, several conditional expression systems can be used in yeast or in mammalian cells, such as the Tet-OFF/ON system, or inducible promoters as Gal1 in yeast. However, expression experiments in *E. coli* have showed that the main factor influencing the expression rate and solubility of the IEP was the growth temperature (and by extension the growth rate) of cells before and during induction of IEP expression (See Fig R-6). In this context, *S. cerevisiae* could be a good candidate for the expression of soluble IEP, as the growth rate of yeast is easily adjustable by changing the incubation temperature.

1.2 - PURIFICATION OF TAGGED IEP

According to the results obtained for the expression of soluble IEP, I attempted to purify active GST- or His-tagged IEP expressed in *E. coli*. However, both attempts were unsuccessful. Although partially purified protein fractions could be obtained for both GST-IEP (See Fig. R-9) and HisV5-IEP (See Fig. R-20, R-24, and R-26), the results of RT assays using these purified protein fractions were not conclusive.

1.2.1 - Contamination of GST-IEP purified fractions

In the case of the GST-IEP purification, both wild-type GST-IEP and mutants GST-IEP IEP (mtDD- and ΔRT5) fractions showed RT activity (See Fig. R-12). Both mutants are expected to be RT-defective so that the activity highlighted with those latter fractions necessarily relies on a contaminating protein co-eluted during the purification. We speculated that a contaminant *E. coli* reverse transcriptase, potentially corresponding to a retron reverse transcriptase (See Fig. R-13) was co-eluted with the GST-tagged, biasing RT reactions. To confirm this hypothesis, it would be interesting to determine if this retron reverse transcriptase could be identified in those fractions, for instance by a mass spectrometry analysis.

To avoid those contaminations, an additional purification step could have been tested, such as gel filtration. However, the contaminant protein was shown to be co-eluted specifically in presence of the IEP, as no RT activity was detected using GST purified fraction (See Fig. R-12). This result could be explained by the presence of a binding (direct or indirect) between the contaminant protein and the IEP, which therefore render useless this additional step. One could speculate that the IEP and the *E. coli* contaminant, being both RNA binding proteins, could be co-eluted through indirect binding to nucleic acids. The precipitation of nucleic acids, for instance using polyethyleneimine, as an initial step prior to purification could release IEP from nucleic acids and limit contaminations with *E. coli* proteins that interact with RNAs.

Those experiments demonstrating the co-purification of a contaminant reverse transcriptase protein highlighted the necessity of identifying this protein. Indeed, the presence of an active reverse transcriptase in *E. coli* should be carefully tested as it could bias biochemical studies and induce false positive results when appropriate controls are missing.

1.2.2 - Purification of HisV5-IEP

With regards to the results obtained with GST-IEP purification, we thought that the use of a different tag could overcome the contamination encountered. The 6xHis and V5 tags were chosen because of their small length, potentially inducing a different tridimensional conformation of the fusion protein (See Fig. R-17) with regards to GST-IEP conformation. In contrast to the results obtained with the purification of GST-IEP by affinity chromatography, we showed that no RT activity was detected using HisV5-IEP purified fractions obtained by IMAC. Several purification conditions were tested but none allowed obtaining active HisV5-IEP (See Fig. R-22, R-25, and R-27). It is likely that those purification fractions are free of the *E. coli* contaminant found in GST-IEP purified fractions and this could be due to a different tridimensional conformation between GST-IEP and HisV5-IEP. However, we can also speculate that the conditions used during IMAC purification are deleterious to the *E. coli* contaminant, impeding it to retain its RT activity. It could be interesting to analyze the HisV5-IEP purified fraction by a mass spectrometry analysis to determine the presence or absence of a contaminating RT protein.

The demonstration of the RT activity of the HisV5-IEP purified by ultracentrifugation in a sucrose cushion (See Fig. 2 and 3, article 2) revealed that the absence of HisV5-IEP activity when purified by

IMAC is not due to the presence of the HisV5 tag, but to the experimental conditions used during the IMAC process. It would have been interesting to purify a control reverse transcriptase, such as the SuperScript II reverse transcriptase, under the same conditions and assay its RT activity to evaluate the potential deleterious impacts of the process on a different protein. Several factors are determinant to successfully purify active proteins, such as the temperature during the process, the length of the process, the composition of all buffers used, the presence of contaminant proteins, etc. We showed that experimental conditions used during IMAC were not appropriate to purify active IEP, in contrast to the ones used during ultracentrifugation in sucrose. Other methods could thus be tested in order to purify active IEP.

1.2.3 - Alternative methods of purification

Among existing purification methods, ion exchange chromatography or affinity chromatography on heparin columns can be considered for IEP purification.

Ion-exchange chromatography could be an efficient method to purify IEP expressed in *E. coli*. Indeed, the isoelectric point (pI) of the Pl.LSU/2 IEP is predicted to be around 10.5 ((EMBOSS iep application, (Rice P et al. 2000)), which could be an advantage for purification from background *E. coli* proteins. Most *E. coli* proteins have a theoretical pI around 6 (Fig. D-1) and therefore will not be positively charged at pH 8.75. At this pH, the Pl.LSU/2 IEP should bind a cation exchange column while the majority of *E. coli* proteins would come straight through or eluate at low NaCl concentrations.

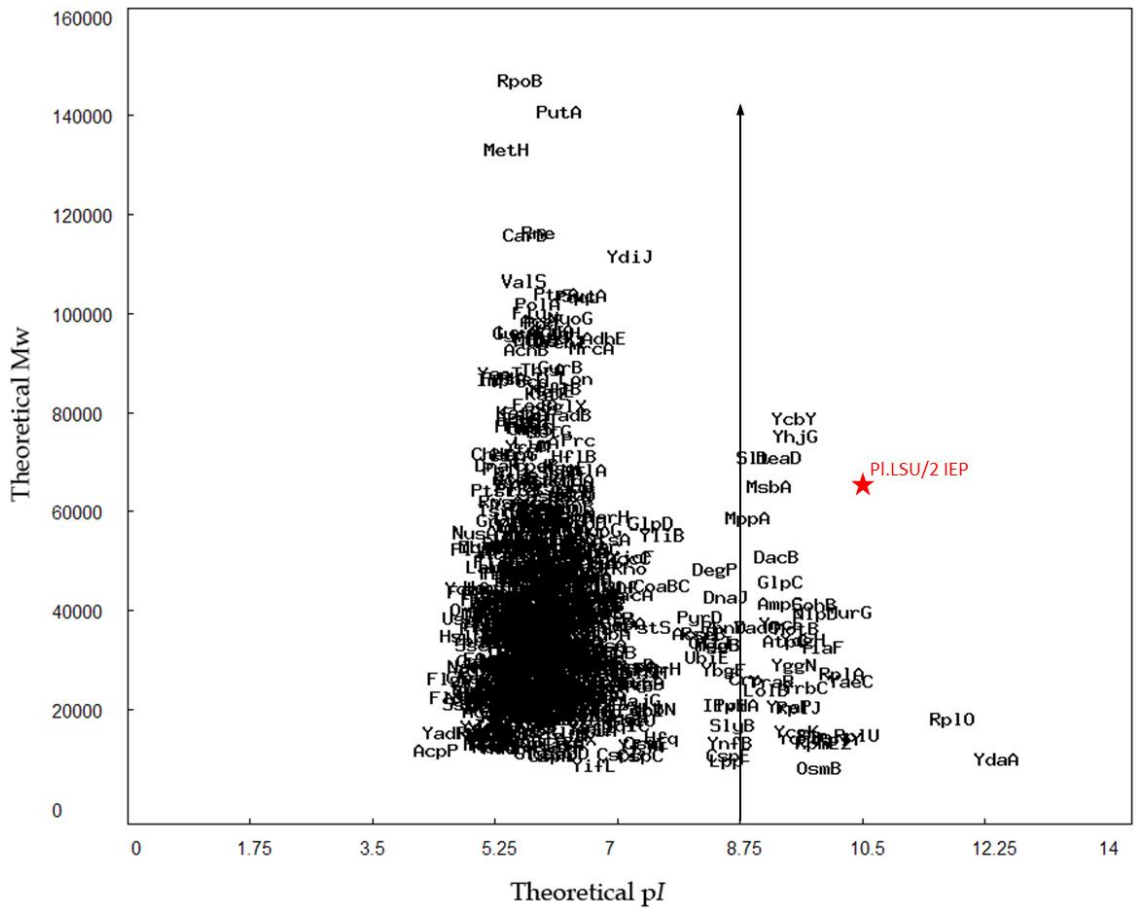


Figure D-1: Scatter graph plotting theoretical molecular weight against theoretical pI of *E. coli* proteins and Pl.LSU/2 IEP.

The majority of *E. coli* proteins have a theoretical pI value around 6.0. Pl.LSU/2 IEP (65 kDa) is depicted by a red star, and has a theoretical pI value of 10.5. Black arrow indicates the 8.75 threshold. Mw: molecular weight (in Daltons). ((EcoProDB online protein database, (Yun H et al. 2007)).

An anion-exchange column can also be used. Due to its high pI value, the IEP should not bind to this column at pH 8.75 and thus be present in the flow-through, in contrast to the majority of the *E. coli* proteins and negatively-charged nucleic acids. This method could thus serve as a purification step whereby nucleic acids are removed along with many contaminating proteins. However, it is also possible that the binding of the IEP to nucleic acids alters the overall charge of the protein, thus impeding its purification through those processes. In this case, a release of the IEP from nucleic acids would thus be required prior to downstream ion-exchange purification.

The binding of the Pl.LSU/2 IEP to nucleic acids could also be used as an advantage for its purification. Heparin columns could make use of this property: the IEP, along with other nucleic acid binding proteins in the protein lysate should bind heparin column, while other proteins will pass through. Elution could subsequently be carried out with a NaCl gradient in an attempt to separate the IEP from other nucleic acid binding proteins. However, if the IEP has a higher affinity to the nucleic acids in the protein lysate over the affinity to the heparin column, this could prevent an efficient binding. In this case, nucleic acid removal might be required as an initial step before purification.

1.2.4 - Pl.LSU/2 IEP activity: nucleic acid binding required for stability?

The finding that IEP displays an RT activity when purified by ultracentrifugation in sucrose cushion but not when affinity chromatography is used suggested that the IEP catalytic conformation was not achieved during this latter purification process. This could be due to several factors such as non-adapted purification conditions that impede the correct folding and/or stability of the IEP and/or to the absence of the Pl.LSU/2 intron RNA as "chaperone". Indeed, a previous study showed that the *Lactococcus lactis* Ll.LtrB group II intron-encoded protein LtrA was fully active when Ll.LtrB intron RNA was co-expressed, as a result of stabilization of the protein in its active conformation (Matsuura M et al. 1997).

According to this latter hypothesis, we coexpressed the HisV5-IEP with the Pl.LSU/2 intron in *E. coli* and purified the RNP particles theoretically formed in bacteria using both IMAC and ultracentrifugation in a sucrose cushion. We showed that HisV5-IEP contained in RNP particles purified by IMAC does not display RT activity (See Fig. R-33) in contrast to HisV5-IEP contained in RNP particles purified using ultracentrifugation in sucrose cushion (See Fig. 2, article 2). These results demonstrated the deleterious effects of the IMAC purification conditions on the activity of the IEP. These results led us to evaluate the RT activity of HisV5-IEP expressed alone in *E. coli* and purified by sucrose cushion ultracentrifugation. The purification experiment showed that nearly pure HisV5-IEP could be purified in those conditions (See Fig. 3, article 2). We also observed that this fraction contained non negligible amount of nucleic acids (data not shown). We can speculate that HisV5-IEP is able to bind nucleic acids of *E. coli* lysate, and that the complexes formed can be isolated by classical sucrose cushion centrifugation. It was subsequently showed that HisV5-IEP has the RT activity *in vitro*. It thus seems that this is not presence of the Pl.LSU/2 intron RNA that is required for the stability of HisV5-IEP, but that more generally nucleic acid can stabilize the HisV5-IEP activity.

Interestingly, RNP particles were shown to have RT activity only in presence of RNase A in the reaction medium (See Fig. R-34). This result indicates that the IEP has to be released from the endogenous RNA to access the exogenous RNA template used in RT reactions. In contrast, the Ll.LtrB IEP contained in RNP does not require RNase A to display a RT activity (Matsuura M et al. 1997). We can speculate that the Pl.LSU/2 IEP has a higher affinity for its intron RNA than Ll.LtrB IEP for its intron RNA. It could be interesting to perform similar experiment on the purified fraction containing HisV5-IEP alone to determine if the protein is able to display RT activity in absence of RNase A. Moreover, nucleic acids in HisV5-IEP purified fraction obtained by sucrose cushion ultracentrifugation could be removed to determine if HisV5-IEP is able to retain RT activity when formulated in nucleic acid-free storage buffer. It is possible that removal of nucleic acids may be deleterious to the stability of the IEP, as nucleic acids (and more probably RNA) seem to act as "chaperone" on HisV5-IEP.

It was previously shown that the Ll.LtrB IEP has RT activity only when copurified with nucleic acids, which could be either Ll.LtrB intron RNA coexpressed with the IEP, or artificial poly(rA)-oligo(dT) supplemented during the purification process (Matsuura M et al. 1997; Saldanha R et al. 1999; San Filippo J and Lambowitz AM 2002). It is likely that this IEP needs high amount of RNA in

its environment to be correctly folded and retain biochemical activity. In contrast to Ll.LtrB IEP, the purification of active Pl.LSU/2 IEP does not require additional RNA to be added during the process. This could potentially support the hypothesis that Pl.LSU/2 IEP has higher affinity to RNA than Ll.LtrB IEP. To test this hypothesis, biochemical analyses on the interaction between RNA and Pl.LSU/2 IEP or Ll.LtrB IEP could be achieved, as well as a study of the interaction kinetic.

To conclude, we demonstrated the RT activity of the *Pylaiella littoralis* Pl.LSU/2 IEP in RNP particles, as it was shown for the *S. cerevisiae* aI1 and aI2 (Kennell JC et al. 1993), the *L. lactis* Ll.LtrB (Matsuura M et al. 1997) and the *S. meliloti* RmInt1 (Martínez-Abarca F et al. 1999) IEPs. Moreover, we demonstrated the RT activity of the Pl.LSU/2 IEP expressed without its intron RNA, as for the *G. stearothermophilus* G.st.I1 IEP (Vellore J et al. 2004), and without the need of supplemental RNA during the purification, in contrast to Ll.LtrB IEP (Saldanha R et al. 1999).

All those results constitute the first functional characterization of the Pl.LSU/2 IEP.

2 - SPLICING OF THE Pl.LSU/2 INTRON

2.1 - SPLICING IN YEAST

To evaluate the splicing of the Pl.LSU/2 intron and the maturase activity of the IEP *in vivo*, we used a strategy based on a Pl.LSU/2 splicing-activated selectable marker in *S. cerevisiae* (See Fig. 4, article 2). A plasmid has been designed to contain the Pl.LSU/2 intron flanked by its two exons and located between an HA tag and the sequence encoding the Ura3p protein. The strategy relies on the expression of the fusion protein, composed of the flanking exons of Pl.LSU/2 intron and the Ura3p protein, which can occur only upon precise Pl.LSU/2 intron splicing. A mutant strain of *S. cerevisiae* in which the URA3 gene is defective has been used, and the idea is that the splicing of the Pl.LSU/2 intron should allow the growth of *S. cerevisiae* on selective medium lacking uracil. To evaluate the maturase activity of the Pl.LSU/2 IEP, the protein should be conditionally expressed. For this purpose, the intron domain IV, containing the IEP ORF, was partially deleted, and the IEP encoding sequence was cloned on a separate plasmid downstream of an inducible promoter. According to our aim to use this intron in genomic targeting, the IEP was here addressed to the yeast nucleus by using nuclear localization signals.

In this splicing assay, we demonstrated the Pl.LSU/2 intron splicing *in vivo* in yeast (See Fig. 5, article 2). The IEP was shown to promote the splicing of Pl.LSU/2 intron in *S. cerevisiae*. In spite of evidence for splicing properties of Pl.LSU/2 intron, the system that we designed to directly detect the splicing using an Ura3 complementation system was unsuccessful (See Fig. 5, article 2). Indeed, the fusion protein theoretically translated from the spliced mRNA was not detected by western blot, explaining the failure of the Ura3p complementation (See, Fig. 5, article 2). Similar results were also obtained in a previous study using the Ll.LtrB intron with another reporter gene (Chalamcharla VR et al. 2010). The authors proposed a model in which transcripts bearing the group II intron are subjected to nonsense-mediated decay that lowers the amount of available transcripts for splicing and translation. They also speculated that translation of spliced transcripts could be blocked by a binding of the intron lariat to the spliced transcripts via EBS/IBS interactions. In their study, it appears that the Ll.LtrB splicing is predominantly cytoplasmic, although the possibility that some group II intron splicing could occur in the nucleus could not be eliminated. It is thus intriguing that they did not show any improvement of the Ll.LtrB intron splicing when the IEP was not addressed to the nucleus. The more probable hypothesis that could explain the absence of detection of translated protein from spliced mRNA is that intron splicing efficiency could be too low to generate sufficient amount of spliced mRNA, thus keeping translated protein yield too low to be detected by western blot.

Nevertheless, it was clearly demonstrated that the Pl.LSU/2 intron could splice in yeast, and that this splicing was improved by the maturase activity of the IEP probably by enhancing the proper folding of the intron into its catalytically active structure. The IEP thus seems to be able to fold correctly in yeast and bind the intron RNA.

2.2 - SPLICING IN HUMAN CELLS

According to the results obtained in yeast, the strategy used to assay the intron splicing in human HCT 116 cells was not based on a splicing-activated selectable marker. The splicing of Pl.LSU/2 intron was here evaluated by RNA analysis. Again, to evaluate the IEP maturase activity, the IEP was expressed separately and the Pl.LSU/2 intron domain IV was partially deleted or not. Different sizes of domain IV deletion were used in order to evaluate the requirement in domain IV sequences in the IEP-mediated Pl.LSU/2 intron splicing (See Fig. 1, article 2) to determine if some regions of the domain IV participate in the binding to the IEP, as it was shown for the Ll.LtrB group II intron. To ensure a correct expression level of the different forms of the intron in the cells, we used intron-expressing lentiviral vectors to establish stable cell lines. IEP-expressing lentiviral vectors, encoding either the fusion GFP-IEP or the IEP both addressed to the nucleus, were subsequently used to transduce intron-expressing stable cell lines. The analysis of RNA by quantitative RT-PCR revealed that all precursor mRNA were expressed in these cells, however we did not detected any trace of spliced mRNA, even upon IEP expression. This absence of detectable spliced mRNA could be explained by a failure of the protein to fold correctly in human cells, or by differential IEP and intron RNA nuclear localizations that impede their binding. Precise evaluation of IEP and intron RNA nuclear localizations should be conducted to address this question. It is likely that the intron splicing efficiency in human cells is lower than in yeast cells, impeding the detection of spliced mRNA by quantitative RT-PCR analysis. It would be interesting to determine if an analysis of RT-PCR followed by an additional nested PCR could enhance the sensitivity sufficiently to allow the detection of spliced mRNA. Indeed, this technique of nested PCR was already used in a previous study to show Ll.LtrB homing on a plasmid target in human cells (Guo H et al. 2000).

2.3 - ENHANCING THE EFFICIENCY OF PL.LSU/2 INTRON SPLICING IN VIVO ?

In spite of a probably low efficient splicing in yeast, it is worth noting that residual spliced mRNA is observed in absence of IEP expression (See Fig. 5B, article 2). In our case, it is difficult to fully establish the intron ability to self-splice in yeast in the absence of IEP. Indeed, this intron is particularly active *in vitro* even at very low Mg^{2+} concentrations (Costa M et al. 1997b). It is therefore possible that spliced mRNA detected by RT-PCR and RT-qPCR in RNA extracts from yeast cells that were either not transformed by the IEP-expressing plasmid or in which the IEP expression was repressed could have been generated by an intron splicing *in vitro* after RNA extraction. However, the fact that no spliced mRNA was observed in RNA extracts from human HTC 116 cells expressing or not the IEP using very similar experiment conditions suggests that the residual splicing observed in yeast in absence of the IEP has likely occurred into the cells. This result is in fact consistent with the highly catalytic activity of the Pl.LSU/2 intron *in vitro* under stringent conditions. It could thus indicate that the Pl.LSU/2 intron is able to fold into its catalytically active conformation in an eukaryotic cell even in absence of its encoded protein. This feature could thus be of a great interest if the efficiency of the splicing reaction could be improved.

We designed the strategy based on a splicing-activated selectable marker to easily detect Pl.LSU/2 intron splicing in yeast with the attempt to perform selection of more efficient ribozyme mutants using

random mutagenesis (Fig. D-2). The strategy was based on error-prone PCR amplification of the Pl.LSU/2 intron from the pEgpIIE-URA3 plasmid used in the yeast splicing assay using the GeneMorph II random mutagenesis kit (Agilent Technologies). The error rate of this amplification step could be adjusted, and a library of mutated introns should be obtained. The transformation of yeasts by linear pEgpIE-URA3 plasmid (Fig. D-2; NcoI-digested pEgpIIE-URA3) along with PCR products lead to the insertion of mutated intron sequence into the linear plasmid by homologous recombination.

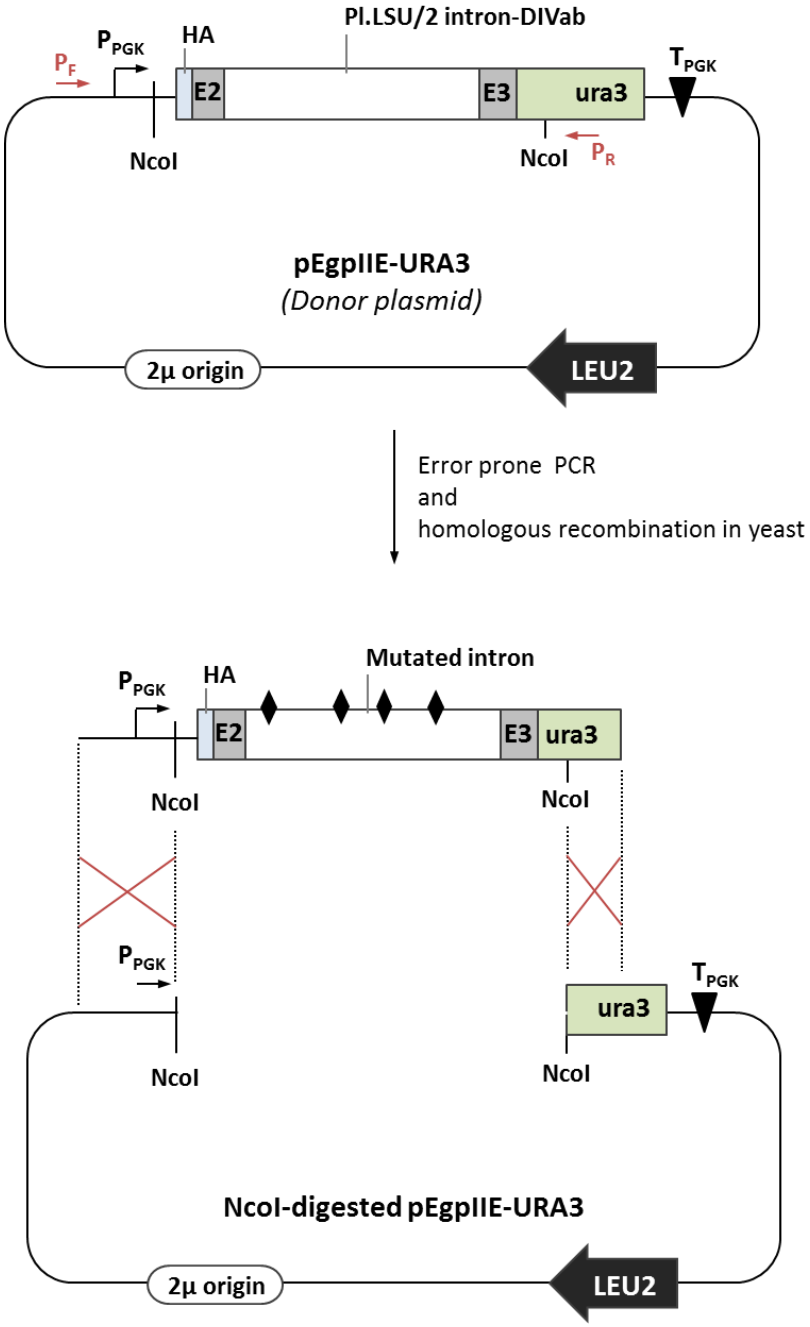


Figure D-2: Random mutagenesis strategy for selection of efficient Pl.LSU/2 ribozyme in yeast.

The plasmid pEgpIIE-URA3, called donor plasmid, is used in the yeast splicing assay. The error-prone PCR amplification is performed using primers Pf and Pr (red arrows) and introduces some mutations during DNA synthesis (black diamonds). The mutated intron is then integrated by homologous recombination (represented by

red crosses). P_{PGK}: PGK promoter; NcoI: NcoI restriction site; HA: stretch of 2 HA tags; Pl.LSU/2 intron-DIVab form: deleted form of Pl.LSU/2 intron (See Fig. 1B, article 2); E2: 50 last nt of exon 2; E3: 71 first nt of exon 3; URA3: Ura3p encoding sequence; T_{PGK}: PGK transcription terminator; LEU2: Leucine encoding selection gene; 2µ origin: 2µ high-copy replication origin.

Some preliminary experiments were conducted during this thesis (data not shown), unfortunately without success. However, as it is possible that spliced mRNA is subjected to a translation blockade, it seems required to design a new system to select mutant forms of the Pl.LSU/2 intron by their ability of efficient splicing *in vivo*. A high throughput analysis of RNA by quantitative RT-PCR could be considered. Recently, such a method has been developed and consist in a genome-wide reverse genetic screen in *S. cerevisiae* that couples robotic RNA isolation and cDNA synthesis, followed by quantitative PCR (Albulescu LO et al. 2012). Alternatively, the Pl.LSU/2 IEP could also be randomly mutated with the aim of select highly catalytic IEP.

To conclude, the splicing of Pl.LSU/2 intron has been demonstrated in *S. cerevisiae* and the Pl.LSU/2 IEP was found to display a maturase activity by promoting the intron splicing. However, we failed to detect the fusion protein theoretically translated from the spliced mRNA by western blot, so that Ura3p complementation could not been achieved. In contrast, the Pl.LSU/2 intron splicing could not be demonstrated in human cells. Although further experiments and optimizations are required to allow efficient splicing of the Pl.LSU/2 in eukaryotic cells, this study brings the first proof of the ribozyme activity of the intron and the maturase activity of the IEP *in vivo*.

3 - HOMING OF THE Pl.LSU/2 INTRON

3.1 - HOMING IN *E. COLI*

To further characterize the Pl.LSU/2 intron, we first evaluated its capacity to transpose into its natural DNA target by retrohoming in *E. coli*. For this purpose, we developed a strategy based on a retrohoming-activated selectable marker adapted from a commercial kit based on Ll.LtrB intron homing (See Fig. R-35). A donor plasmid was constructed and contained the Pl.LSU/2 intron flanked by its two exons and placed just upstream of the HisV5-IEP encoding sequence. The IEP encoding sequence located in the intron domain IV was partially deleted and the kanamycin resistance gene (Kan^R) was cloned in replacement in the opposite orientation with regards to the Pl.LSU/2 intron orientation. The Kan^R gene is disrupted by the self-splicing *td* group I intron cloned in the same orientation than those of the Pl.LSU/2 intron. The splicing of the *td* intron from the precursor RNA restitutes the frame of the Kan^R gene, which can be expressed upon Pl.LSU/2 intron transposition into DNA. The donor plasmid was co-transformed in *E. coli* with an acceptor plasmid containing the natural E2-E3 DNA target site of Pl.LSU/2 intron homing. In this way, homing events can subsequently be detected by the presence of bacterial clones on selective medium containing kanamycin.

In spite of HisV5-IEP expression in *E. coli* (See Fig. R-36), we did not observe any clones on the kanamycin selective medium. Plasmidic DNA were extracted from *E. coli* culture before kanamycin selection and analyzed by restriction digestion (See Fig. R-37). Again, we could not detect the presence of homing events using this technique.

The absence of clones on kanamycin selective medium suggests that either the Pl.LSU/2 intron does not transpose into *E. coli*, or the homing efficiency is too low to be detectable. A PCR analysis of plasmid DNA extracted from *E. coli* may enhance the retrohoming detection. Indeed, this amplification technique should be more sensitive than the analysis of plasmid DNA by restriction digestion. To further enhance the homing detection, an additional step could also be included in the protocol. After induction of both RAM-containing Pl.LSU/2 intron and IEP expression, plasmid DNA could be extracted and digested with enzymes that specifically cut the donor plasmid. Digested plasmid extracts would thus be used to transform *E. coli*, and phenotypic analysis, as well as PCR analysis would be performed. This way, homing products could be enriched, potentially allowing their subsequent detection. This method has already been used to evaluate the homing of Ll.LtrB intron into a plasmid target in *E. coli* (Cousineau B et al. 1998). The authors found that the Ll.LtrB intron retrohoming could be detected with an efficiency of 1.3×10^{-3} (homing product per recipient).

The *E. coli* homing assay should also be performed in *E. coli* Rosetta-gami B (DE3). Indeed, several results suggest that the IEP is correctly folded in Rosetta-gami B (DE3), as showed by the demonstration of the HisV5-IEP RT activity after expression in this strain. Moreover, there is some indications that the protein could adopt a different tridimensional conformation when expressed in BL21 Star (DE3) or in Rosetta-gami B (DE3). Indeed, HisV5-IEP expressed in Rosetta-gami B (DE3)

was able to bind a Ni^{2+}-charge resin during IMAC purification under natives conditions (See Fig. R-26), in contrast to HisV5-IEP expressed in BL21 Star (DE3) (See Fig. R-19), The absence of retrohoming products in BL21 Star (DE3) could thus be due to an incorrect folding of HisV5-IEP in this strain.

We speculated that the exon sequence requirement for Pl.LSU/2 intron retrohoming would be limited to the RNP recognition sequence of the homing site, as shown for *Lactococcus lactis* Ll.LtrB intron (Guo H et al. 1997; Matsuura M et al. 1997; Cousineau B et al. 1998) and yeast aI1 and aI2 introns (Yang J et al. 1998). They span from -25 to +10 for the Ll.LtrB intron, from -21 to +9 for the aI1 intron, and from -19 to +6 for the aI2 intron (See Fig. I-21). So far, those sequences are not defined for the Pl.LSU/2 intron. We postulated that the size of the Pl.LSU/2 intron target used in our homing assay, from -50 to + 71, would be sufficient to allow the homing site recognition by RNPs. However, the length of the target site could be further optimized.

It could also be necessary to first demonstrate the ability of Pl.LSU/2 intron to splice in *E. coli* before assaying its homing capacity. Indeed, we have not demonstrated the Pl.LSU/2 intron splicing ability in *E. coli* while the IEP-dependent splicing of Pl.LSU/2 intron was observed in *S. cerevisiae.*

All those adjustments could contribute to the development of a more relevant protocol that could permit the detection of the Pl.LSU/2 intron homing in *E. coli* if it exists.

3.2 - HOMING IN S. CEREVISIAE

Although the homing of Pl.LSU/2 intron could not been demonstrated in *E. coli*, the finding that Pl.LSU/2 intron is able to splice in *S. cerevisiae* led us to postulate that the formation of the RNP particle could occur in yeast and induce the Pl.LSU/2 intron retrohoming into its natural target. We thus developed a yeast homing assay directly based on the use of the two constructions that allowed the detection of spliced mRNA in yeast (See Fig. R-38). A homing acceptor plasmid was added in this assay and contains the E2-E3 natural DNA target site of Pl.LSU/2 intron. Yeasts were cotransformed by the acceptor plasmid and the donor plasmid expressing the Pl.LSU/2 intron. Recombinant yeast cells were subsequently transformed or not with the IEP expressing plasmid and grown in presence of glucose or galactose to repress or induce the IEP expression, respectively.

Homing events in yeast were analyzed by PCR on plasmid DNA extracted from the cells. We observed the presence of the amplification product specific to a homing event in all conditions, even in absence of IEP expression in yeast cells, where homing could not have occurred. We thus speculated that these plasmids, which was showed to contain integrated Pl.LSU/2 intron, could be the result of homologous recombination events between the donor and the acceptor plasmids, which both share homologous regions of at least 50 nt. Indeed, homologous recombination is a very efficient mechanism in *S. cerevisiae*. The acceptor plasmid contains the last 50 nucleotides of exon 2 and the first 71 nucleotides of exon 3, exactly as the donor plasmid does, except for the mutation of the 16^{th} nucleotide of exon 3 (A), which is replaced by a thymidine in the donor plasmid. These homologous

regions are sufficiently long to induce highly efficient homologous recombination (Manivasakam P et al. 1995).

The adaptation of the twintron strategy used in the *E. coli* homing assay would have overcome this issue. The insertion of the *td* intron sequence into Pl.LSU/2 intron might help to distinguish between retrohoming products and homologous recombination. Indeed, *td* intron splice from an RNA intermediate. The retrohoming of the Pl.LSU/2 intron into its DNA target would thus be performed from RNA in which the *td* group I intron would have been excised. The detection of this *td*-less Pl.LSU/2 intron integrated copy would thus permit to establish the Pl.LSU/2 retrohoming capacity over homologous recombination integration. The use of this twintron strategy in *E. coli* with the Ll.LtrB group II intron and *td* group I intron has showed that almost 80% of integrated events detected were free of group I intron *td* (Saldanha R et al. 1999).

The design of the strategy used for determining the homing in *S. cerevisiae* should thus be carefully assess to overcome these issues due to the high efficiency of homologous recombination in this organism.

4 - CONCLUSION

In this work, I have further characterized the *Pylaiella littoralis* Pl.LSU/2 group II intron with the aim of developing a novel site-specific vector for genomic targeting.

During this thesis work, I have conducted several studies, and in some cases, additional or alternative experiments could not have been performed due to a lack of time.

Nevertheless, recombinant Pl.LSU/2 IEP was expressed and purified both alone and with intronic RNA, and its RT activity was demonstrated. The splicing of the Pl.LSU/2 intron was showed in *S. cerevisiae*, and the IEP maturase activity was found to improve the intron splicing efficiency. However, we have not been able to detect any splicing of the Pl.LSU/2 intron in human cells, and the homing of the intron in *E. coli* and *S. cerevisiae* was not evidenced.

The results presented herein provide new information on the behavior of the Pl.LSU/2 group II intron and contribute to a further comprehension of this catalytic ribozyme. While confirming some findings obtained with the extensively studied Ll.LtrB intron, it emphasize the fact that the use of those evolutionary conserved ribozymes in gene therapy still requires much optimizations, research and time.

The ideal gene therapy vector, that could combine efficiency and safety, remains to be developed.

MATERIALS AND METHODS

1 - CELLULAR BIOLOGY

1.1 - BACTERIA

1.1.1 - *Escherichia coli* strains

- One Shot® TOP10 (Life Technologies; Invitrogen):

Genotype: F$^-$ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 ΔlacX74 *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G λ$^-$.

This strain was used for routinely subcloning. Incubation temperature for growth was 37°C.

- XL10-Gold® (Agilent Technologies; Stratagene):

Genotype: Tet^R Δ(*mcrA*)*183* Δ(*mcrCB-hsdSMR-mrr*)*173 endA1 supE44 thi-1 recA1gyrA96 relA1 lac* Hte [F' *proAB lacI*^{q}ZΔ*M15* Tn*10* (Tet^R) Amy Cam^R].

This strain was used for cloning of large DNA molecules and all final steps of cloning. Incubation temperature for growth was 32°C.

- MAX Efficiency® Stbl2™ (Life Technologies; Invitrogen):

Genotype: F$^-$ *mcr*A Δ(*mcr*BC-*hsd*RMS-*mrr*) *rec*A1 *end*A1*lon gyr*A96 *thi sup*E44 *rel*A1 λ$^-$ Δ(*lac-pro*AB).

This strain was used for cloning of retroviral sequences, as well as Pl.LSU/2 intron and IEP sequences. Incubation temperature for growth was 30°C.

- MAX Efficiency® DH10Bac™ (Life Technologies; Invitrogen):

Genotype: F$^-$ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K λ$^-$ *rps*L *nup*G /bMON14272 / pMON7124.

This strain was used to produce recombinant bacmid DNA for expression of Pl.LSU/2 IEP by the baculovirus/Sf9 system. It contains the parent bacmid (bMON14272) and the helper plasmid (pMON7124). Incubation temperature for growth was 37°C.

- One Shot® BL21 Star™ (DE3) (Life Technologies; Invitrogen):

Genotype: F$^-$ *ompT hsdS*$_B$ (r_B^- m_B^-) *gal dcm rne131* (DE3).

This strain was used to express the fusion GST-IEP protein. Incubation temperature for growth and protein expression was 37°C.

- Rosetta™ (EMD Biosciences; Novagen):

Genotype: F$^-$ *ompT hsdS*$_B$ (r_B^- m_B^-) *gal dcm* pRARE (Cam^R).

This strain was used to purify the pRARE plasmid encoding tRNAs for AGG, AGA, AUA, CUA, CCC and GCA codons. Incubation temperature for growth was 37°C.

- ArcticExpress™ (DE3)RIL (Agilent Technologies; Stratagene):

Genotype : F$^-$ *ompT hsdS* (r_B^- m_B^-) *dcm*$^+$ Tet^R *gal* λ(DE3) *endA* Hte [*cpn10 cpn60* Gent^R] [*argU ileY leuW* Str^R].

This strain was used to express GST-tagged proteins. Incubation temperature for growth and protein expression was 32°C and 15°C, respectively.

- Rosetta-gami™ B (DE3) (EMD Biosciences; Novagen):

Genotype: F^- *ompT hsdS*$_B$ ($r_B^- m_B^-$) *gal dcm lac*Y1 *ahp*C (DE3) gor522::Tn*10 trx*B pRARE (Cam^R, Kan^R, Tet^R).

This strain was used to express HisV5-tagged proteins and RNP particles. Incubation temperature for growth and protein expression was 32°C and 30°C, respectively.

1.1.2 - Growth and maintenance

All *E. coli* strains were grown in LB (Luria-Bertoni) broth with shaking at 180 rpm or streaked on solid LB agar plates and maintained at appropriate temperature. LB broth or agar containing appropriate antibiotics were used (50 µg/ml of carbenicillin, 34 µg/ml of chloramphenicol, 50 µg/ml of kanamycin, 7 µg/ml of gentamicin, 10 µg/ml of tetracycline). When blue/white selection could be performed, 30 µg/ml of X-Gal and IPTG were added on agar plates.

1.1.3 - Production of chemically competent *E. coli* BL21 Star (DE3) pRARE strain

BL21 Star (DE3) strain was transformed with the pRARE plasmid purified from Rosetta strain. A 10 ml starter culture of LB broth containing chloramphenicol was made by inoculation with a single clone of BL21 Star (DE3) pRARE and shaken at 37°C overnight. Eight ml of the overnight culture was diluted 1/100 on LB broth containing chloramphenicol and shaken at 37°C until the culture reached an OD_{600nm} of approximately 0.5. The culture was then chilled on ice for 10 min and centrifuged at 6,000*g* for 20 min at 4°C. The cell pellet was resuspended in 1/2 original volume of ice-cold 50 mM $CaCl_2$ sterile-filtered, incubated at 4°C for 5 min, and centrifuged as before. The cell pellet was resuspended in 1/40 original volume of ice-cold 80 mM $CaCl_2$ containing 15% glycerol and sterile-filtered, and 300 µl aliquots of competent cells were then snap frozen in liquid nitrogen and stored at -80°C.

1.1.4 - Transformation of *E. coli*

- 100 pg of plasmid DNA or 1-5 µl of ligation mixture were gently mixed with a 50µl aliquot of One Shot TOP10 chimiocompetent cells and stored on ice for 30 min. The cells were heated at 42°C for 30 sec and chilled on ice for 2 min. 250 µl of SOC medium sterile-filtered (2% Tyrptone, 0.5% Yeast Extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose) were added and cells were shaken at 37°C for 1 hr before plating on selective medium. Plates were incubated at 37°C overnight.
- 4 µl of β-mercaptoethanol mix supplied with XL10-Gold cells were added to 100 µl aliquot of XL10-Gold chimiocompetent cells. After 10 min of incubation on ice, 1 ng of plasmid DNA or 2 µl of ligation mixture were gently mixed with the cells and stored on ice for 30 min. The cells were heated at 42°C for 30 sec and chilled on ice for 2 min. 500 µl of SOC medium were added and cells were shaken at 32°C for 1 hr before plating on selective medium. Plates were incubated at 32°C overnight.
- 50 pg of plasmid DNA or 10 ng of ligation products were gently mixed with a 100µl aliquot of MAX Efficiency Stbl2 chimiocompetent cells, and incubated on ice for 30 min. The cells were heated at 42°C for 25 sec and chilled on ice for 2 min. 900 µl of SOC medium were added and cells were

shaken at 30°C for 1 hr or 1 hr 30 min for plasmid and ligation products transformation, respectively. Cells were then plating on selective medium and incubated at 30°C overnight.

- 1 ng of plasmid DNA was gently mixed with a 100µl aliquot of MAX Efficiency DH10Bac chimiocompetent cells placed on 14-ml round-bottom tubes, and stored on ice for 30 min. The cells were heated at 42°C for 45 sec and chilled on ice for 2 min. 900 µl of SOC medium were added and cells were shaken at 37°C for 4 hrs before plating on selective medium and incubated at 37°C for at least 24 hrs.
- 10 ng of plasmid DNA were gently mixed with a 50µl aliquot of One Shot BL21 Star (DE3) chimiocompetent cells and stored on ice for 30 min. The cells were heated at 42°C for 30 sec and chilled on ice for 2 min. 250 µl of SOC medium were added and cells were shaken at 37°C for 1 hr before plating on selective medium. Plates were then incubated at 37°C overnight.
- 10 ng of plasmid DNA were gently mixed with a 100µl aliquot of BL21 Star (DE3) pRARE or 50 µl aliquot of Rosetta-gami B (DE3) chimiocompetent cells placed on 14-ml round-bottom tubes, and stored on ice for 30 min. The cells were heated at 42°C for 30 sec and chilled on ice for 2 min. 500 µl of SOC medium were added and cells were shaken at 37°C for 1 hr before plating on selective medium. Plates were then incubated at 37°C overnight.
- 2 µl of 1/10 dilution of β-mercaptoethanol mix supplied with ArcticExpress (DE3)RIL cells were added to 100 µl aliquot of ArcticExpress (DE3)RIL chimiocompetent cells placed on 14-ml round-bottom tubes. After 10 min of incubation on ice, 10 ng of plasmid DNA were gently mixed with the cells and stored on ice for 30 min. The cells were heated at 42°C for 20 sec and chilled on ice for 2 min. 900 µl of SOC medium were added and cells were shaken at 37°C for 1 hr before plating on selective medium. Plates were incubated at 37°C overnight.

1.2 - YEAST

1.2.1 - Strain

In vivo homing of Pl.LSU/2 intron was evaluated in *S. cerevisiae* BY4742 MATα *his3*Δ1 *leu2*Δ0 *lys2*Δ0 *ura3*Δ0 (S288C).

1.2.2 - Growth and maintenance

BY4742 *S. cerevisiae* strain was grown in YPD (Yeast extract Peptone Dextrose, Clontech Laboratories) medium with shaking at 220 rpm or streaked on solid YPD agar plates (2% (w/v) agar dissolved in YPD by heating) and maintained at 30°C.

For selection of recombinant yeast cells, cells were cultured in minimal SD (Synthetic Dextrose) medium containing either glucose or galactose/raffinose (Clontech Laboratories) and appropriate dropout (DO) supplement mix prepared according to (Sambrook J and Russell DW 2001), or streaked on solid SD agar (2% (w/v) agar) plates containing appropriate DO supplement mix. For long term storage of yeast strains, a sample of large inoculum from a freshly grown plate was resuspended in 1 ml of sterile 15% (v/v) glycerol and stored at -80°C.

1.2.3 - Transformation of *S. cerevisiae*

Yeast cells were transformed using the yeast transformation kit from Sigma Aldrich, according to the manufacturer's instructions.

1.3 - INSECT CELLS

1.3.1 - Sf9 cells

Spodoptera frugiperda Sf9 cells (Life Technologies, Invitrogen) adapted to serum-free suspension culture in Sf-900 II SFM (Serum-Free Medium; Life Technologies, Invitrogen) were used to produce Pl.LSU/2 IEP. Sf9 cells were cultured in Sf-900 II SFM at 27°C either in suspension with 150 rpm using magnetic agitation in 500 ml spinners (Bellco), or in adherence in cell culture plates and flasks.

2 - MOLECULAR BIOLOGY

2.1 - OLIGONUCLEOTIDES

Oligonucleotides used in this thesis were provided by Sigma Aldrich and are listed below (Table M-1).

Name	Sequence (5' to 3')	Length (nt)	Tm (°C)
PlLSU2F-BamHI	CGCGGATCCATGAGTATTCCATATATAATT	30	68.9
PlLSU2R-NotI	ATAAGAATGCGGCCGCTTAAATGTTCAAGATCTTGCC	37	79.3
YAAA-F	TGGTAAGGTATGCGGCTGCCTTCGTCGTTACCGC	34	83.5
YAAA-R	GCGGTAACGACGAAGGCAGCCGCATACCTTACCA	34	83.5
DeltaIEP-F	ATGGTTTATTCGTCGTTACCGCTGCAACAAAACG	34	78.4
DeltaIEP-R	ACGACGAATAAACCATCCAACGTCATGTTTGCAATTAAAG	40	78.6
PlLSU2F2	CACCATGAGTATTCCATATATA	22	53.2
PlLSU2R2	TTAAATGTTCAAGATCTTGC	20	54.4
IEP-XbaImut-F	GGTTAAAAGCAGGTGCTCTGGAAACAACAACTCAGGAG	38	78.8
IEP-XbaImut-R	CTCCTGAGTTGTTGTTTCCAGAGCACCTGCTTTTAACC	38	78.8
PlLSU2-XbaI-F	GCTCTAGAACTAGTGGATCCCCCGGGCTGCA	31	80.5
PlLSU2-XbaI-R	GCTCTAGAGTTTTCAAAATGATTTCCTTAGAGCAAG	36	71.0
PlLSU2-XhoI-F	AAACGAATACTCGAGGATATAGTGAAACCG	30	68.7
PlLSU2-SacI-R	CGAGCTCTCGATAAGCTTTACCTGCCG	27	74.0
DM1	AGGATCCCAGCTTTTATCTTTGACACA	27	69.0
DM2	GTAGCTTTCGAAGCTTTACCTGCCGGCACC	30	77.9
DM3	GGTAAAGCTTCGAAAGCTACATATAAGGAA	30	66.6
DM4	GAATTCAGTTTTTTAGTTTTGCTGG	25	62.7
NH27	GGTCTCTGCTATCTCGCAAGAGGATGTATAGGGACTGACACCTGCC	46	83.7
NH28	GGCAGGTGTCAGTCCCTATACATCCTCTTGCGAGATAGCAGAGACC	46	83.7
MutHtoB-F	GAATACTCAAGCTATGGATCCCAGCTTTTATCTTTGACACA	41	76.2
MutHtoB-R	TGTGTCAAAGATAAAAGCTGGGATCCATAGCTTGAGTATTC	41	76.2
B-Ko-F	GGGATCCACCATGAGTATTCCTTACATAATTCCG	34	74.5
E-Stop-R	AGAATTCCTAGGCAGCGCCGTTCAG	25	74.1
SalI-P+T-F	GTCGACCTAAACTCACAAATTAGAGCTTC	29	66.6
XbaI-P+T-R	TCTAGATGTGAGTTAGCTCACTCATTAG	28	61.9
N-DIVa-F	CATATGAAGCTTTTATCTTTGACACAAAATCGGGGGTGGCGAC	43	82.1
NH42	TGATTTAGTGTGCCGCGGTAACTAAACCAGAATCACCTTTT	41	79.1
N-RBS-DIVa-R	CATATGTATATCTCCTTCTAAGCTTTACCTGCCGGCACCG	40	78.2
NH43	TGGTTTAGTTACCGCGGCACACTAAATCAAGAAGCCTTTT	40	79.2
S-Kan^{R}*td*-F	CCGCGGCTAAAACAATTCATCCAGTAA	27	72.8
S-Kan^{R}*td*-R	CCGCGGTTCAAAATCGGCTCCGTC	24	79.3

IDO_1747	CTCCCAGTTCAATTACAGC	19	57.3
IDO_848	TTCACTGCATTCTAGTTGTGG	21	60.4

Table M-1: Oligonucleotides used.
The name, sequence from 5' to 3', length, and melting temperature (Tm) are indicated.

2.2 - NUCLEIC ACID PURIFICATION AND ANALYSES

2.2.1 - Plasmid DNA purification

Recombinant *E. coli* cells were cultured by shaking in LB broth containing relevant antibiotic at the appropriate temperature overnight. Plasmid DNA was extracted from late-log phase cultures by alkaline lysis using the NucleoSpin® plasmid (Macherey-Nagel, < 25 µg), or NucleoBond® Xtra Maxi Plus (Macherey-Nagel, < 1000 µg), according to the manufacturer's instructions.

Plasmid DNA from yeast cells was extracted from 5 ml of mid-log phase cultures in SD minimal medium containing appropriate DO supplement mix using the Yeast plasmid isolation kit (Clontech Laboratories, < 4 µg), according to the manufacturer's instructions.

DNA concentrations were quantified by measuring the OD_{260} on a NanoDrop™ ND-8000 spectrophotometer (Thermo Scientific). The ratio OD_{280}/OD_{260} was 1.85-1.9, revealing a high DNA purity.

2.2.2 - DNA precipitation

DNA in solution was precipitated by adding 1/10 volume of 3 M sodium acetate pH 5.2 and 2 volumes of cold 100% ethanol. Precipitation was performed at -20°C overnight before centrifugation at 13,000 rpm in a microcentrifuge for 30 min. The pellet was washed with 70% ethanol, centrifuged as before, dried, and resuspended in Milli-Q water or TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA).

2.2.3 - DNA electrophoresis

0.5-1.5% (w/v) high melting temperature agarose or 2-3% (w/v) low melting temperature agarose was dissolved in 1X TAE buffer (40 mM Tris-acetate, 5 mM EDTA) by heating, and 1X SYBR® Safe DNA gel stain (Life Technologies; Invitrogen). DNA samples were loaded with DNA loading dye (Thermo Scientific; Fermentas: 0.005% bromophenol blue, 0.005% xylene cyanol FF, 10% glycerol, 10 mM EDTA, and 1.7 mM Tris-HCl pH 7.6). A broad range DNA ladder (GeneRuler DNA ladder mix; Thermo Scientific, Fermentas) was added as a standard and DNA fragments were separated by electrophoresis in 1X TAE at 80-120V. Fluorescent visualization of DNA fragments was performed by exposure to ultraviolet light using a GBox-HR gel documentation system (Syngene).

2.2.4 - Agarose gel extraction

DNA fragments separated by electrophoresis were excised from the gel using a scalpel and purified using the NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel), according to the manufacturer's instructions.

2.2.5 - Enzymatic restriction digestion

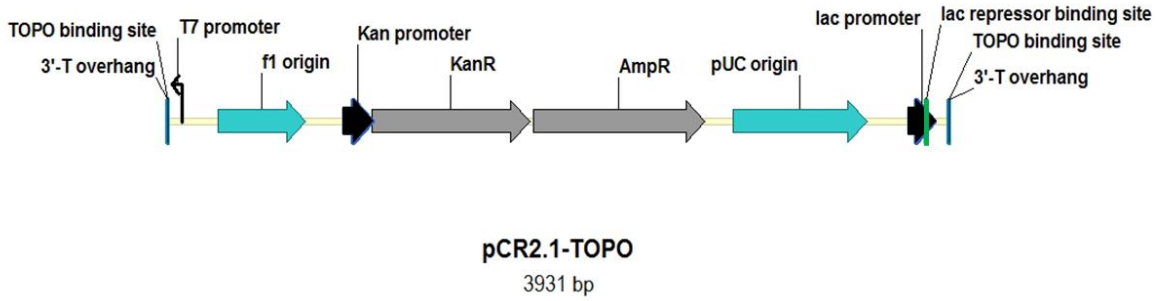
For screening of recombinant plasmid DNA, 500 ng to 2 µg of DNA were digested with 1 to 5 U of one or two restriction endonuclease (New England Biolabs; < 10% final volume) in a final volume of 20 µl. For extraction of digested DNA fragment from agarose gel and cloning, up to 10 µg of DNA were digested in a final volume of 50 µl. Digestions were performed for 1 hr at appropriate temperature for the enzyme used.

2.3 - CLONING

2.3.1 - TOPO cloning

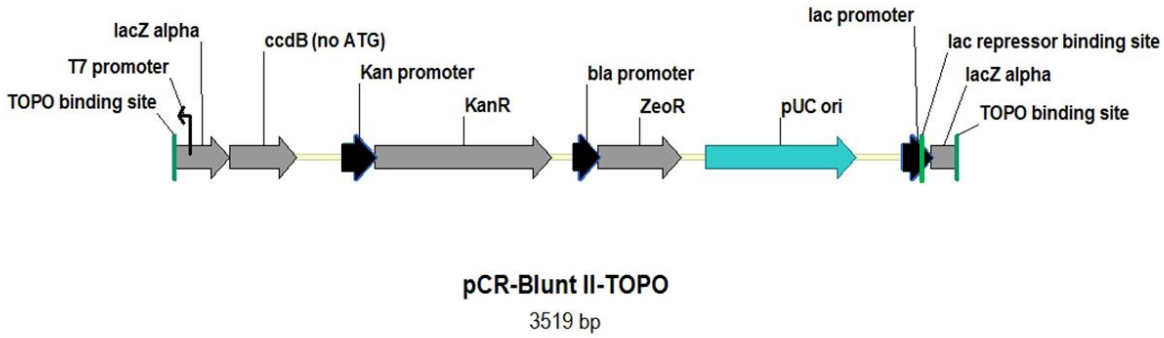
TOPO cloning allows rapid cloning, either directional or not, of PCR products into vectors. The ligation of PCR inserts into linearized TOPO® vectors (Life Technologies; Invitrogen) is achieved through the action of the Topoisomerase I covalently bound to the vector.

TOPO TA Cloning® is a non-directional cloning system for PCR products with 3' dATPs overhang residues, which are ligated into the linearized pCR®2.1-TOPO® vectors, containing 3' dTTP overhang residues.



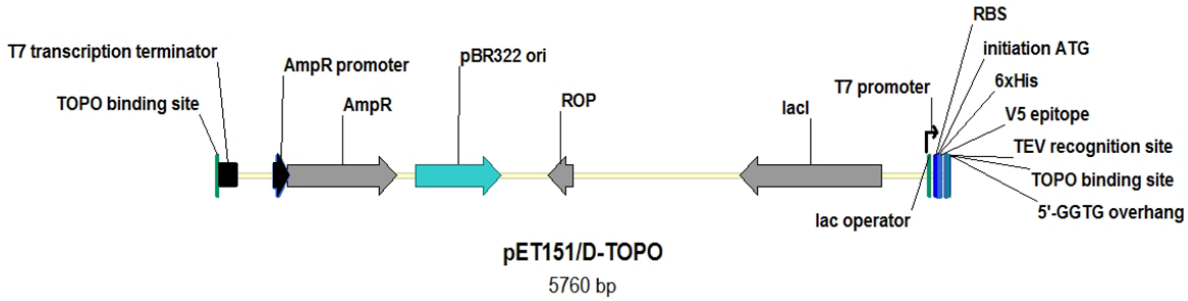
PCR inserts are amplified using a high-fidelity Taq DNA polymerase presenting a proofreading activity (See Materials and methods section 2.4 -). Addition of dATPs to the 3' end of PCR inserts is then performed using the AmpliTaq Gold DNA polymerase (Life Technologies; Applied Biosystems), which has a nontemplate-dependent terminal transferase activity (See Materials and methods section 2.4 -). 1-4 µl of AmpliTaq Gold polymerase-amplified PCR products are incubated with 10 ng of pCR2.1-TOPO vectors in presence of 200 mM NaCl and 10 mM $MgCl_2$ in a final volume of 6 µl at room temperature for 10 min. Ligation products were immediately transformed into competent *E. coli* cells.

Zero Blunt® TOPO® PCR cloning is a non-directional cloning system for blunt-end PCR products, which are ligated into the linearized pCR®-BluntII-TOPO® vector.



1-4 µl of Phusion High-fidelity DNA polymerase-amplified PCR products are incubated with 10 ng of pCR-BluntII-TOPO vectors in presence of 200 mM NaCl and 10 mM $MgCl_2$ in a final volume of 6 µl at room temperature for 10 min. Ligation products were immediately transformed into competent *E. coli* cells.

Champion™ pET Directional TOPO® system allows directional cloning of blunt-end PCR products into *E. coli* expression vectors. The PCR product is amplified using a forward primer whose sequence begins by CACC, and inserted into the linearized pET151/D-TOPO® vector, which have a GTGG overhang.



1-4 µl of Phusion High-fidelity DNA polymerase-amplified PCR products are incubated with 10 ng of pET151/D-TOPO vectors in presence of 200 mM NaCl and 10 mM $MgCl_2$ in a final volume of 6 µl at room temperature for 10 min. Ligation products were immediately transformed into competent *E. coli* cells.

2.3.2 - Dephosphorylation and ligation

5' phosphate groups were removed from up to 5 pmol digested DNA ends using 0.5 units of Shrimp Alkaline Phosphatase (SAP) (Affymetrix; USB) per pmol 5'-ends in 1X SAP reaction buffer in a final volume of 25 µl. Incubation was performed at 37°C for 30 min. SAP was inactivated by incubation at 65°C for 15 min.

50 ng of digested and dephosphorylated plasmid backbone was ligated to insert DNA in 1:1, 1:3 or 1:6 molar ratios using 2,000 units of Quick T4 DNA ligase in 1X Quick Ligation buffer (New England Biolabs) in a final volume of 20 µl. Ligation was performed for 5 min at room temperature and chilled on ice before transformation of ligated products in competent *E. coli*.

2.3.3 - Site-directed mutagenesis

Site-directed mutagenesis of plasmid DNA was performed with the QuickChange II XL site-directed mutagenesis kit (Agilent Technologies) that uses *PfuUltra* high-fidelity (HF) DNA polymerase for

mutagenic primer-directed replication of both plasmid strands. For point mutation, the primers were designed to contain the desired mutation (Fig M-2A; black star), according to manufacturer's guidelines. For deletion, primers were designed to partially hybridize plasmid DNA on either side of the sequence to delete, and to contain a sequence of 8 nt-long at their 5' end that is complementary to the other primer (Fig. M-2B), so that overlapping between the two primers is 16 nt-long.

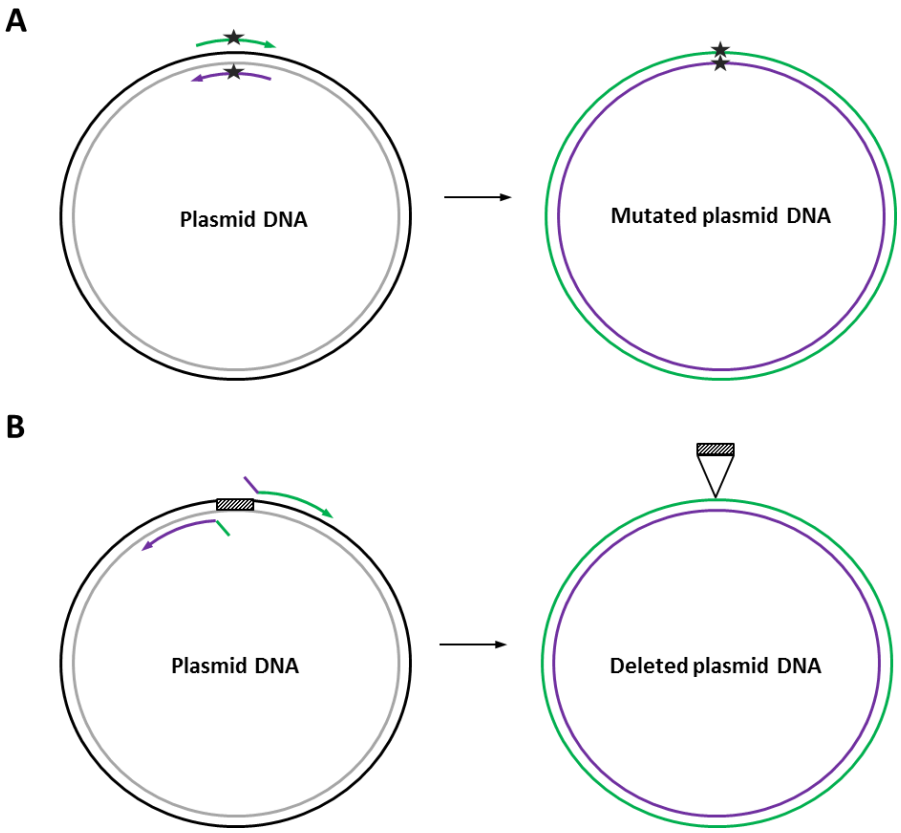
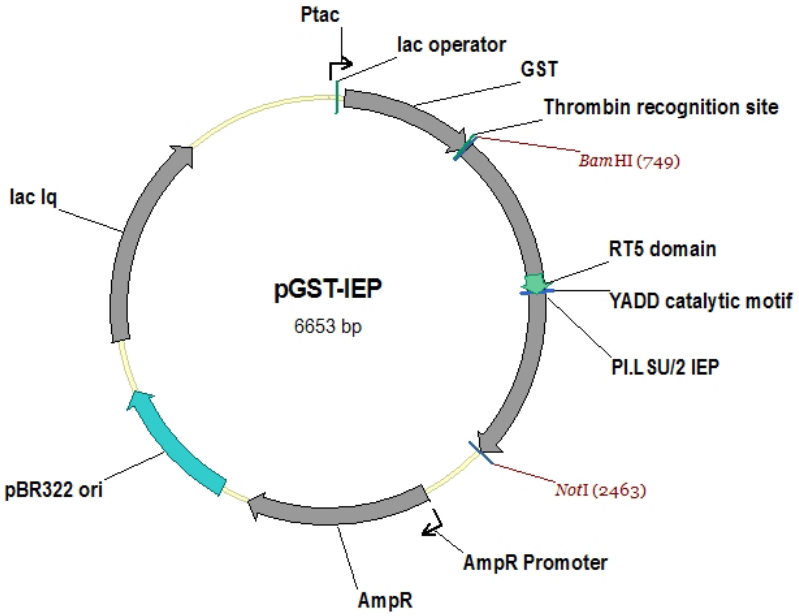


Figure M-2: Schematic overview of site-directed mutagenesis method.
(***A***) Site-directed point mutation. Green and purple arrows: primers used; gray star: desired mutation. (***B***) Site-directed deletion. Purple and green short lines: 5' end of primers that do not hybridize with the parental plasmid DNA and correspond to a 8 nt-long sequence complementary to the other primer; green and purple arrows: primer sequences hybridizing the parental plasmid DNA on either side of the region to delete; hatched rectangle: region to delete. After the first cycle of PCR amplification, the two newly synthetized DNA strands can pair together at primer sequences, which share a 16 nt-long overlap.

10-30 ng of plasmid DNA were amplified using 2.5 units of *PfuUltra* HF DNA polymerase, 11-15 pmole of each FPLC (Fast polynucleotide liquid chromatography)-purified primers, 1 µl of dNTP mix (provided with the kit), 1X reaction buffer (provided with the kit) in a final volume of 50 µl. Initial denaturation step was performed at 95°C for 30 sec, followed by 16 or 18 amplification cycles for point mutation or deletion, respectively, consisting of denaturation at 95°C for 30 sec, primer annealing at 55°C for 1 min (the annealing temperature is set according to the primer melting temperature, which must be ≥ 78°C), and extension at 68°C for 1 min 20 sec/kb. Reactions were then cool to 37°C and 1 unit of *Dpn* I restriction enzyme was added to digest the parental (non-mutated) DNA. Incubation was performed at 37°C for 1 hr and chilled on ice before transformation of 1 µl of *Dpn* I-treated mutation products in XL10-Gold competent *E. coli* cells.

2.3.4 - Description of plasmid cloning

- pGST-IEP



The Pl.LSU/2 IEP ORF was first amplified by PCR from the pPlLSU/2 plasmid (Appendix x; (Costa M et al. 1997b)) using PlLSU2F-BamHI and PlLSU2R-NotI oligonucleotides. The resulting PCR fragment, containing BamHI and NotI restriction sites, was ligated in pCR2.1-TOPO by TOPO TA cloning. After BamHI and NotI restriction digestion of the recombinant plasmid obtained, the BamHI/NotI IEP ORF restriction fragment was ligated with the BamHI/NotI digested pGEX-4T1 plasmid (GE Healthcare Life Sciences; Appendix x). The resulting pGST-IEP plasmid contains the IEP ORF in frame with the GST ORF, and should express the GST-IEP fusion protein.

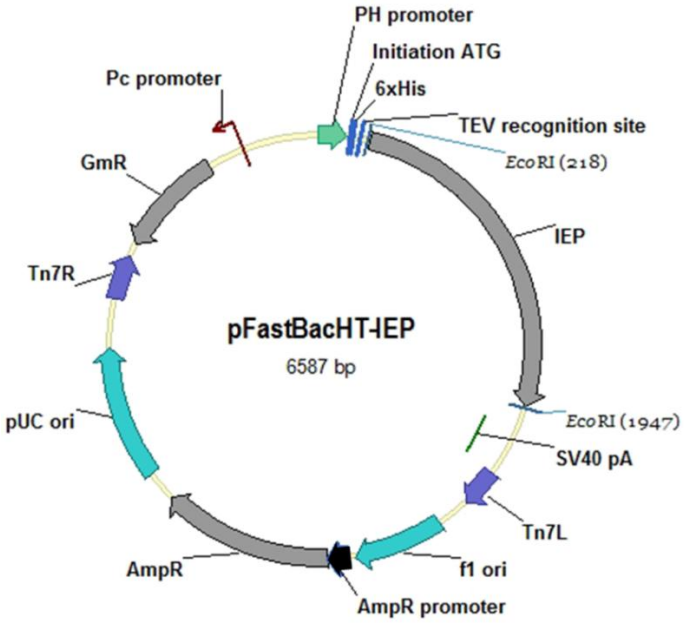
- pGST-IEPmtDD-

The RT catalytic motif YADD of the IEP sequence was mutated in YAAA by site-directed mutagenesis of the pGST-IEP plasmid using YAAA-F and YAAA-R oligonucleotides. The resulting plasmid should express a RT-defective GST-IEPmtDD- fusion protein.

- pGST-IEPΔRT5

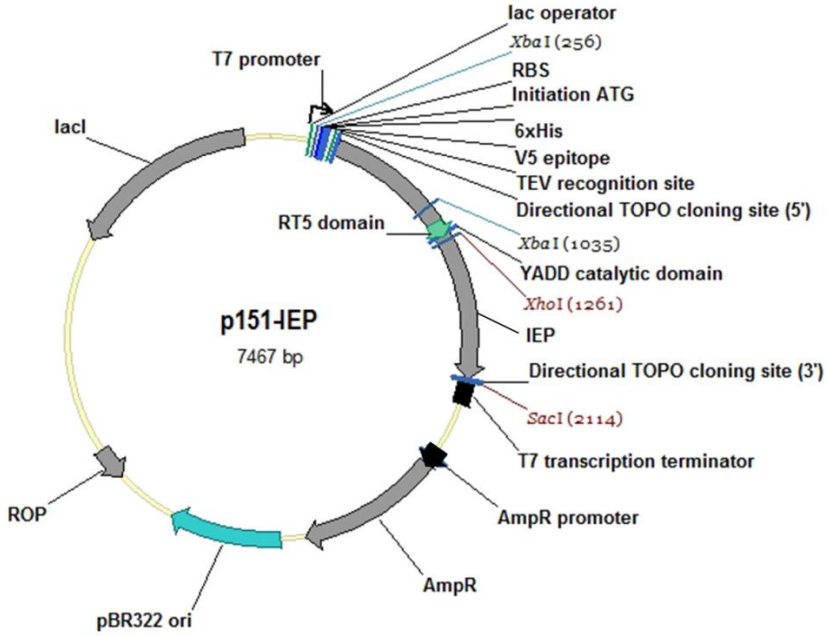
The RT5 domain of the IEP sequence was deleted by site-directed mutagenesis of the pGST-IEP plasmid using DeltaIEP-F and DeltaIEP-R oligonucleotides. The resulting plasmid should express a RT-defective GST-IEPΔRT5 fusion protein.

- pFastBacHT-IEP



The Pl.LSU/2 IEP ORF was first amplified by PCR from pPlLSU/2 plasmid using PlLSU2F2 and PlLSU2R2 oligonucleotides and ligated into pCR2.1-TOPO by TOPO TA cloning. The resulting recombinant plasmid was digested with EcoRI and ligated into EcoRI-digested pFastBac-HT C plasmid (Life Sciences, Invitrogen; Appendix x). The resulting pFastBacHT-IEP plasmid contains the IEP ORF in frame with the 6xHis tag of pFastBacHT C plasmid, and should express the His-IEP fusion protein.

- p151-IEP



The Pl.LSU/2 IEP ORF was amplified by PCR from the pPlLSU/2 plasmid using PlLSU2F2 and PlLSU2R2 oligonucleotides and ligated in pET151/D-TOPO by directional Champion pET TOPO

cloning. The resulting p151-IEP plasmid contains the IEP sequence in frame with the 6xHis tag and the V5 epitope of the pET151/D-TOPO vector, and should express the HisV5-IEP fusion protein.

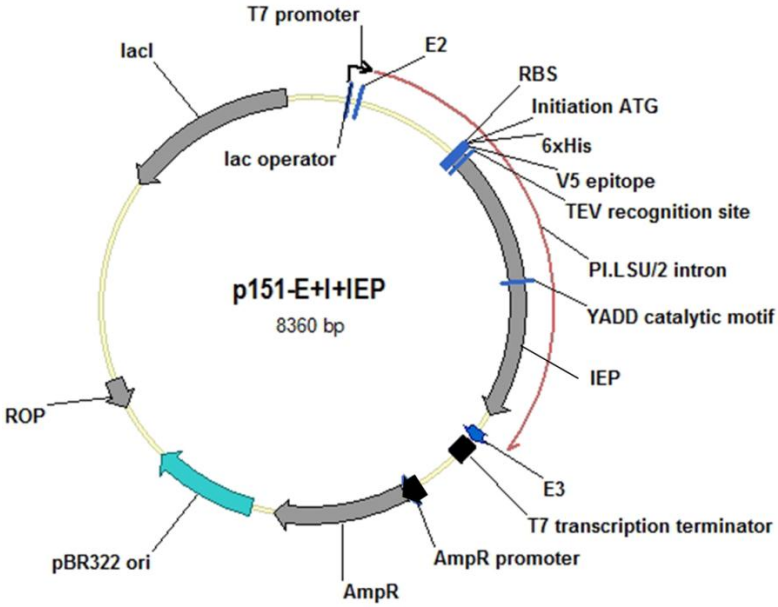
- p151-IEPmtDD-

The RT catalytic motif YADD of the IEP sequence was mutated in YAAA by site-directed mutagenesis of the p151-IEP plasmid using YAAA-F and YAAA-R oligonucleotides. The resulting plasmid should express a RT-defective HisV5-IEPmtDD- fusion protein.

- p151- IEPΔRT5

A part of the IEP sequence containing the RT5 domain was deleted by site-directed mutagenesis of the p151-IEP plasmid using DeltaIEP-F and DeltaIEP-R oligonucleotides. The resulting plasmid should express a RT-defective HisV5-IEPΔRT5 fusion protein.

- p151-E+I+IEP



The cloning of p151-E+I+IEP plasmid was performed in 3 steps (Fig. M-3).

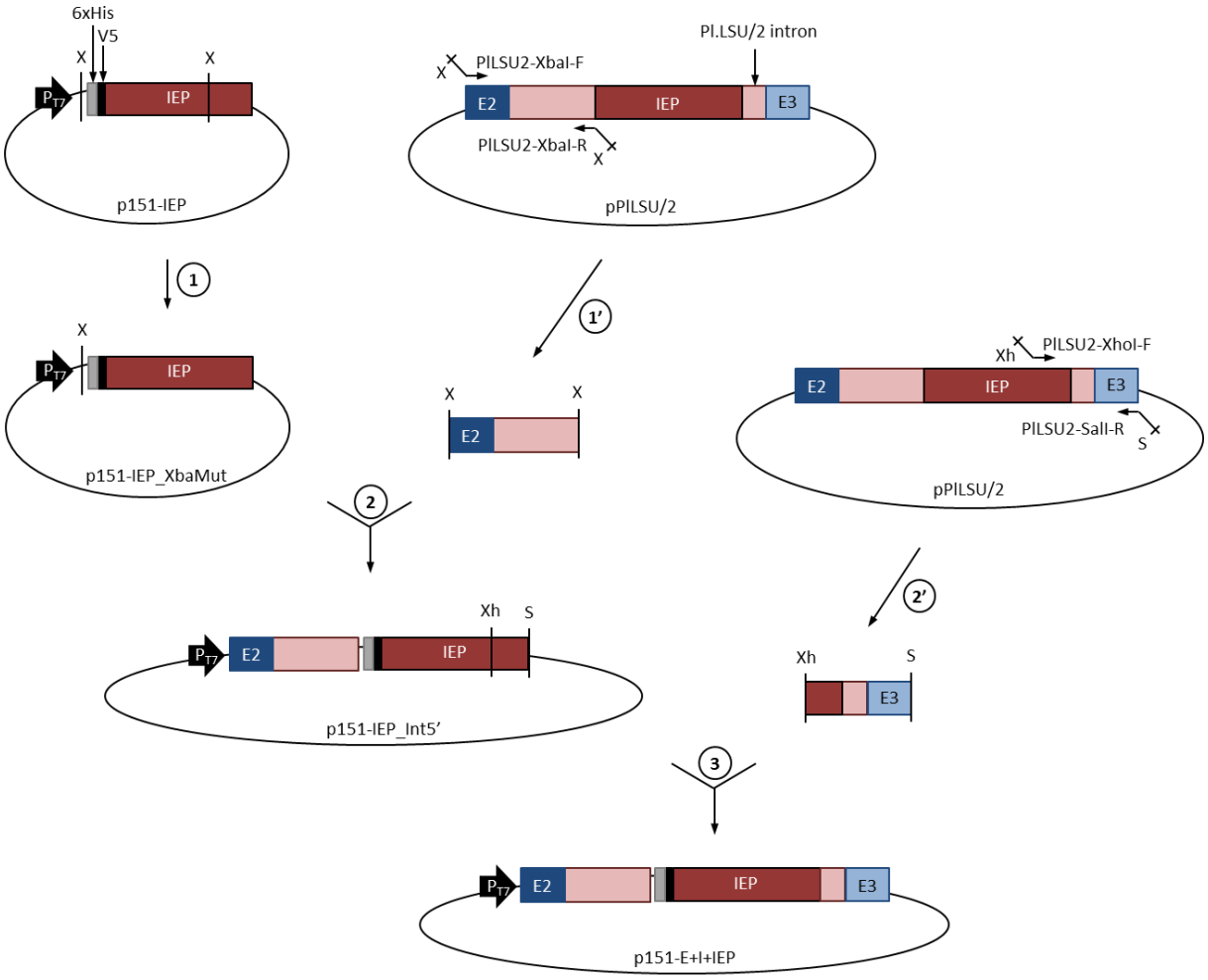


Figure M-3: Schematic representation of p151-E+I+IEP cloning.

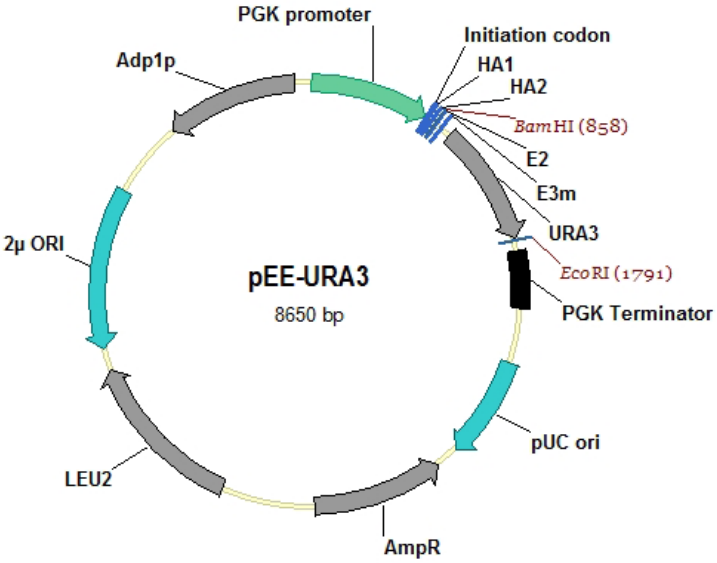
The cloning strategy of p151-E+I+IEP is represented here and described below.

The XbaI site (Fig. M-3; X) located in the IEP sequence of p151-IEP plasmid was first eliminated by site-directed mutagenesis using IEP-XbaImut-F and IEP-XbaImut-R oligonucleotides (Fig. M-3; step 1). The exon 2 and the intron sequence upstream the IEP ORF located in the pPlLSU/2 plasmid were then amplified by PCR using PlLSU2-XbaI-F and PlLSU2-XbaI-R oligonucleotides (Fig. M-3; step 1') and ligated in XbaI digested p151-IEP_XbaMut (Fig. M-3; step 2). The resulting plasmid was digested with XhoI (Fig. M-3; Xh) and SacI (Fig. M-3; S) and ligated with the insert containing the 5'end of the IEP ORF, the intron sequence downstream of the IEP ORF and the Exon 3 (Fig. M-3; step 3), amplified by PCR using PlLSU2-XhoI-F and PlLSU2-SacI-R oligonucleotides (Fig. M-3; step 2'). The resulting p151-E+I+IEP plasmid contains the Pl.LSU/2 intron, flanked by its two exons and containing the IEP sequence in frame with the 6xHis tag and V5 epitope in its domain IV, and should express both the intron and the HisV5-IEP fusion protein.

- p151-E+I+IEPmtDD-

The YADD catalytic motif of the IEP sequence was mutated in YAAA by site-directed mutagenesis of the p151-E+I+IEP plasmid using YAAA-F and YAAA-R oligonucleotides. The resulting plasmid should express both the Pl.LSU/2 intron and a RT-defective HisV5-IEPmtDD- fusion protein.

- pEE-URA3



The cloning of pEE-URA3 plasmid was performed in 5 steps (Fig. M-4).

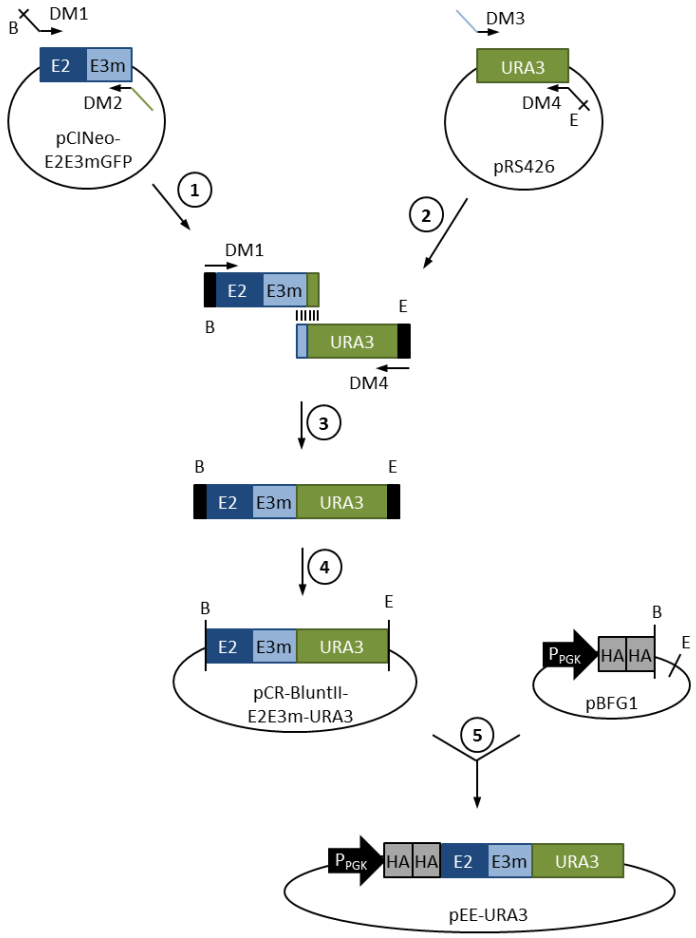
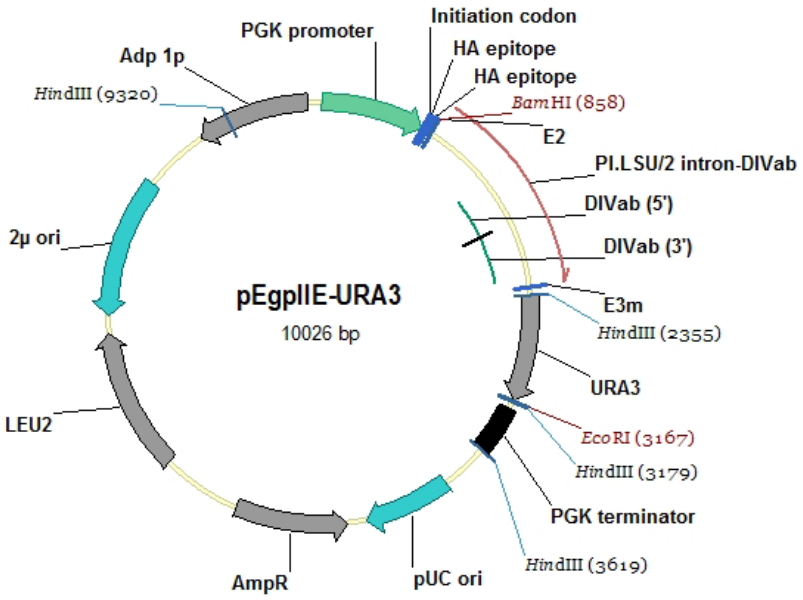


Figure 5.4: Schematic representation of pEE-URA3 cloning.

The cloning strategy of pEE-URA3 is represented here and described below.

Sequences of exon 2 (50 last nucleotides of exon 2; Fig. M-4, E2) and exon 3, in which the 16th nucleotide was changed in thymine to eliminate a Stop codon (the first 71 nucleotides of exon 3; Fig. M-4, E3m), located in the pCINeo-E2E3mGFP plasmid (previously constructed by a collaborator, Appendix x), were amplified by PCR (Fig. M-4; step 1) using oligonucleotides DM1, which contains a BamHI restriction site (Fig. M-4; B) and DM2, which contains the beginning of the URA3 sequence (Fig. M-4; green line). In parallel, the URA3 ORF sequence located in the pRS426 plasmid (Appendix x) was amplified by PCR (Fig. M-4; step 2) using oligonucleotides DM3, which contains the end of the E3m sequence (Fig. M-4; light blue line) and DM4, which contains an EcoRI restriction site (Fig. M-4; E). These two PCR products were subsequently mixed and amplified by PCR using DM1 and DM4 oligonucleotides (Fig. M-4; step 3). The resulting PCR product was cloned into pCR-BluntII-TOPO vector by TOPO cloning (Fig. M-4; step 4), forming the pCR-BluntII-E2E3m-URA3 plasmid. The E2E3m-URA3 sequence was isolated by digestion of pCR-BluntII-E2E3m-URA3 with EcoRI and BamHI and ligated into the BamHI/EcoRI-digested pBFG1 plasmid ((Yelin R et al. 1999); Appendix x) (Fig. M-4; step 5), resulting in the formation of the pEE-URA3 plasmid. This plasmid contains the E2E3m-URA3 sequence cloned in frame of two HA tags (Fig. M-4; gray rectangles), and should express the HA-E2E3m-URA3 fusion protein.

- pEgpIIE-URA3



The cloning of pEgpIIE-URA3 plasmid was performed in 4 steps (Fig. M-5).

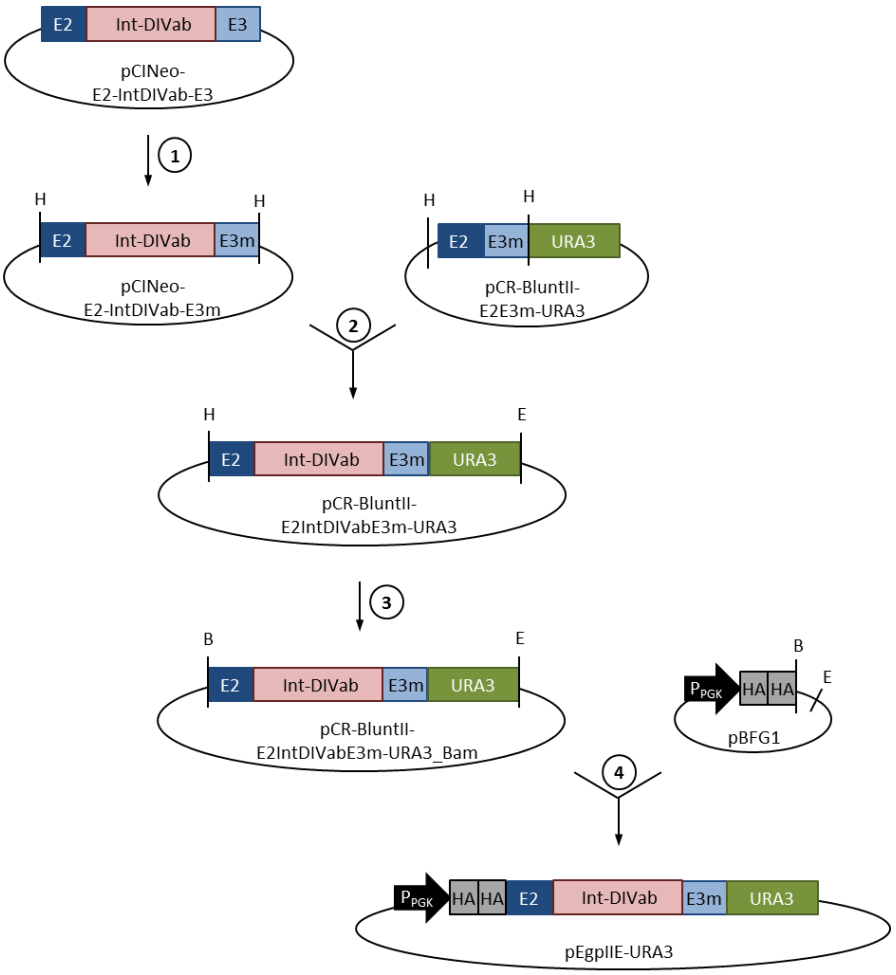
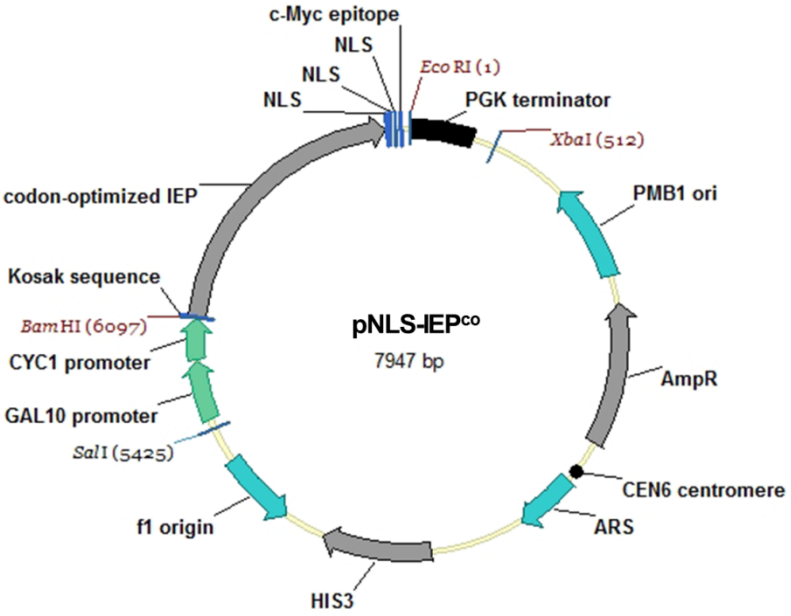


Figure M-5: Schematic representation of pEgpIIE-URA3 cloning.
The cloning strategy of pEgpIIE-URA3 is represented here and described below.

The 16th nucleotide of the exon 3 sequence (Fig. M-5; E3) located in the pCINeo-E2-IntDIVab-E3 plasmid (previously constructed by a collaborator, Appendix x) was changed in thymine (Fig. M-5; E3m) by site-directed mutagenesis using NH27 and NH28 oligonucleotides (Fig. M-5; step 1). The resulting pCINeo-E2-IntDIVab-E3m was digested with HindIII (Fig. M-5; H) and the E2-IntDIVab-E3m restriction fragment was isolated and ligated into the HindIII-digested pCR-BluntII-E2E3m-URA3 plasmid (Fig. M-5; step 2). The HindIII site of the resulting pCR-BluntII-E2IntDIVanE3m-URA3 plasmid was mutated into a BamHI site (Fig. M-5; B) by site-directed mutagenesis using MutHtoB-F and MutHtoB-R oligonucleotides (Fig. M-5; step 3). The resulting pCR-BluntII-E2IntDIVabE3m-URA3_Bam plasmid was digested with BamHI and EcoRI (Fig. M-5; E) and the restriction fragment containing E2IntDIVabE3m-URA3 was ligated into the BamHI/EcoRI-digested pBFG1 plasmid (Fig. M-5; step 4). This plasmid is the splicing reporter construct used in the yeast splicing assay, and should express the HA-E2E3m-URA3 fusion protein upon precise Pl.LSU/2 intron splicing.

- pNLS-IEPco



The cloning of pEgpIIE-URA3 plasmid was performed in 4 steps (Fig. M-6).

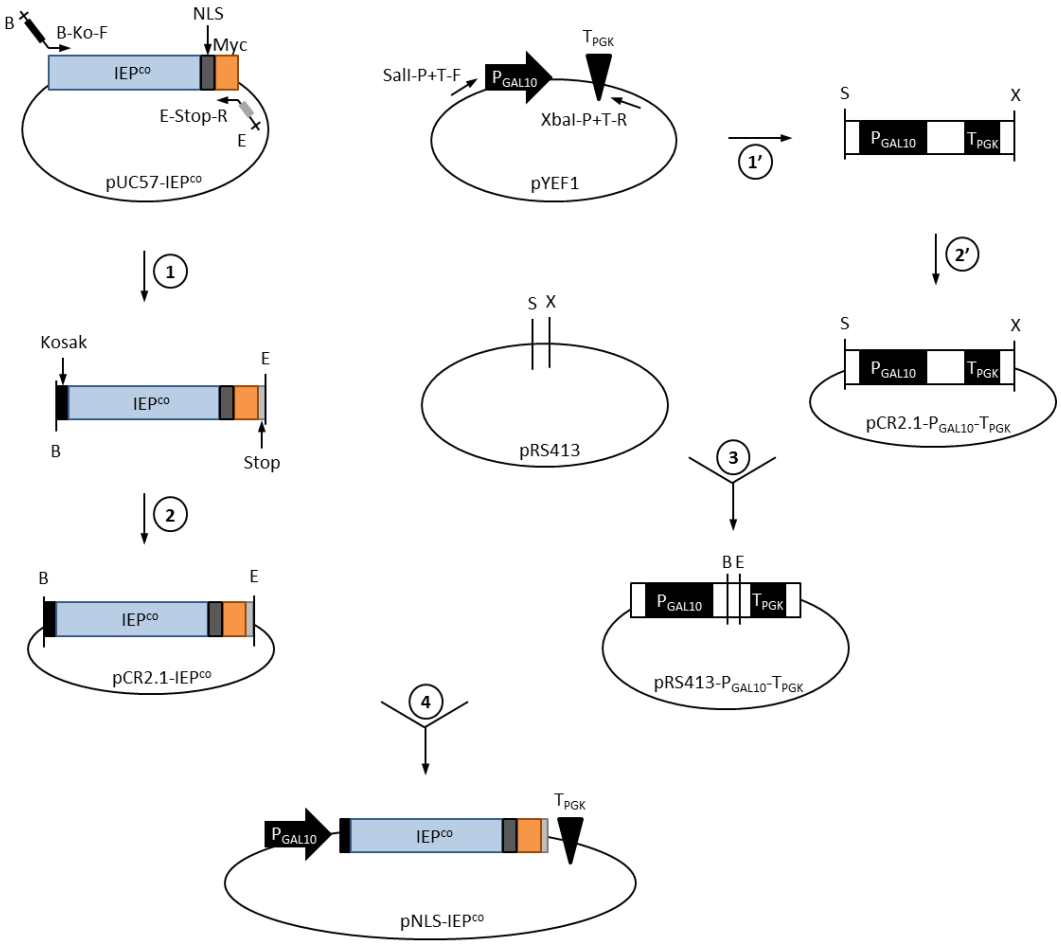


Figure M-6: Schematic representation of pNLS-IEP^{co} cloning.

The cloning strategy of pNLS-IEP^{co} is represented here and described below.

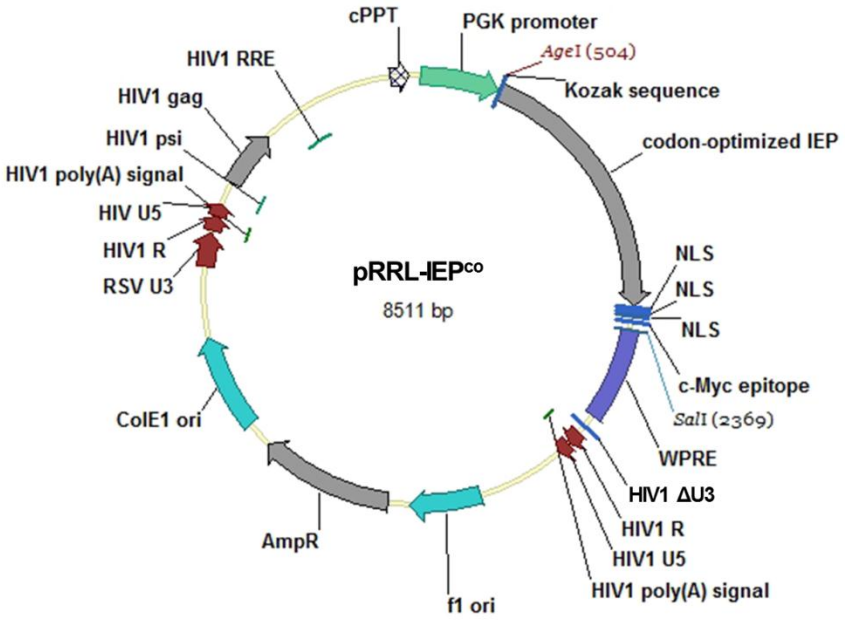
The IEP^{co} sequence followed by 3 nuclear localization signals (Fig. M-6; NLS, dark gray rectangle) and a c-Myc epitope (Fig. M-6; Myc, orange rectangle), located in the pUC57-IEP^{co} plasmid (In frame IEP, 3 NLS and c-Myc epitope sequences, codon-optimized for translation in human cells; Genescript; Appendix x), were amplified by PCR (Fig. M-6; step 1) using the oligonucleotides B-Ko-F, which contains a BamHI restriction site (Fig. M-6; B) and the Kosak sequence (Fig. M-6; black rectangle), and E-Stop-R, which contain an EcoRI restriction site (Fig. M-6; E) and a stop codon (Fig. M-6; light gray rectangle). The resulting PCR product was cloned into pCR2.1-TOPO by TOPO TA cloning (Fig. M-6; step 2). In parallel, the sequence of pYEF1 plasmid ((Cullin C and Minvielle-Sebastia L 1994); Appendix x) containing the GAL10 promoter (Fig. M-6; P_{GAL10}) and the PGK terminator (Fig. M-6; T_{PGK}) was amplified by PCR (Fig. M-6; step 1') using oligonucleotides SalI-P+T-F, which contains a SalI restriction site (Fig. M-6; S), and XbaI-P+T-R, which contains a XbaI restriction site (Fig. M-6; X). The resulting PCR product was cloned into pCR2.1-TOPO by TOPO TA cloning (Fig. M-6; step 2'). The resulting pCR2.1-P_{GAL10}-T_{PGK} was digested with SalI and XbaI and the restriction fragment containing the P_{GAL10} and T_{PGK} was ligated into the SalI/XbaI-digested pRS413 plasmid ((Sikorski RS and Hieter P 1989; Christianson TW et al. 1992); Appendix x) (Fig. M-6; step 3). The plasmid pCR2.1-IEP^{co} obtained after step 2 was digested with BamHI and EcoRI and the resulting restriction fragment containing the Kosak, IEP^{co}, NLS, c-Myc, and stop codon sequences was ligated into the BamHI/EcoRI-digested pRS413-P_{GAL10}-T_{PGK} plasmid (Fig. M-6; step 4). The resulting pNLS-IEP^{co} plasmid contains the codon-optimized IEP ORF sequence for translation in human cells downstream of a Kosak sequence and in frame with 3 NLS and a c-Myc epitope. This plasmid should express the NLS-IEP^{co} protein.

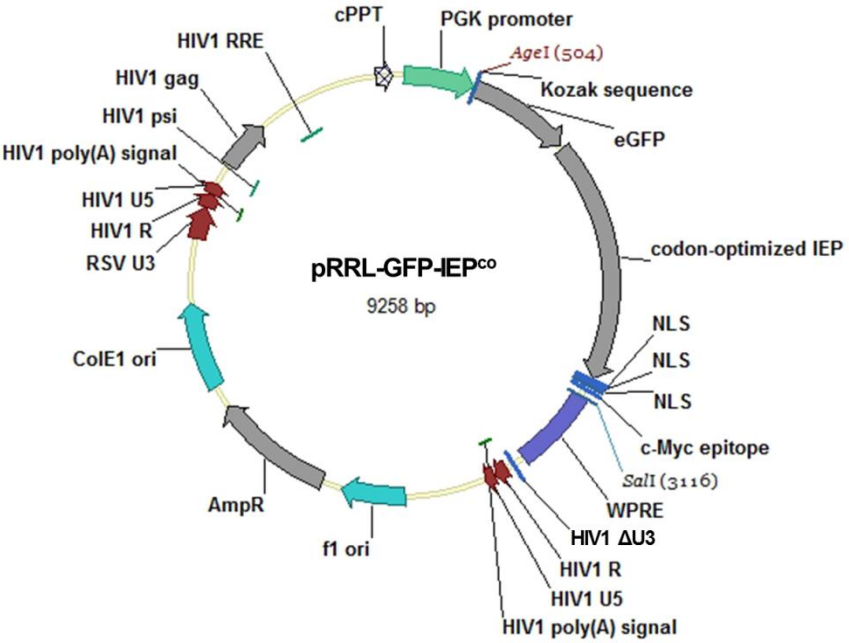
- pRRL-intron-ΔDIV / pRRL-intron-DIVa / pRRL-intron-DIVab / pRRL-intron-Full

figure pRRL-intron-…

These cloning were performed by collaborators. The plasmids pCINeo-E2-IntΔDIV-E3, pCINeo-E2-IntDIVa-E3 (Appendix x), pCINeo-E2-IntDIVab-E3 (Appendix x), and pCINeo-E2-IntFull-E3 (Appendix x), previously constructed by a collaborator, were digested with SalI and partially digested with PstI, and the restriction fragments corresponding to E2-IntΔDIV-E3, E2-IntDIVa-E3, E2-IntDIVab-E3, and E2-IntFull-E3, respectively, were isolated and ligated in the PstI/SalI-digested pRRLSINPPT-PGKmcs-WPRE plasmid ((Follenzi A et al. 2000; Charrier S et al. 2005; Charrier S et al. 2007); Appendix x), forming the pRRL-intron-ΔDIV, pRRL-intron-DIVa, pRRL-intron-DIVab, and pRRL-intron-Full plasmids, respectively. These plasmids correspond to gene transfer vectors for lentiviruses production, used to establish stable human cell lines expressing the intron ΔDIV, DIVa, DIVab or full length forms, flanked by its two exons.

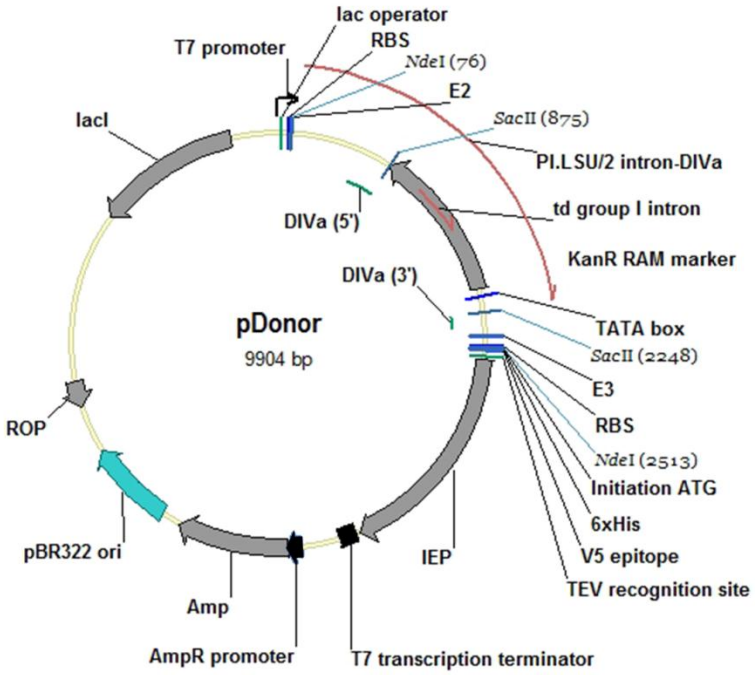
- pRRL-GFP / pRRL-GFP-IEP^{co}





These cloning were performed by collaborators. The plasmids pC1-IEP^{co} and pGFP-IEP^{co}, previously constructed by a collaborator (Appendix x and Appendix x), were digested with AgeI and MfeI. The restriction fragments containing the IEP^{co} and the GFP-IEP^{co} sequences, respectively, were isolated and partially digested with SalI. The restriction fragments containing the IEP^{co} and the GFP-IEP^{co} sequences, respectively, were isolated and ligated in the AgeI/SalI-digested pRRLSINPPT-PGKGFP-WPRE plasmid ((Follenzi A et al. 2000; Charrier S et al. 2005; Charrier S et al. 2007); Appendix x), forming the pRRL-GFP and pRRL-GFP-IEP^{co} plasmids, respectively. These plasmids correspond to gene transfer vectors for lentiviruses production, used to express the NLS-IEP^{co} fused or not in N-terminal to the GFP protein in stable human cell lines previously established.

- pDonor



The cloning of pDonor plasmid was performed in 5 steps (Fig. M-7).

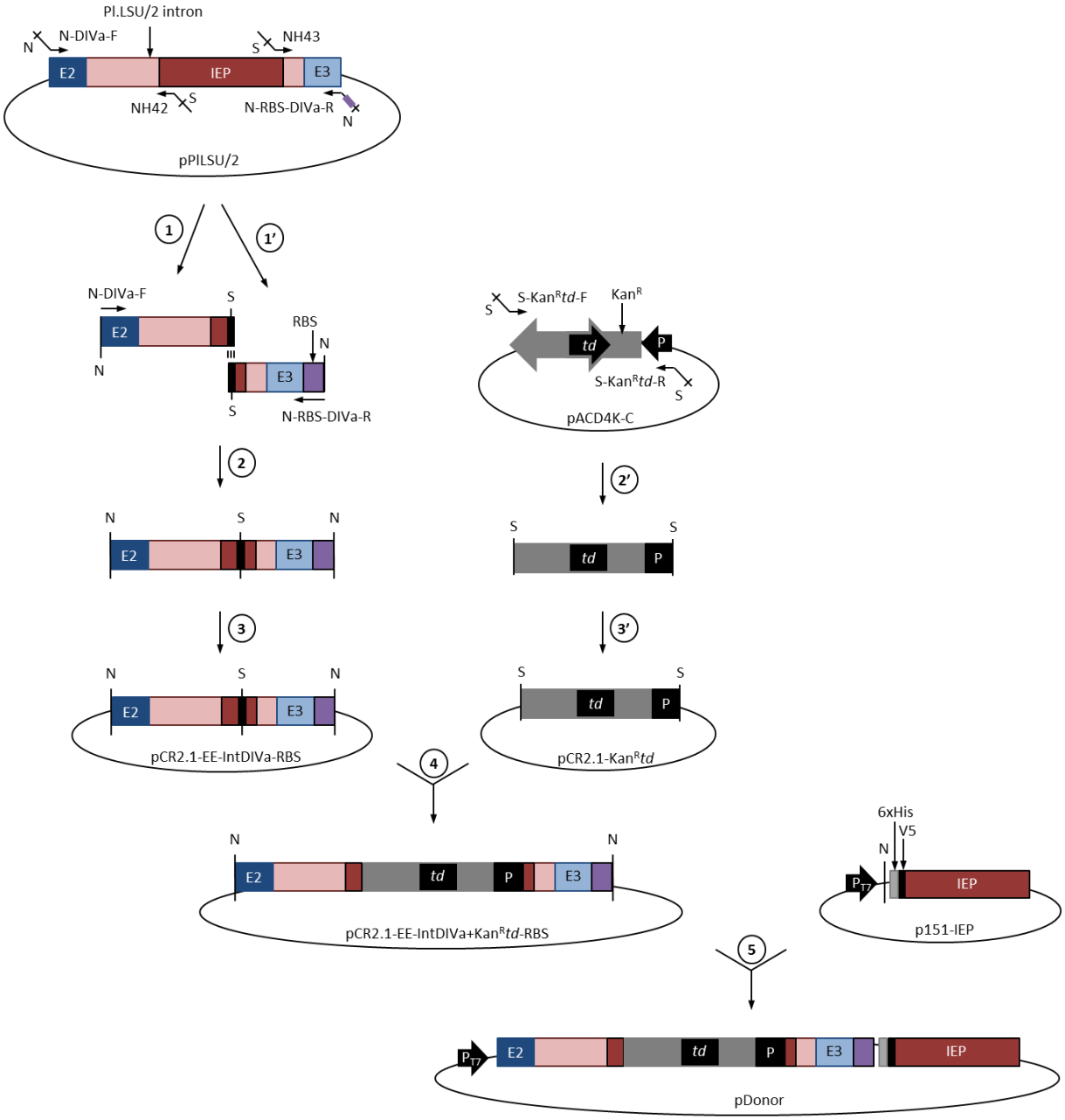
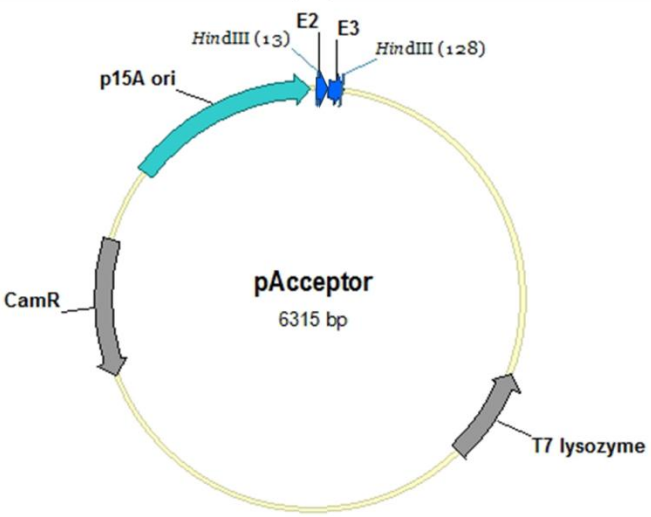


Figure M-7: Schematic representation of the pDonor cloning.
The cloning strategy of pDonor is represented here and described below.

The exon 2 sequence and the beginning of intron sequence (until the DIVa 5' region), located in pPl.LSU/2 plasmid, were amplified by PCR using oligonucleotides N-DIVa-F and NH42 (Fig. M-7; step 1). In parallel, the end of the intron (from the DIVa 3' region) and the exon 3 sequence were amplified using oligonucleotides NH43 and N-RBS-DIVa-R (Fig. M-7; step 1'). These two PCR products were mixed and amplified using N-DIVa-F and N-RBS-DIVa-R (Fig. M-7; step 2). The resulting PCR product was cloned into pCR2.1-TOPO by TOPO TA cloning (Fig. M-7; step 3), leading to the formation of the pCR2.1-EE-IntDIVa-RBS plasmid. In parallel, the Kan^R gene, containing the *td* group I intron sequence, and its promoter were amplified from the pACD4K-C plasmid (TargeTron® Gene Knockout System, Sigma Aldrich; Appendix x) using the oligonucleotides S-Kan^R*td*-F and S-Kan^R*td*-R (Fig. M-7; step 2'). The resulting PCR product was cloned into pCR2.1-

TOPO by TOPO TA cloning (Fig. M-7; step 3'), leading to the formation of the pCR2.1-Kan^R*td* plasmid. The Kan^R*td* and its promoter sequences were isolated by SacII (Fig. M-7; S) digestion and ligated in the SacII-digested pCR2.1-EE-IntDIVa-RBS plasmid (Fig. M-7; step 4). The restriction fragment containing E2, Intron DIVa with Kan^R*td* and its promoter, E3, and a RBS, obtained by NdeI (Fig. M-7; N) digestion of the pCR2.1-EE-IntDIVa+Kan^R*td*-RBS plasmid, was ligated in the NdeI-digested p151-IEP plasmid (Fig. M-7; step 5), leading to the formation of the pDonor plasmid.

- pAcceptor



The cloning of pAcceptor plasmid was performed in 3 steps (Fig. M-8).

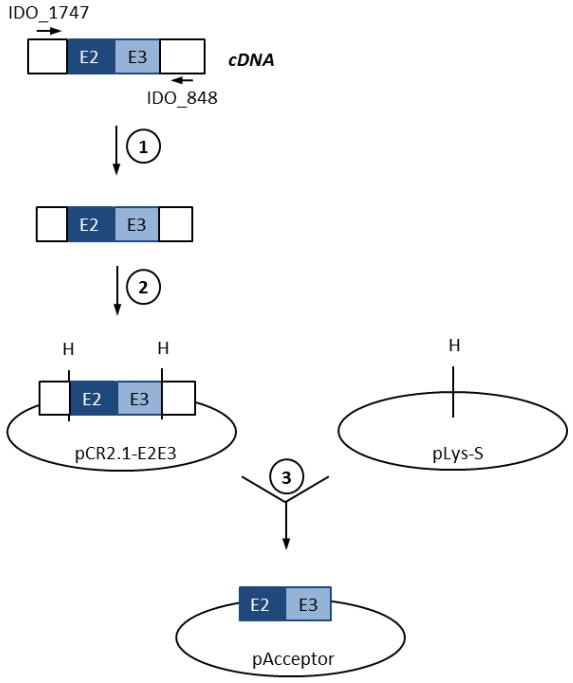
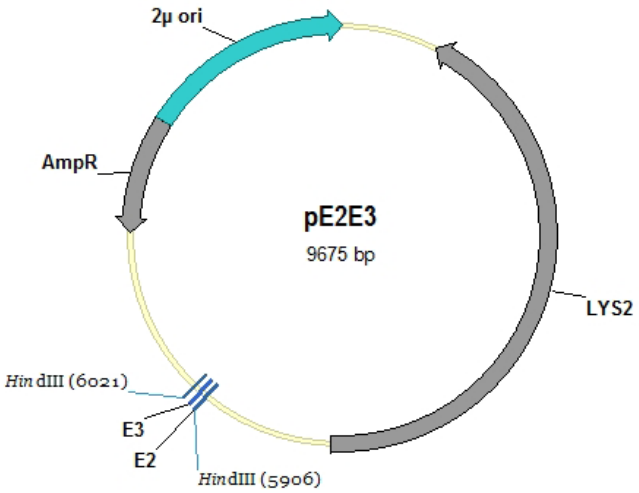


Figure M-8: Schematic representation of the pAcceptor cloning.
The cloning strategy of pAcceptor is represented here and described below.

The sequences of exon 2 and exon 3 was amplified by PCR using oligonucleotides IDO_1747 and IDO_848 on spliced cDNA template, obtained by *in vitro* splicing of the Pl.LSU/2 intron from precursor cDNA transcribed in HEK 293 cells transfected with pCINeo-E2-IntΔDIV-E3 plasmid (Fig. M-8; step 1). The resulting PCR product was cloned into pCR2.1-TOPO by TOPO TA cloning (Fig. M-8; step 2), leading to the pCR2.1-E2E3 plasmid. The fragment containing E2 and E3 was isolated from pCR2.1-E2E3 by digestion with HindIII (Fig. M-8; H) and ligated into the HindIII-digested pLysS plasmid (Fig. M-8; step 3) (Appendix x), leading to the pAcceptor plasmid.

- pE2E3



The fragment containing E2 and E3 was isolated from pCR2.1-E2E3 by digestion with HindIII (Fig. M-8; H) and ligated into the HindIII-digested pRS327 plasmid ((Eriksson P et al. 2004); Appendix x), leading to the pE2E3 plasmid.

2.4 - PCR AMPLIFICATIONS

- PCR amplifications of fragments for cloning were performed using 20 ng of plasmid DNA template, 2 units of Phusion® High-fidelity DNA polymerase (Thermo Scientific; Finnzymes), 0.5 µM of each primer, 400 µM of each dNTPs, and 1X Phusion HF buffer (supplied with the enzyme) in a final volume of 50 µl. Initial denaturation was performed at 98°C for 1 min and followed by 30 amplification cycles consisting of denaturation at 98°C for 10 sec, primer annealing at *X*°C for 20 sec (the annealing *X* temperature was adjusted to 3°C higher than the lower primer melting temperature), extension at 72°C for 20 sec/kb, and a final extension step of 5 min at 72°C. PCR products were incubated with 1 unit of *Dpn* I restriction enzyme at 37°C for 1 hr to digest the plasmid DNA, and gel purified after electrophoresis on agarose gel.

For TOPO TA cloning, addition of dATPs at 3' ends of PCR products was performed using 20 µl of purified PCR products, 2.5 units of AmpliTaq Gold DNA polymerase, and 3 mM of dATPs in 1X PCR Buffer II (supplied with the enzyme) in a final volume of 30 µl at 72°C for 15 min.

- Screening of recombinant *E. coli* colonies and analysis of purified recombinant plasmid DNA by PCR were performed using 1 unit of RedTaq DNA polymerase (Sigma Aldrich), 1X reaction buffer (supplied with the enzyme), 200 µM of each dNTPs, 0.2 µM of each primer, and either a bacterial colony picked from an LB-agar plate or 10 ng of plasmid DNA in a final volume of 50 µl. Initial

denaturation was performed at 94°C for 5 min and followed by 30 amplification cycles consisting of denaturation at 94°C for 40 sec, primer annealing at *X*°C for 1 min (the annealing *X* temperature was adjusted to 5°C lower than the lower primer melting temperature), extension at 72°C for 1 min/kb, and a final extension step of 10 min at 72°C.

3 - PROTEIN EXPRESSION, PURIFICATION AND ANALYSES

3.1 - PROTEIN EXPRESSION

3.1.1 - Cell-free expression system

The Expressway Cell-Free *E. coli* Expression System (Life Technologies; Invitrogen) was used to express the Pl.LSU/2 IEP *in vitro*. The plasmid p151-IEP was used as DNA template. It was initially designed and constructed for the expression of Pl.LSU/2 IEP tagged in N-terminal to a 6xHis tag and a V5 epitope (HisV5-IEP) in *E. coli*. However, the plasmid configuration is close to the recommended characteristics (Fig. M-9). Indeed, it presents required features, regardless to the position of the ribosome binding site (RBS), HisV5-IEP encoding sequence, and T7 terminator (Fig. M-9). The only exception is the length between the promoter and the RBS (Fig M-9; 57 nt instead of recommended 15-20 nt). However, the length between promoter and RBS in the control pEXP5-NT/CALML3 template is similar (Fig. M-9; 51 nt). The p151-IEP plasmid also differs from recommended template design by the presence of a *lac* operator (*lac*O), which should not have any influence on the protein synthesis *in vitro*.

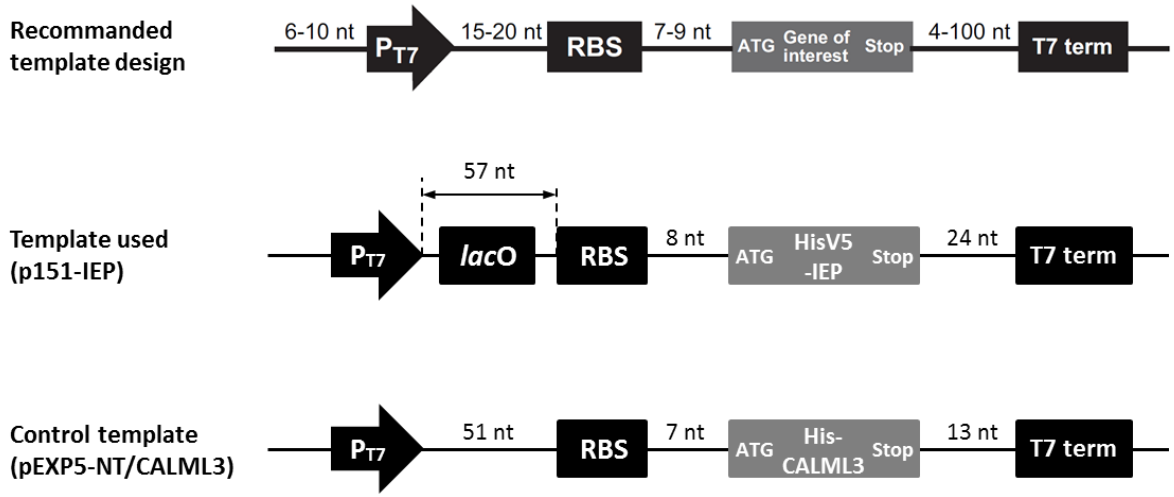


Figure M-9: Template design for Cell-Free expression system.

The manufacturer recommended template design for optimal protein synthesis is indicated. The recommended lengths between promoter, RBS, gene of interest and terminator are indicated. The templates used to express Pl.LSU/2 IEP (p151-IEP) and the control protein CALML3 are represented.

Template DNA was prepared according to the manufacturer's instructions: a sample of previously purified p151-IEP plasmid DNA was precipitated using RNase-free reagents and consumables. Template DNA concentration was set at 500 ng/µl in RNase-free DEPC-treated water.

In vitro protein synthesis was performed according to the manufacturer's instructions using 1 µg of DNA template in a thermomixer at 1,200 rpm for 30 min at 30°C, before adding the Feed buffer and continuing the incubation at 1,200 rpm for 5 hrs 30 min at 30°C. Proteins were then precipitated with acetone and mixed with 3.3X protein sample buffer (composed by 5/6 of 4X XT sample buffer and 1/6 of 20X XT reducing agent; Bio-Rad). Protein samples were heated at 95°C for 5 min and stored at -20°C until SDS-PAGE analysis.

3.1.2 - Insect cell expression

(a) Production of recombinant bacmids

DH10Bac *E. coli* cells were transformed with either pFastBacHT-IEP or pFastBacHT-CAT. Clones selected by blue/white screening were analyzed by PCR using M13(-40)F and BacSeqControl oligonucleotides to determine successful transposition of His-IEP or His-CAT expressing cassettes, and recombinant bacmids were purified from positive clones.

(b) Production of recombinant baculoviruses

- P1 baculovirus stock:

1 x 10^6 Sf9 cells per well were seeded in a 6-well plate in 2 ml of Sf-900 II SFM. The cells were incubated for 1 hr at 27°C to allow the cell adherence before adding 0.8 ml of Sf-900 II SFM containing bacmid DNA:lipid complexes. Bacmid DNA:lipid complexes were prepared by mixing 1 µg of bacmid DNA diluted in 100 µl of Sf-900 II SFM with 9 µl of Cellfectin I reagent (Life Technologies, Invitrogen) diluted in 100 µl of Sf-900 II SFM followed by incubation at room temperature for 30 min. Cells were then incubated at 27°C for 5 hrs, and the transfection medium was replaced by 2 ml of Sf-900 II SFM before continue the incubation at 27°C for 96 hrs. P1 baculoviral stock was then collected by harvesting the medium which was clarified by centrifugation at 500*g* for 5 min.

- Isolation of baculovirus from a single viral clone:

1 x 10^6 Sf9 cells per well were seeded in a 6-well plate in 1 ml of Sf-900 II SFM and incubated at 27°C for 1 hr. 200 µl of P1 viral stock diluted at 10^{-5} to 10^{-7} in Sf-900 II SFM were added to the cells followed by incubation at room temperature for 3 hrs. The medium was then removed and 2 ml of plaquing medium (4% solubilized agarose gel and 1X final Sf-900 II SFM) was added before incubation at room temperature for 20 min. The cells were then placed in a sealed plastic box with moist paper towels and incubated at 27°C for 8 days, or until plaques are formed. Baculovirus from a single clone was purified by collecting agarose from an isolated plaque, and resuspended in 500 µl of Sf-900 II SFM.

- Generation of the baculovirus P1p stock:

7.5 x 10^6 Sf9 cells seeded in a T75 cell culture flask in 10 ml of Sf-900 II SFM were infected with 250 µl of previously isolated baculovirus and incubated at 27°C for 72 °C. P1p baculoviral stock was then collected by harvesting the medium which was clarified by centrifugation at 500*g* for 5 min. P1p stock was titrated as described in section thereafter.

- Generation of the baculovirus P2 stock:

400 ml of Sf-900 II SFM were inoculated with Sf9 cells at 1 x 10^6 cells/ml in 500 ml spinner and infected with P1p baculovirus stock at MOI 0.1. Cells were cultured in suspension for 72 hrs at 27°C. P2 baculoviral stock was then recovered by collecting the medium which was clarified by centrifugation at 1,300*g* for 15 min, and filtrated on 0.45 µm cellulose acetate filter. P2 stock was titrated as described in section thereafter.

- Titration of baculovirus stocks:

1 x 10^6 Sf9 cells per well were seeded in a 6-well plate in 1 ml of Sf-900 II SFM and incubated at 27°C for 1 hr. 100 µl of P1p or P2 viral stocks diluted at 10^{-7} to 10^{-9} in Sf-900 II SFM were added to the cells followed by incubation at room temperature for 3 hrs. The medium was then removed and 2 ml of plaquing medium (4% solubilized agarose gel and 1X final Sf-900 II SFM) were added before incubation at room temperature for 20 min. The cells were then placed in a sealed plastic box with moist paper towels and incubated at 27°C for 8 days, or until plaques are formed. Viral titers (pfu/ml; plaque forming units/ml) were then determined by counting the number of plaques for each dilution and using the following formula:

$$baculovirus\ stock\ titer = \frac{number\ of\ plaques\ (pfu)}{dilution\ factor\ \times\ volume\ (ml)\ of\ inoculum\ used}$$

(c) ***Expression of proteins***

200 ml of Sf-900 II SFM were inoculated with Sf9 cells at 1.09 x 10^6 cells/ml in a 500 ml spinner and infected with P2 baculovirus stock at MOI 5. Cells were cultured in suspension for 72 hrs at 27°C, pelleted by centrifugation at 1,300*g* for 10 min, and pellets were stored at -80°C overnight.

3.1.3 - Bacterial expression

E. coli strain BL21 Star (DE3) was transformed with the appropriate expression plasmid and three single colonies were inoculated into 5 ml of LB containing appropriate antibiotics. Precultures were shaken at 37°C overnight, inoculated into 110 ml of LB medium with antibiotics and grown at 37°C for 3-6 hrs, until OD_{600} reached 0.5. Induction was started by addition of IPTG (0.1 mM final) to 100 ml of the culture, and 10 ml of the culture was non-induced. The incubation was continued for 3 hrs at 37°C. Cells were then collected by centrifugation (1,900*g* for 10 min at 4°C), and washed once with 1X PBS. The washed cell pellet was stored at -80°C overnight before protein extraction.

E. coli strain BL21 Star (DE3) pRARE was transformed with the appropriate expression plasmid and three single colonies were inoculated into 5 ml of LB containing appropriate antibiotics. Precultures were shaken at 37°C overnight, inoculated into 10-210 ml of LB medium with antibiotics and grown at 37°C for 3-6 hrs, until OD_{600} reached 0.5-0.7. Cultures were cooled or not at 18°C for 20 min, and induction were started by addition of IPTG (0.1-1 mM final) to 2-200 ml of the culture. As a control, 2-50 ml of the culture was non-induced. The incubation was continued for 3 hrs at 18°C, 32°C, or 37°C. Cells were then collected by centrifugation (1,900*g* for 10 min at 4°C), and washed once with 1X PBS. The washed cell pellet was stored at -80°C overnight before protein extraction.

E. coli strain ArcticExpress (DE3)RIL was transformed with the appropriate expression plasmid and three single colonies were inoculated into 10 ml of LB containing appropriate antibiotics. Precultures were shaken at 37°C overnight, inoculated into 410 ml of LB medium without antibiotics and grown at 30°C for 3-6 hrs, until OD_{600} reached 0.6. Cultures were cooled or not at 15°C for 25 min, and induction was started by addition of IPTG (0.1 mM final) to 400 ml of the culture. 10 ml of the culture were non-induced. The incubation was continued for 18 hrs at 15°C. Cells were then collected by centrifugation (1,900*g* for 10 min at 4°C), and washed once with 1X PBS. The washed cell pellet was stored at -80°C overnight before protein extraction.

E. coli strain Rosetta-gami B (DE3) was transformed with the appropriate expression plasmid and three single colonies were inoculated into 5 ml of LB containing appropriate antibiotics. Precultures were shaken at 32°C overnight, inoculated into 105 ml of LB medium without antibiotics and grown at 32°C for 3-6 hrs, until OD_{600} reached 0.6-0.8. Induction was started by addition of IPTG (1 mM final) to 100 ml of the culture, and 5 ml of the culture were non-induced as a control. The incubation was continued for 4 hrs at 30°C. Cells were then collected by centrifugation (1,900*g* for 10 min at 4°C), and washed once with 1X PBS. The washed cell pellet was stored at -80°C overnight before protein extraction.

3.2 - PROTEIN EXTRACTION

All total and soluble protein samples collected during protein extractions were mixed with 3.3 X protein sample buffer (composed by 5/6 of 4X XT sample buffer and 1/6 of 20X XT reducing agent; Bio-Rad). All insoluble protein samples collected were mixed with 1X protein sample buffer in 1/5 of the volume that was used to resuspend the original cell pellet. All protein samples were heated at 95°C for 5 min and stored at -20°C until SDS-PAGE analysis. The remaining soluble protein fractions were subsequently used for protein purification.

3.2.1 - Sf9 protein extraction

Sf9 cell pellet was washed in ice-cold 1X Phosphate buffered saline (PBS) and centrifuged at 1,300*g* for 10 min. Cells were then resuspended in 8 ml of ice-cold TEN buffer (50 mM Tris pH 7.5 and 100 mM NaCl), and incubated 5 min at 4°C before being centrifuged as before. The pellet was then resuspended in 8 ml of ice-cold 250 mM Tris pH 7.5 supplemented with protease inhibitor cocktail (cOmplete EDTA-free protease inhibitor cocktail, Roche Applied Science) and cells were lysed by three cycles of freeze/thawing between -180°C and 37°C. A sample of total protein fraction was collected before pelleting cell debris and insoluble proteins by centrifugation at 10,000 rpm on a microcentrifuge for 10 min at 4°C. A sample of soluble protein fraction from the supernatant was subsequently collected and the pellet containing insoluble proteins was also conserved for SDS-PAGE analysis.

3.2.2 - Bacterial protein extraction

BL21 Star (DE3) and BL21 Star (DE3) pRARE cell pellets were thawed on ice and resuspended in 1 volume of BugBuster® Master mix (Novagen) supplemented with protease inhibitor cocktail for 20 volumes of *E. coli* pelleted culture. The lysis was performed at room temperature for 20 min on a rotary wheel before replacing the lysate at 4°C. A sample of total protein fraction was collected before pelleting cell debris and insoluble proteins by centrifugation at 16,000*g* for 20 min at 4°C. A sample of soluble protein fraction from the supernatant was subsequently collected and the pellet containing insoluble proteins was also conserved for SDS-PAGE analysis.

For His-tagged protein purification under denaturing conditions, BL21 Star (DE3) pRARE cell pellets were thawed on ice and resuspended in 1 volume of buffer M (50 mM Tris-HCl at pH 8.0, 1 M NaCl) with either 6 M Guanidine hydrochloride (GuHCl) or 8 M Urea for 20 volumes of *E. coli* pelleted culture. After incubation at room temperature for 10 min on a rotary shaker, cells were lysed by sonication (3 cycles of 6 sec “on” at 200 W and 6 sec “off”). A sample of total protein fraction was collected before pelleting cell debris and insoluble proteins by centrifugation at 16,000*g* for 20 min at

4°C. A sample of soluble protein fraction from the supernatant was subsequently collected and the pellet containing insoluble proteins was also conserved for SDS-PAGE analysis. Total and soluble protein samples containing guanidine hydrochloride were precipitated (See following section) before being mixed with protein sample buffer.

ArcticExpress (DE3)RIL cell pellets were thawed on ice and resuspended in 1 volume of ice-cold buffer Mb (50 mM Tris-HCl at pH 8.0, 1 M NaCl, and 4 mM β-mercaptoethanol) supplemented with protease inhibitor cocktail for 30 volumes of *E. coli* pelleted culture. Lysozyme (BioUltra, Sigma Aldrich) was added to a final concentration of 1 mg/ml, and after 45 min at 4°C cells were lysed by sonication (6 cycles of 12 sec "on" at 200 W and 12 sec "off"). RNase A and DNase I were added at 10 µg/ml and 5 µg/ml respectively, followed by incubation at 4°C for 15 min. A sample of total protein fraction was collected before pelleting cell debris and insoluble proteins by centrifugation at 16,000*g* for 20 min at 4°C. A sample of soluble protein fraction from the supernatant was subsequently collected and the pellet containing insoluble proteins was also conserved for SDS-PAGE analysis.

Rosetta-gami B (DE3) cell pellets were processed as for ArcticExpress (DE3)RIL with two differences: the volume of buffer used to resuspend the cell pellet was 1/80 of the *E. coli* pelleted culture volume, and for lysates that are used for protein purification under non-denaturing conditions in presence of CHAPS, the Mb buffer was supplemented with 10 mM of CHAPS (MbC buffer).

3.3 - PROTEIN PRECIPITATION

Proteins were precipitated with 2.5 volumes of acetone for 1 hr at 4°C and recovered by centrifugation at 18,000*g* for 15 min at 4°C. Protein pellet was then washed with 100% ethanol, dried, and resuspended in 0.5 volumes of 50 mM Tris-HCl at pH 8.5, 8 M urea, and 10 mM DTT.

3.4 - PROTEIN PURIFICATION BY AFFINITY CHROMATOGRAPHY

During all purifications, samples were collected at each step for SDS-PAGE analysis. All samples were mixed in 3.3X protein sample buffer, except the resin sample after elution(s) that was dried using gel-loading pipet tips and resuspended in one bed volume of 1X protein sample buffer. Protein samples were heated at 95°C for 5 min and stored at -20°C until SDS-PAGE analysis.

3.4.1 - GST-tagged protein purification

Soluble protein fractions were incubated with glutathione-charged agarose beads (Glutathione Sepharose® 4B, GE Healthcare) using a bed volume corresponding to either 1/100 or 1/30 of the lysate volume. The binding of proteins was performed for either 30 min at room temperature or overnight at 4°C on a rotary wheel. Beads were then washed three times with 10-20 bed volumes of ice-cold 1X PBS supplemented with 1 mM PMSF. GST-tagged proteins were eluted by incubation either for 15 min at room temperature or overnight at 4°C on a rotary shaker with one bed volume of either ice-cold buffer Gst1 (100 mM HEPES pH 8.0 and 10 mM reduced glutathione), ice-cold buffer Gst5 (100 mM HEPES pH 8.0 and 50 mM reduced glutathione), or ice-cold buffer Gst5N (100 mM HEPES pH 8.0, 120 mM NaCl, and 50 mM reduced glutathione), supplemented with protease inhibitor cocktail. Eluted protein fraction was dialyzed or not using dialysis cassette (Slide-A-Lyzer dialysis cassettes, 20K MWCO; Thermo Scientific, Pierce) twice for 2 hrs and then overnight at 4°C against 500 ml of buffer 100 mM HEPES pH 8.0, 120 mM NaCl, and 1 mM PMSF. Eluted and/or

dialyzed protein fractions were stored in 50% (v/v) glycerol with 1 mM DTT and 0.1 mM EDTA at -20°C.

3.4.2 - Histidine-tagged protein and RNP particles purification

- Native and non-denaturing conditions:

Soluble protein fractions from BL21 Star (DE3) pRARE were incubated with Ni^{2+}-charged agarose beads (HIS-Select® HF Nickel affinity gel, Sigma Aldrich) using a bed volume corresponding to 1/30 of the lysate volume. The binding of proteins was performed for 15 min at 4°C on a rotary wheel. Beads were then washed twice with 20 bed volumes of ice-cold buffer EW (50 mM Sodium phosphate, pH 8.0, 300 mM NaCl, and 10 mM imidazole) supplemented with protease inhibitor cocktail. His-tagged IEP was eluted by incubation 15 min at 4°C on a rotary shaker with 2 bed volumes of ice-cold buffer E (50 mM Sodium phosphate, pH 8.0, 300 mM NaCl, and 250 mM imidazole) supplemented with protease inhibitor cocktail. Eluted protein fraction was stored in 50% (v/v) glycerol with 1 mM DTT and 0.1 mM EDTA at -20°C.

Soluble protein fractions from Sf9 pellets were incubated on-column with Ni^{2+}-charged agarose beads (Ni-NTA agarose, Life Technologies) using a bed volume corresponding to 1/16 of the lysate volume. Binding was performed for 2 hrs at 4°C on a rotary wheel. The columns were then washed four times with 16 bed volumes of ice-cold buffer N (50 mM NaH2PO4 pH 8.0, 500 mM NaCl) with 20 mM imidazole. Elution was performed at 4°C using 3.5 ml of ice-cold buffer N with 250 mM imidazole. Seven 500 µl fractions of eluted proteins were collected, and stored in 50% (v/v) glycerol with 1 mM DTT and 0.1 mM EDTA at -20°C.

Soluble protein fractions from Rosetta-gami B (DE3) were incubated with Ni^{2+}-charged agarose beads (Ni-NTA agarose, Qiagen) using a bed volume corresponding to 1/12 of the lysate volume. The binding was performed 4°C overnight on a rotary wheel. Beads were then washed once with 15 bed volumes of ice-cold buffer Mb (for natives conditions) or MbC (for non-denaturing conditions) with 1 mM PMSF, then three times with 15 bed volumes of ice-cold buffer Mb or MbC with 1 mM PMSF and 20 mM imidazole, twice with 15 bed volumes ice-cold buffer Mb or MbC with 1 mM PMSF and 40 mM imidazole and finally once with 15 bed volumes with ice-cold buffer Mb or MbC with 1 mM PMSF and 60 mM imidazole. His-tagged IEP was eluted by incubation overnight at 4°C on a rotary wheel with one bed volume of ice-cold buffer Mb or MbC supplemented with protease inhibitor cocktail and 1 M imidazole. Eluted protein fractions were then dialyzed using dialysis cassette (Slide-A-Lyzer dialysis cassettes, 10K MWCO; Thermo Scientific, Pierce) twice for 2 hrs and then overnight at 4°C against 500 ml of buffer Mb with 1 mM PMSF. Proteins were finally stored in 50% (v/v) glycerol with 1 mM DTT and 0.1 mM EDTA at -20°C.

- Denaturing conditions:

Soluble protein fraction from BL21 Star (DE3) pRARE was incubated with Ni^{2+}-charged agarose beads (Ni-NTA agarose, Life Technologies) using a bed volume corresponding to 1/8 of the lysate volume. Binding was performed for 30 min at room temperature on a rotary wheel. Beads were then washed once with 10 volumes of ice-cold buffer M with 1 mM PMSF and either 6 M GuHCl (for GuHCl denaturing conditions) or 8 M Urea (for urea denaturing conditions); three times with 10 volumes of ice-cold buffer M with 1 mM PMSF, 20 mM imidazole, and 6 M GuHCl or 8 M Urea; twice with 10 volumes ice-cold buffer M with 1 mM PMSF, 40 mM imidazole, and 6 M GuHCl or 8

M Urea, and finally once with 10 volumes with ice-cold buffer M with 1 mM PMSF, 60 mM imidazole, and 6 M GuHCl or 8 M Urea. His-tagged IEP was eluted by incubation overnight at 4°C on a rotary shaker with one volume of ice-cold buffer M supplemented with protease inhibitor cocktail, 1 M imidazole, and 6 M GuHCl or 8 M Urea. Eluted protein fractions were then subjected to a refolding step using a multi-step dialysis process: eluted fractions were dialyzed using dialysis cassettes at 4°C for 2 hrs against 500 ml of buffer Gu_{-1} (M with 5 M GuHCl) or buffer U_{-1} (M with 7 M Urea) with 1 mM PMSF. First dialysis buffer was then replaced by buffer Gu_{-2} (M with 4 M GuHCl) or buffer U_{-2} (M with 6 M Urea) with 1 mM PMSF and dialysis was continued for 2 hrs at 4°C. This process was reproduced with Gu_{-3} (M with 3 M GuHCl) to Gu_{-5} (M with 1 M GuHCl) and with U_{-3} (M with 5 M Urea) to U_{-7} (M with 1 M Urea), and a final dialysis at 4°C overnight was performed in 500 ml of M buffer with 1 mM PMSF. Proteins were finally stored in 50% (v/v) glycerol with 1 mM DTT and 0.1 mM EDTA at -20°C.

3.5 - PROTEIN ANALYSES

3.5.1 - Protein SDS-PAGE gel staining

Proteins were resolved on a denaturing 10% polyacrylamide-SDS gel (Criterion XT Bis-Tris gel, Bio-Rad) and the Precision Plus Protein All Blue standards (Bio-Rad) was added as a standard. Proteins were separated by electrophoresis in 1X MOPS buffer (XT MOPS Running buffer; Bio-Rad) at 150V. The gel was then washed three times in Milli-Q water and proteins were colored with Coomassie blue stain (Bio-Safe Coomassie stain; Bio-Rad) for 1 hr at room temperature with gentle agitation. Stain was removed and the gel was washed twice for 1 hr at room temperature and once overnight at room temperature with Milli-Q water. Coomassie blue stained protein gels were finally dried on a filter paper (Whatman) using a vacuum gel dryer for 2 hrs at 65°C.

3.5.2 - Western blotting

Proteins were resolved on a denaturing 10% polyacrylamide-SDS gel (Criterion XT Bis-Tris gels, Biorad) and transferred onto a nitrocellulose membrane (Hybond ECL, Amersham) in 20 mM Tris, 150 mM Glycine and 20% (v/v) ethanol at 400 mA for 2 hrs at 4°C.

Immunoblots of GST-tagged proteins were blocked in 5% (w/v) non-fat dried milk in PBST (1X PBS containing 0.1% (v/v) Tween-20) overnight at 4°C. Membranes were probed with primary HRP-conjugate mouse anti-GST antibody (1:25,000 dilution, Amersham) in 5% (w/v) non-fat dried milk in PBST for 1 hr at room temperature, and washed three times with PBST for 10 min and two times with 1X PBS for 10 min at room temperature. Immunoreactive bands were detected by using the Super signal West Dura chemiluminescent substrate (Pierce), according to the manufacturer's instructions and revealed on X-ray autoradiography Hyperfilm ECL (Amersham).

Immunoblots of HisV5-tagged proteins were blocked in Odyssey® blocking buffer (LI-COR) overnight at 4°C. Membranes were probed with primary mouse anti-V5 antibody (1:5,000 dilution, Life technologies, Invitrogen) in Odyssey blocking buffer supplemented with 0.1% (v/v) Tween-20 for 1 hr at room temperature, and washed three times with PBST for 10 min. Membranes were then probed with secondary IRDye 680-conjugated goat anti-mouse antibody (1:8,000 dilution, LI-COR Biosciences) in Odyssey blocking buffer supplemented with 0.1% (v/v) Tween-20 and 0.01% (v/v) SDS for 45 min at room temperature, and washed three times with PBST for 10 min and two times

with 1X PBS for 10 min. Immunoreactive bands were detected with the Odyssey infrared scanner (LI-COR).

3.5.3 - Mass spectrometry

His-tagged purified protein fractions obtained by both IMAC and centrifugation in sucrose gradient were separated on SDS-PAGE gel, followed by coloration with Coomassie blue stain. The 69-kDa band detected in samples purified by sucrose gradient, and both 69-kDa and 75-kDa bands detected in samples purified by IMAC were excised from the gel using a scalpel and washed first in 500 µl of methanol for 10 min, then in 500 µl of Milli-Q water for 5 min, and finally in 500 µl of methanol for 10 min. Proteins were then digested in 100 mM NH_4CO_3 pH 7.9 with 5 ng/µl of trypsin (Sequence Grade Trypsin, Promega) in a final volume of 30 µl at 37°C overnight. Peptides were desalted and concentrated by using ZipTip® µC18 (Millipore) pre-equilibrated in 50% (v/v) acetonitrile and 0.5 % formic acid (v/v) and washed with formic acid 1% (v/v). Peptides were eluted in 1 µl of 80% (v/v) acetonitrile and 0.2% (v/v) formic acid containing 3 mg/ml of α-cyano-4-hydroxy-cinnamic acid (HCCA) matrix, and spotted on a metallic MALDI target plate (Applied Biosystems). Tandem mass spectra were acquired in reflectron mode with acceleration voltage of 25 kV on a 4800 *Plus* MALDI-TOF/TOF™ analyzer (AB SCIEX) equipped with a YAG 200-Hz laser (355 nm). MS and MS/MS data sets were retrieved using the Data Explorer® software (Applied Biosystems), and trypsin autolysis products were used to calibrate spectra. Data sets analyses are described in section 4 -.

3.5.4 - Reverse transcriptase activity assay

RT activity was assayed at 37°C in 14 µl of reaction medium containing 10 mM KCl, 10 mM $MgCl_2$, 50 mM Tris-HCl at pH 8.0, 5 mM DTT, 0.05% NP40, 1 µg of poly(rA)-oligo(dT)12-18 (Amersham), and 10 µCi of [α-32P]dTTP (3,000 Ci/mmole, Perkin-Elmer). Reactions were started by the addition of either 5-8 µl of purified protein sample, 0.03-0.5 units of SuperScript II reverse transcriptase (Life Technologies, Invitrogen) diluted in equivalent volume of the same buffer than those in which are stored purified proteins used in the assay, or 5-8 µl of the buffer in which purified proteins are stored. Reactions were performed for 10-120 min and stopped by addition of EDTA (50 mM final). Radioactive products were spotted on a DE81 filter (Whatman) which was washed twice in 2X SSC. After drying the membrane and expose it overnight on a phosphor screen (Molecular Dynamics PhosphorImager System; GE Healthcare Bio-Sciences), radioactive spots were detected by the Storm system (GE-Healthcare Bio-Sciences) and data were analyzed with ImageQuant software (GE Healthcare Life Sciences).

RT activity of His-tagged proteins contained in RNP particles were assayed as described above with either 0.1 OD_{260} units of RNPs, the equivalent volume of buffer in which RNPs are stored, or 0.03 units of SuperScript II reverse transcriptase diluted in equivalent volume of the buffer in which RNPs are stored. Reactions were performed in 14 µl of previously described reaction medium supplemented or not with 1 µg of RNase A (Sigma Aldrich).

4 - BIOINFORMATICS

The Pl.LSU/2 IEP amino acid sequence was used as a query in a BLASP (version 2.2.26) research against the complete genome sequence of *E. coli* BL21 (DE3). The expected threshold was settled at 10, and the filter of low complexity regions was applied.

ClustalW (version 2.0.12) was used to align amino acid sequences of Pl.LSU/2 IEP and retron EC86 reverse transcriptase with default settings.

The tridimensional structure of GST-IEP and HisV5-IEP were predicted using the I-TASSER server (http://zhanglab.ccmb.med.umich.edu/I-TASSER/). The 3D structure models presenting the higher confidence level were retrieved for each protein (C-scores of -1.51 for GST-IEP and -1.00 for HisV5-IEP models), and analyzed using Jmol software (version 13.0.1) and the 3D molecule viewer component of Vector NTI® Advance 11.0 (VNTI; Life Technologies). The molecule surface was calculated in VNTI by using the Connolly surface method (Connolly ML 1983; Connolly ML 1993).

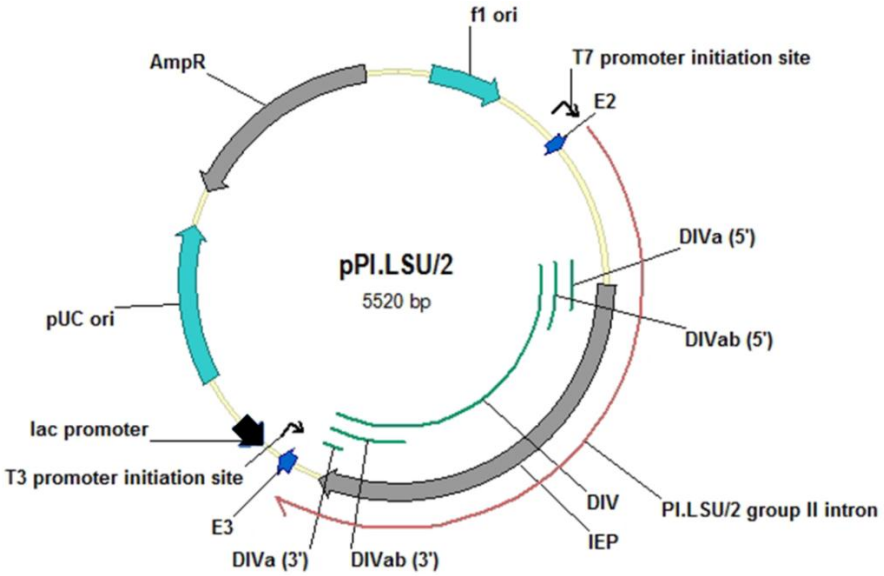
The analysis of each MALDI-TOF/TOF data set obtained for the 69-kDa bands of HisV5-IEP (WT and mtDD-) purified by IMAC and for the 69-kDa and 75-kDa bands purified by sucrose centrifugation consisted in peptide fingerprint mass mapping using the search program MS-Fit (http://prospector2.ucsf.edu). The UniProtKB/SwissProt annotated database (without restriction in species; 517787 entries in the database in the version dated 2011.07.06) was used and the settings were one missed cleavage allowed, requirement of minimum 4 peptides that match with the input data, and a mass tolerance of 20 ppm.

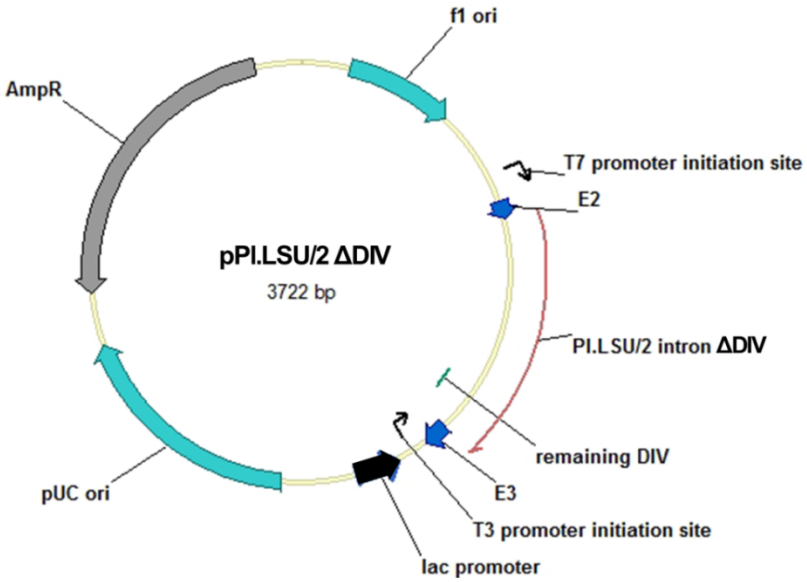
Predicted MS data set for the sequence of Pl.LSU/2 IEP was generated by *in silico* tryptic digestion using the MS-digest program (http://prospector2.ucsf.edu) using default settings. For each predicted parental ion, the program also generates predicted MS/MS data set theoretically obtained after parental ion fragmentation. Experimental MALDI-TOF/TOF MS spectra (MS data) were then manually compared to the MS data set from *in silico* tryptic digestion of Pl.LSU/2 IEP, and experimental MS/MS spectra of 4 selected parental ions were then manually compared to theoretical MS/MS data set from the corresponding *in silico* predicted parental ions.

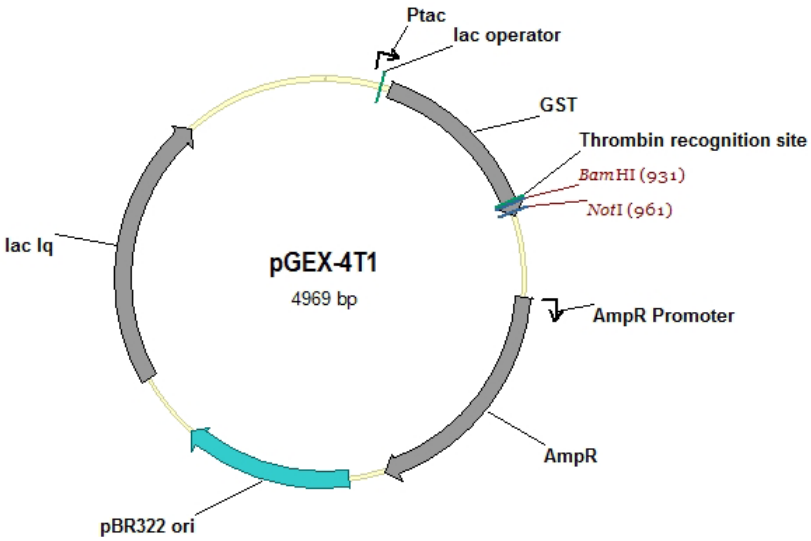
The isoelectric point of the Pl.LSU/2 IEP was predicted using the iep application of the EMBOSS software, version 6.3.1 (Rice P et al. 2000).

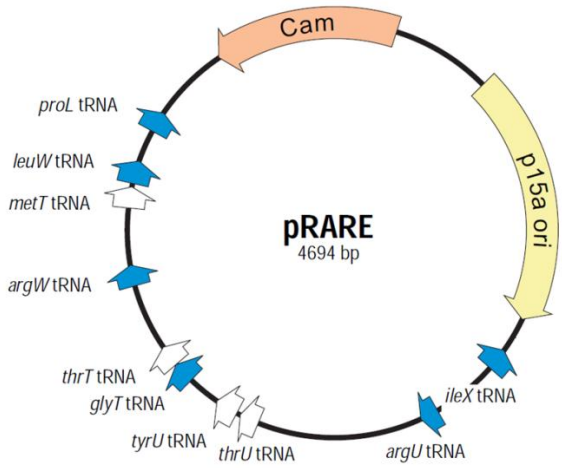
APPENDIX

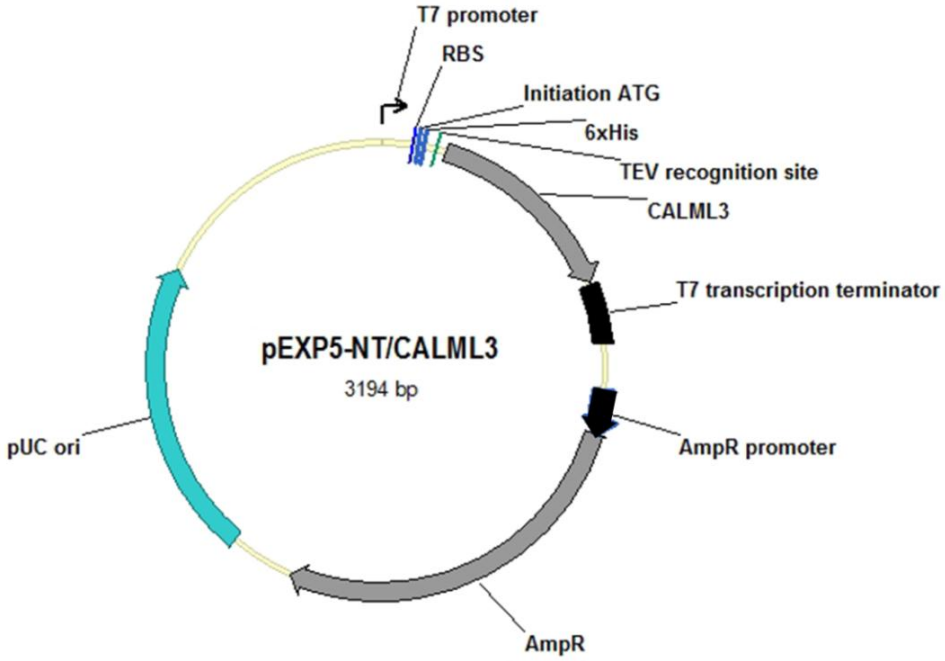
1 - PLASMIDS MAPS

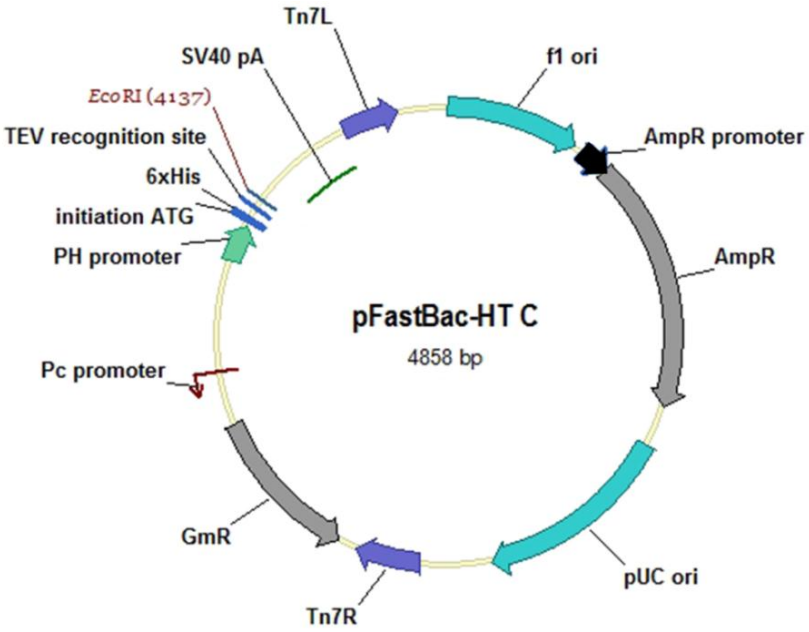


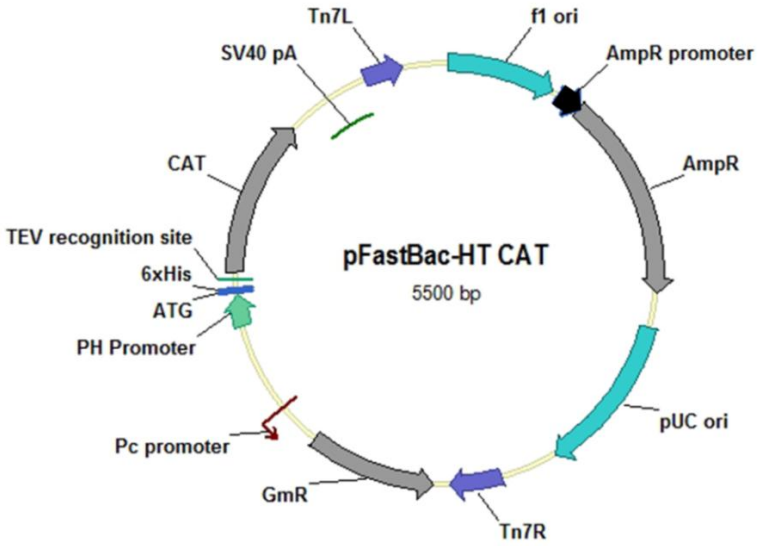


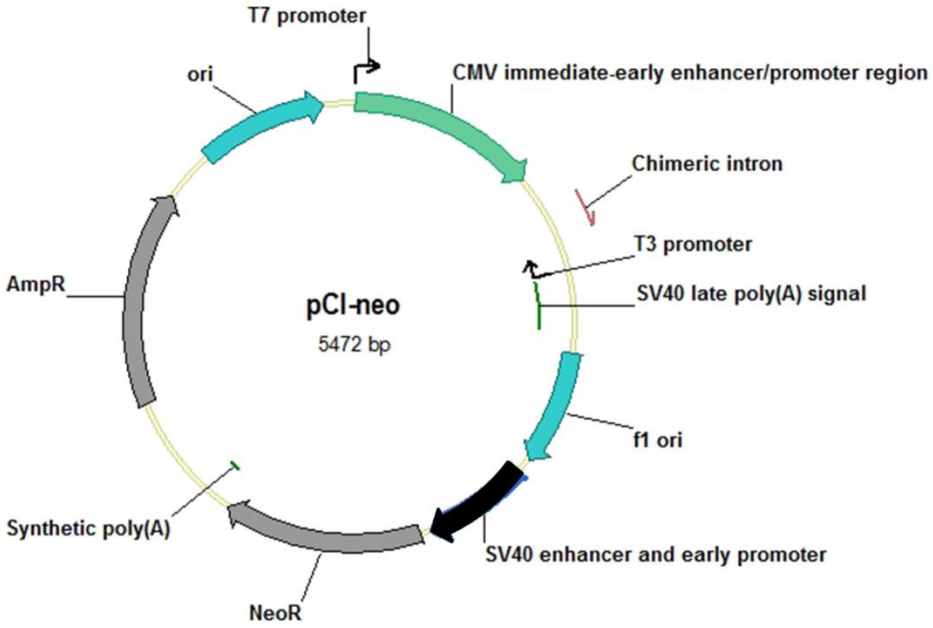


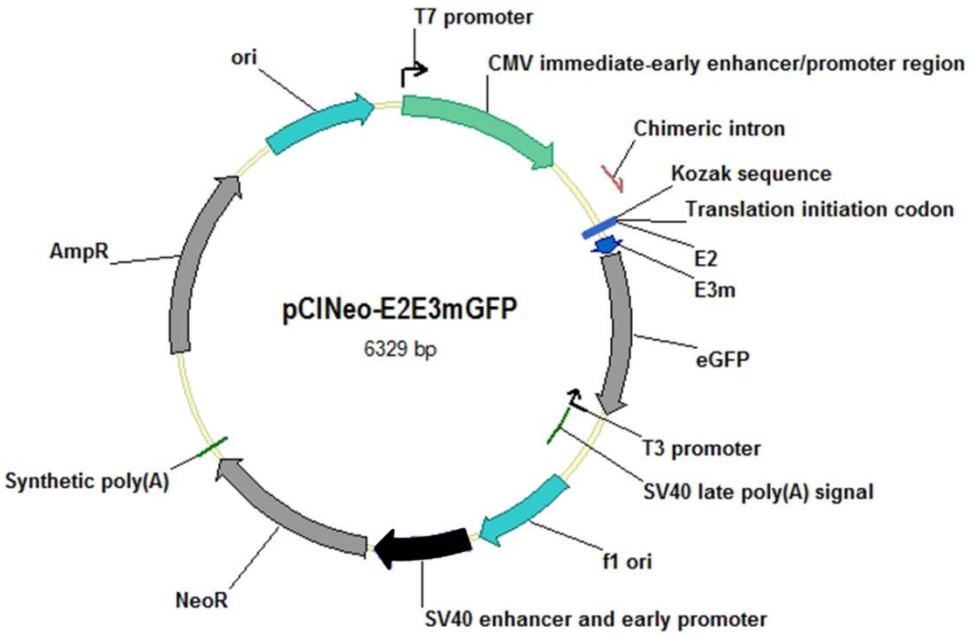

Cam
p15a ori
proL tRNA
leuW tRNA
metT tRNA
argW tRNA
thrT tRNA
glyT tRNA
tyrU tRNA
thrU tRNA
argU tRNA
ileX tRNA
pRARE
4694 bp

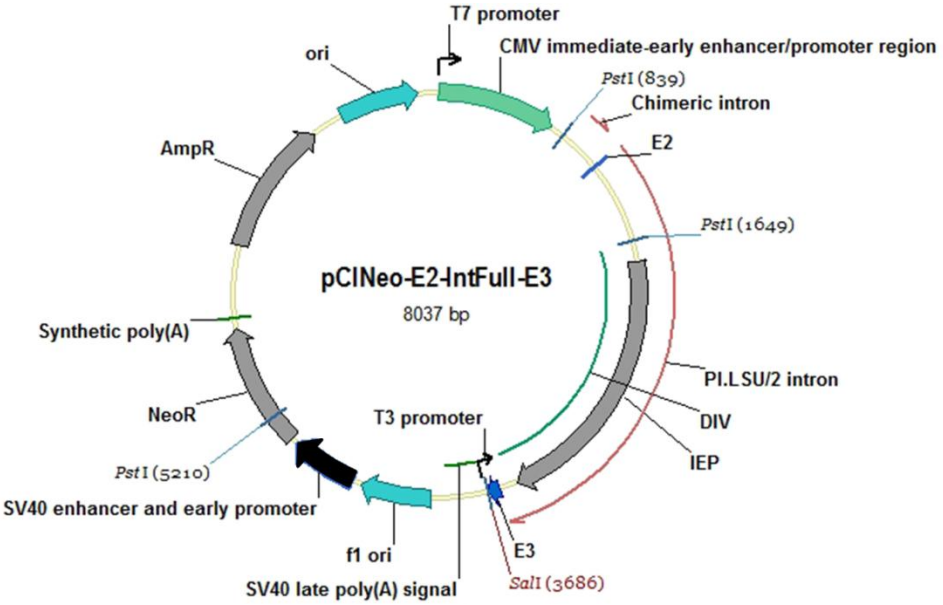

T7 promoter
RBS
Initiation ATG
6xHis
TEV recognition site
CALML3
T7 transcription terminator
AmpR promoter
AmpR
pUC ori
pEXP5-NT/CALML3
3194 bp

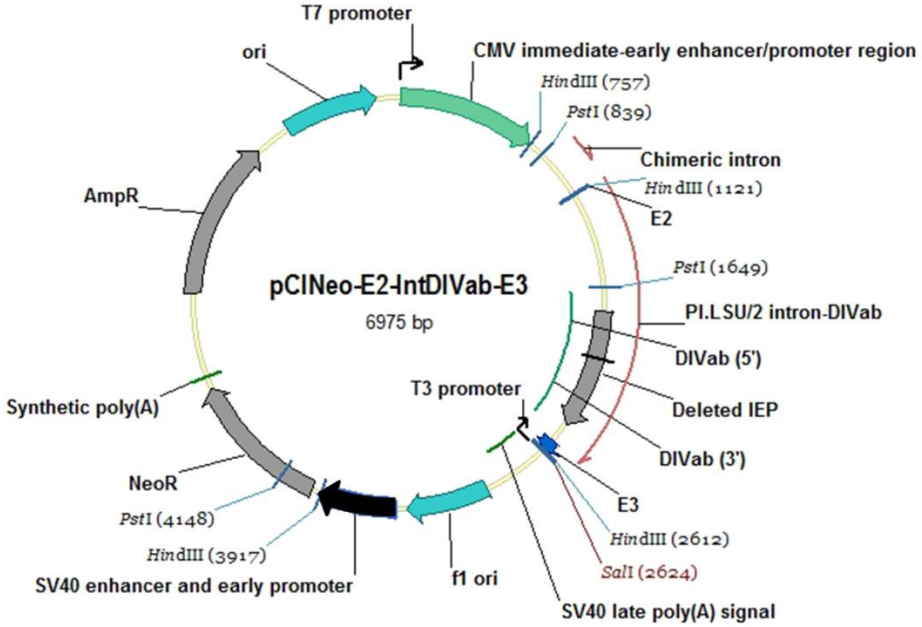

Tn7L
SV40 pA
f1 ori
EcoRI (4137)
TEV recognition site
AmpR promoter
6xHis
initiation ATG
AmpR
PH promoter
pFastBac-HT C
4858 bp
Pc promoter
GmR
pUC ori
Tn7R

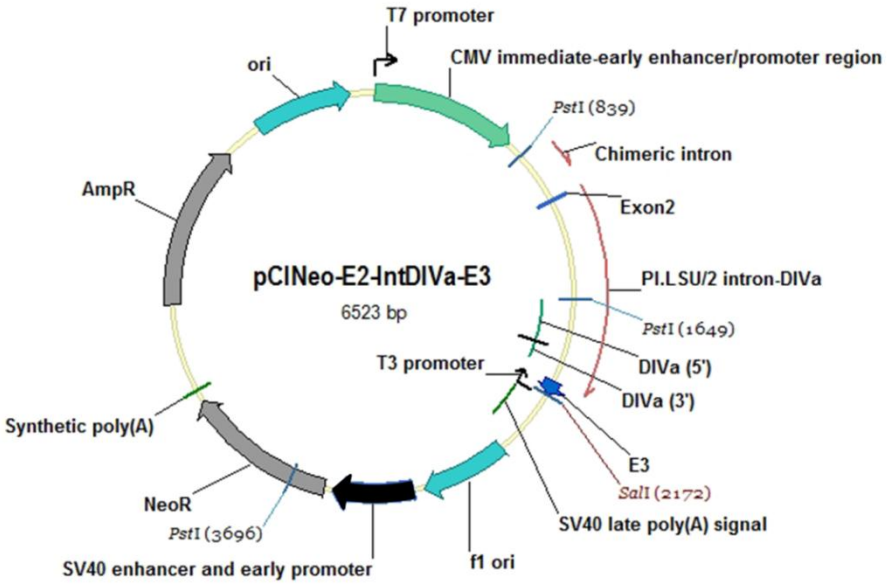

Tn7L
f1 ori
SV40 pA
AmpR promoter
CAT
AmpR
TEV recognition site
pFastBac-HT CAT
5500 bp
6xHis
ATG
PH Promoter
pUC ori
Pc promoter
GmR
Tn7R

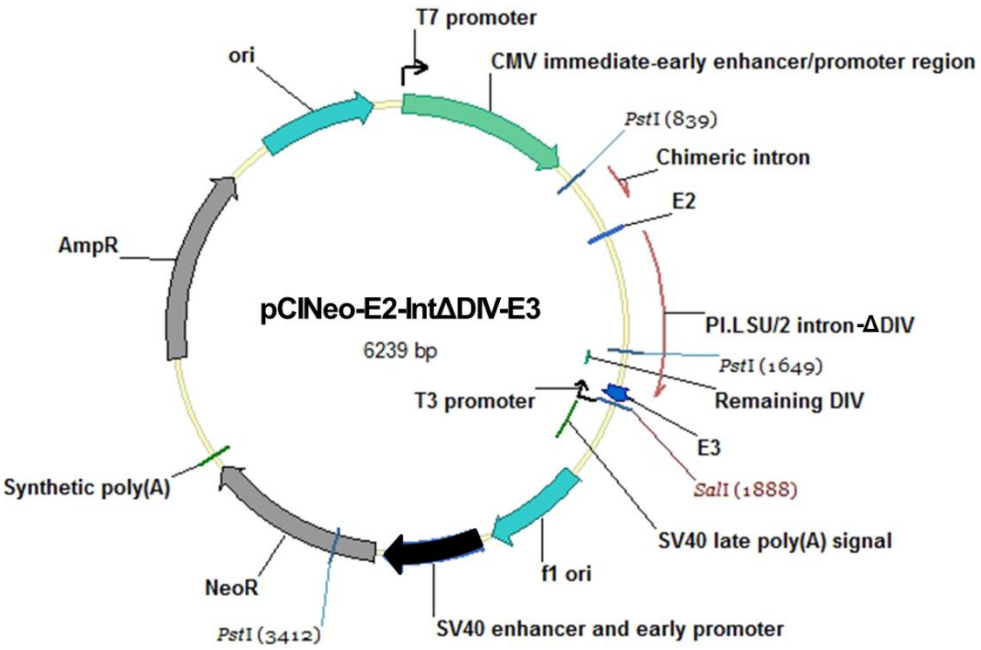

T7 promoter
ori
CMV immediate-early enhancer/promoter region
Chimeric intron
T3 promoter
AmpR
pCI-neo
5472 bp
SV40 late poly(A) signal
f1 ori
Synthetic poly(A)
SV40 enhancer and early promoter
NeoR

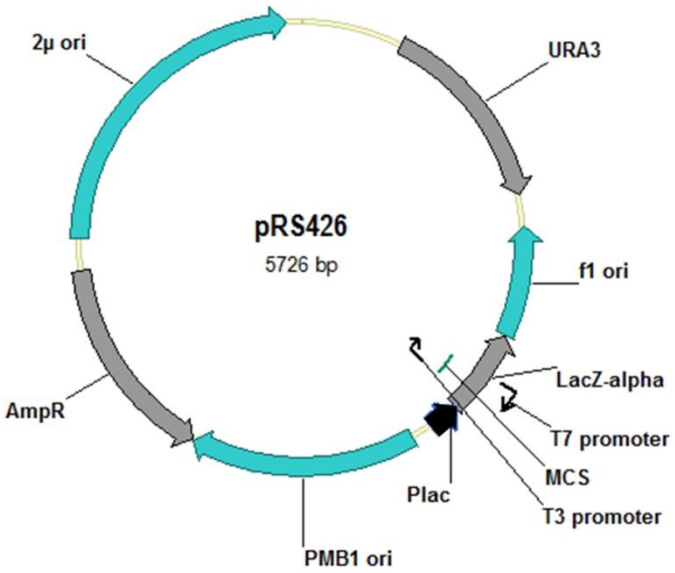

T7 promoter
ori
CMV immediate-early enhancer/promoter region
Chimeric intron
Kozak sequence
Translation initiation codon
AmpR
E2
E3m
pCINeo-E2E3mGFP
6329 bp
eGFP
T3 promoter
Synthetic poly(A)
SV40 late poly(A) signal
f1 ori
NeoR
SV40 enhancer and early promoter

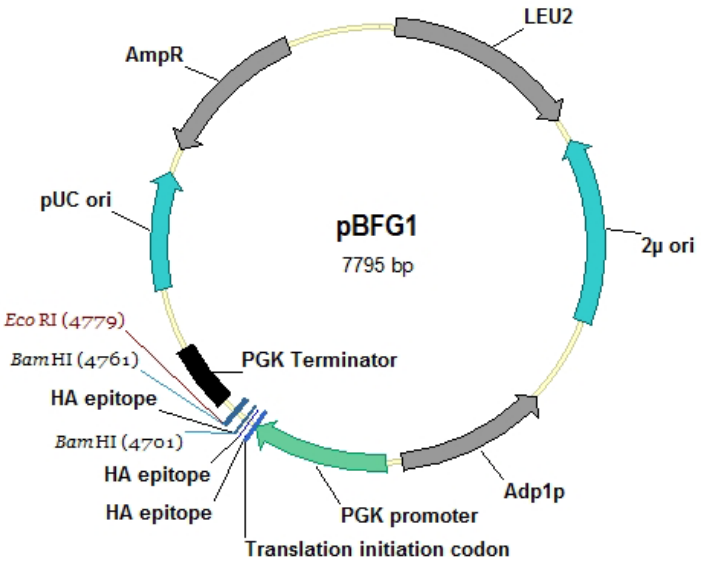

T7 promoter
ori
CMV immediate-early enhancer/promoter region
PstI (839)
Chimeric intron
AmpR
E2
PstI (1649)
pCINeo-E2-IntFull-E3
8037 bp
Synthetic poly(A)
PI.LSU/2 intron
NeoR
T3 promoter
DIV
IEP
PstI (5210)
SV40 enhancer and early promoter
f1 ori
E3
SV40 late poly(A) signal
SalI (3686)

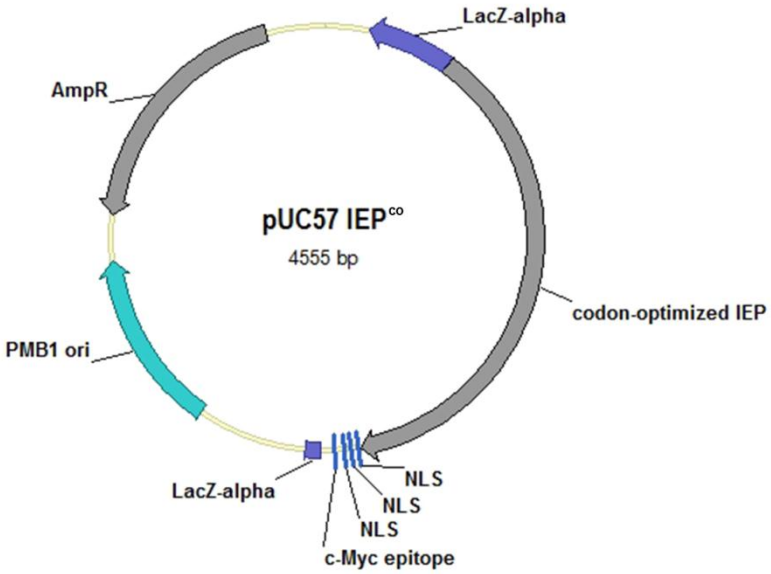

T7 promoter
ori
CMV immediate-early enhancer/promoter region
HindIII (757)
PstI (839)
Chimeric intron
AmpR
HindIII (1121)
E2
PstI (1649)
pCINeo-E2-IntDIVab-E3
6975 bp
PI.LSU/2 intron-DIVab
DIVab (5')
T3 promoter
Synthetic poly(A)
Deleted IEP
DIVab (3')
NeoR
E3
PstI (4148)
HindIII (2612)
HindIII (3917)
SalI (2624)
SV40 enhancer and early promoter
f1 ori
SV40 late poly(A) signal

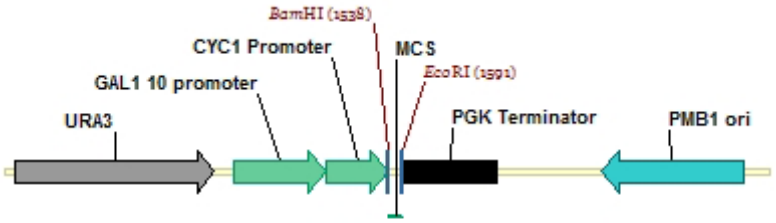

T7 promoter
ori
CMV immediate-early enhancer/promoter region
PstI (839)
Chimeric intron
AmpR
Exon2
pCINeo-E2-IntDIVa-E3
6523 bp
PI.LSU/2 intron-DIVa
PstI (1649)
T3 promoter
DIVa (5')
DIVa (3')
Synthetic poly(A)
E3
SalI (2172)
NeoR
SV40 late poly(A) signal
PstI (3696)
SV40 enhancer and early promoter
f1 ori


T7 promoter
ori
CMV immediate-early enhancer/promoter region
PstI (839)
Chimeric intron
E2
AmpR
pClNeo-E2-IntΔDIV-E3
6239 bp
PI.LSU/2 intron-ΔDIV
PstI (1649)
Remaining DIV
T3 promoter
E3
SalI (1888)
Synthetic poly(A)
SV40 late poly(A) signal
f1 ori
NeoR
PstI (3412)
SV40 enhancer and early promoter


2µ ori
URA3
pRS426
5726 bp
f1 ori
LacZ-alpha
AmpR
T7 promoter
MCS
Plac
T3 promoter
PMB1 ori

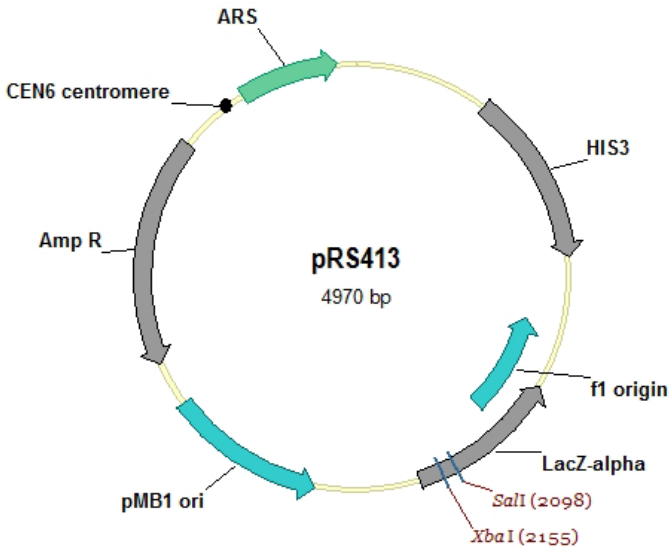

LEU2
AmpR
pUC ori
pBFG1
7795 bp
2µ ori
Eco RI (4779)
BamHI (4761)
PGK Terminator
HA epitope
BamHI (4701)
HA epitope
Adp1p
HA epitope
PGK promoter
Translation initiation codon

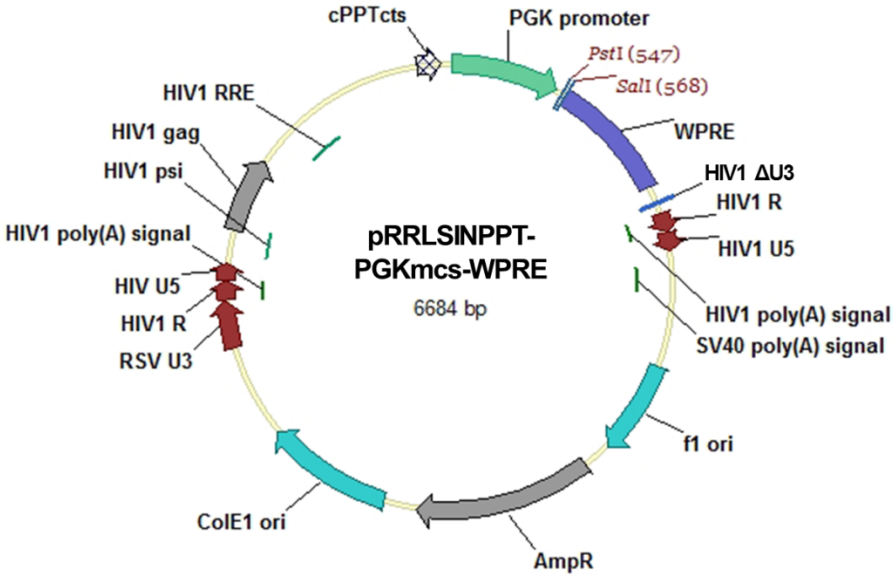


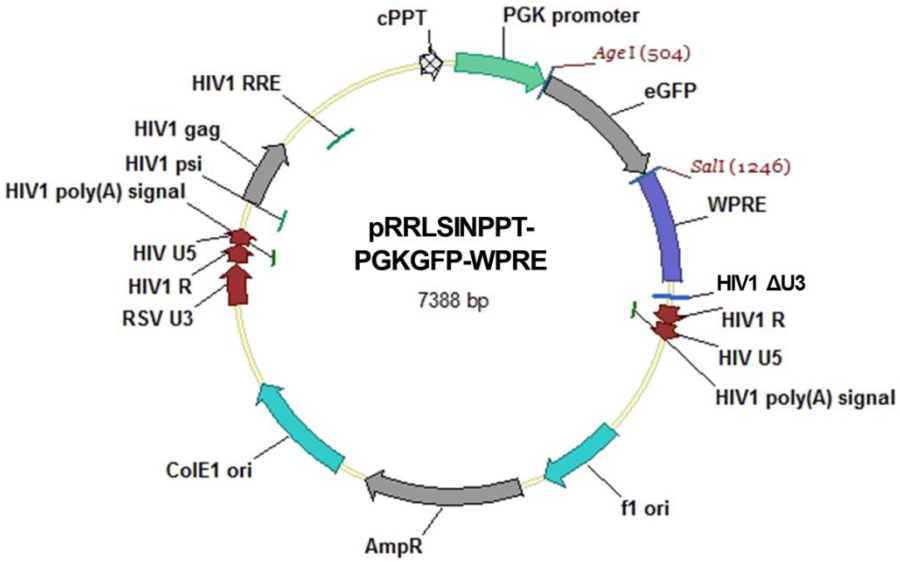


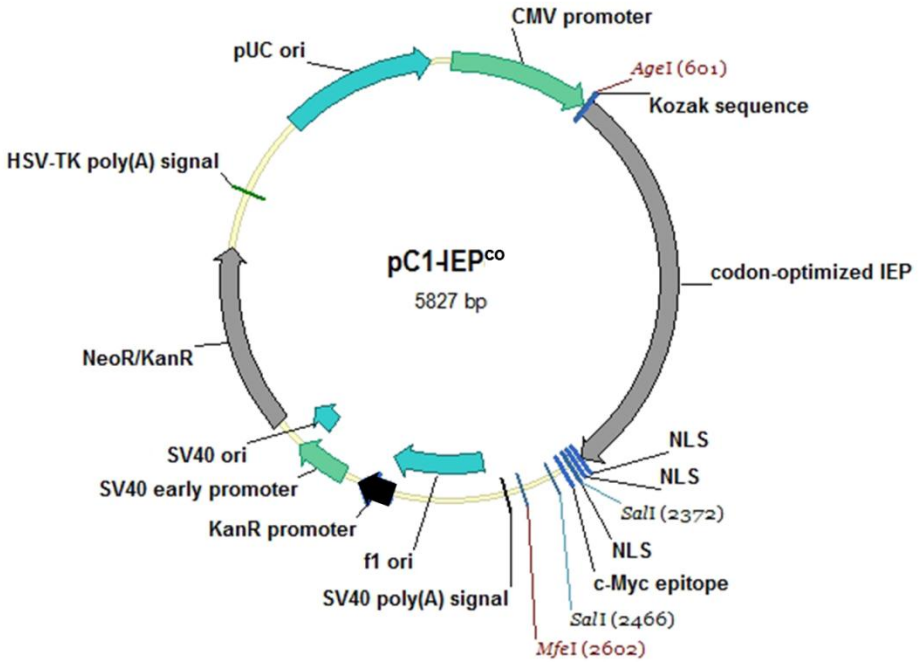
pYEF1 fragment

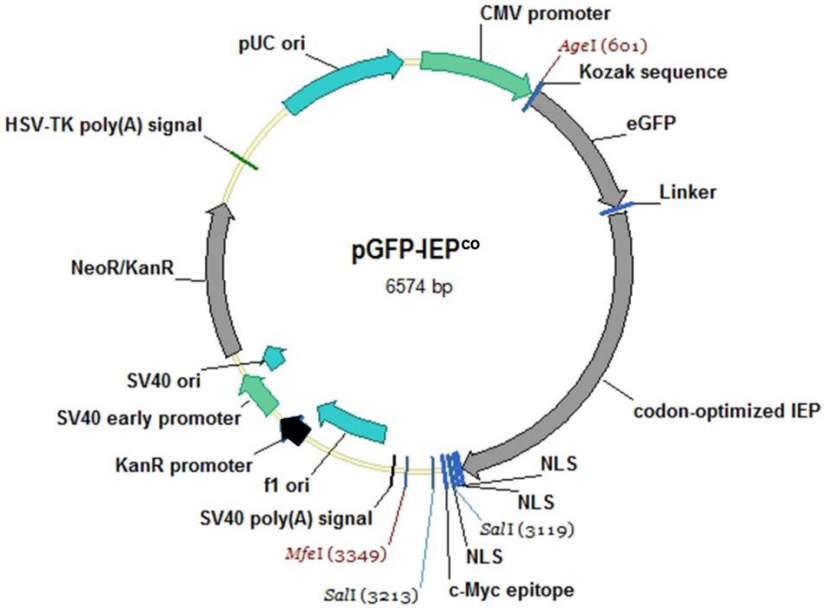
3069 bp

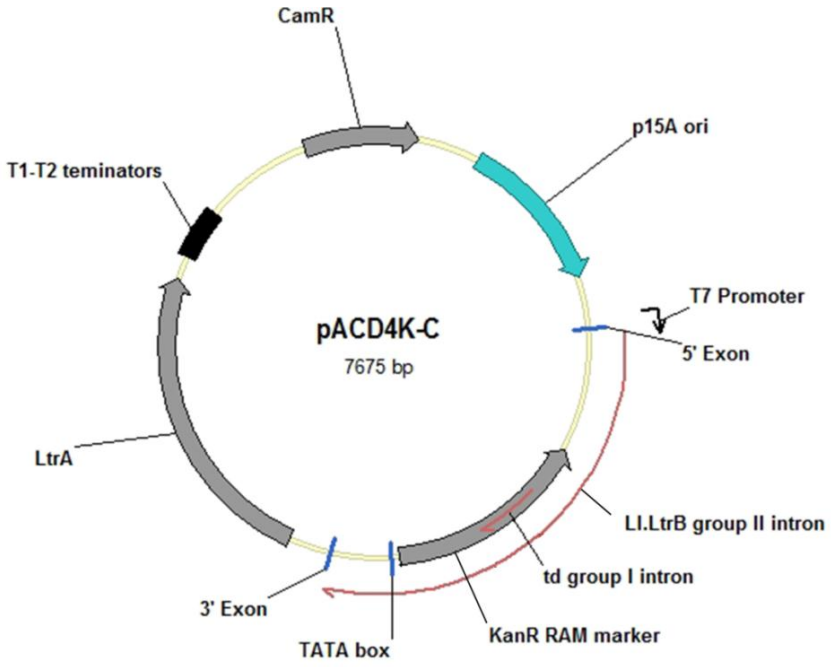


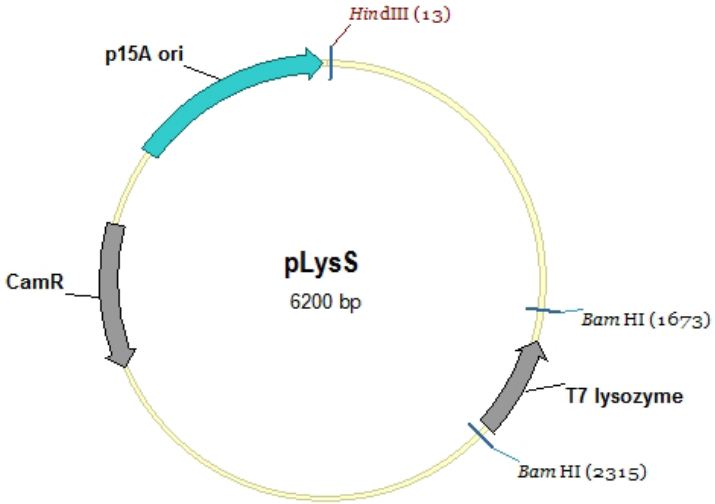

cPPTcts
PGK promoter
PstI (547)
SalI (568)
WPRE
HIV1 ΔU3
HIV1 R
HIV1 U5
HIV1 poly(A) signal
SV40 poly(A) signal
f1 ori
AmpR
ColE1 ori
RSV U3
HIV1 R
HIV U5
HIV1 poly(A) signal
HIV1 psi
HIV1 gag
HIV1 RRE
pRRLSINPPT-
PGKmcs-WPRE
6684 bp

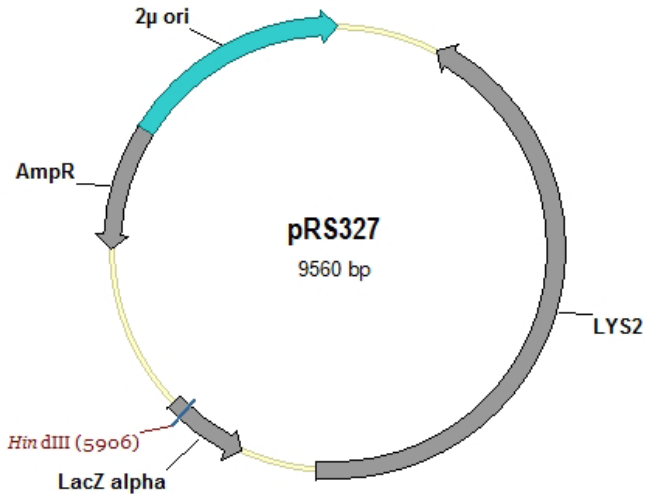

cPPT
PGK promoter
AgeI (504)
eGFP
SalI (1246)
WPRE
HIV1 ΔU3
HIV1 R
HIV U5
HIV1 poly(A) signal
f1 ori
AmpR
ColE1 ori
RSV U3
HIV1 R
HIV U5
HIV1 poly(A) signal
HIV1 psi
HIV1 gag
HIV1 RRE
pRRLSINPPT-
PGKGFP-WPRE
7388 bp


CMV promoter
AgeI (601)
Kozak sequence
codon-optimized IEP
NLS
NLS
SalI (2372)
NLS
c-Myc epitope
SalI (2466)
MfeI (2602)
SV40 poly(A) signal
f1 ori
KanR promoter
SV40 early promoter
SV40 ori
NeoR/KanR
HSV-TK poly(A) signal
pUC ori
pC1-IEPco
5827 bp


CMV promoter
pUC ori
AgeI (601)
Kozak sequence
HSV-TK poly(A) signal
eGFP
Linker
pGFP-IEPco
6574 bp
NeoR/KanR
SV40 ori
SV40 early promoter
codon-optimized IEP
KanR promoter
f1 ori
NLS
NLS
SV40 poly(A) signal
SalI (3119)
MfeI (3349)
NLS
SalI (3213)
c-Myc epitope


CamR
p15A ori
T1-T2 teminators
T7 Promoter
pACD4K-C
7675 bp
5' Exon
LtrA
LI.LtrB group II intron
td group I intron
3' Exon
KanR RAM marker
TATA box


HindIII (13)
p15A ori
pLysS
6200 bp
CamR
BamHI (1673)
T7 lysozyme
BamHI (2315)


2µ ori
AmpR
pRS327
9560 bp
LYS2
Hin dIII (5906)
LacZ alpha

REFERENCES

Abelson J, Trotta CR, Li H. 1998. tRNA splicing. *The Journal of biological chemistry* **273**: 12685-12688.

Aiuti A, Cassani B, Andolfi G, Mirolo M, Biasco L, Recchia A, Urbinati F, Valacca C, Scaramuzza S, Aker M et al. 2007. Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. *The Journal of clinical investigation* **117**: 2233-2240.

Aiuti A, Cattaneo F, Galimberti S, Benninghoff U, Cassani B, Callegaro L, Scaramuzza S, Andolfi G, Mirolo M, Brigida I et al. 2009. Gene therapy for immunodeficiency due to adenosine deaminase deficiency. *The New England journal of medicine* **360**: 447-458.

Aiuti A, Vai S, Mortellaro A, Casorati G, Ficara F, Andolfi G, Ferrari G, Tabucchi A, Carlucci F, Ochs HD et al. 2002. Immune reconstitution in ADA-SCID after PBL gene therapy and discontinuation of enzyme replacement. *Nature medicine* **8**: 423-425.

Aizawa Y, Xiang Q, Lambowitz AM, Pyle AM. 2003. The pathway for DNA recognition and RNA integration by a group II intron retrotransposon. *Molecular cell* **11**: 795-805.

Aker M, Tubb J, Groth AC, Bukovsky AA, Bell AC, Felsenfeld G, Kiem HP, Stamatoyannopoulos G, Emery DW. 2007. Extended core sequences from the cHS4 insulator are necessary for protecting retroviral vectors from silencing position effects. *Human gene therapy* **18**: 333-343.

Albulescu LO, Sabet N, Gudipati M, Stepankiw N, Bergman ZJ, Huffaker TC, Pleiss JA. 2012. A quantitative, high-throughput reverse genetic screen reveals novel connections between Pre-mRNA splicing and 5' and 3' end transcript determinants. *PLoS genetics* **8**: e1002530.

Alkhatib G, Combadiere C, Broder CC, Feng Y, Kennedy PE, Murphy PM, Berger EA. 1996. CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. *Science* **272**: 1955-1958.

Allen C, Miller CA, Nickoloff JA. 2003. The mutagenic potential of a single DNA double-strand break in a mammalian chromosome is not influenced by transcription. *DNA repair* **2**: 1147-1156.

Almarza D, Bussadori G, Navarro M, Mavilio F, Larcher F, Murillas R. 2011. Risk assessment in skin gene therapy: viral-cellular fusion transcripts generated by proviral transcriptional read-through in keratinocytes transduced with self-inactivating lentiviral vectors. *Gene therapy* **18**: 674-681.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**: 3389-3402.

Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schaffer AA, Yu YK. 2005. Protein database searches using compositionally adjusted substitution matrices. *The FEBS journal* **272**: 5101-5109.

Alwin S, Gere MB, Guhl E, Effertz K, Barbas CF, 3rd, Segal DJ, Weitzman MD, Cathomen T. 2005. Custom zinc-finger nucleases for use in human cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **12**: 610-617.

Anderson WF, Blaese RM, Culver K. 1990. The ADA human gene therapy clinical protocol: Points to Consider response with clinical protocol, July 6, 1990. *Human gene therapy* **1**: 331-362.

Andrews JL, Kadan MJ, Gorziglia MI, Kaleko M, Connelly S. 2001. Generation and characterization of E1/E2a/E3/E4-deficient adenoviral vectors encoding human factor VIII. *Molecular therapy : the journal of the American Society of Gene Therapy* **3**: 329-336.

Arai T, Takada M, Ui M, Iba H. 1999. Dose-dependent transduction of vesicular stomatitis virus G protein-pseudotyped retrovirus vector into human solid tumor cell lines and murine fibroblasts. *Virology* **260**: 109-115.

Arlt H, Steglich G, Perryman R, Guiard B, Neupert W, Langer T. 1998. The formation of respiratory chain complexes in mitochondria is under the proteolytic control of the m-AAA protease. *The EMBO journal* **17**: 4837-4847.

Arnould S, Chames P, Perez C, Lacroix E, Duclert A, Epinat JC, Stricher F, Petit AS, Patin A, Guillier S et al. 2006. Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. *Journal of molecular biology* **355**: 443-458.

Arnould S, Perez C, Cabaniols JP, Smith J, Gouble A, Grizot S, Epinat JC, Duclert A, Duchateau P, Paques F. 2007. Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. *Journal of molecular biology* **371**: 49-65.

Aronovich EL, Bell JB, Belur LR, Gunther R, Koniar B, Erickson DC, Schachern PA, Matise I, McIvor RS, Whitley CB et al. 2007. Prolonged expression of a lysosomal enzyme in mouse liver after Sleeping Beauty transposon-mediated gene delivery: implications for non-viral gene therapy of mucopolysaccharidoses. *The journal of gene medicine* **9**: 403-415.

Arumugam PI, Higashimoto T, Urbinati F, Modlich U, Nestheide S, Xia P, Fox C, Corsinotti A, Baum C, Malik P. 2009. Genotoxic potential of lineage-specific lentivirus vectors carrying the beta-globin locus control region. *Molecular therapy : the journal of the American Society of Gene Therapy* **17**: 1929-1937.

Arumugam PI, Scholes J, Perelman N, Xia P, Yee JK, Malik P. 2007. Improved human beta-globin expression from self-inactivating lentiviral vectors carrying the chicken hypersensitive site-4 (cHS4) insulator element. *Molecular therapy : the journal of the American Society of Gene Therapy* **15**: 1863-1871.

Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, Jr., Stoddard BL, Baker D. 2006. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**: 656-659.

Avedillo Diez I, Zychlinski D, Coci EG, Galla M, Modlich U, Dewey RA, Schwarzer A, Maetzig T, Mpofu N, Jaeckel E et al. 2011. Development of novel efficient SIN vectors with improved safety features for Wiskott-Aldrich syndrome stem cell based gene therapy. *Molecular pharmaceutics* **8**: 1525-1537.

Baca AM, Hol WG. 2000. Overcoming codon bias: a method for high-level overexpression of Plasmodium and other AT-rich parasite genes in Escherichia coli. *International journal for parasitology* **30**: 113-118.

Baekelandt V, Eggermont K, Michiels M, Nuttin B, Debyser Z. 2003. Optimized lentiviral vector production and purification procedure prevents immune response after transduction of mouse brain. *Gene therapy* **10**: 1933-1940.

Bainbridge JW, Smith AJ, Barker SS, Robbie S, Henderson R, Balaggan K, Viswanathan A, Holder GE, Stockman A, Tyler N et al. 2008. Effect of gene therapy on visual function in Leber's congenital amaurosis. *The New England journal of medicine* **358**: 2231-2239.

Balciunas D, Wangensteen KJ, Wilber A, Bell J, Geurts A, Sivasubbu S, Wang X, Hackett PB, Largaespada DA, McIvor RS et al. 2006. Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. *PLoS genetics* **2**: e169.

Baneyx F. 1999. Recombinant protein expression in Escherichia coli. *Current opinion in biotechnology* **10**: 411-421.

Beauregard A, Curcio MJ, Belfort M. 2008. The take and give between retrotransposable elements and their hosts. *Annual review of genetics* **42**: 587-617.

Belhocine K, Mak AB, Cousineau B. 2008. Trans-splicing versatility of the Ll.LtrB group II intron. *RNA* **14**: 1782-1790.

Benabdellah K, Cobo M, Munoz P, Toscano MG, Martin F. 2011. Development of an all-in-one lentiviral vector system based on the original TetR for the easy generation of Tet-ON cell lines. *PloS one* **6**: e23734.

Biasco L, Ambrosi A, Pellin D, Bartholomae C, Brigida I, Roncarolo MG, Di Serio C, von Kalle C, Schmidt M, Aiuti A. 2011. Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. *EMBO molecular medicine* **3**: 89-101.

Bibikova M, Golic M, Golic KG, Carroll D. 2002. Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* **161**: 1169-1175.

Biffi A, Bartolomae CC, Cesana D, Cartier N, Aubourg P, Ranzani M, Cesani M, Benedicenti F, Plati T, Rubagotti E et al. 2011. Lentiviral vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection. *Blood* **117**: 5332-5339.

Bitinaite J, Wah DA, Aggarwal AK, Schildkraut I. 1998. FokI dimerization is required for DNA cleavage. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 10570-10575.

Bjork BC, Fujiwara Y, Davis SW, Qiu H, Saunders TL, Sandy P, Orkin S, Camper SA, Beier DR. 2010. A transient transgenic RNAi strategy for rapid characterization of gene function during embryonic development. *PloS one* **5**: e14375.

Blaese RM, Culver KW, Chang L, Anderson WF, Mullen C, Nienhuis A, Carter C, Dunbar C, Leitman S, Berger M et al. 1993. Treatment of severe combined immunodeficiency disease (SCID) due to adenosine deaminase deficiency with CD34+ selected autologous peripheral blood cells transduced with a human ADA gene. Amendment to clinical research project, Project 90-C-195, January 10, 1992. *Human gene therapy* **4**: 521-527.

Blaese RM, Culver KW, Miller AD, Carter CS, Fleisher T, Clerici M, Shearer G, Chang L, Chiang Y, Tolstoshev P et al. 1995. T lymphocyte-directed gene therapy for ADA- SCID: initial trial results after 4 years. *Science* **270**: 475-480.

Blocker FJ, Mohr G, Conlan LH, Qi L, Belfort M, Lambowitz AM. 2005. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA* **11**: 14-28.

Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, Lahaye T, Nickstadt A, Bonas U. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**: 1509-1512.

Bogerd HP, Echarri A, Ross TM, Cullen BR. 1998. Inhibition of human immunodeficiency virus Rev and human T-cell leukemia virus Rex function, but not Mason-Pfizer monkey virus constitutive transport element activity, by a mutant human nucleoporin targeted to Crm1. *Journal of virology* **72**: 8627-8635.

Bonini C, Ferrari G, Verzeletti S, Servida P, Zappone E, Ruggieri L, Ponzoni M, Rossini S, Mavilio F, Traversari C et al. 1997. HSV-TK gene transfer into donor lymphocytes for control of allogeneic graft-versus-leukemia. *Science* **276**: 1719-1724.

Bordignon C, Notarangelo LD, Nobili N, Ferrari G, Casorati G, Panina P, Mazzolari E, Maggioni D, Rossi C, Servida P et al. 1995. Gene therapy in peripheral blood lymphocytes and bone marrow for ADA- immunodeficient patients. *Science* **270**: 470-475.

Bosch ML, Earl PL, Fargnoli K, Picciafuoco S, Giombini F, Wong-Staal F, Franchini G. 1989. Identification of the fusion peptide of primate immunodeficiency viruses. *Science* **244**: 694-697.

Boudvillain M, de Lencastre A, Pyle AM. 2000. A tertiary interaction that links active-site domains to the 5' splice site of a group II intron. *Nature* **406**: 315-318.

Boudvillain M, Pyle AM. 1998. Defining functional groups, core structural features and inter-domain tertiary contacts essential for group II intron self-splicing: a NAIM analysis. *The EMBO journal* **17**: 7091-7104.

Bouma G, Burns SO, Thrasher AJ. 2009. Wiskott-Aldrich Syndrome: Immunodeficiency resulting from defective cell migration and impaired immunostimulatory activation. *Immunobiology* **214**: 778-790.

Bour S, Schubert U, Strebel K. 1995. The human immunodeficiency virus type 1 Vpu protein specifically binds to the cytoplasmic domain of CD4: implications for the mechanism of degradation. *Journal of virology* **69**: 1510-1520.

Boussif O, Lezoualc'h F, Zanta MA, Mergny MD, Scherman D, Demeneix B, Behr JP. 1995. A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 7297-7301.

Bowers WJ, Mastrangelo MA, Howard DF, Southerland HA, Maguire-Zeiss KA, Federoff HJ. 2006. Neuronal precursor-restricted transduction via in utero CNS gene delivery of a novel bipartite HSV amplicon/transposase hybrid vector. *Molecular therapy : the journal of the American Society of Gene Therapy* **13**: 580-588.

Boztug K, Schmidt M, Schwarzer A, Banerjee PP, Diez IA, Dewey RA, Bohm M, Nowrouzi A, Ball CR, Glimm H et al. 2010. Stem-cell gene therapy for the Wiskott-Aldrich syndrome. *The New England journal of medicine* **363**: 1918-1927.

Briggs AW, Rios X, Chari R, Yang L, Zhang F, Mali P, Church GM. 2012. Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic acids research.*

Briggs JA, Simon MN, Gross I, Krausslich HG, Fuller SD, Vogt VM, Johnson MC. 2004. The stoichiometry of Gag protein in HIV-1. *Nature structural & molecular biology* **11**: 672-675.

Briggs JA, Wilk T, Fuller SD. 2003. Do lipid rafts mediate virus assembly and pseudotyping? *The Journal of general virology* **84**: 757-768.

Brown BD, Cantore A, Annoni A, Sergi LS, Lombardo A, Della Valle P, D'Angelo A, Naldini L. 2007. A microRNA-regulated lentiviral vector mediates stable correction of hemophilia B mice. *Blood* **110**: 4144-4152.

Brown BD, Venneri MA, Zingale A, Sergi Sergi L, Naldini L. 2006. Endogenous microRNA regulation suppresses transgene expression in hematopoietic lineages and enables stable gene transfer. *Nature medicine* **12**: 585-591.

Brown PO, Bowerman B, Varmus HE, Bishop JM. 1987. Correct integration of retroviral DNA in vitro. *Cell* **49**: 347-356.

Brun S, Faucon-Biguet N, Mallet J. 2003. Optimization of transgene expression at the posttranscriptional level in neural cells: implications for gene therapy. *Molecular therapy : the journal of the American Society of Gene Therapy* **7**: 782-789.

Buck CB, Shen X, Egan MA, Pierson TC, Walker CM, Siliciano RF. 2001. The human immunodeficiency virus type 1 gag gene encodes an internal ribosome entry site. *Journal of virology* **75**: 181-191.

Bukrinsky MI, Sharova N, McDonald TL, Pushkarskaya T, Tarpley WG, Stevenson M. 1993. Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 6125-6129.

Burns JC, Friedmann T, Driever W, Burrascano M, Yee JK. 1993. Vesicular stomatitis virus G glycoprotein pseudotyped retroviral vectors: concentration to very high titer and efficient gene transfer into mammalian and nonmammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 8033-8037.

Bushey AM, Dorman ER, Corces VG. 2008. Chromatin insulators: regulatory mechanisms and epigenetic inheritance. *Molecular cell* **32**: 1-9.

Bushman F, Lewinski M, Ciuffi A, Barr S, Leipzig J, Hannenhalli S, Hoffmann C. 2005. Genome-wide analysis of retroviral DNA integration. *Nature reviews Microbiology* **3**: 848-858.

Bushman FD. 1994. Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 9233-9237.

Bushman FD, Miller MD. 1997. Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. *Journal of virology* **71**: 458-464.

Campbell TL. 1966. Reflections on research and the future of medicine. *Science* **153**: 442-449.

Capecchi MR. 2001. Generating mice with targeted mutations. *Nature medicine* **7**: 1086-1090.

Carignani G, Groudinsky O, Frezza D, Schiavon E, Bergantino E, Slonimski PP. 1983. An mRNA maturase is encoded by the first intron of the mitochondrial gene for the subunit I of cytochrome oxidase in S. cerevisiae. *Cell* **35**: 733-742.

Carson M, Johnson DH, McDonald H, Brouillette C, Delucas LJ. 2007. His-tag impact on structure. *Acta crystallographica Section D, Biological crystallography* **63**: 295-301.

Cartier N, Hacein-Bey-Abina S, Bartholomae CC, Veres G, Schmidt M, Kutschera I, Vidaud M, Abel U, Dal-Cortivo L, Caccavelli L et al. 2009. Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* **326**: 818-823.

Cattoglio C, Facchini G, Sartori D, Antonelli A, Miccio A, Cassani B, Schmidt M, von Kalle C, Howe S, Thrasher AJ et al. 2007. Hot spots of retroviral integration in human CD34+ hematopoietic cells. *Blood* **110**: 1770-1778.

Cattoglio C, Maruggi G, Bartholomae C, Malani N, Pellin D, Cocchiarella F, Magnani Z, Ciceri F, Ambrosi A, von Kalle C et al. 2010a. High-definition mapping of retroviral integration sites defines the fate of allogeneic T cells after donor lymphocyte infusion. *PloS one* **5**: e15688.

Cattoglio C, Pellin D, Rizzi E, Maruggi G, Corti G, Miselli F, Sartori D, Guffanti A, Di Serio C, Ambrosi A et al. 2010b. High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. *Blood* **116**: 5507-5517.

Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, Gross F, Yvon E, Nusbaum P, Selz F, Hue C, Certain S, Casanova JL et al. 2000. Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288**: 669-672.

Cavazzana-Calvo M, Payen E, Negre O, Wang G, Hehir K, Fusil F, Down J, Denaro M, Brady T, Westerman K et al. 2010. Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. *Nature* **467**: 318-322.

Cermak T, Doyle EL, Christian M, Wang L, Zhang Y, Schmidt C, Baller JA, Somia NV, Bogdanove AJ, Voytas DF. 2011. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research* **39**: e82.

Cesana D, Sgualdino J, Rudilosso L, Merella S, Naldini L, Montini E. 2012. Whole transcriptome characterization of aberrant splicing events induced by lentiviral vector integrations. *The Journal of clinical investigation* **122**: 1667-1676.

Chalamcharla VR, Curcio MJ, Belfort M. 2010. Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. *Genes & development* **24**: 827-836.

Chalberg TW, Portlock JL, Olivares EC, Thyagarajan B, Kirby PJ, Hillman RT, Hoelters J, Calos MP. 2006. Integration specificity of phage phiC31 integrase in the human genome. *Journal of molecular biology* **357**: 28-48.

Chambers SP, Austen DA, Fulghum JR, Kim WM. 2004. High-throughput screening for soluble recombinant expressed kinases in Escherichia coli and insect cells. *Protein expression and purification* **36**: 40-47.

Chan RT, Robart AR, Rajashankar KR, Pyle AM, Toor N. 2012. Crystal structure of a group II intron in the pre-catalytic state. *Nature structural & molecular biology* **19**: 555-557.

Chanfreau G, Jacquier A. 1994. Catalytic site components common to both splicing steps of a group II intron. *Science* **266**: 1383-1387.

-. 1996. An RNA conformational change between the two chemical steps of group II self-splicing. *The EMBO journal* **15**: 3466-3476.

Charneau P, Clavel F. 1991. A single-stranded gap in human immunodeficiency virus unintegrated linear DNA defined by a central copy of the polypurine tract. *Journal of virology* **65**: 2415-2421.

Charrier S, Dupre L, Scaramuzza S, Jeanson-Leh L, Blundell MP, Danos O, Cattaneo F, Aiuti A, Eckenberg R, Thrasher AJ et al. 2007. Lentiviral vectors targeting WASp expression to hematopoietic cells, efficiently transduce and correct cells from WAS patients. *Gene therapy* **14**: 415-428.

Charrier S, Ferrand M, Zerbato M, Precigout G, Viornery A, Bucher-Laurent S, Benkhelifa-Ziyyat S, Merten OW, Perea J, Galy A. 2011. Quantification of lentiviral vector copy numbers in individual hematopoietic colony-forming cells shows vector dose-dependent effects on the frequency and level of transduction. *Gene therapy* **18**: 479-487.

Charrier S, Stockholm D, Seye K, Opolon P, Taveau M, Gross DA, Bucher-Laurent S, Delenda C, Vainchenker W, Danos O et al. 2005. A lentiviral vector encoding the human Wiskott-Aldrich syndrome protein corrects immune and cytoskeletal defects in WASP knockout mice. *Gene therapy* **12**: 597-606.

Chen Y, McClane BA, Fisher DJ, Rood JI, Gupta P. 2005. Construction of an alpha toxin gene knockout mutant of Clostridium perfringens type A by use of a mobile group II intron. *Applied and environmental microbiology* **71**: 7542-7547.

Chen Z, Zhao H. 2005. A highly sensitive selection method for directed evolution of homing endonucleases. *Nucleic acids research* **33**: e154.

Cheng J, Saigo H, Baldi P. 2006. Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins* **62**: 617-629.

Cherepanov P, Devroe E, Silver PA, Engelman A. 2004. Identification of an evolutionarily conserved domain in human lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) that binds HIV-1 integrase. *The Journal of biological chemistry* **279**: 48883-48892.

Chevalier BS, Stoddard BL. 2001. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic acids research* **29**: 3757-3774.

Chin K, Pyle AM. 1995. Branch-point attack in group II introns is a highly reversible transesterification, providing a potential proofreading mechanism for 5'-splice site selection. *RNA* **1**: 391-406.

Choe H, Farzan M, Sun Y, Sullivan N, Rollins B, Ponath PD, Wu L, Mackay CR, LaRosa G, Newman W et al. 1996. The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. *Cell* **85**: 1135-1148.

Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF. 2010. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**: 757-761.

Christianson TW, Sikorski RS, Dante M, Shero JH, Hieter P. 1992. Multifunctional yeast high-copy-number shuttle vectors. *Gene* **110**: 119-122.

Chu VT, Liu Q, Podar M, Perlman PS, Pyle AM. 1998. More than one way to splice an RNA: branching without a bulge and splicing without branching in group II introns. *RNA* **4**: 1186-1202.

Ciuffi A, Diamond TL, Hwang Y, Marshall HM, Bushman FD. 2006. Modulating target site selection during human immunodeficiency virus DNA integration in vitro with an engineered tethering factor. *Human gene therapy* **17**: 960-967.

Ciuffi A, Llano M, Poeschla E, Hoffmann C, Leipzig J, Shinn P, Ecker JR, Bushman F. 2005. A role for LEDGF/p75 in targeting HIV DNA integration. *Nature medicine* **11**: 1287-1289.

Clever JL, Miranda D, Jr., Parslow TG. 2002. RNA structure and packaging signals in the 5' leader region of the human immunodeficiency virus type 1 genome. *Journal of virology* **76**: 12381-12387.

Coffin JM, Hughes SH, Varmus H. 1997. *Retroviruses*. Cold Spring Harbor Laboratory Press, Plainview, N.Y.

Colleaux L, D'Auriol L, Galibert F, Dujon B. 1988. Recognition and cleavage site of the intron-encoded omega transposase. *Proceedings of the National Academy of Sciences of the United States of America* **85**: 6022-6026.

Colman PM, Lawrence MC. 2003. The structural biology of type I viral membrane fusion. *Nature reviews Molecular cell biology* **4**: 309-319.

Connolly ML. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**: 709-713.

-. 1993. The molecular surface package. *Journal of molecular graphics* **11**: 139-141.

Corbi N, Libri V, Fanciulli M, Tinsley JM, Davies KE, Passananti C. 2000. The artificial zinc finger coding gene 'Jazz' binds the utrophin promoter and activates transcription. *Gene therapy* **7**: 1076-1083.

Cornu TI, Cathomen T. 2007. Targeted genome modifications using integrase-deficient lentiviral vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **15**: 2107-2113.

Cornu TI, Thibodeau-Beganny S, Guhl E, Alwin S, Eichtinger M, Joung JK, Cathomen T. 2008. DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**: 352-358.

Coros CJ, Landthaler M, Piazza CL, Beauregard A, Esposito D, Perutka J, Lambowitz AM, Belfort M. 2005. Retrotransposition strategies of the Lactococcus lactis Ll.LtrB group II intron are dictated by host identity and cellular environment. *Molecular microbiology* **56**: 509-524.

Costa M, Christian EL, Michel F. 1998. Differential chemical probing of a group II self-splicing intron identifies bases involved in tertiary interactions and supports an alternative secondary structure model of domain V. *RNA* **4**: 1055-1068.

Costa M, Deme E, Jacquier A, Michel F. 1997a. Multiple tertiary interactions involving domain II of group II self-splicing introns. *Journal of molecular biology* **267**: 520-536.

Costa M, Fontaine JM, Loiseaux-de Goer S, Michel F. 1997b. A group II self-splicing intron from the brown alga Pylaiella littoralis is active at unusually low magnesium concentrations and forms populations of molecules with a uniform conformation. *Journal of molecular biology* **274**: 353-364.

Costa M, Michel F. 1995. Frequent use of the same tertiary motif by self-folding RNAs. *The EMBO journal* **14**: 1276-1285.

-. 1999. Tight binding of the 5' exon to domain I of a group II self-splicing intron requires completion of the intron active site. *The EMBO journal* **18**: 1025-1037.

Costa M, Michel F, Westhof E. 2000. A three-dimensional perspective on exon binding by a group II self-splicing intron. *The EMBO journal* **19**: 5007-5018.

Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M. 1998. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* **94**: 451-462.

Craigie R, Fujiwara T, Bushman F. 1990. The IN protein of Moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration in vitro. *Cell* **62**: 829-837.

Cronin J, Zhang XY, Reiser J. 2005. Altering the tropism of lentiviral vectors through pseudotyping. *Current gene therapy* **5**: 387-398.

Cuchet D, Potel C, Thomas J, Epstein AL. 2007. HSV-1 amplicon vectors: a promising and versatile tool for gene delivery. *Expert opinion on biological therapy* **7**: 975-995.

Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM. 2004. A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to promote RNA splicing. *Journal of molecular biology* **340**: 211-231.

Cui Z, Geurts AM, Liu G, Kaufman CD, Hackett PB. 2002. Structure-function analysis of the inverted terminal repeats of the sleeping beauty transposon. *Journal of molecular biology* **318**: 1221-1235.

Cullin C, Minvielle-Sebastia L. 1994. Multipurpose vectors designed for the fast generation of N- or C-terminal epitope-tagged proteins. *Yeast* **10**: 105-112.

D'Agostino DM, Felber BK, Harrison JE, Pavlakis GN. 1992. The Rev protein of human immunodeficiency virus type 1 promotes polysomal association and translation of gag/pol and vpu/env mRNAs. *Molecular and cellular biology* **12**: 1375-1386.

Daboussi F, Zaslavskiy M, Poirot L, Loperfido M, Gouble A, Guyot V, Leduc S, Galetto R, Grizot S, Oficjalska D et al. 2012. Chromosomal context and epigenetic mechanisms control the efficacy of genome editing by rare-cutting designer endonucleases. *Nucleic acids research* **40**: 6367-6379.

Daecke J, Fackler OT, Dittmar MT, Krausslich HG. 2005. Involvement of clathrin-mediated endocytosis in human immunodeficiency virus type 1 entry. *Journal of virology* **79**: 1581-1594.

Dai L, Chai D, Gu SQ, Gabel J, Noskov SY, Blocker FJ, Lambowitz AM, Zimmerly S. 2008. A three-dimensional model of a group II intron RNA and its interaction with the intron-encoded reverse transcriptase. *Molecular cell* **30**: 472-485.

Dai L, Toor N, Olson R, Keeping A, Zimmerly S. 2003. Database for mobile group II introns. *Nucleic acids research* **31**: 424-426.

Dai L, Zimmerly S. 2002a. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic acids research* **30**: 1091-1102.

-. 2002b. The dispersal of five group II introns among natural populations of Escherichia coli. *RNA* **8**: 1294-1307.

Dalgleish AG, Beverley PC, Clapham PR, Crawford DH, Greaves MF, Weiss RA. 1984. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature* **312**: 763-767.

Daniels DL, Michels WJ, Jr., Pyle AM. 1996. Two competing pathways for self-splicing by group II introns: a quantitative analysis of in vitro reaction rates and products. *Journal of molecular biology* **256**: 31-49.

Dave UP, Jenkins NA, Copeland NG. 2004. Gene therapy insertional mutagenesis insights. *Science* **303**: 333.

Dayie KT. 2005. Resolution enhanced homonuclear carbon decoupled triple resonance experiments for unambiguous RNA structural characterization. *Journal of biomolecular NMR* **32**: 129-139.

de Felipe P, Izquierdo M, Wandosell F, Lim F. 2001. Integrating retroviral cassette extends gene delivery of HSV-1 expression vectors to dividing cells. *BioTechniques* **31**: 394-402, 404-395.

de Lencastre A, Hamill S, Pyle AM. 2005. A single active-site region for a group II intron. *Nature structural & molecular biology* **12**: 626-627.

de Lencastre A, Pyle AM. 2008. Three essential and conserved regions of the group II intron are proximal to the 5'-splice site. *RNA* **14**: 11-24.

de Silva S, Bowers WJ. 2011. Targeting the central nervous system with herpes simplex virus / Sleeping Beauty hybrid amplicon vectors. *Current gene therapy* **11**: 332-340.

de Soultrait VR, Caumont A, Durrens P, Calmels C, Parissi V, Recordon P, Bon E, Desjobert C, Tarrago-Litvak L, Fournier M. 2002. HIV-1 integrase interacts with yeast microtubule-associated proteins. *Biochimica et biophysica acta* **1575**: 40-48.

Deichmann A, Brugman MH, Bartholomae CC, Schwarzwaelder K, Verstegen MM, Howe SJ, Arens A, Ott MG, Hoelzer D, Seger R et al. 2011. Insertion sites in engrafted cells cluster within a limited repertoire of genomic areas after gammaretroviral vector gene therapy. *Molecular therapy : the journal of the American Society of Gene Therapy* **19**: 2031-2039.

Deichmann A, Hacein-Bey-Abina S, Schmidt M, Garrigue A, Brugman MH, Hu J, Glimm H, Gyapay G, Prum B, Fraser CC et al. 2007. Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. *The Journal of clinical investigation* **117**: 2225-2232.

Del Campo M, Mohr S, Jiang Y, Jia H, Jankowsky E, Lambowitz AM. 2009. Unwinding by local strand separation is critical for the function of DEAD-box proteins as RNA chaperones. *Journal of molecular biology* **389**: 674-693.

Del Campo M, Tijerina P, Bhaskaran H, Mohr S, Yang Q, Jankowsky E, Russell R, Lambowitz AM. 2007. Do DEAD-box proteins promote group II intron splicing without unwinding RNA? *Molecular cell* **28**: 159-166.

Delacote F, Perez C, Guyot V, Mikonio C, Potrel P, Cabaniols JP, Delenda C, Paques F, Duchateau P. 2011. Identification of genes regulating gene targeting by a high-throughput screening approach. *Journal of nucleic acids* **2011**: 947212.

Demaison C, Parsley K, Brouns G, Scherr M, Battmer K, Kinnon C, Grez M, Thrasher AJ. 2002. High-level transduction and gene expression in hematopoietic repopulating cells using a human immunodeficiency [correction of imunodeficiency] virus type 1-based lentiviral vector containing an internal spleen focus forming virus promoter. *Human gene therapy* **13**: 803-813.

Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, Di Marzio P, Marmon S, Sutton RE, Hill CM et al. 1996. Identification of a major co-receptor for primary isolates of HIV-1. *Nature* **381**: 661-666.

Derry JM, Ochs HD, Francke U. 1994. Isolation of a novel gene mutated in Wiskott-Aldrich syndrome. *Cell* **79**: following 922.

Di Matteo M, Matrai J, Belay E, Firdissa T, Vandendriessche T, Chuah MK. 2012. PiggyBac toolbox. *Methods Mol Biol* **859**: 241-254.

Di Nunzio F, Maruggi G, Ferrari S, Di Iorio E, Poletti V, Garcia M, Del Rio M, De Luca M, Larcher F, Pellegrini G et al. 2008. Correction of laminin-5 deficiency in human epidermal stem cells by transcriptionally targeted lentiviral vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**: 1977-1985.

Dickson L, Huang HR, Liu L, Matsuura M, Lambowitz AM, Perlman PS. 2001. Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 13207-13212.

Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T. 2005. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**: 473-483.

Doetsch NA, Thompson MD, Hallick RB. 1998. A maturase-encoding group III twintron is conserved in deeply rooted euglenoid species: are group III introns the chicken or the egg? *Molecular biology and evolution* **15**: 76-86.

Doetschman T, Gregg RG, Maeda N, Hooper ML, Melton DW, Thompson S, Smithies O. 1987. Targetted correction of a mutant HPRT gene in mouse embryonic stem cells. *Nature* **330**: 576-578.

Doherty JE, Huye LE, Yusa K, Zhou L, Craig NL, Wilson MH. 2012. Hyperactive piggyBac gene transfer in human cells and in vivo. *Human gene therapy* **23**: 311-320.

Donahue RE, Kessler SW, Bodine D, McDonagh K, Dunbar C, Goodman S, Agricola B, Byrne E, Raffeld M, Moen R et al. 1992. Helper virus induced T cell lymphoma in nonhuman primates after retroviral mediated gene transfer. *The Journal of experimental medicine* **176**: 1125-1135.

Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *Journal of molecular biology* **260**: 649-663.

Doranz BJ, Rucker J, Yi Y, Smyth RJ, Samson M, Peiper SC, Parmentier M, Collman RG, Doms RW. 1996. A dual-tropic primary HIV-1 isolate that uses fusin and the beta-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. *Cell* **85**: 1149-1158.

Doyle EL, Booher NJ, Standage DS, Voytas DF, Brendel VP, Vandyk JK, Bogdanove AJ. 2012. TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic acids research* **40**: W117-122.

Doyon Y, Vo TD, Mendel MC, Greenberg SG, Wang J, Xia DF, Miller JC, Urnov FD, Gregory PD, Holmes MC. 2011. Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nature methods* **8**: 74-79.

Dragic T, Litwin V, Allaway GP, Martin SR, Huang Y, Nagashima KA, Cayanan C, Maddon PJ, Koup RA, Moore JP et al. 1996. HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature* **381**: 667-673.

Dreier B, Fuller RP, Segal DJ, Lund CV, Blancafort P, Huber A, Koksch B, Barbas CF, 3rd. 2005. Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors. *The Journal of biological chemistry* **280**: 35588-35597.

Du T, Zamore PD. 2005. microPrimer: the biogenesis and function of microRNA. *Development* **132**: 4645-4652.

Du Y, Jenkins NA, Copeland NG. 2005a. Insertional mutagenesis identifies genes that promote the immortalization of primary bone marrow progenitor cells. *Blood* **106**: 3932-3939.

Du Y, Spence SE, Jenkins NA, Copeland NG. 2005b. Cooperating cancer-gene identification through oncogenic-retrovirus-induced insertional mutagenesis. *Blood* **106**: 2498-2505.

Dujon B. 1989. Group I introns as mobile genetic elements: facts and mechanistic speculations--a review. *Gene* **82**: 91-114.

Dull T, Zufferey R, Kelly M, Mandel RJ, Nguyen M, Trono D, Naldini L. 1998. A third-generation lentivirus vector with a conditional packaging system. *Journal of virology* **72**: 8463-8471.

Dupre L, Marangoni F, Scaramuzza S, Trifari S, Hernandez RJ, Aiuti A, Naldini L, Roncarolo MG. 2006. Efficacy of gene therapy for Wiskott-Aldrich syndrome using a WAS promoter/cDNA-containing lentiviral vector and nonlethal irradiation. *Human gene therapy* **17**: 303-313.

Dupre L, Trifari S, Follenzi A, Marangoni F, Lain de Lera T, Bernad A, Martino S, Tsuchiya S, Bordignon C, Naldini L et al. 2004. Lentiviral vector-mediated gene transfer in T cells from Wiskott-Aldrich syndrome patients leads to functional correction. *Molecular therapy : the journal of the American Society of Gene Therapy* **10**: 903-915.

Eastberg JH, McConnell Smith A, Zhao L, Ashworth J, Shen BW, Stoddard BL. 2007. Thermodynamics of DNA target site recognition by homing endonucleases. *Nucleic acids research* **35**: 7209-7221.

Ehrhardt A, Yant SR, Giering JC, Xu H, Engler JA, Kay MA. 2007. Somatic integration from an adenoviral hybrid vector into a hot spot in mouse liver results in persistent transgene expression levels in vivo. *Molecular therapy : the journal of the American Society of Gene Therapy* **15**: 146-156.

Eklund JL, Ulge UY, Eastberg J, Monnat RJ, Jr. 2007. Altered target site specificity variants of the I-PpoI His-Cys box homing endonuclease. *Nucleic acids research* **35**: 5839-5850.

Eldho NV, Dayie KT. 2007. Internal bulge and tetraloop of the catalytic domain 5 of a group II intron ribozyme are flexible: implications for catalysis. *Journal of molecular biology* **365**: 930-944.

Ellis J, Yao S. 2005. Retrovirus silencing and vector design: relevance to normal and cancer stem cells? *Current gene therapy* **5**: 367-373.

Emery DW, Yannaki E, Tubb J, Stamatoyannopoulos G. 2000. A chromatin insulator protects retrovirus vectors from chromosomal position effects. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 9150-9155.

Ems SC, Morden CW, Dixon CK, Wolfe KH, dePamphilis CW, Palmer JD. 1995. Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant Epifagus virginiana. *Plant molecular biology* **29**: 721-733.

Engelman A, Cherepanov P. 2008. The lentiviral integrase binding protein LEDGF/p75 and HIV-1 replication. *PLoS pathogens* **4**: e1000046.

Engelman A, Craigie R. 1992. Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro. *Journal of virology* **66**: 6361-6369.

Engelman A, Hickman AB, Craigie R. 1994. The core and carboxyl-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding. *Journal of virology* **68**: 5911-5917.

Epinat JC, Arnould S, Chames P, Rochaix P, Desfontaines D, Puzin C, Patin A, Zanghellini A, Paques F, Lacroix E. 2003. A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic acids research* **31**: 2952-2962.

Eriksson P, Thomas LR, Thorburn A, Stillman DJ. 2004. pRS yeast vectors with a LYS2 marker. *BioTechniques* **36**: 212-213.

Eskes R, Yang J, Lambowitz AM, Perlman PS. 1997. Mobility of yeast mitochondrial group II introns: engineering a new site specificity and retrohoming via full reverse splicing. *Cell* **88**: 865-874.

Evans-Galea MV, Wielgosz MM, Hanawa H, Srivastava DK, Nienhuis AW. 2007. Suppression of clonal dominance in cultured human lymphoid cells by addition of the cHS4 insulator to a lentiviral vector. *Molecular therapy : the journal of the American Society of Gene Therapy* **15**: 801-809.

Fajardo-Sanchez E, Stricher F, Paques F, Isalan M, Serrano L. 2008. Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences. *Nucleic acids research* **36**: 2163-2173.

Fassati A, Goff SP. 2001. Characterization of intracellular reverse transcription complexes of human immunodeficiency virus type 1. *Journal of virology* **75**: 3626-3635.

Fedorova O, Mitros T, Pyle AM. 2003. Domains 2 and 3 interact to form critical elements of the group II intron active site. *Journal of molecular biology* **330**: 197-209.

Fedorova O, Pyle AM. 2005. Linking the group II intron catalytic domains: tertiary contacts and structural features of domain 3. *The EMBO journal* **24**: 3906-3916.

Fedorova O, Waldsich C, Pyle AM. 2007. Group II intron folding under near-physiological conditions: collapsing to the near-native state. *Journal of molecular biology* **366**: 1099-1114.

Fedorova O, Zingler N. 2007. Group II introns: structure, folding and splicing mechanism. *Biological chemistry* **388**: 665-678.

Fehse B, Kustikova OS, Bubenheim M, Baum C. 2004. Pois(s)on--it's a question of dose. *Gene therapy* **11**: 879-881.

Felzien LK, Woffendin C, Hottiger MO, Subbramanian RA, Cohen EA, Nabel GJ. 1998. HIV transcriptional activation by the accessory protein, VPR, is mediated by the p300 co-activator. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 5281-5286.

Feng Y, Broder CC, Kennedy PE, Berger EA. 1996. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science* **272**: 872-877.

Ferat JL, Le Gouar M, Michel F. 1994. Multiple group II self-splicing introns in mobile DNA from Escherichia coli. *Comptes rendus de l'Academie des sciences Serie III, Sciences de la vie* **317**: 141-148.

Ferat JL, Michel F. 1993. Group II self-splicing introns in bacteria. *Nature* **364**: 358-361.

Ferrer M, Chernikova TN, Yakimov MM, Golyshin PN, Timmis KN. 2003. Chaperonins govern growth of Escherichia coli at low temperatures. *Nature biotechnology* **21**: 1266-1267.

Follenzi A, Ailles LE, Bakovic S, Geuna M, Naldini L. 2000. Gene transfer by lentiviral vectors is limited by nuclear translocation and rescued by HIV-1 pol sequences. *Nature genetics* **25**: 217-222.

Fontaine JM, Goux D, Kloareg B, Loiseaux-de Goer S. 1997. The reverse-transcriptase-like proteins encoded by group II introns in the mitochondrial genome of the brown alga Pylaiella littoralis belong to two different lineages which apparently coevolved with the group II ribosyme lineages. *Journal of molecular evolution* **44**: 33-42.

Fontaine JM, Rousvoal S, Leblanc C, Kloareg B, Loiseaux-de Goer S. 1995. The mitochondrial LSU rDNA of the brown alga Pylaiella littoralis reveals alpha-proteobacterial features and is split by four group IIB introns with an atypical phylogeny. *Journal of molecular biology* **251**: 378-389.

Fortin JF, Cantin R, Tremblay MJ. 1998. T cells expressing activated LFA-1 are more susceptible to infection with human immunodeficiency virus type 1 particles bearing host-encoded ICAM-1. *Journal of virology* **72**: 2105-2112.

Frazier CL, San Filippo J, Lambowitz AM, Mills DA. 2003. Genetic manipulation of Lactococcus lactis by using targeted group II introns: generation of stable insertions without selection. *Applied and environmental microbiology* **69**: 1121-1128.

Frecha C, Costa C, Negre D, Gauthier E, Russell SJ, Cosset FL, Verhoeyen E. 2008. Stable transduction of quiescent T cells without induction of cycle progression by a novel lentiviral vector pseudotyped with measles virus glycoproteins. *Blood* **112**: 4843-4852.

Friedmann T. 1992. A brief history of gene therapy. *Nature genetics* **2**: 93-98.

Funke S, Maisner A, Muhlebach MD, Koehl U, Grez M, Cattaneo R, Cichutek K, Buchholz CJ. 2008. Targeted cell entry of lentiviral vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**: 1427-1436.

Gabriel R, Lombardo A, Arens A, Miller JC, Genovese P, Kaeppel C, Nowrouzi A, Bartholomae CC, Wang J, Friedman G et al. 2011. An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nature biotechnology* **29**: 816-823.

Gallaher WR. 1987. Detection of a fusion peptide sequence in the transmembrane protein of human immunodeficiency virus. *Cell* **50**: 327-328.

Galvan DL, Nakazawa Y, Kaja A, Kettlun C, Cooper LJ, Rooney CM, Wilson MH. 2009. Genome-wide mapping of PiggyBac transposon integrations in primary human T cells. *J Immunother* **32**: 837-844.

Galy A, Roncarolo MG, Thrasher AJ. 2008. Development of lentiviral gene therapy for Wiskott Aldrich syndrome. *Expert opinion on biological therapy* **8**: 181-190.

Galy A, Thrasher AJ. 2011. Gene therapy for the Wiskott-Aldrich syndrome. *Current opinion in allergy and clinical immunology* **11**: 545-550.

Gao K, Wong S, Bushman F. 2004. Metal binding by the D,DX35E motif of human immunodeficiency virus type 1 integrase: selective rescue of Cys substitutions by Mn2+ in vitro. *Journal of virology* **78**: 6715-6722.

Garcia-Rodriguez FM, Barrientos-Duran A, Diaz-Prado V, Fernandez-Lopez M, Toro N. 2011. Use of RmInt1, a group IIB intron lacking the intron-encoded protein endonuclease domain, in gene targeting. *Applied and environmental microbiology* **77**: 854-861.

Garcia JV, Miller AD. 1991. Serine phosphorylation-independent downregulation of cell-surface CD4 by nef. *Nature* **350**: 508-511.

Garrison BS, Yant SR, Mikkelsen JG, Kay MA. 2007. Postintegrative gene silencing within the Sleeping Beauty transposition system. *Molecular and cellular biology* **27**: 8824-8833.

Garrus JE, von Schwedler UK, Pornillos OW, Morham SG, Zavitz KH, Wang HE, Wettstein DA, Stray KM, Cote M, Rich RL et al. 2001. Tsg101 and the vacuolar protein sorting pathway are essential for HIV-1 budding. *Cell* **107**: 55-65.

Gasior SL, Olivares H, Ear U, Hari DM, Weichselbaum R, Bishop DK. 2001. Assembly of RecA-like recombinases: distinct roles for mediator proteins in mitosis and meiosis. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 8411-8418.

Gaspar HB, Cooray S, Gilmour KC, Parsley KL, Adams S, Howe SJ, Al Ghonaium A, Bayford J, Brown L, Davies EG et al. 2011. Long-term persistence of a polyclonal T cell repertoire after gene therapy for X-linked severe combined immunodeficiency. *Science translational medicine* **3**: 97ra79.

Gaspar HB, Parsley KL, Howe S, King D, Gilmour KC, Sinclair J, Brouns G, Schmidt M, Von Kalle C, Barington T et al. 2004. Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet* **364**: 2181-2187.

Gaussin A, Modlich U, Bauche C, Niederlander NJ, Schambach A, Duros C, Artus A, Baum C, Cohen-Haguenauer O, Mermod N. 2012. CTF/NF1 transcription factors act as potent genetic insulators for integrating gene transfer vectors. *Gene therapy* **19**: 15-24.

Gennari F, Lopes L, Verhoeyen E, Marasco W, Collins MK. 2009. Single-chain antibodies that target lentiviral vectors to MHC class II on antigen-presenting cells. *Human gene therapy* **20**: 554-562.

Geurts AM, Hackett CS, Bell JB, Bergemann TL, Collier LS, Carlson CM, Largaespada DA, Hackett PB. 2006. Structure-based prediction of insertion-site preferences of transposons into chromosomes. *Nucleic acids research* **34**: 2803-2811.

Geurts AM, Yang Y, Clark KJ, Liu G, Cui Z, Dupuy AJ, Bell JB, Largaespada DA, Hackett PB. 2003. Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Molecular therapy : the journal of the American Society of Gene Therapy* **8**: 108-117.

Gijsbers R, Ronen K, Vets S, Malani N, De Rijck J, McNeely M, Bushman FD, Debyser Z. 2010. LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Molecular therapy : the journal of the American Society of Gene Therapy* **18**: 552-560.

Gimble FS, Moure CM, Posey KL. 2003. Assessing the plasticity of DNA target site recognition of the PI-SceI homing endonuclease using a bacterial two-hybrid selection system. *Journal of molecular biology* **334**: 993-1008.

Goldman E, Rosenberg AH, Zubay G, Studier FW. 1995. Consecutive low-usage leucine codons block translation only when near the 5' end of a message in Escherichia coli. *Journal of molecular biology* **245**: 467-473.

Goncalves MA, Holkers M, Cudre-Mauroux C, van Nierop GP, Knaan-Shanzer S, van der Velde I, Valerio D, de Vries AA. 2006. Transduction of myogenic cells by retargeted dual high-capacity hybrid viral vectors: robust dystrophin synthesis in duchenne muscular dystrophy muscle cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **13**: 976-986.

Goncalves MA, Holkers M, van Nierop GP, Wieringa R, Pau MG, de Vries AA. 2008. Targeted chromosomal insertion of large DNA into the human genome by a fiber-modified high-capacity adenovirus-based vector system. *PloS one* **3**: e3084.

Goncalves MA, van Nierop GP, Tijssen MR, Lefesvre P, Knaan-Shanzer S, van der Velde I, van Bekkum DW, Valerio D, de Vries AA. 2005. Transfer of the full-length dystrophin-coding sequence into muscle cells by a dual high-capacity hybrid viral vector with site-specific integration ability. *Journal of virology* **79**: 3146-3162.

Gotte M, Maier G, Onori AM, Cellai L, Wainberg MA, Heumann H. 1999. Temporal coordination between initiation of HIV (+)-strand DNA synthesis and primer removal. *The Journal of biological chemistry* **274**: 11159-11169.

Gould SJ, Booth AM, Hildreth JE. 2003. The Trojan exosome hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 10592-10597.

Goulding CW, Perry LJ. 2003. Protein production in Escherichia coli for structural studies by X-ray crystallography. *Journal of structural biology* **142**: 133-143.

Grabundzija I, Irgang M, Mates L, Belay E, Matrai J, Gogol-Doring A, Kawakami K, Chen W, Ruiz P, Chuah MK et al. 2010. Comparative analysis of transposable element vector systems in human cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **18**: 1200-1209.

Granlund M, Michel F, Norgren M. 2001. Mutually exclusive distribution of IS1548 and GBSi1, an active group II intron identified in human isolates of group B streptococci. *Journal of bacteriology* **183**: 2560-2569.

Gregan J, Bui DM, Pillich R, Fink M, Zsurka G, Schweyen RJ. 2001a. The mitochondrial inner membrane protein Lpe10p, a homologue of Mrs2p, is essential for magnesium homeostasis and group II intron splicing in yeast. *Molecular & general genetics : MGG* **264**: 773-781.

Gregan J, Kolisek M, Schweyen RJ. 2001b. Mitochondrial Mg(2+) homeostasis is critical for group II intron splicing in vivo. *Genes & development* **15**: 2229-2237.

Grez M, Reichenbach J, Schwable J, Seger R, Dinauer MC, Thrasher AJ. 2011. Gene therapy of chronic granulomatous disease: the engraftment dilemma. *Molecular therapy : the journal of the American Society of Gene Therapy* **19**: 28-35.

Griesenbach U, Alton EW. 2009. Gene transfer to the lung: lessons learned from more than 2 decades of CF gene therapy. *Advanced drug delivery reviews* **61**: 128-139.

Grindley ND, Whiteson KL, Rice PA. 2006. Mechanisms of site-specific recombination. *Annual review of biochemistry* **75**: 567-605.

Grivell LA. 1995. Nucleo-mitochondrial interactions in mitochondrial gene expression. *Critical reviews in biochemistry and molecular biology* **30**: 121-164.

Grizot S, Epinat JC, Thomas S, Duclert A, Rolland S, Paques F, Duchateau P. 2010. Generation of redesigned homing endonucleases comprising DNA-binding domains derived from two different scaffolds. *Nucleic acids research* **38**: 2006-2018.

Grizot S, Smith J, Daboussi F, Prieto J, Redondo P, Merino N, Villate M, Thomas S, Lemaire L, Montoya G et al. 2009. Efficient targeting of a SCID gene by an engineered single-chain homing endonuclease. *Nucleic acids research* **37**: 5405-5419.

Groth AC, Olivares EC, Thyagarajan B, Calos MP. 2000. A phage integrase directs efficient site-specific integration in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 5995-6000.

Gruter P, Tabernero C, von Kobbe C, Schmitt C, Saavedra C, Bachi A, Wilm M, Felber BK, Izaurralde E. 1998. TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. *Molecular cell* **1**: 649-659.

Gu SQ, Cui X, Mou S, Mohr S, Yao J, Lambowitz AM. 2010. Genetic identification of potential RNA-binding regions in a group II intron-encoded reverse transcriptase. *RNA* **16**: 732-747.

Guirouilh-Barbat J, Huck S, Bertrand P, Pirzio L, Desmaze C, Sabatier L, Lopez BS. 2004. Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells. *Molecular cell* **14**: 611-623.

Guo H, Karberg M, Long M, Jones JP, 3rd, Sullenger B, Lambowitz AM. 2000. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* **289**: 452-457.

Guo H, Zimmerly S, Perlman PS, Lambowitz AM. 1997. Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *The EMBO journal* **16**: 6835-6848.

Haas DL, Case SS, Crooks GM, Kohn DB. 2000. Critical factors influencing stable transduction of human CD34(+) cells with HIV-1-derived lentiviral vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **2**: 71-80.

Hacein-Bey-Abina S, Garrigue A, Wang GP, Soulier J, Lim A, Morillon E, Clappier E, Caccavelli L, Delabesse E, Beldjord K et al. 2008. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *The Journal of clinical investigation* **118**: 3132-3142.

Hacein-Bey-Abina S, Hauer J, Lim A, Picard C, Wang GP, Berry CC, Martinache C, Rieux-Laucat F, Latour S, Belohradsky BH et al. 2010. Efficacy of gene therapy for X-linked severe combined immunodeficiency. *The New England journal of medicine* **363**: 355-364.

Hacein-Bey-Abina S, von Kalle C, Schmidt M, Le Deist F, Wulffraat N, McIntyre E, Radford I, Villeval JL, Fraser CC, Cavazzana-Calvo M et al. 2003a. A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *The New England journal of medicine* **348**: 255-256.

Hacein-Bey-Abina S, Von Kalle C, Schmidt M, McCormack MP, Wulffraat N, Leboulch P, Lim A, Osborne CS, Pawliuk R, Morillon E et al. 2003b. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**: 415-419.

Hackett PB, Largaespada DA, Cooper LJ. 2010. A transposon and transposase system for human application. *Molecular therapy : the journal of the American Society of Gene Therapy* **18**: 674-683.

Halls C, Mohr S, Del Campo M, Yang Q, Jankowsky E, Lambowitz AM. 2007. Involvement of DEAD-box proteins in group I and group II intron splicing. Biochemical characterization of Mss116p, ATP hydrolysis-dependent and -independent mechanisms, and general RNA chaperone activity. *Journal of molecular biology* **365**: 835-855.

Hamer DH, Leder P. 1979. Expression of the chromosomal mouse Beta maj-globin gene cloned in SV40. *Nature* **281**: 35-40.

Hamill S, Pyle AM. 2006. The receptor for branch-site docking within a group II intron active site. *Molecular cell* **23**: 831-840.

Hammarstrom M, Hellgren N, van Den Berg S, Berglund H, Hard T. 2002. Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. *Protein science : a publication of the Protein Society* **11**: 313-321.

Hanawa H, Kelly PF, Nathwani AC, Persons DA, Vandergriff JA, Hargrove P, Vanin EF, Nienhuis AW. 2002. Comparison of various envelope proteins for their ability to pseudotype lentiviral vectors and transduce primitive hematopoietic cells from human blood. *Molecular therapy : the journal of the American Society of Gene Therapy* **5**: 242-251.

Hargrove PW, Kepes S, Hanawa H, Obenauer JC, Pei D, Cheng C, Gray JT, Neale G, Persons DA. 2008. Globin lentiviral vector insertions can perturb the expression of endogenous genes in beta-thalassemic hematopoietic cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**: 525-533.

Harris-Kerr CL, Zhang M, Peebles CL. 1993. The phylogenetically predicted base-pairing interaction between alpha and alpha' is required for group II splicing in vitro. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 10658-10662.

Hartl FU, Hayer-Hartl M. 2002. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**: 1852-1858.

Hartley JL, Donelson JE. 1980. Nucleotide sequence of the yeast plasmid. *Nature* **286**: 860-865.

Hausl MA, Zhang W, Muther N, Rauschhuber C, Franck HG, Merricks EP, Nichols TC, Kay MA, Ehrhardt A. 2010. Hyperactive sleeping beauty transposase enables persistent phenotypic correction in mice and a canine model for hemophilia B. *Molecular therapy : the journal of the American Society of Gene Therapy* **18**: 1896-1906.

Hayashita Y, Osada H, Tatematsu Y, Yamada H, Yanagisawa K, Tomida S, Yatabe Y, Kawahara K, Sekido Y, Takahashi T. 2005. A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer research* **65**: 9628-9632.

Hayes F. 2003. Transposon-based strategies for microbial functional genomics and proteomics. *Annual review of genetics* **37**: 3-29.

He J, Choe S, Walker R, Di Marzio P, Morgan DO, Landau NR. 1995. Human immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of the cell cycle by inhibiting p34cdc2 activity. *Journal of virology* **69**: 6705-6711.

He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ et al. 2005. A microRNA polycistron as a potential human oncogene. *Nature* **435**: 828-833.

Heap JT, Pennington OJ, Cartman ST, Carter GP, Minton NP. 2007. The ClosTron: a universal gene knock-out system for the genus Clostridium. *Journal of microbiological methods* **70**: 452-464.

Heath PJ, Stephens KM, Monnat RJ, Jr., Stoddard BL. 1997. The structure of I-Crel, a group I intron-encoded homing endonuclease. *Nature structural biology* **4**: 468-476.

Heggestad AD, Notterpek L, Fletcher BS. 2004. Transposon-based RNAi delivery system for generating knockdown cell lines. *Biochemical and biophysical research communications* **316**: 643-650.

Heinzinger NK, Bukinsky MI, Haggerty SA, Ragland AM, Kewalramani V, Lee MA, Gendelman HE, Ratner L, Stevenson M, Emerman M. 1994. The Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 7311-7315.

Heister T, Heid I, Ackermann M, Fraefel C. 2002. Herpes simplex virus type 1/adeno-associated virus hybrid vectors mediate site-specific integration at the adeno-associated virus preintegration site, AAVS1, on human chromosome 19. *Journal of virology* **76**: 7163-7173.

Henaut A, Danchin A. 1996. in *Escherichia coli and Salmonella typhimurium cellular and molecular biology, vol 2* (ed. ASo Microbiology), pp. 2047-2066. Neidhardt, F.; Curtiss III, R.; Ingraham, J.; Lin, E.; Low, B.; Magasanik, B.; Reznikoff, W.; Riley, M.; Schaechter, M. and Umbarger, H.

Henderson LE, Krutzsch HC, Oroszlan S. 1983. Myristyl amino-terminal acylation of murine retrovirus proteins: an unusual post-translational proteins modification. *Proceedings of the National Academy of Sciences of the United States of America* **80**: 339-343.

Henriksen NM, Davis DR, Cheatham Iii TE. 2012. Molecular dynamics re-refinement of two different small RNA loop structures using the original NMR data suggest a common structure. *Journal of biomolecular NMR* **53**: 321-339.

Hermonat PL, Muzyczka N. 1984. Use of adeno-associated virus as a mammalian DNA cloning vector: transduction of neomycin resistance into mammalian tissue culture cells. *Proceedings of the National Academy of Sciences of the United States of America* **81**: 6466-6470.

Hindmarsh P, Leis J. 1999. Retroviral DNA integration. *Microbiology and molecular biology reviews : MMBR* **63**: 836-843, table of contents.

Hioki H, Kameda H, Nakamura H, Okunomiya T, Ohira K, Nakamura K, Kuroda M, Furuta T, Kaneko T. 2007. Efficient gene transduction of neurons by lentivirus with enhanced neuron-specific promoters. *Gene therapy* **14**: 872-882.

Hockemeyer D, Soldner F, Beard C, Gao Q, Mitalipova M, DeKelver RC, Katibah GE, Amora R, Boydston EA, Zeitler B et al. 2009. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature biotechnology* **27**: 851-857.

Hockemeyer D, Wang H, Kiani S, Lai CS, Gao Q, Cassady JP, Cost GJ, Zhang L, Santiago Y, Miller JC et al. 2011. Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**: 731-734.

Holman AG, Coffin JM. 2005. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 6103-6107.

Holt N, Wang J, Kim K, Friedman G, Wang X, Taupin V, Crooks GM, Kohn DB, Gregory PD, Holmes MC et al. 2010. Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nature biotechnology* **28**: 839-847.

Hovanessian AG, Briand JP, Said EA, Svab J, Ferris S, Dali H, Muller S, Desgranges C, Krust B. 2004. The caveolin-1 binding domain of HIV-1 glycoprotein gp41 is an efficient B cell epitope vaccine candidate against virus infection. *Immunity* **21**: 617-627.

Howe SJ, Mansour MR, Schwarzwaelder K, Bartholomae C, Hubank M, Kempski H, Brugman MH, Pike-Overzet K, Chatters SJ, de Ridder D et al. 2008. Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *The Journal of clinical investigation* **118**: 3143-3150.

Huang HR, Chao MY, Armstrong B, Wang Y, Lambowitz AM, Perlman PS. 2003. The DIVa maturase binding site in the yeast group II intron aI2 is essential for intron homing but not for in vivo splicing. *Molecular and cellular biology* **23**: 8809-8819.

Huang HR, Rowe CE, Mohr S, Jiang Y, Lambowitz AM, Perlman PS. 2005. The splicing of yeast mitochondrial group I and group II introns requires a DEAD-box protein with RNA chaperone function. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 163-168.

Huang X, Guo H, Tammana S, Jung YC, Mellgren E, Bassi P, Cao Q, Tu ZJ, Kim YC, Ekker SC et al. 2010. Gene transfer efficiency and genome-wide integration profiling of Sleeping Beauty, Tol2, and piggyBac transposons in human primary T cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **18**: 1803-1813.

Hug P, Lin HM, Korte T, Xiao X, Dimitrov DS, Wang JM, Puri A, Blumenthal R. 2000. Glycosphingolipids promote entry of a broad range of human immunodeficiency virus type 1 isolates into cell lines expressing CD4, CXCR4, and/or CCR5. *Journal of virology* **74**: 6377-6385.

Hunke S, Betton JM. 2003. Temperature effect on inclusion body formation and stress response in the periplasm of Escherichia coli. *Molecular microbiology* **50**: 1579-1589.

Ichiyanagi K, Beauregard A, Belfort M. 2003. A bacterial group II intron favors retrotransposition into plasmid targets. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 15742-15747.

Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M. 2002. Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Molecular microbiology* **46**: 1259-1272.

Ikeda R, Kokubu C, Yusa K, Keng VW, Horie K, Takeda J. 2007. Sleeping beauty transposase has an affinity for heterochromatin conformation. *Molecular and cellular biology* **27**: 1665-1676.

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of molecular biology* **146**: 1-21.

Ikuta K, Kawai H, Muller DG, Ohama T. 2008. Recurrent invasion of mitochondrial group II introns in specimens of Pylaiella littoralis (brown alga), collected worldwide. *Current genetics* **53**: 207-216.

Im DS, Muzyczka N. 1990. The AAV origin binding protein Rep68 is an ATP-dependent site-specific endonuclease with DNA helicase activity. *Cell* **61**: 447-457.

Inouye M, Inouye S. 1991. msDNA and bacterial reverse transcriptase. *Annual review of microbiology* **45**: 163-186.

Inouye S, Inouye M. 1995. Structure, function, and evolution of bacterial reverse transcriptase. *Virus genes* **11**: 81-94.

Ivics Z, Hackett PB, Plasterk RH, Izsvak Z. 1997. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**: 501-510.

Ivics Z, Katzer A, Stuwe EE, Fiedler D, Knespel S, Izsvak Z. 2007. Targeted Sleeping Beauty transposition in human cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **15**: 1137-1144.

Ivics Z, Li MA, Mates L, Boeke JD, Nagy A, Bradley A, Izsvak Z. 2009. Transposon-mediated genome manipulation in vertebrates. *Nature methods* **6**: 415-422.

Izsvak Z, Hackett PB, Cooper LJ, Ivics Z. 2010. Translating Sleeping Beauty transposition into cellular therapies: victories and challenges. *BioEssays : news and reviews in molecular, cellular and developmental biology* **32**: 756-767.

Izsvak Z, Ivics Z, Plasterk RH. 2000. Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *Journal of molecular biology* **302**: 93-102.

Izsvak Z, Khare D, Behlke J, Heinemann U, Plasterk RH, Ivics Z. 2002. Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition. *The Journal of biological chemistry* **277**: 34581-34588.

Izsvak Z, Stuwe EE, Fiedler D, Katzer A, Jeggo PA, Ivics Z. 2004. Healing the wounds inflicted by sleeping beauty transposition by double-strand break repair in mammalian somatic cells. *Molecular cell* **13**: 279-290.

Jacks T, Power MD, Masiarz FR, Luciw PA, Barr PJ, Varmus HE. 1988. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331**: 280-283.

Jacobson SG, Cideciyan AV, Ratnakaram R, Heon E, Schwartz SB, Roman AJ, Peden MC, Aleman TS, Boye SL, Sumaroka A et al. 2012. Gene therapy for leber congenital amaurosis caused by RPE65 mutations: safety and efficacy in 15 children and adults followed up to 3 years. *Archives of ophthalmology* **130**: 9-24.

Jacquier A. 1990. Self-splicing group II and nuclear pre-mRNA introns: how similar are they? *Trends in biochemical sciences* **15**: 351-354.

Jacquier A, Dujon B. 1985. An intron-encoded protein is active in a gene conversion process that spreads an intron into a mitochondrial gene. *Cell* **41**: 383-394.

Jacquier A, Jacquesson-Breuleux N. 1991. Splice site selection and role of the lariat in a group II intron. *Journal of molecular biology* **219**: 415-428.

Jacquier A, Michel F. 1987. Multiple exon-binding sites in class II self-splicing introns. *Cell* **50**: 17-29.

-. 1990. Base-pairing interactions involving the 5' and 3'-terminal nucleotides of group II self-splicing introns. *Journal of molecular biology* **213**: 437-447.

Jarrell KA, Peebles CL, Dietrich RC, Romiti SL, Perlman PS. 1988. Group II intron self-splicing. Alternative reaction conditions yield novel products. *The Journal of biological chemistry* **263**: 3432-3439.

Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi SH, Couloux A, Lee SW, Yoon SH, Cattolico L et al. 2009. Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). *Journal of molecular biology* **394**: 644-652.

Jimenez-Zurdo JI, Garcia-Rodriguez FM, Barrientos-Duran A, Toro N. 2003. DNA target site requirements for homing in vivo of a bacterial group II intron encoding a protein lacking the DNA endonuclease domain. *Journal of molecular biology* **326**: 413-423.

Johnson LA, Morgan RA, Dudley ME, Cassard L, Yang JC, Hughes MS, Kammula US, Royal RE, Sherry RM, Wunderlich JR et al. 2009. Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood* **114**: 535-546.

Jones JP, 3rd, Kierlin MN, Coon RG, Perutka J, Lambowitz AM, Sullenger BA. 2005. Retargeting mobile group II introns to repair mutant genes. *Molecular therapy : the journal of the American Society of Gene Therapy* **11**: 687-694.

Jurica MS, Monnat RJ, Jr., Stoddard BL. 1998. DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI. *Molecular cell* **2**: 469-476.

Kadyk LC, Hartwell LH. 1992. Sister chromatids are preferred over homologs as substrates for recombinational repair in Saccharomyces cerevisiae. *Genetics* **132**: 387-402.

Kaji K, Norrby K, Paca A, Mileikovsky M, Mohseni P, Woltjen K. 2009. Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature* **458**: 771-775.

Kane JF. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Current opinion in biotechnology* **6**: 494-500.

Kang EM, Choi U, Theobald N, Linton G, Long Priel DA, Kuhns D, Malech HL. 2010. Retrovirus gene therapy for X-linked chronic granulomatous disease can achieve stable long-term correction of oxidase activity in peripheral blood neutrophils. *Blood* **115**: 783-791.

Kang Y, Zhang X, Jiang W, Wu C, Chen C, Zheng Y, Gu J, Xu C. 2009. Tumor-directed gene therapy in mice using a composite nonviral gene delivery system consisting of the piggyBac transposon and polyethylenimine. *BMC cancer* **9**: 126.

Kaplitt MG, Feigin A, Tang C, Fitzsimons HL, Mattis P, Lawlor PA, Bland RJ, Young D, Strybing K, Eidelberg D et al. 2007. Safety and tolerability of gene therapy with an adeno-associated virus (AAV) borne GAD gene for Parkinson's disease: an open label, phase I trial. *Lancet* **369**: 2097-2105.

Karberg M, Guo H, Zhong J, Coon R, Perutka J, Lambowitz AM. 2001. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nature biotechnology* **19**: 1162-1167.

Katz RA, Merkel G, Skalka AM. 1996. Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles. *Virology* **217**: 178-190.

Kaushik R, Ratner L. 2004. Role of human immunodeficiency virus type 1 matrix phosphorylation in an early postentry step of virus replication. *Journal of virology* **78**: 2319-2326.

Kawakami K. 2007. Tol2: a versatile gene transfer vector in vertebrates. *Genome biology* **8 Suppl 1**: S7.

Kawakami K, Noda T. 2004. Transposition of the Tol2 element, an Ac-like element from the Japanese medaka fish Oryzias latipes, in mouse embryonic stem cells. *Genetics* **166**: 895-899.

Kay S, Hahn S, Marois E, Hause G, Bonas U. 2007. A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science* **318**: 648-651.

Keating KS, Toor N, Perlman PS, Pyle AM. 2010. A structural analysis of the group II intron active site and implications for the spliceosome. *RNA* **16**: 1-9.

Keating KS, Toor N, Pyle AM. 2008. The GANC tetraloop: a novel motif in the group IIC intron structure. *Journal of molecular biology* **383**: 475-481.

Kennell JC, Moran JV, Perlman PS, Butow RA, Lambowitz AM. 1993. Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. *Cell* **73**: 133-146.

Keravala A, Chavez CL, Hu G, Woodard LE, Monahan PE, Calos MP. 2011. Long-term phenotypic correction in factor IX knockout mice by using PhiC31 integrase-mediated gene therapy. *Gene therapy* **18**: 842-848.

Keravala A, Lee S, Thyagarajan B, Olivares EC, Gabrovsky VE, Woodard LE, Calos MP. 2009. Mutational derivatives of PhiC31 integrase with increased efficiency and specificity. *Molecular therapy : the journal of the American Society of Gene Therapy* **17**: 112-120.

Keren I, Bezawork-Geleta A, Kolton M, Maayan I, Belausov E, Levy M, Mett A, Gidoni D, Shaya F, Ostersetzer-Biran O. 2009. AtnMat2, a nuclear-encoded maturase required for splicing of group-II introns in Arabidopsis mitochondria. *RNA* **15**: 2299-2311.

Khorchid A, Javanbakht H, Wise S, Halwani R, Parniak MA, Wainberg MA, Kleiman L. 2000. Sequences within Pr160gag-pol affecting the selective packaging of primer tRNA(Lys3) into HIV-1. *Journal of molecular biology* **299**: 17-26.

Kido M, Yamanaka K, Mitani T, Niki H, Ogura T, Hiraga S. 1996. RNase E polypeptides lacking a carboxyl-terminal half suppress a mukB mutation in Escherichia coli. *Journal of bacteriology* **178**: 3917-3925.

Kim S, Lee SB. 2008. Soluble expression of archaeal proteins in Escherichia coli by using fusion-partners. *Protein expression and purification* **62**: 116-119.

Kim VN, Mitrophanous K, Kingsman SM, Kingsman AJ. 1998. Minimal requirement for a lentivirus vector based on human immunodeficiency virus type 1. *Journal of virology* **72**: 811-816.

Kim YG, Cha J, Chandrasegaran S. 1996. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 1156-1160.

Kim YG, Chandrasegaran S. 1994. Chimeric restriction endonuclease. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 883-887.

Kim YG, Smith J, Durgesha M, Chandrasegaran S. 1998. Chimeric restriction enzyme: Gal4 fusion to FokI cleavage domain. *Biological chemistry* **379**: 489-495.

Kingsman SM, Mitrophanous K, Olsen JC. 2005. Potential oncogene activity of the woodchuck hepatitis post-transcriptional regulatory element (WPRE). *Gene therapy* **12**: 3-4.

Kipp M, Gohring F, Ostendorp T, van Drunen CM, van Driel R, Przybylski M, Fackelmayer FO. 2000. SAF-Box, a conserved protein domain that specifically recognizes scaffold attachment region DNA. *Molecular and cellular biology* **20**: 7480-7489.

Klammt C, Schwarz D, Lohr F, Schneider B, Dotsch V, Bernhard F. 2006. Cell-free expression as an emerging technique for the large scale production of integral membrane protein. *The FEBS journal* **273**: 4141-4153.

Knaan-Shanzer S, van de Watering MJ, van der Velde I, Goncalves MA, Valerio D, de Vries AA. 2005. Endowing human adenovirus serotype 5 vectors with fiber domains of species B greatly enhances gene transfer into human mesenchymal stem cells. *Stem Cells* **23**: 1598-1607.

Knaan-Shanzer S, Van Der Velde I, Havenga MJ, Lemckert AA, De Vries AA, Valerio D. 2001. Highly efficient targeted transduction of undifferentiated human hematopoietic cells by adenoviral vectors displaying fiber knobs of subgroup B. *Human gene therapy* **12**: 1989-2005.

Knight S, Zhang F, Mueller-Kuller U, Bokhoven M, Gupta A, Broughton T, Sha S, Antoniou MN, Brendel C, Grez M et al. 2012. Safer, silencing-resistant lentiviral vectors: Optimization of ubiquitous chromatin opening element (UCOE) through elimination of aberrant splicing. *Journal of virology*.

Knoop V, Altwasser M, Brennicke A. 1997. A tripartite group II intron in mitochondria of an angiosperm plant. *Molecular & general genetics : MGG* **255**: 269-276.

Koch JL, Boulanger SC, Dib-Hajj SD, Hebbar SK, Perlman PS. 1992. Group II introns deleted for multiple substructures retain self-splicing activity. *Molecular and cellular biology* **12**: 1950-1958.

Kohler D, Schmidt-Gattung S, Binder S. 2010. The DEAD-box protein PMH2 is required for efficient group II intron splicing in mitochondria of Arabidopsis thaliana. *Plant molecular biology* **72**: 459-467.

Kohlstaedt LA, Wang J, Friedman JM, Rice PA, Steitz TA. 1992. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* **256**: 1783-1790.

Kool J, Berns A. 2009. High-throughput insertional mutagenesis screens in mice to identify oncogenic networks. *Nature reviews Cancer* **9**: 389-399.

Kost TA, Condreay JP, Jarvis DL. 2005. Baculovirus as versatile vectors for protein expression in insect and mammalian cells. *Nature biotechnology* **23**: 567-575.

Kostriken R, Strathern JN, Klar AJ, Hicks JB, Heffron F. 1983. A site-specific endonuclease essential for mating-type switching in Saccharomyces cerevisiae. *Cell* **35**: 167-174.

Kotin RM, Siniscalco M, Samulski RJ, Zhu XD, Hunter L, Laughlin CA, McLaughlin S, Muzyczka N, Rocchi M, Berns KI. 1990. Site-specific integration by adeno-associated virus. *Proceedings of the National Academy of Sciences of the United States of America* **87**: 2211-2215.

Kotsopoulou E, Kim VN, Kingsman AJ, Kingsman SM, Mitrophanous KA. 2000. A Rev-independent human immunodeficiency virus type 1 (HIV-1)-based vector that exploits a codon-optimized HIV-1 gag-pol gene. *Journal of virology* **74**: 4839-4852.

Kren BT, Unger GM, Sjeklocha L, Trossen AA, Korman V, Diethelm-Okita BM, Reding MT, Steer CJ. 2009. Nanocapsule-delivered Sleeping Beauty mediates therapeutic Factor VIII expression in liver sinusoidal endothelial cells of hemophilia A mice. *The Journal of clinical investigation* **119**: 2086-2099.

Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**: 147-157.

Kuhstoss S, Rao RN. 1991. Analysis of the integration function of the streptomycete bacteriophage phi C31. *Journal of molecular biology* **222**: 897-908.

Kurland C, Gallant J. 1996. Errors of heterologous protein expression. *Current opinion in biotechnology* **7**: 489-493.

Kustikova O, Brugman M, Baum C. 2010. The genomic risk of somatic gene therapy. *Seminars in cancer biology* **20**: 269-278.

Kustikova OS, Geiger H, Li Z, Brugman MH, Chambers SM, Shaw CA, Pike-Overzet K, de Ridder D, Staal FJ, von Keudell G et al. 2007. Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways. *Blood* **109**: 1897-1907.

Kwakman JH, Konings DA, Hogeweg P, Pel HJ, Grivell LA. 1990. Structural analysis of a group II intron by chemical modifications and minimal energy calculations. *Journal of biomolecular structure & dynamics* **8**: 413-430.

Laakso MM, Sutton RE. 2006. Replicative fidelity of lentiviral vectors produced by transient transfection. *Virology* **348**: 406-417.

Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. *Annual review of genetics* **38**: 1-35.

-. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harbor perspectives in biology* **3**: a003616.

Lampson BC, Inouye M, Inouye S. 2005. Retrons, msDNA, and the bacterial genome. *Cytogenetic and genome research* **110**: 491-499.

Lanchy JM, Keith G, Le Grice SF, Ehresmann B, Ehresmann C, Marquet R. 1998. Contacts between reverse transcriptase and the primer strand govern the transition from initiation to elongation of HIV-1 reverse transcription. *The Journal of biological chemistry* **273**: 24425-24432.

Larder BA, Kemp SD, Purifoy DJ. 1989. Infectious potential of human immunodeficiency virus type 1 reverse transcriptase mutants with altered inhibitor sensitivity. *Proceedings of the National Academy of Sciences of the United States of America* **86**: 4803-4807.

Lehmann K, Schmidt U. 2003. Group II introns: structure and catalytic versatility of large natural ribozymes. *Critical reviews in biochemistry and molecular biology* **38**: 249-303.

Lewinski MK, Bisgrove D, Shinn P, Chen H, Hoffmann C, Hannenhalli S, Verdin E, Berry CC, Ecker JR, Bushman FD. 2005. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *Journal of virology* **79**: 6610-6619.

Lewis PF, Emerman M. 1994. Passage through mitosis is required for oncoretroviruses but not for the human immunodeficiency virus. *Journal of virology* **68**: 510-516.

Li-Pook-Than J, Bonen L. 2006. Multiple physical forms of excised group II intron RNAs in wheat mitochondria. *Nucleic acids research* **34**: 2782-2790.

Li CF, Costa M, Michel F. 2011. Linking the branchpoint helix to a newly found receptor allows lariat formation by a group II intron. *The EMBO journal* **30**: 3040-3051.

Li CL, Emery DW. 2008. The cHS4 chromatin insulator reduces gammaretroviral vector silencing by epigenetic modifications of integrated provirus. *Gene therapy* **15**: 49-53.

Li CL, Xiong D, Stamatoyannopoulos G, Emery DW. 2009. Genomic and functional assays demonstrate reduced gammaretroviral vector genotoxicity associated with use of the cHS4 chromatin insulator. *Molecular therapy : the journal of the American Society of Gene Therapy* **17**: 716-724.

Li H, Haurigot V, Doyon Y, Li T, Wong SY, Bhagwat AS, Malani N, Anguela XM, Sharma R, Ivanciu L et al. 2011. In vivo genome editing restores haemostasis in a mouse model of haemophilia. *Nature* **475**: 217-221.

Li H, Pellenz S, Ulge U, Stoddard BL, Monnat RJ, Jr. 2009. Generation of single-chain LAGLIDADG homing endonucleases from native homodimeric precursor proteins. *Nucleic acids research* **37**: 1650-1662.

Li L, Wu LP, Chandrasegaran S. 1992. Functional domains in Fok I restriction endonuclease. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 4275-4279.

Li T, Huang S, Zhao X, Wright DA, Carpenter S, Spalding MH, Weeks DP, Yang B. 2011. Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic acids research* **39**: 6315-6325.

Li Y, Luo L, Rasool N, Kang CY. 1993. Glycosylation is necessary for the correct folding of human immunodeficiency virus gp120 in CD4 binding. *Journal of virology* **67**: 584-588.

Li Z, Dullmann J, Schiedlmeier B, Schmidt M, von Kalle C, Meyer J, Forster M, Stocking C, Wahlers A, Frank O et al. 2002. Murine leukemia induced by retroviral gene marking. *Science* **296**: 497.

Liang F, Han M, Romanienko PJ, Jasin M. 1998. Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 5172-5177.

Lim D, Maas WK. 1989. Reverse transcriptase in bacteria. *Molecular microbiology* **3**: 1141-1144.

Lin CL, Sewell AK, Gao GF, Whelan KT, Phillips RE, Austyn JM. 2000. Macrophage-tropic HIV induces and exploits dendritic cell chemotaxis. *The Journal of experimental medicine* **192**: 587-594.

Liu J, Jeppesen I, Nielsen K, Jensen TG. 2006. Phi c31 integrase induces chromosomal aberrations in primary human fibroblasts. *Gene therapy* **13**: 1188-1190.

Liu L, Mah C, Fletcher BS. 2006. Sustained FVIII expression and phenotypic correction of hemophilia A in neonatal mice using an endothelial-targeted sleeping beauty transposon. *Molecular therapy : the journal of the American Society of Gene Therapy* **13**: 1006-1015.

Liu Q, Perez CF, Wang Y. 2006. Efficient site-specific integration of large transgenes by an enhanced herpes simplex virus/adeno-associated virus hybrid amplicon vector. *Journal of virology* **80**: 1672-1679.

Llano M, Saenz DT, Meehan A, Wongthida P, Peretz M, Walker WH, Teo W, Poeschla EM. 2006. An essential role for LEDGF/p75 in HIV integration. *Science* **314**: 461-464.

Lobel LI, Murphy JE, Goff SP. 1989. The palindromic LTR-LTR junction of Moloney murine leukemia virus is not an efficient substrate for proviral integration. *Journal of virology* **63**: 2629-2637.

Lochelt M, Flugel RM, Aboud M. 1994. The human foamy virus internal promoter directs the expression of the functional Bel 1 transactivator and Bet protein early after infection. *Journal of virology* **68**: 638-645.

Lohe AR, Hartl DL. 1996. Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Molecular biology and evolution* **13**: 549-555.

Lombardo A, Genovese P, Beausejour CM, Colleoni S, Lee YL, Kim KA, Ando D, Urnov FD, Galli C, Gregory PD et al. 2007. Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nature biotechnology* **25**: 1298-1306.

Lopez PJ, Marchand I, Joyce SA, Dreyfus M. 1999. The C-terminal half of RNase E, which organizes the Escherichia coli degradosome, participates in mRNA degradation but not rRNA processing in vivo. *Molecular microbiology* **33**: 188-199.

Luckow VA. 1993. Baculovirus systems for the expression of human gene products. *Current opinion in biotechnology* **4**: 564-572.

Madhani HD, Guthrie C. 1992. A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell* **71**: 803-817.

Maeder ML, Thibodeau-Beganny S, Osiak A, Wright DA, Anthony RM, Eichtinger M, Jiang T, Foley JE, Winfrey RJ, Townsend JA et al. 2008. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Molecular cell* **31**: 294-301.

Maguire AM, High KA, Auricchio A, Wright JF, Pierce EA, Testa F, Mingozzi F, Bennicelli JL, Ying GS, Rossi S et al. 2009. Age-dependent effects of RPE65 gene therapy for Leber's congenital amaurosis: a phase 1 dose-escalation trial. *Lancet* **374**: 1597-1605.

Malhotra M, Srivastava S. 2008. An ipdC gene knock-out of Azospirillum brasilense strain SM and its implications on indole-3-acetic acid biosynthesis and plant growth promotion. *Antonie van Leeuwenhoek* **93**: 425-433.

Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. *Molecular biology and evolution* **16**: 793-805.

Malik P, Arumugam PI, Yee JK, Puthenveetil G. 2005. Successful correction of the human Cooley's anemia beta-thalassemia major phenotype using a lentiviral vector flanked by the chicken hypersensitive site 4 chromatin insulator. *Annals of the New York Academy of Sciences* **1054**: 238-249.

Malim MH, Emerman M. 2008. HIV-1 accessory proteins--ensuring viral survival in a hostile environment. *Cell host & microbe* **3**: 388-398.

Malim MH, Hauber J, Le SY, Maizel JV, Cullen BR. 1989. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* **338**: 254-257.

Manes S, del Real G, Lacalle RA, Lucas P, Gomez-Mouton C, Sanchez-Palomino S, Delgado R, Alcami J, Mira E, Martinez AC. 2000. Membrane raft microdomains mediate lateral assemblies required for HIV-1 infection. *EMBO reports* **1**: 190-196.

Manivasakam P, Weber SC, McElver J, Schiestl RH. 1995. Micro-homology mediated PCR targeting in Saccharomyces cerevisiae. *Nucleic acids research* **23**: 2799-2800.

Mann R, Mulligan RC, Baltimore D. 1983. Construction of a retrovirus packaging mutant and its use to produce helper-free defective retrovirus. *Cell* **33**: 153-159.

Manno CS, Pierce GF, Arruda VR, Glader B, Ragni M, Rasko JJ, Ozelo MC, Hoots K, Blatt P, Konkle B et al. 2006. Successful transduction of liver in hemophilia by AAV-Factor IX and limitations imposed by the host immune response. *Nature medicine* **12**: 342-347.

Mantovani J, Charrier S, Eckenberg R, Saurin W, Danos O, Perea J, Galy A. 2009. Diverse genomic integration of a lentiviral vector developed for the treatment of Wiskott-Aldrich syndrome. *The journal of gene medicine* **11**: 645-654.

Maragathavally KJ, Kaminski JM, Coates CJ. 2006. Chimeric Mos1 and piggyBac transposases result in site-directed integration. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **20**: 1880-1882.

Marangoni F, Bosticardo M, Charrier S, Draghici E, Locci M, Scaramuzza S, Panaroni C, Ponzoni M, Sanvito F, Doglioni C et al. 2009. Evidence for long-term efficacy and safety of gene therapy for Wiskott-Aldrich syndrome in preclinical models. *Molecular therapy : the journal of the American Society of Gene Therapy* **17**: 1073-1082.

Marblestone JG, Edavettal SC, Lim Y, Lim P, Zuo X, Butt TR. 2006. Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein science : a publication of the Protein Society* **15**: 182-189.

Mariani R, Chen D, Schrofelbauer B, Navarro F, Konig R, Bollman B, Munk C, Nymark-McMahon H, Landau NR. 2003. Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell* **114**: 21-31.

Markosyan RM, Cohen FS, Melikyan GB. 2003. HIV-1 envelope proteins complete their folding into six-helix bundles immediately after fusion pore formation. *Molecular biology of the cell* **14**: 926-938.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440**: 41-45.

Martinez-Abarca F, Barrientos-Duran A, Fernandez-Lopez M, Toro N. 2004. The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. *Nucleic acids research* **32**: 2880-2888.

Martínez-Abarca F, García-Rodríguez FM, Muñoz E, Toro N. 1999. Biochemical demonstration of reverse transcriptase activity and splicing efficiency of a bacterial group II intron from Sinorhizobium meliloti. *Nucleic Acids Symp Series*: 117-119.

Martinez-Abarca F, Toro N. 2000. Group II introns in the bacterial world. *Molecular microbiology* **38**: 917-926.

Mastroianni M, Watanabe K, White TB, Zhuang F, Vernon J, Matsuura M, Wallingford J, Lambowitz AM. 2008. Group II intron-based gene targeting reactions in eukaryotes. *PloS one* **3**: e3121.

Mates L, Chuah MK, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, Grzela DP, Schmitt A, Becker K, Matrai J et al. 2009. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nature genetics* **41**: 753-761.

Matsuura M, Noah JW, Lambowitz AM. 2001. Mechanism of maturase-promoted group II intron splicing. *The EMBO journal* **20**: 7259-7270.

Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes & development* **11**: 2910-2924.

Mavilio F, Pellegrini G, Ferrari S, Di Nunzio F, Di Iorio E, Recchia A, Maruggi G, Ferrari G, Provasi E, Bonini C et al. 2006. Correction of junctional epidermolysis bullosa by transplantation of genetically modified epidermal stem cells. *Nature medicine* **12**: 1397-1402.

McCarty DM, Young SM, Jr., Samulski RJ. 2004. Integration of adeno-associated virus (AAV) and recombinant AAV vectors. *Annual review of genetics* **38**: 819-845.

McConnell Smith A, Takeuchi R, Pellenz S, Davis L, Maizels N, Monnat RJ, Jr., Stoddard BL. 2009. Generation of a nicking enzyme that stimulates site-specific gene conversion from the I-AniI LAGLIDADG homing endonuclease. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 5099-5104.

McDonald D, Vodicka MA, Lucero G, Svitkina TM, Borisy GG, Emerman M, Hope TJ. 2002. Visualization of the intracellular behavior of HIV in living cells. *The Journal of cell biology* **159**: 441-452.

Meir YJ, Weirauch MT, Yang HS, Chung PC, Yu RK, Wu SC. 2011. Genome-wide target profiling of piggyBac and Tol2 in HEK 293: pros and cons for gene discovery and gene therapy. *BMC biotechnology* **11**: 28.

Mendell JR, Campbell K, Rodino-Klapac L, Sahenk Z, Shilling C, Lewis S, Bowles D, Gray S, Li C, Galloway G et al. 2010. Dystrophin immunity in Duchenne's muscular dystrophy. *The New England journal of medicine* **363**: 1429-1437.

Merten OW, Charrier S, Laroudie N, Fauchille S, Dugue C, Jenny C, Audit M, Zanta-Boussif MA, Chautard H, Radrizzani M et al. 2011. Large-scale manufacture and characterization of a lentiviral vector produced for clinical ex vivo gene therapy application. *Human gene therapy* **22**: 343-356.

Meunier B, Tian GL, Macadre C, Slonimski PP, Lazowska J. 1990. Group II introns transpose in yeast mitochondria. in *Structure Function and Biogenesis of Energy Transfer Systems* (eds. E Quagliariello, S Papa, F Palmieri, C Saccone), pp. 169-174. Elsevier, Amsterdam.

Meyer BE, Meinkoth JL, Malim MH. 1996. Nuclear transport of human immunodeficiency virus type 1, visna virus, and equine infectious anemia virus Rev proteins: identification of a family of transferable nuclear export signals. *Journal of virology* **70**: 2350-2359.

Michel F, Costa M, Westhof E. 2009. The ribozyme core of group II introns: a structure in want of partners. *Trends in biochemical sciences* **34**: 189-199.

Michel F, Ferat JL. 1995. Structure and activities of group II introns. *Annual review of biochemistry* **64**: 435-461.

Michel F, Jacquier A, Dujon B. 1982. Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie* **64**: 867-881.

Michel F, Lang BF. 1985. Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses. *Nature* **316**: 641-643.

Michel F, Umesono K, Ozeki H. 1989. Comparative and functional anatomy of group II catalytic introns--a review. *Gene* **82**: 5-30.

Michels WJ, Jr., Pyle AM. 1995. Conversion of a group II intron into a new multiple-turnover ribozyme that selectively cleaves oligonucleotides: elucidation of reaction mechanism and structure/function relationships. *Biochemistry* **34**: 2965-2977.

Mikkelsen JG, Yant SR, Meuse L, Huang Z, Xu H, Kay MA. 2003. Helper-Independent Sleeping Beauty transposon-transposase vectors for efficient nonviral gene delivery and persistent gene expression in vivo. *Molecular therapy : the journal of the American Society of Gene Therapy* **8**: 654-665.

Miller DG, Petek LM, Russell DW. 2003. Human gene targeting by adeno-associated virus vectors is enhanced by DNA double-strand breaks. *Molecular and cellular biology* **23**: 3550-3557.

Miller DG, Trobridge GD, Petek LM, Jacobs MA, Kaul R, Russell DW. 2005. Large-scale analysis of adeno-associated virus vector integration sites in normal human cells. *Journal of virology* **79**: 11434-11442.

Miller JC, Holmes MC, Wang J, Guschin DY, Lee YL, Rupniewski I, Beausejour CM, Waite AJ, Wang NS, Kim KA et al. 2007. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* **25**: 778-785.

Miller JC, Tan S, Qiao G, Barlow KA, Wang J, Xia DF, Meng X, Paschon DE, Leung E, Hinkley SJ et al. 2011. A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29**: 143-148.

Miller MD, Farnet CM, Bushman FD. 1997. Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *Journal of virology* **71**: 5382-5390.

Mills DA, McKay LL, Dunny GM. 1996. Splicing of a group II intron involved in the conjugative transfer of pRS01 in lactococci. *Journal of bacteriology* **178**: 3531-3538.

Mingozzi F, Meulenberg JJ, Hui DJ, Basner-Tschakarjan E, Hasbrouck NC, Edmonson SA, Hutnick NA, Betts MR, Kastelein JJ, Stroes ES et al. 2009. AAV-1-mediated gene transfer to skeletal muscle in humans results in dose-dependent activation of capsid-specific T cells. *Blood* **114**: 2077-2086.

Miskey C, Izsvak Z, Plasterk RH, Ivics Z. 2003. The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. *Nucleic acids research* **31**: 6873-6881.

Mitsuyasu RT, Merigan TC, Carr A, Zack JA, Winters MA, Workman C, Bloch M, Lalezari J, Becker S, Thornton L et al. 2009. Phase 2 gene therapy trial of an anti-HIV ribozyme in autologous CD34+ cells. *Nature medicine* **15**: 285-292.

Modlich U, Kustikova OS, Schmidt M, Rudolph C, Meyer J, Li Z, Kamino K, von Neuhoff N, Schlegelberger B, Kuehlcke K et al. 2005. Leukemias following retroviral transfer of multidrug resistance 1 (MDR1) are driven by combinatorial insertional mutagenesis. *Blood* **105**: 4235-4246.

Modlich U, Navarro S, Zychlinski D, Maetzig T, Knoess S, Brugman MH, Schambach A, Charrier S, Galy A, Thrasher AJ et al. 2009. Insertional transformation of hematopoietic cells by self-inactivating lentiviral and gammaretroviral vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **17**: 1919-1928.

Moehle EA, Rock JM, Lee YL, Jouvenot Y, DeKelver RC, Gregory PD, Urnov FD, Holmes MC. 2007. Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 3055-3060.

Mohr G, Perlman PS, Lambowitz AM. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic acids research* **21**: 4991-4997.

Mohr G, Smith D, Belfort M, Lambowitz AM. 2000. Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes & development* **14**: 559-573.

Mohr S, Matsuura M, Perlman PS, Lambowitz AM. 2006. A DEAD-box protein alone promotes group II intron splicing and reverse splicing by acting as an RNA chaperone. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 3569-3574.

Moiani A, Mavilio F. 2012. Alternative splicing caused by lentiviral integration in the human genome. *Methods in enzymology* **507**: 155-169.

Moiani A, Paleari Y, Sartori D, Mezzadra R, Miccio A, Cattoglio C, Cocchiarella F, Lidonnici MR, Ferrari G, Mavilio F. 2012. Lentiviral vector integration in the human genome induces alternative splicing and generates aberrant transcripts. *The Journal of clinical investigation* **122**: 1653-1666.

Moldt B, Miskey C, Staunstrup NH, Gogol-Doring A, Bak RO, Sharma N, Mates L, Izsvak Z, Chen W, Ivics Z et al. 2011. Comparative genomic integration profiling of Sleeping Beauty transposons mobilized with high efficacy from integrase-defective lentiviral vectors in primary human cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **19**: 1499-1510.

Moldt B, Yant SR, Andersen PR, Kay MA, Mikkelsen JG. 2007. Cis-acting gene regulatory activities in the terminal regions of sleeping beauty DNA transposon-based vectors. *Human gene therapy* **18**: 1193-1204.

Molina-Sanchez MD, Martinez-Abarca F, Toro N. 2006. Excision of the Sinorhizobium meliloti group II intron RmInt1 as circles in vivo. *The Journal of biological chemistry* **281**: 28737-28744.

Mondor I, Ugolini S, Sattentau QJ. 1998. Human immunodeficiency virus type 1 attachment to HeLa CD4 cells is CD4 independent and gp120 dependent and requires cell surface heparans. *Journal of virology* **72**: 3623-3634.

Monnat RJ, Jr., Hackmann AF, Cantrell MA. 1999. Generation of highly site-specific DNA double-strand breaks in human cells by the homing endonucleases I-PpoI and I-CreI. *Biochemical and biophysical research communications* **255**: 88-93.

Montini E, Cesana D. 2012. Genotoxicity assay for gene therapy vectors in tumor prone Cdkn2a(-)/(-) mice. *Methods in enzymology* **507**: 171-185.

Montini E, Cesana D, Schmidt M, Sanvito F, Bartholomae CC, Ranzani M, Benedicenti F, Sergi LS, Ambrosi A, Ponzoni M et al. 2009. The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy. *The Journal of clinical investigation* **119**: 964-975.

Montini E, Cesana D, Schmidt M, Sanvito F, Ponzoni M, Bartholomae C, Sergi Sergi L, Benedicenti F, Ambrosi A, Di Serio C et al. 2006. Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. *Nature biotechnology* **24**: 687-696.

Montini E, Held PK, Noll M, Morcinek N, Al-Dhalimy M, Finegold M, Yant SR, Kay MA, Grompe M. 2002. In vivo correction of murine tyrosinemia type I by DNA-mediated transposition. *Molecular therapy : the journal of the American Society of Gene Therapy* **6**: 759-769.

Moolten FL, Cupples LA. 1992. A model for predicting the risk of cancer consequent to retroviral gene therapy. *Human gene therapy* **3**: 479-486.

Moran JV, Mecklenburg KL, Sass P, Belcher SM, Mahnke D, Lewin A, Perlman P. 1994. Splicing defective mutants of the COXI gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. *Nucleic acids research* **22**: 2057-2064.

Moran JV, Zimmerly S, Eskes R, Kennell JC, Lambowitz AM, Butow RA, Perlman PS. 1995. Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. *Molecular and cellular biology* **15**: 2828-2838.

Morbitzer R, Elsaesser J, Hausner J, Lahaye T. 2011. Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic acids research* **39**: 5790-5799.

Moretz SE, Lampson BC. 2010. A group IIC-type intron interrupts the rRNA methylase gene of Geobacillus stearothermophilus strain 10. *Journal of bacteriology* **192**: 5245-5248.

Morral N, Parks RJ, Zhou H, Langston C, Schiedner G, Quinones J, Graham FL, Kochanek S, Beaudet AL. 1998. High doses of a helper-dependent adenoviral vector yield supraphysiological levels of alpha1-antitrypsin with negligible toxicity. *Human gene therapy* **9**: 2709-2716.

Morsy MA, Gu M, Motzel S, Zhao J, Lin J, Su Q, Allen H, Franlin L, Parks RJ, Graham FL et al. 1998. An adenoviral vector deleted for all viral coding sequences results in enhanced safety and extended expression of a leptin transgene. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 7866-7871.

Moscou MJ, Bogdanove AJ. 2009. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**: 1501.

Muller MW, Stocker P, Hetzer M, Schweyen RJ. 1991. Fate of the junction phosphate in alternating forward and reverse self-splicing reactions of group II intron RNA. *Journal of molecular biology* **222**: 145-154.

Mullineux ST, Costa M, Bassi GS, Michel F, Hausner G. 2010. A group II intron encodes a functional LAGLIDADG homing endonuclease and self-splices under moderate temperature and ionic conditions. *RNA* **16**: 1818-1831.

Murphy SJ, Chong H, Bell S, Diaz RM, Vile RG. 2002. Novel integrating adenoviral/retroviral hybrid vector for gene therapy. *Human gene therapy* **13**: 745-760.

Murray HL, Mikheeva S, Coljee VW, Turczyk BM, Donahue WF, Bar-Shalom A, Jarrell KA. 2001. Excision of group II introns as circles. *Molecular cell* **8**: 201-211.

Mussolino C, Cathomen T. 2011. On target? Tracing zinc-finger-nuclease specificity. *Nature methods* **8**: 725-726.

Mussolino C, Morbitzer R, Lutge F, Dannemann N, Lahaye T, Cathomen T. 2011. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids research* **39**: 9283-9293.

Nakagawa N, Sakurai N. 2006. A mutation in At-nMat1a, which encodes a nuclear gene having high similarity to group II intron maturase, causes impaired splicing of mitochondrial NAD4 transcript and altered carbon metabolism in Arabidopsis thaliana. *Plant & cell physiology* **47**: 772-783.

Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic acids research* **28**: 292.

Nakanishi H, Higuchi Y, Kawakami S, Yamashita F, Hashida M. 2010. piggyBac transposon-mediated long-term gene expression in mice. *Molecular therapy : the journal of the American Society of Gene Therapy* **18**: 707-714.

Naldini L, Blomer U, Gallay P, Ory D, Mulligan R, Gage FH, Verma IM, Trono D. 1996. In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272**: 263-267.

Nathwani AC, Davidoff AM, Linch DC. 2005. A review of gene therapy for haematological disorders. *British journal of haematology* **128**: 3-17.

Netzer WJ, Hartl FU. 1997. Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* **388**: 343-349.

Neumann E, Schaefer-Ridder M, Wang Y, Hofschneider PH. 1982. Gene transfer into mouse lyoma cells by electroporation in high electric fields. *The EMBO journal* **1**: 841-845.

Nickoloff JA, Chen EY, Heffron F. 1986. A 24-base-pair DNA sequence from the MAT locus stimulates intergenic recombination in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 7831-7835.

Nisole S, Krust B, Callebaut C, Guichard G, Muller S, Briand JP, Hovanessian AG. 1999. The anti-HIV pseudopeptide HB-19 forms a complex with the cell-surface-expressed nucleolin independent of heparan sulfate proteoglycans. *The Journal of biological chemistry* **274**: 27875-27884.

Niu Y, Tenney K, Li H, Gimble FS. 2008. Engineering variants of the I-SceI homing endonuclease with strand-specific and site-specific DNA-nicking activity. *Journal of molecular biology* **382**: 188-202.

Noah JW, Lambowitz AM. 2003. Effects of maturase binding and Mg2+ concentration on group II intron RNA folding investigated by UV cross-linking. *Biochemistry* **42**: 12466-12480.

O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT. 2005. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* **435**: 839-843.

Ohlfest JR, Frandsen JL, Fritz S, Lobitz PD, Perkinson SG, Clark KJ, Nelsestuen G, Key NS, McIvor RS, Hackett PB et al. 2005. Phenotypic correction and long-term expression of factor VIII in hemophilic mice by immunotolerization and nonviral gene transfer using the Sleeping Beauty transposon system. *Blood* **105**: 2691-2698.

Olivares EC, Hollis RP, Chalberg TW, Meuse L, Kay MA, Calos MP. 2002. Site-specific genomic integration produces therapeutic Factor IX levels in mice. *Nature biotechnology* **20**: 1124-1128.

Olschowka JA, Bowers WJ, Hurley SD, Mastrangelo MA, Federoff HJ. 2003. Helper-free HSV-1 amplicons elicit a markedly less robust innate immune response in the CNS. *Molecular therapy : the journal of the American Society of Gene Therapy* **7**: 218-227.

Ortiz-Urda S, Lin Q, Yant SR, Keene D, Kay MA, Khavari PA. 2003. Sustainable correction of junctional epidermolysis bullosa via transposon-mediated nonviral gene transfer. *Gene therapy* **10**: 1099-1104.

Osborne BI, Baker B. 1995. Movers and shakers: maize transposons as tools for analyzing other plant genomes. *Current opinion in cell biology* **7**: 406-413.

Ott DE. 2002. Potential roles of cellular proteins in HIV-1. *Reviews in medical virology* **12**: 359-374.

Ott MG, Schmidt M, Schwarzwaelder K, Stein S, Siler U, Koehl U, Glimm H, Kuhlcke K, Schilz A, Kunkel H et al. 2006. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nature medicine* **12**: 401-409.

Oudot-Le Secq MP, Fontaine JM, Rousvoal S, Kloareg B, Loiseaux-De Goer S. 2001. The complete sequence of a brown algal mitochondrial genome, the ectocarpale Pylaiella littoralis (L.) Kjellm. *Journal of molecular evolution* **53**: 80-88.

Padgett RA, Podar M, Boulanger SC, Perlman PS. 1994. The stereochemical course of group II intron self-splicing. *Science* **266**: 1685-1688.

Pai SY, Notarangelo LD. 2010. Hematopoietic cell transplantation for Wiskott-Aldrich syndrome: advances in biology and future directions for treatment. *Immunology and allergy clinics of North America* **30**: 179-194.

Pal R, Hoke GM, Sarngadharan MG. 1989. Role of oligosaccharides in the processing and maturation of envelope glycoproteins of human immunodeficiency virus type 1. *Proceedings of the National Academy of Sciences of the United States of America* **86**: 3384-3388.

Papanikolaou E, Georgomanoli M, Stamateris E, Panetsos F, Karagiorga M, Tsaftaridis P, Graphakos S, Anagnou NP. 2012. The new self-inactivating lentiviral vector for thalassemia gene therapy combining two HPFH activating elements corrects human thalassemic hematopoietic stem cells. *Human gene therapy* **23**: 15-31.

Papapetrou EP, Kovalovsky D, Beloeil L, Sant'angelo D, Sadelain M. 2009. Harnessing endogenous miR-181a to segregate transgenic antigen receptor expression in developing versus post-thymic T cells in murine hematopoietic chimeras. *The Journal of clinical investigation* **119**: 157-168.

Paques F, Duchateau P. 2007. Meganucleases and DNA double-strand break-induced recombination: perspectives for gene therapy. *Current gene therapy* **7**: 49-66.

Parks RJ, Chen L, Anton M, Sankar U, Rudnicki MA, Graham FL. 1996. A helper-dependent adenovirus vector system: removal of helper virus by Cre-mediated excision of the viral packaging signal. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 13565-13570.

Pattanayak V, Ramirez CL, Joung JK, Liu DR. 2011. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods* **8**: 765-770.

Patterson RM, Selkirk JK, Merrick BA. 1995. Baculovirus and insect cell gene expression: review of baculovirus biotechnology. *Environmental health perspectives* **103**: 756-759.

Pavletich NP, Pabo CO. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**: 809-817.

Pearson MM, Mobley HL. 2007. The type III secretion system of Proteus mirabilis HI4320 does not contribute to virulence in the mouse model of ascending urinary tract infection. *Journal of medical microbiology* **56**: 1277-1283.

Peebles CL, Benatan EJ, Jarrell KA, Perlman PS. 1987. Group II intron self-splicing: development of alternative reaction conditions and identification of a predicted intermediate. *Cold Spring Harbor symposia on quantitative biology* **52**: 223-232.

Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL. 1986. A self-splicing RNA excises an intron lariat. *Cell* **44**: 213-223.

Pereira LA, Bentley K, Peeters A, Churchill MJ, Deacon NJ. 2000. A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter. *Nucleic acids research* **28**: 663-668.

Perez EE, Wang J, Miller JC, Jouvenot Y, Kim KA, Liu O, Wang N, Lee G, Bartsevich VV, Lee YL et al. 2008. Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology* **26**: 808-816.

Perez LG, Davis GL, Hunter E. 1987. Mutants of the Rous sarcoma virus envelope glycoprotein that lack the transmembrane anchor and cytoplasmic domains: analysis of intracellular transport and assembly into virions. *Journal of virology* **61**: 2981-2988.

Perutka J, Wang W, Goerlitz D, Lambowitz AM. 2004. Use of computer-designed group II introns to disrupt Escherichia coli DExH/D-box protein and DNA helicase genes. *Journal of molecular biology* **336**: 421-439.

Picard-Maureau M, Kreppel F, Lindemann D, Juretzek T, Herchenroder O, Rethwilm A, Kochanek S, Heinkelein M. 2004. Foamy virus--adenovirus hybrid vectors. *Gene therapy* **11**: 722-728.

Piccirilli JA. 2008. Biochemistry. Toward understanding self-splicing. *Science* **320**: 56-57.

Pichlmair A, Diebold SS, Gschmeissner S, Takeuchi Y, Ikeda Y, Collins MK, Reis e Sousa C. 2007. Tubulovesicular structures within vesicular stomatitis virus G protein-pseudotyped lentiviral vector preparations carry DNA and stimulate antiviral responses via Toll-like receptor 9. *Journal of virology* **81**: 539-547.

Pike-Overzet K, van der Burg M, Wagemaker G, van Dongen JJ, Staal FJ. 2007. New insights and unresolved issues regarding insertional mutagenesis in X-linked SCID gene therapy. *Molecular therapy : the journal of the American Society of Gene Therapy* **15**: 1910-1916.

Plasterk RH. 1996. The Tc1/mariner transposon family. *Current topics in microbiology and immunology* **204**: 125-143.

Plasterk RH, Izsvak Z, Ivics Z. 1999. Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends in genetics : TIG* **15**: 326-332.

Poch O, Sauvaget I, Delarue M, Tordo N. 1989. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *The EMBO journal* **8**: 3867-3874.

Podar M, Chu VT, Pyle AM, Perlman PS. 1998a. Group II intron splicing in vivo by first-step hydrolysis. *Nature* **391**: 915-918.

Podar M, Perlman PS, Padgett RA. 1995. Stereochemical selectivity of group II intron splicing, reverse splicing, and hydrolysis reactions. *Molecular and cellular biology* **15**: 4466-4478.

Podar M, Zhuo J, Zhang M, Franzen JS, Perlman PS, Peebles CL. 1998b. Domain 5 binds near a highly conserved dinucleotide in the joiner linking domains 2 and 3 of a group II intron. *RNA* **4**: 151-166.

Porteus MH. 2006. Mammalian gene targeting with designed zinc finger nucleases. *Molecular therapy : the journal of the American Society of Gene Therapy* **13**: 438-446.

Porteus MH, Baltimore D. 2003. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**: 763.

Porteus MH, Cathomen T, Weitzman MD, Baltimore D. 2003. Efficient gene targeting mediated by adeno-associated virus and DNA double-strand breaks. *Molecular and cellular biology* **23**: 3558-3565.

Pullen KA, Ishimoto LK, Champoux JJ. 1992. Incomplete removal of the RNA primer for minus-strand DNA synthesis by human immunodeficiency virus type 1 reverse transcriptase. *Journal of virology* **66**: 367-373.

Purcell DF, Martin MA. 1993. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *Journal of virology* **67**: 6365-6378.

Pyle AM. 2010. The tertiary structure of group II introns: implications for biological function and evolution. *Critical reviews in biochemistry and molecular biology* **45**: 215-232.

Pyle AM, Fedorova O, Waldsich C. 2007. Folding of group II introns: a model system for large, multidomain RNAs? *Trends in biochemical sciences* **32**: 138-145.

Qin PZ, Pyle AM. 1998. The architectural organization and mechanistic function of group II intron structural elements. *Current opinion in structural biology* **8**: 301-308.

Quenneville SP, Chapdelaine P, Rousseau J, Tremblay JP. 2007. Dystrophin expression in host muscle following transplantation of muscle precursor cells modified with the phiC31 integrase. *Gene therapy* **14**: 514-522.

Rabbitts TH. 1998. LMO T-cell translocation oncogenes typify genes activated by chromosomal translocations that alter transcription and developmental processes. *Genes & development* **12**: 2651-2657.

Ramezani A, Hawley TS, Hawley RG. 2000. Lentiviral vectors for enhanced gene expression in human hematopoietic cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **2**: 458-469.

Raper SE, Chirmule N, Lee FS, Wivel NA, Bagg A, Gao GP, Wilson JM, Batshaw ML. 2003. Fatal systemic inflammatory response syndrome in a ornithine transcarbamylase deficient patient following adenoviral gene transfer. *Molecular genetics and metabolism* **80**: 148-158.

Raposo G, Moore M, Innes D, Leijendekker R, Leigh-Brown A, Benaroch P, Geuze H. 2002. Human macrophages accumulate HIV-1 particles in MHC II compartments. *Traffic* **3**: 718-729.

Ratnasabapathy R, Sheldon M, Johal L, Hernandez N. 1990. The HIV-1 long terminal repeat contains an unusual element that induces the synthesis of short RNAs from various mRNA and snRNA promoters. *Genes & development* **4**: 2061-2074.

Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, Doran ER, Rafalski JA, Whitehorn EA, Baumeister K et al. 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature* **313**: 277-284.

Rausch H, Lehmann M. 1991. Structural analysis of the actinophage phi C31 attachment site. *Nucleic acids research* **19**: 5187-5189.

Rausch JW, Le Grice SF. 2004. 'Binding, bending and bonding': polypurine tract-primed initiation of plus-strand DNA synthesis in human immunodeficiency virus. *The international journal of biochemistry & cell biology* **36**: 1752-1766.

Re F, Braaten D, Franke EK, Luban J. 1995. Human immunodeficiency virus type 1 Vpr arrests the cell cycle in G2 by inhibiting the activation of p34cdc2-cyclin B. *Journal of virology* **69**: 6859-6864.

Recchia A, Parks RJ, Lamartina S, Toniatti C, Pieroni L, Palombo F, Ciliberto G, Graham FL, Cortese R, La Monica N et al. 1999. Site-specific integration mediated by a hybrid adenovirus/adeno-associated virus vector. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 2615-2620.

Recchia A, Perani L, Sartori D, Olgiati C, Mavilio F. 2004. Site-specific integration of functional transgenes into the human genome by adeno/AAV hybrid vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **10**: 660-670.

Reiser J. 2000. Production and concentration of pseudotyped HIV-1-based gene transfer vectors. *Gene therapy* **7**: 910-913.

Relander T, Johansson M, Olsson K, Ikeda Y, Takeuchi Y, Collins M, Richter J. 2005. Gene transfer to repopulating human CD34+ cells using amphotropic-, GALV-, or RD114-pseudotyped HIV-1-based vectors from stable producer cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **11**: 452-459.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG* **16**: 276-277.

Richard E, Douillard-Guilloux G, Batista L, Caillaud C. 2008. Correction of glycogenosis type 2 by muscle-specific lentiviral vector. *In vitro cellular & developmental biology Animal* **44**: 397-406.

Robart AR, Seo W, Zimmerly S. 2007. Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 6620-6625.

Robart AR, Zimmerly S. 2005. Group II intron retroelements: function and diversity. *Cytogenetic and genome research* **110**: 589-597.

Rodriguez SA, Yu JJ, Davis G, Arulanandam BP, Klose KE. 2008. Targeted inactivation of francisella tularensis genes by group II introns. *Applied and environmental microbiology* **74**: 2619-2626.

Roe T, Reynolds TC, Yu G, Brown PO. 1993. Integration of murine leukemia virus DNA depends on mitosis. *The EMBO journal* **12**: 2099-2108.

Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. *Biology direct* **7**: 11.

Roitzsch M, Pyle AM. 2009. The linear form of a group II intron catalyzes efficient autocatalytic reverse splicing, establishing a potential for mobility. *RNA* **15**: 473-482.

Romer P, Strauss T, Hahn S, Scholze H, Morbitzer R, Grau J, Bonas U, Lahaye T. 2009. Recognition of AvrBs3-like proteins is mediated by specific binding to promoters of matching pepper Bs3 alleles. *Plant physiology* **150**: 1697-1712.

Rosen LE, Morrison HA, Masri S, Brown MJ, Springstubb B, Sussman D, Stoddard BL, Seligman LM. 2006. Homing endonuclease I-CreI derivatives with novel DNA target specificities. *Nucleic acids research* **34**: 4791-4800.

Rothkamm K, Kruger I, Thompson LH, Lobrich M. 2003. Pathways of DNA double-strand break repair during the mammalian cell cycle. *Molecular and cellular biology* **23**: 5706-5715.

Rouet P, Smih F, Jasin M. 1994. Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 6064-6068.

Rousseau J, Chapdelaine P, Boisvert S, Almeida LP, Corbeil J, Montpetit A, Tremblay JP. 2011. Endonucleases: tools to correct the dystrophin gene. *The journal of gene medicine* **13**: 522-537.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols* **5**: 725-738.

Saldanha R, Chen B, Wank H, Matsuura M, Edwards J, Lambowitz AM. 1999. RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. *Biochemistry* **38**: 9069-9083.

Sambrook J, Russell DW. 2001. *Molecular cloning: A Laboratory Manual*.

San Filippo J, Lambowitz AM. 2002. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *Journal of molecular biology* **324**: 933-951.

Sanjana NE, Cong L, Zhou Y, Cunniff MM, Feng G, Zhang F. 2012. A transcription activator-like effector toolbox for genome engineering. *Nature protocols* **7**: 171-192.

Sayeed S, Uzal FA, Fisher DJ, Saputo J, Vidal JE, Chen Y, Gupta P, Rood JI, McClane BA. 2008. Beta toxin is essential for the intestinal virulence of Clostridium perfringens type C disease isolate CN3685 in a rabbit ileal loop model. *Molecular microbiology* **67**: 15-30.

Scalley-Kim M, McConnell-Smith A, Stoddard BL. 2007. Coevolution of a homing endonuclease and its host target sequence. *Journal of molecular biology* **372**: 1305-1319.

Scaramuzza S, Biasco L, Ripamonti A, Castiello MC, Loperfido M, Draghici E, Hernandez RJ, Benedicenti F, Radrizzani M, Salomoni M et al. 2012. Preclinical Safety and Efficacy of Human CD34(+) Cells Transduced With Lentiviral Vector for the Treatment of Wiskott-Aldrich Syndrome. *Molecular therapy : the journal of the American Society of Gene Therapy*.

Schaefer-Ridder M, Wang Y, Hofschneider PH. 1982. Liposomes as gene carriers: efficient transformation of mouse L cells by thymidine kinase gene. *Science* **215**: 166-168.

Schambach A, Galla M, Maetzig T, Loew R, Baum C. 2007. Improving transcriptional termination of self-inactivating gamma-retroviral and lentiviral vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **15**: 1167-1173.

Schiedner G, Morral N, Parks RJ, Wu Y, Koopmans SC, Langston C, Graham FL, Beaudet AL, Kochanek S. 1998. Genomic DNA transfer with a high-capacity adenovirus vector results in improved in vivo gene expression and decreased toxicity. *Nature genetics* **18**: 180-183.

Schmidt M, Hacein-Bey-Abina S, Wissler M, Carlier F, Lim A, Prinz C, Glimm H, Andre-Schmutz I, Hue C, Garrigue A et al. 2005. Clonal evidence for the transduction of CD34+ cells with lymphomyeloid differentiation potential and self-renewal capacity in the SCID-X1 gene therapy trial. *Blood* **105**: 2699-2706.

Schuesler T, Reeves L, Kalle C, Grassman E. 2009. Copy number determination of genetically-modified hematopoietic stem cells. *Methods Mol Biol* **506**: 281-298.

Schwarzwaelder K, Howe SJ, Schmidt M, Brugman MH, Deichmann A, Glimm H, Schmidt S, Prinz C, Wissler M, King DJ et al. 2007. Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution in vivo. *The Journal of clinical investigation* **117**: 2241-2249.

Seetharaman M, Eldho NV, Padgett RA, Dayie KT. 2006. Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA. *RNA* **12**: 235-247.

Segal DJ, Dreier B, Beerli RR, Barbas CF, 3rd. 1999. Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 2758-2763.

Selvaggi TA, Walker RE, Fleisher TA. 1997. Development of antibodies to fetal calf serum with arthus-like reactions in human immunodeficiency virus-infected patients given syngeneic lymphocyte infusions. *Blood* **89**: 776-779.

Seraphin B, Simon M, Boulet A, Faye G. 1989. Mitochondrial splicing requires a protein from a novel helicase family. *Nature* **337**: 84-87.

Sharp PA. 1985. On the origin of RNA splicing and introns. *Cell* **42**: 397-400.

-. 1991. "Five easy pieces". *Science* **254**: 663.

Shayakhmetov DM, Papayannopoulou T, Stamatoyannopoulos G, Lieber A. 2000. Efficient gene transfer into human CD34(+) cells by a retargeted adenovirus vector. *Journal of virology* **74**: 2567-2583.

Shearman C, Godon JJ, Gasson M. 1996. Splicing of a group II intron in a functional transfer gene of Lactococcus lactis. *Molecular microbiology* **21**: 45-53.

Shimotohno K, Temin HM. 1981. Formation of infectious progeny virus after insertion of herpes simplex thymidine kinase gene into DNA of an avian retrovirus. *Cell* **26**: 67-77.

Shioda T, Shibuta H. 1990. Production of human immunodeficiency virus (HIV)-like particles from cells infected with recombinant vaccinia viruses carrying the gag gene of HIV. *Virology* **175**: 139-148.

Shirano Y, Shibata D. 1990. Low temperature cultivation of Escherichia coli carrying a rice lipoxygenase L-2 cDNA produces a soluble and active enzyme at a high level. *FEBS letters* **271**: 128-130.

Shub DA, Goodrich-Blair H, Eddy SR. 1994. Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *Trends in biochemical sciences* **19**: 402-404.

Shukla GC, Padgett RA. 2002. A catalytically active group II intron domain 5 can function in the U12-dependent spliceosome. *Molecular cell* **9**: 1145-1150.

Sigel RK, Sashital DG, Abramovitz DL, Palmer AG, Butcher SE, Pyle AM. 2004. Solution structure of domain 5 of a group II intron ribozyme reveals a new RNA motif. *Nature structural & molecular biology* **11**: 187-192.

Sigel RK, Vaidya A, Pyle AM. 2000. Metal ion binding sites in a group II intron core. *Nature structural biology* **7**: 1111-1116.

Sikorski RS, Hieter P. 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics* **122**: 19-27.

Silva GH, Belfort M, Wende W, Pingoud A. 2006. From monomeric to homodimeric endonucleases and back: engineering novel specificity of LAGLIDADG enzymes. *Journal of molecular biology* **361**: 744-754.

Silvers RM, Smith JA, Schowalter M, Litwin S, Liang Z, Geary K, Daniel R. 2010. Modification of integration site preferences of an HIV-1-based vector by expression of a novel synthetic protein. *Human gene therapy* **21**: 337-349.

Simon DM, Clarke NA, McNeil BA, Johnson I, Pantuso D, Dai L, Chai D, Zimmerly S. 2008. Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA* **14**: 1704-1713.

Simon DM, Kelchner SA, Zimmerly S. 2009. A broadscale phylogenetic analysis of group II intron RNAs and intron-encoded reverse transcriptases. *Molecular biology and evolution* **26**: 2795-2808.

Simonelli F, Maguire AM, Testa F, Pierce EA, Mingozzi F, Bennicelli JL, Rossi S, Marshall K, Banfi S, Surace EM et al. 2010. Gene therapy for Leber's congenital amaurosis is safe and effective through 1.5 years after vector administration. *Molecular therapy : the journal of the American Society of Gene Therapy* **18**: 643-650.

Singh NN, Lambowitz AM. 2001. Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. *Journal of molecular biology* **309**: 361-386.

Sivalingam J, Krishnan S, Ng WH, Lee SS, Phan TT, Kon OL. 2010. Biosafety assessment of site-directed transgene integration in human umbilical cord-lining cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **18**: 1346-1356.

Smagulova F, Maurel S, Morichaud Z, Devaux C, Mougel M, Houzet L. 2005. The highly structured encapsidation signal of MuLV RNA is involved in the nuclear export of its unspliced RNA. *Journal of molecular biology* **354**: 1118-1128.

Smith DB, Johnson KS. 1988. Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. *Gene* **67**: 31-40.

Smith J, Bibikova M, Whitby FG, Reddy AR, Chandrasegaran S, Carroll D. 2000. Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. *Nucleic acids research* **28**: 3361-3369.

Smith J, Grizot S, Arnould S, Duclert A, Epinat JC, Chames P, Prieto J, Redondo P, Blanco FJ, Bravo J et al. 2006. A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic acids research* **34**: e149.

Smith MC, Thorpe HM. 2002. Diversity in the serine recombinases. *Molecular microbiology* **44**: 299-307.

Smyth DR, Mrozkiewicz MK, McGrath WJ, Listwan P, Kobe B. 2003. Crystal structures of fusion proteins with large-affinity tags. *Protein science : a publication of the Protein Society* **12**: 1313-1322.

Sollu C, Pars K, Cornu TI, Thibodeau-Beganny S, Maeder ML, Joung JK, Heilbronn R, Cathomen T. 2010. Autonomous zinc-finger nuclease pairs for targeted chromosomal deletion. *Nucleic acids research* **38**: 8269-8276.

Sorensen HP, Sperling-Petersen HU, Mortensen KK. 2003. Production of recombinant thermostable proteins expressed in Escherichia coli: completion of protein synthesis is the bottleneck. *Journal of chromatography B, Analytical technologies in the biomedical and life sciences* **786**: 207-214.

Spiess C, Beil A, Ehrmann M. 1999. A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein. *Cell* **97**: 339-347.

Stahley MR, Strobel SA. 2006. RNA splicing: group I intron crystal structures reveal the basis of splice site selection and metal ion catalysis. *Current opinion in structural biology* **16**: 319-326.

Staunstrup NH, Moldt B, Mates L, Villesen P, Jakobsen M, Ivics Z, Izsvak Z, Mikkelsen JG. 2009. Hybrid lentivirus-transposon vectors with a random integration profile in human cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **17**: 1205-1214.

Stein S, Ott MG, Schultze-Strasser S, Jauch A, Burwinkel B, Kinner A, Schmidt M, Kramer A, Schwable J, Glimm H et al. 2010. Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. *Nature medicine* **16**: 198-204.

Steitz TA, Smerdon S, Jager J, Wang J, Kohlstaedt LA, Friedman JM, Beese LS, Rice PA. 1993. Two DNA polymerases: HIV reverse transcriptase and the Klenow fragment of Escherichia coli DNA polymerase I. *Cold Spring Harbor symposia on quantitative biology* **58**: 495-504.

Stephen SL, Sivanandam VG, Kochanek S. 2008. Homologous and heterologous recombination between adenovirus vector DNA and chromosomal DNA. *The journal of gene medicine* **10**: 1176-1189.

Sternberg N, Sauer B, Hoess R, Abremski K. 1986. Bacteriophage P1 cre gene and its regulatory region. Evidence for multiple promoters and for regulation by DNA methylation. *Journal of molecular biology* **187**: 197-212.

Steuer S, Pingoud V, Pingoud A, Wende W. 2004. Chimeras of the homing endonuclease PI-SceI and the homologous Candida tropicalis intein: a study to explore the possibility of exchanging DNA-binding modules to obtain highly specific endonucleases with altered specificity. *Chembiochem : a European journal of chemical biology* **5**: 206-213.

Su LJ, Waldsich C, Pyle AM. 2005. An obligate intermediate along the slow folding pathway of a group II intron ribozyme. *Nucleic acids research* **33**: 6674-6687.

Sullivan KE, Mullen CA, Blaese RM, Winkelstein JA. 1994. A multiinstitutional survey of the Wiskott-Aldrich syndrome. *The Journal of pediatrics* **125**: 876-885.

Sung P, Klein H. 2006. Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nature reviews Molecular cell biology* **7**: 739-750.

Suzuki M, Kasai K, Saeki Y. 2006. Plasmid DNA sequences present in conventional herpes simplex virus amplicon vectors cause rapid transgene silencing by forming inactive chromatin. *Journal of virology* **80**: 3293-3300.

Swierczek M, Izsvak Z, Ivics Z. 2012. The Sleeping Beauty transposon system for clinical applications. *Expert opinion on biological therapy* **12**: 139-153.

Szczepek M, Brondani V, Buchel J, Serrano L, Segal DJ, Cathomen T. 2007. Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nature biotechnology* **25**: 786-793.

Tabin CJ, Hoffmann JW, Goff SP, Weinberg RA. 1982. Adaptation of a retrovirus as a eucaryotic vector transmitting the herpes simplex virus thymidine kinase gene. *Molecular and cellular biology* **2**: 426-436.

Takata M, Sasaki MS, Sonoda E, Morrison C, Hashimoto M, Utsumi H, Yamaguchi-Iwai Y, Shinohara A, Takeda S. 1998. Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *The EMBO journal* **17**: 5497-5508.

Tan W, Dong Z, Wilkinson TA, Barbas CF, 3rd, Chow SA. 2006. Human immunodeficiency virus type 1 incorporated with fusion proteins consisting of integrase and the designed polydactyl zinc finger protein E2C can bias integration of viral DNA into a predetermined chromosomal region in human cells. *Journal of virology* **80**: 1939-1948.

Tan W, Zhu K, Segal DJ, Barbas CF, 3rd, Chow SA. 2004. Fusion proteins consisting of human immunodeficiency virus type 1 integrase and the designed polydactyl zinc finger protein E2C direct integration of viral DNA into specific sites. *Journal of virology* **78**: 1301-1313.

Tatum EL. 1964. The Determinants and Evolution of Life. Genetic Determinants. *Proceedings of the National Academy of Sciences of the United States of America* **51**: 908-915.

Thomas CE, Schiedner G, Kochanek S, Castro MG, Lowenstein PR. 2000. Peripheral infection with adenovirus causes unexpected long-term brain inflammation in animals injected intracranially with first-generation, but not with high-capacity, adenovirus vectors: toward realistic long-term neurological gene therapy for chronic diseases. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 7482-7487.

Thornhill SI, Schambach A, Howe SJ, Ulaganathan M, Grassman E, Williams D, Schiedlmeier B, Sebire NJ, Gaspar HB, Kinnon C et al. 2008. Self-inactivating gammaretroviral vectors for gene therapy of X-linked severe combined immunodeficiency. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**: 590-598.

Thorpe HM, Smith MC. 1998. In vitro site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 5505-5510.

Titeux M, Pendaries V, Zanta-Boussif MA, Decha A, Pironon N, Tonasso L, Mejia JE, Brice A, Danos O, Hovnanian A. 2010. SIN retroviral vectors expressing COL7A1 under human promoters for ex vivo gene therapy of recessive dystrophic epidermolysis bullosa. *Molecular therapy : the journal of the American Society of Gene Therapy* **18**: 1509-1518.

Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* **7**: 1142-1152.

Toor N, Keating KS, Fedorova O, Rajashankar K, Wang J, Pyle AM. 2010. Tertiary architecture of the Oceanobacillus iheyensis group II intron. *RNA* **16**: 57-69.

Toor N, Keating KS, Pyle AM. 2009. Structural insights into RNA splicing. *Current opinion in structural biology* **19**: 260-266.

Toor N, Keating KS, Taylor SD, Pyle AM. 2008a. Crystal structure of a self-spliced group II intron. *Science* **320**: 77-82.

Toor N, Rajashankar K, Keating KS, Pyle AM. 2008b. Structural basis for exon recognition by a group II intron. *Nature structural & molecular biology* **15**: 1221-1222.

Toor N, Robart AR, Christianson J, Zimmerly S. 2006. Self-splicing of a group IIC intron: 5' exon recognition and alternative 5' splicing events implicate the stem-loop motif of a transcriptional terminator. *Nucleic acids research* **34**: 6461-6471.

Toro N. 2003. Bacteria and Archaea Group II introns: additional mobile genetic elements in the environment. *Environmental microbiology* **5**: 143-151.

Toro N, Jimenez-Zurdo JI, Garcia-Rodriguez FM. 2007. Bacterial group II introns: not just splicing. *FEMS microbiology reviews* **31**: 342-358.

Towers GJ. 2007. The control of viral infection by tripartite motif proteins and cyclophilin A. *Retrovirology* **4**: 40.

Trinh AT, Ball BG, Weber E, Gallaher TK, Gluzman-Poltorak Z, Anderson F, Basile LA. 2009. Retroviral vectors encoding ADA regulatory locus control region provide enhanced T-cell-specific transgene expression. *Genetic vaccines and therapy* **7**: 13.

Trobridge GD. 2011. Genotoxicity of retroviral hematopoietic stem cell gene therapy. *Expert opinion on biological therapy* **11**: 581-593.

Turlure F, Devroe E, Silver PA, Engelman A. 2004. Human cell proteins and human immunodeficiency virus DNA integration. *Frontiers in bioscience : a journal and virtual library* **9**: 3187-3208.

Tuschong L, Soenen SL, Blaese RM, Candotti F, Muul LM. 2002. Immune response to fetal calf serum by two adenosine deaminase-deficient patients after T cell gene therapy. *Human gene therapy* **13**: 1605-1610.

Uren AG, Kool J, Berns A, van Lohuizen M. 2005. Retroviral insertional mutagenesis: past, present and future. *Oncogene* **24**: 7656-7672.

Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC. 2005. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**: 646-651.

Valadkhan S. 2007. The spliceosome: a ribozyme at heart? *Biological chemistry* **388**: 693-697.

Valles Y, Halanych KM, Boore JL. 2008. Group II introns break new boundaries: presence in a bilaterian's genome. *PloS one* **3**: e1488.

Valton J, Daboussi F, Leduc S, Redondo P, Molina R, Macmaster R, Paques F, Montoya G, Duchateau P. 2012. CpG Methylation Impacts the Activity of Natural and Engineered Meganucleases. *The Journal of biological chemistry*.

van den Berg B, Ellis RJ, Dobson CM. 1999. Effects of macromolecular crowding on protein folding and aggregation. *The EMBO journal* **18**: 6927-6933.

van der Loo JC, Swaney WP, Grassman E, Terwilliger A, Higashimoto T, Schambach A, Baum C, Thrasher AJ, Williams DA, Nordling DL et al. 2012a. Scale-up and manufacturing of clinical-grade self-inactivating gamma-retroviral vectors by transient transfection. *Gene therapy* **19**: 246-254.

van der Loo JC, Swaney WP, Grassman E, Terwilliger A, Higashimoto T, Schambach A, Hacein-Bey-Abina S, Nordling DL, Cavazzana-Calvo M, Thrasher AJ et al. 2012b. Critical Variables affecting clinical-grade production of the self-inactivating gamma-retroviral vector for the treatment of X-linked severe combined immunodeficiency. *Gene therapy* **19**: 872-876.

van der Veen R, Arnberg AC, van der Horst G, Bonen L, Tabak HF, Grivell LA. 1986. Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro. *Cell* **44**: 225-234.

Van Doren K, Hanahan D, Gluzman Y. 1984. Infection of eucaryotic cells by helper-independent recombinant adenoviruses: early region 1 is not obligatory for integration of viral DNA. *Journal of virology* **50**: 606-614.

Van Dyck E, Jank B, Ragnini A, Schweyen RJ, Duyckaerts C, Sluse F, Foury F. 1995. Overexpression of a novel member of the mitochondrial carrier family rescues defects in both DNA and RNA metabolism in yeast mitochondria. *Molecular & general genetics : MGG* **246**: 426-436.

Van Maele B, Busschots K, Vandekerckhove L, Christ F, Debyser Z. 2006. Cellular co-factors of HIV-1 integration. *Trends in biochemical sciences* **31**: 98-105.

Van Maele B, De Rijck J, De Clercq E, Debyser Z. 2003. Impact of the central polypurine tract on the kinetics of human immunodeficiency virus type 1 vector transduction. *Journal of virology* **77**: 4685-4694.

Vargas J, Jr., Gusella GL, Najfeld V, Klotman ME, Cara A. 2004. Novel integrase-defective lentiviral episomal vectors for gene transfer. *Human gene therapy* **15**: 361-372.

Varmus HE, Heasley S, Kung HJ, Oppermann H, Smith VC, Bishop JM, Shank PR. 1978. Kinetics of synthesis, structure and purification of avian sarcoma virus-specific DNA made in the cytoplasm of acutely infected cells. *Journal of molecular biology* **120**: 55-82.

Vellore J, Moretz SE, Lampson BC. 2004. A group II intron-type open reading frame from the thermophile Bacillus (Geobacillus) stearothermophilus encodes a heat-stable reverse transcriptase. *Applied and environmental microbiology* **70**: 7140-7147.

Vigdal TJ, Kaufman CD, Izsvak Z, Voytas DF, Ivics Z. 2002. Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. *Journal of molecular biology* **323**: 441-452.

Villa T, Pleiss JA, Guthrie C. 2002. Spliceosomal snRNAs: Mg(2+)-dependent chemistry at the catalytic core? *Cell* **109**: 149-152.

Vink CA, Gaspar HB, Gabriel R, Schmidt M, McIvor RS, Thrasher AJ, Qasim W. 2009. Sleeping beauty transposition from nonintegrating lentivirus. *Molecular therapy : the journal of the American Society of Gene Therapy* **17**: 1197-1204.

Vogel J, Borner T. 2002. Lariat formation and a hydrolytic pathway in plant chloroplast group II intron splicing. *The EMBO journal* **21**: 3794-3803.

Vogel J, Borner T, Hess WR. 1999. Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic acids research* **27**: 3866-3874.

Wagner R, Graf M, Bieler K, Wolf H, Grunwald T, Foley P, Uberla K. 2000. Rev-independent expression of synthetic gag-pol genes of human immunodeficiency virus type 1 and simian immunodeficiency virus: implications for the safety of lentiviral vectors. *Human gene therapy* **11**: 2403-2413.

Wakefield JK, Morrow CD. 1996. Mutations within the primer binding site of the human immunodeficiency virus type 1 define sequence requirements essential for reverse transcription. *Virology* **220**: 290-298.

Walisko O, Schorn A, Rolfs F, Devaraj A, Miskey C, Izsvak Z, Ivics Z. 2008. Transcriptional activities of the Sleeping Beauty transposon and shielding its genetic cargo with insulators. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**: 359-369.

Wang CL, Wang BB, Bartha G, Li L, Channa N, Klinger M, Killeen N, Wabl M. 2006. Activation of an oncogenic microRNA cistron by provirus integration. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 18680-18684.

Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD. 2007. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome research* **17**: 1186-1194.

Wang H, Lieber A. 2006. A helper-dependent capsid-modified adenovirus vector expressing adeno-associated virus rep78 mediates site-specific integration of a 27-kilobase transgene cassette. *Journal of virology* **80**: 11699-11709.

Wang Y, Camp SM, Niwano M, Shen X, Bakowska JC, Breakefield XO, Allen PD. 2002. Herpes simplex virus type 1/adeno-associated virus rep(+) hybrid amplicon vector improves the stability of transgene expression in human cells by site-specific integration. *Journal of virology* **76**: 7150-7162.

Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM. 1999. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Molecular cell* **4**: 239-250.

Watanabe K, Lambowitz AM. 2004. High-affinity binding site for a group II intron-encoded reverse transcriptase/maturase within a stem-loop structure in the intron RNA. *RNA* **10**: 1433-1443.

Watts JK, Corey DR. 2012. Silencing disease genes in the laboratory and the clinic. *The Journal of pathology* **226**: 365-379.

Wei CM, Gibson M, Spear PG, Scolnick EM. 1981. Construction and isolation of a transmissible retrovirus containing the src gene of Harvey murine sarcoma virus and the thymidine kinase gene of herpes simplex virus type 1. *Journal of virology* **39**: 935-944.

Weinstock DM, Richardson CA, Elliott B, Jasin M. 2006. Modeling oncogenic translocations: distinct roles for double-strand break repair pathways in translocation formation in mammalian cells. *DNA repair* **5**: 1065-1074.

Weiss RA. 2006. The discovery of endogenous retroviruses. *Retrovirology* **3**: 67.

Weitzman MD, Kyostio SR, Kotin RM, Owens RA. 1994. Adeno-associated virus (AAV) Rep proteins mediate complex formation between AAV DNA and its integration site in human DNA. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 5808-5812.

Wiesenberger G, Waldherr M, Schweyen RJ. 1992. The nuclear gene MRS2 is essential for the excision of group II introns from yeast mitochondrial transcripts in vivo. *The Journal of biological chemistry* **267**: 6963-6969.

Wilber A, Frandsen JL, Geurts JL, Largaespada DA, Hackett PB, McIvor RS. 2006. RNA as a source of transposase for Sleeping Beauty-mediated gene insertion and expression in somatic cells and tissues. *Molecular therapy : the journal of the American Society of Gene Therapy* **13**: 625-630.

Wilk T, Gowen B, Fuller SD. 1999. Actin associates with the nucleocapsid domain of the human immunodeficiency virus Gag polyprotein. *Journal of virology* **73**: 1931-1940.

Wilson MH, Coates CJ, George AL, Jr. 2007. PiggyBac transposon-mediated gene transfer in human cells. *Molecular therapy : the journal of the American Society of Gene Therapy* **15**: 139-145.

Wodrich H, Bohne J, Gumz E, Welker R, Krausslich HG. 2001. A new RNA element located in the coding region of a murine endogenous retrovirus can functionally replace the Rev/Rev-responsive element system in human immunodeficiency virus type 1 Gag expression. *Journal of virology* **75**: 10670-10682.

Woltjen K, Michael IP, Mohseni P, Desai R, Mileikovsky M, Hamalainen R, Cowling R, Wang W, Liu P, Gertsenstein M et al. 2009. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **458**: 766-770.

Woodson SA. 2005. Structure and assembly of group I introns. *Current opinion in structural biology* **15**: 324-330.

Wu SC, Meir YJ, Coates CJ, Handler AM, Pelczar P, Moisyadi S, Kaminski JM. 2006. piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 15008-15013.

Wu X, Li Y, Crise B, Burgess SM, Munroe DJ. 2005. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *Journal of virology* **79**: 5211-5214.

Xiong Y, Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO journal* **9**: 3353-3362.

Xue X, Huang X, Nodland SE, Mates L, Ma L, Izsvak Z, Ivics Z, LeBien TW, McIvor RS, Wagner JE et al. 2009. Stable gene transfer and expression in cord blood-derived CD34+ hematopoietic stem and progenitor cells by a hyperactive Sleeping Beauty transposon system. *Blood* **114**: 1319-1330.

Yamashita M, Emerman M. 2004. Capsid is a dominant determinant of retrovirus infectivity in nondividing cells. *Journal of virology* **78**: 5670-5678.

-. 2005. The cell cycle independence of HIV infections is not determined by known karyophilic viral elements. *PLoS pathogens* **1**: e18.

Yang J, Mohr G, Perlman PS, Lambowitz AM. 1998. Group II intron mobility in yeast mitochondria: target DNA-primed reverse transcription activity of aI1 and reverse splicing into DNA transposition sites in vitro. *Journal of molecular biology* **282**: 505-523.

Yang J, Zimmerly S, Perlman PS, Lambowitz AM. 1996. Efficient integration of an intron RNA into double-stranded DNA by reverse splicing. *Nature* **381**: 332-335.

Yang L, Bailey L, Baltimore D, Wang P. 2006. Targeting lentiviral vectors to specific cell types in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 11479-11484.

Yant SR, Ehrhardt A, Mikkelsen JG, Meuse L, Pham T, Kay MA. 2002. Transposition from a gutless adeno-transposon vector stabilizes transgene expression in vivo. *Nature biotechnology* **20**: 999-1005.

Yant SR, Huang Y, Akache B, Kay MA. 2007. Site-directed transposon integration in human cells. *Nucleic acids research* **35**: e50.

Yant SR, Kay MA. 2003. Nonhomologous-end-joining factors regulate DNA repair fidelity during Sleeping Beauty element transposition in mammalian cells. *Molecular and cellular biology* **23**: 8505-8518.

Yant SR, Wu X, Huang Y, Garrison B, Burgess SM, Kay MA. 2005. High-resolution genome-wide mapping of transposon integration in mammals. *Molecular and cellular biology* **25**: 2085-2094.

Yao J, Lambowitz AM. 2007. Gene targeting in gram-negative bacteria by use of a mobile group II intron ("Targetron") expressed from a broad-host-range vector. *Applied and environmental microbiology* **73**: 2735-2743.

Yao J, Zhong J, Fang Y, Geisinger E, Novick RP, Lambowitz AM. 2006. Use of targetrons to disrupt essential and nonessential genes in Staphylococcus aureus reveals temperature sensitivity of Ll.LtrB group II intron splicing. *RNA* **12**: 1271-1281.

Yelin R, Rotem D, Schuldiner S. 1999. EmrE, a small Escherichia coli multidrug transporter, protects Saccharomyces cerevisiae from toxins by sequestration in the vacuole. *Journal of bacteriology* **181**: 949-956.

Yoder KE, Bushman FD. 2000. Repair of gaps in retroviral DNA integration intermediates. *Journal of virology* **74**: 11191-11200.

Yuan J, Wang J, Crain K, Fearns C, Kim KA, Hua KL, Gregory PD, Holmes MC, Torbett BE. 2012. Zinc-finger nuclease editing of human cxcr4 promotes HIV-1 CD4(+) T cell resistance and enrichment. *Molecular therapy : the journal of the American Society of Gene Therapy* **20**: 849-859.

Yun H, Lee JW, Jeong J, Chung J, Park JM, Myoung HN, Lee SY. 2007. EcoProDB: the Escherichia coli protein database. *Bioinformatics* **23**: 2501-2503.

Yusa K, Rad R, Takeda J, Bradley A. 2009. Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nature methods* **6**: 363-369.

Yusa K, Zhou L, Li MA, Bradley A, Craig NL. 2011. A hyperactive piggyBac transposase for mammalian applications. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 1531-1536.

Zaiss AK, Son S, Chang LJ. 2002. RNA 3' readthrough of oncoretrovirus and lentivirus: implications for vector safety and efficacy. *Journal of virology* **76**: 7209-7219.

Zanta-Boussif MA, Charrier S, Brice-Ouzet A, Martin S, Opolon P, Thrasher AJ, Hope TJ, Galy A. 2009. Validation of a mutated PRE sequence allowing high and sustained transgene expression while abrogating WHV-X protein synthesis: application to the gene therapy of WAS. *Gene therapy* **16**: 605-619.

Zavada J. 1982. The pseudotypic paradox. *The Journal of general virology* **63 (Pt 1)**: 15-24.

Zayed H, Izsvak Z, Walisko O, Ivics Z. 2004. Development of hyperactive sleeping beauty transposon vectors by mutational analysis. *Molecular therapy : the journal of the American Society of Gene Therapy* **9**: 292-304.

Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. 2011. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **29**: 149-153.

Zhang L, Doudna JA. 2002. Structural insights into group II intron catalysis and branch-site selection. *Science* **295**: 2084-2088.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* **9**: 40.

Zheng C, Baum BJ, Iadarola MJ, O'Connell BC. 2000. Genomic integration and gene expression by a modified adenoviral vector. *Nature biotechnology* **18**: 176-180.

Zhong J, Karberg M, Lambowitz AM. 2003. Targeted and random bacterial gene disruption using a group II intron (targetron) vector containing a retrotransposition-activated selectable marker. *Nucleic acids research* **31**: 1656-1664.

Zhong J, Lambowitz AM. 2003. Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *The EMBO journal* **22**: 4555-4565.

Zhou Q, Sharp PA. 1995. Novel mechanism and factor for regulation by HIV-1 Tat. *The EMBO journal* **14**: 321-328.

Zhou W, Parent LJ, Wills JW, Resh MD. 1994. Identification of a membrane-binding domain within the amino-terminal region of human immunodeficiency virus type 1 Gag protein which interacts with acidic phospholipids. *Journal of virology* **68**: 2556-2569.

Zhuang F, Karberg M, Perutka J, Lambowitz AM. 2009a. EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. *RNA* **15**: 432-449.

Zhuang F, Mastroianni M, White TB, Lambowitz AM. 2009b. Linear group II intron RNAs can retrohome in eukaryotes and may use nonhomologous end-joining for cDNA ligation. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 18189-18194.

Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM. 1995a. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell* **83**: 529-538.

Zimmerly S, Guo H, Perlman PS, Lambowitz AM. 1995b. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**: 545-554.

Zimmerly S, Hausner G, Wu X. 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic acids research* **29**: 1238-1250.

Zimmerly S, Moran JV, Perlman PS, Lambowitz AM. 1999. Group II intron reverse transcriptase in yeast mitochondria. Stabilization and regulation of reverse transcriptase activity by the intron RNA. *Journal of molecular biology* **289**: 473-490.

Zoschke R, Nakamura M, Liere K, Sugiura M, Borner T, Schmitz-Linneweber C. 2010. An organellar maturase associates with multiple group II introns. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 3245-3250.

Zou J, Maeder ML, Mali P, Pruett-Miller SM, Thibodeau-Beganny S, Chou BK, Chen G, Ye Z, Park IH, Daley GQ et al. 2009. Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell stem cell* **5**: 97-110.

Zufferey R, Donello JE, Trono D, Hope TJ. 1999. Woodchuck hepatitis virus posttranscriptional regulatory element enhances expression of transgenes delivered by retroviral vectors. *Journal of virology* **73**: 2886-2892.

Zufferey R, Dull T, Mandel RJ, Bukovsky A, Quiroz D, Naldini L, Trono D. 1998. Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery. *Journal of virology* **72**: 9873-9880.

Zychlinski D, Schambach A, Modlich U, Maetzig T, Meyer J, Grassman E, Mishra A, Baum C. 2008. Physiological promoters reduce the genotoxic risk of integrating gene vectors. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**: 718-725.

ABSTRACT

Integrating vectors are widely used in gene therapy for stable and long-term transgene expression. In *ex vivo* hematopoietic gene therapy approaches, HIV-1-derived lentiviral vectors can thus be used to transduce hematopoietic progenitors. The biological potency of the vector is expected to correlate positively with the frequency of transduced cells and also with the number of integration (VCN, vector copy number) per cell. However, the use of these vectors that cannot target transgene integration into host chromosome may lead to insertional mutagenesis. In this regard, the safety of integrating vectors remains a significant concern in clinical applications. We first evaluated the level of transduction of hematopoietic progenitor cells at the single-cell level by measuring VCN in individual colony-forming cell units using an adapted quantitative PCR method. We showed that the frequency of transduced progenitor cells and the distribution of VCN in hematopoietic colonies may depend upon experimental conditions including features of vectors.

On the other hand, the use of vectors that can target the integration of the transgene into a specific-site of the host genome would overcome genotoxicity issues related to integrating vectors. While site-specific integrative approaches based on engineered nucleases such as Zinc-finger nucleases or Meganucleases are currently developed, we evaluated the use of a group II intron for genomic targeting. Group II introns are self-splicing mobile elements found in prokaryotes and eukaryotic organelles. They can integrate into precise genomic locations by homing, following assembly of a ribonucleoprotein complex containing the intron-encoded protein (IEP) and the spliced intron RNA. Engineered group II introns are commonly used tools for targeted genomic modifications in prokaryotes but not in eukaryotes, probably due limited catalytic activation of currently known group II introns in eukaryotic cells. The brown algae *Pylaiella littoralis* Pl.LSU/2 group II intron is uniquely capable of *in vitro* ribozyme activity at unusually low level of magnesium. As this intron remains poorly characterized, we purified recombinant Pl.LSU/2 IEP expressed in *Escherichia coli* and showed that the protein displays a reverse transcriptase activity either alone or associated with intronic RNA. The Pl.LSU/2 intron could be engineered to splice accurately in *Saccharomyces cerevisiae* and splicing efficiency was increased by the maturase activity of the IEP. However, spliced transcripts were not expressed. Although intron splicing was not detected in human cells, and homing of Pl.LSU/2 in *E. coli* and *S. cerevisiae* could not be demonstrated, these data provide the first functional characterization of the PI.LSU/2 IEP and the first evidence that the Pl.LSU/2 group II intron splicing occurs *in vivo* in eukaryotes in an IEP-dependent manner.

KEYWORDS

Gene therapy; integrating vector; group II intron; Intron-encoded protein; splicing; homing; *Pylaiella littoralis*