Université d'Évry-Val d'Essonne
Ecole Doctorale GAO :
Des Génomes Aux Organismes

THÈSE

Présentée pour obtenir le grade de Docteur en sciences
De l'université d'Évry-Val d'Essonne
SPÉCIALITÉ : Bioinformatique et analyse des génomes

par
Mathieu CHARLES
le 4 janvier 2010

Évolution des génomes du blé (genres *Aegilops* et *Triticum*) au sein des *Poaceae*

Dynamique rapide de l'espace occupé par les éléments transposables et conservation relative des gènes

JURY

Dr. Richard Cooke	**Rapporteur**
Pr. Alan Schulman	**Rapporteur**
Pr. Francis Quétier	**Président du Jury**
Dr. Joseph Jahier	**Examinateur**
Dr. Carène Rizzon	**Examinateur**
Dr. Patrick Wincker	**Examinateur**
Dr. Boulos Chalhoub	**Directeur de thèse**

Remerciements

La thèse, quelle aventure à la fois unique, passionnante et éprouvante. Pas moins de cinq années se sont écoulées depuis mes premiers pas à l'URGV et la fin de la rédaction de ce document. Je ne pourrais dire si ces années se sont écoulées vite ou lentement, mais une chose est sûre, elles auront été remplies de moments et de rencontres inoubliables..

A commencer par celle avec Boulos qui a passé ces cinq dernières années à m'encadrer, avec la bonne dose de patience et de courage. Tu m'as fait confiance pour défendre un sujet de thèse mûri depuis des années qui te tenait à cœur. Pour cette confiance, ta bonne humeur et ton encadrement, je te remercie très sincèrement.

Je tiens à remercier mes rapporteurs Alan Schulman et Richard Cooke ainsi que les autres membres du jury d'avoir accepté de se plonger quelque temps dans mon sujet de recherche, et particulièrement Francis Quétier sans qui cette thèse n'aurait pas été possible.

Je remercie également Michel Caboche et Heribert Hirt de m'avoir accueilli dans l'Unité de Recherche en Génomique Végétale me permettant d'intégrer cette inoubliable équipe OEPG guidée par Boulos.

Pendant toutes ces années, bien du monde est passé dans cette équipe, de façon éphémère ou plus durable. Tous y ont laissé leur empreinte, et il n'est pas difficile de se rappeler toutes les personnes qui m'ont accompagnées.

Jérémy 'JJ', comment ne pas commencer par toi après tout ce que tu as donné à cette équipe pendant ces années. Ton récent départ à Rennes laisse un vide qui ne se comblera probablement jamais vraiment. Tu es probablement l'une des personnes les plus intéressantes que j'ai eu la chance de rencontrer, ne change rien.

Ma petite Imen, 'elle …' comme dirait Ben. Ta gentillesse et ta bonne humeur ont fait de toi une collègue et amie précieuse pendant toutes ces années. Ton aventure à toi s'achève bientôt et n'hésite pas à solliciter Erwan 'le bûcheron fantastique' que nous avons eu la chance de rencontrer et que tu as eu le bonheur de garder au plus proche de toi.

Gwen ‘pitchoune’, comme je regrette le temps où le couloir résonnait de tes pas discrets annonçant ton arrivée… Bon ok, on continue de se voir toutes les semaines avec grand plaisir alors ce n’est pas si grave. Grâce à toi, j’ai pu aussi rencontrer le Ben, et on peut dire que ça valait le coup. J’ai une pensée bien particulière pour vous deux qui n’avez pas été épargnés par les épreuves cette année et je vous transmets toutes mes amitiés en espérant que l’année à venir sera bien meilleure.

Charlotte ‘Blondy-Poultry…’, tu n’es passée que rapidement à l’URGV, mais cela nous a suffit pour créer une relation d’amitié. Merci pour tous ces moments de détente passer ensemble avec Fabrice, si précieux pendant des périodes de stress.

Je pense aussi à un dernier couple ayant concrétisé leur engagement : Clémence ‘Marie-Do’ et Yannick. Cette fin d’année nous a un peu éloigné mais ce n’est que pour mieux se retrouver prochainement.

Je remercie aussi nos gentils permanents Nathalie, Harry et Cécile pour tous les bons moments passés ensemble et en compagnie de vos adorables familles (je n’oublie pas Marie-Rose, Camille et Marine non plus). Bon, c’est vrai qu’on n’a pas trop vu ta famille, Cécilou, mais je suis sûr qu’elle est adorable.

Je n’oublie pas non plus ma fidèle comparse de la première heure, Aurélie ‘Ozrélie’, exilée en Australie. Merci pour ton soutien et pour les rudiments de biologie humide que tu as essayé de m’inculquer. Tu te lances à ton tour dans une thèse et je ne peux que te souhaiter bonne chance et bon courage. Je suis très content qu’on ait réussi à garder le contact, ton amitié m’est très chère même si la distance n’aide pas toujours à l’entretenir.

Magalie ‘Coco’, Julie ‘Juliagotchi la …’, Marco ‘Ma qué Catzooo’ et Soazic ‘auteur des soazismes ©’, nous nous sommes malheureusement perdus de vue au fil des ans, mais les moments passés ensemble ne sont pas oubliés

Je me contenterai malheureusement d’un remerciement global à tous les autres collègues et amis que j’ai eu la chance de connaître à l’URGV en espérant n’oublier personne. Alors merci à tous les volleyeurs : Véro ‘Mamie/Quelqu’un ?’, Clément ‘Jeune’, Nelly,

‘l’incroyable’ Jean-Philippe, Laure ‘Tigrou’, Alex, Lulu, les Sandra, les Sandrine, les Juliens, Virginie ‘TATAWA’ mais aussi Aloïs ‘Bob’ mon thésard référé, partenaire de squash, Cléa ma thésarde petite-petite-référée qui a quand même bien choisi son sujet, le vieil homme et l’ordinateur, Fred, Véro ‘super chat’, Fabien, David et Karim et j’en oublie sûrement, qu’ils m’excusent.

Je remercie aussi Toufik et Alex, deux amis précieux que je n’ai pas la chance de voir souvent, même si c’est toujours avec le même plaisir.

Les remerciements touchent à leur fin, et c’est le moment de remercier ma famille, en commençant par mes parents qui m’ont toujours soutenu dans ces études qui n’en finissaient pas en m’apportant leur soutien indéfectible. Je pense aussi à toi Aude, à vous trois vous m’avez offert une vie de famille heureuse, que demandez de plus ? Une famille qui ne cesse d’ailleurs de s’agrandir, de ton coté avec le petit Emile et toute la famille BEZIN mais aussi du mien, avec mes parents d’adoption d’Annecy et mon Karibouchou.

Mes derniers remerciements iront bien évidemment à Odile, ma compagne depuis plus longtemps que la thèse, c’est dire. Je te dédis cette thèse à laquelle tu as largement participé que ce soit en la relisant ou en me supportant pendant tout ce temps.

Sommaire

Avant propos

Le blé est l'une des principales céréales cultivées dans le monde, avec le riz, le maïs, l'orge et le sorgho. Elles fournissent plus de 60% des calories et des apports en protéines de l'alimentation humaine. Une des particularités du blé réside dans la forte teneur en amidon (70%) et en gluten (15%) de ses grains. Le blé est au centre de l'alimentation humaine en tant qu'ingrédient principal pour la fabrication du pain, de la semoule, des biscuits et des pâtes. Sa bonne tolérance au froid est un de ses atouts : elle lui permet d'être cultivé aussi bien en zone tempérée (blé d'hiver, semé à l'automne) que dans des régions au climat plus rigoureux comme le Canada ou la Sibérie (blé de printemps, semé au printemps).

L'espèce majoritairement cultivée (>90% des cultures) est le blé tendre *Triticum aestivum* ssp. *aestivum*, utilisé principalement pour la fabrication du pain. Le blé dur *Triticum turgidum* ssp. *durum* est utilisé pour la fabrication des pâtes alimentaires et des semoules (5% de la production de blé). C'est la différence de dureté du grain (dur ou tendre) qui les destine à ces utilisations différentes. L'origine de la culture du blé se confond avec les débuts de l'agriculture, il y a plus de 10.000 ans dans le croissant fertile. Il a été la première espèce végétale domestiquée et sélectionnée par l'homme.

L'équipe OEPG, qui m'a accueilli, est pionnière dans la caractérisation de l'organisation et l'évolution des génomes du blé. Peu avant mon arrivée, elle a étudié le locus de la dureté de la graine (*Ha*) et sa dynamique dans les génomes du blé. Ces résultats ont constitué le point de départ de ma thèse, qui a finalement englobé un sujet plus vaste au fil de mes travaux.

Dans ce manuscrit de thèse, les résultats sont en partie présentés sous la forme d'articles publiés dans des revues à comité de lecture. Deux des trois parties des résultats reposent ainsi principalement sur ces articles écrits en anglais. Ils sont encadrés par une introduction, des résultats complémentaires et une conclusion (en français). Aussi, je prie le lecteur de bien vouloir excuser cette alternance de français et d'anglais. Chaque partie est ainsi structurée sous la forme : introduction, matériel et méthode, résultats et discussion. Je présente dans un chapitre à part les méthodes d'annotation des séquences génomiques ainsi que le programme que j'ai développé à cet effet.

Les références bibliographiques regroupent l'ensemble des références associées aux articles et aux parties rédigées en français et sont présentées de façon homogène. L'ensemble des tableaux et des figures présentés de ce document est listé dans les Annexes 1 et 2. Trois autres articles auxquels j'ai contribué qui ont été publiés dans des revues à comité de lecture sont joints en Annexes 3, 4 et 5.

Sigles et abréviations

aa : acide aminé

ADN : acide désoxyribonucléique

ADNc : ADN complémentaire d'un ARNm

ADNg : ADN génomique

ARN : acide ribonucléique

ARNm : ARN messager

ARNi : ARN interférant

BAC (Bacterial Artificial Chromosome) : chromosome artificiel bactérien

EST (Expressed Sequenced Tag) : étiquettes de transcrits ou d'ARNm

Gb : 10^9 paires de bases

kb : 10^3 paires de bases

LTR (Long Terminal Repeat) : longues répétitions directes des extrémités

Ma : 10^6 années

Mb : 10^6 paires de bases

Mt : 10^6 tonnes

MHa : 10^6 hectares

ORF (Open Reading Frame) : cadre ouvert de lecture

pb : paires de bases

pg : 10^{-12} grammes

PCR (Polymerisation Chain Reaction) : réaction en chaîne de polymérisation

SNP (Single Nucleotide Polymorphism) : mutation ponctuelle

TE (Transposable Element) : élément transposable

TIR (Tandem Inversed Repeat) : répétitions en tandem inversées

TSD (Target Site Duplication) : duplication du site cible

Étude bibliographique

	Production			Surface		
	Tonnes / an	% céréales	% plantes	Hectares	% céréales	% plantes
MONDE						
Maïs	791 794 584	33,7%	11,1%	158 034 025	22,7%	12,6%
Riz	659 590 623	28,1%	9,2%	155 811 821	22,4%	12,4%
Blé	605 994 942	25,8%	8,5%	214 207 581	30,8%	17,1%
Orge	133 431 341	5,7%	1,9%	55 441 486	8,0%	4,4%
Sorgho	63 375 602	2,7%	0,9%	46 928 032	6,7%	3,7%
Mil	33 949 456	1,4%	0,5%	34 963 783	5,0%	2,8%
Avoine	24 897 095	1,1%	0,3%	11 597 407	1,7%	0,9%
Seigle	14 741 248	0,6%	0,2%	6 307 272	0,9%	0,5%
Triticale	11 973 031	0,5%	0,2%	3 661 240	0,5%	0,3%
Autres	11 648 502	0,5%	0,2%	8 645 920	1,2%	0,7%
Total (céréales)	2 351 396 424	100%	32,9%	695 598 567	100%	55,4%
Total (plantes cultivées)	7 155 007 807	-	100%	1 255 909 717	-	100%
FRANCE						
Blé	32 769 900	55,0%	26,8%	5 238 000	57,6%	38,9%
Maïs	14 528 000	24,4%	11,9%	1 530 700	16,8%	11,4%
Orge	9 475 100	15,9%	7,7%	1 699 100	18,7%	12,6%
Triticale	1 476 000	2,5%	1,2%	324 200	3,6%	2,4%
Autres	1 287 948	2,2%	1,1%	303 845	3,3%	2,3%
Total (céréales)	59 536 948	100%	48,6%	9 095 845	100%	67,5%
Total (plantes cultivées)	122 481 996	-	100%	13 479 097	-	100%

Tableau 1. Production et surface occupées par les principales céréales en 2007, dans le monde et en France en comparaison avec l'ensemble des autres plantes cultivées pour l'alimentation (source FAO : http://faostat.fao.org/ au 20/10/2009).

Figure 1. (A) Évolution de la consommation, de la production et des stocks de blé depuis 1998 (source CIC / Arvalis-Institut du végétal). (B) Évolution des prix mondiaux, aux USA et en France, du blé tendre en dollars par tonne (source ONIGC / Arvalis-Institut du végétal).

I Présentation du blé et de sa famille (les *Poaceae*)

I.1 Importance économique du blé

Le blé est une céréale aux enjeux économiques très importants. Il occupe le troisième rang mondial des céréales, en volume récolté, avec 606 millions de tonnes (Mt) en 2007, derrière le maïs (792 Mt) et le riz (660 Mt) (Tableau 1, source FAO : http://faostat.fao.org/ au 20/10/2009). Ces céréales représentent, à elles trois, plus de 85% de la production céréalière mondiale (Tableau 1) et constituent la base de la nutrition humaine. Ce sont aussi d'importantes sources potentielles d'énergie renouvelable sous la forme de biocarburants.

La culture du blé s'étend sur 214 millions d'hectares (MHa) au niveau mondial, soit plus de 17% de la surface cultivée totale pour les plantes (Tableau 1). C'est aussi la céréale la plus importante en France, avec 32,8 Mt récoltées en 2007 pour une surface cultivée de 5,2 MHa (Tableau 1). Les rendements très élevés en France, en Allemagne et au Royaume-Uni (4 à 11 tonnes par hectare), assurent à l'Union Européenne une place de leader mondial dans la production de blé (190 Mt en 2007).

Cependant, ces niveaux de production peinent à satisfaire la demande mondiale qui ne cesse de s'accroître avec l'augmentation de la population (Figure 1A). Les très mauvaises récoltes de ces dernières années, causées par des aléas climatiques, associées aux utilisations non alimentaires ont particulièrement contribué à la diminution du stock mondial de blé, entraînant une hausse des prix sans précédent (de 100 €/tonne en juin 2006 à 300 €/tonne en avril 2008, avec des pics à 500 €/tonne au début 2008) (Figure 1B). L'amélioration de la production de blé est donc un sujet d'actualité. La compréhension de ses génomes et de leur évolution est un des moyens d'y contribuer.

I.2 Evolution du blé et des autres espèces de *Poaceae*

I.2.1 Positionnement taxonomique du blé parmi les *Poaceae*

L'appellation blé regroupe de nombreuses espèces qui appartiennent, selon la classification hiérarchique des espèces, aux angiospermes (plantes à fleurs) monocotylédones

Triticum			***Aegilops***		
Espèce	Génome	Autre appellation	Espèce	Génome	Autre appellation
Diploïdes :			Diploïdes :		
Triticum monococcum			***Aegilops markgrafii***	C	*Aegilops caudata*
ssp. monococcum	A^m	*Triticum monococcum*	***Aegilops tauschii***	D	
ssp. aegilopoides	A^u	*Triticum boeoticum*	***Aegilops comosa***	M	
Triticum urartu	A^u		***Aegilops uniaristata***	N	
Tétraploïdes :			***Aegilops speltoides***	S	
Triticum turgidum			***Aegilops bicornis***	S^b	
ssp. durum	BA^u	*Triticum durum*	***Aegilops longissima***	S^l	
ssp. dicoccoides	BA^u	*Triticum dicoccoides*	***Aegilops sharonensis***	S^{sh}	
ssp. dicoccon	BA^u	*Triticum dicoccon*	***Aegilops searsii***	S^s	
ssp. carthlicum	BA^u	*Triticum carthlicum*	***Aegilops umbellulata***	U	
ssp. turanicum	BA^u	*Triticum turanicum*	Tétraploïdes :		
ssp. polonicum	BA^u	*Triticum polonicum*	***Aegilops triuncialis***	UC	
ssp. paleocolchicum	BA^u	*Triticum karamyschevii*	***Aegilops cylindrica***	CD	
Triticum timopheevii			***Aegilops ventricosa***	DN	
ssp. timopheevii	GA^u	*Triticum timopheevii*	***Aegilops crassa***	D^cM	
ssp. armeniacum	GA^u	*Triticum araraticum*	***Aegilops biuncialis***	UM	
Hexaploïdes :			***Aegilops columnaris***	UM	
Triticum aestivum			***Aegilops kotschyi***	US	
ssp. aestivum	BA^uD	*Triticum aestivum*	***Aegilops peregrina***	SU	*Aegilops variabilis*
ssp. compactum	BA^uD	*Triticum compactum*	***Aegilops geniculata***	MU	*Aegilops ovata*
ssp. macha	BA^uD	*Triticum macha*	Hexaploïdes :		
ssp. spelta	BA^uD	*Triticum spelta*	***Aegilops crassa***	D^cDM	
ssp. sphaerococcum	BA^uD	*Triticum sphaerococcum*	***Aegilops trivialis***	DMS	
ssp. vavilovii	BA^uD	*Triticum vavilovii*	***Aegilops juvenalis***	DMU	
Triticum zhukovskyi	GA^mA^u		***Aegilops vavilovii***	DMS	
			Aegilops neglecta	UMN	*Aegilops recta*

Tableau 2. Les différentes espèces et sous espèces de blé appartenant aux genres *Triticum* et *Aegilops* et leurs génomes correspondants. La classification GrainTax (http://wheat.pw.usda.gov/ggpages/GrainTax/index.shtml), qui s'appuie sur 12 classifications antérieures, a été utilisée comme référence. Certaines espèces ont plusieurs appellations venant de méthodes de classification différentes. Les lettres (A,B,D,G) pour les espèces du genre *Triticum* et (C,D,M,N,S) pour celles du genre *Aegilops* représentent les génome diploïdes correspondants. Dans le cas des polyploïdes, la (les) lettre(s) représentant le génome maternel est (sont) indiquée(s) en premier. Une lettre supplémentaire, en exposant, peut être utilisée pour préciser l'espèce associée au génome. Les deux espèces les plus cultivées, *T. aestivum* ssp. *aestivum* et *T. turgidum* ssp. *durum*, sont soulignées.

de la famille des *Poaceae* (poacées en français, anciennement graminées), de la sous-famille des *Pooideae* et de la tribu des *Triticeae*.

La famille des *Poaceae* compte plus de 600 genres et 10.000 espèces, poussant sous des latitudes et des climats diversifiés (Kellogg 2001). De nombreuses espèces de cette famille ont été domestiquées et représentent un intérêt agronomique majeur : le riz (genre *Oryza*), le maïs (genre *Zea*), le sorgho (genre *Sorghum*), l'avoine (genre *Avena*), le seigle (genre *Secale*), l'orge (genre *Hordeum*) et le blé (genres *Triticum* et *Aegilops*) (Figure 2).

L'organisation des différentes espèces du blé, appartenant donc à deux genres, n'est pas triviale. En effet, il existe de nombreuses classifications différentes, très inconsistantes selon qu'elles se basent sur des critères botaniques ou génétiques. La classification GrainTax (http://wheat.pw.usda.gov/ggpages/GrainTax/index.shtml) a été créée dans le but d'unifier et de compiler les informations des différentes classifications existantes. J'ai utilisé cette classification pour présenter les principales espèces de blé (Tableau 2). Je n'ai cependant pas mentionné tous les croisements effectués récemment en laboratoire et déposés comme nouvelles espèces (par exemple le triticale, croisement entre le blé et le seigle). On obtient alors six espèces (19 sous-espèces) de blé du genre *Triticum* et 24 espèces du genre *Aegilops*. Comme pour d'autres espèces, les génomes du blé sont classifiés et désignés par des lettres différentes de l'alphabet (A, B, C, D…). Les génomes des espèces très proches sont symbolisés par la même lettre.

I.2.2 Evolution des génomes du blé au sein des *Poaceae*

Divergence des espèces, synténie et polyploïdisations communes

L'intérêt économique d'un grand nombre d'espèces de *Poaceae* en a fait une famille assez étudiée en génomique. Les premières études comparatives ont montré une bonne conservation des gènes et de leur ordre à l'échelle des chromosomes (on parle de synténie) (Moore *et al.* 1995). Elles ont aussi montré que les espèces de cette famille ont divergé d'un ancêtre commun il y a environ 70 Ma (Moore *et al.* 1995, Prasad *et al.* 2005).

Le séquençage complet des génomes de plusieurs espèces de *Poaceae* a permis une meilleure compréhension de l'évolution de leurs génomes. A ce jour, les génomes complets d'une espèce de riz (*Oriza sativa* : International Rice Genome Sequencing Project 2005), de sorgho (*Sorghum bicolor* : Paterson *et al.* 2009) et de *Brachypodium* (*Brachypodium distachyon*, The International Brachypodium Initiative 2010), couvrant trois tribus différentes (Figure 2), sont disponibles. Le séquençage du maïs est également sur le point d'être terminé

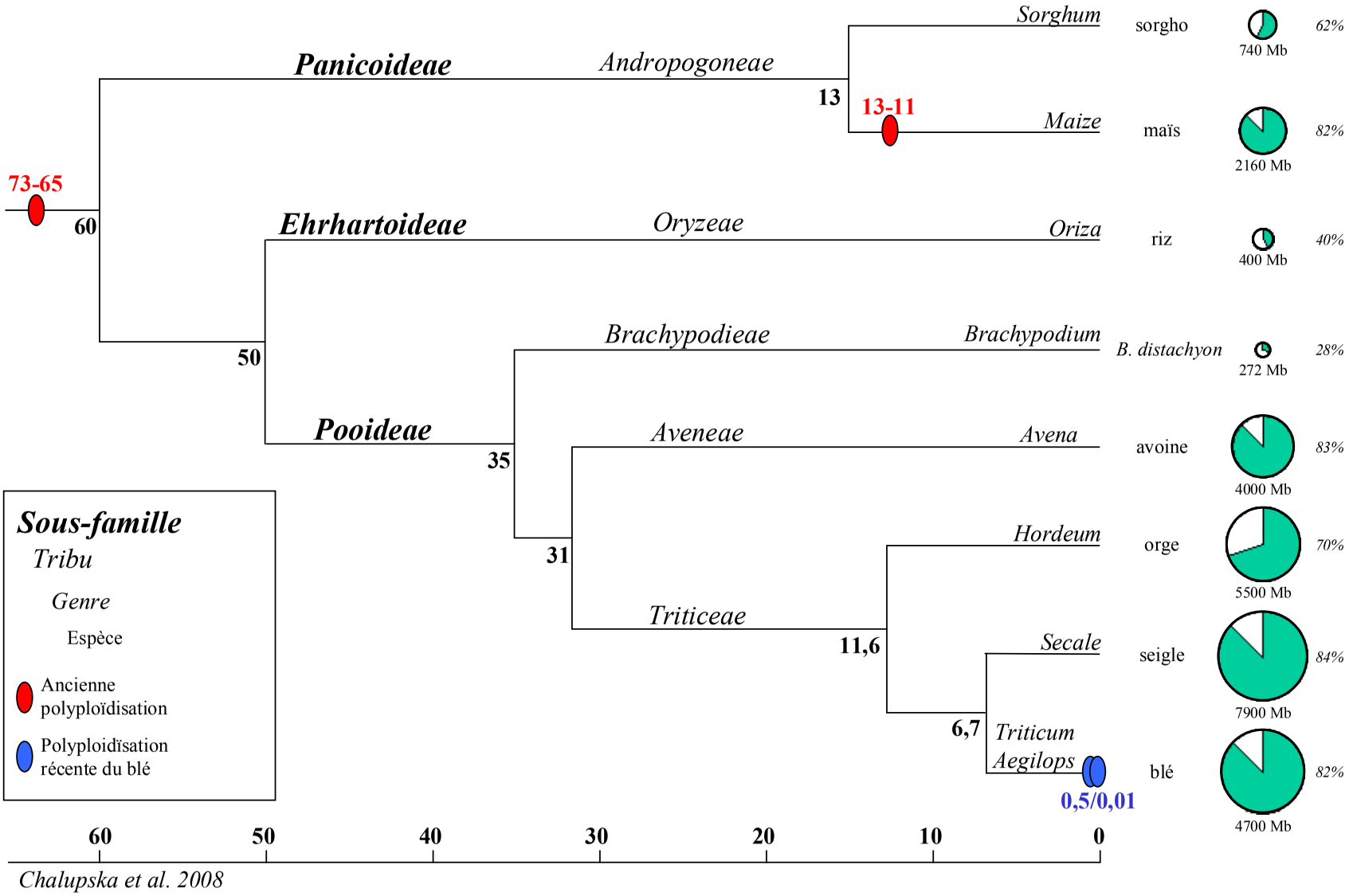


Figure 2. Divergence des espèces de la famille des *Poaceae* et évènements de polyploïdisation. Les divergences indiquées entre les espèces sont tirées de Chalupska *et al.* (2008). Les tailles des différents génomes ainsi que leurs proportion en TEs (portion verte) viennent de Paterson *et al.* (2009) pour le sorgho, le maïs et le riz, de Voghel *et al.* (*2010*) pour *B. distachyon,* Paux *et al.* (2006) et Charles *et al.* (2008) pour le blé, Bartos *et al.* (2008) pour le seigle, Bennett et Smith (1976) pour l'avoine et Wicker *et al.* (2009b) pour l'orge.

(Consortium for Maize Genomics, http://www.maizegenome.org/) et servira de modèle pour le séquençage de grands génomes comme le blé.

Les espèces de *Poaceae* ont divergé relativement rapidement à partir d'un ancêtre commun, il y a environ 60-70 millions d'années (Ma) dominant des systèmes écologiques et agricoles assez variés (Kellogg 2001, Gaut 2002, Prasad *et al.* 2005). Les dates de divergence de plusieurs espèces de *Poaceae* ont été estimées sur la bases de comparaisons des séquences des gènes orthologues d'acetyl-CoA carboxylase (*Acc-1* et *Acc-2*) présents généralement en une seule copie dans les *Poaceae* (Huang *et al.* 2002, Chalupska *et al.* 2008). La divergence entre les *Panicoideae* (mais / sorgho) et les *Ehrhartoideae* (riz), estimée à 60 Ma, a servi de référence pour calibrer la divergence moléculaire et calculer les divergences entre les espèces de *Poaceae*. Chalupska *et al.* (2008) ont ainsi estimé la divergence du blé avec le riz à 50 Ma, l'avoine à 31 Ma, l'orge à 11,6 Ma et le seigle à 6,7 Ma (Figure 2). Il est généralement admis que les différentes espèces du blé portant les génomes A, B et D, ont divergé il y a 2,5-4 Ma (Huang *et al.* 2002, Chalupska *et al.* 2008) (Figure 2).

Les séquences complètes des génomes du sorgho, du riz et de *Brachypodium* ont permis de confirmer la très bonne synténie entre les différentes espèces de *Poaceae* (Figure 3A). Les analyses des duplications dans les génomes complets et des ESTs ont également confirmé qu'au moins une duplication complète du génome (polyploïdisation) est commune à tous les *Poaceae*, ayant eu lieu il y a 65-73 Ma (The International Brachypodium Initiative 2010). La présence d'un autre événement plus ancien (>200 Ma), commun aux dicotylédones reste à confirmer (Adams et Wendel 2005, Tang *et al.* 2008b).

Evolution des chromosomes des Poaceae à partir de cinq chromosomes ancestraux

Les analyses comparatives ont également permis la reconstruction de l'évolution des chromosomes et la compréhension de la diversification des espèces des *Poaceae.* Il a ainsi été montré que leurs chromosomes ont évolué à partir d'un set de cinq chromosomes ancestraux (Salse *et al.* 2008a, Figure 3B). Les analyses du génome de *Brachypodium* ont révélé que l'insertion de chromosomes entiers dans les centromères d'autres chromosomes représente un mécanisme majeur de la variation du nombre de chromosomes chez les *Poaceae*. (The International Brachypodium Initiative 2010)

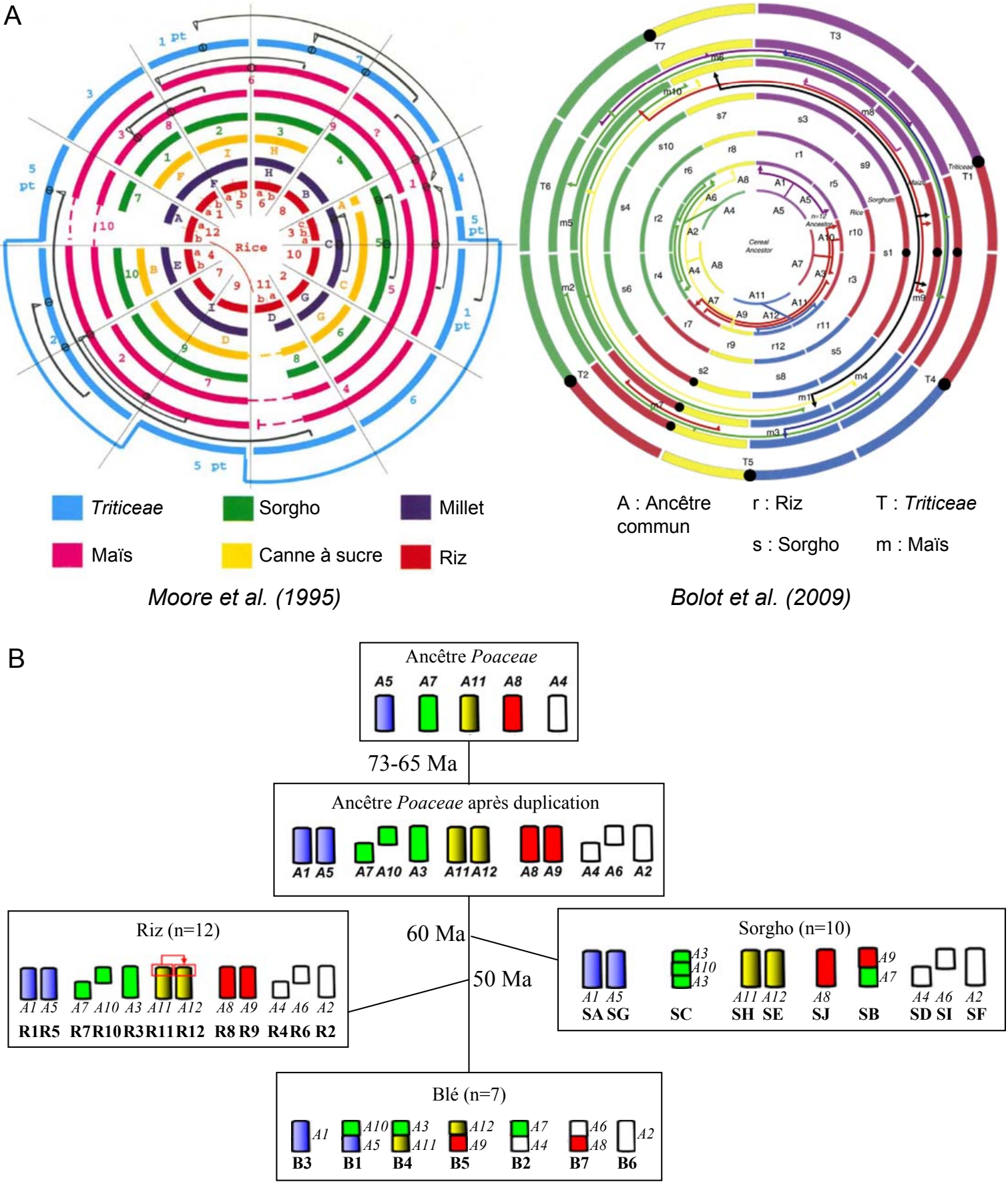


Figure 3. Conservation de synténie entre les chromosomes des différentes espèces de *Poaceae*. (A) Les cercles représentent l'alignement des chromosomes de différentes espèces de *Poaceae*, basé sur des marqueurs génétiques (Moore *et al.* 1995) et sur les données de séquences (Bolot *et al.* 2009). (B) Évolution de la structure des chromosomes du blé, du riz et du sorgho à partir des 5 chromosomes ancestraux (Salse *et al.* 2008a).

Prolifération des éléments transposables des Poaceae et polyploïdisations récentes dans le blé

Les génomes des différentes espèces des *Poaceae* montrent une grande diversité de tailles. Ces espèces partagent pourtant les mêmes événements de polyploïdisation ancestraux. Les importantes différences de tailles observées (400 Mb pour le riz contre 16,7 Gb pour le blé hexaploïde) sont donc dues à des événements de polyploïdisations récents et/ou à la prolifération différentielle des éléments transposables (TEs). En comparant les tailles haploïdes des espèces de *Poaceae* (Figure 2) et leur teneur en TEs, on observe que les grands génomes (>2 Gb) d'espèces de *Pooideae* (blé, avoine, orge) et de *Panicoideae* (maïs) montrent une prolifération des TEs plus importante que les petits génomes (<800Mb) d'autres espèces de ces même tribus (*Brachypodium* et sorgho respectivement). De plus, les familles de TEs rencontrées dans les différentes espèces ne sont pas les mêmes (Wicker *et al.* 2009a, Paterson *et al.* 2009). Ces comparaisons suggèrent une prolifération différentielle et indépendante des TEs dans les espèces de *Poaceae* (Parties I et II de Résultats).

Les espèces du blé ont eu de nombreux événements récents d'allopolyploïdisation, aboutissant à des espèces de blé tétraploïdes et hexaploïdes (Tableau 2). Combiné avec une prolifération très importante des TEs dans ces génomes (80% de la séquence), on obtient des génomes de très grande taille, allant de 4,7 Gb pour l'espèce diploïde *T. urartu* (AA) à 16,7 Gb pour le blé tendre hexaploïde *T. aestivum* ssp. *aestivum* (BBAADD).

La prolifération des TEs et la polyploïdie sont donc les deux forces majeures de l'évolution des génomes du blé et la suite de l'étude bibliographique les décrit en détail.

II Les éléments transposables

II.1 Présentation

Les éléments transposables (TEs) sont des fragments d'ADN génomique répétés, qui ont la capacité de se déplacer et de modifier le nombre de leurs copies au sein de leur génome hôte. Ils ont été historiquement découverts par Barbara McClintock (McClintock, 1950) dans le maïs avec le couple d'éléments mobiles (*Activator* /*Dissociator*) provoquant des cassures chromosomiques. Ils sont alors appelés 'éléments régulateurs'.

Ils ont depuis été détectés dans la plupart des génomes eucaryotes et procaryotes. Depuis les années 80, de nombreux éléments très différents ont été découverts dans les génomes, rendant indispensable leur classification.

II.1.1 Classification

On trouve des TEs dans tous les organismes vivants, mais leur importance relative dans le génome et leur diversité peuvent grandement varier selon l'espèce considérée. La première classification proposée (Finnegan 1990, Capy *et al.* 1998) repose sur les différents modes et mécanismes de réplication observés (transposition). On distingue ainsi les TEs de classe I qui utilisent un intermédiaire ARN pour transposer et les TEs de classe II qui utilisent un intermédiaire ADN.

Ces deux mécanismes correspondent aussi à deux modes différents de propagation. Les éléments de classe I génèrent une nouvelle copie de l'élément à partir de la copie originale qui va s'insérer à une autre place dans le génome (mécanisme de 'copier-coller'). Pour les éléments de classe II, la copie originale est excisée du génome puis réintégrée à un endroit différent (mécanisme de 'couper-coller').

Les TEs de classe I sont appelés rétroélements en rapport avec la transcriptase inverse dont ils se servent pour copier l'ARN en ADN. On distingue deux catégories de rétroéléments : les rétrotransposons et les rétroposons. Les rétrotransposons sont reconnaissables par leurs LTRs (Long terminal Repeat) qui correspondent à des longues séquences répétées au début et à la fin de l'élément. Les éléments de classe II sont appelés transposons à ADN.

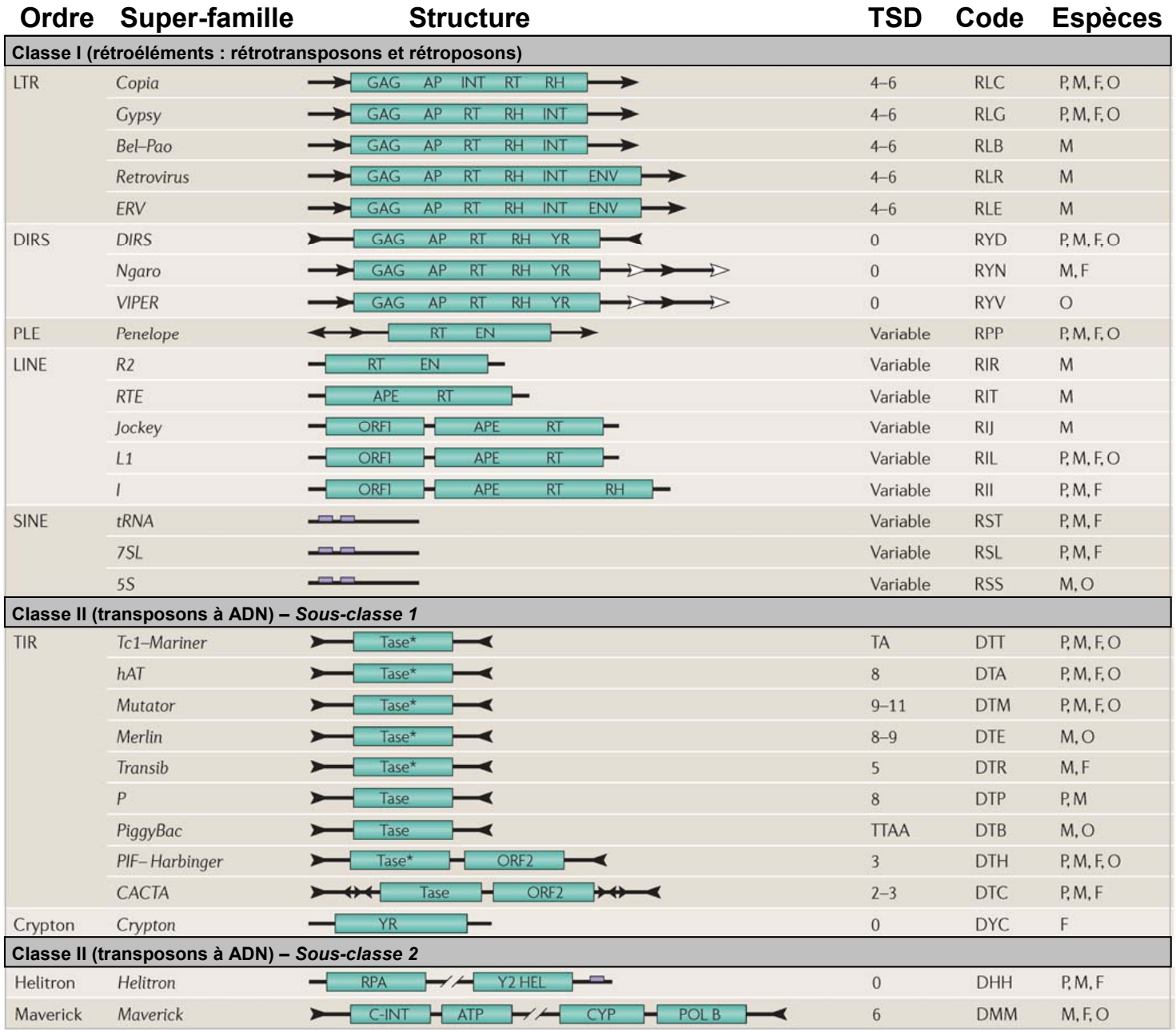


Figure 4. Classification des différents éléments transposables sur 5 niveaux : Classe, Sous-classe, Ordre, Super-famille, Famille d'après Wicker *et al.* (2007a). Le dernier niveau, sous-famille, n'est pas illustré sur cette figure. Les éléments caractéristiques de chaque catégorie sont mentionnés, ainsi que les espèces dans lesquelles ils ont été trouvés. La colonne TSD indique la taille de ceux-ci pour les éléments de cette famille. Code indique le préfixe à faire figurer devant l'élément pour sa nomenclature. ORF1 et ORF2 représentent des protéines dont la fonction est inconnue

L'arrivée massive de séquences depuis une dizaine d'année, provenant de nombreux génomes différents, a permis l'identification de nombreux nouveaux éléments. Un nouveau modèle de classification plus organisé est devenu nécessaire, notamment avec la découverte dans le maïs d'éléments utilisant un nouveau mécanisme de transposition (dit par cercle coulant ou 'rolling-circle') (Kapitonov et Jurka 2001).

Un modèle hiérarchique (Figure 4) a été récemment proposé (Wicker *et al.* 2007a) pour classer les TEs d'eucaryotes en fonction de leur mode de transposition (au niveau mécanistique et enzymatique). Il prend en compte d'autres modèles (Jurka *et al.* 2005), mais aussi des règles pratiques venant des retours d'experts en annotation de TEs. On peut citer la règle dite '80-80-80' basée sur des similarités de séquences : deux éléments appartiennent à la même famille s'ils ont plus de 80% d'identité sur 80% de leur séquence sur un minimum de 80 pb. Le système binaire (classe I / classe II) a été conservé et enrichi de 5 niveaux : sous-classe, ordre, super-famille, famille et sous-famille.

- Les classes restent définies par la présence ou non d'un intermédiaire ARN lors de la transposition (transposons à ARN ou transposons à ADN).
- Les sous-classes séparent les transpositions par 'copier-coller' des transpositions par 'couper-coller'. Tous les éléments de classe I appartiennent donc à la même sous-classe ('copier-coller'), alors que les éléments de classe II peuvent appartenir à l'une ou à l'autre.
- Les ordres séparent les transpositions ayant des caractéristiques enzymatiques et organisationnelles différentes.
- Les super-familles différencient des éléments ayant une même stratégie de réplication, mais une conservation au niveau protéique très limitée. (Figure 4, super-familles des *copia* et des *gypsy*)
- Les éléments d'une même famille ont une forte conservation au niveau protéique (>80%).
- Certaines familles ont des membres se regroupant en sous-familles sur des critères de similarité (conservation nucléique ou regroupement dans des arbres phylogénétiques). Exemple : *Wis* et *Angela* sont deux sous-familles de *BARE-1*.

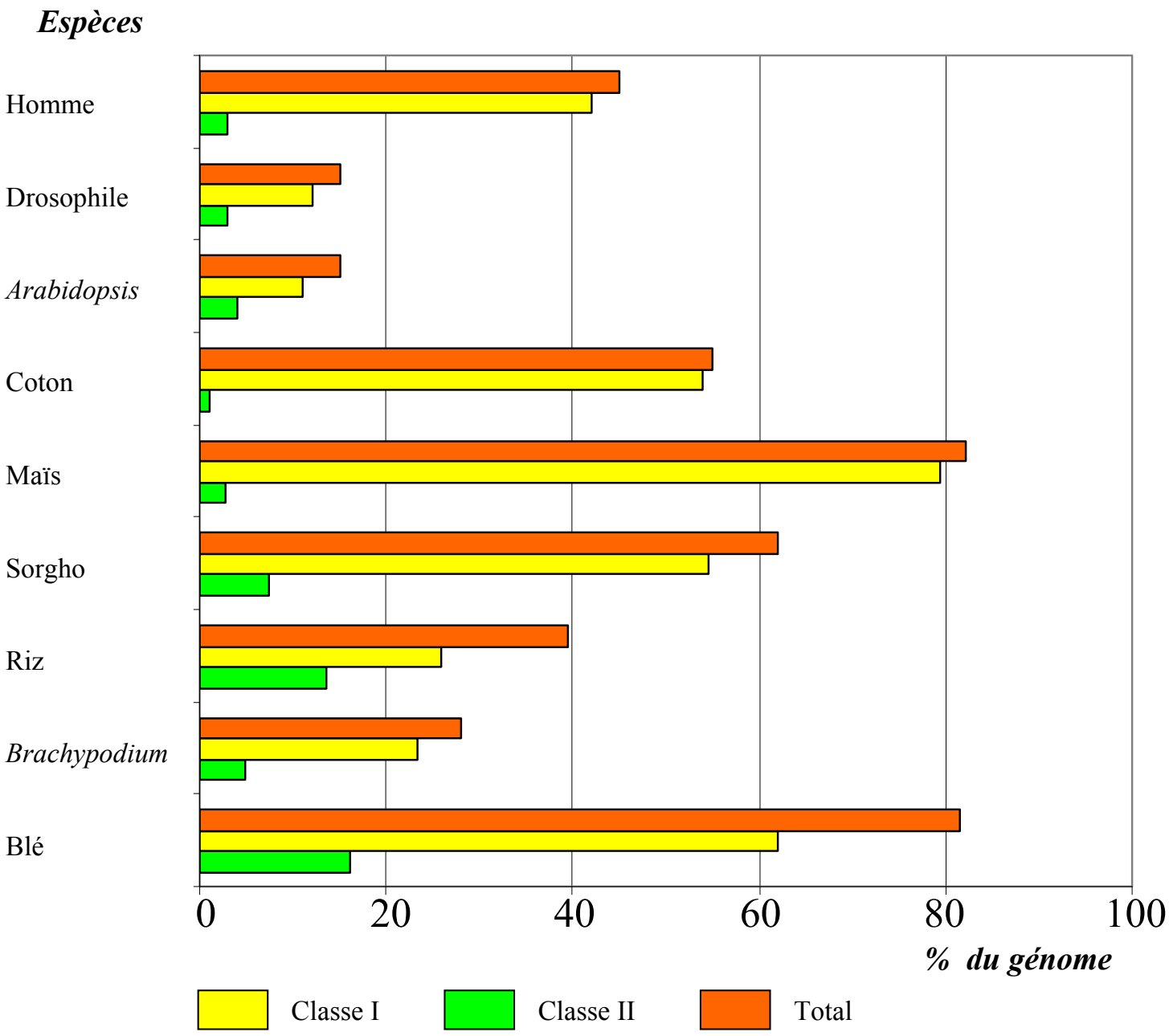


Figure 5. Proportion des TEs de classe I et II dans différentes espèces animales et végétales. Ces proportions viennent des données de séquencage (Aradidopsis Genome Initiative 2000, Lander *et al.* 2001, Kaminker *et al.* 2002, International Rice Genome Sequencing Project 2005, Paterson *et al.* 2009, Voghel *et al.* Soumis) ou d'études représentatives (Hawkins *et al.* 2006, Paux *et al.* 2006, Piegu *et al.* 2006, Charles *et al.* 2008). Ces données ont ensuite été complétées par les données d'une revue récente des TEs dans les eucaryotes (Pritham 2009), en particulier pour les proportions de classe I et II.

Cette proposition de classification, même si elle n'est pas exempte de défauts, permet d'organiser simplement les TEs et aide à annoter les nouveaux éléments (décrits pour la première fois).

II.1.2 Importance des TEs dans le blé et les autres espèces

Les génomes du blé sont particulièrement riches en TEs représentant plus de 80% de leur séquence génomique (Paux *et al.* 2006, Charles *et al.* 2008 publié dans le cadre de cette thèse). Les TEs sont des éléments très importants et dynamiques des génomes du blé. On observe ainsi des variations de taille très importantes entre les différentes espèces de blé pouvant atteindre plusieurs centaines de Mb pour des espèces ayant le même niveau de ploïdie (Bennett et Smith 1976, 1991, http://data.kew.org/cvalues/homepage.html).

La proportion des TEs est aussi très variable chez les plantes dicotylédones [le coton (Hawkins *et al.* 2006) et *Arabidopsis* (Arabidopsis Genome Initiative 2000)] (Figure 5) et les espèces animales [l'homme (Lander *et al.* 2001) et la drosophile (Kaminker *et al.* 2002, Hoskins *et al.* 2007)]. Par exemple, ils occupent 45% de la séquence du génome humain, et environ 15% du génome de la drosophile (Figure 5).

La proportion des TEs de classes I et II est également très variable selon les espèces (revue dans Pritham 2009). Les éléments de classe I représentent plus de 60% des génomes du blé mais seulement 10% du génome de la drosophile. Ils sont globalement très abondants dans les plantes (Figure 5). Les éléments de classe II sont près de 10 fois plus importants dans le génome du blé (20%) que celui du maïs (3%).

La taille des TEs peut également varier de quelques dizaines de paires de bases (pb) à quelques dizaines de milliers de paires de bases (kb), selon la famille des éléments. Certaines espèces contiennent plusieurs centaines de familles d'éléments alors qu'une seule famille d'éléments peut représenter la grande majorité des TEs présents dans un génome (élément *Alu* chez l'homme, *BARE* chez l'orge).

II.2 Dynamique des éléments transposables

II.2.1 Prolifération des TEs

Les variations de la composition en TEs trouvées dans les espèces illustrent des dynamiques de prolifération des TEs très différentes. Cette prolifération est la résultante de

deux forces d'évolution antagonistes : l'activité insertionnelle des TEs et leur élimination (SanMiguel *et al.* 1996, Bennetzen et Kellogg 1997, Bennetzen 2000b, 2002a, Petrov *et al.* 2000, Petrov 2002a, Kidwell 2002, Wendel *et al.* 2002, Bennetzen *et al.* 2005, Hawkins *et al.* 2006, Piegu *et al.* 2006, Zuccolo *et al.* 2007). La disponibilité et l'abondance des séquences de TEs à l'échelle génomique ont permis une caractérisation de ces deux forces (SanMiguel *et al.* 1998, 2002, Wicker *et al.* 2003b, 2005, Gao *et al.* 2004, Ma *et al.* 2004, Du *et al.* 2006, Piegu *et al.* 2006, Wicker et Keller 2007 ainsi que mes travaux de thèse). Il est en effet devenu possible de déterminer :

- la proportion des TEs et des différentes familles dans le génome
- la proportion des copies complètes par rapport aux copies tronquées
- la distribution des TEs le long des chromosomes
- l'estimation des dates d'insertion des rétrotransposons ayant leur deux LTRs

La prolifération des TEs n'est pas constante au cours de l'évolution des espèces, ni homogène le long des chromosomes. On distingue des périodes d'activité très fortes et des périodes de plus faible intensité. De plus, elle semble différente d'une espèce à une autre et d'une famille de TEs à une autre. Chez les angiospermes, les estimations des dates d'insertions des rétrotransposons n'excèdent pas 3 millions d'années alors que chez les gymnospermes, la plupart des estimations sont supérieures à 35 Ma (De Paoli 2009, Morgante *communications personnelles*). De plus, un grand nombre de copies apparaissent tronquées suggérant des mécanismes d'élimination (délétion) des TEs.

Il est difficile de séparer ces deux forces lorsque l'on étudie la dynamique globale des TEs dans un génome. Par exemple, des pic apparents d'insertions de TEs ('bursts') dans un génome peuvent résulter d'une forte activité insertionnelle et/ou d'une faible vitesse d'élimination. Pendant ma thèse, j'ai pu caractériser la prolifération des TEs et le taux de remplacement de l'espace occupé par les TEs dans le blé, résultant de ces deux forces d'évolution (Partie I et II).

II.2.2 Mécanismes d'insertion des TEs

Les éléments transposables qui peuvent coder les protéines essentielles pour leur transposition sont dits autonomes. Ils ont néanmoins besoin de la machinerie de la cellule hôte pour exprimer ces protéines. Lorsqu'un élément autonome est actif, il peut transposer mais

aussi induire la transposition d'éléments de la même famille dits non-autonomes qui n'encodent pas toutes les protéines essentielles pour leur transposition. Un élément non-autonome peut avoir conservé les sites de reconnaissance (peu spécifiques) de la machinerie de transposition et être mobilisé en 'trans'.

Une région codante parfaitement conservée ne garantit pas l'activation et la transposition de l'élément et inversement une région codante dégradée ne va pas forcément empêcher la transposition. Un faible nombre de copies autonomes actives dans un génome peut permettre la transposition d'un grand nombre de copies non-autonomes ('trans' activation). Les mécanismes de réplication des TEs de classes I et II sont très différents. Les éléments de classe I passent par un intermédiaire ARN. Le mode de transposition par 'copier-coller' des éléments de classe I les rend potentiellement très invasifs pour un génome. Il n'est donc pas surprenant de les voir représenter parfois une grande partie d'un génome. C'est particulièrement le cas des rétrotransposons, plus fréquents chez les plantes que chez les animaux, occupant jusqu'à 60% des grands génomes de céréales (blé, orge ou maïs). Il est intéressant de constater que quelques familles de TEs peuvent représenter une grande partie de la multitude d'éléments transposables trouvés dans un génome.

Je décris dans cette partie les mécanismes supposés de transpositions des principaux types d'éléments transposables de classe I et II.

Mécanisme de transposition des rétrotransposons

Les rétrotransposons sont les éléments les plus répandus dans les génomes du blé (prés de 60% de la séquence). Les sous-familles *Wis* et *Angela* de la famille *BARE-1 (Copia)* sont les plus fréquentes et représentent à elles seules de 10 à 20% des génomes du blé (Charles *et al.* 2008). Les familles *Sabrina* (*Athila*) et *Fatima* (*Gypsy*) occupent chacune près de 7% des génomes.

Les rétrotransposons autonomes encodent une polyprotéine comprenant deux domaines : *GAG* et *Pol*. Les deux domaines sont transcrits en une fois (Figure 6A, étape 1). La traduction du messager donne la protéine GAG (protéine capside) et le complexe de protéines Pol, divisé en 4 protéines distinctes au niveau post-traductionnel (RT : transcriptase inverse, INT : intégrase, RH : RnaseH, AP : protéase aspartique) (Figure 6A, étape 2). Les protéines GAG vont se polymériser dans le cytoplasme pour former des VLPs (Virus Like Particules) (Figure 6A, étape 3). Une partie de l'ARNm va entrer dans ces VLPs, se dimériser et, sous l'action de RH et RT, être retro-transcrit en ADN double brin (Figure 6A, étape 4). L'intégrase (INT) va former une cassure double brin de l'ADN génomique et y intégrer

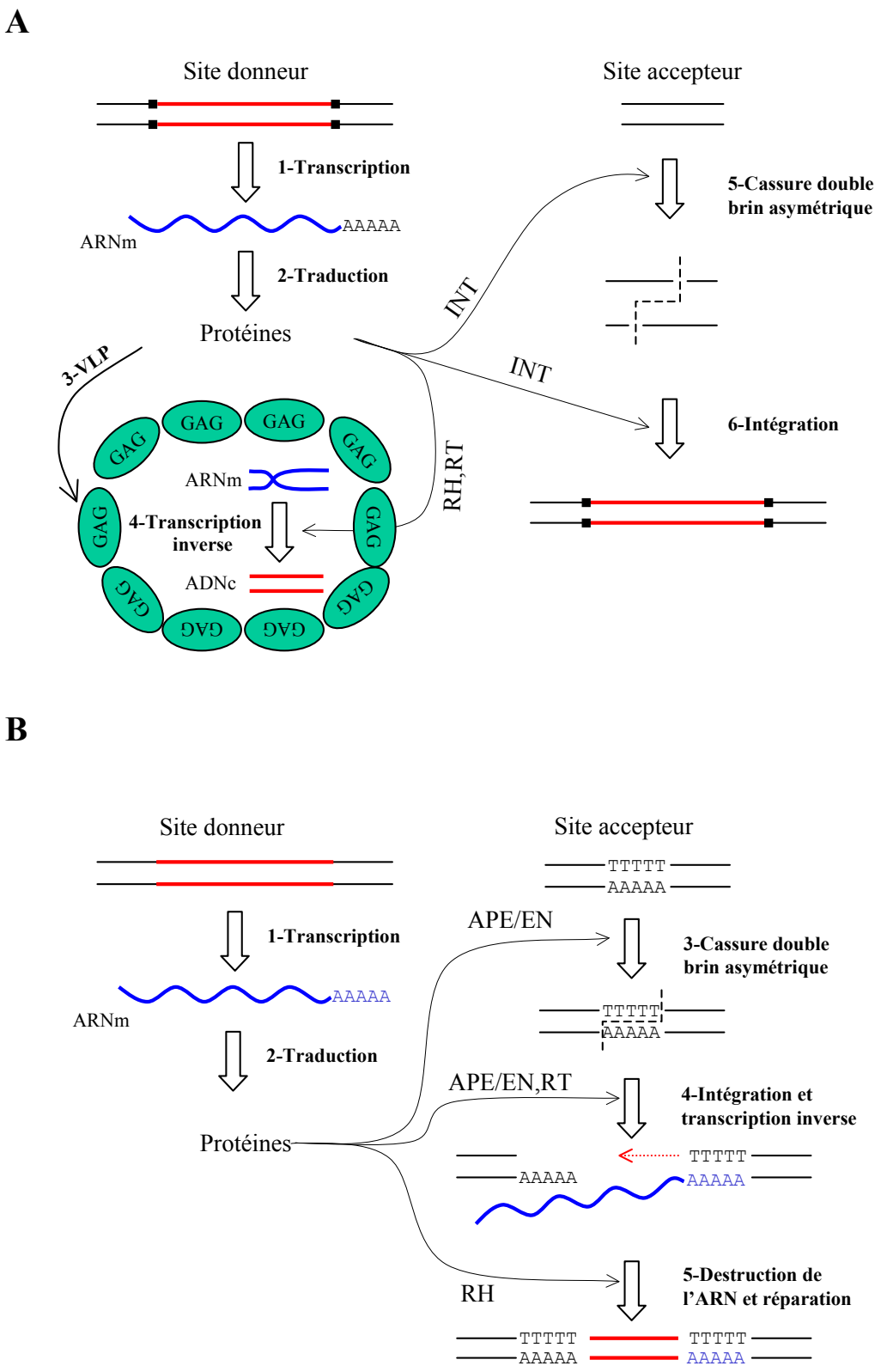


Figure 6. Mécanismes de transposition des principaux TEs de classe I, d'après les schémas de Sabot *et al.* (2004). (A) Transposition des rétrotransposons. Les carrés noirs indiquent les TSDs. (B) Transposition des rétroposons. Les différentes étapes des transpositions sont détaillées dans le texte correspondant. VLP : Virus like Particule, GAG : protéine capside.

l'ADN du rétrotransposon (Figure 6A, étape 5 et 6). La réparation de cette cassure asymétrique va former les TSDs (Target Site Duplication), signatures caractéristiques laissées par l'insertion de la plupart des TEs.

Le cycle de transposition supposé des rétrotransposons est calqué sur celui des rétrovirus. Ils ont en effet une origine commune et gardent de nombreux points communs, même si l'absence de protéine d'enveloppe chez les rétrotransposons reste une différence notable, les confinant dans la cellule et empêchant leur transfert horizontal.

Mécanisme de transposition des rétroposons

Les rétroposons sont peu fréquents dans les génomes du blé (autour de 2%) et sont essentiellement de la famille des *LINEs.*

Les rétroposons autonomes (comme les *LINEs*) ont un mécanisme de transposition assez différent de celui des rétrotransposons. Pourtant, le principe général de la transposition reste le même : transcription puis traduction des parties codantes de l'élément par la machinerie de l'hôte, cassure asymétrique double brin de l'ADN au site cible suivi par l'intégration de l'élément. Une des différences principales entre les deux mécanismes vient de la transcription inverse qui se fait sur le lieu d'insertion pour les rétroposons, et pas dans une VLP. Un rétroposon complet a généralement deux cadres ouverts de lecture (ORFs) : ORF1 et ORF2. Le premier code pour une protéine pouvant se lier à l'ADN et le deuxième pour une endonucléase (APE ou EN), une transcriptase inverse (RT) et parfois une RnaseH (RH). L'endonucléase va permettre de lier un ARNm de l'élément avec la partie 3' d'un fragment d'ADN libre après une cassure double brin asymétrique, provoquée ou non par l'endonucléase (Figure 6B, étapes 3 et 4). La transcription inverse (action de RT) commence alors sur place, dans le sens 3' vers 5' (Figure 6B, étape 4). Elle se déroule cependant rarement en entier (10% des cas), produisant donc de nombreuses copies partielles de l'élément d'origine. Une fois la transcription terminée, l'ARNm est digéré (action de RH) et la cassure double brin se répare, dupliquant ainsi le rétroposon sur le deuxième brin d'ADN (Figure 6B, étape 5).

Mécanisme de transposition des transposons à ADN

Les transposons à ADN sont généralement moins invasifs que les rétrotransposons dans les génomes de plantes, car ils transposent par un système de 'couper-coller'. Cependant, ils constituent dans le blé jusqu'à 16% du génome et appartiennent quasiment tous (99%) à la superfamille des *CACTA*. Pour expliquer cette amplification importante, il a été suggéré que

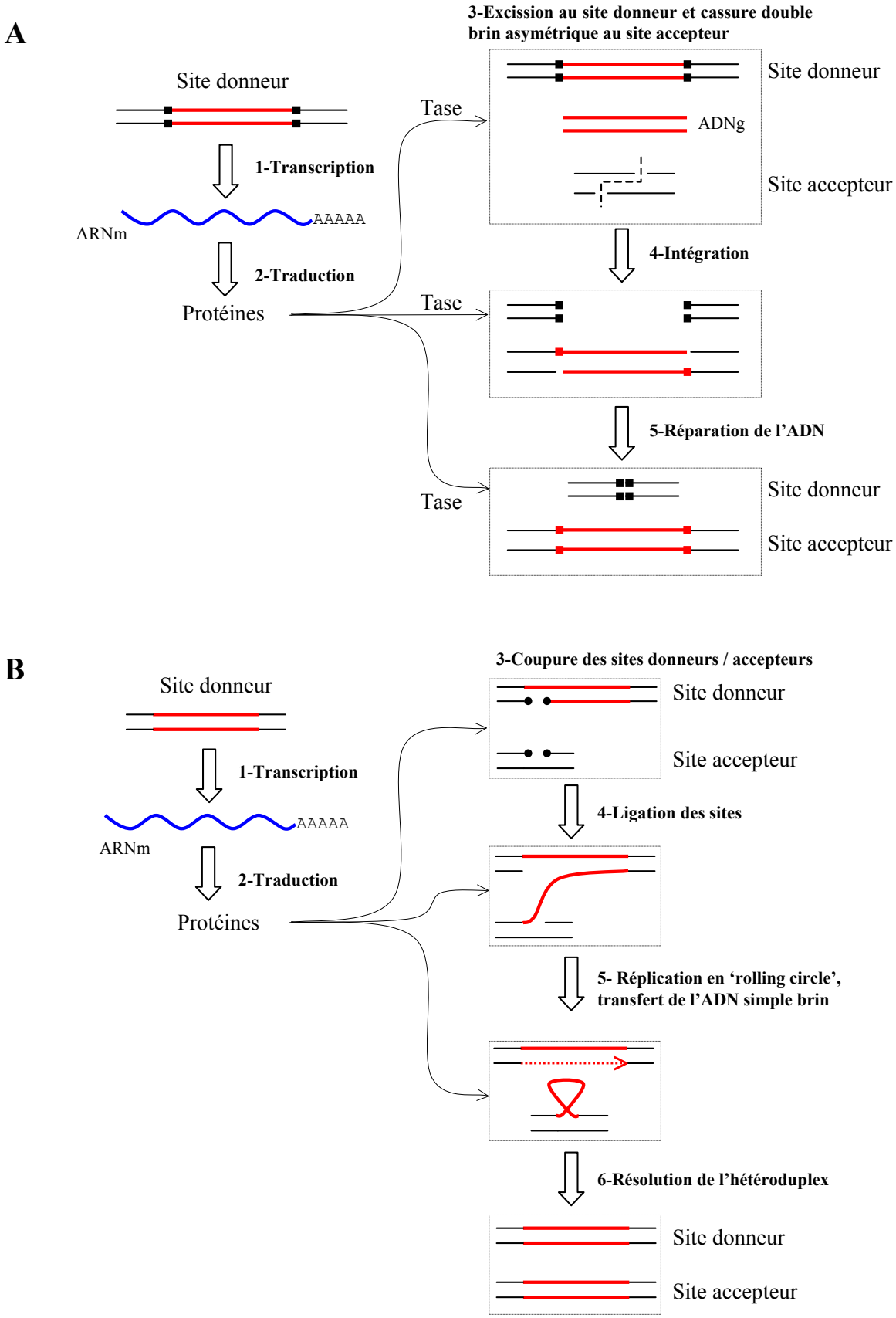


Figure 7. Mécanismes de transposition des principaux TEs de classe II, d'après les schémas de Sabot *et al.* (2004) et Kapitnov et Jurka (2007). (A) Transposition des transposons à ADN de type *CACTA*. Les carrés noirs et rouges indiquent les TIRs. (B) Transposition des helitrons. Les points noirs indiquent le site de cleavage entre A et T. Les différentes étapes des transpositions sont détaillées dans le texte correspondant.

leur transposition à des moments précis du cycle cellulaire permettrait d'augmenter leur nombre (décrit ci-après).

Les transposons à ADN complet de type *CACTA* ont deux ORFs, le premier encodant pour une transcriptase (Tase) et le second pour une protéine pouvant se lier à l'ADN mais dont le rôle reste à préciser (Wicker *et al.* 2003a). La transcriptase reconnaît spécifiquement les TIRs (Tandem Inverted Repeat) présents aux extrémités de ces TEs et catalyse toutes les étapes de la transposition, de l'excision à l'intégration (Figure 7A, étape 3, 4 et 5). La transposition se fait donc par un mécanisme de 'couper-coller', conservatif au niveau du nombre de copies. Mais si la transposition se produit en phase S du cycle cellulaire, en aval de la fourche de réplication, la cassure double brin provoquée par le transposon va se réparer en utilisant la chromatide sœur comme modèle possédant encore l'élément. L'élément est donc copié à un autre endroit du génome tout en gardant une copie à sa position d'origine.

Mécanisme de transposition des helitrons

Les helitrons, découverts relativement récemment (Kapitonov et Jurka 2001), utilisent un mécanisme singulier, dit de 'rolling-circle' en référence à un mécanisme similaire observé chez les bactéries, pour leur transposition. Ces éléments, transposant sans intermédiaire ARN, par un mécanisme original, et par 'copier-coller', étaient difficiles à décrire selon les critères de la première classification. Ils ont ainsi contribué à la formation d'une nouvelle classification, présentée précédemment. Contrairement à la plupart des autres éléments de classe II, ils n'ont pas de TIRs, motifs reconnus spécifiquement par les transposases. Ils commencent par TC, finissent par CTRR (R étant A ou G) et ont aussi une séquence palindromique, de 16-20 pb situé une dizaine de pb en amont du CTRR. Ils s'insèrent entre les bases A et T d'un site accepteur. Selon les espèces, la composition des helitrons autonomes peut être assez variable, mais ils ont tous en commun un ORF codant pour la protéine 'RepHel' (Kapitonov et Jurka 2007), constituée des domaines *Rep* (Réplicase) et *Hel* (Helicase). Pour la transposition d'un helitron, le domaine *Rep* se lie aux sites donneur et accepteur, coupe l'ADN (simple brin) entre les bases A et T des deux sites (Figure 7B, étape 3) et fait une ligation entre la partie 3' du site donneur et la partie 5' du site accepteur (Figure 7B, étape 4). Le domaine *Hel* catalyse ensuite la synthèse de l'helitron par l'ADN polymérase de l'hôte au niveau du site donneur (Figure 7B, étape 5). En fin de synthèse, l'ADN simple brin correspondant à l'helitron original est transféré au site accepteur formant ainsi un hétéroduplex (Figure 7B, étape 5). Lors d'une prochaine réplication de l'ADN, cet hétéroduplex va être résolu par la réplication de l'helitron sur le deuxième brin du site

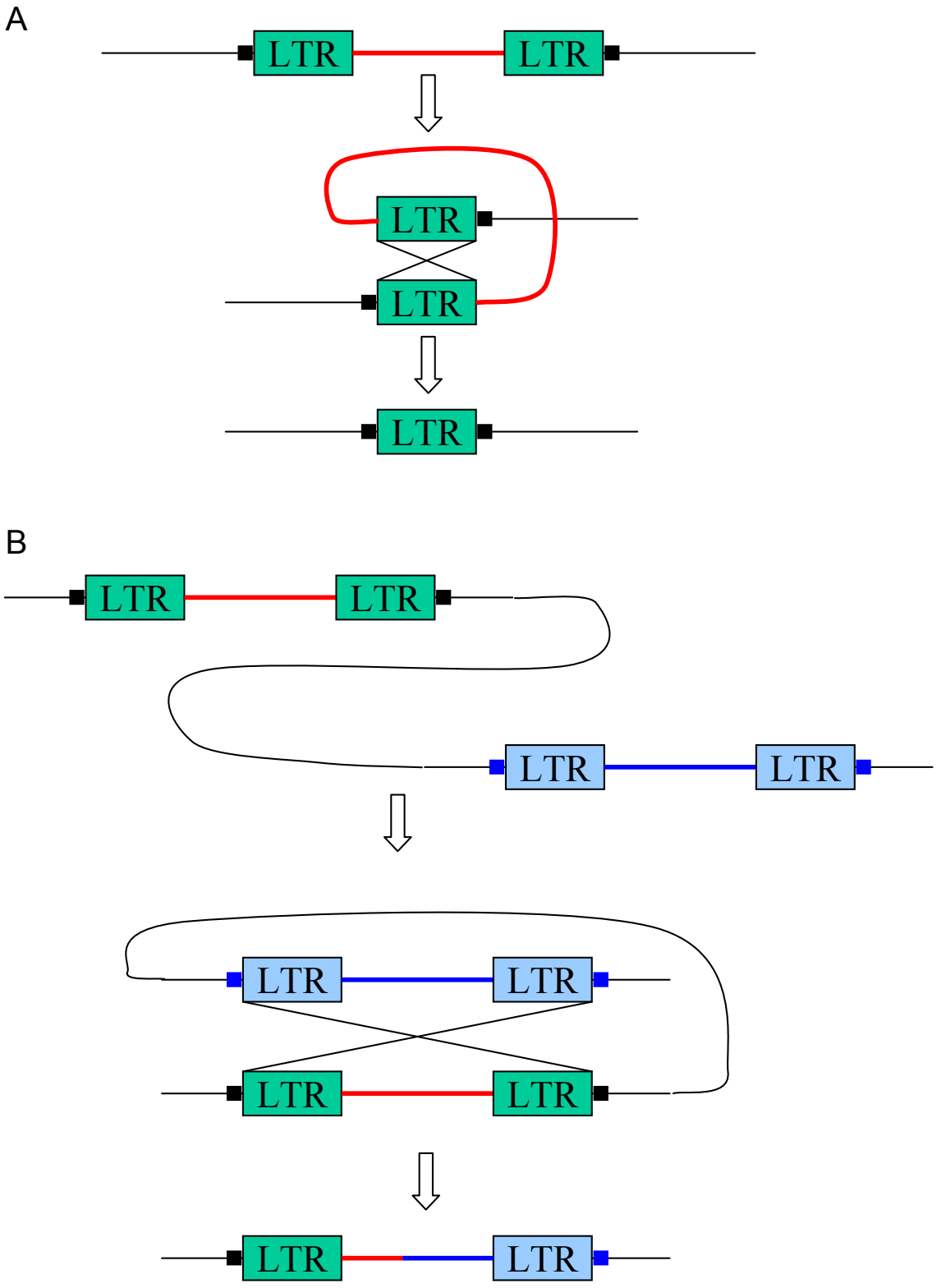


Figure 8. Mécanismes schématiques de délétion par recombinaison homologue inégale. (A) Formation d'un Solo-LTR avec TSD par recombinaison homologue inégale entre les deux LTRs d'un rétrotransposon sur le même brin d'ADN. (B) Formation d'un rétrotransposon complet mais chimérique, sans TSD, par recombinaison entre deux éléments suffisamment similaires, sur le même brin d'ADN. Toute la séquence d'ADN entre les deux éléments est éliminée.

donneur (Figure 7B, étape 6). Le palindrome à la fin de l'helitron joue le rôle de terminateur de la réplication par 'rolling-circle'. S'il n'est pas bien reconnu, la synthèse continue sur le site donneur, si bien que le brin d'ADN transféré au site accepteur peut transporter non seulement l'helitron, mais aussi des séquences en 3' du site donneur. Ce phénomène a été notamment observé dans le maïs (Morgante *et al.* 2005) où des fragments de gènes voire des gènes entiers sont ainsi transportés à d'autres endroits du génome.

II.2.3 Mécanismes d'élimination des TEs

Les éléments transposables ne restent pas intacts dans le génome hôte après leur insertions, comme le montre le nombre important de copies tronquées et dégénérées trouvées dans les séquences génomiques disponibles. Les deux principaux mécanismes responsables de cette élimination des TEs sont les recombinaisons homologues inégales et les recombinaisons illégitimes (Devos *et al.* 2002, Ma *et al.* 2004, Ma et Bennetzen 2004, Vitte et Bennetzen 2006).

Les recombinaisons homologues inégales font intervenir deux séquences suffisamment longues (généralement plusieurs centaines de pb) et similaires (>85-100% d'identité, selon la taille). Les rétrotransposons sont donc particulièrement concernés par ce type de recombinaison puisque leur LTRs remplissent très bien ces critères. La recombinaison peut ainsi avoir lieu entre les deux LTRs d'un même élément (formation d'un Solo-LTR) (Figure 8A) ou des LTRs d'éléments de la même famille (Figure 8B), aboutissant à la délétion plus ou moins importante de toute la séquence entre les deux éléments. Le taux de recombinaisons homologues, révélées par la présence de Solo-LTR et d'éléments complets sans TSD, n'est pas très important dans le blé (taux de 1/50, Charles *et al.* 2008 publiés dans le cadre de cette thèse).

Les recombinaisons qui ne sont pas homologues sont dites 'illégitimes'. Le mécanisme des ces recombinaisons illégitimes est encore mal caractérisé. Elles impliquent souvent des motifs de quelques pb, conservés dans des orientations variables (directs, complémentaires, anti-sens). On observe ainsi des délétions allant de quelques pb à plusieurs dizaines de kb (Chantret *et al.* 2005). Ces recombinaisons sont très fréquemment associées avec les TEs (intra-éléments ou inter-éléments). J'aborde en détail les différents mécanismes observés et les conséquences des recombinaisons homologues inégales et illégitimes dans les Résultats (Partie I et II).

II.3 Rôle des TEs dans les génomes

Au début des années 80, les TEs étaient considérés comme des parasites de l'ADN ou du génome, de l'ADN poubelle égoïste (Doolittle et Sapienza 1980, Orgel et Crick 1980) ne servant en rien la cellule ou l'organisme et n'ayant que peu d'effets notables au niveau phénotypique. Ils avaient pourtant été désignés comme 'éléments régulateurs' à leur découverte (McClintock 1950). Ce n'est qu'à la fin des années 90, après la découverte de leur quasi-ubiquité et l'analyse de leurs effets sur les génomes grâce à la disponibilité croissante de leur séquence, que leur statut a peu à peu changé de 'parasites' à partenaire symbiotique.

Une des conséquences de leur activité est la génération de diversité au sein d'une population et d'une espèce par une action directe ou indirecte. L'insertion d'un TE dans la séquence codante d'un gène va le plus souvent l'inactiver. Son insertion dans la zone promotrice du gène ou la présence d'un LTR, contenant des régions promotrices peut aussi altérer l'expression du gène (Kashkush *et al.* 2003). La methylation par l'organisme d'un TE peut également s'étendre jusqu'à un gène proche et conduire à l'inactivation transcriptionnelle ('silencing') de celui-ci. Les TEs peuvent aussi participer à la création de nouveaux exons ou même de nouveaux gènes. En effet, des études ont montré que certains éléments transposables (*Helitrons*, *PACK-Mule*, *CACTA* éléments de classe II) peuvent emporter des exons (ou même la totalité) des gènes adjacents lorsqu'ils transposent (Jiang *et al.* 2004, Lai *et al.* 2005, Zabala et Vodkin 2005). Ce phénomène 'd'exon-shuffling', aboutit le plus souvent à des pseudo gènes. Parfois, ces fragments sont capturés par un gène existant (modifiant la fonction du gène) ou peuvent même former, dans de rares cas, un gène complètement nouveau.

La première 'domestication' (recrutement) d'un TE par l'organisme a été identifié chez *Drosophila* (Levis *et al.* 1993). Les TEs de type LINE (*HET-A* et *TART*) sont laissés libres de s'insérer dans les télomères remplaçant ainsi l'activité d'une télomérase. Depuis, de nombreux autres cas ont été détectés chez les animaux comme chez les végétaux où l'organisme utilise les TEs d'une facon qui lui est bénéfique (Agrawal *et al.* 1998, Lynch et Tristem 2003, Kapitonov et Jurka 2004, 2005, Gao et Voytas 2005, Muehlbauer *et al.* 2006). Chez l'homme, ils ont contribué à la formation de notre système immunitaire (Agrawal *et al.* 1998).

Plus les TEs sont étudiés, plus on découvre le rôle important qu'ils ont joué ou jouent dans les organismes. Les études récentes ont montré leur implication au niveau de la forme et de la fonction des chromosomes par leur insertion dans l'hétérochromatine autour des centromères et des télomères mais aussi au niveau de la régulation de l'expression et les modifications de la chromatine par leurs liens avec les RNAi (revue dans Slotkin et Martienssen 2007).

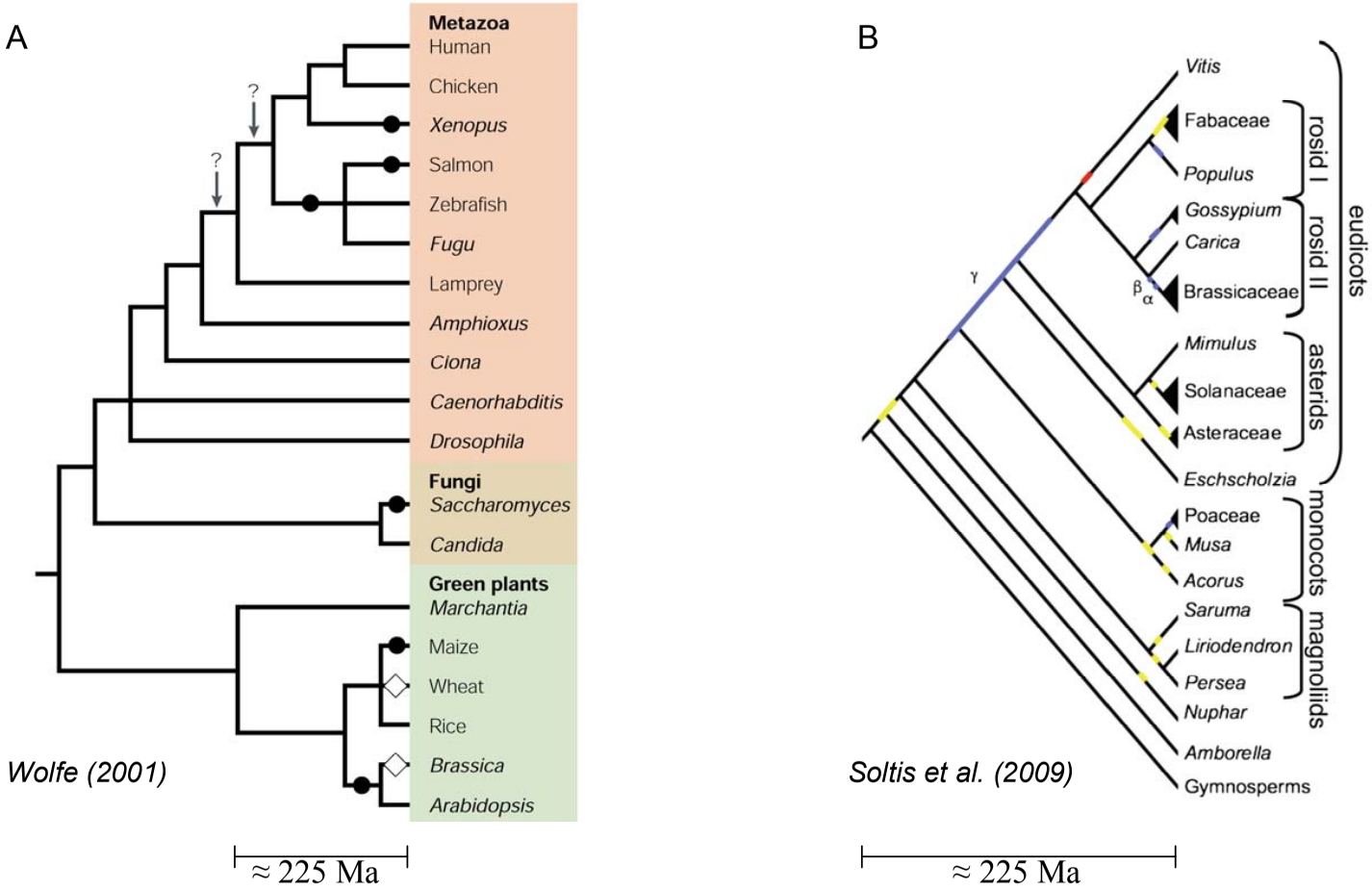


Figure 9. Evènements de polyploïdisation au cours de l'évolution des eucaryotes et des angiospermes. (A) Polyploïdisation chez les eucaryotes par Wolfe (2001). Les points d'interrogation marquent les emplacements possibles de 2 cycles de duplication globale (2R, Hughes 1999). Un cercle noir indique un évènement de polyploïdisation (allo- ou auto-) tandis qu'un diamant indique 2 évènements.(B) Polyploïdisation chez les angiospermes par Soltis *et al.* (2009). Les barres bleus et jaunes indiquent des duplications entières de génome montrées par l'analyse de la séquence de génome complet ou d'EST. La barre rouge indique une duplication alternative de la vigne (Velasco *et al.* 2007). Le moment précis de la duplication gamma est incertain, notamment pour savoir si elle est commune ou non aux plantes monocotylédones.

III La polyploïdie

La polyploïdie, ou l'assortiment de plusieurs jeux complets de chromosomes dans un noyau, est un facteur important dans l'évolution des génomes des eucaryotes. Elle est fréquente chez les plantes, particulièrement chez les angiospermes (Wolfe 2001, Adams et Wendel 2005, Soltis *et al.* 2009) (Figure 9). On distingue l'**autopolyploïdie** (les jeux de chromosomes proviennent de la même espèce) et l'**allopolyploïdie** (les jeux de chromosomes viennent d'espèces différentes, mais suffisamment proches pour s'hybrider, on parle de chromosomes homéologues).

Certains événements de polyploïdisation détectés chez les *Poaceae* sont anciens (on parle alors de paléopolyploïdie) et donc communs à toutes les espèces de cette famille. D'autres événements sont plus récents et donc restreints à certaines tribus / espèces (Figure 2). C'est le cas pour les génomes du blé où de nombreux événements de polyploïdisation récents et récurrents ont été détectés (Tableau 2). Certains de ces évènements ont abouti aux espèces modernes de blé cultivées : le blé dur *T. turgidum* ssp. *durum* tétraploïde (2n=4x=28, BBAA) et le blé tendre *T. aestivum* ssp. *aestivum* hexaploïde (2n=6x=42, BBAADD) (Tableau 2, souligné).

Ce chapitre décrit les différents aspects de la polyploïdie (formation, fréquence) ainsi que son rôle dans l'organisation, l'évolution et le fonctionnement des génomes, en particulier ceux du blé.

III.1 Mécanismes de formation des polyploïdes

Les organismes avec des cellules contenant deux copies de chaque chromosome sont dits diploïdes (2n=2x). Leur méiose produit des gamètes haploïdes (n) et la fusion de ces gamètes forme un embryon diploïde. Les cellules avec plus de deux jeux de chromosomes sont dites polyploïdes (tétraploïde : 2n=4x, hexaploïde : 2n=6x …). Le niveau de ploïdie d'un organisme est celui de l'ensemble de ses cellules, en excluant les quelques cas particuliers ayant un nombre de chromosomes différents comme les gamètes et les cellules endomitotiques.

Deux mécanismes majeurs permettent d'expliquer la formation des polyploïdes : le doublement somatique du stock chromosomique et la fusion de gamètes non réduites trouvant leur origine dans des erreurs de division cellulaire.

III.1.1 Doublement somatique du stock chromosomique

L'absence de division cellulaire lors de la mitose, après la réplication des chromosomes, aboutit à la formation d'une cellule somatique polyploïde. En principe, elle ne conduit pas à la formation d'un individu polyploïde. Cependant, si le doublement chromosomique se produit dans une des cellules se différenciant en gamètes (pré-zygotique), dans l'œuf ou dans une des cellules du jeune embryon (post-zygotique), il peut être à l'origine de la formation d'un individu polyploïde viable.

Ce type de doublement peut être induit artificiellement par un traitement chimique comme la colchicine. Ce procédé est utilisé couramment en laboratoire pour produire des polyploïdes synthétiques. Cependant, il est peu probable que les polyploïdes se forment naturellement par cette voie.

III.1.2 Formation et fusion de gamètes non réduites

Les polyploïdes naturels se formeraient plutôt par l'intermédiaire de gamètes non-réduites (Harlan et De Wet 1975, Bretagnolle et Thompson 1995, Ramsey et Schemske 1998). Cette voie de formation implique l'autre grande division cellulaire : la méiose. Une méiose normale débute dans une cellule avec des paires de chromosomes homologues possédant chacun deux chromatides sœurs. Lors de la première division méiotique, les paires de chromosomes se séparent. Se sont ensuite les chromatides sœurs qui se séparent pendant la deuxième division pour former finalement quatre gamètes haploïdes (n) (Figure 10). On dit que ces gamètes sont réduites car elles ont chacune un seul jeu de chromosomes (à une chromatide). Une erreur lors de la méiose peut aboutir à la formation de gamètes diploïdes non réduites par des mécanismes encore mal connus. On en distingue quatre types différents :

- les gamètes FDR (First Division Restitution) : il n'y a pas de première division méiotique. La méiose aboutit à la formation de deux gamètes diploïdes et hétérozygotes contenant chacune les deux jeux de chromosomes avec une seule chromatide (Figure 10). C'est le type de gamètes non réduites le plus souvent observé chez le blé (Jauhar *et al.* 2007).

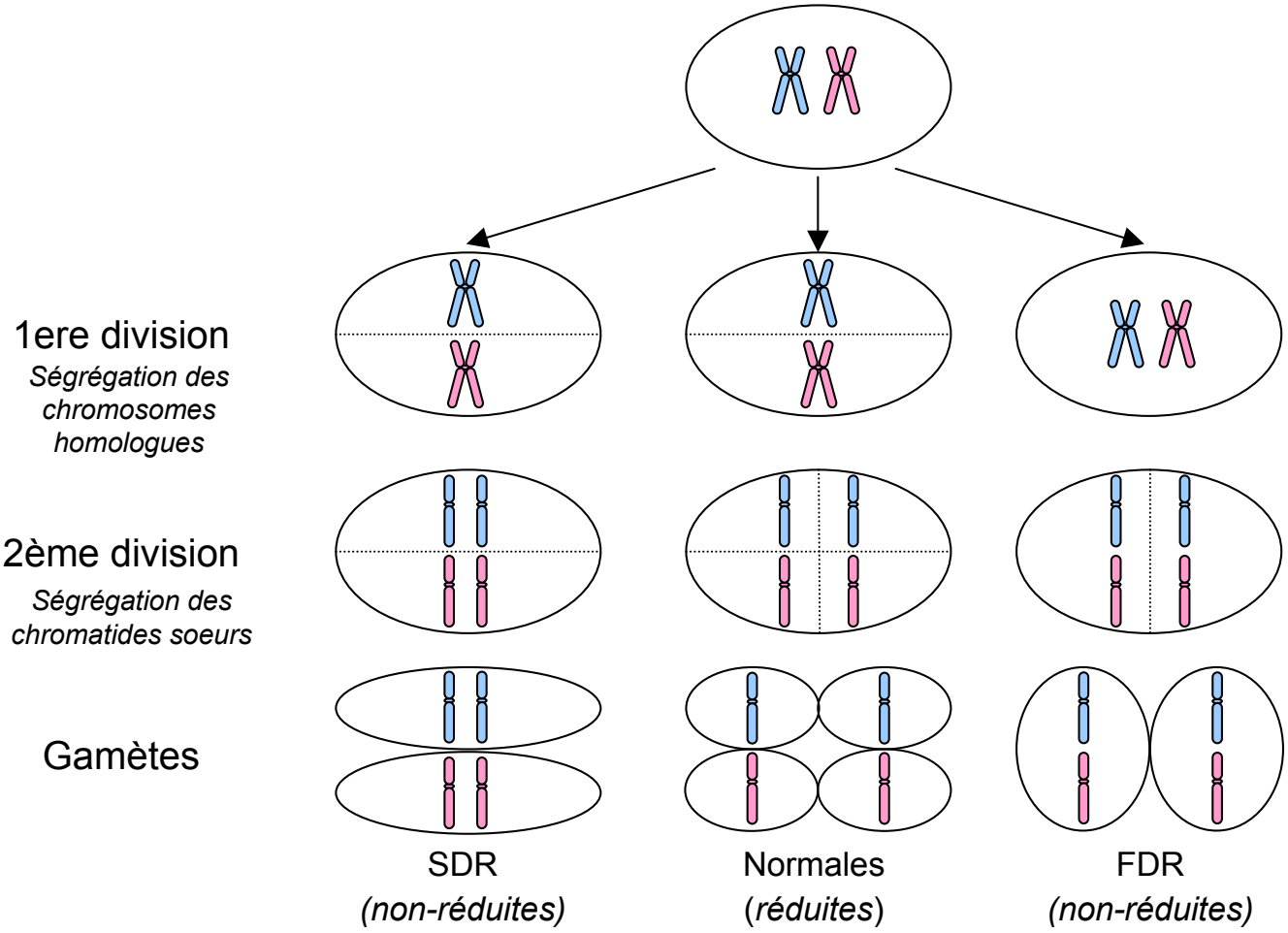


Figure 10. Formation des gamètes non-reduites. FDR / SDR : First / Second Division Restitution

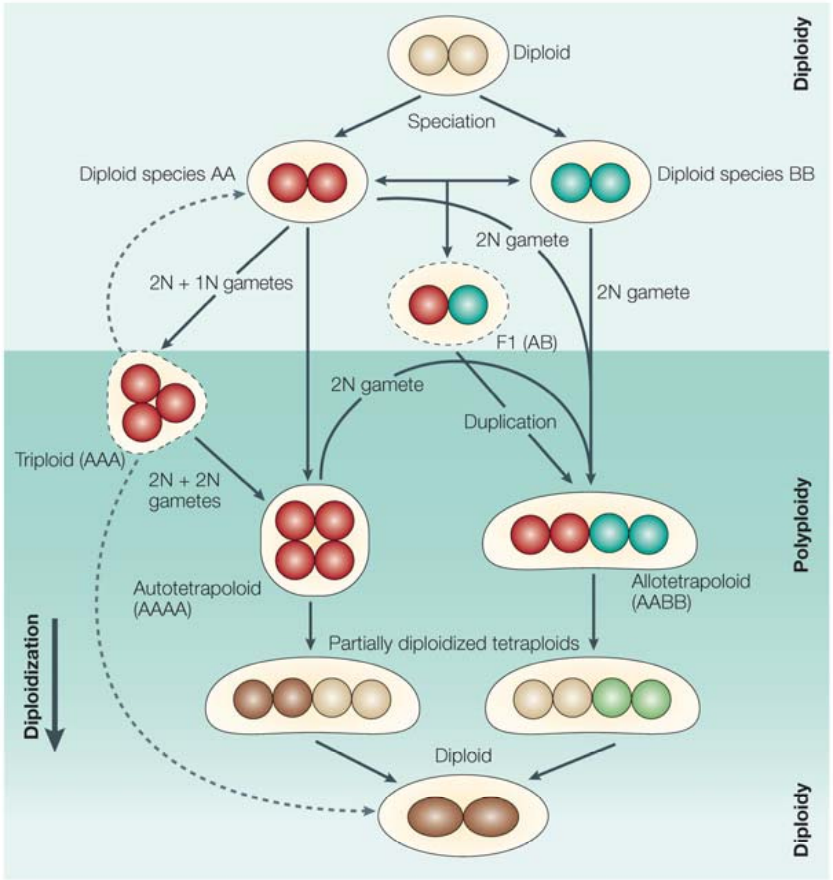


Figure 11. Différentes voies de formation des polyploïdes et de rediploïdisation d'après Comai (2005).

- les gamètes SDR (Second Division Restitution) : il n'y a pas de deuxième division méiotique. Les gamètes formées sont homozygotes et contiennent chacune une paire de chromosomes à une chromatide appartenant à l'un des deux jeux d'origine (Figure 10).
- les gamètes de type IMR (Indeterminate Meiotic Restitution) : certains chromosomes subissent une FDR et d'autres une SDR.
- les gamètes sans division : il n'y a aucune division pendant la méiose et le gamète obtenu possède les deux paires de chromosomes homologues d'origine ayant chacun deux chromatides.

Une fusion de ces gamètes non réduites (2n), lors de l'hybridation, aboutit donc à la formation d'un individu polyploïde. Si les gamètes viennent de la même espèce voire du même individu (autofécondation), on obtiendra un autopolyploïde. Si les gamètes viennent d'individus d'espèces différentes ou d'un hybride interspécifique, on obtiendra un allopolyploïde combinant deux génomes différents.

Les gamètes non réduites (2n) peuvent également fusionner avec des gamètes réduites (n) formant des polyploïdes avec un nombre impair de jeux de chromosomes (triploïde, pentaploïde…). Ils sont normalement stériles car la méiose ne peut se dérouler correctement (problème d'appariement des chromosomes) dans ces polyploïdes. Cependant, un doublement somatique dans ces individus ou la fusion de leur gamète non réduite (sans division) entre-elles ou avec des gamètes parentales (n) peut leur permettre de former des polyploïdes stables et fertiles (Figure 11). Les polyploïdes avec un nombre impair de chromosomes, bien que généralement stériles, peuvent former un pont entre des espèces avec des niveaux de ploïdie différents, et donc normalement isolées.

Par exemple, le blé tendre hexaploïde est issu d'un évènement de polyploïdisation entre un blé tétraploïde et un blé diploïde (Feldman *et al.* 1995) (détaillé ci-après). En laboratoire, ce croisement donne un hybride F1 triploïde stérile qui, une fois doublé (spontanément ou avec la colchicine), aboutit à l'hexaploïde fertile.

III.2 Fréquence de polyploïdisation

Au cours des 10 dernières années, les divers projets de séquençage complet des génomes ou d'ESTs ont révélé des événements de polyploïdisation fréquents et récurrents chez les plantes et particulièrement dans les angiospermes (Wolfe 2001, Blanc & Wolfe 2004,

Schlueter *et al.* 2004, Adams et Wendel 2005, Cui *et al.* 2006, Jaillon *et al.* 2007, Tang *et al.* 2008b, Paterson *et al.* 2009), qui n'avaient pas été détectés auparavant par les moyens de cytologie et de génétique. Ainsi, toutes les espèces d'angiospermes ont probablement subi des événements de polyploïdisation récurrents, anciens (paléopolyploïdes) et/ou récents, au cours de leur évolution (Figure 9A, 9B).

En comparaison, très peu d'espèces animales ont un récent passé polyploïde, à l'exception de quelques espèces de poissons et d'amphibiens (Wolfe 2001, Jaillon *et al.* 2004). La nécessité d'une seule paire de chromosomes sexuels chez les animaux explique en partie cette différence. La bonne tolérance des angiospermes à la formation de triploïdes et le rôle d'intermédiaire que ces derniers peuvent jouer dans la formation de polyploïdes stables est également une explication.

Le haut niveau de formation de polyploïdes naturels dans les angiospermes (1/100000 plantes) vient du taux élevé de gamètes non réduites produites (0,56%) dans ces espèces (Ramsey et Schemske 1998). Cette abondance de polyploïdes et de paléopolyploïdes (80-100%) semble très spécifique des angiospermes. Les gymnospermes ne montrent que 5% d'espèces polyploïdes alors que c'est le sous-embranchement le plus proche des angiospermes dans la classification des espèces.

III.3 Effets de la polyploïdie

La duplication complète du génome, et donc de tous les gènes, n'est pas sans effet sur l'organisme. Il y a déjà près de 40 ans, Ohno (Ohno 1970) suggérait le rôle évolutif majeur des gènes dupliqués, qui faciliterait l'émergence de nouvelles fonctions et la diversification d'une espèce. La polyploïdie offre donc une réserve de gènes accrue, augmentant potentiellement les possibilités d'évolution et de spécialisation fonctionnelle. Ainsi, les polyploïdes observés dans la nature sont en général bien adaptés à leur environnement. Cependant, la polyploïdisation reste un choc génomique et la présence de plusieurs génomes dans les cellules n'est pas sans inconvénient. Les espèces polyploïdes doivent surmonter rapidement les effets à court terme pour survivre. Je décris dans ce chapitre les principaux effets de la polyploïdie observés à court et à long terme.

III.3.1 Effets à court terme

Pour observer les effets à court terme de la polyploïdie, l'idéal serait de comparer un polyploïde naturel nouvellement formé (néopolyploïde) à ses parents (progéniteurs). Malheureusement, les polyploïdes et leurs progéniteurs (parfois difficiles à identifier précisément) trouvés dans la nature ont plus ou moins divergé depuis l'événement de polyploïdisation. Plus l'événement est ancien, plus il est difficile de séparer les effets de la polyploïdie de ceux de la divergence des génomes. Une autre approche consiste à comparer un polyploïde synthétisé en conditions de laboratoire aux géniteurs utilisés pour le produire. En pratique, on étudie le plus souvent des allopolyploïdes obtenus par croisement inter-spécifique. Si le doublement chromosomique ne s'est pas fait naturellement, on utilise alors la colchicine pour induire un doublement somatique et obtenir un polyploïde. Cette approche permet de comparer directement le polyploïde à ses progéniteurs mais fait souvent intervenir un agent chimique dont les effets seront donc ajoutés à ceux de la polyploïdie.

Au cours des 15 dernières années, la caractérisation des effets à court terme de la polyploïdie, utilisant principalement l'étude de polyploïdes synthétiques, a montré des changements au niveau structural (Song *et al.* 1995, Ozkan *et al.* 2001, Rieseberg 2001, Shaked *et al.* 2001, Gaeta *et al.* 2007) et fonctionnel (Kashkush *et al.* 2003, Adams & Wendel 2005, Wang *et al.* 2006, Flagel *et al.* 2008, Hovav *et al.* 2008, Rapp *et al.* 2009). L'importance des effets observés est variable selon les espèces étudiées et le modèle polyploïde choisi. Je présente ci-dessous une description non exhaustive des principaux effets observés à court terme de la polyploïdie.

Effets au niveau de la méiose

La polyploïdie fait cohabiter plusieurs génomes dans une seule cellule ce qui n'est pas sans inconvénients pour la méiose. Pour distribuer équitablement le matériel génétique à la méiose entre les gamètes, il faut que les chromosomes homologues s'apparient deux à deux. La présence de plus de deux jeux de chromosomes similaires (homologues et homéologues) complique cette phase d'appariement. Il peut se former des regroupements de plus de deux chromosomes aboutissant à la formation de gamètes avec un nombre variable de chromosomes. La production de ces gamètes, dites aneuploïdes, qui ne sont généralement pas viables entraîne donc une baisse de fertilité des polyploïdes. Ces regroupements anormaux peuvent donc se produire quand plus de deux chromosomes se 'ressemblent' suffisamment au niveau des séquences pour s'apparier pendant la méiose.

Chez les allopolyploïdes, une méiose régulière est donc un prérequis pour leur stabilité. L'appariement des chromosomes est alors restreint aux chromosomes homologues. Plusieurs mécanismes interviennent pour la stabilisation des appariements chez les polyploïdes. La divergence des chromosomes homéologues par des réarrangements structuraux rapides dans les néopolyploïdes, notamment dans des polyploïdes synthétiques du blé, serait un mécanisme permettant de réduire l'appariement entre ces chromosomes (Levy et Feldman 2004, Feldman et Levy 2005). Cependant, Comai *et al.* (2003) ont montré que les polyploïdes synthétiques d'*Arabidopsis* montrent un appariement homologue régulier dès les premières générations sans changements structuraux suggérant un autre mécanisme de stabilisation par contrôle génétique. Des gènes contrôlant l'appariement homéologue ont ainsi été identifiés dans plusieures espèces de polyploïdes stables. Chez le blé par exemple, le gène *Ph1* (Pairing homoeologous 1) empêche tout appariement homéologue dans les polyploïdes naturels et stables du blé comme *T. turgidum*, *T. timopheveii* et *T. aestivum* (Riley et Chapman, 1958).

Changements structuraux

De nombreux réarrangements structuraux (particulièrement des délétions d'ADN) ont été observés dans des études portant sur des polyploïdes synthétiques de blé (Feldman *et al.* 1997, Shaked *et al.* 2001), de *Brassica* (Song *et al.* 1995) ainsi que sur des polyploïdes très récents de *Tragopogon* (Tate *et al.* 2006). Ces modifications peuvent concerner indifféremment des séquences géniques et non géniques. Le relâchement de la pression de sélection au niveau des régions dupliquées, autorise des réarrangements rapides et importants. Dans les génomes riches en TEs, la duplication de ces éléments associée au relâchement de la pression de sélection, multiplie les possibilités de recombinaisons homologues inégales.

Il a été suggéré que les changements structuraux, partagés par les chromosomes homologues, vont les faire diverger des chromosomes homéologues correspondants, parfois rapidement, et donc contribuer à stabiliser les polyploïdes (voir le point précédent concernant les effets de la polyploïdie sur la méiose).

Les modifications structurales n'apparaissent pas forcément immédiatement dans les polyploïdes synthétiques. En effet, à court terme, il semblerait que les modifications fonctionnelles (décrites ci-après) soient prédominantes. Ces dernières sont réversibles dans une certaine mesure et donc plus flexibles en termes de possibilités d'évolution pour les polyploïdes pendant la période d'adaptation suivant l'évènement de polyploïdisation.

Changements fonctionnels

La polyploïdie augmente brutalement le nombre de copies des gènes de l'organisme et implique donc une reprogrammation de leur expression. La comparaison des transcriptomes (Puce à ADN) de polyploïdes d'*Arabidopsis*, de blé, de coton, de *Senecio*, de *Tragopogon* (He *et al.* 2003, Adams *et al.* 2003, Hegarty *et al.* 2005, Tate *et al.* 2006, Wang *et al.* 2006) et d'hybrides intra-spécifiques de maïs (Swanson-Wagner *et al.* 2006) avec leurs parents respectifs a montré que la grande majorité des gènes (de 82,5% à 95%) avait un niveau d'expression correspondant à la valeur moyenne de celle de leurs parents (on parle de profil additif).

Ces différentes comparaisons ont aussi montré que plus les parents étaient divergents, plus la proportion de profils non additifs était importante, les autopolypoïdes ne montrant ainsi quasiment que des profils additifs (Wang *et al.* 2006). Les profils non additifs correspondent donc à un niveau d'expression du gène dans le polyploïde différent de la moyenne des parents. Le gène peut avoir le niveau d'expression d'un seul de ses parents (dans le cas de profils parentaux différents) ou un niveau d'expression différent des deux parents. Les profils non additifs sont donc des indicateurs de modification dans l'expression du gène du polyploïde. Ces modifications peuvent avoir une origine structurale (modification / perte du gène) ou fonctionnelle (sur-expression, 'silencing', dominance génétique).

Les modifications de l'expression ne concernent pas que les gènes. La polyploïdie est un stress important pour l'organisme, et les TEs peuvent être réactivés dans ces conditions (revue dans Capy *et al.* 2000). Cette réactivation des TEs a été observée dans des polyploïdes synthétiques d'*Arabidopsis* (Madlung *et al.* 2005) et de blé (Kashkush *et al.* 2003).

Allopolyploïdie et hétérosis

Les différentes copies homéologues sont co-exprimées dans les polyploïdes. Les allopolyploïdes fixent ainsi l'hétérozygotie pouvant exister entre les parents. C'est une des raisons suggérées de l'effet hétérosis souvent observé dans les espèces allopolyploïdes qui montrent une vigueur plus importante que leurs espèces progénitrices. Certains gènes importants, en particulier au niveau du développement, pourraient avoir une expression optimale en condition hétérozygote (modèle surdominant). On trouve cet effet hétérosis à l'origine de la sélection des plantes cultivées pour leur intérêt agronomique, qui comportent ainsi de nombreuses espèces polyploïdes relativement récentes (canne à sucre, colza, blé, maïs).

III.3.2 Effets à long terme de la polyploïdie

Les effets à long terme sont observables naturellement chez les espèces où la polyploïdisation est relativement ancienne (>1 Ma) et sont discernables par leur analyse précise et la comparaison avec leurs progéniteurs. Ces espèces sont généralement bien intégrées dans leur environnement et peuvent même se révéler supérieures à leurs espèces progénitrices sur le plan de la vigueur et des capacités d'adaptation, leur permettant d'investir de nouvelles niches écologiques. Ces anciens polyploïdes ont retrouvé un comportement de diploïdes à plusieurs niveaux, en particulier celui de la méiose, et sont à l'origine d'une importante diversification des espèces.

Rediploïdisation

Le plus souvent, l'organisme n'a besoin que d'une seule copie du gène pour garder la fonction originale. L'autre copie va donc être libre d'accumuler des mutations sans pression de sélection. L'évolution des gènes en plusieurs copies a fait l'objet de nombreuses études (Lynch et Force 2000, Lynch et Connery 2000). Le modèle d'évolution résultant donne trois voies d'évolution pour des gènes dupliqués : la **non-fonctionnalisation**, la **néo-fonctionnalisation** et la **sub-fonctionnalisation**.

Dans la plupart des cas, les mutations accumulées aboutissent à la perte de la fonction et l'élimination de la copie dupliquée (**non-fonctionnalisation**). Le choix de la copie éliminée semble aléatoire pendant les premières générations. Dans certains allopolyploïdes, les gènes d'un des génomes parentaux sont souvent préférentiellement conservés (Rapp *et al.* 2009). Dans de rares cas, les mutations sur la copie divergente sont avantageuses et confèrent une nouvelle fonction (**néo-fonctionnalisation**). La **sub-fonctionnalisation** correspond à une évolution plus 'concertée' des copies : elles gardent chacune une partie de la fonction originale ou se diversifient en s'exprimant dans des tissus différents par exemple (Adams *et al.* 2003, Hovav *et al.* 2008). Les origines de ces différentes évolutions peuvent être structurales (délétions ou mutations dans les séquences codantes ou promotrices) mais aussi fonctionnelles et épigénétique (régulation de l'expression).

Il faut ajouter à ce modèle que certains gènes restent conservés en plusieurs copies après des millions d'années d'évolution (Sémon et Wolfe 2007). La présence de deux copies peut avoir un effet avantageux pour l'espèce au niveau de l'expression (dosage de gènes,

fixation de l'hétérozygotie) et un effet tampon pour des gènes importants, susceptibles d'être perdus par mutations/délétions.

Les chromosomes homéologues vont progressivement se différencier par l'élimination plus ou moins rapide de la redondance génomique et/ou par des réarrangements structuraux affectant l'ensemble du génome. Cette différentiation va réduire les chances d'appariements multiples pendant les divisions cellulaires et ainsi stabiliser le polyploïde. Il retourne donc à une organisation diploïde (avec un nombre de chromosomes différent de ses progéniteurs), si bien qu'il est parfois difficile d'identifier d'anciennes polyploïdisations dans une espèce apparemment diploïde.

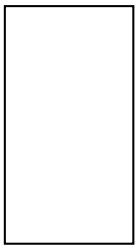
Effet de polyploïdie sur la diversité des espèces

La polyploïdie peut paraître comme un goulet d'étranglement pour la diversité. En effet, elle implique un nombre limité d'individus et les croisements progéniteurs / polyploïdes sont souvent stériles, isolant ces derniers au niveau reproductif. Par contre, on constate souvent une très forte augmentation du nombre d'espèces après un évènement de polyploïdisation dans les arbres phylogénétiques (Soltis *et al*. 2009). La polyploïdie est donc paradoxalement une source de diversification et de spéciation très importante.

Les génomes polyploïdes offrent des possibilités pour des changements génomiques importants et rapides par des modifications structurales et fonctionnelles. La polyploïdie augmente aussi considérablement la réserve génétique et les possibilités d'évolution des gènes (**néo-fonctionalisation**, **sub-fonctionalisation**). Après un événement de polyploïdisation, il va se créer des populations possédant des modifications différentes aboutissant à plus long terme à la formation de nouvelles espèces.

III.4 Polyploïdie et domestication des blés

La culture moderne du blé est essentiellement basée sur deux espèces allopolyploïdes : le blé dur allotétraploïde (2n=4x=28, BBAA) et le blé tendre allohexaploïde (2n=6x=42, BBAADD). La formation de ces espèces de blé a impliqué plusieurs étapes de polyploïdisation et de domestication décrites dans ce chapitre (Figure 12).



0.5 Ma

X

Ae. speltoides

(SS)

(2n=2x=14)

?

(BB)

(2n=2x=14)

T. urartu

(AA)

(2n=2x=14)

Amidonnier sauvage

T. turgidum

ssp *dicoccoides*

(BBAA) **(2n=4x=28)**

Amidonnier

T. turgidum

ssp *dicoccum*

(BBAA)

(2n=4x=28)

0.01 Ma

X

Ae. tauschii

(DD)

(2n=2x=14)

Blé dur

T. turgidum

ssp *durum*

(BBAA)

(2n=4x=28)

Blé tendre

T. aestivum

ssp *aestivum*

(BBAADD)

(2n=6x=42)

Figure 12. Évènements de polyploïdisation ayant abouti à la formation des blés cultivés tétraploïdes et hexaploïdes. Si les progéniteurs des génomes A et D sont clairement identifiés, le donneur du génome B reste pour l'instant introuvable (Kihara 1994, McFaden et Sears 1946, Feldman *et al.* 1995, Nesbitt et Samuel 1996, Blake *et al.* 1999, Feldman 2001, Huang *et al.* 2002)

L'amidonnier sauvage est une espèce de blé tétraploïde *T. turgidum* ssp. *dicoccoides* (2n=4x=28, BBAA) qui s'est formée il y a environ 0,5 Ma par un événement d'allopolyploïdisation entre l'espèce diploïde *T. urartu* (2n=2x=14, A^uA^u), progénitice du génome A et une espèce de la section *Sitopsis*, progénitrice du génome B, qui reste à identifier (Feldman *et al.* 1995, Blake *et al.* 1999, Feldman 2001, Huang *et al.* 2002, Dvorak *et al.* 2006) (Tableau 2). L'espèce la plus proche dans cette section est *Aegilops speltoides* (2n=2x=14, SS).

Il y a environ 10.000-15.000 ans, l'homme a commencé à cueillir puis cultiver et sélectionner les espèces pour correspondre à ses besoins : c'est la domestication. Dans les premières formes cultivées du blé, les grains se détachaient facilement du rachis comme dans les formes sauvages. Ce caractère assurant la dissémination naturelle des espèces du blé causait des pertes de récoltes importantes. La première sélection du blé s'est faite par rapport à ce critère. Les plantes qui gardaient le plus de grains attachés au rachis étaient sélectionnées, aboutissant progressivement aux formes domestiquées. L'engrain *T. monococcum* (2n=2x=14, A^mA^m) et l'amidonnier *T. turgidum* spp. *dicoccum* (2n=4x=28, BBAA) sont ainsi les premières espèces domestiquées du blé. Elles viennent respectivement de l'engrain sauvage *T. boeoticum* (2n=2x=14, A^mA^m) et de l'amidonnier sauvage (BBAA).

Ces formes domestiquées vont être disséminées progressivement à travers l'Asie, l'Europe et l'Afrique. L'amidonnier (BBAA) sera ainsi mis en contact avec l'espèce diploïde sauvage *Aegilops tauschii* (2n=2x=14, DD) dans la région géographique allant de l'Arménie à la côte sud-ouest de la mer Caspienne (Feldman *et al.* 1995, Dvorak *et al.* 1998, Dubcowsky et Dvorak 2007). Leur hybridation et doublement chromosomique a donné le blé tendre allohexaploïde actuellement cultivé *T. aestivum* ssp. *aestivum* (2n=6x=42, BBAADD) (Kihara 1944, McFaden et Sears 1946, Feldman *et al.* 1995, Nesbitt et Samuel 1996). Les premiers restes archéologiques du blé hexaploïde datent d'environ 10.000 ans (Willcox 1996). Il s'est vite disséminé et adapté aux climats plus froids. Il occupe actuellement plus de 95% des surfaces cultivées de blé. Cette allopolyploïdisation, favorisée par l'homme au cours de ses premiers pas dans l'agriculture, constitue donc un événement fondateur majeur des civilisations humaines.

Le blé dur actuellement cultivé *T. turgidum* ssp. *durum* (2n=4x=28, BBAA) apparaît avec l'empire romain il y a environ 3000-4000 ans. Son origine et sa relation avec les premières espèces domestiquées tétraploïdes du blé (*T. turgidum* spp. *dicoccum*) restent mal caractérisées (M. Feldman, communications personnelles).

IV Contexte et objectifs de la thèse

Ma thèse s'est déroulée à l'URGV (Unité de Recherche en Génomique Végétale) dans l'équipe OEPG (Organisation et Evolution des Génomes des Plantes) sous la direction de Boulos Chalhoub. Les thématiques de l'équipe sont l'analyse de l'organisation et de l'évolution des génomes du blé et des *Brassica* en relation avec la polyploïdie et les éléments transposables, utilisant notamment des approches de génomique comparée. Mes travaux de thèse s'inscrivent donc complètement dans cette thématique.

Le blé a une importance économique de tout premier plan à l'échelle mondiale. Son génome se distingue par sa richesse en éléments transposables (>80% du génome) et des évènements de polyploïdisation récurrents et récents. La question centrale de ma thèse porte sur l'évolution et l'organisation des génomes du blé dans ce contexte.

Les tailles importantes de ces génomes ont jusque là limité et conditionné les approches génomiques adéquates à leur étude. Mon travail de thèse a pu bénéficier d'un projet de séquençage comparatif (APCNS2003 : http://www.cns.fr/spip/Triticum-ssp-comparative-genome.html), explorateur de plusieurs régions d'intérêt dans différentes espèces du blé, que mon laboratoire d'accueil a entrepris en collaboration avec l'Institut de Génomique (Centre National de Séquençage - CNS). Par ailleurs, grâce à la participation du laboratoire à des consortiums internationaux, j'ai bénéficié des initiatives de séquençage des génomes d'autres espèces de *Poaceae* comme *Brachypodium* (The International Brachypodium Initiative 2010) et le sorgho (Paterson *et al.* 2009). Ceci m'a permis, non seulement une participation modeste à ces consortiums, mais surtout de pouvoir élargir mes approches de génomique comparative afin d'apprécier l'évolution du génome du blé par rapport à d'autres espèces de *Poaceae*.

La plupart des travaux que j'ai réalisés repose donc sur des démarches d'analyse de séquences de régions précises ou représentatives des génomes du blé et de génomique comparative entre des régions orthologues dans des génotypes et des espèces plus ou moins éloignées (même famille, même genre ou même espèce).

Après une description de la méthode d'annotation mise au point et utilisée pour l'ensemble des séquences génomiques de ma thèse, je présente, dans une première partie, mes travaux sur l'analyse de la dynamique et de la prolifération des TEs dans les génomes A et B

du blé. Au début de ma thèse, les séquences génomiques disponibles pour les génomes du blé étaient réduites à quelques clones BAC (revu par Sabot *et al.* 2005, Stein 2007 et disponibles sur http://genome.jouy.inra.fr/triannot/index.php et http://www.ncbi.nlm.nih.gov/). Elles avaient néanmoins permis de confirmer la richesse des génomes du blé en éléments transposables (Smith et Flavell 1975, Vedel et Delseny 1987) et d'en identifier les principaux types (Wicker *et al.* 2002, Sabot *et al.* 2005). La prolifération des différents TEs dans les génomes du blé, leur contribution aux variations des tailles des génomes ainsi que leur distribution le long des chromosomes n'étaient donc pas encore explorés. En analysant toutes les séquences génomiques disponibles et en réalisant du séquençage complémentaire représentatif, j'ai pu analyser pour la première fois la dynamique et la prolifération différentielle des TEs dans les génomes A et B du blé. Ces travaux ont été publiés en 2008 dans la revue *Genetics* et sont présentés en partie I des résultats.

La prolifération des TEs est la résultante de deux forces d'évolution antagonistes : leur activité insertionnelle et leur élimination. La première partie de mon travail de thèse m'a permis de caractériser l'activité insertionnelle des TEs dans les génomes A et B du blé grâce à l'analyse de séquences représentatives de ces génomes. L'élimination des TEs est plus difficile à caractériser, puisqu'il faut non seulement la séquence présentant une délétion et d'une séquence orthologue ne la présentant pas. La caractérisation de la force d'élimination est d'autant plus précise que l'on dispose d'un nombre important de séquences orthologues à comparer. Le locus de la dureté de la graine *Ha* (Hardness locus) est certainement l'un des plus séquencés dans différentes espèces du blé. L'étude de la dynamique de ce locus est un des travaux pionniers de mon équipe (Chantret *et al.* 2005). Ils ont montré que son absence dans les génomes A et B des blés durs et tendres était due à sa délétion par une recombinaison illégitime impliquant des TEs. Ce locus constituait donc un cadre idéal pour étudier les éliminations des TEs dans un locus. Dans le cadre du projet APCNS2003, les séquences de clone BACs portant ce locus dans de nouvelles espèces sont venues s'ajouter à celles déjà disponibles. Au total, l'analyse que j'ai menée a porté sur pas moins de 16 haplotypes différents des génomes A, B, S et D. La variabilité haplotypique de ce locus et la dynamique des TEs dans le blé par des comparaisons aux niveaux inter-spécifique (inter-génomique et intra-génomique) et intra-spécifique ont été ainsi étudiés pour la première fois. Nous avons ainsi révélé d'importantes différences entre les génomes, présentées dans la partie II des résultats, et qui feront l'objet d'une publication dans un futur proche.

A une échelle plus large (>4 Ma), j'ai aussi étudié l'évolution de la région couvrant le locus *Ha* en comparant le locus entre les différentes espèces de *Poaceae* pour lesquelles les

séquences génomiques étaient disponibles. A cette échelle d'évolution, la génomique comparée concerne uniquement les gènes puisque les éléments transposables ne sont pas conservés. Les gènes du locus *Ha* ne sont pas présents dans les *Panicoideae* (sorgho et maïs) alors qu'une étude de génomique comparée a montré l'existence d'un petit segment d'un de ces gènes dans le riz (Chantret *et al.* 2004). En réalisant du séquençage génomique complémentaire dans d'autres espèces de *Poaceae*, comme *Brachypodium sylvaticum* mais aussi en analysant les régions orthologues d'autres espèces entièrement séquencées (riz, *Brachypodium distachyon* et sorgho), j'ai essayé d'élucider la difficile question de l'émergence de ce locus au cours de l'évolution des *Poaceae*. Les résultats de ces travaux, que j'ai publiés dans la revue Molecular Biology and Evolution en 2009, sont présentés en partie III des résultats.

Les publications auxquelles j'ai contribué dans le cadre de projets de séquençage comparatif (Gu *et al.* 2006, Salse *et al.* 2008b) et de génome complet (The International Brachypodium Initiative 2010) sont présentées en Annexe 3, 4 et 5.

Matériels et méthodes d'annotation de séquences génomiques

I Introduction

Une large partie de mon travail de thèse repose sur l'annotation de séquences génomiques. La précision dans l'annotation de ces séquences est déterminante pour comprendre leur évolution, particulièrement dans un contexte de génomique comparative. Pour identifier et caractériser les gènes et les TEs constitutifs de la séquence, deux approches sont utilisées de manière complémentaire : la recherche par similarité qui se base sur des séquences déjà référencées et la recherche par structure qui se base sur certaines caractéristiques 'physiques' propres aux séquences que l'on cherche à identifier.

Au début de ma thèse, l'annotation des éléments transposables se faisait en grande partie manuellement dans le laboratoire et constituait une étape limitante pour l'annotation rapide de séquences génomiques, surtout pour les génomes riches en TEs comme celui du blé. Le manque d'outils adaptés et l'important volume de séquences à annoter pour les études de génomique comparée m'ont poussé à améliorer cette partie de l'annotation par la conception d'un programme automatisé d'aide à l'annotation des TEs.

Depuis, de nombreux outils performants d'annotation sont disponibles pour l'annotation des gènes [FGENESH http://www.softberry.com, Eugene (Shiex *et al.* 2001)] et d'éléments répétés de génomes complets [Recon (Bao et Eddy 2002), Blaster et Matcher (Bergman et Quesneville 2007)]. Cependant, en ce qui concerne l'annotation des TEs du blé, qui n'est pas un génome séquencé, je n'ai pas trouvé de programme plus adapté que celui que j'ai conçu. En effet, il repose sur des paramètres spécialement optimisés pour l'analyse des TEs du blé. Ce programme devrait s'intégrer au pipeline d'annotation TriAnnot développé par l'équipe de Philippe Leroy à Clermont-Ferrand et qui est disponible sur le site http://urgi.versailles.inra.fr/projects/TriAnnot/.

Je vais maintenant préciser les différentes étapes que j'ai suivies pour annoter les séquences génomiques

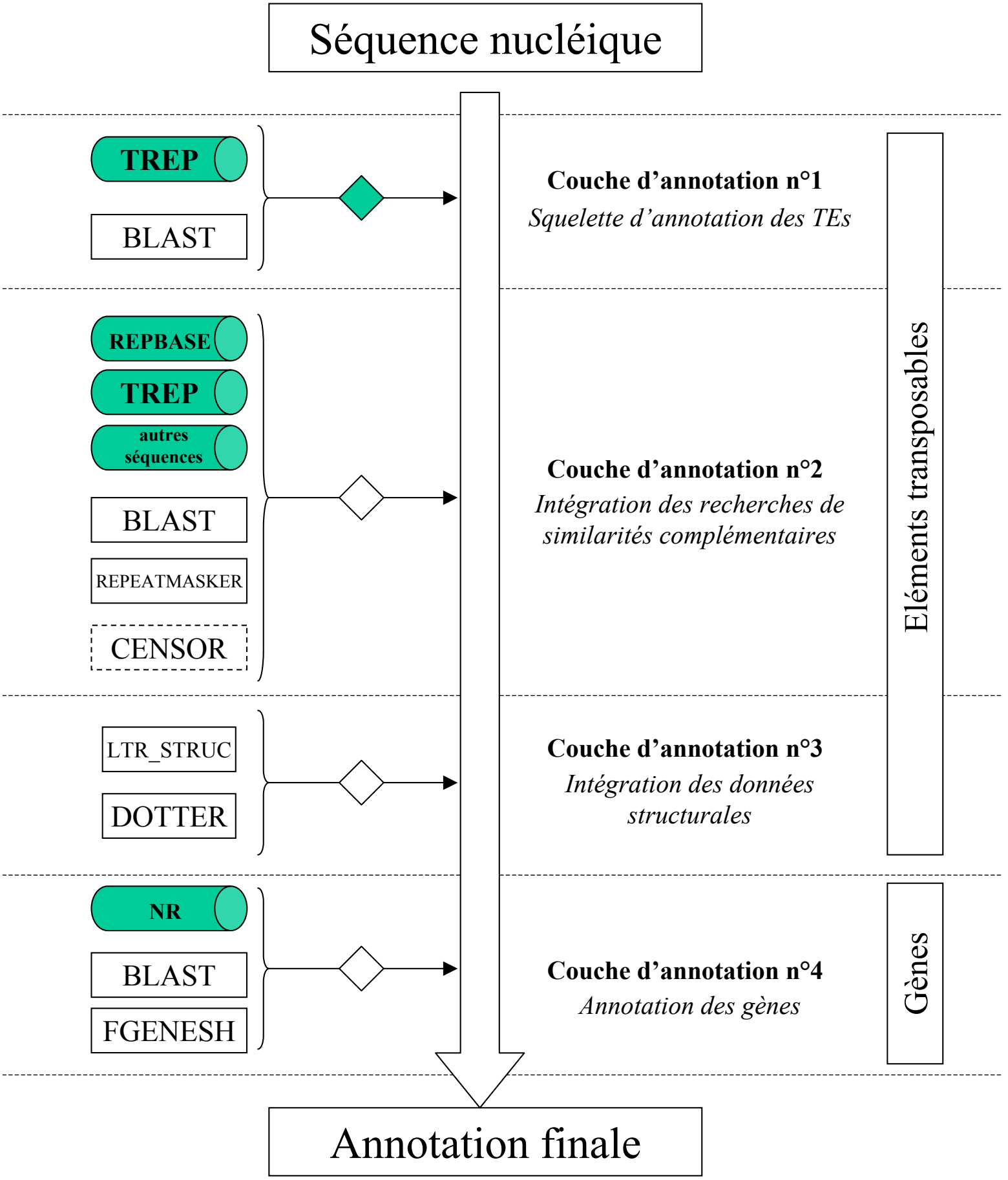


Figure 13. Procédure d'annotation par couches successives. Chaque couche ajoute de l'information et précise l'annotation. Les cylindres verts représentent les bases de données, les rectangles représentent les logiciels. Un diamant vert correspond à une étape automatisée et les diamants blancs à des étapes expertisées.

II Annotation des séquences génomiques

Le programme que j'ai développé est à la base d'une procédure d'annotation par couches successives, optimisée pour les séquences du blé. Chaque couche rajoute des informations et précise l'annotation (Figure 13). Les éléments transposables, constituant souvent plus de 80% des séquences de blé, sont annotés en premier (couches 1 à 3). Cela permet de limiter les fausses prédictions de gènes correspondant en réalité à des portions codantes de TEs qui ne sont pas encore référencés. On recherche ensuite les gènes en dehors de l'espace TE dans la dernière couche d'annotation (couche 4).

II.1 Couche n°1 : le programme

L'objectif du programme est de faciliter l'annotation, tout en prenant en compte certaines particularités des génomes riches en TEs :

- Les TEs référencés dans la base TREP peuvent être très proches (par exemple s'ils appartiennent à la même famille), générant lors d'une recherche par similarité une multitude de résultats.
- Les TEs divergent rapidement après leur insertion, en particulier par indels, ce qui complique leur détection par similarité en fragmentant les résultats.
- Les TEs sont souvent imbriqués les uns dans les autres ('nested insertions'), formant de larges régions d'éléments transposables. Ces régions contiennent donc des fragments d'un même élément qui peuvent être éloignés de dizaines voire des centaines de kb, ne facilitant pas leur reconstruction.

Il est basé sur des recherches par similarités, en comparant la séquence à annoter avec la séquence d'éléments déjà répertoriés dans des Bases de données (BdD). J'ai donc commencé par choisir une Base de Données (BdD) de référence et un programme de recherche de similarité qui, utilisé avec des paramètres adaptés, assurerait une détection optimale des TEs. Mon programme va intégrer l'ensemble de ces résultats, en tenant compte des difficultés liées aux TEs évoquées précédemment, pour fournir différents fichiers 'sorties' constituant la base de l'annotation.

II.1.1 Choix de la base de données de référence

Plusieurs bases de données recensent des TEs de différentes espèces :

- REPBASEupdate (Jurka *et al.* 2000, 2005) contient des consensus (ou des éléments pour les petites familles) venant de l'ensemble des eucaryotes.

http://www.girinst.org/repbase/index.html

- Triticeae Repeat Sequence Database (TREP, Wicker *et al.* 2002) recense les TEs trouvés dans des *Triticeae* (blé et orge principalement).

http://wheat.pw.usda.gov/ITMI/Repeats/

- TIGR Plant Repeat Databases (Ouyang et Buell 2004) sont des bases de données contenant des éléments répétés de différentes familles de plantes.

http://plantrepeats.plantbiology.msu.edu/

La grande majorité des séquences à annoter pour mes travaux provenant du blé, le choix de TREP comme base de référence était particulièrement indiqué. Cependant, les résultats obtenus sur REPBASEupdate et sur les bases du TIGR n'ont pas été ignorés : ils sont utilisés en complément pour la couche d'annotation n°2.

II.1.2 Choix des programmes de recherche par similarité et leur paramétrage

Les programmes de recherche par similarité vont comparer la séquence étudiée aux séquences des éléments présents dans des BdD. L'efficacité de la recherche par similarités dépend donc directement des BdD (exhaustivité, qualité d'annotation) et de la pertinence des programmes de recherche (sensibilité, spécificité).

Plusieurs programmes sont couramment utilisés pour la recherche par similarité entre séquences :

- BLAST (Altschul *et al.* 1990, 1997)
- CENSOR (Jurka *et al.* 1996, 2005)
- REPEATMASKER (http://www.repeatmasker.org/)

On évalue les programmes de recherche de similarités par rapport à deux valeurs : la spécificité et la sensibilité. Pour illustrer ces valeurs, prenons un exemple. On dispose de 10 séquences de référence dont six correspondent à des TEs. Un programme de recherche par similarité trouve que cinq de ces séquences correspondent à des TEs. En réalité, seulement

quatre de ces cinq séquences sont effectivement des TEs et les deux autres séquences de TEs n'ont pas été détectées.

Cela nous donne :

- vrai positif (VP) : 4 (detecté et TE)
- faux positif (FP) : 1 (detecté mais pas TE)
- vrai négatif (VN) : 3 (non detecté et pas TE)
- faux négatif (FN) : 2 (non detecté mais TE)

De façon simplifiée, la sensibilité représente le taux de détection des valides VP/(VP+FN) = 4/6 = 67% et la spécificité évalue la fiabilité de la détection VN/(VN+FP) = 3/4 = 75%.

Les programmes de recherche de similarités sont généralement paramétrables par l'utilisateur, permettant de modifier dans une certaine mesure leur sensibilité et leur spécificité. Avec les paramètres par défaut (voir ci-après), BLAST est le plus spécifique et le moins sensible, CENSOR est le plus sensible et le moins spécifique et REPEATMASKER est un compromis entre les deux.

Cette première couche d'annotation repose sur la construction d'un 'squelette' d'annotation sur lequel vont venir se greffer des couches d'annotation supplémentaires. Pour cette étape, j'ai privilégié la spécificité pour obtenir un squelette 'robuste' (avec le moins d'erreurs possibles), les couches suivantes se chargeant de compléter l'annotation. Après plusieurs essais, nous avons retenu le programme BLAST avec des paramètres ajustés pour augmenter sa sensibilité aux TEs :

- q -2 (pénalité de 'mismatch' de 2, 3 par défaut)
- F F (pas de filtrage des séquences répétées).

Pour tenir compte de la fragmentation des TEs, nous avons autorisé une plus grande tolérance aux gaps avec les réglages :

- G 3 (3 comme pénalité d'ouverture de gap, 5 par défaut)
- E 1 (1 comme pénalité d'extension de gap, 3 par défaut)
- X 200 (regroupement des matchs séparés par moins de 200 pb)

Avec ces paramètres, BLAST détecte des niveaux d'identité de séquences compris entre 80% et 100% ce qui est particulièrement adapté pour la recherche de TEs.

Les résultats de REPEATMASKER et de CENSOR, obtenus avec les paramètres par défaut, sont utilisés dans la couche d'annotations n°2. Pour un gain de temps substantiel, on peut se passer du programme CENSOR : la grande majorité des résultats spécifiques à CENSOR sont le plus souvent des faux positifs.

II.1.3 Création du programme d'annotation

Je me suis servi de mes compétences en informatique et de mon expérience d'annotateur pour mettre au point un programme automatisant au maximum l'identification des TEs par similarité (avec les paramètres décrits précédemment), et gardant une qualité optimum d'annotation. Les fichiers d'entrée du programme sont donc des résultats de BLAST de la séquence sur la BdD TREP qui vont être transformés, de façon automatique, en plusieurs fichiers de sorties correspondant à une première annotation plus ou moins détaillée suivant le fichier. Cette annotation est disponible sous la forme de plusieurs fichiers qui correspondent à différents degrés d'expertise automatique au format du logiciel Artemis (Rutherford *et al.* 2000).

Le principe de base du programme est de regrouper et d'intégrer les informations de différents résultats de BLAST de façon automatique. Ces regroupements transforment une multitude de résultats en une seule entité plus facile à utiliser et à analyser. Ils sont réalisés par des sous-programmes d'assemblage constituant la base de ce programme.

Le programme se déroule en un ensemble d'étapes que je vais maintenant détailler. Son fonctionnement dépend de nombreux paramètres. Afin de garder une certaine flexibilité, les paramètres indiqués **en gras** dans le texte sont facilement modifiables pour adapter le programme à différents types de séquences et de contraintes.

Transformation du fichier d'entrée

Le fichier de résultats de BLAST, utilisé comme entrée, contient tous les 'matchs' ou HSPs (Hit Scoring Pair) trouvés et les alignements correspondants, associant des coordonnées de la séquence analysée (*query*) avec des coordonnées de séquences trouvées dans les bases de données (*subject*).

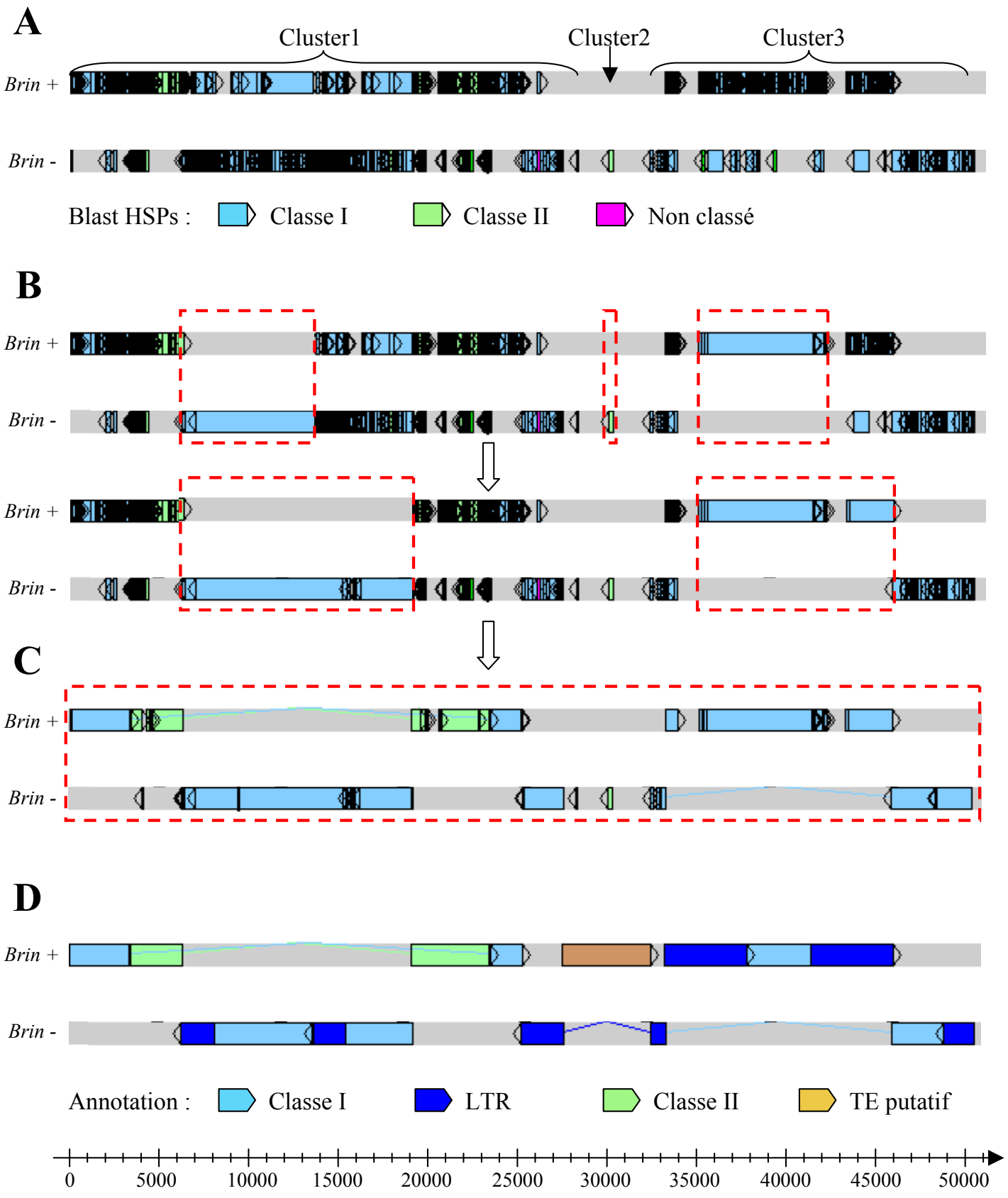
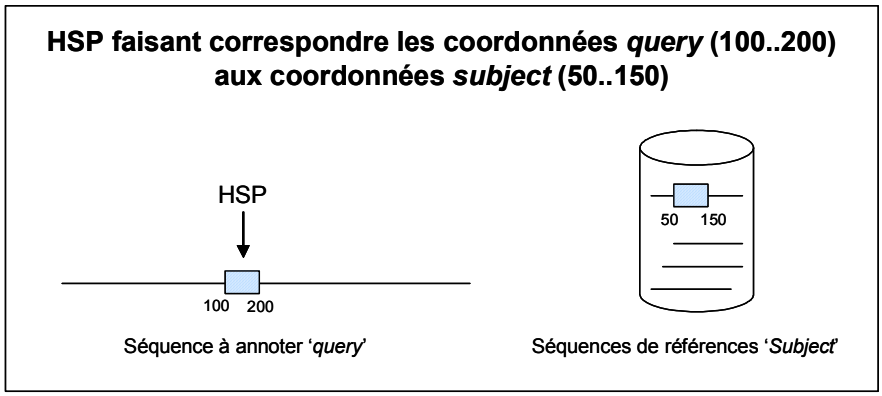


Figure 14. Exemple réel des différentes étapes d'annotation, visualisées dans le logiciel Artemis. (A) Clusterisation des HSPs de BLAST. (B) Les deux premières itérations de la procédure de création et de selection des regroupements. Les rectangles rouges représentent les séquences entre les bornes 'début' et 'fin'. Les meilleurs regroupements (identité et longueur) sont identifiés. (C) Résultat donné par le programme d'annotation à la fin de la procédure de création et de selection des regroupements. (D) Résultat final de l'annotation de la séquence après intégration de l'ensemble des données et expertise manuelle.

Ce fichier est transformé en une version condensée contenant, pour chaque HSP, uniquement les informations utilisées pour l'annotation : les coordonnées de début et de fin des couples *query/subject*, la longueur de l'alignement associé, le nombre de bases/acides aminés effectivement alignés (donc hors gap), pourcentage d'identité/similarité, le score et l'e-value. Un fichier sortie est généré (appelé <*TEbase*>) à cette étape : il contient l'ensemble des HSPs de BLAST associés aux informations précédentes.



Clusterisation des HSPs

Les HSPs sont ensuite répartis en clusters : tous les HSPs distants de moins de **1000** bp sont regroupés (Figure 14A). Cette étape a été conçue pour refléter la tendance des TEs à s'imbriquer les uns dans les autres, formant des clusters de TEs entourant des gènes. Elle a aussi l'avantage d'accélérer certaines phases de calcul.

Regroupement des HSPs

Pour faire les regroupements, les HSPs 'compatibles' d'un même cluster sont assemblés. Pour expliquer la compatibilité entre HSPs, prenons un exemple. Une séquence de référence (*query*) à deux HSPs (HSP1, HSP2) se référant au même TE (*TE*) d'une certaine longueur (*length*). Ces HSPs associent deux parties de la séquence *query* (query_debut1..query1_fin) et (query_debut2..query_fin2) à deux parties de la séquence de *TE* (TE_debut1..TE_fin1) et (TE_debut2..TE_fin2). Un certain nombre de conditions doivent être remplies pour les considérer comme compatibles :

(1) Les deux HSPs ont le même TE de référence.

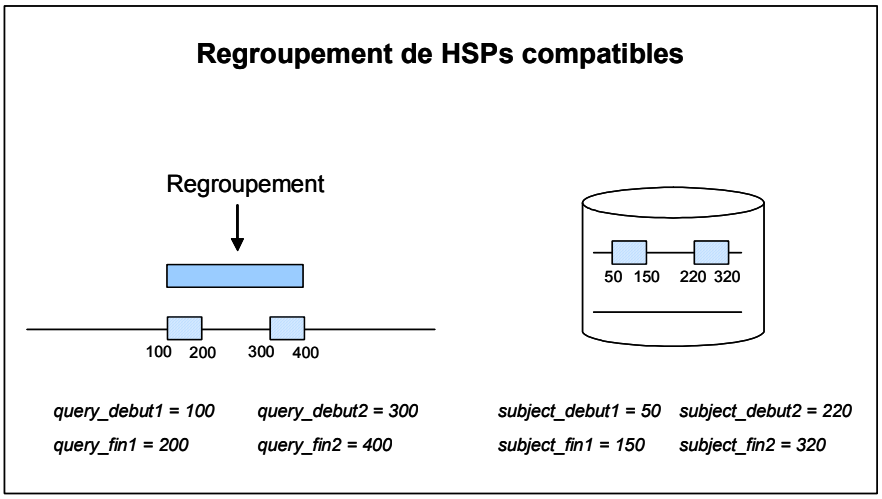
(2) Les deux HSPs doivent être dans la même orientation

(3) La distance entre les HSPs doit rester dans les limites de la taille de l'élément de référence (**+20%**) : query_fin2 – query_debut1 < (length + 20%).

(4) On n'accepte pas d'indels de plus de **300** paires de bases : query_debut2 - query_fin1 < 300 et subject_debut2 - subject_fin1 < 300.

(5) Les distances entre les deux HSPs sur la séquence de référence et sur le TE de référence doivent être proches (**+/-20%**) : 0,8<(query_fin2-query_debut1) / (TE_fin2 - TE_debut1)<1,2.

Pour chaque HSP, on va chercher d'autres HSPs compatibles dans son cluster en les testant de proche en proche. Tous les HSPs compatibles sont regroupés ensemble.



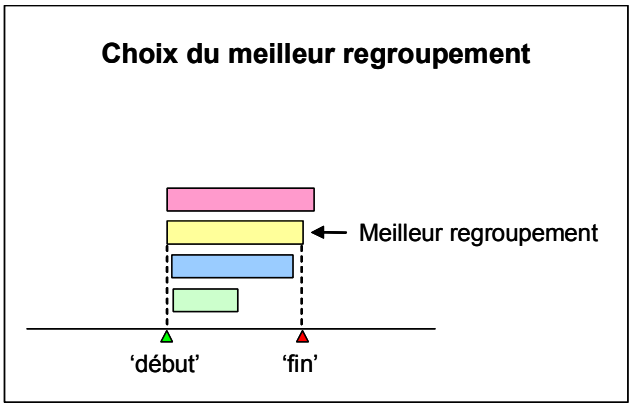
Rappelons que ces regroupements représentent une partie continue d'un TE dans la séquence. On leur associe des coordonnées de début et de fin sur la séquence et sur le TE de référence (coordonnées de début du premier HSP et de fin du dernier HSP de l'assemblage), une longueur d'alignement (somme des longueurs des alignements des HSPs) et une longueur effective (somme des paires de bases/acides aminés réellement alignés dans les HSPs).

Après cette étape, on va progressivement réduire le nombre de regroupements considérés en ne conservant que les plus pertinents pour l'annotation. Cependant, l'ensemble des informations brutes reste disponible dans le premier fichier de sortie <*TEbase*>.

Choix du meilleur regroupement d'un cluster

Dans chaque cluster, le programme choisit le regroupement le mieux conservé (longueur effective) parmi les assemblages les plus longs (longueur d'alignement **>90%** de celle du plus long regroupement). Cet assemblage va servir de point de départ pour l'annotation du reste du cluster. Cette procédure reflète le travail d'annotation d'éléments transposables imbriqués les uns dans les autres : on commence par annoter les éléments les

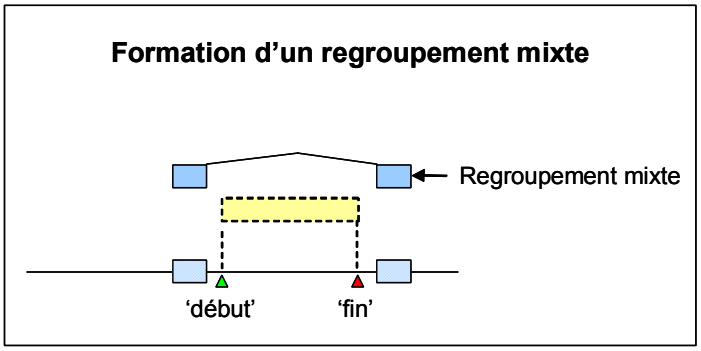
plus récents (les mieux conservés et les plus longs) puis on progresse vers les éléments les plus anciens. Cet élément le plus conservé va définir des bornes 'début' et 'fin' correspondant à ses coordonnées de début et de fin sur la séquence analysée (Figure 14B).



Tous les regroupements situés entre 'début' et 'fin' (**+/- 100 pb**) et qui ont une longueur d'alignement effective suffisante (**>50%** de l'alignement du meilleur assemblage) sont conservés et stockés dans le fichiers de sortie <*TE*>. Les HSPs correspondant sont disponibles dans le fichier de sortie <*TEdetail*> sous une forme jointe mais non regroupée.

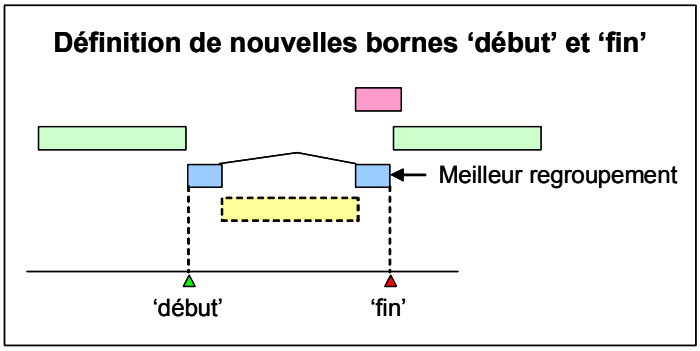
Identification des regroupements de façon récursive.

Le programme va de nouveau lancer la procédure pour faire des regroupements dans chaque cluster, mais en ne considérant que les HSPs en dehors des bornes 'début' et 'fin'. Pour cette étape, tout se passe comme si la séquence entre les deux bornes 'n'existait pas'. Auparavant, certains HSPs, situés de part et d'autre des bornes 'début' et 'fin', étaient trop éloignés pour pouvoir être regroupés. En considérant que la séquence entre les bornes n'existe pas, ils peuvent être assemblés et former des regroupements 'mixtes'.



Une fois les regroupements constitués, le programme va chercher le meilleur regroupement (même critères que précédemment) parmi les plus proches des bornes ‘début’ et ‘fin’.

Si le meilleur regroupement est en amont de la borne ‘début’, sa coordonnée de départ va définir une nouvelle borne ‘début’, s’il est en aval sa coordonnée de fin va définir une nouvelle borne ‘fin’. Un regroupement mixte va définir des nouvelles bornes ‘début’ et ‘fin’.



Tous les regroupements situés entre ces nouvelles bornes et d’une longueur suffisante (>50% de l’alignement du meilleur regroupement) sont conservés dans le fichier <TE>, et leurs HSPs correspondant sont conservés dans <*TEdetail*>.

On procède ainsi par itérations successives, en décalant les bornes ‘début’ et ‘fin’ jusqu'à couvrir tout le cluster (Figure 14B, 14C).

II.2 Couche n°2 : analyses de similarités complémentaires

Pour cette couche d’annotation, j’utilise l’ensemble des autres recherches par similarité (Figure13) pour compléter les résultats du programme d’annotation. Cette étape se fait de façon manuelle, en ajustant éventuellement les coordonnées des TEs de l’annotation de base ou en ajoutant des éléments (non détectés par le BLAST sur TREP) dans Artemis.

En plus des résultats obtenus par les recherches par similarité sur les banques de TEs, on va également comparer la séquence à l’ensemble des séquences non-gènes de la même espèce (ou d’espèces proches) à notre disposition. Cette comparaison peut faire apparaître des séquences répétées qu’une analyse complémentaire (structurale) identifiera ou non comme étant des TEs.

II.3 Couche n°3 : analyses structurales

A cette étape du processus d'annotation, on a terminé l'identification d'éléments transposables par comparaisons avec des éléments déjà connus. La couche d'annotation n°3 s'intéresse aux caractéristiques structurales des TEs. En effet, les TEs de classe I à LTRs (>50% du génome) et de classes II de la superfamille *CACTA* (>20% du génome) ont de nombreuses caractéristiques structurales facilement identifiables par la méthode DOT-PLOT (Maizel et Lenk 1981).

Les LTRs des rétrotransposons forment deux longues répétitions au début et à la fin de l'élément très visibles par DOT-PLOT. De plus, ces LTRs commencent/finissent par le dinucléotide TG/CA et sont encadrés par une répétition directe de 5 pb (TSD : Target Site Duplication) causée par leur insertion.

De la même façon, les éléments *CACTA* (16% des génomes du blé) commencent/finissent par le motif CACTA/TAGTG et ont des TSD de 3 pb. Ce motif précède/suit une zone répétée inversée en tandem (TIRs : Tandem Inversed Repeat) très caractéristique.

Nous avons utilisé les logiciels comme DOTTER (Sonnhammer and Durbin 1995) ou LTR_STRUC (McCarthy *et al.* 2003) pour les analyses structurales

La prise en compte de ces éléments structuraux aide à caractériser et à compléter l'annotation des éléments obtenue par la recherche de similarité. Elle permet aussi d'identifier des éléments qui n'ont pas encore été décrits, uniquement grâce à leur signature structurale. Cela permet aussi de mettre parfois en évidence des incohérences entre les informations de structure et de similarité, soulignant le stade encore jeune des bases de données, incluant un certain nombre d'erreurs.

La combinaison des approches par similarité et structure permet une grande sensibilité dans la détection des TEs des *Triticeae*. L'annotation délimite ainsi de larges zones remplies de TEs. On trouve parfois des 'trous' non annotés au milieu de ces clusters de TEs. Ces zones (séquences d'ADN non assignées) correspondent le plus souvent à des éléments transposables qui n'ont pas encore été décrits et sont une source importante de détection de nouveaux éléments complets possédant ou non des caractéristiques structurales identifiables.

Regrouper l'ensemble des trois couches d'annotations des TEs en une version finale reste une étape très délicate, basée principalement sur l'expérience de l'annotateur. En effet, il n'est pas rare d'avoir plusieurs annotations possibles, mais l'expérience permet de garder la plus probable. Quand il est difficile de départager deux annotations possibles, on laisse les deux possibilités, afin de ne pas induire en erreur des personnes réutilisant ces informations. C'est aussi pour cette raison que le programme d'aide à l'annotation que j'ai développé automatise certaines étapes, mais laisse certaines prises de décision importantes à l'annotateur.

II.4 Couche n°4 : prédictions et annotation des gènes

Dans cette dernière couche, on recherche les gènes dans toutes les zones non-TE. On se sert également d'une recherche par similarité et par structure. La prédiction de gène ab-initio est faite par FGENESH (http://www.softberry.com) paramétré pour les monocotylédones. C'est un programme très puissant qui détecte plus de 90% des gènes dans les séquences de blé et prédit exactement la structure des introns et des exons dans plus de 50% des cas. On utilise en complément une recherche de similarité avec différents algorithmes de BLAST (BLASTn, BLASTx, tBLASTx) contre des banques de protéines (UNIPROT) ou d'ESTs (dbEST) mais aussi une recherche de synténie avec les séquences d'espèces proches disponibles (riz, sorgho, *Brachypodium*). Nous avons annoté la plupart des séquences dans un contexte de génomique comparée entre espèces suffisamment proches pour avoir leurs gènes très bien conservés et suffisamment lointaines pour que les TEs ne le soient pas. Ainsi, la détection de séquences conservées entre plusieurs espèces indique généralement la présence d'un gène, même si ce ne sont pas les seules séquences conservées. Cela ajouté à la puissance de FGENESH rend la détection de gènes relativement aisée. Il reste à faire l'annotation des gènes, consistant principalement au bon positionnement des sites donneur /accepteur des introns (consensus GT / AG) pour bien délimiter la structure introns/exons.

Nous avons utilisé sept catégories pour définir les gènes : quatre catégories pour des gènes potentiellement fonctionnels (pas de codon stop, pas de décalage du cadre de lecture), trois catégories pour les gènes qui ne le sont plus.

Les gènes potentiellement fonctionnels doivent être au moins partiellement prédits par FGENESH. Ensuite, selon leurs 'matchs' avec des banques de protéines / d'ESTs, ou leur conservation avec une espèce proche (synténie), on les classifie différemment. Les gènes

ayant des ‘matchs’ sur les banques de protéines ou par synténie sont classés en **gène putatif**, les gènes ayant des ‘matchs’ sur les banques d’ESTs sont classés en **gène de fonction inconnue**, les gènes ayant les deux types de ‘matchs’ précédents sont classés en **gène de fonction connue** et l’absence des deux types de ‘matchs’ est classé en **gène hypothétique**.

Si un gène contient des codons stop ou des décalages du cadre de lecture, il est classé comme **pseudogène**. Lorsqu’une partie d’un gène est perdu à la suite d’une délétion ou d’une insertion de TEs, on le classe en **gène tronqué**. Et enfin un gène tronqué, avec une séquence très divergente (introduction de codon stop, de décalage du cadre de lecture) est dit **gène relique**.

Résultats

Partie I : Dynamique et prolifération différentielle des éléments transposables dans les génomes A et B du blé.

Partie II : Caractérisation de l'élimination active des TEs dans les génomes du blé : analyse de variabilité haplotypique inter- et intra-génomique..

Partie III : Évolution du caractère 'grain tendre' dans les *Poaceae* au cours des 60 derniers Ma : émergence dans l'ancêtre commun des *Erhrartoideae* et des *Pooideae*, après leur divergence avec les *Panicoideae*

Partie I : Dynamique et prolifération différentielle des TEs dans les génomes A et B du blé

I Introduction

Les espèces de la famille des *Poaceae* ont tous en commun au moins un évènement ancien de polyploïdisation (Adams et Wendel 2005, Salse *et al.* 2008a). Pourtant, la taille de leur génome haploïde est très variable selon les espèces (Figure 2). Ainsi, les génomes du blé, de l'orge (tribu des *Triticeae*) et du maïs (*Panicoideae*) ont un grand génome (>2 Gb) (Bennett et Smith 1991) tandis que d'autres comme le riz (*Ehrhartoideae*) et *Brachypodium* (*Brachypodieae*) ont comparativement un petit génome (<500 Mb) (Sasaki *et al.* 2005, The International Brachypodium Initiative 2010). Le génome du sorgho, espèce appartenant à la même tribu que le maïs (*Panicoideae*), présente une taille intermédiaire (790 Mb) (Paterson *et al.* 2009). Ces différences de taille entre les génomes des *Poaceae* s'expliquent principalement par leurs différentes teneurs en TEs (Figure 2). Pour les espèces du blé (genre *Triticum* et *Aegilops*), une prolifération importante des TEs est évidente : ils représentent plus de 80% de la séquence (Paux *et al.* 2006, Charles *et al.* 2008 dans cette partie).

Les génomes des espèces diploïdes du blé présentent également d'importantes variations de taille. On observe ainsi des variations pouvant atteindre des centaines de Mb (Bennett et Smith 1976, 1991, http://data.kew.org/cvalues/homepage.html). Par exemple, la taille du génome de *T. monococcum* (6,23 picogrammes - pg) est 1,3 pg plus grande que celle de *T. urartu* (4,93 pg) (Bennett et Smith 1976, 1991), alors que ces deux espèces ont divergé il y a moins de 1,5 Ma (Dvorak *et al.* 1993, Huang *et al.* 2002, Wicker *et al.* 2003b).

Une caractérisation de la composition en éléments transposables, leur dynamique ainsi que leur prolifération dans les différents génomes du blé est alors intéressante.

L'importance de leur taille et le manque d'outils adaptés ne permettaient pas d'envisager le séquençage d'un ou plusieurs génomes complets du blé. L'arrivée des séquenceurs haut débit de nouvelle génération, comme le 454 (Margulies *et al.* 2005), permettront peut être de disposer de ces séquences dans quelques années. En attendant, les efforts se sont concentrés sur le séquençage de régions ciblées, sélectionnées comme couvrant un ou plusieurs gènes d'intérêt, appelé locus. Quelques clones BAC couvrant la région correspondante dans un ou plusieurs génomes du blé ont été ainsi séquencés. Les séquences disponibles pendant la réalisation des mes travaux de thèse couvraient une dizaine de locus (Tableau I-1).

Locus	Publications associées
Ha	Chantret *et al.* 2004, 2005, 2008
Vrn1, Vrn2	Yan *et al.* 2002, 2003, 2004
HMW-Glu	Gu *et al.* 2004, 2006, Kong *et al.* 2004
LMW-Glu	Wicker *et al.* 2003, Gao *et al.* 2007
Lr10	Isidore *et al.* 2005
Lr34	Bossolini *et al.* 2007, Wicker *et al.* 2009
Xpsr920	Dvorak *et al.* 2006
Ph1	Griffith *et al.* 2006
Q	Faris *et al.* 2008
SPA	Salse *et al.* 2008
Acc	Chalupska *et al.* 2008

Tableau I-1 Locus étudiés avant et pendant ma thèse avec le séquençage de clones BAC de blé et les publications associées

Si la quantité des séquences génomiques disponibles était alors limitée, les premiers efforts de séquençage et d'annotation (revu par Sabot *et al.* 2005, Stein 2007, http://genome.jouy.inra.fr/triannot/index.php et http://www.ncbi.nlm.nih.gov/, Tableau I.1) ont néanmoins permis de confirmer la richesse des génomes du blé en éléments transposables (Smith et Flavell 1975, Vedel et Delseny 1987) et permis d'identifier les types les plus abondants de ces TEs (Wicker *et al.* 2002, Sabot *et al.* 2005). Il est alors apparu qu'aucune insertion de TE n'est conservée entre des régions orthologues des différents génomes du blé, indiquant leur dynamique importante sur la courte échelle de temps de divergence de ces génomes (2,5-4 Ma) (Wicker *et al.* 2003b, Chantret *et al.* 2005, 2008, Dvorak *et al.* 2006, Gu *et al.* 2006, Salse *et al.* 2008b, Charles *et al.* 2008 présenté dans cette partie).

Néanmoins, la dynamique et la prolifération des différents types de TEs, au cours des trois derniers millions d'années, ainsi que leur distribution et leur contribution à l'organisation et aux variations des tailles des différents génomes du blé, n'étaient pas encore explorées du fait de l'absence de séquences représentatives de ces génomes.

A mon arrivée, mon équipe d'accueil commençait un projet de séquençage comparatif de 10 régions dans différentes espèces diploïdes et polyploïdes de blé couvrant des gènes d'intérêt (APCNS2003, http://www.cns.fr/spip/Triticum-ssp-comparative.html). Ce projet continue de représenter, à ce jour, le projet de séquençage comparatif le plus important entrepris chez le blé.

J'ai pu ainsi m'initier à l'annotation des séquences génomiques du blé et j'ai eu l'opportunité de créer un programme permettant d'accélérer leur annotation (voir Matériels et méthodes d'annotation). J'ai aussi modestement participé à l'annotation des séquences de deux locus à l'origine de publications : le locus *HMW-Glu* (Gu *et al.* 2006) et le locus *SPA* (Salse *et al.* 2008b) (Annexe 3 et 4). Cette expérience m'a aussi permis de maîtriser une technique d'estimation des dates d'insertion des rétrotransposons à LTRs en se basant sur la divergence des séquences des LTRs. En effet, les LTRs d'un rétrotransposon à son insertion sont parfaitement identiques, et le calcul de la divergence entre les séquences de ses LTRs permet donc d'estimer leur date d'insertion. On utilise le calcul de la divergence par la méthode Kimura 2 paramètres (Kimura 1980), et une horloge moléculaire de $1{,}3 \times 10^{-8}$ substitutions/site/an (Article 1, Matériel et Méthode).

Je présente dans ce chapitre, l'évaluation de la dynamique et de la prolifération différentielle des TEs dans les génomes A et B du blé. Pour cela, j'ai analysé toutes les

séquences génomiques provenant de clones BAC disponibles publiquement et de clones complémentaires, isolés au laboratoire et séquencés par le Centre National de Séquencage, afin de constituer rapidement un ensemble représentatif de séquences pour les génomes A et B (respectivement 3,6 Mb et 1,98 Mb).

Avec la collaboration d'autres membres du laboratoire, j'ai utilisé une combinaison d'approches basées sur : l'estimation des proportions des copies complètes/incomplètes (tronquées) et le calcul des dates d'insertion des rétrotransposons pour analyser la dynamique de l'espace TEs dans les génomes A et B. De plus, j'ai pu valider la calibration de l'horloge moléculaire utilisée dans la datation des insertions de TEs en croisant les résultats avec une analyse haplotypique *in planta* détectant la présence ou l'absence d'un événement d'insertion dans des collections de lignées représentatives d'espèces diploïdes, tétraploïdes et hexaploïdes. Les anciennes insertions devraient être détectées dans la plupart des lignées alors que les insertions récentes ne devraient être détectées que dans quelques lignées.

Mes travaux ont ainsi permis de montrer pour la première fois que les différents types de TEs ont eu des vagues de proliférations différentielles dans les génomes B et A du blé ce qui a probablement largement contribué à la stabilisation des allopolyploïdes qui les combinent. Je présente mes résultats sous forme d'un article que j'ai publié en 2008 dans la revue *Genetics*. Des résultats importants publiés en ligne dans le même article en tant que 'Supplemental Data' ainsi que l'amplification de rétrotransposons par recombinaisons homologues inégales sont présentés en tant que résultats complémentaires.

II Article 1

Cet article a été publié dans la revue Genetics

Dynamics and Differential Proliferation of Transposable Elements During the Evolution of the B and A Genomes of Wheat

Mathieu Charles, Harry Belcram, Jérémy Just, Cécile Huneau, Agnès Viollet, Arnaud Couloux, Béatrice Segurens, Meredith Carter, Virginie Huteau, Olivier Coriton, Rudi Appels, Sylvie Samain and Boulos Chalhoub

Genetics **180**: 1071-1086, octobre 2008

DOI: 10.1534/genetics.108.092304

Dynamics and Differential Proliferation of Transposable Elements During the Evolution of the B and A Genomes of Wheat

Mathieu Charles,* Harry Belcram,* Jérémy Just,* Cécile Huneau,* Agnès Viollet,† Arnaud Couloux,† Béatrice Segurens,† Meredith Carter,‡ Virginie Huteau,§ Olivier Coriton,§ Rudi Appels,‡ Sylvie Samain† and Boulos Chalhoub*,[1]

**Organization and Evolution of Plant Genomes, Unité de Recherche en Génomique Végétale, UMR: INRA-1165, CNR-S8114, 91057 Evry Cedex, France, †CEA: Institut de Génomique GENOSCOPE, 91057 Evry Cedex, France, ‡State Agricultural Biotechnology Centre and Centre for Comparative Genomics, Murdoch University, Perth, Western Australia 6150, Australia and §Unité Mixte de Recherches INRA, Agrocampus Rennes Amélioration des Plantes et Biotechnologies Végétales, 35653 Le Rheu, France*

Manuscript received June 6, 2008
Accepted for publication August 7, 2008

ABSTRACT

Transposable elements (TEs) constitute >80% of the wheat genome but their dynamics and contribution to size variation and evolution of wheat genomes (Triticum and Aegilops species) remain unexplored. In this study, 10 genomic regions have been sequenced from wheat chromosome 3B and used to constitute, along with all publicly available genomic sequences of wheat, 1.98 Mb of sequence (from 13 BAC clones) of the wheat B genome and 3.63 Mb of sequence (from 19 BAC clones) of the wheat A genome. Analysis of TE sequence proportions (as percentages), ratios of complete to truncated copies, and estimation of insertion dates of class I retrotransposons showed that specific types of TEs have undergone waves of differential proliferation in the B and A genomes of wheat. While both genomes show similar rates and relatively ancient proliferation periods for the *Athila* retrotransposons, the *Copia* retrotransposons proliferated more recently in the A genome whereas *Gypsy* retrotransposon proliferation is more recent in the B genome. It was possible to estimate for the first time the proliferation periods of the abundant *CACTA* class II DNA transposons, relative to that of the three main retrotransposon superfamilies. Proliferation of these TEs started prior to and overlapped with that of the *Athila* retrotransposons in both genomes. However, they also proliferated during the same periods as *Gypsy* and *Copia* retrotransposons in the A genome, but not in the B genome. As estimated from their insertion dates and confirmed by PCR-based tracing analysis, the majority of differential proliferation of TEs in B and A genomes of wheat (87 and 83%, respectively), leading to rapid sequence divergence, occurred prior to the allotetraploidization event that brought them together in *Triticum turgidum* and *Triticum aestivum*, <0.5 million years ago. More importantly, the allotetraploidization event appears to have neither enhanced nor repressed retrotranspositions. We discuss the apparent proliferation of TEs as resulting from their insertion, removal, and/or combinations of both evolutionary forces.

GENOMES of higher eukaryotes, and particularly those of plants, vary extensively in size (Bennett and Smith 1976, 1991; Bennett and Leitch 1997, 2005). This is observed not only among distantly related organisms, but also between species belonging to the same family or genus (Chooi 1971; Jones and Brown 1976). More than 90% of genes are conserved in sequenced plant genomes (Bennetzen 2000a; Sasaki *et al.* 2005; Jaillon *et al.* 2007) and thus differences in gene content explain only a small fraction of the genome size variation. It is widely accepted that whole-genome duplication by polyploidization (Blanc *et al.* 2000; Paterson *et al.* 2004; Adams and Wendel 2005) and differential proliferation of transposable elements (TEs) are the main driving forces of genome size variation. The differential proliferation of TEs results from their transposition (SanMiguel *et al.* 1996; Bennetzen 2000b, 2002a,b; Kidwell 2002; Bennetzen *et al.* 2005; Hawkins *et al.* 2006; Piegu *et al.* 2006; Zuccolo *et al.* 2007) as well as the differential efficiency of their removal (Petrov *et al.* 2000; Petrov 2002a,b; Wendel *et al.* 2002).

Polyploidization and differential proliferation of TEs are particularly obvious in the case of wheat species belonging to the closely related Triticum and Aegilops genera. Rice (*Oryza sativa*), *Brachypodium*, and diploid Triticum or Aegilops species underwent the same whole-genome duplications (Adams and Wendel 2005; Salse

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AM932680, AM932681, AM932682, AM932683, AM932684, AM932685, AM932686, AM932687, AM932688, AM932689.

[1]*Corresponding author:* Organization and Evolution of Plant Genomes, Unité de Recherche en Génomique Végétale, UMR: INRA-1165, CNR-S8114, 2 rue Gaston Crémieux, 91057 Evry Cedex, France.
E-mail: chalhoub@evry.inra.fr

et al. 2008), but Triticum or Aegilops genomes are >10 times larger (Bennett and Smith 1991), mainly due to proliferation of repetitive DNA, which represents >80% of the genome size (Smith and Flavell 1975; Vedel and Delseny 1987). Diploid wheat species can differ in their genome sizes by hundreds or even thousands of megabases (Bennett and Smith 1976, 1991; http://data.kew.org/cvalues/homepage.html). For example, the genome size of *Triticum monococcum* (6.23 pg) is 1.3 pg greater than that of *Triticum urartu* (4.93 pg) (Bennett and Smith 1976, 1991), although these species diverged <1.5 million years ago (MYA) (Dvorak *et al.* 1993; Huang *et al.* 2002; Wicker *et al.* 2003b). Similarly, the calculated size of the B genome of polyploid wheat species (7 pg) is higher than that of any diploid wheat species (http://data.kew.org/cvalues/homepage.html).

The genome size variation within wheat is also accentuated by frequent allopolyploidization events, among which two successive events have led to the formation of the allohexaploid bread wheat *Triticum aestivum* ($2n = 6x = 42$, AABBDD). The first event led to the formation of the allotetraploid *Triticum turgidum* ($2n = 4x = 28$, AABB) and occurred <0.5–0.6 MYA between the diploid species *T. urartu* ($2n = 2x = 14$, AA), donor of the A genome, and an unidentified diploid species of the Sitopsis section, donor of the B genome (Feldman *et al.* 1995; Blake *et al.* 1999; Huang *et al.* 2002; Dvorak *et al.* 2006). The second allopolyploidization event occurred 7000–12,000 years ago, between the early domesticated tetraploid *T. turgidum* ssp. *dicoccum* and the diploid species *Aegilops tauschii* ($2n = 14$), donor of the D genome, resulting in hexaploid wheat (Feldman *et al.* 1995).

The amount of available wheat genomic sequences is very limited, compared to other organisms (reviewed by Sabot *et al.* 2005; Stein 2007; http://genome.jouy.inra.fr/triannot/index.php and http://www.ncbi.nlm.nih.gov/). Individual bacterial artificial chromosome (BAC) clones, selected primarily because they contained genes of agronomic interest, have been sequenced. Analyses of randomly chosen BAC clones from wheat have been also performed (Devos *et al.* 2005), and 2.9 Mb of sequences from a whole-genome shotgun library of *Ae. tauschii* were analyzed by Li *et al.* (2004). More recently, a detailed analysis of 19,400 BAC-end sequences of chromosome 3B, representing a cumulative sequence length of nearly 11 Mb (1.1% of the estimated chromosome length) was reported (Paux *et al.* 2006). Altogether, these sequencing efforts have confirmed previous estimates of the amount of repetitive DNA in the wheat genome (~80%) (Smith and Flavell 1975; Vedel and Delseny 1987) and have identified the major types of TEs (Wicker *et al.* 2002; Sabot *et al.* 2005).

Because of the limited genomic sequence information, the extent to which various TEs contribute to the wheat genome and affect its size variation, or how they are distributed among different genomes, remains unexplored. Little is known about the dynamics of TEs, their proliferation processes, and whether they proliferated gradually or in waves of sudden bursts of insertions. In this study, 10 genomic regions from wheat chromosome 3B were sequenced and used to constitute, along with three other genomic sequences, 1.98 Mb of sequence from the wheat B genome. Transposable element dynamics and proliferation in these B-genome sequences were analyzed and compared to those in 3.63 Mb of sequence from 19 genomic regions of the wheat A genome. Our study provides novel insights into the dynamics and differential proliferation of TEs as well as their important role in the evolution and divergence of the wheat B and A genomes.

MATERIALS AND METHODS

Plant material and genomic DNA isolation: Hexaploid wheat deletion lines used to map the 10 BAC clones on different deletion bins of chromosome 3B (see RESULTS) were originally described by Qi *et al.* (2003) and kindly provided by Catherine Feuillet (INRA, Clermont-Ferrand, France). Hexaploid wheat genotypes were kindly provided by Joseph Jahier (INRA, Rennes, France). Tetraploid wheat genotypes were kindly provided by Moshe Feldman (Weizemann Institute). Genomic DNA was extracted from leaves as described by Graner *et al.* (1990).

Primer design and PCR-based tracing of retrotransposon insertions: The program Primer3 (Rozen and Skaletsky 2000) was used to design oligonucleotide primers on the basis of TE–TE or TE-unassigned DNA junctions. We often designed and used several couples (including nested) of PCR primers. Internal controls (PCR primers designed within the TE) were also used. Primer sequences are given in supplemental Table 1. PCR reactions were carried out in a final volume of 10 μl with 200 μM of each dNTP, 500 nM each of forward and reverse primers, 0.2 units Taq polymerase (Perkin Elmer). PCR amplification was conducted using the following "touchdown" procedure: 14 cycles (30 sec 95°, 30 sec 72° minus 1° for each cycle, 30 sec 72°), 30 cycles (30 sec 95°, 30 sec 55°, 30 sec 72°), and one additional cycle of 10 min 72°. Amplification products were visualized using standard 2% agarose gels.

BAC sequencing, sequence assembly, and annotation: BAC shotgun sequencing was performed at the Centre National de Sequencage (Evry, France) essentially as described by Chantret *et al.* (2005). Genes, TEs, and other repeats were identified by computing and integrating results on the basis of BLAST algorithms (Altschul *et al.* 1990, 1997), predictor programs, and different software and procedures, detailed below. Cross-analysis of the information obtained for genes and TEs as well as for repeats and unassigned DNA was integrated into ARTEMIS (Rutherford *et al.* 2000). Sequence annotation and analysis were performed as described in supplemental Method 1. The 10 BAC clone sequences were submitted to EMBL and under the following accession nos.: TA3B54F7, AM932680; TA3B63B13, AM932681; TA3B63B7, AM932682; TA3B81B7, AM932683; TA3B95C9, AM932684; TA3B95F5, AM932685; TA3B95G2, AM932686; TA3B63C11, AM932687; TA3B63E4, AM932688; TA3B63N2, AM932689. Accession numbers for the three publicly available genomic sequences

from the wheat B genome (Sabot *et al.* 2005; Gu *et al.* 2006; Dvorak *et al.* 2006) are CT009588, AY368673, DQ267103.

Publicly available genomic sequences from the wheat A genome: The retained publicly available A-genome sequences consist of 19 sequenced and well annotated BAC clones or contigs (SanMiguel *et al.* 2002; Yan *et al.* 2002, 2003; Wicker *et al.* 2003b; Chantret *et al.* 2005; Isidore *et al.* 2005; Dvorak *et al.* 2006; Gu *et al.* 2006; Miller *et al.* 2006), representing >3.5 Mb. Accession numbers for the analyzed BAC sequences are the following: diploid A genome—AF326781, AF488415, AY146588, AY188331, AY188332, AY188333, AY491681, AY951944, AY951945, DQ267106, AF459639; tetraploid A genome—AY146587, AY485644, AY663391, CT009587, DQ267105; hexaploid A genome—AY663392, CT009586, DQ537335.

Chromosome 3B BAC clones and fluorescent *in situ* hybridization: The 10 BAC clones and/or their subclones were originally mapped by fluorescence *in situ* hybridization (FISH) on flow-sorted 3B chromosomes using the Cot-1 fraction as blocking DNA to suppress hybridization of repeated sequences (Dolezel *et al.* 2004; Safar *et al.* 2004; M. Kubalakova and J. Dolezel, personal communication). Further FISH hybridization experiments were conducted, without Cot-1 DNA, on mitotic metaphase chromosomes of hexaploid wheat (*T. aestivum*) cv. Chinese Spring. The FISH hybridization protocol is presented in supplemental Method 2.

Estimation of Long Terminal Repeat-retrotransposon insertion dates: For all genomic sequences of the B and A genomes of wheat, retrotransposon copies with both 5′ and 3′ long terminal repeats (LTRs), and target-site duplications (TSD) were considered as corresponding to original insertions and analyzed by comparing their 5′ and 3′ LTR sequences. The two LTRs were aligned and the number of transition and transversion mutations was calculated using MEGA3 software (Kumar *et al.* 2004). A mutation rate of 1.3×10^{-8} substitutions/site/year (SanMiguel *et al.* 1998; Ma *et al.* 2004; Ma and Bennetzen 2004; Wicker *et al.* 2005; Gu *et al.* 2006) was used. The insertion dates and their standard errors (SE) were estimated using the formula $T = K2P/2r$ (Kimura 1980).

Statistical analysis: All statistical analyses and the different tests (Kolmogorov–Smirnov, Bootstrap, and probability density functions) were done with the R-package (http://www.r-project.org). Kolmogorov–Smirnov tests (Férignac 1962) were applied to check whether the distribution of insertion dates of retrotransposons deviates from uniformity, and whether they are different when comparing different TE families or superfamilies within and between the B and A genomes. Probability density of TE insertion dates was estimated using Gaussian kernel density estimation (Silverman 1986), taking into account measured standard deviation for each individual insertion date (Kimura 1980).

RESULTS

Constitution of a genomic sequence data set representative of the wheat B genome—analysis of 10 BAC sequences from the wheat chromosome 3B: Only three large well-annotated genomic sequences (BAC clones), representing 0.55 Mb of sequence, were available for the wheat B genome (Sabot *et al.* 2005; Dvorak *et al.* 2006; Gu *et al.* 2006). To obtain more representative genomic sequences, we sequenced and annotated 10 BAC clones of wheat chromosome 3B, representing 0.15% of the chromosome length (1.43 Mb) (Figure 1). Detailed annotation files are deposited at EMBL/GenBank Data Libraries.

These sequenced genomic regions show a high proportion of TEs, which represent 79.1% of the cumulative sequence length (Figure 1, supplemental Table 2). Other repeated DNA sequences represent 2.4% and unassigned DNA sequences account for 17.5% of the cumulative sequence length.

We conducted gene prediction analysis for the remaining 18.5% non-TEs and nonrepeated DNA, using different search programs (see supplemental Method 1 and supplemental Text 1 for detailed description). Genes of known and unknown functions or putative genes were defined on the basis of predictions and the existence of rice or other Triticeae homologs. Hypothetical genes were identified on the basis of prediction programs only. Pseudogenes were not well predicted and frameshifts need to be introduced within the coding sequences (CDS) structure to better fit a putative function on the basis of BLASTX (mainly with rice). Truncated pseudogenes (genes disrupted by large insertion or deletion) and highly degenerated CDS sequences were considered as *gene-relics*. Combined together, all these types of gene sequence information (GSI) account for only 1.0% of the sequence and are present in seven BAC clones (one or two genes per clone) while the remaining three BAC clones (TA3B95C9, TA3B95G2, TA3B63N2) contain no genes (indicated in Figure 1A and detailed in supplemental Text 1, supplemental Table 3, and supplemental Table 4).

Six genes (of known or unknown function) and two putative genes were identified using the FGENESH prediction software (http://www.softberry.com) and by identification of homologs in rice (Figure 1A, supplemental Table 3). Six additional "gene-relics" or "pseudogenes" were also identified on the basis of colinearity with rice (Figure 1A, supplemental Table 3). Finally, 10 CDS, designated as "hypothetical genes," were identified according to the FGENESH prediction program only (Figure 1A, supplemental Table 4).

TE prediction, annotation, classification, and nomenclature were performed essentially as suggested by the unified classification system for eukaryotic TEs (Wicker *et al.* 2007) with two modifications. The *Athila* retrotransposons were analyzed separately from the other *Gypsy* retrotransposons (see also supplemental Methods 1). The *Sukkula* retrotransposons were considered as belonging to the *Gypsy* superfamily because of similarities with the *Erika* (*Gypsy*) elements. The 79.1% of TEs were shown to be composed of a wide variety of TEs, distributed as follows: 61.9% class I (171 TEs from 48 families), 16.2% class II (113 TEs from 28 families), and 1.0% unclassified TEs (18 TEs from 9 families) (Figure 1). The *CACTA* TEs represent the majority (96%) of class II TEs. More details about the TE composition in the 10 different BAC clones of wheat chromosome 3B are provided in supplemental Text 2.

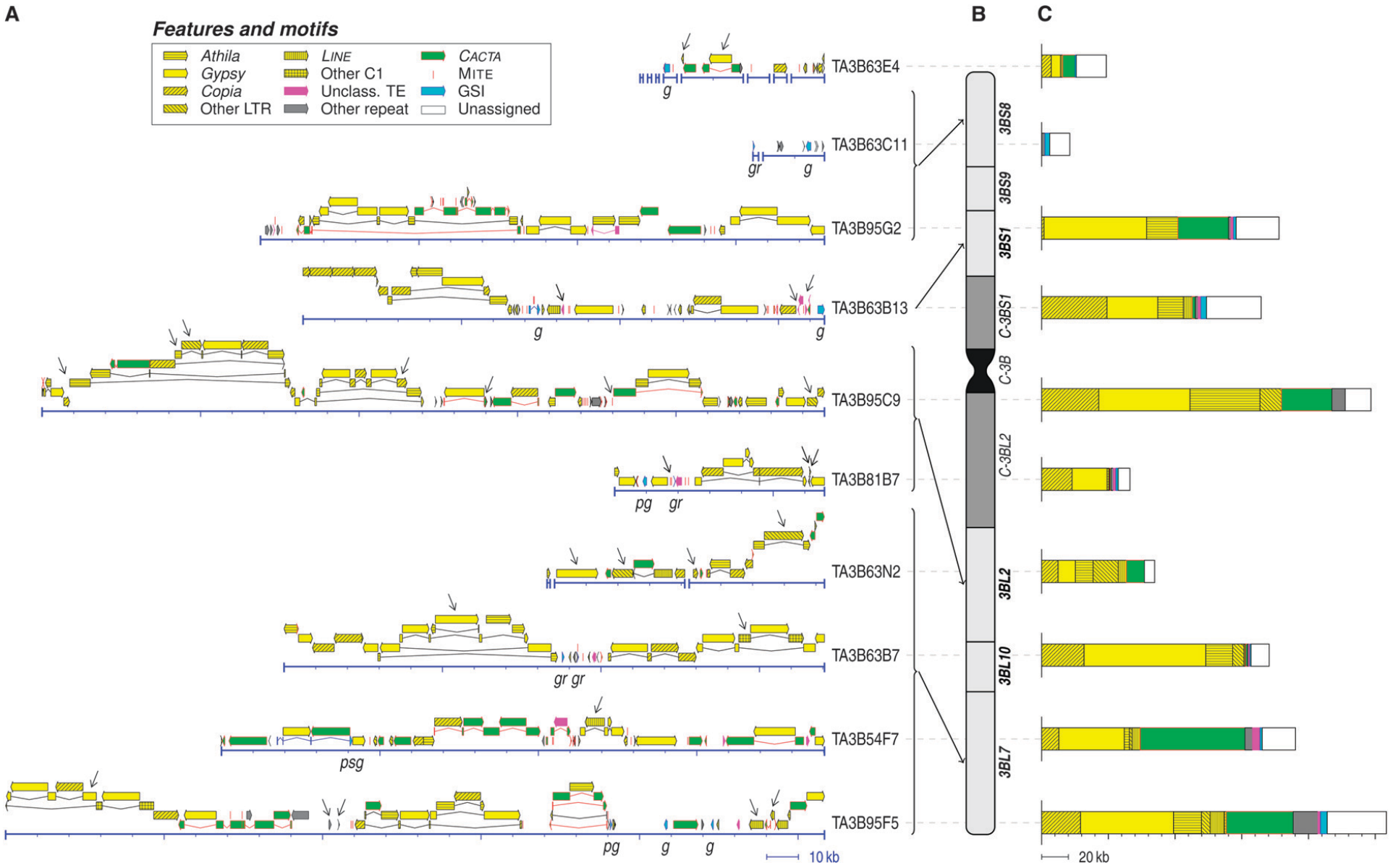


FIGURE 1.—Detailed annotation, BIN map positions, and sequence composition of 10 sequenced BAC clones of wheat chromosome 3B. (A) Detailed annotations of the 10 sequenced BAC clones. Main TEs, other repeats, and gene sequence information (GSI) are represented with distinct features and motifs (detailed in the "features and motifs" key). *g*, genes; *pg*, putative genes; *gr*, gene relics; and *psg*, pseudogenes. For nested insertions of TEs, the newly inserted TE is presented above the split one. Complete reconstruction of split TEs was done and the different parts are linked with a line to visualize the entire element. Some BAC clones are represented by several unordered contigs (TA3B63E4, TA3B63C11, TA3B63N2). EMBL BAC clone references and annotation files are given in MATERIALS AND METHODS. Detailed coding sequence and TE descriptions are supplied in supplemental Text 1 and supplemental Text 2. Arrows indicate novel TEs identified in this study and described in supplemental Text 2 and supplemental Table 5. (B) BIN map position of nine of the BAC clones. The wheat chromosome 3B bins are according to QI *et al.* (2003). Details of the genotyping results are given in supplemental Table 6. (C) Proportions of the main sequence classes and types. See "features and motifs" in A for an explanation of colors. Details are given in supplemental Table 2.

TABLE 1

Details of TEs from the four most represented superfamilies in 13 genomic regions of the wheat B genome, compared to publicly available sequences from 19 genomic regions of the wheat A genome

	13 genomic regions of the wheat B genome (1.98 Mb)[a]				19 publicly available genomic regions of the wheat A genome (3.63 Mb)[b]			
	Athila	*Copia*	*Gypsy*	*CACTA*	*Athila*	*Copia*	*Gypsy*	*CACTA*
Observed number of TEs	54	57	79	70	72	149	123	53
Sequence proportion (means ±SE)%[c]	10.8 ± 1.6	14.2 ± 2.5	28.1 ± 3.8	13.4 ± 3.3	10.4 ± 1.8	21.8 ± 1.8	19.7 ± 2.9	9.4 ± 1.9
Bootstrap means deviation[d]	−0.07	+0.02	+0.02	−0.05	+0.01	−0.02	−0.03	−0.09
Complete TEs with TSD (%)	13	18	39	19	19	60	38	32
Incomplete (truncated) TEs	41	39	40	51	53	89	85	21
LTR-mediated homologous recombination								
Entire TE without TSD	3	7	0	—	0	4	0	—
Solo LTR	4	2	2	—	5	15	9	—
Illegitimate recombination	34	30	38	51	48	70	76	21
Complete TEs/incomplete (truncated) TEs	0.32	0.46	0.98	0.37	0.36	0.67	0.45	1.52

[a] This corresponds to 1.43 Mb from the 10 genomic regions sequenced in this study and 0.55 Mb from three other publicly available genomic regions from Sabot *et al.* (2005), Gu *et al.* (2006), and Dvorak *et al.* (2006). See materials and methods for BAC clone sequence references.

[b] Nineteen genomic regions available for the A genome (SanMiguel *et al.* 2002; Yan *et al.* 2002, 2003; Wicker *et al.* 2003b; Chantret *et al.* 2005; Isidore *et al.* 2005; Dvorak *et al.* 2006; Gu *et al.* 2006; Miller *et al.* 2006). See materials and methods for BAC clone sequence references.

[c] Relative to cumulative sequence length. SE, standard errors for estimated means.

[d] Differences between arithmetic means (line above) and bootstrap analysis (Efron 1979) with 10,000 resamplings.

Twenty-one transposable element families, some of which are present in several copies, were identified for the first time in this study (Figure 1A, indicated by arrows). They account for 9.8% by number and 7.9% by length of the overall sequences. Class I retrotransposons are the category for which we found the majority of novel TE families (17). Description of these novel TEs, their features, and the suggested nomenclature are presented in supplemental Text 2 and supplemental Table 5.

The 10 sequenced BAC clones or their subclones were originally mapped by FISH on flow-sorted 3B chromosomes, using the $C_0t - 1$ fraction as blocking DNA to suppress hybridization of repeated sequences (Dolezel *et al.* 2004; Safar *et al.* 2004; M. Kubalakova and J. Dolezel, personal communication). As described by Devos *et al.* (2005) and Paux *et al.* (2006), specific PCR markers, based on TE–TE or TE-unassigned DNA junctions, were used to confirm the different BAC clone map positions on the deletion bins (Qi *et al.* 2003) of chromosome 3B (except TA3B63E4) (Figure 1B). Details of PCR markers and genotyping results are given in supplemental Table 6.

Representation of transposable elements and the wheat B genome: Five BAC clone sequences were publicly available from the B genome of wheat (Sabot *et al.* 2005; Dvorak *et al.* 2006; Gu *et al.* 2006). Four of these were sequenced for two orthologous regions in tetraploid and hexaploid wheat species (one BAC clone per region and per species) (Sabot *et al.* 2005; Gu *et al.* 2006). As they share nearly identical sequences (99%) with common TE insertions, they were considered as redundant in our study and only the longest BAC clone sequences (three in total) were counted in calculation and appreciation of TE proliferation. These, added to the above-described 10 genomic region sequences of wheat chromosome 3B, constitute 1.98 Mb of sequence from the wheat B genome. Four main TE superfamilies occupy 66.5% of the analyzed B-genome loci: the *Athila* superfamily (54 elements), the *Copia* superfamily (57 elements), the *Gypsy* superfamily (79 elements), and the *CACTA* superfamily (70 elements) (Table 1). Interestingly, proportions of the *Athila*, *Copia*, and *Gypsy* retrotransposons (respectively, 10.8, 14.2, and 28.1%) (Table 1) are very similar to estimates based on 11 Mb of the chromosome 3B sequence BAC end (Paux *et al.* 2006). The major deviation concerns the proportion of *CACTA* class II TEs, which is higher in the 13 genomic regions (13.4%) than in the overall BAC-end sequences (4.9%), probably due to their clustering in some BAC clones that we have sequenced, such as TA3B54F7 (40.5% of CACTA TEs) (Figure 1).

The 13 sequences represent only ~0.03% of the B genome. However, statistical tests, using SE as well as a bootstrap analysis with 10,000 resamplings, confirm the robustness of estimations of sequence proportions of the *Gypsy*, *Copia*, *Athila*, and *CACTA* TE superfamilies (Table 1). We also evaluated the variation of mean

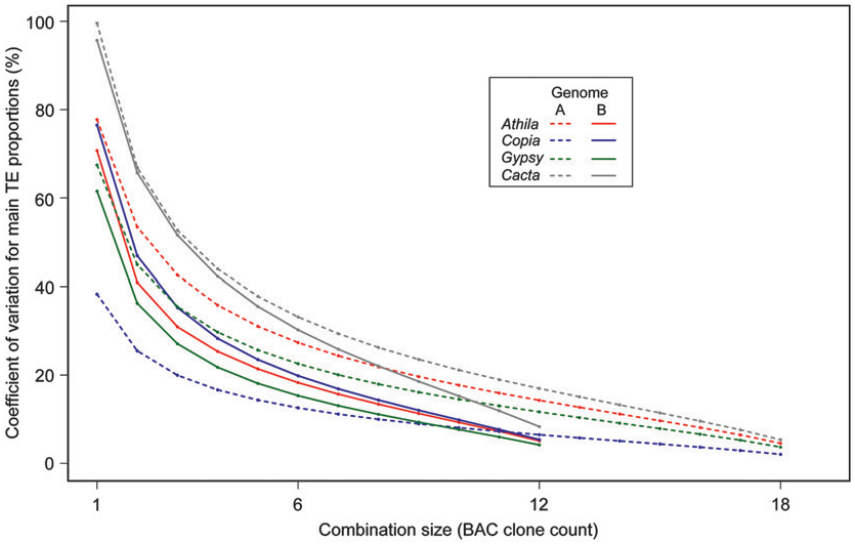


FIGURE 2.—Changes of the coefficient of variation of proportions (in percentages) of the main transposable element superfamilies calculated over all possible BAC clone combinations and simulated over a size varying from 1 to 12 BAC clones for the wheat B genome and 1 to 18 for the wheat A genome (combination size). For each number of considered BAC clones (*x*-axis), sequence proportions (in percentages) were calculated for all possible BAC clone combinations, and the coefficient of variation between these proportions was calculated (*y*-axis).

sequence proportions estimated for the four TE superfamilies by comparing all possible clone number representations and combinations (from 1 to 12 BAC clones) (Figure 2). Results show that representing the wheat B genome with a low number of BAC clones results in very variable proportions of the TE sequences (Figure 2). These variations decrease significantly by increasing the number of considered BAC clones (Figure 2). This confirms the usefulness of our effort in sequencing more BAC clones for better representation of the wheat B genome.

It is also interesting to note that direct FISH hybridization, using the whole BAC clone as a probe, resulted in dispersed and mostly homogenous signals across all wheat chromosomes for 8 of all 10 BAC clones of wheat chromosome 3B (except TA3B63C11 and TA3B54F7) (SAFAR *et al.* 2004 and supplemental Figure 1), thus confirming sequencing results that show high TE composition.

Constitution of a genomic sequence data set representative of the wheat A genome: The publicly available A-genome sequences that we were able to use are more abundant and consist of 20 sequenced and well-annotated BAC clones or contigs. Ten of these were comparatively sequenced for five orthologous regions of the wheat A genome at the diploid, tetraploid, and/or hexaploid levels and were partially overlapping (WICKER *et al.* 2003b; CHANTRET *et al.* 2005; ISIDORE *et al.* 2005; DVORAK *et al.* 2006; GU *et al.* 2006), while others were determined at only one ploidy level (mostly diploid) (SANMIGUEL *et al.* 2002; YAN *et al.* 2002, 2003; MILLER *et al.* 2006). Comparisons show that no shared TE insertions were observed between orthologous regions (from two ploidy levels), except in the region of the high-molecular-weight (HMW) glutenin gene, the sequences of which were nearly identical at the tetraploid and hexaploid levels (GU *et al.* 2006). Thus, we used only the sequence from hexaploid wheat to represent the HMW glutenin gene region and considered all the other different orthologous regions (from different ploidy levels) separately. This led to 19 BAC clones, representing 3.63 Mb of sequence, that were analyzed for the wheat A genome.

The *Gypsy* TEs were found to occupy 19.7%, the *Athila* TEs 10.4%, the *Copia* TEs 21.8%, and the *CACTA* TEs 9.4% of the cumulative sequence length (Table 1). Similarly, for the B-genome sequences, we also analyzed and validated the robustness of the estimation of sequence proportions of the main TE superfamilies and their representation of the A genome (Figure 2). Similar proportions of the *Gypsy*, *Copia*, *Athila*, and *CACTA* TEs were found whether the 11 genomic sequences from the diploid A genome or those determined from A genomes of tetraploid (six regions) and hexaploid (three regions) wheat species were considered separately or combined (data not shown).

Comparison of TE sequence proportions and ratios of complete to truncated copies: Our analysis showed a significantly higher number of *Gypsy* retrotransposons in the wheat B-genome sequences than in the A genome (Table 1). Conversely, a higher proportion of *Copia* retrotransposons is observed in genomic sequences of the wheat A genome than in the B genome (Table 1). Proportions of the *Athila* and *CACTA* TEs were not statistically different between the two genomes (Table 1).

Major differences were found between the three main retrotransposon superfamilies in the ratio of complete (intact) copies, defined as having both LTRs and target-site TSD, as compared to degenerated and truncated copies that resulted from LTR-mediated unequal homologous recombinations or illegitimate DNA recombination (DEVOS *et al.* 2002; MA *et al.* 2004; MA and BENNETZEN 2004; VITTE and BENNETZEN 2006) (Table 1). In the B-genome sequences, the *Athila* and *Copia* retrotransposons show low ratios of complete to incomplete retrotransposons (respectively, 0.32 and 0.46),

whereas the *Gypsy* retrotransposons show the highest ratio (0.98) (Table 1). In comparison, the 3.63 Mb of genomic sequence of the wheat A genome shows a lower ratio (0.45) of complete to incomplete *Gypsy* retrotransposons whereas proportions of intact *Copia* retrotransposons are relatively higher than those observed in the B genome (0.67) (Table 1). The *Athila* retrotransposon ratio in the A genome is comparable to the ratio in the B genome (0.36 and 0.32, respectively).

CACTA TE original insertions are characterized by the "CACTA" sequence and 3-bp TSD sequence motifs surrounding terminal inverted repeats (TIR) at both ends. We used these signatures to define complete *CACTA* copies, where the "CACTA," TIR, and TSD sequence motifs are observed at both ends, and truncated copies, where the "CACTA" and TSD motifs are absent from one or both ends. The ratio of complete to incomplete copies of the *CACTA* class II TEs was about five times lower in the wheat B genome (ratio of 0.37) than in the A genome (ratio of 1.52) (Table 1).

Insertion dates and proliferation of LTR retrotransposons: To understand differences in sequence proportions and the ratios of complete to truncated copies between retrotransposon superfamilies, as well as between the B and A genomes, we compared TE proliferation periods and rates.

The two LTRs are identical at the time of retrotransposon insertion and their sequence divergence reflects time lapsed since the insertion (SanMiguel *et al.* 1998). Several studies have shown that LTRs evolve at approximately twice the rate of genes and UTR regions, and we used a rate of 1.3×10^{-8} substitutions/site/year (Ma *et al.* 2004; Ma and Bennetzen 2004; Wicker *et al.* 2005; Gu *et al.* 2006).

We calculated the LTR divergence and dates of insertion of the *Athila*, *Copia*, and *Gypsy* retrotransposon (complete copies with both LTRs and TSD) found in the wheat B and A genomes (Figure 3). Such TE insertion dates offer a very important insight into the relative timing of various events, regardless of the approaches used to estimate nucleotide substitution rates or the molecular clock calibration points used in these calculations.

The vast majority of complete retrotransposons in the B and A genomes of wheat (86 and 92%, respectively) were estimated to be <3 million years old (Figure 2) in agreement with several previous studies of grasses and other plants species (SanMiguel *et al.* 1998, 2002; Wicker *et al.* 2003b, 2005; Gao *et al.* 2004; Ma *et al.* 2004; Du *et al.* 2006; Piegu *et al.* 2006; Wicker and Keller 2007). This is explained by the fact that LTR retrotransposons are continuously removed by unequal homologous recombination and illegitimate DNA recombination as new ones are inserted (Vicient *et al.* 1999; Devos *et al.* 2002; Ma *et al.* 2004; Pereira 2004). Insertion of the *Egug* element (RLGa_*Egug_TA3B95C9-1* ~5 MYA; divergence of 0.131) is the oldest such event found in our study and the most recent one is the *Sukkula* insertion (RLG_*Sukkula_TA3B63B7-2*) for which only a 1-base indel differentiates the two LTRs of 4192/4193 bp.

Comparison of LTR divergence dates revealed that different LTR–retrotransposon superfamilies and families proliferated at different periods and rates during evolution of the wheat B and A genomes (Figure 3). We applied Kolmogorov–Smirnov tests to check whether within the last 3 million years (0.078 divergence) the distribution of insertion dates of retrotransposons deviates from uniformity (thus confirming a burst of higher proliferation), and whether these dates are different when comparing different retrotransposon families or superfamilies within and between the wheat B and A genomes (thus illustrating differential proliferation). This was done for all complete copies of the three main retrotransposon superfamilies as well as for the most abundant retrotransposon families (nine) that have five or more complete copies in the B and/or A genomes (Figure 3).

Superfamily level comparison: The combination of all complete retrotransposon copies at the superfamily level (Figure 3A) indicated that the distribution of the *Gypsy* retrotransposon insertion dates in both B and A genomes and that of *Copia* retrotransposons in the A genome were significantly different from uniform (*P*-value <0.01) because of their higher proliferation during the last 2 million years (Figure 3A). Proliferation of the *Copia* retrotransposons in the B genome was uniform and low all across the 3-million-year period, whereas proliferation of the *Athila* retrotransposons was different from a uniform distribution in both genomes at *P*-value <0.1.

One possible reason for the non-uniform distributions of retrotransposon insertion dates within the 3-million-year period is because older insertions are more likely to be removed (completely or partially) from the genome (see above). Therefore, we checked whether distributions of insertions are significantly different from a uniform distribution for the most recent period of evolution during which the impact of DNA removal should be lower. To carry out this analysis, we divided the LTR–retrotransposon insertions according to the median (of their distribution) that varies depending on the retrotransposon superfamily and family (Figure 3A, gray circle). Kolmogorov–Smirnov (Férignac 1962) tests were then conducted on half of the complete copies, which show the most recent insertion dates. Distribution of insertion dates of the *Gypsy* retrotransposons in the wheat B genome and that of the *Copia* retrotransposons in the B and A genomes can be considered as uniform (*P*-value >0.05, Figure 3A), indicating that they have constantly proliferated during this most recent period. In contrast, the distribution of *Athila* retrotransposons in the wheat B and A genomes and that of *Gypsy* retrotransposons in the A genome

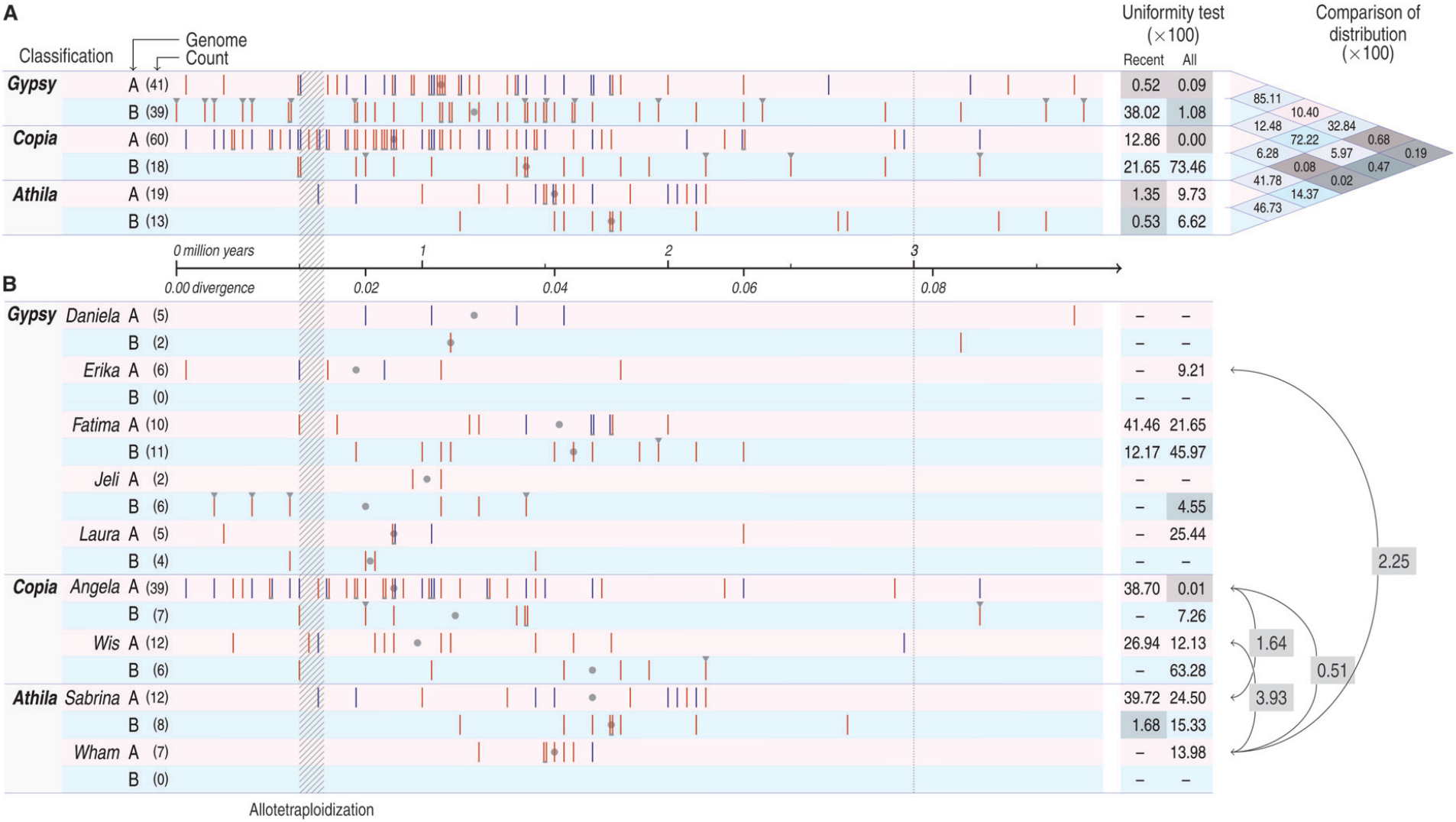


FIGURE 3.—Distribution of insertion dates estimated for LTR retrotransposons in the B and A-genome sequences of wheat (divergence and MYA). (A) All dated LTR retrotransposons combined at the three main superfamily levels (*Athila*, *Gypsy*, and *Copia*). (B) The most abundant retrotransposon families, showing five or more dated copies in at least one of the A or B wheat genomes. Mean insertion dates calculated for retrotransposons are represented by vertical bars. For the A-genome sequences, blue indicates retrotransposons detected from the diploid and red from the polyploid genomic sequences. The genomic sequences of the B genome (red) were obtained from the polyploid wheat. Copies of a given retrotransposon superfamily or family showing identical mean insertion dates are presented by adjacent vertical bars that are joined with a lower horizontal gray bar. The number within parentheses corresponds to the total number of considered retrotransposon copies. Gray triangles indicate retrotransposon insertions that have been traced using PCR in a collection of genotypes of *T. aestivum* and *T. turgidum*. The interval period of the allotetraploidization event (0.5–0.6 MYA, divergence 0.013–0.016) is highlighted in gray. "Uniformity test" refers to Kolmogorov–Smirnov (FÉRIGNAC 1962) tests determining probabilities (*P*-value) that the distribution of insertion dates of retrotransposons deviates from uniformity (thus confirming a burst of higher proliferation): "All" refers to the last 3 million years (0.078 divergence); "Recent" refers to the most recent periods, estimated when dividing the LTR–retrotransposon insertions by the median (indicated by gray circles). Tests were done on families that show five copies or more. "Comparison of distribution" indicates the same Kolmogorov–Smirnov tests determining probabilities that distributions of insertion dates for the last 3 million years (0.078 divergence) are different in the retrotransposon superfamilies and families as well as the in B and A genomes of wheat.

are not uniform (*P*-value <0.05, Figure 3A), consistent with a decreasing proliferation during the most recent period.

Comparison of the proliferation of the three retrotransposon superfamilies shows that distribution of the *Athila* retrotransposons is statistically different from that of the *Gypsy* retrotransposons (Figure 3A, *P*-value <0.05) in the B genome. The *Athila* distribution is significantly different from that of the *Gypsy* and *Copia* retrotransposons (Figure 3A, *P*-value <0.05) in the A genome.

Comparison of the distributions of the three retrotransposon superfamilies between the B and A genomes shows that *Copia* distributions are significantly different (Figure 3A, *P*-value = 0.628) due to their higher proliferation and more recent insertions in the A genome. Both genomes show similar old distribution of the *Athila* retrotransposons (Figure 3A). Distributions of the *Gypsy* retrotransposons were not statistically different between the two genomes for the entire 3-million-year period (Figure 3A, *P*-value >0.05). However, separate Kolmogorov–Smirnov tests for the most recent period show that these have proliferated less in the wheat A genome (*P*-value = 0.052, Figure 3A), unlike in the wheat B genome (*P*-value = 0.38, Figure 3A).

Distribution of the most abundant retrotransposon families: Some specific retrotransposon families were abundant in the B and/or A genomes. This is the case of the *Angela* and *Wis* families, together representing 72 and 85% of the *Copia* superfamily in the B and A genomes, respectively (Figure 3B). This is also the case of the *Sabrina* family representing 62 and 63% of the *Athila* superfamily in the B and A genomes, respectively (Figure 3B). There are more families that compose the *Gypsy* retrotransposon superfamily, the most abundant being *Fatima*, representing 25% in both genomes (Figure 3B).

Kolmogorov–Smirnov tests show nonsignificant deviations (*P*-value >0.05) from uniform distributions for all nine retrotransposon families (with five or more observed complete copies in at least one genome), with the exception of the *Jeli* (*Gypsy*) elements in the B genome and the *Angela* (*Copia*) elements in the A genome, which have more recently proliferated (Figure 3B). Separate analysis for the most recent period, corresponding to half of the complete copies, shows that, as expected from the superfamily-level analysis, the *Wham* family in the A genome and the *Sabrina* family in the B genome have not recently proliferated (*P*-value <0.05, Figure 3B).

Distribution of insertion dates of the *Wham* and *Sabrina* families is different from almost all the other seven families within and between the B and A genomes (*P*-value <0.05). Distribution of insertion dates of the *Angela* family in the wheat A genome is statistically different (*P*-value <0.05) from that of the *Fatima* family in both genomes. Distributions of insertion dates of the remaining families do not show statistical differences (*P*-value >0.05) within and between the wheat B and A genomes (Figure 3B).

Moreover, some retrotransposon families were abundant and present in several complete copies in only one genome (*Romani*, *Daniela*, *Erika*, and *Wham* for the A genome; *Egug* and *Jeli* for the B genome) but absent or presenting few copies in the other (Figure 3B). It is likely that this corresponds to differential proliferation of the considered retrotransposons, as different copies were detected in different genomic regions of wheat B or A genomes.

LTR–retrotransposon proliferation was neither enhanced nor repressed by the allotetraploidization event: The allotetraploidization event that brought the B and A genomes of wheat together in one nucleus was estimated to occur no more than 0.5–0.06 MYA (Huang *et al.* 2002; Dvorak *et al.* 2006; Chalupska *et al.* 2008). This corresponds to a divergence interval of 0.013–0.016, using the corrected rate of 1.3×10^{-8} substitutions/site/year for more rapid divergence of LTRs (Ma *et al.* 2004; Ma and Bennetzen 2004; Dvorak *et al.* 2006).

Comparisons show that retrotransposon insertions continued in wheat B and A genomes during the last 0.5–0.6 million years, apparently without being enhanced nor repressed by the allotetraploidization event (Figure 3). For example, analysis of genomic sequences available from the three ploidy levels of the A genome does not show differences in proliferation periods and rates of retrotransposons (Figure 3).

To check the accuracy of these observations and to calibrate the divergence rate used for coding sequences, on one hand, and that used for LTRs of retrotransposons, on the other hand, we traced several retrotransposons for their insertion prior or posterior to the allopolyploidization event. A PCR-based tracing strategy, derived from the retrotransposon-based insertion polymorphism method (Flavell *et al.* 1998; Devos *et al.* 2005; Paux *et al.* 2006), was developed for 21 retrotransposon insertions from the B genome, sampled as having different estimated insertion dates (Figure 3, indicated by gray triangles). It simply relies on primers designed in both the retrotransposon and its flanking sequences (either unassigned DNA or an older preinserted TE sequence) so that PCR amplification will be specific to the retrotransposon insertion. As the diploid wheat species donor of the B genome is unknown (Feldman *et al.* 1995; Blake *et al.* 1999; Huang *et al.* 2002), we analyzed the occurrence (*i.e.*, presence or absence) of the 21 retrotransposon insertions in hexaploid (*T. aestivum*) and tetraploid (*T. turgidum*) wheat genotypes, which carry the wheat B genome. Examples of PCR-based tracing of the 21 original retrotransposon insertions in the wheat genotypes compared with their estimated insertion dates (±SE) are presented in Figure 4. Full tracing results are supplied in supplemental Table 7 and sequences of

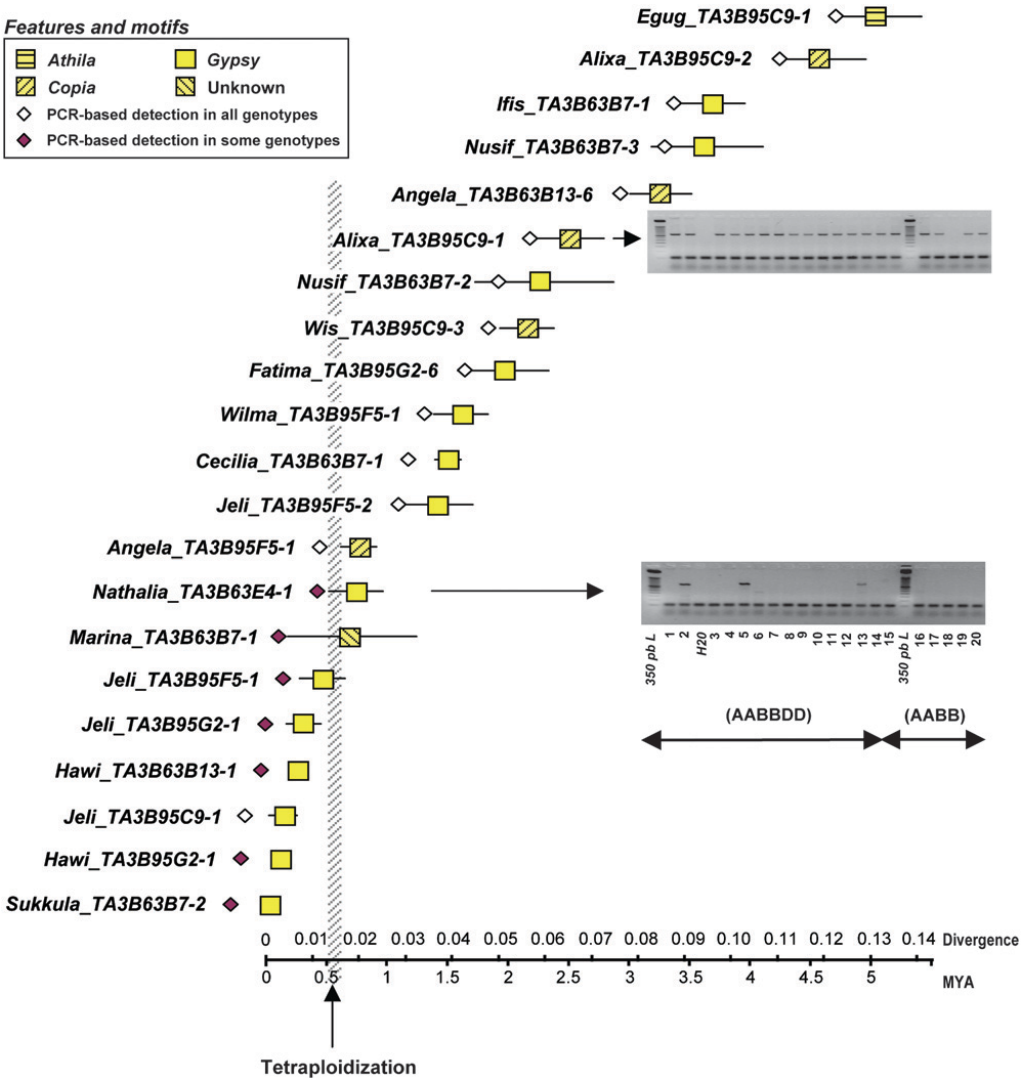


FIGURE 4.—PCR-based tracing of series of retrotransposons, inserted at different dates in wheat chromosome 3B across a collection of wheat tetraploid (*T. turgidum*) and hexaploid (*T. aestivum*) genotypes. Averages and intervals (means ±SE) of retrotransposon insertion dates are presented. The interval of the allotetraploidization event (0.5–0.6 MYA), calculated according to gene sequence divergence (HUANG *et al.* 2002; DVORAK *et al.* 2006; CHALUPSKA *et al.* 2008), is in gray. Retrotransposons for which insertion dates were estimated on the basis of divergence of their LTR prior to the tetraploidization event were generally detected in almost all genotypes, whereas those posterior to the tetraploidization event were detected in only some genotypes. Gels show PCR-based detection of insertion of RLC_*Alixa_TA3B95C9-1* (into DTC_*Caspar_TA3B95C9-1*) in all tested genotypes, except one and insertion of RLG_*Nathalia_TA3B63E4-1* (into DTC_*Vincent_TA3B63E4-2*) in some genotypes. Primer sequences, details of insertion dates (averages and intervals), and PCR-based detection in different wheat genotypes are given in supplemental Table 1 and supplemental Table 7. AABBDD (hexaploid wheat accessions): -1—*T. aestivum* cv. Renan; -2—*T. aestivum* cv. Chinese Spring; -3—*T. aestivum* spelta, Erge 27216; -4—*T. aestivum* spelta, Erge 2776; -5—*T. aestivum* spelta, Erge 2771; -6—*T. aestivum* spelta Rouquin, Erge 6329; -7—*T. aestivum* macha 1793, Erge 27240; -8—*T. aestivum* compactum rufulum 71V, Erge 26786; -9—*T. aestivum* compactum crebicum 72V, Erge 26787; -10—*T. aestivum* compactum clavatum 73V, Erge 26788; -11—*T. aestivum* compactum icterinum 74V, Erge 26789; -12—*T. aestivum* compactum erinaceum 75V, Erge 26790; -13—*T. aestivum* sphaerococcum tumidum perciv globosum, Erge 27016; -14—*T. aestivum* cv. Soisson. AABB (*T. turgidum*, tetraploid wheat accessions): -15—*T. turgidum* durum cv. Langdon; -16—*T. turgidum* durum; -17—*T. turgidum* dicoccum, -18—*T. turgidum* dicoccoides; -19—*T. turgidum* polinicum; -20—*T. turgidum* turgidum.

the PCR primers in supplemental Table 1. With the exception of *Jeli_TA3B95C9-1*, all the other 7 most recently inserted retrotransposons, which have calculated insertion date intervals (means ±SE) equal to or less than the 0.5–0.6 MYA interval (divergence 0.013–0.016), were detected in some but not all genotypes carrying the B genome, suggesting their occurrence after the tetraploidization event (Figure 4 and supplemental Table 7). In contrast, all 13 retrotransposon insertions, which have calculated insertion intervals (means ±SE) >0.7 MYA, were detected in all tested genotypes carrying the B genome, suggesting their occurrence prior to the allotetraploidization event (Figure 4 and supplemental Table 7). Given the uncertainty in calculating intervals of insertion dates, the PCR-based tracing method confirms the calibration of LTR divergence on that of gene divergence. More importantly, it also confirms that retrotranspositions (insertions) were not enhanced or repressed by the alloteraploidization event.

Relative proliferation periods of the *CACTA* class II transposable elements: The *CACTA* class II DNA TEs represent an important proportion of the B- and A-genome sequences (13.4 and 9.4%, respectively). As for the main LTR–retrotransposon superfamilies, ratios of complete to truncated copies are very different for B (0.37) and A (1.52) genomes (Table 1). In contrast to LTR retrotransposons, the *CACTA* TEs do not have long repeats or other features, which would allow determination of their insertion dates on the basis of sequence divergence. Therefore, their proliferation periods and rates were evaluated indirectly, relative to their level of insertions into or by other *CACTA* TEs and, more

TABLE 2

Associations of *CACTA* transposable elements with the four most represented TE superfamilies and other DNA sequence classes in 13 genomic regions of the wheat B genome and 19 publicly available genomic sequences of the wheat A genome

	13 genomic regions of the wheat B genome (1.98 Mb)[a]		19 publicly available genomic regions of the wheat A genome (3.63 Mb)[b]	
DNA sequence classes	*CACTA* TEs inserted into other DNA sequences[c]	Other DNA sequences inserted into *CACTA* TEs[c]	*CACTA* TEs inserted into other DNA sequences	Other DNA sequences inserted into *CACTA* TEs
Athila TEs	12: 4/8	6: 0/6	7: 7/0	0: 0/0
Copia TEs	1: 0/1	5: 0/5	10: 7/3	8: 6/2
Gypsy TEs	1: 1/0	6: 0/6	5: 4/1	6: 5/1
CACTA TEs	7: 4/3	6: 1/5	4: 4/0	3: 3/0
Other TEs	3: 3/0	0: 0/0	2: 1/1	0: 0/0
Unclear TE associations[d]	2: 0/2	2: 0/2	4: 0/4	4: 0/0
Unassigned DNA	44: 7/37	—	21: 9/12	—
Total	70: 19/51	25: 1/24	53: 32/21	21: 14/7

[a] This corresponds to 1.43 Mb from the 10 genomic regions sequenced in this study and 0.55 Mb from three other publicly available genomic regions from Sabot *et al.* (2005), Gu *et al.* (2006), and Dvorak *et al.* (2006). See materials and methods for BAC clone sequence references.

[b] Nineteen genomic regions available for the A genome (SanMiguel *et al.* 2002; Yan *et al.* 2002, 2003; Wicker *et al.* 2003b; Chantret *et al.* 2005; Isidore *et al.* 2005; Dvorak *et al.* 2006; Gu *et al.* 2006; Miller *et al.* 2006). See materials and methods for BAC clone sequence references.

[c] Results are as follows: total *CACTA* TE copies: complete *CACTA* TE copies/truncated *CACTA* TE copies.

[d] From cases where we cannot be certain that a *CACTA* TE is inserted into or by another TE element.

importantly, by elements of the three main LTR–retrotransposon superfamilies for which proliferation periods and rates were evaluated on the basis of the dates of insertions (described above). This was calculated for all *CACTA* TE copies as well as for complete and truncated copies separately (Table 2).

In the wheat B genome, the majority of *CACTA* TE insertions (mainly those detected as truncated copies) occurred in DNA annotated as unassigned (Table 2). For the rest, significantly higher insertions of *CACTA* TEs into *Athila* and other *CACTA* TEs than into *Copia* and *Gypsy* retrotransposons were observed. The two latter retrotransposon superfamilies were significantly more inserted into, rather than by, *CACTA* TEs (Table 2). These observations indicate that proliferation of the *CACTA* TEs in the B genome of wheat started before, and continued during and after *Athila* retrotransposon proliferation, whereas very few insertions occurred during the last waves of high proliferation of *Copia* and *Gypsy*.

Similarly, a high level of insertions into unassigned DNA was observed for the *CACTA* TEs in the A genome. However, for the remaining insertions, no clear period of proliferation could be determined as these show similar levels of insertions into or by all other TE superfamilies (Table 2). These observations, combined with the observed higher level of complete copies (Table 1), suggest that the *CACTA* TE proliferation continued in the wheat A genome during the last waves of proliferation of *Copia* and *Gypsy*, unlike those in the B genome.

DISCUSSION

To constitute representative genomic sequences of the wheat B genome, in this study we have sequenced 10 BAC clones of the chromosome 3B, representing the most important number of genomic regions sequenced for a single wheat chromosome and a cumulative sequence length of 1.429 Mb (0.15% of the chromosome length). As expected, TE proliferation was pronounced (representing 79.1%). Five of these were revealed as gene-containing BAC clones at a density of one or two genes per clone; two other BAC clones contain gene relics or pseudogenes, whereas the three remaining BAC clones were missing genes. This confirms the previous conclusion about the more random distribution of genes on the wheat genome (Devos *et al.* 2005). Interestingly and in comparison with rice, a high level of "truncated genes" was revealed [six gene relic or pseudogenes, several of which because of TE insertions (three confirmed cases)]. If the confirmed gene number (excluding hypothetical genes) identified in the 1.43-Mb sequences (eight) is extrapolated to the whole wheat chromosome 3B of 1 Gb estimated size, then 5594 genes might be present. A slightly higher number (6000) was calculated from BAC-end sequence analysis (Paux *et al.* 2006).

Representation of transposable elements: In this study, TE dynamics, proliferation, and evolutionary pathways were analyzed and compared in 1.98 Mb of sequence from 13 BAC clones of the wheat B genome and 3.63 Mb of sequence from 19 BAC clones of the

wheat A genome. These genomic sequences represent very small fractions (<0.03%) of their respective genomes. Nevertheless, it has been argued that, for studying abundant repeats, sequencing and annotation of a small proportion of the genome can be representative (Brenner *et al.* 1993; Vitte and Bennetzen 2006; Liu *et al.* 2007). We have been able to confirm the adequate representation where less variation in the proportion of the main TE superfamilies was observed when analyzing a large number of BAC clones (Figure 2). Interestingly, TE proportions observed in the 13 genomic regions of the B genome of wheat are similar to those obtained from 11 Mb of BAC-end sequences of wheat chromosome 3B (Paux *et al.* 2006). Similarly, TE proportions were not significantly different for the wheat A genome when they were compared with the different ploidy levels (see results).

Although they are representative of abundant wheat TEs available in the TREP database (Wicker *et al.* 2002; http://wheat.pw.usda.gov.ITMI/Repeats), the class I and class II TEs observed in the genomic sequences of the wheat B and A genomes may not cover all wheat TEs. It is expected that more wheat TEs will be identified, as more wheat genomic sequences will become available. This is particularly supported by the identification in this study of >21 different novel TE families, most of which (17) are retrotransposons. We also believe that low-copy TEs and those that tend to "compartmentalize" in specific regions, such as pericentromeric heterochromatin regions (which is not the case in our regions), would be missed, over-, or underrepresented in this study (Ma and Bennetzen 2006; Liu *et al.* 2007). This could be the case for the *CACTA* TEs, which show the highest variation in sequence proportion between regions because they tend to be clustered in the Triticeae genomes (our unpublished results and Wicker *et al.* 2003a, 2005).

Transposable elements proliferated differentially in the B and A genomes of wheat: Abundance of TEs varies widely across different organisms. Human (*Homo sapiens*) DNA is composed of 45% (Lander *et al.* 2001) repetitive sequences, *Drosophila melanogaster* of 3.9% (Kaminker *et al.* 2002), and maize of 67% (Haberer *et al.* 2005; Liu *et al.* 2007) whereas TE content in the wheat genomic sequences analyzed in this study or in other studies (Li *et al.* 2004; Gu *et al.* 2006; Paux *et al.* 2006) is ~80%. Proportions of different classes of TEs also vary among organisms. Class II TEs are almost >10 times less abundant than class I TEs and constitute a small fraction (<2%) of the human, rice (Piegu *et al.* 2006), maize (Kronmiller and Wise 2008), Arabidopsis, and cotton (Hawkins *et al.* 2006) genomes. In comparison, class II TE abundance is relatively high in the wheat B and A genomes (14.1 and 9.9%, respectively), the majority of which (95%) are *CACTA* TEs, which are particularly abundant in the Triticeae genomes (Wicker *et al.* 2003a, 2005). Class I retrotransposon abundance is relatively high in several plant genomes, 58.7 and 56.6% estimated in this study for the wheat B and A genomes, respectively; 40–50% in cotton species (Hawkins *et al.* 2006); 35–60% in rice species (Piegu *et al.* 2006); and 64% in maize (Liu *et al.* 2007; Kronmiller and Wise 2008).

In this study, combination of TE sequence analysis and classification, comparison of proportions of complete to incomplete copies, TE insertion date estimations, and PCR-based tracing of insertions allow us to compare TE proliferation periods and rates in the wheat B and A genomes (Figure 5). It is evident that TEs appear to proliferate differentially in waves of high activity followed by periods of low activity (Figure 5). Both genomes show similar rates and relatively old proliferation periods for the *Athila* retrotransposons (Figure 5). However, the *Copia* retrotransposons have proliferated relatively more recently in the A genome whereas a more recent *Gypsy* proliferation is observed in the B genome. Due to their biology and replication mechanism, it was not possible to directly estimate the *CACTA* class II TE insertion dates. We have estimated their proliferation periods and rates relative to that of the three main LTR retrotransposon superfamilies. In the wheat B genome, the *CACTA* TE high proliferation period started before and overlaps with that of the *Athila* retrotransposons. In the wheat A genome, in addition to the relatively old proliferation similar to that in the B genome, *CACTA* TEs continued to proliferate during the same period as *Gypsy* and *Copia* retrotransposons. Determining the ancient proliferation periods of CACTA TEs partially explains why *CACTA* TEs often tend to be clustered together (see results and Wicker *et al.* 2003a, 2005), although they were detected in almost all analyzed BAC clones. Differential proliferation of TEs provides a valid explanation for the size variation of closely related wheat genomes (Bennett and Smith 1976, 1991; http://data.kew.org/cvalues/homepage.html).

Four families (*Angela*, *Wis*, *Sabrina*, *Fatima*) were abundant, representing the majority of LTR retrotransposons in the B and A genomes of wheat, some of which proliferated differentially (see results). Proliferation of specific types of TEs in specific genomes (or species), leading to rapid genome size variation and sequence divergence, has also been observed in other plant species. Analysis of maize (*Zea mays*) genomic sequences suggests that the high percentage of LTR retrotransposons is due to proliferation of only a few families of TEs (Meyers *et al.* 2001; Liu *et al.* 2007; Kronmiller and Wise 2008). Similarly, comparison of TE proportions between various cotton species (Gossypium species) revealed differential lineage-specific expansion of various LTR–retrotransposon superfamilies and families, leading to threefold genome size differences (Hawkins *et al.* 2006). Species-specific differential retrotransposon expansions are also the

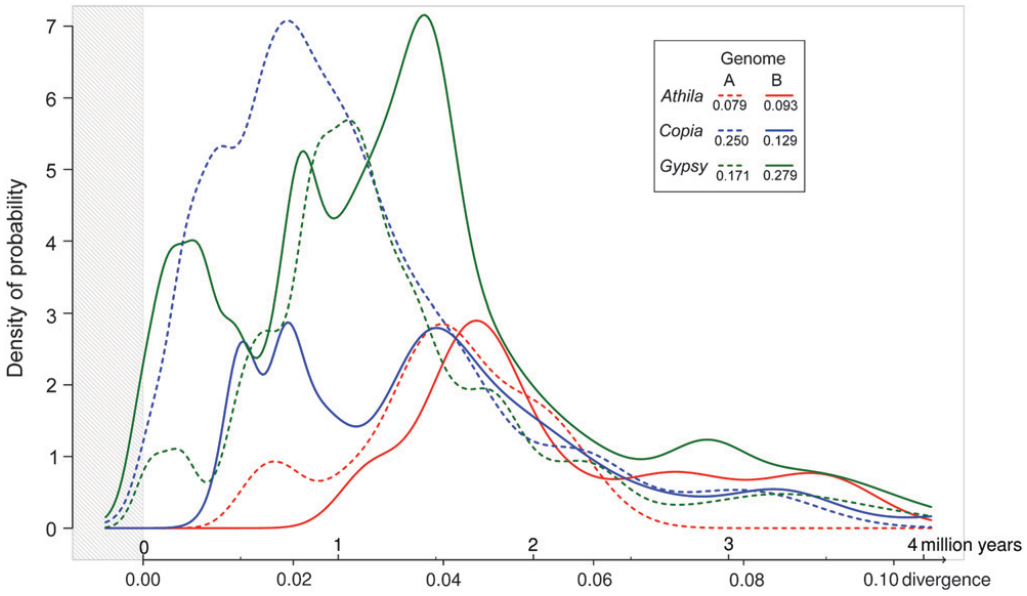


FIGURE 5.—Proliferation periods and rates of the main retrotransposon superfamilies in the wheat B and A genomes. Expressed as probability density functions, where the area under each curve was calculated on the basis of the estimated insertion dates of retrotransposons (in Figure 3) and their corresponding standard errors, using Gaussian kernel density estimation (SILVERMAN 1986). The curves have been scaled with respect to the number of observations, so that the sum of their areas (given for each retrotransposon superfamily in the key) equals the probability of 1 and comparisons between genomes and retrotransposon superfamilies can be performed. When calculated standard errors were very low, a minimum value of 80,000 years (corresponding to 0.002 divergence) was used. The shaded field is due to uncertainty in very recent insertion date estimations.

cause of the size doubling of the *Oryza australiensis* genome as compared to cultivated rice (*O. sativa*) (PIEGU *et al.* 2006).

This is the first time that dynamics as well as proliferation periods and rates of TEs have been compared between two closely related wheat genomes. This was possible only because in this study we sequenced 10 different genomic regions that constituted a genomic sequence data set representative of the wheat B genome. For the wheat A genome, more representative genomic sequence data were rendered publicly available. There have been initial attempts to evaluate TE proliferation in the wheat genomes. LI *et al.* (2004) analyzed the D genome of the diploid *Ae. tauschii* and showed that the copy number of most TEs have increased gradually following polyploidization. However, they used dot blots, which are not very accurate. SABOT *et al.* (2005) have updated TE annotation in wheat genomic sequences and reported their composition and distribution in relation to genes. They suggested that *Copia* TEs have been most active in the wheat A, B, and D genomes, combined together (SABOT *et al.* 2005). Accurate comparison of dynamics as well as proliferation periods and rates between individual genomes of wheat could not be conducted in the study of SABOT *et al.* (2005) as, in the genomic sequences available at that time, the A genome was overrepresented whereas the B genome was underrepresented. By using more representative genomic sequences in this study, we showed the more recent activation of the *Copia* and *CACTA* TEs in the wheat A genome but not in the B genome in which a more recent *Gypsy* proliferation is observed. Overrepresentation of the A-genome sequences in the study of SABOT *et al.* (2005) may explain the reason why they found that *Copia* TEs have been most active in the wheat A, B, and D genomes combined together. Thus our analysis, using representative sequence data sets, for the first time shows differential proliferation of TEs between the wheat A and B genomes and illustrates the inadequacy of combining sequence data sets from different genomes as was previously done.

Neither enhancement nor repression of transposable element proliferation following allotetraploidization: As estimated from their insertion dates and confirmed by PCR-based tracing analysis, the majority of the differential proliferation of TEs in B and A genomes of wheat (87% and 83, respectively) occurred prior to the allotetraploidization event that brought them together in *T. turgidum* and *T. aestivum* <0.5 MYA (HUANG *et al.* 2002; DVORAK *et al.* 2006; CHALUPSKA *et al.* 2008). More importantly, the allotetraploidization event appears to have neither enhanced nor repressed retrotranspositions. We suggest that, in addition to the *Ph1* gene preventing homeologous recombination (GRIFFITHS *et al.* 2006), differential proliferation of TEs has also contributed to the rapid divergence of the B and A genomes of the wheat diploid progenitors and the relative stability of the natural wheat allopolyploids that occurs thereafter.

Different levels of stability, estimated as elimination of DNA sequences, were observed in newly synthesized wheat allopolyploids, depending on wheat genome combinations (FELDMAN and LEVY 2005 and our unpublished results). The natural wheat allopolyploids combining the B and A genomes are relatively stable

and cannot be exactly resynthesized because the diploid progenitor of the B genome is unidentified (Feldman *et al.* 1995; Blake *et al.* 1999; Huang *et al.* 2002; Dvorak *et al.* 2006). Nevertheless, by studying a synthetic wheat allotetraploid combining the A and S genomes (the closest identified diploid relatives to the progenitors of the A and of the B genomes of natural wheat polyploids), Kashkush *et al.* (2003) reported on transcriptional activation of the *Wis* LTR retrotransposon but not its transposition following allotetraploidization. This is in agreement with the lack of enhancement of transpositions observed in this study in wheat natural allopolyploids combining the A and B genomes. Comparatively, less TE proliferation, estimated as the increased rate of deletions and the decreased rate of insertions, was recently observed in the cotton polyploid species *Gossypium hirsutum* as compared to its diploid progenitors *Gossypium arboretum* and *Gossypium raimondii* (Grover *et al.* 2008).

Apparent transposable element proliferation as a balance between two evolutionary forces: TEs "transposition" and also their removal: As in this study, the vast majority of complete retrotransposons studied so far were also estimated to be <3 million years old (SanMiguel *et al.* 1998, 2002; Wicker *et al.* 2003b, 2005; Gao *et al.* 2004; Ma *et al.* 2004; Du *et al.* 2006; Piegu *et al.* 2006; Wicker and Keller 2007). These findings imply that there are mechanisms of active deletion of LTR retrotransposons from the genome, such as unequal homologous recombination and illegitimate recombination (Vicient *et al.* 1999; Devos *et al.* 2002; Ma *et al.* 2004; Pereira 2004). Proliferation periods and rates estimated for TEs at a given evolutionary period are the result of both antagonist evolutionary forces: TE insertion activity (transpositions) (Bennetzen and Kellogg 1997) and the removal of TEs (Petrov *et al.* 2000; Petrov 2002a). Thus, it is not clear whether the insertions and/or truncation (removal) rates of TEs are constant or vary during genome evolution. The "burst of insertions" described for TEs could correspond to periods of (i) high insertion activity, (ii) low rates of TE removal, and/or (iii) combinations of both evolutionary forces.

The fact that *Copia* retrotransposons have been active until recently in the *Arabidopsis thaliana* genome allowed Pereira (2004) to calculate the rate of their elimination (or half-life) as 472,000 years, outside of centromeric regions. Using this method and assuming that repetitive sequences are removed from the genome at a constant rate, a higher half-life (79,000 years) was calculated for *Copia* removal in rice (Wicker and Keller 2007). As the insertion-date distribution of *Copia* retrotransposons in Triticeae (wheat and barley) is not exponential, Wicker and Keller (2007) suggested that their half-life is much longer than in rice, thus representing a major difference between small and large genomes of plants. Similar distributions are observed in our study for all three retrotransposon superfamilies in both B and A genomes of wheat. Our analysis suggests that lower proliferation of the LTR retrotransposons during the most recent period could account for these apparent nonexponential distributions of insertion dates (including *Copia* retrotransposons) (Figure 5).

Our study clearly shows that, during their evolution, specific types of TEs have undergone differential proliferation in specific wheat genomes (or species) but not in others, leading to rapid sequence divergence. Little is known about the mechanistic causes that lead to differential proliferation of a single or related group of TEs across the genome of a specific species. These rapid TE expansions could correspond to periods of relaxed selection pressure such as genome duplication, interspecific hybridizations (although this was not revealed in our study), or stress conditions. It is also possible that TE proliferation could be caused by advantageous mutations in the TE sequence. A third alternative is differential deregulation of epigenetic silencing that allows specific TE families to proliferate in specific genomes.

We sincerely thank J. Dolezel and M. Kubalakova (Institute of Experimental Botany, Olomouc, Czech Republic) for providing FISH mapping information for BAC clones B95G2, B95C9, B63B7, and B54F7; Joseph Jahier [Institut National de la Recherche Agronomique (INRA), Rennes, France] and Moshe Feldman (Weizmann Institute of Science) for valuable discussions and for providing wheat genotypes; Catherine Feuillet (INRA, Clermont-Ferrand, France) for providing the wheat deletion lines; Thomas Wicker (Zurich University) for valuable advice on novel transposable element classifications and *CACTA* TE evolution; Piotr Gornicki (University of Chicago) and anonymous reviewers for valuable discussion and constructive criticisms; and Heather McKhann (Centre National de Génotypage, Etude du Polymorphisme Génomique Vegetal-INRA, Evry, France) for valuable discussion and revision of the manuscript. This project was supported by the National Center for Sequencing (Centre National de Séquençage-Génoscope)/APCNS2003-Project: Triticum species comparative genome sequencing in wheat (http://www.genoscope.cns.fr/externe/English/). PCR-based tracing of retrotransposons insertions was funded by the Agence Nationale pour la Recherche Biodiversité Project (ANR-05-BDIV-015) and the ANR-05-Blanc project-ITEGE.

LITERATURE CITED

Adams, K. L., and J. F. Wendel, 2005 Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. **8:** 135–141.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Bennett, M. D., and I. J. Leitch, 1997 Nuclear DNA amounts in angiosperms: 583 new estimates. Ann. Bot. **80:** 169–196.

Bennett, M. D., and I. J. Leitch, 2005 Plant genome size research: a field in focus. Ann. Bot. **95:** 1–6.

Bennett, M. D., and J. B. Smith, 1976 Nuclear DNA amounts in angiosperms. Philos. Trans. R. Soc. Lond. B Biol. Sci. **274:** 227–274.

Bennett, M. D., and J. B. Smith, 1991 Nuclear DNA amounts in angiosperms. Philos. Trans. R. Soc. Lond. B Biol. Sci. **334:** 309–345.

Bennetzen, J. L., 2000a Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell **12:** 1021–1029.

Bennetzen, J. L., 2000b Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. **42:** 251–269.

Bennetzen, J. L., 2002a Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115:** 29–36.
Bennetzen, J. L., 2002b The rice genome: opening the door to comparative plant biology. Science **296:** 60–63.
Bennetzen, J. L., and E. A. Kellogg, 1997 Do plants have a one-way ticket to genomic obesity? Plant Cell **9:** 1509–1514.
Bennetzen, J. L., J. Ma and K. M. Devos, 2005 Mechanisms of recent genome size variation in flowering plants. Ann. Bot. **95:** 127–132.
Blake, N. K., B. R. Lehfeldt, M. Lavin and L. E. Talbert, 1999 Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat. Genome **42:** 351–360.
Blanc, G., A. Barakat, R. Guyot, R. Cooke and M. Delseny, 2000 Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12:** 1093–1101.
Brenner, S., G. Elgar, R. Sandford, A. Macrae, B. Venkatesh *et al.*, 1993 Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature **366:** 265–268.
Chalupska, D., H. Y. Lee, J. D. Faris, A. Evrard, B. Chalhoub *et al.*, 2008 Acc homoeoloci and the evolution of wheat genomes. Proc. Natl. Acad. Sci. USA **105:** 9691–9696.
Chantret, N., J. Salse, F. Sabot, S. Rahman, A. Bellec *et al.*, 2005 Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). Plant Cell **17:** 1033–1045.
Chooi, W. Y., 1971 Variation in nuclear DNA content in the genus Vicia. Genetics **68:** 195–211.
Devos, K. M., J. K. Brown and J. L. Bennetzen, 2002 Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. **12:** 1075–1079.
Devos, K. M., J. Ma, A. C. Pontaroli, L. H. Pratt and J. L. Bennetzen, 2005 Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. Proc. Natl. Acad. Sci. USA **102:** 19243–19248.
Dolezel, J., M. Kubalakova, J. Bartos and J. Macas, 2004 Flow cytogenetics and plant genome mapping. Chromosome Res. **12:** 77–91.
Du, C., Z. Swigonova and J. Messing, 2006 Retrotranspositions in orthologous regions of closely related grass species. BMC Evol. Biol. **6:** 62.
Dvorak, J., P. Diterlizzi, H.-B. Zhang and P. Resta, 1993 The evolution of polyploid wheats: identification of the A genome donor species. Genome **36:** 21–31.
Dvorak, J., E. D. Akhunov, A. R. Akhunov, K. R. Deal and M. C. Luo, 2006 Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. Mol. Biol. Evol. **23:** 1386–1396.
Efron, B., 1979 Bootstrap methods: another look at the jackknife. Ann. Stat. **7:** 1–26.
Feldman, M., and A. A. Levy, 2005 Allopolyploidy: a shaping force in the evolution of wheat genomes. Cytogenet. Genome Res. **109:** 250–258.
Feldman, M., F. G. H. Lupton and T. E. Miller, 1995 Wheats, pp.184–192 in *Evolution of Crops*, Ed. 2, edited by J. Smartt and N. W. Simmonds. Longman Scientific, London.
Férignac, P., 1962 Test de Kolmogorov-Smirnov sur la validité d'une fonction de distribution. Rev. Stat. Appl. **10:** 13–32.
Flavell, A. J., M. R. Knox, S. R. Pearce and T. H. Ellis, 1998 Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J. **16:** 643–650.
Gao, L., E. M. McCarthy, E. W. Ganko and J. F. McDonald, 2004 Evolutionary history of Oryza sativa LTR retrotransposons: a preliminary survey of the rice genome sequences. BMC Genomics **5:** 18.
Graner, A., H. Siedler, A. Jahoor, R. G. Herrman and G. Wenzal, 1990 Assessment of the degree and the type of restriction fragment length polymorphism in barley (*Hordeum vulgare*). Theor. Appl. Genet. **80:** 826–832.
Griffiths, S., R. Sharp, T. N. Foote, I. Bertin, M. Wanous *et al.*, 2006 Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. Nature **439:** 749–752.
Grover, C. E., Y. Yu, R. A. Wing, A. H. Paterson and J. F. Wendel, 2008 A phylogenetic analysis of indel dynamics in the cotton genus. Mol. Biol. Evol. **25:** 1415–1428.
Gu, Y. Q., J. Salse, D. Coleman-Derr, A. Dupin, C. Crossman *et al.*, 2006 Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. Genetics **174:** 1493–1504.
Haberer, G., S. Young, A. K. Bharti, H. Gundlach, C. Raymond *et al.*, 2005 Structure and architecture of the maize genome. Plant Physiol. **139:** 1612–1624.
Hawkins, J. S., H. Kim, J. D. Nason, R. A. Wing and J. F. Wendel, 2006 Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res. **16:** 1252–1261.
Huang, S., A. Sirikhachornkit, X. Su, J. Faris, B. Gill *et al.*, 2002 Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proc. Natl. Acad. Sci. USA **99:** 8133–8138.
Isidore, E., B. Scherrer, B. Chalhoub, C. Feuillet and B. Keller, 2005 Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels. Genome Res. **15:** 526–536.
Jaillon, O., J. M. Aury, B. Noel, A. Policriti, C. Clepet *et al.*, 2007 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467.
Jones, R. N., and L. M. Brown, 1976 Chromosome evolution and DNA variation in *Crepis*. Heredity **36:** 91–104.
Kaminker, J. S., C. M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas *et al.*, 2002 The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. **3:** RESEARCH0084.
Kashkush, K., M. Feldman and A. A. Levy, 2003 Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat. Genet. **33:** 102–106.
Kidwell, M. G., 2002 Transposable elements and the evolution of genome size in eukaryotes. Genetica **115:** 49–63.
Kimura, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.
Kronmiller, B. A., and R. P. Wise, 2008 TE nest: automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol. **146:** 45–59.
Kumar, S., K. Tamura and M. Nei, 2004 MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. **5:** 150–163.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.
Li, W., P. Zhang, J. P. Fellers, B. Friebe and B. S. Gill, 2004 Sequence composition, organization, and evolution of the core Triticeae genome. Plant J. **40:** 500–511.
Liu, R., C. Vitte, J. Ma, A. A. Mahama, T. Dhliwayo *et al.*, 2007 A GeneTrek analysis of the maize genome. Proc. Natl. Acad. Sci. USA **104:** 11844–11849.
Ma, J., and J. L. Bennetzen, 2004 Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl. Acad. Sci. USA **101:** 12404–12410.
Ma, J., and J. L. Bennetzen, 2006 Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc. Natl. Acad. Sci. USA **103:** 383–388.
Ma, J., K. M. Devos and J. L. Bennetzen, 2004 Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. **14:** 860–869.
Meyers, B. C., S. V. Tingey and M. Morgante, 2001 Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. **11:** 1660–1676.
Miller, A. K., G. Galiba and J. Dubcovsky, 2006 A cluster of 11 CBF transcription factors is located at the frost tolerance locus Fr-Am2 in Triticum monococcum. Mol. Genet. Genomics **275:** 193–203.
Paterson, A. H., J. E. Bowers and B. A. Chapman, 2004 Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc. Natl. Acad. Sci. USA **101:** 9903–9908.
Paux, E., D. Roger, E. Badaeva, G. Gay, M. Bernard *et al.*, 2006 Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J. **48:** 463–474.

Pereira, V., 2004 Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. Genome Biol. **5:** R79.
Petrov, D. A., 2002a Mutational equilibrium model of genome size evolution. Theor. Popul. Biol. **61:** 531–544.
Petrov, D. A., 2002b DNA loss and evolution of genome size in Drosophila. Genetica **115:** 81–91.
Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl and K. L. Shaw, 2000 Evidence for DNA loss as a determinant of genome size. Science **287:** 1060–1062.
Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Saniyal *et al.*, 2006 Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. **16:** 1262–1269.
Qi, L., B. Echalier, B. Friebe and B. S. Gill, 2003 Molecular characterization of a set of wheat deletion stocks for use in chromosome bin mapping of ESTs. Funct. Integr. Genomics **3:** 39–55.
Rozen, S., and H. Skaletsky, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. **132:** 365–386.
Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice *et al.*, 2000 Artemis: sequence visualization and annotation. Bioinformatics **16:** 944–945.
Sabot, F., R. Guyot, T. Wicker, N. Chantret, B. Laubin *et al.*, 2005 Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. Mol. Genet. Genomics **274:** 119–130.
Safar, J., J. Bartos, J. Janda, A. Bellec, M. Kubalakova *et al.*, 2004 Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. Plant J. **39:** 960–968.
Salse, J., S. Bolot, M. Throude, V. Jouffe, B. Piegu *et al.*, 2008 Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell **20:** 11–24.
SanMiguel, P., A. Tikhonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768.
SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima and J. L. Bennetzen, 1998 The paleontology of intergene retrotransposons of maize. Nat. Genet. **20:** 43–45.
SanMiguel, P. J., W. Ramakrishna, J. L. Bennetzen, C. S. Busso and J. Dubcovsky, 2002 Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). Funct. Integr. Genomics **2:** 70–80.
Sasaki, T., W. Jianzhong, T. Itoh and T. Matsumoto, 2005 [Complete rice genome sequence information: the key for elucidation of Rosetta stones of other cereal genome] Tanpakushitsu Kakusan Koso **50:** 2167–2173.
Silverman, B. W., 1986 *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London/New York.
Smith, D., and R. Flavell, 1975 Characterization of the wheat genome by renaturation kinetics. Chromosoma **50:** 223–242.
Stein, N., 2007 Triticeae genomics: advances in sequence analysis of large genome cereal crops. Chromosome Res. **15:** 21–31.
Vedel, F., and M. Delseny, 1987 Repetivity and variability of higher plant genomes. Plant Physiol. Biochem. **25:** 191–210.
Vicient, C. M., A. Suoniemi, K. Anamthawat-Jonsson, J. Tanskanen, A. Beharav *et al.*, 1999 Retrotransposon BARE-1 and its role in genome evolution in the genus Hordeum. Plant Cell **11:** 1769–1784.
Vitte, C., and J. L. Bennetzen, 2006 Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc. Natl. Acad. Sci. USA **103:** 17638–17643.
Wendel, J. F., R. C. Cronn, J. S. Johnston and H. J. Price, 2002 Feast and famine in plant genomes. Genetica **115:** 37–47.
Wicker, T., and B. Keller, 2007 Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res. **17:** 1072–1081.
Wicker, T., D. Matthews and B. Keller, 2002 TREP: a database for Triticeae repetitive elements. Trends Plant. Sci. **7:** 561–562.
Wicker, T., R. Guyot, N. Yahiaoui and B. Keller, 2003a CACTA transposons in Triticeae: a diverse family of high-copy repetitive elements. Plant Physiol. **132:** 52–63.
Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z. D. Liu *et al.*, 2003b Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. Plant Cell **15:** 1186–1197.
Wicker, T., W. Zimmermann, D. Perovic, A. H. Paterson, M. Ganal *et al.*, 2005 A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley Hv-eIF4E locus: recombination, rearrangements and repeats. Plant J. **41:** 184–194.
Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy *et al.*, 2007 A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8:** 973–982.
Yan, L., V. Echenique, C. Busso, P. SanMiguel, W. Ramakrishna *et al.*, 2002 Cereal genes similar to Snf2 define a new subfamily that includes human and mouse genes. Mol. Genet. Genomics **268:** 488–499.
Yan, L., A. Loukoianov, G. Tranquilli, M. Helguera, T. Fahima *et al.*, 2003 Positional cloning of the wheat vernalization gene VRN1. Proc. Natl. Acad. Sci. USA **100:** 6263–6268.
Zuccolo, A., A. Sebastian, J. Talag, Y. Yu, H. Kim *et al.*, 2007 Transposable element distribution, abundance and role in genome size variation in the genus Oryza. BMC Evol. Biol. **7:** 152.

Communicating editor: J. A. Birchler

III Résultats complémentaires

III.1 'Supplemental Data' en ligne de l'Article 1

En complément du texte principal, l'Article 1 présente également de nombreux résultats complémentaires, sous la forme de 'Supplemental Data', mises à disposition sur le site de l'éditeur. J'ai repris ici ces résultats en les présentant selon trois axes : l'annotation des séquences des 10 clones BAC du chromosome 3B, la cartographie des clones BAC par hybridation cytogénétique *in situ* (FISH) et les résultats des analyses PCR portant sur le traçage de la dynamique insertionnelle des TEs dans une collection de lignées de blés et sur le 'mapping' des clones BACs sur les BINs de délétions du chromosome 3B.

III.1.1 Annotation des séquences de 10 clones BAC du chromosome3B

Les séquences des 10 clones BAC du chromosome 3B ont été annotées selon la méthode décrite dans le chapitre 'Matériels et Méthodes d'annotation' de la thèse. Ces 'Supplemental Data' présentent le détail de l'ensemble de ces annotations et des résultats correspondant aux niveaux des gènes (**Supplemental Text 1**) et des TEs (**Supplemental Text 2**). Ils s'appuient sur des tableaux présentant la proportion précise des différentes types de séquences (**Supplemental Table 2**), le détail de l'annotation des gènes et de leur classification (**Supplemental Table 3** et **Supplemental Table 4**) et la caractérisation des nouveaux TEs identifiés dans ces séquences (**Supplemental Table 5**).

DNA Sequence classes	Count (number)	Cumulative size (bp)	Proportion (%)
Transposable elements			
ClassI	171	884,291	61.9%
Athila	37	154,791	10.8%
Copia	45	209,605	14.7%
Gypsy	57	439,932	30.8%
LTR (unknown)	10	46,259	3.2%
LINE	17	31,402	2.2%
Other (unknown)	5	2,302	0.2%
Class II	113	231,687	16.2%
CACTA	54	221,563	15.5%
MITE	56	7,156	0.5%
LITE	3	2,968	0.2%
Unclassified	18	14,271	1.0%
Other Repeat	**58**	**34,373**	**2.4%**
Gene related sequences (GRS)	**23**	**13,826**	**1.0%**
Known	3	3,811	0.3%
Unknown	3	2,157	0.2%
Putative	2	1,066	0.1%
Pseudogenes	1	798	0.1%
Relics	4	1,371	0.1%
Hypothetical	10	4,623	0.3%
Unassigned DNA		**250,376**	**17.5%**
Total Length		**1,432,824**	

Supplemental Table 2A. Overall sequence class proportion of 10 sequenced BAC clones of wheat chromosome 3B.

Supplemental Text 1. Sequence annotation of 10 BAC clones from wheat chromosome 3B: Gene prediction, description and synteny with rice.

We conducted gene prediction analysis for the remaining 18.5% non-TEs and non-repeated DNA, using different search programs (see Supplemental Method 1 for detailed annotation method). *Genes of known and unknown functions*, or putative genes were defined based on predictions and the existence of rice or other *Triticeae* homologs. *Hypothetical genes* were identified based on prediction programs only. *Pseudogenes* were not well predicted and frameshifts need to be introduced within the CDS structure to better fit a putative function based on BLASTX (mainly with rice). *Truncated pseudogenes* (genes disrupted by large insertion or deletion) and highly degenerated CDS sequences were considered as *gene-relics*.

Combined together, all these types of gene sequence information (GSI) account for only 1.0% of the sequence and are present in seven BAC clones (one or two genes per clone) while the remaining three BAC clones (TA3B95C9, TA3B95G2, TA3B63N2) contain no genes (indicated in Figure 1A, Supplemental Table 3 and Supplemental Table 4).

Six genes (of known and unknown functions), and 2 putative genes were detected on 5 of the BAC clones (indicated on Figure 1A and detailed in Supplemental Table 3): BAC clone TA3B63B13 contains two genes of known functions, one of which was incompletely sequenced (located on the end of the BAC clone), BAC clone TA3B81B7 one putative gene, BAC clone TA3B95F5 one putative and two other genes of unknown functions, BAC clone TA3B63C11 one known gene and BAC clone TA3B63E4 one incompletely sequenced gene of unknown function.

In addition to genes (of known or unknown functions) and putative genes, the search for sequence homologies between the whole 18.5% non-TE and non-repeated DNA sequences and the rice genome sequence (http://www.tigr.org/tdb/e2k1/osa1/), allowed us to detect several conserved sequences between wheat and rice. As summarized, one pseudogene and four gene-relics detected in (respectively) the BAC clones TA3B54F7 (one pseudogene), TA3B63B7 (two gene-relics), TA3B81B7 (one gene-relic) and TA3B63C11 (one gene-relic) (Supplemental Table 3), could not be predicted with the CDS prediction program (FGENESH), as they show frameshifts, stop mutations, TE

BAC clone	Start ORF position	Stop ORF position	Best BLASTx homologies	Classification (relative function)	Additional information
TA3B54F7	19,392	41,098	**67%** on all length ($e=10^{-74}$) with Rice *OS09G08190*	Pseudogene (Sulfotransferase)	Two insertions of TEs 55% BLASTx ($e=10^{-39}$) *with OS01G20950*
TA3B63B13	71,534	74,358	**81%** on all length ($e=10^{-68}$) with Rice *OS01G07850*	Gene of known function (Glyoxalase)	-
	162,472	164,455	**94%** on all length (e=0) with Rice *OS01G07870*	Gene of known function (ABC transporter)	Incompletely sequenced gene
TA3B63B7	87,799	88,372	**55%** on 75 a.a. ($e=10^{-7}$) with Rice *OS03G09850*	Gene-relic (Beta-mono-oxygenase)	-
	92,763	92,963	**93%** on all length ($e=10^{-24}$) with Wheat *maturaseK*	Gene-relic (Maturase K)	Truncated by a class II MITE TE
TA3B81B7	8,909	10,055	**53%** on 125 a.a ($e=10^{-7}$) with Rice *OS01G41880*	Putative gene	-
	19,057	19,170	**86%** on all length ($e=10^{-12}$) with Rice *OS01G55940*	Gene-relic (GH3 auxine responser)	Truncated by an unclassified TE
TA3B95F5	190,794	191,264	**50%** on 3' 110 a.a ($e=10^{-7}$) with Rice OS05G12000	Putative gene	-
	208,034	208,723	**65%** on all length ($e=10^{-55}$) with Rice *OSA01G0930500*	Gene of unknown function	-
	222,584	223,264	**70%** on 5' 180 a.a ($e=10^{-48}$) with Rice *OS10G10700*	Gene of unknown function	-
TA3B63C11	65	536	**69%** on all length ($e=10^{-36}$) with Rice *OSA11G0306400*	Gene-relic (O-Methyltransferase)	Probably an incomplete gene duplication 54% BLASTx ($e=10^{-15}$) *with OS01G54970*
	13,637	15,095	**81%** on all length ($e=10^{-145}$) with Rice *OSA11G0306400*	Gene of known function (O-Methyltransferase)	57% BLASTx ($e=10^{-72}$) *with OS01G54970*
TA3B63E4	1	1,963	**80%** on all length ($e=10^{-101}$) with Rice *OS04G55970*	Gene of unknown function	Incompletely sequenced gene 92% BLASTx ($e=10^{-57}$) *with OS01G67410*

Supplemental Table 3. Details of complete and truncated genes predicted from 10 sequenced BAC clones of wheat chromosome 3B and their classification (according to criteria developed in Supplemental Method 1)

Genomic regions (BAC clones)	Start position	Stop position	ORF
TA3B54F7	129,689	130,084	115 a.a
TA3B63B13	83,789	84,424	106 a.a
	115,241	116,350	133 a.a
TA3B81B7	22,440	22,934	164 a.a
TA3B95F5	99,613	99,981	122 a.a
	163,372	163,968	198 a.a
	201,662	202,147	161 a.a
	204,781	205,206	141 a.a
TA3B63C11	2,514	5,586	197 a.a
	12,926	13,510	194 a.a

Supplemental Table 4. Hypothetical genes (predicted based on FGENESH prediction only) found in 10 sequenced BAC clones of wheat chromosome 3B.

insertions and/or large indels, and are probably no longer functional (Supplemental Table 2). Three of these five truncated genes (pseudogenes and gene-relics) have resulted from TEs insertions (Supplemental Table 3).

The wheat chromosome 3B is homologous to the rice chromosome 1. For orthology and synteny analysis, we considered the rice chromosome 1 and its duplicated segments that are found on other chromosomes (GUYOT *et al.* 2004 and TIGR site http://www.tigr.org/tdb/e2k1/osa1/segmental_dup/). Three BAC clones (TA3B63B13, TA3B81B7, TA3B95F5) have one or two of their orthologous rice genes that can be mapped on the rice chromosome 1 and were considered as confirmed in their synteny (Table 1). It is interesting to note that the two genes of known functions, separated by 88,114 bp on the BAC clone TA3B63B13 (Figure 1A) have their respective orthologs separated by 22,816 bp on rice chromosome 1. Thus, for this intergenic region, there is four-fold size difference between rice and wheat since their divergence from a common ancestor. Three other BAC clones (TA3B54F7, TA3B63C11 and TA3B63E4) also have homologs on rice chromosome 1, but the best match was observed with genes mapped on other rice chromosomes (Supplemental Table 3). BAC clone TA3B63B7 shows, for its putative gene and pseudogene, homologies with rice genes located on rice chromosome other than chromosome 1 (Supplemental Table 3).

No GSI or orthologous rice regions could be assigned to the three remaining BAC clones (TA3B95C9, TA3B95G2, TA3B63N2).

Finally 10 hypothetical genes were identified based on gene prediction only in the BAC clones TA3B54F7 (one), TA3B63B13 (two), TA3B81B7 (one), TA3B95F5 (four), TA3B63C11 (two) (Supplemental Table 4).

	TA3B54F7 (190,249bp)			TA3B63B13 (164,504bp)			TA3B63C11 (21,231bp)			TA3B95C9 (246,833 bp)			TA3B95G2 (177,914 bp)		
DNA sequence classes	**Count**	**Size (bp)**	**(%)**	**Count**	**Size (bp)**	**(%)**	**Count**	**Size (bp)**	**(%)**	**Count**	**Size (bp)**	**(%)**	**Count**	**Size (bp)**	**(%)**
Transposable elements															
Class I	16	73,914	38.9%	23	112,838	68.6%	0	0	0.0%	32	173,594	70.3%	16	102,213	57.5%
Athila	2	3,787	2.0%	5	19,295	11.7%	0	0	0.0%	10	52,446	21.2%	5	23,448	13.2%
Copia	3	12,896	6.8%	10	48,952	29.8%	0	0	0.0%	10	42,774	17.4%	1	1,747	1.0%
Gypsy	8	49,136	25.8%	3	38,096	23.2%	0	0	0.0%	9	68,413	27.7%	10	77,018	43.3%
LTR-Unknown	2	2,005	1.1%	0	0	0.0%	0	0	0.0%	3	9,961	4.0%	0	0	0.0%
LINE	1	6,090	3.2%	5	6,495	3.9%	0	0	0.0%	0	0	0.0%	0	0	0.0%
Others (unknown)	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%
Class II	24	78,450	41.2%	14	2,348	1.4%	0	0	0.0%	19	37,766	15.3%	16	37,831	21.3%
CACTA	15	77,056	40.5%	1	427	0.3%	0	0	0.0%	13	36,929	14.9%	6	36,631	19.6%
MITE	8	901	0.5%	12	1,662	1.0%	0	0	0.0%	6	837	0.3%	10	1,200	0.6%
LITE	1	493	0.3%	1	259	0.2%	0	0	0.0%	0	0	0.0%	0	0	0.0%
Unclassified	3	5,651	3.0%	8	2081	1.3%	0	0	0.0%	0	0	0.0%	2	1,873	1.1%
Other Repeat	6	5,455	2.9%	9	1,606	1.0%	8	2,505	11.8%	11	6,932	2.8%	9	4,110	2.3%
														4,110	2.3%
All Repeat	**49**	**163,470**	**85.9%**	**54**	**118,873**	**72.3%**	**8**	**2,505**	**11.8%**	**62**	**218,292**	**88.4%**	**43**	**146,027**	**82.1%**
Gene related sequences	**2**	**1,146**	**0.6%**	**4**	**3,439**	**2.1%**	**4**	**2,745**	**13.0%**	**0**	**0**	**0.0%**	**0**	**0**	**0.0%**
Known	0	0	0.0%	2	2,716	1.6%	1	1,095	5.2%	0	0	0.0%	0	0	0.0%
Unknown	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%
Putative	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%
Pseudo	1	798	0.4%	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%
Relic	0	0	0.0%	0	0	0.0%	1	471	2.2%	0	0	0.0%	0	0	0.0%
Hypothetical	1	348	0.2%	2	723	0.4%	2	1,179	5.6%	0	0	0.0%	0	0	0.0%
Unassigned DNA		**25,633**	**13.5%**		**42,192**	**25.6%**		**15,971**	**75.2%**		**28,541**	**11.6%**		**31,887**	**17.9%**

Supplemental Table 2B. Detailled sequence class proportion of 10 sequenced BAC clones of wheat chromosome 3B.

Supplemental Text 2. Sequence annotation of 10 BAC clones from wheat chromosome 3B: Transposable element prediction, annotation, classification and composition.

TEs prediction, annotation, classification and nomenclature were performed essentially as suggested by the unified classification system for eukaryotic TEs (WICKER *et al.* 2007) with two exceptions. The *Sukkula* were considered as *Gypsy* because of similarities with the *Erika* (*Gypsy*) elements. The *Athila* retrotransposons were analyzed separately from the other *Gypsy* retrotransposons (see also Supplemental Method 1 for detailed classification and annotation method).

The 79.1% of TEs space were shown to be composed of a wide variety of TEs, distributed as follows: 61.9% for class I (171 TEs from 48 families), 16.2% for class II (113 TEs from 28 families) and 1% for unclassified TEs (18 TEs of nine families) (Figure 1). Transposable elements distribution is not homogeneous or random across the 10 sequenced genomic regions, which map to different locations of the chromosome 3B. While class I retrotransposons constitute the highest TEs proportion of eight sequenced regions, BAC clone TA3B54F7 shows the highest proportion of *CACTA* class II (40.5%), while the smallest BAC clone TA3B63C11 (21.23 kb) carry no TEs (Figure 1). On the other hand, there are no clear relationships between sequence composition of the 10 genomic regions and their BIN map position on the chromosome 3B (Figure 1). For example BAC clones TA3B63B7, TA3B95F5, TA3B63N2 and TA3B54F7, which map on the , deletion BIN 3BL7 of the long arm of chromosome 3B, show different sequence classes and TEs proportions.

Class I transposable elements

The 61.9% class I TEs were composed of 171 TEs belonging to 48 families. Three main retrotransposon superfamilies constitute the majority of class I TE DNA sequences, as follows: 10.8% *Athila*- (37 TEs from four families), 30.8% *Gypsy*- (57 TEs of 14 families) and 14.7% *Copia*- (45 TE from 10 families) like 'long terminal repeats (LTR)'-retrotransposons (Figure1, see also supplemental Table 1 for details). With the exception

DNA sequence Classes	TA3B63B7 (170,495 bp)			TA3B63N2 (84,905bp)			TA3B81B7 (66,174bp)			TA3B95F5 (258,333bp)			TA3B63E4 (49,038bp)		
	Count	Size (bp)	(%)	Count	Size (bp)	(%)	Count	Size (bp)	(%)	Count	Size (bp)	(%)	Count	Size (bp)	(%)
Transposable elements															
Class I	24	152,852	89.7%	13	63,527	75.0%	12	51,214	77.4%	26	138,384	53.6%	9	15,755	32.1%
Athila	5	20,157	11.8%	5	13,310	15.7%	1	1,419	2.1%	4	20,929	8.1%	0	0	0.0%
Copia	4	31,799	18.7%	4	12,330	14.6%	3	22,756	34.4%	4	29,159	14.3%	6	7,192	14.7%
Gypsy	11	91,235	53.5%	1	12,986	15.3%	6	26,303	39.7%	8	69,708	27.0%	1	7,037	14.4%
LTR-Unknown	1	8,422	4.9%	2	18,913	22.3%	1	369	0.6%	1	6,589	2.6%	0	0	0.0%
LINE	2	853	0.5%	1	5,988	7.1%	0	0	0.0%	6	10,450	4.0%	2	1,526	3.1%
Others (unknown)	1	386	0.2%	0	0	0.0%	1	367	0.6%	3	1,549	0.6%	0	0	0.0%
Class II	6	1,458	0.9%	8	13,512	16.0%	5	747	1.1%	13	50,118	19.4%	8	9,457	19.3%
CACTA	3	1,147	0.4%	6	13,267	15.7%	1	130	0.2%	7	47,243	18.5%	2	8,733	17.8%
MITE	3	311	0.0%	2	245	0.3%	4	617	0.9%	5	659	0.3%	6	724	1.5%
LITE	0	0	0.0%	0	0	0.0%	0	0	0.0%	1	2,216	0.9%	0	0	0.0%
Unclassified	1	908	0.5%	0	0	0.0%	1	1,856	2.8%	3	1,902	0.7%	0	0	0.0%
Other Repeats	3	750	0.4%	0	0	0.0%	0	0	0.0%	11	12,210	4.7%	1	805	1.6%
All repeats	**34**	**155,968**	**91.5%**	**21**	**77,039**	**90.7%**	**18**	**53,817**	**81.3%**	**53**	**202,614**	**78.4%**	**18**	**26,017**	**53.1%**
Gene related	**2**	**786**	**0.5%**	**0**	**0**	**0.0%**	**3**	**1,204**	**1.9%**	**7**	**3,720**	**1.4%**	**1**	**786**	**1.6%**
Known	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%
Unknown	0	0	0.0%	0	0	0.0%	0	0	0.0%	2	1,371	0.5%	1	786	1.6%
Putative	0	0	0.0%	0	0	0.0%	1	595	0.9%	1	471	0.2%	0	0	0.0%
Pseudo	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%	0	0	0.0%
Relic	2	786	0.5%	0	0	0.0%	1	114	0.2%	0	0	0.0%	0	0	0.0%
Hypothetical	0	0	0.0%	0	0	0.0%	1	495	0.7%	4	1,878	0.7%	0	0	0.0%
Unassigned DNA		**13,741**	**8.0%**		**7,866**	**9.3%**		**11,153**	**16.8%**		**51,999**	**20.2%**		**22,235**	**45.3%**

Supplemental Table 2B. (continued) Detailled sequence class proportion of 10 sequenced BAC clones of wheat chromosome 3B.

of BAC clones TA3B63C11 (no TE detected), class I TEs composition range between 32.5% and 89.7%, depending on the sequenced region (Figure 1C).

Class II transposable elements

Class II TEs represent 16.2% of the cumulative sequence length (113 TEs from 28 families) and are composed of 54 *CACTA*, 56 MITEs and 3 LITEs. In term of sequence representation, the *CACTA* TEs represent the majority (96%) of class II DNA sequences. As in previous studies with *Triticeae* (WICKER *et al.* 2003a, 2005) the *CACTA* transposons were often found clustered in the genome. This is particularly the case for BAC clone TA3B54F7 where 15 *CACTA* TEs (complete and truncated) were found, representing 40.5% of the 190 kb (Figure1 and Supplemental Table 1). It is also the case of BAC clones TA3B95G2, TA3B95C9, TA3B95F5 and TA3B63N2, each containing 6-13 *CACTA*-like elements (complete and truncated) representing 15-20% of the BACs sequence lengths. The other five BAC clones are relatively *CACTA*-poor regions containing 0 to 2 *CACTA* TEs.

Novel transposable elements

Twenty-one transposable element families were identified for the first time in this study (Figure 1, indicated by arrows), four of which are present in several copies. Description of these novel TEs, their features, characteristics as well as the suggested nomenclature are presented in Supplemental Table 5. They account for 9.8% by number and 7.9% by length of the overall sequences.

Class I retrotransposons are the category for which we found the majority of novel TE families (17). From these, 11 novel LTR class I retrotransposon families were identified. Three novel LTR retrotransposons show stretches of weak similarities with known *Copia*-like and three other with known *Gypsy* retrotransposon families. They were designated with new family names, based on the TE classification guidelines (WICKER *et al.* 2007), and considered as belonging to the same super-families of the referenced TE with which they show the highest similarity (Supplemental Table 5).

We were not able to assign five of the novel LTR retrotransposon families to any of the three LTR retrotransposon superfamilies. Three of them (*Marina*, *Camillia*, and *Cathia*)

	Suggested family name	Novel TEs		Similarities with other TEs		Structural properties				
		BAC clones	Size	Family	BLAST	TSD	Nested	LTR size (%id)	PBS	PPT
Class I (LTR)										
Copia	*Alixa*	TA3B95C9	7,938	*BARE-1*	79%/452 aa	X	X	1421 (93%)	X	X
		TA3B95C9	6,958	*BARE-1*	77%/527 aa	X	X	1431 (88%)	-	-
		TA3B63E4	607	*Alixa*	86%/607 bp	-	-	-	-	-
		TA3B63N2	1,433	*Alixa*	90%/1433 bp	X	-	-	-	-
	Ambra	TA3B95F5	6,589	*Eugene*	71%/2121 bp	-	-	2704	-	-
	Verona	TA3B95C9	1,820	*Maximus*	61%/891 bp	-	-	-	-	-
Gypsy	*Cecilia*	TA3B63B7	13,558	*Laura*	70%/1213 bp	X	-	3828 (91%)	-	-
	Gwenaella	TA3B63N2	12,976	*Romani*	72%/666 bp	X	X	3638 (86%)	-	-
	Nathalia	TA3B63E4	7,027	*Jela*	89%/148 bp	X	X	643 (91%)	-	-
Undetermined	*Marina*	TA3B63B7	8,412	-	-	-	-	1603 (97%)	X	X
	Camillia	TA3B63N2	12,355	-	-	-	-	5590 (96%)	X	X
	Magella	TA3B81B7	534	*Unknown*	84%/401 bp	-	-	-	-	-
	Cathia	TA3B95C9	6,281	-	-	-	-	1729 (94%)	-	-
	Melina	TA3B63N2	6,538	*Unknown*	81%/634 bp	X	-	1023 (90%)	X	X
		TA3B95C9	3,306	*Unknown*	79%/1450 bp	-	-	1020	-	-
Class I (non-LTR)										
LINE	*Rachana*	TA3B95F5	4,418	*Mara*	67%/1281 bp	-	-			
	Pierrina	TA3B95F5	1,193	*Mara*	68%/1193 bp	-	-			
	Imena	TA3B54F7	6,085	*Isabelle*	69%/2958 bp	X	X			
Class I Undetermined										
	Stella	TA3B81B7	291	*polyprotein*	62%/98 aa	-	-			
	Odila	TA3B63B13	714	*polyprotein*	58%/157 aa	-	-			
	Soazica	TA3B95F5	481	*polyprotein*	66%/65 aa	-	-			
		TA3B95F5	481	*polyprotein*	66%/65 aa	-	-			
Class II										
CACTA	*Harry*	TA3B95C9	949	*Jorge*	73%/459 bp	-	-			
MITE	*Boulos*	TA3B95C9	241	*Polyphemus*	60%/110 bp	X	-			
	Mathieu	TA3B81B7	355	*Xenon*	80%/52 bp	X	-			
Unclassified										
	Aurelie	TA3B63B13	240	*Unclassified*	69%/180bp	-	-			
		TA3B63B13	236	*Unclassified*	77%/180 bp	-	-			

Supplemental Table 5. Characteristics of novel transposable elements (TEs) identified in the 10 sequenced genomic regions of wheat chromosome 3B. The similarities (BLASTN / BLASTX) and structural properties found are also presented.

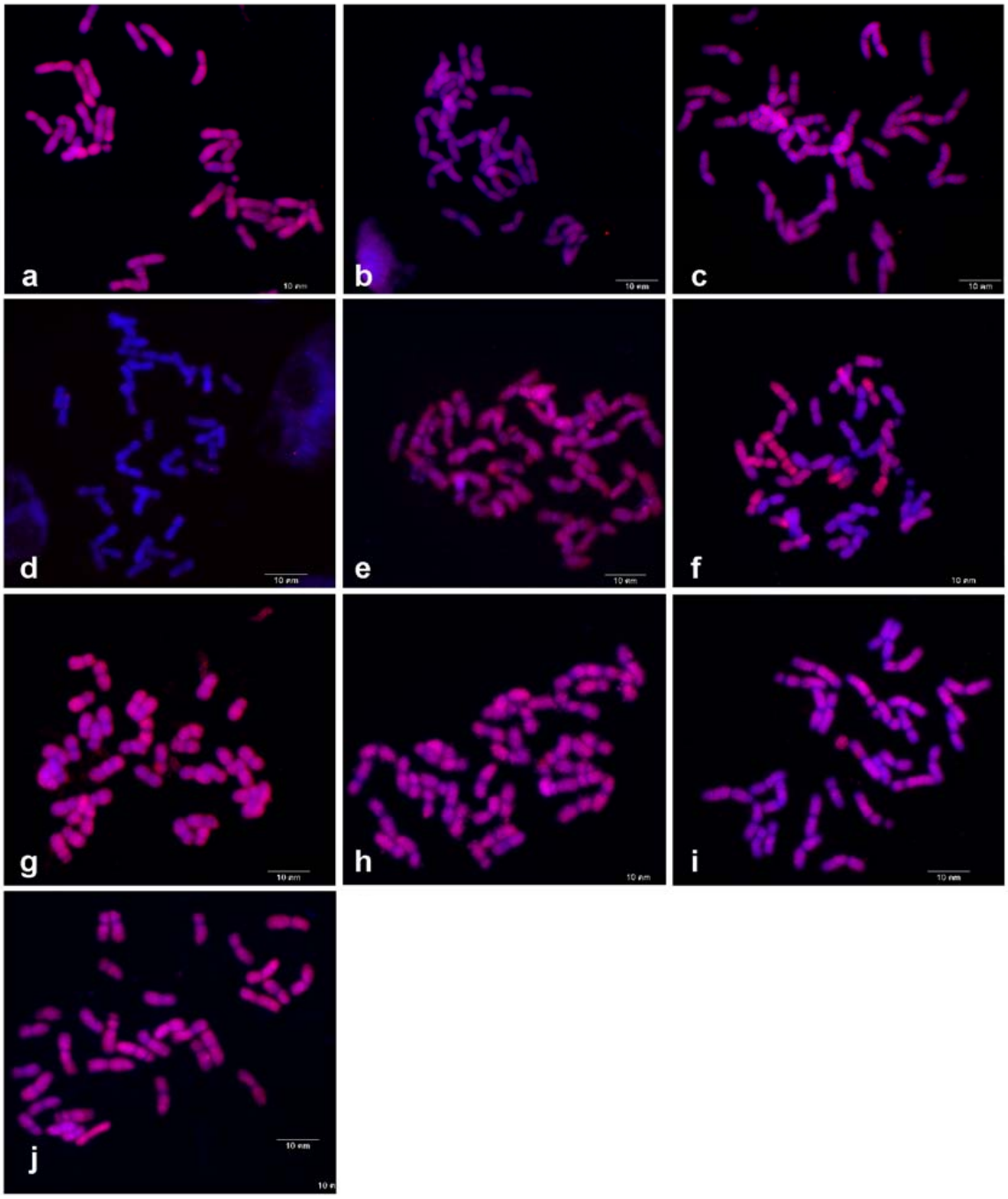
do not show matches in their LTR or internal domains with known LTR retrotransposon families (and superfamilies) and were identified based on their structural features (Supplemental Table 4). Two of these seem complete and the LTR_STRUC program (McCarthy *et al.* 2003) predicts two LTRs, with target site duplications (TSD) as well as predicted polynucleotide binding site (PBS) and a polypurine tract (PPT) signatures. The novel retrotransposon *Cathia* has its 3' LTR truncated, but LTR_STRUC predicts a putative PBS and a PPT (Supplemental Table 5) after adjustments of parameters.

Overall, seven novel LTR-retrotransposons (from six novel families) have 5' and 3' LTRs and target site duplication (TSD) motifs, which allow the estimation of their insertion dates (*see Results*: Insertions dates and proliferation of LTR-retrotransposons).

Three novel non-LTR class I retrotransposons were identified as *LINE* families and the remaining three novel class I TEs show weak similarities with class I TE polyproteins and could not be assigned as LTR or non-LTR class I TEs.

The three novel class II TE families include one *CACTA* and two MITEs of new families, sharing weak homologies with known *CACTA* and MITE.

The two copies of the novel unclassified element that we named *Aurelie* share a stretch of weak homology with other *Triticeae* unclassified transposable elements (Supplemental Table 5).



Supplemental Figure 1. Fluorescence in situ hybridization (FISH) pattern observed for 10 sequenced BAC clones from wheat chromosome 3B on mitotic metaphase chromosomes of *Triticum aestivum cv.* Chinese Spring (CS). BAC clone TA3B63B13 (**a**), TA3B63N2 (**b**),TA3B 81B7 (**c**),TA3B63C11 (**d**), TA3B63B7 (**e**), TA3B54F7 (**f**), TA3B95C9 (**g**), TA3B63E4 (**h**), TA3B95G2 (**i**), TA3B95F5 (**j**). The BAC DNA labelled with biotin-14-dCTP and detected with Texas-red avidin DCS. Chromosomes were counterstained with DAPI**. a, b, c, e, g, h, i** and **j** FISH pattern of BAC clones, which strongly hybridized to all chromosomes of CS and the hybridization signal covered the entire chromosome length. **d** BAC TA3B63C11 did no give a clear hybridization signal on any wheat chromosome. **f** FISH pattern of BAC TA3B54F7 with different hybridization intensities and also hybridized more strongly seven pairs of chromosomes. Bars represent 10 mm·Details of FISH experiment is supplied in Supplemental Method 2

III.1.2 Analyses FISH (Fluorescent In Situ Hybridisation).

Nous avons utilisé la méthode FISH pour observer l'hybridation des 10 clones BAC sur les chromosomes. Les résultats complémentaires de cette partie présentent les détails de la méthode utilisée (**Supplemental Method 2**) et les résultats des hybridations (**Supplemental Figure 1**).

Supplemental Method 2. Fluorescent in situ hybridisation (FISH) protocol

The root tips of the hexaploid wheat *Triticum aestivum* cv. 'Chinese Spring' were treated in a saturated water solution of α-bromonaphtalene at 4°C for 24 h and fixed in 1:3 acetic-ethanol. After the root tips were placed on aceto-carmine during 20 minutes and squashed in 45% acetic acid. The BAC clones were labelled by random priming with biotin-14-dUTP (Invitrogen, life technologies). Chromosome preparations were incubated in Rnase A (100 ng/µL) and pepsin (0.05%) in 10 mmol HCl, then fixed with paraformaldehyde (1%), dehydrated in an ethanol series (70%, 90% and 100%) and air-dried. The hybridization mixture consisted of 50% deionized formamide, 10% dextran sulfate, 2 X SSC, 1% SDS and labelled probe (100 ng per slide). Chromosome preparations and pre-denatured (92°C for 6 min) probe were denatured at 85°C for 10 min. In situ hybridization was carried out overnight in a moist chamber at 37°C. After hybridization, slides were washed for 5 min in 50% formamide in 2 X SSC at 42°C, followed by several washes in 4 X SSC-Tween20. Biotinylated probes were detected with Texas-red avidin (Vector Laboratories). The chromosomes were mounted and counterstained in Vectashield (Vector Laboratories) containing 2.5µg/mL 4',6-diamidino-2-phenylindole (DAPI). Fluorescence images were captured using a CoolSnap HQ camera (Photometrics, Tucson, Ariz) on an Axioplan 2 microscope (Zeiss, Oberkochen, Germany) and analyzed using MetaVue^{TM} (Universal Imaging corporation, Downington, PA).

BAC	Code	OLIGO_NAME	PCR PRODUCT (pb)	FORWARD SEQUENCE	REVERSE SEQUENCE	POSITION 1	POSITION 2
TA3B54F7	3	B54F7-HB-F/R-3	342	5' GAGGAGCTTGTGCCTGATACAC 3'	5' CATAGAGATCGATCATACATAGACG 3'	40775	41117
TA3B63B7	8	B63B7-HB-F/R-1	208	5' CGCCCTACATCAGCTTCCATCTTAG 3'	5' CGGGCTGTTTTCTGTTACAGATCTA 3'	163914	164122
TA3B63B7	9	B63B7-HB-F/R-2	135	5' TGGGGAAAAATCCAACCGTTATGC 3'	5' TCCTATTGGTTCTCCAGCCCTCACC 3'	24928	25063
TA3B63B7	11	B63B7-HB-F/R-4	239	5' GCCATATCCCCAACACCCAGTAATA 3'	5' TTTTGTTTGCTCGTGCGTTAGATCG 3'	84224	84463
TA3B95F5	23	B95F5-HB-F/R-3	197	5' CAAATCCTACAGTTTTCCTTTGCTC 3'	5' CACAAAACTATCTATCCCCT 3'	56284	56481
TA3B95G2	25	B95G2-HB-F/R-1	257	5' CCTTCTCCGAGAGATTGGTC 3'	5' TCGTTGCTAGTATCTCCATAGATTG 3'	11774	12031
TA3B95G2	27	B95G2-HB-F/R-3	150	5' GATGTTTCAACAACCGCCTTGACTT 3'	5' CCCGGAAGCTACACCACCATCTCA 3'	81027	81177
TA3B63C11	28	B63C11-HB-F/R-1	253	5' CTCGTGCTCTGCTTGACCTC 3'	5' AAACATGGCACGTACATAGATACAC 3'	423	676
TA3B63C11	30	B63C11-HB-F/R-3	302	5' ATACAACCTTGGCCATCCTGTCGC 3'	5' ACCTATGCCGGACCCTGTGAATAGT 3'	12479	12781
TA3B63E4	34	B63E4-HB-F/R-3	321	5' CCACCGCACCTCCAGTGATC 3'	5' AAGAAGAAACAATCTTAGGGCTAGA 3'	8704	9025
TA3B54F7	57	B54F7-DHB-F/R-6	475	5' CATAGCTGGCAGGACCTTCATACTG 3'	5' GCTGATGACCGACACGTTCGATG 3'	167751	168226
TA3B54F7	58	B54F7-DHB-F/R-7	330	5' TTTGTCTTCCCATACACGCC 3'	5' GTGCCACTTTGGGTTTACGACTAGC 3'	167840	168170
TA3B63B7	62	B63B7-DHB-F/R-10	334	5' AACGGAAGAAAAAGCGGAACCCA 3'	5' CCTCGCACTAGGACATTTGCCGTGA 3'	30032	30366
TA3B63B7	63	B63B7-DHB-F/R-11	725	5' TTTTCCGAAGGACAACACCACCGCA 3'	5' GTCCAGCAGCTCGGCAACCCGTTC 3'	45422	46147
TA3B63B7	64	B63B7-DHB-F/R-12	430	5' TCTTAGCGAGGCATTGAGGCGTGGG 3'	5' TTGCGCGTGTGTGGTGCGGTTG 3'	45607	46037
TA3B95F5	67	B95F5-DHB-F/R-7	705	5' TAGACCGCCTGTTGGATGGTATTG 3'	5' ATCTTCGTGATCGTTCCTTAGCCGT 3'	15470	16175
TA3B95F5	68	B95F5-DHB-F/R-8	266	5' CATGCATGGAGACAAGGCGTAGG 3'	5' GACGTACGACTACATCAACCGCATT 3'	15661	15927
TA3B95C9	69	B95C9-DHB-F/R-7	620	5' CTGGGAAGAAGGTGGGGAGGCGTTG 3'	5' TTGTGGGCCCCTCAGGCATCGTC 3'	86185	86805
TA3B95C9	70	B95C9-DHB-F/R-8	450	5' CCGTTGAAATCACCACTACCCGTAA 3'	5' GCCCCTCAGGCATCGTCTCA 3'	86352	86802
TA3B95C9	72	B95C9-DHB-F/R-10	272	5' ATTTTTAGCCGATTACGACCAA 3'	5' TGGATCAAGAAGGAGGAGACGTCCC 3'	147940	148212
TA3B95G2	73	B95G2-DHB-F/R-6	713	5' CAGTCAATGGAAATACAAACCGCTC 3'	5' AAGATTGAAGACTCCGTCCCGTGT 3'	87540	88253
TA3B95G2	74	B95G2-DHB-F/R-7	296	5' GCAACTACATCCCCGAACACCGTCA 3'	5' CTTGTCTCCTGTTACCATCCGCC 3'	87641	87937
TA3B95G2	75	B95G2-DHB-F/R-8	703	5' GCGAAGAAGAAGACCGACTCCGAA 3'	5' GATATTGTTGCTTCCGCTGACTCGT 3'	150679	151382
TA3B95G2	76	B95G2-DHB-F/R-9	349	5' GAATATCATCTACCCGAAGCATACC 3'	5' TGTTGTGATATCTTCCCGTGAGTCC 3'	150923	151272
TA3B63B7	81	B63B7-NHB-F/R-15	351	5' CAAAAAGATCGCTCCCGAAG 3'	5' CAAAGGACTATCCAGGGGTGTTGAG 3'	47718	48069
TA3B63E4	83	B63E4-NHB-F/R-4	392	5' TATATATGTTGAAGGGGCGGTTTAC 3'	5' CGGGCTCAGGGAAGGACGCAATG 3'	15498	15890
TA3B63N2	84	B63N2-NHB-F/R-6	378	5' CGGGCCATTTATTAGTCGGGT 3'	5' GCGGCAGTGGAACAAAATCGTAGTG 3'	13045	13423
TA3B95C9	89	B95C9-NHB-F/R-12	351	5' AATTTCAAACCGGGACTAAT 3'	5' TTTTAGGCCAGTCATCATACAATGC 3'	33812	34163
TA3B95C9	92	B95C9-NHB-F/R-15	372	5' ATCCCTAGTTAATATCAAGCGA 3'	5' TCTTCAACTAACATAGGTGCCATAG 3'	98515	98887
TA3B63B13	111	B63B13-BHB-F/R-14	375	5' GATCGCGAAGAAGTTCGACTA 3'	5' CTTGGAATTATTTTTTCCGAAAGTG 3'	58770	59145
TA3B63B7	112	B63B7-BHB-F/R-17	394	5' ATAAAGCGGTAAAGCTAGGA 3'	5' ATGTAGAAAGTGAAGGTTCGGAGAG 3'	143206	143600
TA3B63B7	113	B63B7-BHB-F/R18	420	5' AGACGGTGCTACGGCTACATC 3'	5' AAGAATATATGGCAAAGTTAGAGGC 3'	147098	147518
TA3B63B7	116	B63B7-BHB-F/R21	443	5' CCCCTTGCAGTGAAGGTAAAG 3'	5' TCGCAATTTTTATATGGGTGTTACT 3'	83832	84275
TA3B54F7	124	B54F7_NES-F/R-15	268	5' CTAGTCCAGTTGCTACAGCCCGGAT 3'	5' CTGTCAGGACCCCGATTCTATGCCA 3'	143308	143576
TA3B63B13	127	B63B13_NES-F/R-15	220	5' AATTAGGTCAAACGGAGTTGCC 3'	5' ATTAACCTAGTGCCCTGTGTTAGTT 3'	107610	107830
TA3B63B13	128	B63B13_NES-F/R16	262	5' AACGCTTCCGCTTTCGGTCTA 3'	5' CAATCAATAACAGCCCCTGTAGTAT 3'	58809	59071
TA3B63B7	132	B63B7_NES-F/R-23	221	5' CGTAATACTTCATCCCGCAACTA 3'	5' TGAGATTATGCAACTCCCGAATACC 3'	15216	15437
TA3B63B7	134	B63B7_NES-F/R-25	281	5' TGTTTCTTGCCCTGAACCGTC 3'	5' GTTGGGGATATGGCTATTAGGTATG 3'	83932	84213
TA3B63E4	140	B63E4_NES-F/R7	248	5' TTTACCCCTCTGGAGCCCCTAAATC 3'	5' ATAGCCGGTTTACGGCGACT 3'	8747	8995
TA3B95C9	145	B95C9_NES-F/R-17	155	5' TCATAATGTTGGGAAACGTAGCAAT 3'	5' TGAATCGCCACGAGTACGAC 3'	147971	148126
TA3B95C9	148	B95C9_NES-F/R-20	192	5' GGCTCACAAAATTGGGGTGCTAC 3'	5' GATATCCAACATACCTATCCGGTAA 3'	98587	98779
TA3B95G2	153	B95F5_NES-F/R16	204	5' GAGAAGGAGTCGGACAAGGCCGAA 3'	5' TACGATTAAATCGGGGGCGTCACAG 3'	151005	151209
TA3B63B13	155	B63B13_FN-F/R-20	378	5'TGTCCGGACAACATGGCACGTCAAG 3'	5'TTGTGATATCCTCCCGTGAGTCCC 3'	85361	85739
TA3B63B13	156	B63B13_FN-F/R-21	323	5'GAACCATCGCGAGATTAGTAACAGT 3'	5'TTTCTCCCGAGAGCTGTACTACCAT 3'	97711	98034
TA3B63B7	158	B63B7_FN_F/R27	253	5' GGGGACCTTTGTCGCATTTG 3'	5' GATGAAGTCGATGTAGGCGTACTGT 3'	158972	159225
TA3B95C9	163	B95C9_FN-F/R-23	443	5'GCTCTGCCATCCACCAAATTAGATT 3'	5'GGCCTGAAGGACTCGACCAC 3'	126889	127332
TA3B95C9	166	B95C9_FN-F/R-26	337	5'GGGTAAGTTACTTGGATCAGGTC 3'	5'CTGACAAATAGATAGCAATCGAAAA 3'	139385	139722
TA3B95G2	172	B95G2_FN-F/R-15	512	5'AGCGTTAATTATGCCCTGGTCAATG 3'	5'AAGCCCTACGTCTCGGAGCC 3'	148189	148701
TA3B95G2	175	B95G2_FN-F/R-18	249	5'TCCATACCTACCCCCCACTCTACTG 3'	5'CGGGTTCTCCTCGACGATCC 3'	173423	173672
TA3B95F5	189	B95F5_VD-F/R28	375	5'CTGGCATCCGAAGTCAGTAAGTTTA 3'	5'AGAAAGAAGTACGTCGGTGGAC 3'	135928	136303
TA3B95F5	190	B95F5_VD-F/R29	323	5'TTGGGAACAAATCTGATCCGGAGAG 3'	5'ACTCCTTCAACCCCGTTCACT 3'	141372	141695

Supplemental Table 1. Sequences and positions on the corresponding BAC clones of PCR primers used for tracing of the TE insertions (Figure 4 and Supplemental Table 7) as well as for the mapping of the 10 BAC clones on the deletion BINs of wheat chromosome 3B (Figure 1 and Supplemental Table 6).

III.1.3 Analyses PCRs (Polymerase Chain Reaction)

De nombreuses PCRs ont été effectuées dans le cadre de cet article pour préciser certains points ou comme bases d'analyses. Cette partie présente les résultats complémentaires concernant la localisation des 10 clones BAC sur les BINs de délétion du chromosome 3B (**Supplemental Table 1**) ainsi que la vérification de la méthode de datation des rétrotransposons par des PCRs sur des collections de génotypes de blé (**Supplemental Table 6**, **Supplemental Table 7**).

BAC clones	TE reference and name	Divergence (time in MYA)	Standard Error(time in MYA)	Insertion informations	Codes of primers used for PCR-tracing*	1	2	H20	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
TA3B95G2	RLG_Hawi_TA3B95G2-1	0.003 (0.12)	0.001 (0.04)	RLG_Hawi_TA3B95G2-1 into RLG_Jeli_TA3B95G2-1	75-76-nested153	0	1		0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	RLG_Jeli_TA3B95G2-1	0.008 (0.31)	0.004 (0.15)	RLG_Jeli_TA3B95G2-1 - into Unassigned DNA	172-175	1	1		1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0
	RLG_Fatima_TA3B95G2-6	0.051 (1.96)	0.010 (0.38)	RLG_Fatima_TA3B95G2-6 into RLG_Fatima_TA3B95G2-5	73-74	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	RLGa_Egug_TA3B95G2-1	*0.070 (2.70)*	*0.007(0.27)*	*Internal control for RLGa_Egug_TA3B95G2-1*	*27*	*1*	*1*		*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
TA3B63B13	RLG_Hawi_TA3B63B13-1	0.007 (0.27)	0.001(0.04)	RLG_Hawi_TA3B63B13-1 into unassigned DNA	155-156	1	1		1	1	1	1	1	0	0	1	0	1	0	1	1	1	1	0	0	1	1
	RLC_Angela_TA3B63B13-6	0.085 (3.27)	0.007(0.27)	RLC_Angela_TA3B63B13-6 into RLGa_Wham_TA3B63B13-1	111-nested 128	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TA3B95C9	RLG_Jeli_TA3B95C9-1	0.004 (0.15)	0.003(0.12)	RLG_Jeli_TA3B95C9-1 into DTC_Harry_TA3B95C9-1	163-166	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	RLC_Wis_TA3B95C9-3	0.056 (2.15)	0.006(0.24)	RLC_Wis_TA3B95C9-3 into DTC_Vincent_TA3B95C9-2	72-nested 145	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	**RLC_Alixa_TA3B95C9-1	0.065 (2.50)	0.008 (0.31)	RLC_Alixa_TA3B95C9-1 into DTC_Caspar_TA3B95C9-1	89	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
	**RLC_Alixa_TA3B95C9-2	0.119 (4.58)	0.010(0.38)	RLC_Alixa_TA3B95C9-2 into RLC_Wis_TA3B95C9-2	92-nested 148	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	RLGa_Egug_TA3B95C9-1	0.131 (5.04)	0.005(0.20)	RLGa_Egug_TA3B95C9-1 into RLG_Fatima_TA3B95C9-1	69-70-nested	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	RLC_Wis_TA3B95C9-3	*0.056 (2.15)*	*0.006(0.24)*	*Internal control for RLC_Wis_TA3B95C9-3*	*145*	*1*	*1*		*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
	**RLC_Alixa_TA3B95C9-2*	*0.119 (4.58)*	*0.010(0.38)*	*Internal control for RLC_Alixa_TA3B95C9-2*	*148*	*1*	*1*		*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
TA3B63B7	RLG_Sukkula_TA3B63B7-2	0.000 (0.00)	0.000(0.00)	RLG_Sukkula_TA3B63B7-2 into RLX_Marina_TA3B63B7-1	65-113	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	**RLX_Marina_TA3B63B7-1	0.018 (0.69)	0.004(0.20)	RLX_Marina_TA3B63B7-1 into RLG_Wilma_TA3B63B7-1	8-112-158	1	1		1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1
	**RLG_Cecilia_TA3B63B7-1	0.039 (1.50)	0.003(0.12)	RLG_Cecilia_TA3B63B7-1 into RLGa_Sabrina_TA3B63B7-3	81	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	RLG_Nusif_TA3B63B7-2	0.062 (2.38)	0.009(0.35)	RLG_Nusif_TA3B63B7-1 into Nikita_TA3B63B7-1	9	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	RLG_Nusif_TA3B63B7-3	0.095 (3.65)	0.012(0.46)	RLG_Nusif_TA3B63B7-3 into RLG_Nusif_TA3B63B7-2	63-64	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
	RLG_Ifis_TA3B63B7-1	0.096 (3.70)	0.014(0.54)	RLG_Ifis_TA3B63B7-1 into RLGa_Sabrina_TA3B63B7-2	11-62-116-nested134	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	RLG_Ifis_TA3B63B7-1	*0.096 (3.70)*	*0.014(0.54)*	*Internal control for RLG_Ifis_TA3B63B7-1*	*134*	*1*	*1*		*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
TA3B95F5	RLG_Jeli_TA3B95F5-2	0.037 (1.42)	0.008 (0.31)	RLG_Jeli_TA3B95F5-2 into RLC_Oref_TA3B95F5-1	189-190	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	RLC_Angela_TA3B95F5-1	0.020 (0.77)	0.004(0.15)	RLC_Angela_TA3B95F5-1 into RLG_Fatima_TA3B95F5-1	67-68	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	RLG_Wilma_TA3B95F5-1	0.042 (1.62)	0.006(0.24)	RLG_Wilma_TA3B95F5-1 into Jorge_TA3B95F5-1	23	1	1		1	1	1	1	1	1	1	1	1	1		1	1	1	0	1	1	1
TA3B63E4	**RLG_Nathalia_TA3B63E4-1	0.019 (1.73)	0.006(0.24)	RLG_Nathalia_TA3B63E4-1 into DTC_Vincent_TA3B63E4-2	34-83-nested140	0	1		0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
TA3B54F7	*RLG_Laura_TA3B54F7-3*	*0.02 0(0.77)*	*0.003(0.12)*	*Internal control for RLG_Laura_TA3B54F7-3*	*124*	*1*	*1*		*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
TA3B63N2	**RLG_Gwenaella_TA3B63N2-1*	*0.098 (3.77)*	*0.010(0.38)*	*Internal control for RLG_Gwenaella_TA3B63N2-1*	*84*	*1*	*1*		*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*

Supplemental Table 7. PCR-based tracing of series of retrotransposons, inserted at different dates in wheat chromosome 3B, across a collection of genotypes of tetraploid (*T. turgidum*) and hexaploid (*T. aestivum*) wheat species.

-1: detection of the TE insertion: PCR amplification. -0: no PCR amplification is observed. - TE in italics correspond to internal PCR control of the cited TE (done with PCR primers designed within the TE) and showing the presence of TE in almost all tested wheat genotypes and species. -L: 50 bp ladder (Invitrogen).. * Details of PCR primer sequences are given in Supplemental Table 7. ** Those elements correspond to Novel retrotransposons identified within this study.

BAC	Primers code[a]	OLIGO_NAME	1- R	2- C S	3- N3AT3B	4- N3BT3A	5- N3DT3A	6- H2O	7- Dt3BL	8- Dt3DL	9- Dt3DS	10- 3AL-3	11- 3AL-5	12- 3AS-2	13- 3AS-4	14- 3BL-2	15- 3BL-7	16- 3BL-10	17- 3BS-1	18- 3BS-8	19- 3BS-9	20- 3DL-3	21- 3DS-3	22- 3DS-6	BIN map position
TA3B63E4	34	B63E4-HB-F/R-3	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	Not localized (Disperse signal)
	83	B63E4-NHB-F/R-4	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	Disperse signal
TA3B63C11	28	B63C11-HB-F/R-1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	3BS8-0.78-1.00
	30	B63C11-HB-F/R-3	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	3BS8-0.78-1.00
TA3B95G2	25	B95G2-HB-F/R-1	0	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	3BS8-0.78-1.00
	73	B95G2-DHB-F/R-6	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	3BS8-0.78-1.00
	74	B95G2-DHB-F/R-7	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	3BS8-0.78-1.00
	76	B95G2-DHB-F/R-9	0	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	3BS8-0.78-1.00
TA3B63B13	111	B63B13-BHB-F/R-14	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	3BS1-0.33-.57
TA3B95C9	70	B95C9-DHB-F/R-8	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	3BL2-0.22-0.50
	72	B95C9-DHB-F/R-10	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	3BL2-0.22-0.50
	89	B95C9-NHB-F/R-12	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	3BL2-0.22-0.50
	92	B95C9-NHB-F/R-15	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	3BL2-0.22-0.50
TA3B81B7	13	B81B7-HB-F/R-2	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	3BL2-0.22-0.50
TA3B63N2	84	B63N2-NHB-F/R-6	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
TA3B63B7	9	B63B7-HB-F/R-2	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
	62	B63B7-DHB-F/R-10	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
	63	B63B7-DHB-F/R-11	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
	64	B63B7-DHB-F/R-12	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
	81	B63B7-NHB-F/R-15	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
	112	B63B7-BHB-F/R-17	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
TA3B54F7	3	B54F7-HB-F/R-3	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
	57	B54F7-DHB-F/R-6	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
	58	B54F7-DHB-F/R-7	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
TA3B95F5	23	B95F5-HB-F/R-3	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
	67	B95F5-DHB-F/R-7	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00
	68	B95F5-DHB-F/R-8	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	3BL7-0.36-1.00

[a] Sequences of the PCR primers are supplied in Supplemental Table 7.

1 : Amplified with PCR on the deletion line.

0 : not amplified with PCR.

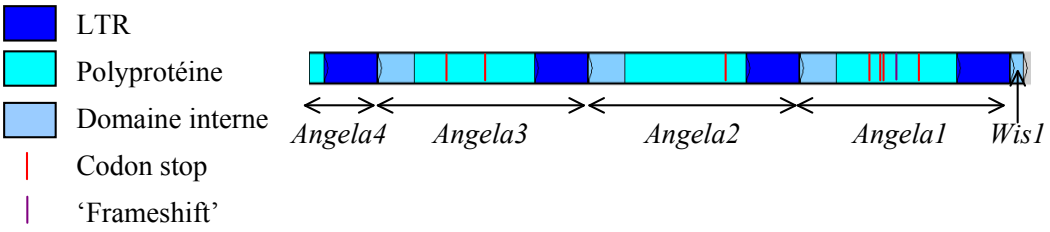
R: hexaploid wheat *cv.* Renan

CS: hexaploid wheat *cv.* Chinese Spring

Deletion lines are from Qi et al. 2003

Supplemental Table 6. Mapping of 10 sequenced BAC clones on wheat chromosome 3B deletion lines, with PCR-specific markers derived from transposable elements (TE)-TE junctions.

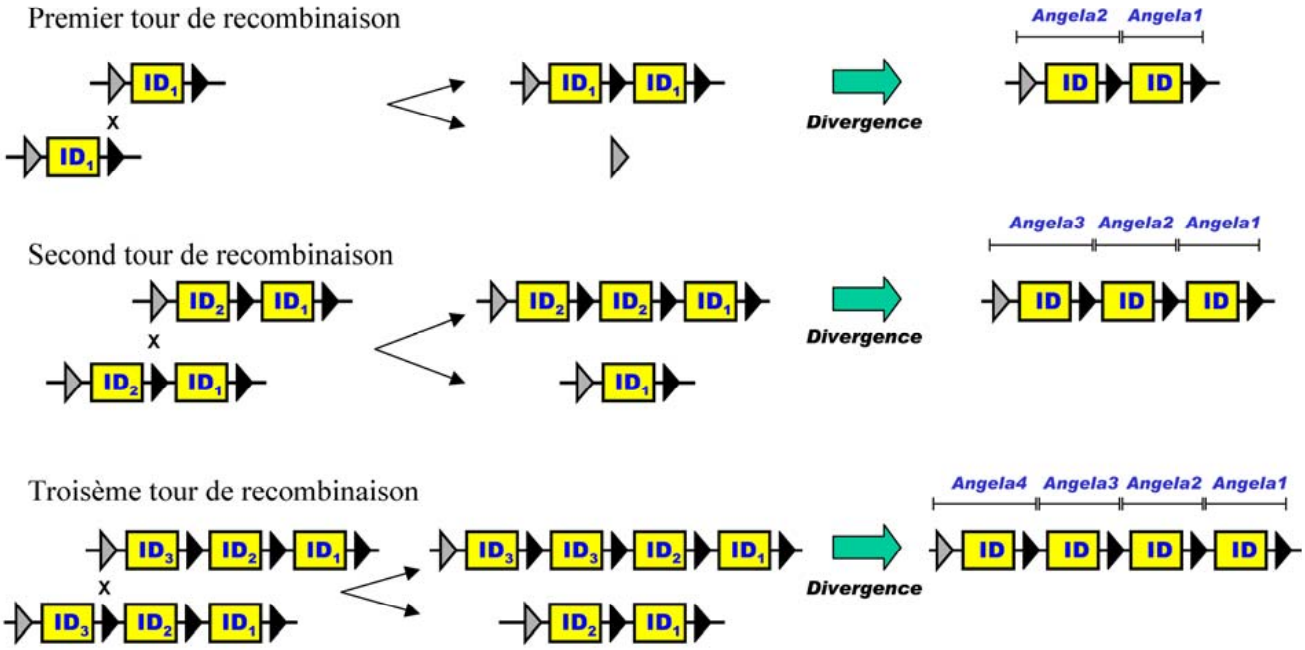
A



B

	Divergence des LTRs			Divergence des domaines internes		
	Angela2	Angela3	Angela4	Angela2	Angela3	Angela4
Angela1	0,023	0,023	0,025	0,008	0,012	0,014
Angela2		0,014	0,015		0,004	0,006
Angela3			0,011			0,006

C



D

BAC_NAME	TEs_REFERENCE	Datation	Commentaires	RESULTS	MAPPING ON CHROMOSOME	NUM	1_R	2_C S	3_H2O	4_ABD_T. spelta épautre barbu	5_ABD_T. spelta épautre noir barbu velu	6_ABD_T. spelta épautre blanc sans barbe	7_ABD_T. spelta Rouquin	8_ABD_T. macha 1793	9_ABD_T. compactum nututum 71V	10_ABD_T. compactum crebicum 72V	11_ABD_T. compactum clavatum 73V	12_ABD_T. compactum icterinum 74V	13_ABD_T. compactum eninaceum 75V	14_ABD_T. sparococcum tumidum percic globosum	15_ABD_T. aestivum Soisson	16_AB_T. turgidum durum cv Langdon	17_AB_T. turgidum durum	18_AB_T. turgidum dicoccum	19_AB_T. turgidum dicoccoides	20_AB_T. turgidum polinicum	21_AB_T. turgidum turgidum	22_AG_T. timophevii militanae	23_AG_T. timophevii (kurdistan?)	24_AG_T. timophevii	25_S_Ae. longissima	26_S_Ae. speltoides	27_S_Ae. sharonensis	28_A_T. urartu	29_A_T. monococcum beotricum	30_D_Ae tauschii	31_D_Ae tauschii	32_D_Ae tauschii
B63B13	Angela_B63B13-4 --> Wis_B63B13-1		Insertion de Angela4 dans Wis1	177-178		177	1	1		?	?	?	?	?			1		1		1	1	1			1	1											
B63B13	Wis_B63B13-1 --> Angela_B63B13-5		Insertion de Angela5 dans Wis1	179		179	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1											

Figure I-1. Détection de recombinaisons homologues inégales dans le BAC B63B13. (A) Structure du cluster d'éléments *Angela*. (B) Divergence entre les différents éléments. (C) Formation du cluster d'*Angela* par une série de recombinaison homologues inégales. (D) Résultats des PCRs sur une collection de lignées de blés (les cases oranges correspondent aux amplifications)

III.2 Prolifération des TEs par recombinaisons homologues inégales

L'analyse des différents TEs présentée dans l'Article 1 montre une proportion plus élevée de copies tronquées par rapport aux copies non-tronquées (Article 1, Table 1). Ces copies tronquées résultent de différents mécanismes de recombinaisons homologues inégales que j'avais schématisés dans mon introduction générale (Figure 8). De plus, la majorité des copies tronquées provient de recombinaisons illégitimes d'ADN. Les copies tronquées par des recombinaisons homologues inégales 'médiées' par les LTRs sont beaucoup moins nombreuses (Article 1, Table 1). Ces recombinaisons forment des Solo-LTRs avec TSD (Target Site Duplication) et des éléments dits 'complets sans TSD' (Figure 8). Cette absence de TSD confirme leur origine par recombinaisons homologues inégales et non par insertion.

En analysant en détail les différentes séquences génomiques, j'ai trouvé que le clone BAC B63B13 présente à son extrémité 5' une structuration intéressante des rétrotransposons de la famille des *Angela* (superfamille des *Copia*, classe I). En effet les premières 23.681 pb (14,4 % de la longueur du clone BAC) sont composées d'un cluster de quatre copies de cet élément que nous avons désignées *Angela4*, *Angela3*, *Angela2*, *Angela1* partant de l'extrémité 5' (Article 1 - Figure 1a, Figure I-1A). La copie *Angela4* à l'extrémité du clone BAC est incomplètement séquencée alors que la copie *Angela1* est flanquée par un fragment LTR de l'élément *Wis1*. Une analyse plus détaillée (Figure I-1A, I-1B) montre que les quatre copies d'*Angela* sont organisées en tandem, avec un seul LTR séparant le domaine interne de chaque copie (essentiellement composé de la polyprotéine). Ces quatre copies sont d'origine relativement récente comme l'atteste la forte conservation de leur polyprotéine, avec quelques mutations stop et/ou décalage du cadre de lecture ('frameshift'), en fonction des copies (Figure I-2B). Cependant, les séquences flanquant les LTRs montrent l'absence de TSDs qui joue le rôle de signature d'une nouvelle insertion (Figure I-1A). Ces quatre LTRs semblent correspondre à des LTR-3', et aucun ne serait le LTR-5' d'une des quatre copies *Angela*. Plus intéressant encore, les comparaisons des séquences des quatre copies montrent que la copie *Angela4* est plus proche d'*Angela3*, elle-même plus proche d'*Angela2* qui est plus proche d'*Angela1* (Figure I-1B), que l'on compare séparément ou ensemble les séquences des polyprotéines et des LTRs (Figure I-1B).

Ces observations suggèrent que les quatre copies d'*Angela* ne résultent pas uniquement d'une activité insertionnelle. Après l'insertion initiale de la copie *Angela1*, les copies *Angela2*, *Angela3* et *Angela4* ont été générées par au moins trois cycles de recombinaisons homologues inégales, médiées par les LTRs comme illustré dans la Figure I-1C. Ces cycles multiples de recombinaisons homologues inégales ont donné quatre copies *Angela* (au moins) dupliquées en tandem (Figure I-1C). Il est vraisemblable que ces événements de recombinaisons homologues inégales ont eu lieu entre chromatides sœurs (Figure I-1C). Les comparaisons des séquences nous ont aussi permis d'évaluer les dates récentes de ces événements. Nous avons pu valider, avec des PCRs, l'origine récente de l'insertion de la copie *Angela1* (détectée seulement dans quelques génotypes portant le génome B) et l'origine plus ancienne du fragment de *Wis1* (détecté dans tous les génotypes) (Figure I-1D). La forte conservation des séquences entre les quatre copies *Angela* ne nous a pas permis de développer des marqueurs PCRs permettant de préciser l'origine récente de leur duplication en tandem.

J'ai également trouvé un événement similaire de recombinaison homologue inégale sur le même clone BAC (Article 1, Figure1a). Cet événement a abouti à la duplication d'un élément de la famille *Wham* (superfamille *Athila*). De la même façon que le cas des duplications en tandem des *Angela*, décrites ci-dessus, deux copies *Wham* (B63B13_Wham2 et B63B13_Wham3) sont séparées par un seul LTR qui ne montre pas de signatures TSD, suggérant que B63B13_Wham3 dériverait d'une duplication par recombinaison homologue inégale de B63B13_Wham2.

Les cas décrits ci-dessus illustrent qu'en plus de leur activité insertionnelle, l'amplification des rétrotransposons dans les génomes du blé s'est aussi faite par des recombinaisons homologues inégales impliquant leurs LTRs et sans impliquer leur machinerie de transposition.

Une amplification en tandem de séquences, concernant des gènes *NBS-LRR*, a été également décrite dans l'orge (Wicker *et al.* 2007b). Cependant, les motifs impliqués étaient bien plus courts (<20 bp) et correspondaient plus à des recombinaisons illégitimes qu'à des recombinaisons homologues inégales.

IV Discussion

L'abondance des TEs varie largement entre les différents organismes (Figure 5). Mon étude présente pour la première fois la dynamique, les taux et les périodes de prolifération des éléments transposables comparés dans deux génomes du blé. Ceci a été rendu possible par le séquençage de 10 régions génomiques du génome B du blé (chromosome 3B) qui ont constitué, avec les séquences disponibles du génome A, des sets de données de séquences représentatives des deux génomes. Il a été suggéré que pour analyser les éléments transposables les plus abondants, le séquençage d'une faible proportion du génome pourrait suffire (Brenner *et al.* 1993, Vitte et Bennetzen 2006, Liu *et al.* 2007, Charles *et al.* 2008). Néanmoins, les TEs de classe I et classe II étudiés sont loin de couvrir l'ensemble des TEs du blé comme l'atteste l'identification de nombreux nouveaux TEs (21 nouvelles familles dont 17 rétrotransposons) dans cette étude. De nombreux autres TEs seront probablement identifiés au fur et à mesure de l'avancée des programmes de séquençage.

La majeure partie des séquences génomiques analysées (80%) pour les génomes A et B du blé est composée d'éléments transposables. Néanmoins, les TEs qui les composent ont évolué différemment dans les deux génomes. Nous avons ainsi montré que les TEs des superfamilles *Copia* (classe I) et *CACTA* (classe II) ont plus récemment proliféré dans le génome A mais pas dans le génome B. Ce dernier a eu une prolifération plus importante des éléments de la superfamille *Gypsy* (classe I). D'autres études ont initialement essayé d'évaluer la prolifération des éléments transposables dans les génomes des blés. Li *et al.* (2004) a analysé le génome D de l'espèce dipoïde *Ae. tauschii* et a conclu que le nombre de copies de la plupart des TEs a augmenté après l'allohexaploïdisation, la méthode utilisée ('dot blots') n'étant cependant pas très précise.

Sabot *et al.* (2005) ont actualisé l'annotation des TEs dans les séquences génomiques disponibles (à l'époque) et étudié leur composition et distribution par rapport aux gènes. Ils ont suggéré que les éléments de la superfamille *Copia* ont été les plus actifs en considérant l'ensemble des trois génomes A, B et D du blé. Ceci n'est en fait valable que pour le génome A qui était sur-représenté dans l'étude de Sabot *et al.* (2005) alors que les génomes B et D étaient sous-représentés.

Mon étude révèle que la majorité des clones BAC du génome B (7/10), séquencés sans a priori d'ancrage par des gènes, portent des gènes complets (5) ou tronqués (mutation ou relique) (2). Ceci confirme des conclusions antérieures montrant que les gènes sont distribués de façon uniforme sur les chromosomes du blé (Devos *et al.* 2005). Il est intéressant de noter que trois des six gènes tronqués l'ont été par des insertions d'éléments transposables (Article 1, Figure 1a) illustrant le rôle de ces derniers dans l'évolution des génomes du blé.

Comme dans l'analyse présentée ici, les dates d'insertion de la majeure partie des copies complètes (ayant les deux LTRs) sont estimées à moins de 3 Ma (SanMiguel *et al.* 1998, 2002, Wicker *et al.* 2003b, 2005, Gao *et al.* 2004, Ma *et al.* 2004, Du *et al.* 2006, Piegu *et al.* 2006, Wicker et Keller 2007). Les éléments plus anciens ne sont pas conservés suggérant des mécanismes d'élimination 'active' des rétrotransposons comme la recombinaison illégitime ou la recombinaison homologue inégale (Vicient *et al.* 1999, Devos *et al.* 2002, Ma *et al.* 2004, Pereira 2004). Mon étude montre que la proportion des TEs complets est plus faible que celle des TEs tronqués (Article 1, Table 1). D'autre part, la majeure partie des éléments a été tronquée par des recombinaisons illégitimes (Article 1, Table 1). De plus, mon étude a révélé des cas d'amplification des rétrotransposons impliquant la recombinaison homologue inégale et sans impliquer leur machinerie de transposition.

Mon étude confirme la pertinence des méthodes de datation des insertions des rétrotransposons basées sur la divergence des séquences de leur LTRs. En effet, du fait de leur mécanisme de réplication (Figure 2), les séquences des deux LTRs d'un rétrotransposon sont normalement identiques quand l'insertion a lieu. Mon étude a révélé une bonne corrélation entre les estimations des dates d'insertion (avant ou après la tétraploïdisation) et les profils PCRs observés pour 24 des insertions datées sur des lignées de polyploïdes : la présence/absence des insertions dans certaines ou dans toutes les lignées permet de vérifier si une insertion a bien eu lieu avant ou après la tétraploïdisation. Aucune de ces 24 insertions étudiées n'était commune aux génomes A, B et D ce qui confirme les observations antérieures sur la non-conservation de l'espace TEs entre les génomes (Wicker *et al.* 2003b, Chantret *et al.* 2005, Isidore *et al.* 2005, Gu *et al.* 2006, Dvorak *et al.* 2006). Par ailleurs, l'étude révèle que la majeure partie (87%) de la prolifération différentielle des TEs dans les génomes A et B du blé s'est produite avant l'événement de polyploïdisation qui les a réunis il y a près de 0,5 Ma.

Les taux et les périodes de prolifération des TEs sont donc la résultante de deux forces d'évolution : leur insertion (transposition) (Bennetzen et Kellogg 1997) mais aussi leur élimination (Petrov *et al.* 2000, Petrov 2002a). Les résultats de l'Article 1 nous ont permis de caractériser la force d'insertion dans les génomes A et B du blé. Le traçage par PCR utilisé dans cette étude a même permis de retracer l'histoire insertionnelle des TEs dans différentes accessions de chaque génome étudié. Cependant, les estimations sont moins précises en ce qui concerne l'autre force d'évolution des TEs, c'est à dire leur élimination par recombinaisons homologues inégales ou illégitimes. Jusqu'à présent, les études comparatives s'intéressaient principalement à la comparaison d'un génotype représentatif de chaque génome (Wicker *et al.* 2003b, Chantret *et al.* 2005, Isidore *et al.* 2005, Gu *et al.* 2006, Dvorak *et al.* 2006). Des analyses comparatives entre différents haplotypes au sein de chaque génome permettraient d'apprécier l'étendue, la vitesse et le taux de variabilité, qui résultent de cette force évolutive majeure (recombinaisons homologues inégales ou illégitimes des TEs).

Partie II : Caractérisation de l'élimination active des TEs dans les génomes du blé : analyse de variabilité haplotypique inter- et intra-génomique

I Introduction

L'étude présentée en Partie I sur l'analyse de séquences représentatives des génomes A et B du blé a montré que la proportion des TEs complets est plus faible que celle des TEs tronqués (Article 1, table 1). De plus, les dates d'insertion estimées pour la plupart des rétrotransposons complets sont inférieures à 3 Ma, comme celles trouvées dans d'autres études (SanMiguel *et al.* 1998, 2002, Wicker *et al.* 2003b, 2005, Gao *et al.* 2004, Ma *et al.* 2004, Du *et al.* 2006, Piegu *et al.* 2006, Wicker et Keller 2007). Ces résultats suggèrent des mécanismes d'élimination rapide des rétrotransposons. Nous avons montré que la recombinaison illégitime est le mécanisme principal de délétion des TEs dans le blé (Article 1). Mes études ont aussi confirmé que les taux et les périodes de prolifération des TEs sont la résultante de deux forces d'évolution : leur insertion (transposition) (Bennetzen et Kellogg 1997) et leur élimination (Petrov *et al.* 2000, Petrov 2002a).

J'ai pu notamment apprécier l'insertion des TEs dans les génomes A et B du blé (Article 1, Figure 3). L'activité insertionnelle est continue au cours du temps : elle n'a été ni activée ni réprimée par les événements de polyploïdisation récents (allotétraploïdisation et allohexaploïdisation) (Article 1, Figure 3 et 4). Cependant, nous avons montré que cette prolifération est différentielle entre les deux génomes. Une stratégie de traçage par des marqueurs PCR des insertions datées de rétrotransposons a permis de vérifier la présence ou l'absence de ces insertions dans différents génotypes des génomes du blé. Cette analyse haplotypique inter- et intra-génomique a donc montré une importante variabilité insertionnelle.

L'élimination des TEs, appréciée sur la base de la proportion des éléments tronqués, semble importante. Je n'ai malheureusement pas pu analyser de façon précise, dans l'étude précédente, l'étendue de cette force d'évolution dans les différents génomes. En effet, l'insertion d'un TE mobilise sa séquence entière tandis qu'une délétion par recombinaison illégitime, principale force d'élimination des TEs observée dans le blé, ne concerne pas précisément la séquence entière d'un TE. Elle peut aussi bien impliquer une partie d'un TE que plusieurs éléments ainsi que l'ADN les séparant. On ne peut pas connaître, a priori, l'étendue d'une délétion et donc utiliser la stratégie de traçage par marqueur PCR pour

distinguer la présence ou l'absence de la délétion, comme cela a été réalisé pour les insertions de TEs dans l'Article 1.

Les comparaisons d'haplotypes permettent d'apprécier l'étendue des insertions et des éliminations des TEs, ce qui est primordial pour la compréhension de l'organisation et de l'évolution des génomes du blé. Cette analyse peut se faire sur trois niveaux de comparaison (inter-génomique, intra-génomique, inter-spécifique) donnant une représentation de l'évolution de l'espace TEs à des échelles de temps différentes.

Il a été aussitôt révélé que l'espace occupé par les TEs (espace TEs) n'est pas du tout conservé entre les génomes du blé (Isidore *et al.* 2005, Chantret *et al.* 2005, Gu *et al.* 2006, Dvorak *et al.* 2006, Salse *et al.* 2008b), qui ont divergé il y a 2,5-4 Ma (Huang *et al.* 2002) indiquant une évolution rapide. Les analyses comparatives inter-génomiques de l'espace TEs ne permettent donc pas de préciser la variabilité haplotypique. Cependant, nous avons montré précédemment qu'une analyse sur un jeu de séquences suffisamment important (représentatif) des génomes permet d'apprécier l'activité insertionnelle et sa variabilité à ce niveau (Article 1).

Les analyses de génomique comparative sur les génomes du blé ont rarement porté sur des accessions ou des génotypes différents de la même espèce (comparaison intra-génomique et intra-spécifique). Ce niveau de comparaison aurait pourtant permis d'évaluer l'étendue de l'insertion et de l'élimination des TEs sur une courte échelle de temps d'évolution. Si la conservation entre les séquences est suffisante, on peut même déduire les bases moléculaires précises des événements de réarrangements (délétions, inversions).

Des études de comparaisons intra-génomiques partielles ont été réalisées sur le blé et ont permis une appréciation préliminaire des forces d'insertion et d'élimination des éléments transposables (Wicker *et al.* 2003b, Chantret *et al.* 2005, Isidore *et al.* 2005, Gu *et al.* 2006, Dvorak *et al.* 2006). Ces études ont essentiellement comparé les séquences d'un locus à différents niveaux de ploïdie dans les génomes A, B et D.

Chantret *et al.* (2005) ont suggéré le rôle des TEs dans l'élimination indépendante du locus *Ha* dans les génomes A et B des blés polyploïdes par recombinaison illégitime. Les TEs favoriseraient donc ce mécanisme de délétion très important dans le blé. Isidore *et al.* (2005) ont aussi montré que la variabilité haplotypique préexistait avant les évènements récents de polyploïdisation. En comparant une ou deux régions orthologues dans trois espèces différentes du génome A du blé (*T. urartu*, *T. monococcum* et *T. turgidum*), Dubcowsky et

Dvorak (2007) ont estimé le taux de remplacement de l'espace TEs (la proportion non conservée) à 62 ± 3% par Ma. Avec une telle dynamique, il n'est pas surprenant que l'espace TEs ne soit pas conservé entre les différents génomes du blé, ayant divergé il y a moins de 4 Ma.

La variabilité haplotypique intra-spécifique a été plus ou moins étudiée dans d'autres espèces de *Poaceae*. Au cours des dernières années, la variabilité du maïs a été intensivement analysée et a montré que la proportion des séquences génomiques non-conservées entre les lignées est très importante, avec moins de 50% de séquences communes (Tikhonov *et al.* 1999, Fu et Dooner 2002, Song et Messing 2003, Brunner *et al.* 2005, Lai *et al.* 2005, Morgante *et al.* 2005, Wang et Dooner 2006, Xu et Messing 2006). La prolifération différentielle des éléments transposables et surtout le mouvement des gènes par les TEs de type *Helitrons*, très actifs dans le maïs, sont les principaux responsables de l'importante variabilité trouvée dans cette espèce (Lai *et al.* 2005, Morgante *et al.* 2005, Wang et Dooner 2006, Xu et Messing 2006). La comparaison de plusieurs génotypes d'orge (Scherrer *et al.* 2005, Wicker *et al.* 2009a) et de riz (Han et Xue, 2003, Piegu *et al.* 2006) révèle que la prolifération différentielle des TEs est la raison principale de cette variabilité dans ces espèces.

Les tailles importantes des génomes du blé et la difficulté de préparer certaines ressources génomiques comme les banques BAC représentaient un frein majeur aux approches comparatives de ce type. A mon arrivée, l'équipe OEPG (Organisation et Evolution des Génomes des Plantes) de Boulos Chalhoub avait acquis et surtout développé de nombreuses ressources génomiques dans le cadre d'un projet de séquençage comparatif sur le blé (APCNS2003). Une attention particulière était portée sur le locus *Ha*. Une première étude de ce locus, menée par cette équipe, a montré le rôle important des TEs dans l'évolution des génomes du blé, en relation avec la polyploïdie (Chantret *et al.* 2005). Le séquençage comparatif de ce locus dans plusieurs génotypes de plusieurs espèces de blé, à différents niveaux de ploïdie était en cours. J'ai donc analysé la variabilité haplotypique de ce locus (le plus séquencé) dans 16 haplotypes des génomes A, B, S et D correspondant à des génotypes différents, et représentant des niveaux de ploïdie différents. Une publication est en cours de préparation (non-soumise) et les principaux résultats et conclusions sont détaillés dans ce chapitre.

	Espèces	Clones BAC	Nom usuel		Genbank	Référence des banques BAC	Individu d'origine
			Clone	Haplotype			
Génome A	*Triticum monococcum* ssp. *monococcum*	109N23	Mono	Mono	AY491681	Lijavetzky *et al.* 1999	DV92
	Triticum urartu	**Ble-EAB-UrH2a14-8K2Ha**	**8K2**	**8K2-8K13**	**-**	**Unpublished**	**TMU138**
	Triticum urartu	**Ble-EAC-UrH2a16-8K13Ha**	**8K13**		**-**	**Unpublished**	**TMU138**
	Triticum urartu	**Ble-GAC-TU_G1812-114L16**	**114L16**	**114L16**	**-**	**Luo and Dvorak, unpublished**	**G1812**
	Triticum turgidum ssp. *durum**	542K11	DurumA	DurumA	CR626933	Cenci *et al.* 2003	Langdon65
	Triticum aestivum ssp. *aestivum***	213F23	RenanA	RenanA	CR626929	Genoplante consortium	Renan
	Triticum aestivum* ssp. *aestivum**	**BAC_Ble_CS259J18HaA**	**CSA**	**CSA**	**-**	**Allouis *et al.* 2003**	**Chinese Spring**
Génome B	*Triticum turgidum* ssp. *durum**	545A13	DurumB	DurumB	CR626932	Cenci *et al.* 2003	Langdon65
	Triticum aestivum ssp. *aestivum***	1793L02	RenanB	RenanB	CR626930	Genoplante consortium	Renan
	Triticum aestivum* ssp. *aestivum**	**BAC_Ble_CS46I23HaB**	**CSB**	**CSB**	**-**	**Allouis *et al.* 2003**	**Chinese Spring**
Génome S	***Aegilops speltoides***	**SAA-ble-Sho44-6H2Ha**	**SAA**	**SAA-SAB**	**-**	**Unpublished**	**Spelt1**
	Aegilops speltoides	**Ble-SAB-Sh128-17J1Ha**	**SAB**		**-**	**Unpublished**	**Spelt1**
	Aegilops speltoides	**Ble-GAA-S134103H01**	**GAA**	**GAA**	**-**	**Akhunov et *al.* 2005**	**134**
Génome D	***Aegilops tauschii***	**Ble-GAD-TAS75137E09**	**137E09**	**137E09**	**-**	**Akhunov et *al.* 2005**	**AS75**
	Aegilops tauschii	BAC10	**Lagudah**	**Lagudah**	CR626926	Moullet *et al.* 1999	Aus18913
	Aegilops tauschii	**L-taus7.2p1**			-	**Moullet *et al.* 1999**	**Aus18913**
	Aegilops tauschii	**BAC_Ble_D_HB041M6**	**41M6**	**41M6**	**-**	**Akhunov et *al.* 2005**	**Tausch2**
	Triticum aestivum ssp. *aestivum***	1611A10	RenanD	RenanD	CR626934	Genoplante consortium	Renan
	Triticum aestivum* ssp. *aestivum**	**BAC_Ble_CS361O08HaD**	**CSD**	**CSD**	**-**	**Allouis *et al.* 2003**	**Chinese Spring**

Tableau II-1. Liste des clones BAC séquencés pour l'étude avec leur référence GenBank quand elle est disponible. Les séquences des clones en noir étaient déjà disponibles, les autres ont été nouvellement séquencées. Avant d'être séquencé par le CNS, les clones en bleu ont été isolés au laboratoire et ceux en vert ont été fournis par des collaborateurs. Les noms usuels des clones BACs et des différents haplotypes correspondent aux dénominations utilisées pour les résultats dans un soucis de lisibilité. Les publications présentant les banques BAC d'où sont issues les clones, et les individus à leur origine sont également indiqués. * tetraploïde **hexaploïde

II Matériels et méthodes

II.1 Ressources génomiques

L'utilisation des séquences génomiques disponibles pour ce locus *Ha* (Chantret *et al.* 2005) (8 clones BAC) et le séquençage complémentaire de 12 clones BAC portant cette région dans différents génotypes des espèces *T. urartu*, *Ae. speltoides*, *Ae. tauschii* et *T. aestivum*, a abouti à la comparaison des séquences génomiques d'un total de 16 haplotypes différents (20 clones BAC) (Tableau II-1) couvrant le locus *Ha* dans les génomes A, B, S et D. Pour simplifier la présentation des résultats, nous avons désigné les séquences et les haplotypes par des noms usuels. Nous avons au final comparé six haplotypes du génome A [3 diploïdes (Mono, 8K2-8K13 et 114L16), 1 tétraploïde (DurumA) et 2 hexaploïdes (RenanA et CSA)], trois haplotypes du génome B [1 tétraploïde (DurumB) et 2 hexaploïdes (RenanB et CSB)], deux haplotypes diploïdes du génome S (SAA-SAB, GAA) et cinq haplotypes du génome D [3 diploïdes (137E09, Lagudah, 41M61) et 2 hexaploïdes (RenanD et CSD)].

Sept des douze 'nouveaux' clones BAC ont été isolés, par PCR selon la stratégie d'Isidore *et al.* (2005), sur des banques BACs construites par le laboratoire qui couvrent les espèces *T. aestivum* ssp. *aestivum* (clones CSA, CSB et CSD), *T. urartu* (clones 8K2 et 8K13) et *Ae. speltoides* (clones SAA et SAB) (Tableau II-1, en bleu). Deux de ces banques BACs (celles couvrant *T. urartu* et *Ae. speltoides*) ont été créées spécialement afin de réaliser le projet de séquençage comparatif (APCNS 2003). Les 5 autres clones BAC ont été fournis par des collaborateurs internationaux (Tableau II-1, en vert). A. Akhunov (UC, Davis, USA) nous a fourni 4 clones venant de banques BACs réalisées à partir des génotypes évalués comme les plus proches des polyploïdes pour les génomes A, B et D [*T. urartu* (clone 114L16), *Ae. speltoides* (clone GAA) et *Ae. tauschii* (clone 41M6)] et d'un génotype de *Ae tauschii* (137E09) évalué comme le plus éloigné du génome D (Akhunov *et al.* 2005). Finalement, E. Lagudah nous a fourni le clone L-tau7.2p1, chevauchant avec le clone (BAC10) du même génotype. Les séquences de ces deux clones ont ensuite été regroupées en une seule.

II.2 Annotation et analyse des séquences génomiques

II.2.1 Annotation et comparaison de séquences

L'annotation des gènes et des éléments transposables a été réalisée de la même façon que pour l'Article 1. Cette méthode est détaillée précisément dans le chapitre général 'Matériels et Méthodes d'annotation' de cette thèse. Chantret *et al.* (2005, 2008) ont fait une première étude comparative du locus *Ha* en se basant sur la séquence de 8 clones BAC représentant 8 haplotypes des génomes A, B et D. Mon analyse comparative porte sur la séquence de 16 haplotypes, et a consisté à confirmer, étendre et préciser les observations sur la conservation des gènes précédemment décrite dans Chantret *et al.* (2005, 2008) et analyser la dynamique (insertion et élimination) de l'espace TEs en comparant les différents haplotypes dans chacun des 4 génomes.

II.2.2 Méthode de datation

J'ai utilisé une combinaison d'approches pour apprécier les dates de divergence des différents génomes et haplotypes analysés. Les divergences inter-génomiques ont été évaluées sur la base de comparaisons des séquences des gènes, en utilisant une horloge moléculaire de $4{,}9 \times 10^{-9}$ substitutions/site/an (Chalupska *et al.* 2008). Les divergences intra-génomiques et intra-spécifiques ont été calculées sur la base des comparaisons de l'espace TEs encore conservé, par la méthode de Kimura 2 paramètres (Kimura 1980) implémentée dans le logiciel MEGA4 (Kumar *et al.* 2004) et en utilisant une horloge moléculaire de $1{,}3 \times 10^{-8}$ substitutions/site/an pour l'espace TEs et pour les deux LTRs d'un même rétrotransposon (SanMiguel *et al.* 1998, Ma *et al.* 2004, Ma et Bennetzen 2004, Wicker *et al.* 2005, Gu *et al.* 2006).

II.2.3 Calcul des taux de remplacement de l'espace TEs

Nous avons utilisé la méthode de Dubcowsky et Dvorak (2007) pour calculer les taux de remplacements. Pour une région donnée, on obtient le taux de remplacement entre deux haplotypes (haplotype1 et haplotype2) en divisant la somme des longueurs des segments inter-géniques non communs dans les haplotype1 et haplotype2 par la somme des longueurs totales des segments (conservés et non conservés) de ces deux haplotypes.

II.2.4 Confirmation et traçage par PCR des principaux événements de réarrangement

Un effort particulier s'est porté sur le développement de marqueurs PCR afin de pouvoir confirmer les réarrangements observés dans les génotypes étudiés. Les oligonucléotides des PCRs ont été dessinés de façon à couvrir les 'breakpoints' détectés en comparant les différents haplotypes. Harry Belcram, ingénieur au laboratoire, s'est chargé de ce développement, de leur utilisation et de l'analyse des résultats des PCRs obtenus.

III Résultats

Nous disposions à l'origine de la séquence de 8 clones BAC couvrant la région *Ha* (la région contenant le locus *Ha*) dans les génomes A, B et D du blé (Chantret *et al.* 2005, 2008). Pour préciser les analyses comparatifs du locus *Ha*, nous avons séquencé 12 autres clones BAC pour obtenir au final 16 haplotypes différents des génomes A (6), B (3), S (2) et D (5) représentés par un ou deux clones BAC (Tableau II-1).

Nous avons donc analysé la variabilité génomique entre ces différents haplotypes aux niveaux inter-génomique, intra-génomique (le même génome à différents niveaux de ploïdie) et intra-spécifique. Les comparaisons sur ces deux derniers niveaux d'analyse sont complémentaires et sont présentées conjointement. Nous avons comparé l'espace gènes (taux de SNPs, divergence) et l'espace TEs (taux de SNPs, divergence, taux de remplacement) dans chacune de ces comparaisons.

III.1 Analyse de la variabilité inter-génomique

La conservation entre les séquences des génomes A, B, S et D concerne uniquement quelques gènes : l'espace TEs est complètement différent (Figures II-1, II-3, II-4, présentées ci-après, en même temps que leurs analyses respectives) comme cela a été précédemment décrit (Chantret *et al.* 2005, Isidore *et al.* 2005, Dvorak *et al.* 2006, Gu *et al.* 2006). L'étude de la variabilité haplotypique inter-génomique est donc restreinte à l'espace gènes.

Nous avons identifié 21 gènes différents représentant 198 copies sur l'ensemble des séquences des 16 haplotypes. Les comparaisons montrent que seuls les gènes *BGGP* (codant pour la Beta-1-3-galactosyl-O-glycosyl-glycoprotéine), *Gsp-1* (Grain Softness Protein-1), *AAA-ATPase* (AAA-ATPase) et *Nodulin* (Nodulin) sont conservés dans tous les haplotypes des 4 génomes. Les gènes d'*AAA-ATPase* et de *Nodulin* sont présents en cluster de copies dupliquées en tandem (Figures II-1, II-3, II-4). Les gènes dupliqués en tandem ont été exclus de la comparaison. Leur évolution est nettement plus dynamique que la plupart des autres gènes, comme l'atteste le nombre de copies tronquées ou 'pseudoisées' des gènes *AAA-ATPase* et *Nodulin* présentes dans les séquences rendant difficile de préciser les relations d'orthologie entre ces copies. Les gènes *Unknown-2*, *CHS* (Chalcone synthase) et *VAMP*

Génomes comparés			*BGGP (1278 pb)*		*Gsp-1 (492 pb)*		*CHS (1251 pb)*		*VAMP (648 pb)*	
			SNP/kb	Divergence (Ma)	SNP/kb	Divergence (Ma)	SNP/kb	Divergence (Ma)	SNP/kb	Divergence (Ma)
CSA	/	CSB	24,3	8,02	56,9	10,10	-	-	-	-
CSA	/	CSD	21,9	7,29	50,8	6,88	33,8	7,6	34,1	8,9
CSB	/	CSD	21,1	7,29	52,8	10,10	-	-	-	-
RenanA	/	RenanB	24,3	8,02	56,9	10,10	-	-	-	-
RenanA	/	RenanD	21,9	7,29	50,8	6,88	-	-	34,1	8,9
RenanB	/	RenanD	21,1	7,29	52,8	10,10	-	-	-	-
DurumA	/	DurumB	27,4	8,65	54,9	10,10	-	-	-	-
CSA	/	SAA	24,3	7,29	48,8	6,88	-	-	-	-
CSB	/	SAA	20,3	7,29	24,4	4,90	-	-	-	-
CSD	/	SAA	21,1	7,29	44,7	7,92	-	-	-	-
RenanA	/	SAA	24,3	7,29	48,8	6,88	-	-	-	-
RenanB	/	SAA	20,3	7,29	24,4	4,90	-	-	-	-
RenanD	/	SAA	21,1	7,29	44,7	7,92	-	-	-	-
DurumA	/	SAA	26,6	8,02	46,7	6,88	-	-	-	-
DurumB	/	SAA	21,1	7,29	24,4	4,90	-	-	-	-

Tableau II-2. Comparaison de la séquence des gènes *BGGP*, *Gsp-1*, *CHS* et *VAMP* entre les différents génomes A, B et D des polyploïdes mais aussi entre ces gènes et ceux du génome S. L'estimation des temps de divergence est obtenue par le calcul de la divergence avec la méthode de Kimura à 2 paramètres (Kimura 1980) utilisant un taux de $4,9x10^{-9}$ substitutions / site / an (Chalupska *et al.* 2008).

(Vesicule Associated Membran Protein) ne sont pas présents dans les génomes B et S, mais ont été trouvés en commun dans certains haplotypes des génomes A et D (Figures II-1, II-3, II-4). Nous avons utilisé leur comparaison en complément, après avoir exclu le gène *Unknown-2* trouvé dupliqué en tandem dans le génome A (Figure II-1).

Au final, nous avons comparé les gènes *BGGP* et *Gsp-1* entre les différents haplotypes des 4 génomes ainsi que *CHS* et *VAMP* entre les génomes A et D. La comparaison précise des copies de ces gènes m'a permis de calculer leur divergence et d'estimer les dates relatives de radiation des différents génomes en utilisant un taux de mutation de $4{,}9 \times 10^{-9}$ mutations / site / an (Chalupska *et al.* 2008) (Tableau II-2). Ces estimations varient, en fonction des gènes et des génomes comparés, de 6,9 Ma à plus de 10 Ma, ce qui est deux à trois fois plus important que la divergence de 2,5-4 Ma couramment admise (Huang *et al.* 2002). De façon intéressante, nos estimations sont très similaires à celles obtenues pour la comparaison des gènes *Acc* (acetyl-CoA carboxylase) qui donnaient une divergence entre les génomes A, B et D comprise entre 2 et 9 Ma (Chaluspka *et al.* 2008), mais qui suggéraient que l'estimation de 2,5-4 Ma était plus fiable. De la même façon, la faible taille de notre échantillon (2 à 4 gènes de moins de 1,5 kb) n'est pas suffisante pour être représentative et les gènes considérés ont probablement évolué plus rapidement que la moyenne des gènes. Nous avons donc représenté la divergence de ces principaux génomes en calibrant leur origine commune à environ 2,5 Ma couramment utilisée (Huang *et al.* 2002, Chalupska *et al.* 2008) (Figure II-6, présentée dans la conclusion). Cette figure est présentée en conclusion car elle inclut l'ensemble des divergences trouvées dans cette étude.

Sur la base des comparaisons des deux gènes *BGGP* et *Gsp-1*, le génome A semble légèrement plus proche du génome D que du génome B. Un résultat similaire a été trouvé au niveau du locus *SPA* (Salse *et al.* 2008b). La comparaison des séquences du génome B avec celles du génome S, connu comme le génome diploïde le plus proche, ont montré un espace TEs complètement différent et un taux de SNPs extrêmement élevé (Tableau II-2) pour les gènes *BGGP* et *Gsp-1* détectés en commun. Les dates de divergence estimées suggèrent que les deux génomes B et S sont bien différents et ont divergé peu après la séparation de leur ancêtre avec les génomes A et D (Tableau II-2).

Comme attendu, les gènes *Pina* et *Pinb* ont été identifiés dans la plupart des génomes des blés diploïdes (quand les clones BAC séquencés les couvraient) et dans les génomes D du blé hexaploïde mais pas dans les génomes A et B des blés polyploïdes où ils ont été délétés (Chantret *et al.* 2005, Figures II-1, II-3, II-4).

III.2 Analyse de la variabilité intra-génomique et intra-spécifique

Les différents haplotypes de chacun des génomes A, B, S et D ont été comparés séparément. A cette échelle de divergence (< 4 Ma), on observe une forte conservation de l'espace gènes, mais également la conservation d'une proportion variable de l'espace TEs. Les espèces *T. aestivum*, *T. urartu*, *Ae. speltoides* et *Ae. tauschii* sont représentées par plusieurs haplotypes que nous avons comparés aux niveaux intra-génomique et intra-spécifique. L'alignement des séquences des différents génomes A, B, S et D sont présentés dans les Figures II-1, II-3 et II-4. Cette étude m'a permis de déterminer l'étendue des réarrangements liés aux insertions et éliminations des TEs et parfois les bases moléculaires des mécanismes correspondants, quand les séquences étaient assez conservées. Les différentes comparaisons de séquences m'ont ainsi permis d'évaluer les taux de SNPs et la divergence entre les espaces TEs et gènes ainsi que le taux de remplacement de l'espace TEs. De plus, de brusques variations du taux de SNPs le long de la séquence permettent de détecter la présence d'haplotypes 'mixtes' ayant pour origine une recombinaison génétique.

III.2.1 Variabilité haplotypique du génome A

Comparaison des séquences géniques

Nous avons comparé les gènes *BGGP*, *Gsp-1* et *VAMP* qui sont présents dans les six haplotypes du génome A. Les gènes sont globalement très proches et nous avons préféré présenter leur divergence par le nombre de SNPs trouvés entre les gènes dans les comparaisons deux à deux. Les gènes des haplotypes 114L16, 8K2-8K13, DurumA, RenanA et CSA sont extrêmement proches entre eux (Tableau II-3A). La distance entre les gènes de ces haplotypes et ceux de Mono est significativement plus importante (Tableau II-3A). Cette divergence plus élevée d'un haplotype de *T. monococcum* est attendue, puisque c'est une espèce distante de *T. urartu* (l'espèce progénitrice du génome A dans les polyploïdes).

Les gènes *CHS* et *Unknown-1* ont pu être comparés entre l'haplotype diploïde 8K2, tétraploïde Durum et hexaploïde CS. La comparaison de *Unknown-1* entre le tétraploïde et l'hexaploïde donne sensiblement le même nombre de SNPs que ceux observés pour *BGGP*, *Gsp-1* et *VAMP* (Tableau II-3A). Par contre, la copie *Unknown-1* du diploïde 8K2 est très divergente comparée aux copies des polyploïdes (Tableau II-3A). Les nombres de SNPs sont

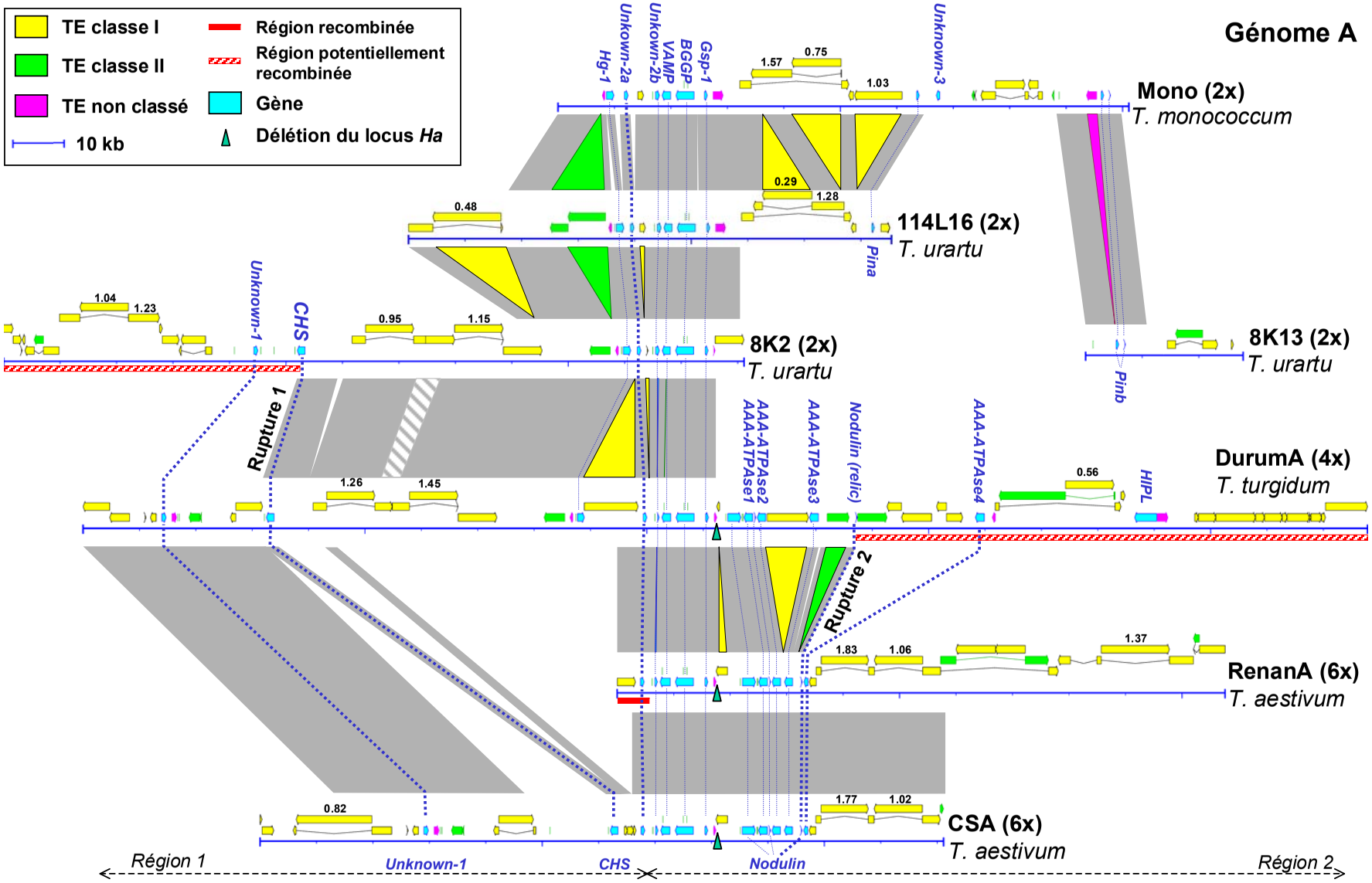


Figure II-1. Comparaison des régions *Ha* des différents haplotypes du génome A. Les portions grisées représentent des séquences similaires. Les gènes orthologues sont reliés par des traits pointillés bleu, plus épais pour les gènes d'un intérêt particulier. Les gènes *Unknown-1* de 8K2 et *AAA-ATPase4* de DurumA ont été reliés à leurs homologues. Les insertions ou les délétions de grande taille (>500 pb) sont indiquées par les triangles de la couleur correspondant au type de la séquence impliquée. La partie grisée hachurée indique une zone avec un probable problème d'assemblage dans la séquence de 8K2 et qui n'a donc pas été considérée. Les rectangles rouges (hachurés rouge) représentent les régions recombinées (potentiellement recombinées).

même bien plus élevés que pour ceux trouvés dans les comparaisons avec les gènes de Mono. Les comparaisons avec le gène *CHS* montrent également des nombres de SNPs plus élevés que ceux observés pour les autres gènes (Tableau II-3A).

Les autres gènes trouvés dans les séquences du génome A présentant des copies tronquées (*Hg-1*) et/ou dupliquées en tandem (*Nodulin*, *AAA-ATPase*) n'ont pas été comparés pour éviter des cas de divergences spécifiques. Les gènes *Pina* et *Pinb* sont absents des haplotypes polyploïdes du génome A du blé à la suite de la délétion du locus *Ha* (Figure II-2, Chantret *et al.* 2005) et n'ont donc pas été comparés.

Les divergences calculées pour les gènes sont globalement si faibles qu'elles ne permettent pas de préciser les dates de séparation des différents haplotypes, même si les gènes de Mono apparaissent clairement comme les plus divergents. Je me suis donc basé sur la divergence des séquences de l'espace TEs encore conservé pour estimer les divergences entre ces haplotypes.

Comparaison des régions inter-géniques

Nous avons donc comparé l'ensemble des différents haplotypes du génome A en séparant ces comparaisons en plusieurs parties. Nous avons ainsi comparé les haplotypes diploïdes entre eux puis avec le tétraploïde et enfin le tétraploïde et les hexaploïdes entre eux. Nous n'avons malheureusement pas pu comparer les haplotypes diploïdes et hexaploïdes car ils ne couvrent que très peu de séquences inter-géniques en commun (<500 bp). Pour l'ensemble des comparaisons, nous avons estimé les taux de SNPs, la divergence et les taux de remplacement (Tableau II-3B).

Comparaison des haplotypes diploïdes (8K2, 114L16 et Mono)

Les séquences inter-géniques des haplotypes 8K2 et 114L16 montrent un taux de SNPs de 13,5 SNPs / kb et une divergence de 0,54 Ma (Tableau II-3B). La divergence entre Mono et celles de ces deux haplotypes est deux fois plus importante (Tableau II-3B), correspondant effectivement à une radiation il y a environ 1 Ma (Huang *et al.* 2002) des deux espèces *T. urartu* et *T. monococcum* portant le génome A.

Les taux de remplacement observés en comparant les haplotypes diploïdes (de 34,3 à 41,8 %) sont dus en majeure partie à des insertions différentielles d'éléments transposables (de 32,4 à 40,2 %) (Tableau II-3B).

A

	BGGP (1278 pb)					*Gsp-1* (492 pb)					*VAMP* (648 pb)					*CHS* (1251 pb)					*Unknown-1* (774 pb)				
	Mono	114L16	8K2	DurumA	CSA	Mono	114L16	8K2	DurumA	CSA	Mono	114L16	8K2	DurumA	CSA	Mono	114L16	8K2	DurumA	CSA	Mono	114L16	8K2	DurumA	CSA
Mono (*T. monoccocum*)																									
114L16 (*T. urartu*)	6					8					12					-					-				
8K2 (*T. urartu*)	9	3				9	1				14	2				-	-				-	-			
DurumA (*T. turgidum*)	8	2	5			10	2	3			11	1	3			-	-	19			-	-	86		
CSA (*T. aestivum*)	5	1	4	3		11	3	4	1		12	2	4	1		-	-	11	12		-	-	89	3	
RenanA (*T. aestivum*)	5	1	4	3	0	11	3	4	0	1	12	2	4	0	1	-	-	-	-	-	-	-	-	-	-

B

	Région commune	Région alignée	SNP / kb	Divergence (Ma)	Taux de remplacement						
					Total (%)	**TEs** (insertions)		**Réarrangements** (délétions)		**Petits indels**	
						Nb	%	Nb	%	Nb	%
Mono / 114L16	89870	26668	27,4	1,08	41,8%	5	40,2%	5	1,5%	82	0,2%
Mono / 8K2	27171	8532	31,3	1,23	34,3%	2	32,4%	2	1,7%	34	0,3%
114L16 / 8K2	60505	17491	13,5	0,54	40,8%	3	39,3%	5	1,5%	19	0,0%
Mono / DurumA	104806	8553	29,9	1,18	84,1%	2	12,6%	4	71,4%	39	0,1%
114L16 / DurumA	138140	17530	13,6	0,53	75,0%	3	20,4%	7	54,5%	25	0,0%
8K2 / DurumA	220995	47813	17,6	0,68	54,9%	2	6,6%	9	48,1%	54	0,1%
DurumA / RenanA	124283	5084	variable	variable	71,4%	2	9,0%	2	62,4%	1	0,0%
DurumA / CSA	175903	32872	6,1	0,23	64,8%	2	6,4%	5	58,2%	32	0,2%
RenanA / CSA	58002	28833	variable	variable	4,9%	0	0,0%	1	4,8%	6	0,1%

C

	Région 1			**Région 2**		
	Région commune	**SNP / kb**	**Divergence (Ma)**	**Région commune**	**SNP / kb**	**Divergence (Ma)**
RenanA / DurumA	3236	0,0	0,00	1848	7,0	0,27
RenanA / CSA	443	18,1	0,81	28390	0,5	0,02

Tableau II-3. Comparaison de la séquence des gènes et de l'espace TEs du génome A. (A) Nombre de SNPs trouvés dans les comparaisons deux à deux des gènes dans les différents haplotypes du génome A. (B) Comparaison des séquences de l'espace TEs du génome A, avec le taux de SNPs, la divergence et les taux de remplacement. Ces derniers sont séparés en trois : importance de l'insertion des TEs, des réarrangements de grande taille (>50 pb) et les petits indels (<50pb). (C) Comparaison des séquences de l'espace TEs par région, dans le cas où le taux de SNPs fluctue de façon importante le long des séquences.

Comparaison des haplotypes diploïdes à celui du tétraploïde (8K2, 114L16 et Mono contre DurumA)

La séquence de Mono est la plus divergente des diploïdes comparées à celle du tétraploïde DurumA, montrant un taux de SNPs de 29,9 SNPs / kb, et une date de divergence estimée à 1,18 Ma. Les comparaisons 8K2/DurumA et 114L16/DurumA montrent un taux de SNPs allant de 13,6 SNPs / kb à 17,6 SNPs / kb et un temps de divergence estimé de 0,53 Ma et 0,68 Ma respectivement. En se basant sur cette analyse, 114L16 semble légèrement plus proche du tétraploïde DurumA que 8K2 et Mono est approximativement deux fois plus divergent de DurumA que 114L16 et 8K2 (Tableau II-3B).

Nous avons remarqué une soudaine rupture de colinéarité entre les séquences de 8K2 et de DurumA en 5' du gène *CHS* (Figure II-2, Rupture 1). Les séquences d'ADN sont en effet complètement différentes sur 51.154 pb dans 8K2 et 31.838 pb dans DurumA. Cette partie de la séquence de DurumA est couverte par la séquence d'un autre haplotype (l'hexaploïde CSA) contrairement à celle de 8K2 qui n'est couverte par aucune autre séquence. Cette rupture de colinéarité pourrait correspondre à une large insertion / délétion qui ne serait pas entièrement couverte par la séquence des clones BAC ou à une recombinaison génétique ayant apporté dans 8K2 une portion de séquence d'un haplotype très divergent.

Le gène *Unknown-1*, identifié dans la séquence de 8K2, montre plus de 90% de similarité avec les gènes *Unknown-1* de DurumA et de l'hexaploïde CSA. Pour expliquer cette rupture de colinéarité par des insertions / délétions, il faut donc faire intervenir au moins deux événements de part et d'autre du gène. De plus, les niveaux de divergence pour ce gène dans les comparaisons 8K2/DurumA et 8K2/CSA sont nettement plus élevés que pour les gènes *BGGP*, *Gsp-1* et *VAMP* (Tableau II-3A). Ces résultats suggèrent plutôt que cette rupture de colinéarité correspond à une recombinaison génétique entre 8K2 et un haplotype très divergent du tétraploïde DurumA. Il faudrait néanmoins une séquence d'un haplotype diploïde présentant une configuration similaire à celle des polyploïdes pour cette région pour confirmer cette hypothèse.

Pour calculer les taux de remplacement, nous avons pris en considération la délétion du locus *Ha* ainsi que la perte de colinéarité évoquée précédemment. Nous les avons respectivement assimilés à une délétion d'au moins 73.397 pb en se basant sur la taille du locus *Ha* dans Mono et à un réarrangement d'au moins 31.838 pb correspondant à la plus petite des deux séquences non-colinéaires. Les taux de remplacement 'minimums' ainsi

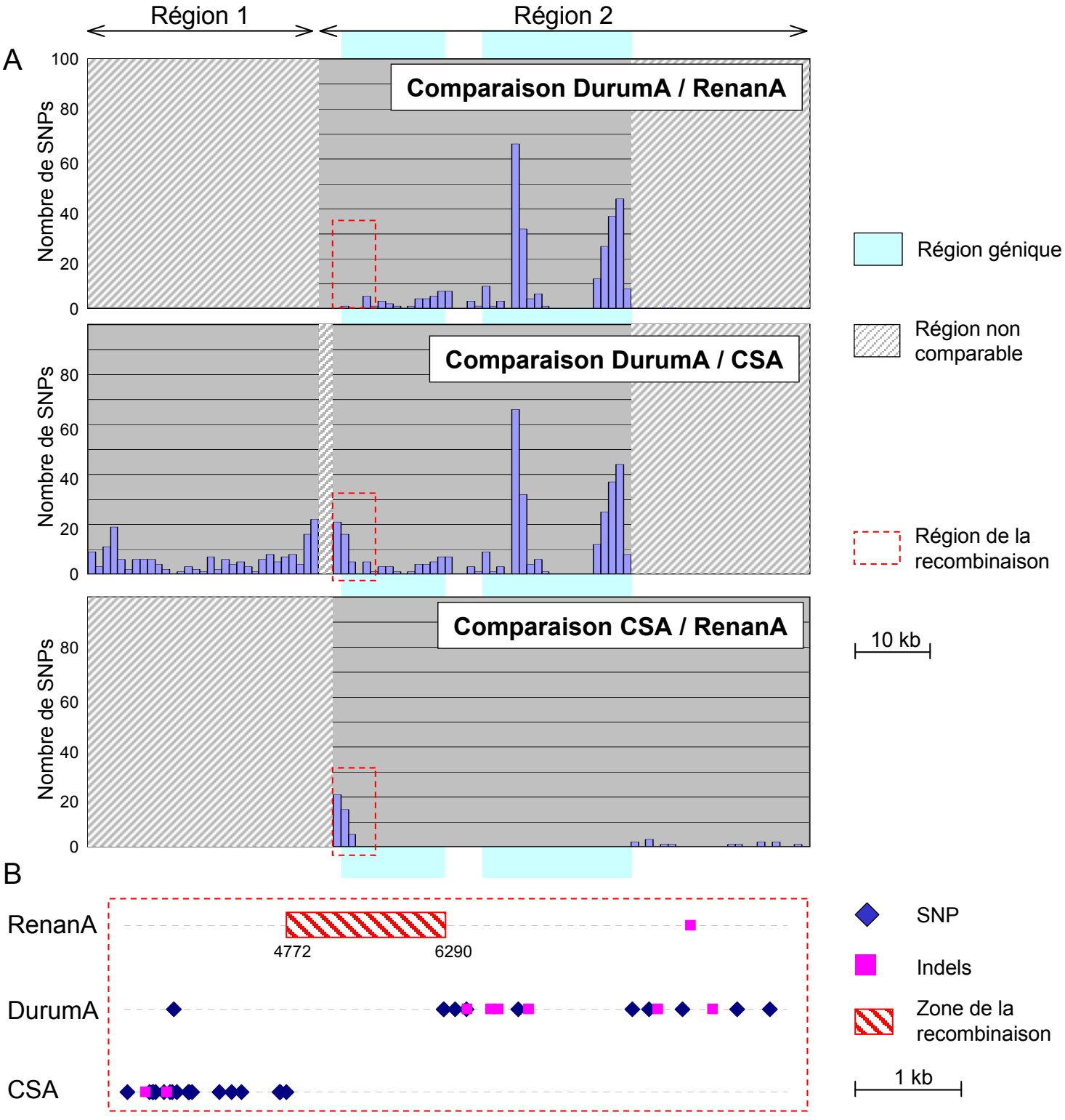


Figure II-2. Variation de la distribution des SNPs dans l'espace TEs de 3 haplotypes polyploïdes du génome A. (A) Les histogrammes montrent le nombre de SNPs dans des fenêtres de 1000 pb entre les séquences trouvées en commun (sans les indels). Ces comparaisons portent à la fois sur les espaces géniques (zones bleues) et TEs (le reste). Les parties hachurées représentent les régions non comparables. La région des rectangles pointillés est détaillée dans la partie (B) (B) Détail des SNPs et des Indels par séquence autour de la zone de recombinaison, indiquée par un rectangle rouge avec les coordonnées correspondantes de la séquence de RenanA.

recombinaison entre les coordonnées 4.772 et 6.290 sur le clone BAC de Renan, soit sur 1,5 kb juste après le gène *Unknown-2a* (Figure II-2B).

Nous avons également observé une rupture de colinéarité supplémentaire avec les comparaisons DurumA/CSA et DurumA/RenanA (Figure II-1, Rupture 2). D'une façon très similaire à ce qu'on a précédemment observé entre 8K2 et DurumA, les 90.422 bp de DurumA et les 74.800 bp de RenanA situées en 3' de la relique de *Nodulin* sont complètement différentes et un gène d'*AAA-ATPase* est trouvé en commun entre les zones 'indépendantes'. Cette partie de la séquence de RenanA est partiellement couverte (25.012 bp) par la séquence de CSA. Dans les hexaploïdes, on trouve un cluster de 4 gènes d'*AAA-ATPases* suivant un gène de *Nodulin* (Figure II-1). La quatrième copie est incomplète, mais est très proche des autres copies du cluster. Dans le tétraploïde, la quatrième copie d'*AAA-ATPase* trouvée dans la région 'indépendante' est complète et relativement divergente des autres *AAA-ATPase* du cluster. Cela suggère de nouveau une recombinaison génétique qui reste à confirmer en mettant en évidence une séquence de diploïde ou de tétraploïde présentant une configuration similaire à celles des hexaploïdes.

Les taux de remplacement entre le tétraploïde et les hexaploïdes, prenant en compte la rupture de colinéarité de la même façon que précédemment, sont très élevés (64,8% et 71,4%). Les insertions différentielles d'éléments transposables sont toujours présentes et comptent pour 6,4% et 9,0% des remplacements, mais ce sont les réarrangements qui ont joué un rôle prédominant dans ces taux de remplacements. En effet, il y a d'une part la rupture de colinéarité, mais aussi des délétions importantes dans la région 1 : 2 dans CSA (6.604 pb et 53.630 bp) et 1 dans DurumA (14.400 pb) (Figure II-1). La séquence de RenanA ne couvre pas entièrement ces délétions, mais la présence de l'élément transposable de la famille *Inga*, au début de la séquence, sous une forme non tronquée, suggère que la délétion de 53.630 bp de CSA n'est pas dans RenanA. Nous avons pu confirmer par PCR sur l'ADN génomique de *T. aestivum* ssp. *aestivum cv* Renan que ces délétions n'étaient effectivement pas présentes dans cet haplotype, confortant encore une fois l'hypothèse d'une recombinaison génétique.

III.2.2 Variabilité haplotypique des génomes B et S

A la différence des génomes A et D, l'espèce progénitrice du génome B n'est toujours pas identifiée. La plus proche espèce connue du progéniteur du génome B, *Ae. speltoides*

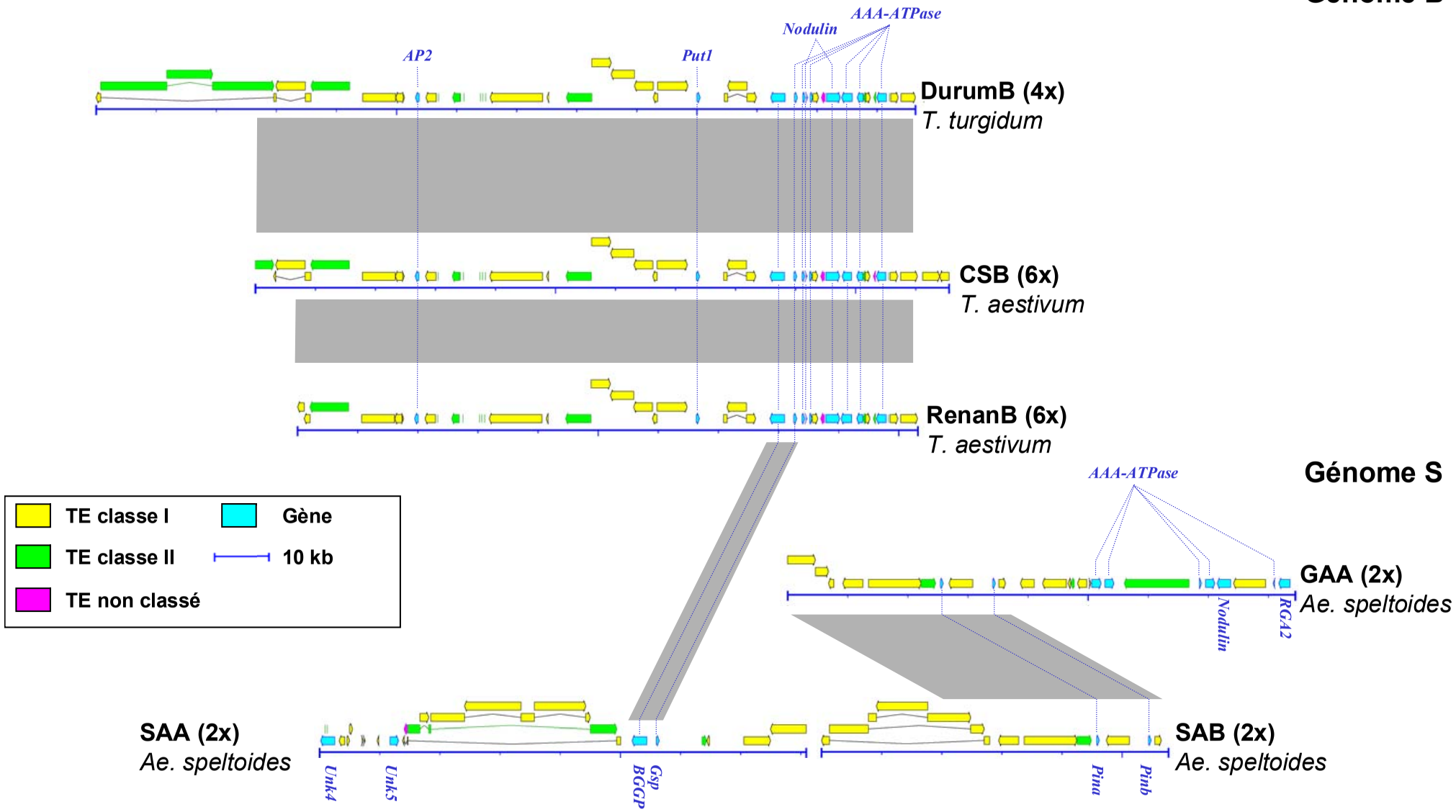


Figure II-3. Comparaison des régions *Ha* des différentes haplotypes du génome B et S. Les portions grisées représentent des séquences similaires. Les gènes orthologues sont reliés par des traits pointillés bleu.

recombinaison entre les coordonnées 4.772 et 6.290 sur le clone BAC de Renan, soit sur 1,5 kb juste après le gène *Unknown-2a* (Figure II-2B).

Nous avons également observé une rupture de colinéarité supplémentaire avec les comparaisons DurumA/CSA et DurumA/RenanA (Figure II-1, Rupture 2). D'une façon très similaire à ce qu'on a précédemment observé entre 8K2 et DurumA, les 90.422 bp de DurumA et les 74.800 bp de RenanA situées en 3' de la relique de *Nodulin* sont complètement différentes et un gène d'*AAA-ATPase* est trouvé en commun entre les zones 'indépendantes'. Cette partie de la séquence de RenanA est partiellement couverte (25.012 bp) par la séquence de CSA. Dans les hexaploïdes, on trouve un cluster de 4 gènes d'*AAA-ATPases* suivant un gène de *Nodulin* (Figure II-1). La quatrième copie est incomplète, mais est très proche des autres copies du cluster. Dans le tétraploïde, la quatrième copie d'*AAA-ATPase* trouvée dans la région 'indépendante' est complète et relativement divergente des autres *AAA-ATPase* du cluster. Cela suggère de nouveau une recombinaison génétique qui reste à confirmer en mettant en évidence une séquence de diploïde ou de tétraploïde présentant une configuration similaire à celles des hexaploïdes.

Les taux de remplacement entre le tétraploïde et les hexaploïdes, prenant en compte la rupture de colinéarité de la même façon que précédemment, sont très élevés (64,8% et 71,4%). Les insertions différentielles d'éléments transposables sont toujours présentes et comptent pour 6,4% et 9,0% des remplacements, mais ce sont les réarrangements qui ont joué un rôle prédominant dans ces taux de remplacements. En effet, il y a d'une part la rupture de colinéarité, mais aussi des délétions importantes dans la région 1 : 2 dans CSA (6.604 pb et 53.630 bp) et 1 dans DurumA (14.400 pb) (Figure II-1). La séquence de RenanA ne couvre pas entièrement ces délétions, mais la présence de l'élément transposable de la famille *Inga*, au début de la séquence, sous une forme non tronquée, suggère que la délétion de 53.630 bp de CSA n'est pas dans RenanA. Nous avons pu confirmer par PCR sur l'ADN génomique de *T. aestivum* ssp. *aestivum cv* Renan que ces délétions n'étaient effectivement pas présentes dans cet haplotype, confortant encore une fois l'hypothèse d'une recombinaison génétique.

III.2.2 Variabilité haplotypique des génomes B et S

A la différence des génomes A et D, l'espèce progénitrice du génome B n'est toujours pas identifiée. La plus proche espèce connue du progéniteur du génome B, *Ae. speltoides*

A

	Pina (444 pb)		*Pinb* (444 pb)	
	SAA	SAB	SAA	SAB
SAA (*Ae. Speltoides*)				
SAB (*Ae. Speltoides*)	-		-	
GAA (*Ae. Speltoides*)	-	4	-	11

B

	Région commune	Région alignée	SNP / kb	Divergence (Ma)	Taux de remplacement					
					Total (%)	**TE (insertions)**		**Réarrangements (délétions)**		**Petits indels**	
						Nb	%	Nb	%	Nb	%
GAA / SAB	33376	31657	21,4	0,83		0	0,0%				

C

	BGGP (1278 pb)		*Gsp-1* (492 pb)	
	Durum	CS	Durum	CS
DurumB *(T. turgidum)*				
CSB *(T. aestivum)*	1		0	
RenanB *(T. aestivum)*	1	0	0	0

D

	Région commune	Région alignée	SNP / kb	Divergence (Ma)	Taux de remplacement						
					Total (%)	**TE (insertions)**		**Réarrangements (délétions)**		**Petits indels**	
						Nb	%	Nb	%	Nb	%
DurumB / RenanB	152391	76182	0,5	0,02	0,0%	0	0,0%	0	0,0%	14	0,0%
DurumB / CSB	166727	83352	0,2	0,01	0,0%	0	0,0%	0	0,0%	15	0,0%
RenanB / CSB	152392	76186	0,5	0,02	0,0%	0	0,0%	0	0,0%	13	0,0%

Tableau II-4. Comparaison de la séquence des gènes et de l'espace TEs des génomes B et S. (A),(C) Nombre de SNPs trouvés dans les comparaisons deux à deux des gènes dans les différents haplotypes du génome B et S. (B),(D) Comparaison des séquences de l'espace TEs du génome B et S, avec le taux de SNPs, la divergence et les taux de remplacement. Ces derniers sont séparés en trois : importance de l'insertion des TEs, des réarrangements de grande taille (>50 pb) et les petits indels (<50pb).

(SS), reste assez éloignée des polyploïdes. En effet, l'espace TEs n'est pas conservé entre les haplotypes du génome S et du génome B et comme nous l'avons vu dans la comparaison inter-génomique, l'espace gènes est relativement divergent (Figure II-3). Nous avons donc analysé les haplotypes de S et de B séparément comme deux génomes différents, et non en tant qu'haplotypes d'un même génome.

Comparaison des gènes des haplotypes S (GAA,SAA,SAB)

Les gènes *Pina* et *Pinb* du locus *Ha* sont trouvés dans les deux haplotypes du génome S. Nous avons trouvé un taux de SNPs important pour ces deux gènes (Tableau II-4A), en particulier pour le gène *Pinb*. Cependant, ces taux restent inférieurs à ceux trouvés pour l'autre gène du locus *Ha* (*Gsp-1*) lors de la comparaison entre les haplotypes du génome B et S (Tableau II-2) confirmant la divergence des haplotypes S après la séparation des deux génomes.

Comparaison de l'espace inter-génique des haplotypes S

Les deux régions inter-géniques des génomes S sont assez divergentes (21,4 SNPs/kb, 0,83 Ma de divergence), malgré l'absence de larges réarrangements (>50bp) ou d'insertions différentielles d'éléments transposables (Tableau II-4B).

Comparaison des gènes des haplotypes B (DurumB, RenanB, CSB)

Les gènes *BGGP* et *Gsp-1* ont été trouvés en commun dans les trois haplotypes du génome B des polyploïdes. Les comparaisons ont montré des taux de SNPs extrêmement bas et stables entre les gènes des trois haplotypes (Tableau II-4C), contrastant avec les taux élevés de SNPs détectés en comparant les gènes des haplotypes S et B. (Tableau II-2).

Comparaison de l'espace inter-génique des haplotypes B

Les séquences inter-géniques du tétraploïde et des hexaploïdes sont extrêmement proches. Les comparaisons entre les trois séquences ont donné sensiblement les mêmes résultats, avec des taux de SNPs très faibles, compris entre 0,2 et 0,5 SNP/kb et des divergences comprises entre 10.000 et 19.000 ans correspondant à moins de 40 SNPs sur plus de 75kb de séquence (Tableau II-4D).

De plus, nous avons trouvé très peu d'indels dans ces différentes comparaisons et aucun réarrangement important. Il y a respectivement 14 et 15 indels correspondant à 27 et 23 bp de différences entre les séquences de DurumB/RenanB et DurumB/CSB, confirmant les

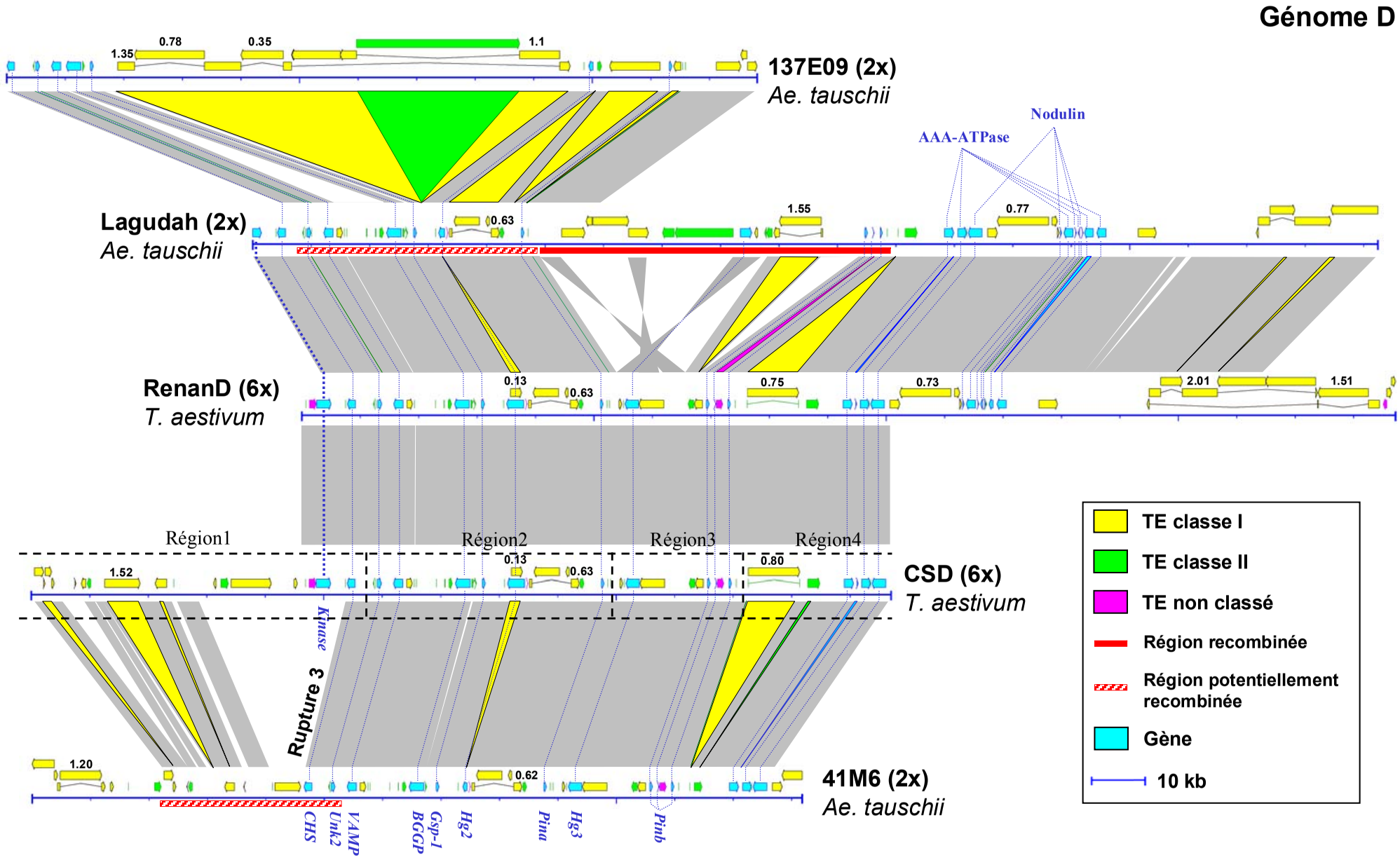


Figure II-4. Comparaison des régions *Ha* des différentes haplotypes du génome D. Les portions grisées représentent des séquences similaires. Les gènes orthologues sont reliés par des traits pointillés bleu, plus épais pour les gènes d'un intérêt particulier. Les insertions ou les délétions de grande taille (>500 pb) sont indiquées par les triangles de la couleur correspondant au type de la séquence impliquée. Les rectangles rouges (hachurés rouge) représentent les régions recombinées (potentiellement recombinées). Le découpage en 4 régions est représenté directement sur la séquence de CSD.

très faibles taux de SNPs et temps de divergence (Tableau II-4D). Les deux séquences des haplotypes hexaploïdes sont extrêmement proches dans le génome B (13 indels, 28 bp).

Cette proximité contraste largement avec les différences observées dans la comparaison du génome A des hexaploïdes.

III.2.3 Variabilité haplotypique du génome D

Comparaisons des gènes

Les gènes *BGGP*, *Gsp-1*, *VAMP* et *CHS* sont dans les séquences de tous les haplotypes du génome D. Nous avons trouvé un taux de polymorphisme très faible entre les gènes des différentes haplotypes, exceptés les gènes de l'haplotype diploïde 137E09 qui montrent des taux de polymorphismes légèrement supérieurs comparés aux gènes des autres haplotypes (Tableau II-5A).

Comparaison de l'espace inter-génique (137E09,Lagudah, 41M6, RenanD, CSD)

Les séquences des deux haplotypes hexaploïdes RenanD et CSD sont extrêmement proches (Figure II-4). Très similairement aux résultats trouvés pour le génome B des hexaploïdes, le taux de SNPs est très faible et stable sur toute la séquence (0,8 SNP/kb et 0,03 Ma de divergence) (Tableau II-5B). Nous avons également calculé ces taux pour les autres comparaisons entre les haplotypes du génome D, mais nous avons trouvé d'importantes variations dans les taux de SNPs le long des séquences pour toutes les autres comparaisons (Figure II-5). De la même façon que pour le génome A, nous avons séparé les séquences en 4 régions pour lesquelles les comparaisons donnent des taux de SNPs relativement stables (Figure II-5).

La région 1 comprend les séquences situées en 5' du gène *VAMP*. Cette région est principalement couverte par les haplotypes CSD et 41M6 : ils ont près de 18kb (17.953 pb) en commun contre moins de 3 kb pour les autres comparaisons de cette région. On obtient un taux de SNPs de 12,3 SNPs/kb pour une divergence de 0,48 Ma (Tableau II-5C). On observe de plus une petite rupture de colinéarité en 3' du gène *CHS* (une séquence de 7.124 pb de 41M6 complètement différente des 25.933 pb dans CSD) dans la comparaison (Figure II-4, Rupture 3). Les autres comparaisons pour cette région donnent des taux de SNPs compris entre 7,5 et 13,5 SNPs/kb et des divergences entre 0,29 et 0,52 Ma (Tableau II-5C).

La région 2 se situe entre le gène *VAMP* et le début de la région 'inversée' en 3' du gène *Pina*, présentée ci-après dans la description de la région 3 (Figure II-4). Cette région est

A

	BGGP (1278 pb)				*Gsp-1* (492 pb)				*VAMP* (648 pb)				*CHS* (1251 pb)			
	137E09	Lag	41M6	CSD	137E09	Lag	41M6	CSD	137E09	Lag	41M6	CSD	137E09	Lag	41M6	CSD
137E09 *(Ae. Tauschii)*																
Lagudah *(Ae. Tauschii)*	6				2				2				7			
41M6 *(Ae. Tauschii)*	6	0			2	0			3	3			6	8		
CSD *(T. aestivum)*	5	1	1		3	1	1		2	2	1		6	5	8	
RenanD *(T. aestivum)*	5	1	1	0	3	1	1	0	2	2	1	0	6	5	8	-

B

	Région commune	Région alignée	SNP / kb	Divergence (Ma)	Taux de remplacement						
					Total (%)	**TEs** (insertions)		**Réarrangements** (délétions)		**Petits indels**	
						Nb	%	Nb	%	Nb	%
RenanD / CSD	111066	28833	0,8	0,03	0,2%	0	0,0%	1	0,2%	12	0,0%
137E09 / Lagudah	120047	17564	variable	variable	90,4%	11	79,6%	11	10,6%	55	0,2%
137E09 / 41M6	119853	10592	variable	variable	89,5%	11	79,7%	8	9,7%	46	0,1%
Lagudah / 41M6	129195	42588	variable	variable	35,6%	2	6,4%	13	29,0%	58	0,2%
137E09 / CSD	121814	10594	variable	variable	90,6%	12	80,1%	14	10,3%	60	0,1%
Lagudah / CSD	240934	80428	variable	variable	24,7%	5	7,9%	20	15,6%	75	0,1%
41M6 / CSD	99578	53612	variable	variable	13,3%	3	10,8%	8	2,4%	47	0,1%

C

	Région 1			**Région 2**			**Région 3**			**Région 4**		
	Région commune	**SNP / kb**	**Divergence (Ma)**	**Région commune**	**SNP / kb**	**Divergence (Ma)**	**Région commune**	**SNP / kb**	**Divergence (Ma)**	**Région commune**	**SNP / kb**	**Divergence (Ma)**
137E09 / Lagudah	1335	13,5	0,52	6114	28,0	1,10	10115	2,2	0,08	-	-	-
137E09 / 41M6	1336	7,5	0,29	6113	28,0	1,10	3143	57,3	2,30	-	-	-
Lagudah / 41M6	1909	13,1	0,51	20918	1,0	0,04	7975	48,4	1,93	11786	5,3	0,20
137E09 / RenanD	1334	12,0	0,47	6116	28,0	1,10	3144	56,3	2,25	-	-	-
Lagudah / RenanD	2575	8,9	0,35	20748	4,1	0,16	7974	48,4	1,93	49131	14,6	0,53
41M6 / CSD	17953	12,3	0,48	20927	3,7	0,14	8004	1,5	0,06	6728	8,8	0,34

Tableau II-5. Comparaison de la séquence des gènes et de l'espace TEs du génome D. (A) Nombre de SNPs trouvés dans les comparaisons deux à deux des gènes dans les différents haplotypes du génome D. (B) Comparaison des séquences de l'espace TEs du génome D, avec le taux de SNPs, la divergence et les taux de remplacement. Ces derniers sont séparés en trois : importance de l'insertion des TEs, des réarrangements de grande taille (>50 pb) et les petits indels (<50pb). (C) Comparaison des séquences de l'espace TEs par région, dans le cas où les taux de SNPs fluctuent de façon importante le long des séquences.

représentée dans tous les haplotypes. La séquence de l'haplotype 137E09 du diploïde *Ae. tauschii* est la plus divergente comparée aux autres (Tableau II-5C). Elle présente le même taux de SNPs (28,0 SNPs/kb) et temps de divergence (1,10 Ma) dans ses différentes comparaisons avec les autres séquences de la région 2. Les séquences des haplotypes diploïdes Lagudah et 41M6 sont très proches pour cette région (1,0 SNP/kb, 0,04 Ma de divergence) mais 41M6 est légèrement plus proche des polyploïdes que Lagudah (3,7 SNPs/kb contre 4,1 SNPs/kb et 0,14 Ma contre 0,16 Ma divergence) (Tableau II-5C). Ces résultats sont conformes aux analyses de marqueurs RFLP qui désignaient 41M6 et 137E09 comme respectivement les haplotypes d'*Ae. tauschii* les plus proches et les plus éloignés des hexaploïdes (Akhunov *et al.* 2005).

La région 3 comprend toute la région 'inversée' ainsi que la suite de la séquence jusqu'au gène *Pinb*. Cette région 'inversée' est orientée dans un sens pour les séquences de 137E09 et de Lagudah et dans l'autre sens pour les séquences de RenanD, CSD et 41M6 (Figure II-4). Cette inversion a été accompagnée de délétions par recombinaisons illégitimes. La séquence de 137E07 ne couvre pas toute la région 'inversée'. Le reste de la séquence de cette région 3, après la région inversée, est dans la même orientation dans tous les haplotypes la couvrant (non couverte par 137E09). Les séquences présentant la même orientation pour l'inversion sont très proches entre elles avec des taux de SNPs de 1,5 et 2,2 SNPs/kb et une divergence de 0,06 et 0,08 Ma pour les comparaisons 137E09/Lagudah et Renan/CSD/41M6 respectivement (Tableau II-5C). Inversement, les comparaisons entre des séquences présentant des orientations différentes montrent une divergence élevée (de 48,4 à 57,3 SNPs/kb et de 1,93 à 2,30 Ma de divergence) (Tableau II-5C). Ces taux de SNPs et ces divergences ne concernent pas seulement la région de l'inversion mais bien toute la séquence de la région 3.

La région 4 couvre les séquences situées après le gène *Pinb*. La comparaison Lagudah/RenanD pour cette région porte sur près de 50 kb et donne un taux de SNP de 14,6 SNPs/kb et un temps de divergence de 0,53 Ma. Les comparaisons 41M6/Lagudah et 41M6/CSD donnent des taux sensiblement plus bas avec 5,3 et 8,8 SNPs/kb et 0,20 et 0,34 Ma de divergence respectivement.

Cette alternance de régions très similaires et très divergentes suggère la présence possible de recombinaisons génétiques ou de régions avec une évolution spécifique dans le génome (pression de sélection accentuée ou relâchée). Pour préciser ces variations de taux de

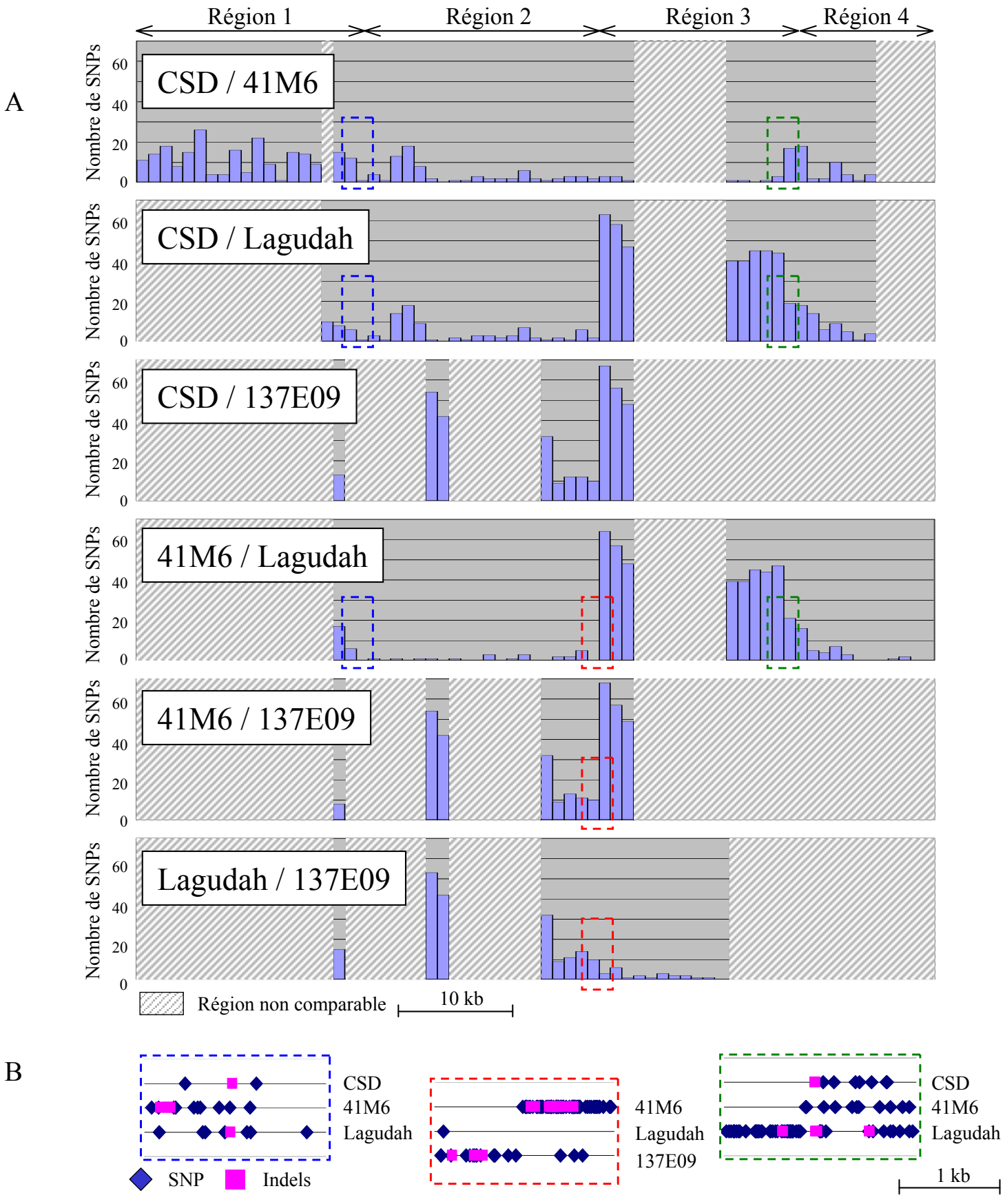


Figure II-5. Variation de la distribution des SNPs dans l'espace TEs de 5 haplotypes du génome D. (A) Les histogrammes montrent le nombre de SNPs dans des fenêtres de 1000 pb entre les séquences trouvées en commun (sans les indels). Ces comparaisons portent uniquement sur les séquence inter-géniques. Les parties hachurées représentent les régions non comparables. Les comparaisons sont divisées en 4 régions. (B) Détail des SNPs et des Indels pour les séquences situées à l'interface de régions.

SNPs dans et entre les différentes régions ainsi que leur origine, nous avons aussi analysé les régions de 2 kb à l'interface des ces régions (Figure II-5B).

La différence de taux de SNPs la plus flagrante entre des régions ressort de la comparaison entre les régions 2 et 3 (Figure II-5). La séquence de Lagudah est d'abord très proche de celle de 41M6/RenanD/CSD en région 2 et très divergente de celle de 137E09. Puis la tendance s'inverse en région 3 et la séquence de Lagudah est très proche de 137E09 et très divergente de celles de 41M6/RenanD/CSD. Ces résultats montrent que l'haplotype Lagudah est un haplotype mixte issu de recombinaison(s). Trois hypothèses sont possibles :

H1 : La séquence de la région 3 de Lagudah vient d'une recombinaison avec un haplotype proche de 137E09. La séquence de la région 2 de Lagudah ne vient pas d'une recombinaison et est très proche de celle de 41M6.

H2 : La séquence de la région 2 de Lagudah vient d'une recombinaison avec un haplotype proche de 41M6. La séquence de la région 3 de Lagudah ne vient pas d'une recombinaison et est très proche de celle de 137E07.

H3 : Les séquences des régions 2et 3 (H3a) ou 3et 4 (H3b) de Lagudah sont issues de recombinaisons.

Les importants niveaux de divergence observés pour la région 3 ne se retrouvent pas pour les régions 2 et 4. Les différences entre les divergences trouvées entre Lagudah/41M6 et Lagudah/CSD (identique à Lagudah/RenanD) pour les régions 2 et 4 ne sont pas flagrantes (0,04 Ma et 0,16 Ma contre 0,20 Ma et 0,53 Ma) et ne permettent pas de favoriser une hypothèse.

Pour la région 4, les trois haplotypes (CSD, Lagudah et 41M6) ne sont pas très proches deux à deux (divergences comprises entre 0,20 et 0,53 Ma). Cette région n'est donc probablement pas recombinante entre les différents haplotypes (excluant H3b). Si la région 4 de Lagudah, modérément divergente de celles de CSD et 41M6, correspond à une partie non recombinée, la région 3, très divergente entre Lagudah/41M6 et Lagudah/CSD, vient certainement d'une recombinaison génétique (on exclut donc l'hypothèse H2).

Il est difficile de trancher entre les 2 hypothèses restantes (H1 et H3a). Lagudah est globalement plus proche de CSD et 41M6 pour la région 2 que pour la région 4. Dans la région 2, la divergence entre Lagudah/41M6 est nettement plus faible que celle entre Lagudah/CSD (0,04 Ma contre 0,16 Ma). Cette très faible divergence entre Lagudah et 41M6 est surprenante et fait penser à une nouvelle recombinaison dans Lagudah (H3a). Ce scénario

fait donc intervenir deux recombinaisons successives dans l'haplotype Lagudah. Cependant, les éléments à notre disposition ne sont pas suffisants pour exclure complètement l'hypothèse H1, où la recombinaison n'a eu lieu que pour la région 3 avec des régions 2 et 4 évoluant à des vitesses sensiblement différentes.

L'analyse des comparaisons de la région 1 apporte une couche supplémentaire de complexité. On remarque ainsi une différence de taux de SNPs et divergence assez nette entre les régions 1 et 2 de la comparaison 41M6/CSD (12,3 SNPs/kb contre 3,7 SNPs/kb et 0,48 Ma contre 0,014 Ma de divergence). De plus, on observe, dans la région 1, une rupture de colinéarité et un nombre important de réarrangements comparé à la région 2. Ces résultats peuvent indiquer une autre recombinaison ou une région plus variable. Le manque de couverture des autres haplotypes ne permet pas de conclure, même si la rupture et les réarrangements nous font légèrement privilégier l'hypothèse d'une recombinaison génétique.

Nous avons calculé les taux de remplacement entre les différentes haplotypes, dont deux sont mixtes (Lagudah et probablement 41M6). Ces taux sont très variables (Tableau II-5B) et reflètent les importantes différences entre les haplotypes, qu'elles soient dues à l'insertion massive d'éléments transposables, à des recombinaisons génétiques et/ou illégitimes.

Les hauts taux de remplacement trouvés dans les comparaisons entre 137E09 et 41M6/RenanD/CSD (de 89,5% à 90,6%) sont principalement dus à une invasion de la séquence par des éléments transposables (dont 6 insertions de TEs les unes dans les autres) (Figure II-4, Tableau II-5B). Le taux de remplacement élevé dans les comparaisons entre Lagudah et 41M6/RenanD/CSD résulte principalement des réarrangements par recombinaison illégitime dans la région 3, apportée par une recombinaison génétique. De la même façon, le taux de remplacement important trouvé entre 41M6/CSD/RenanD trouve son origine dans les réarrangements observés dans la zone variable ou recombinée. Sans considérer cette zone, la différence majeure entre 41M6 et les hexaploïdes est l'insertion de deux éléments transposables (*Wis* et *Morgan*) uniquement dans ces derniers.

IV Discussion

Nous avons analysé pour la première fois plusieurs haplotypes représentants les génomes A, B, S et D du blé. L'étendue des insertions et des éliminations des éléments transposables et leurs conséquences sur l'organisation et l'évolution rapide de ces génomes, sur une courte échelle de temps, ont ainsi été évaluées au niveau du locus *Ha* par des comparaisons haplotypiques aux niveaux inter-génomique, intra-génomique et intra-spécifiques.

Les comparaisons inter-génomiques soulignent l'évolution rapide de l'espace TEs qui est complètement différent entre les principaux génomes A, B et D, comme observé dans des études précédentes (Chantret *et al.* 2005, Isidore *et al.* 2005, Dvorak *et al.* 2006, Gu *et al.* 2006). L'espace TEs est également complètement différent entre les génomes B des blés polyploïdes et le génome S de l'espèce diploïde *Ae. speltoides* confirmant encore une fois des résultats précédemment obtenus pour les locus *SPA* et *Acc1* (Chalupska *et al.* 2008, Salse *et al.* 2008b). Ces observations, combinées à la divergence des séquences des gènes *BGGP* et *Gsp-1* (Tableau II-2), confirment la divergence des génomes B et S très peu de temps après leur divergence des génomes A et D (Huang *et al.* 2002, Chalupska *et al.* 2008, Salse *et al.* 2008b). Les génomes B et S sont donc des génomes distincts et ne représentent pas des haplotypes différents d'un même génome. Nous avons retracé les dates de radiation de ces quatre principaux génomes (Figure II-6), en approximant la divergence des génomes A/D et S/B à 3 Ma (Huang *et al.* 2002, Chalupska *et al.* 2008).

La variabilité haplotypique dans les différentes comparaisons intra-génomiques et inter-specifiques dépend du génome considéré. En effet, une proportion variable de l'espace TEs semble être conservée entre les différents haplotypes d'un même génome. Nous avons constaté que la non-conservation de cet espace TEs est due aux insertions différentielles de TEs mais aussi aux recombinaisons illégitimes et génétiques (Figure II-1, II-3, II-4).

Variabilité haplotypique des génomes B et S

La conservation haplotypique la plus forte est observée pour le génome B. Pour ce génome, les comparaisons entre les haplotypes des polyploïdes (DurumB, RenanB et CSB)

montrent seulement quelques SNPs et Indels tout au long des séquences en commun, soit plus de 75 kb, sans aucune insertion ou élimination différentielle de TEs (Figure II-3, Tableau II-4). En se basant sur la comparaison de l'espace TEs commun, ces trois haplotypes du génome B ont divergé très récemment (de 10.000 à 20.000 ans). Autant ce niveau de divergence est attendu entre les hexaploïdes, autant il est quelque peu surprenant entre le tétraploïde et les hexaploïdes, surtout compte tenu de la divergence observée dans le génome A. Une importante conservation haplotypique du génome B a été également observée au niveau du locus *HMW-Glu* (Gu *et al.* 2006).

Les deux haplotypes du génome S étudiés sont assez divergents (0,83 Ma) mais l'espace TEs en commun est complètement dénué de réarrangements importants (>1000bp) (Figure II-3) ce qui pourrait indiquer une très faible activité des forces d'insertion et d'élimination des TEs dans ce génome.

Variabilité haplotypique du génome D

Les deux haplotypes hexaploïdes du génome D (RenanD et CSD) montrent également une très forte conservation. Leur séquence est quasiment identique à l'exception de quelques SNPs et indels. Leur divergence est estimée à 30.000 ans. (Figure II-4, Tableau II-5), confirmant l'origine commune et unique du génome D pour ces deux haplotypes du blé hexaploïde. Les haplotypes diploïdes du génome D montrent d'importantes différences lorsqu'on les compare entre eux et avec les hexaploïdes. L'haplotype 137E09 est le plus divergent (1,10 Ma) et montre une séquence envahie par des éléments transposables, en grande partie imbriqués les uns dans les autres. Cependant, en supposant que la suite de la séquence de 137E09 ressemble à celle trouvée dans Lagudah, on observerait une contribution relativement équivalente de la force insertionnelle et éliminatrice (par recombinaison illégitime) dans l'évolution dynamique de l'espace TEs. Une ou plusieurs recombinaisons génétiques ont été observées dans les deux autres haplotypes diploïdes (Lagudah et 41M6). En considérant uniquement les portions non recombinées, 41M6 est l'haplotype le plus proche des hexaploïdes (0,14 Ma de divergence) suivi de Lagudah (0,53 Ma de divergence). La principale différence entre ces diploïdes et les hexaploïdes dans les portions non recombinées vient de l'insertion du TE *Wis* dans ces derniers. On notera aussi l'insertion de l'élément *Morgan* dans les hexaploïdes qui n'est pas présent dans 41M6 ni dans Lagudah. La majeure partie des différences dans ces régions vient des nombreuses délétions par recombinaisons illégitimes associées ou non à des recombinaisons génétiques. Les divergences calculées pour

ces trois haplotypes diploïdes concordent avec les résultats obtenus par RFLP qui désignaient 137E09 et 41M6 comme respectivement l'haplotype le plus divergent et le moins divergent des polyploïdes. Néanmoins, la date de divergence trouvée pour la partie non recombinée de 41M6 reste élevée pour l'haplotype le plus proche (Tableau II-5). Une analyse PCR avec des marqueurs spécifiques des insertions différentielles (*Wis* et *Morgan*) sur une collection d'*Ae. tauschii* a permis d'identifier un haplotype ayant l'insertion *Wis* et donc probablement plus proche des polyploïdes que 41M6, pour cette région. Ces comparaisons multiples montrent qu'à l'exception de l'insertion de l'élément *Morgan*, la majorité des réarrangements génomiques et des insertions différentielles de TEs, décrites précédemment entre les haplotypes du génome D diploïdes et hexaploïdes (Chantret *et al.* 2005) semblent avoir eu lieu antérieurement à l'allohexaploïdie, lors de la diversification de l'espèce diploïde *Ae. tauschii*. Ces résultats soulignent l'importance et la complémentarité des comparaisons intra-spécifiques aux comparaisons intra-génomiques dans les études de l'évolution, par génomique comparative, pour prendre en compte la diversité dans les espèces étudiées.

Variabilité haplotypique du génome A

L'insertion différentielle des TEs semble représenter la proportion la plus importante de l'ADN non conservé entre les haplotypes diploïdes du génome A. Par contre, les comparaisons diploïdes/polyploïdes et polyploïdes/polyploïdes montrent une proportion importante des réarrangements par recombinaison illégitimes (Figure II-1, Tableau II-3). En particulier, nous avons observé de nombreux réarrangements de plusieurs kb voire dizaines de kb, en plus de la délétion du locus *Ha* déjà caractérisée (Chantret *et al.* 2005). Tous ces larges réarrangements ont été validés par PCR au niveau des clones BAC mais aussi des génotypes utilisés pour confirmer un séquençage correct de ces clones. Ces réarrangements correspondent probablement tous à des délétions par recombinaison illégitime, même si deux des ces événements, correspondant aux ruptures de colinéarités entre 8K2/DurumA et DurumA/Polyploïdes, restent à préciser. En effet, le manque de couverture et de séquences de références pour ces événements ne nous permet pas de statuer formellement sur leur nature, même si le scénario d'une recombinaison génétique associée à des recombinaisons illégitimes est privilégié.

Les calculs de divergence ont montré que l'haplotype de *T. monococcum* est près de deux fois plus divergent des polyploïdes que les haplotypes de *T. urartu* (1,19 Ma conte 0,53 et 0,67 Ma divergence). Ces estimations sont concordantes avec la date de l'allotétraploïdisation (0,5 Ma) et de la divergence des génomes A de *T. monococcum*, *T.*

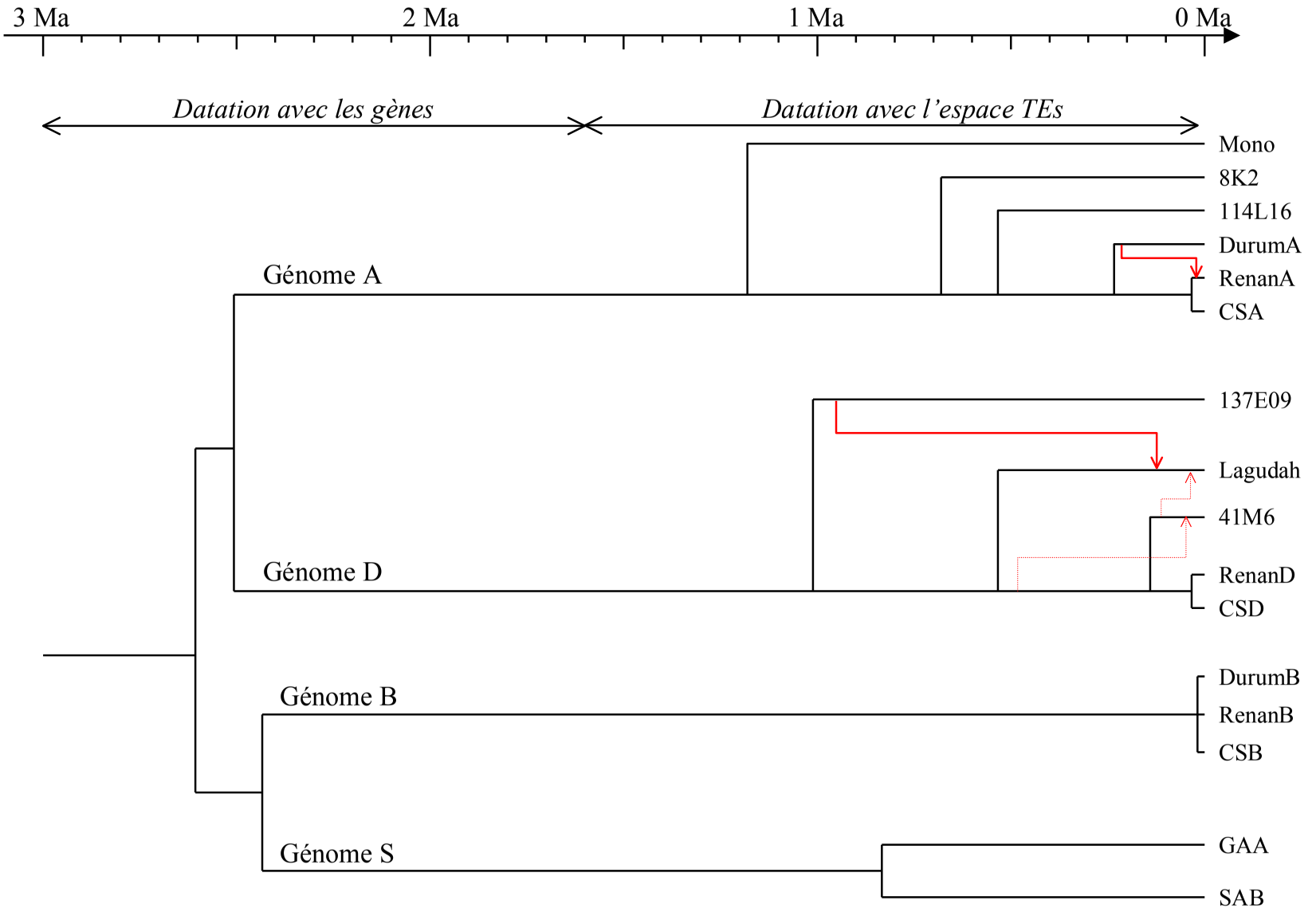


Figure II-6. Arbre phylogénétique reprenant les principaux résultats des comparaisons de 16 haplotypes des génomes A, B, S et D. La divergence des différents génomes a été estimée sur la base de la divergence observée pour le gène *BGGP*, en calibrant à 2,5 Ma la divergence entre les 4 génomes. La divergence des haplotypes a été estimée sur la base des comparaisons de l'espace TEs. Les flèches rouges représentent les recombinaisons génétiques observées entre hapotypes, confirmées (flèche pleine) ou probables (flèche pointillée). Dans le cas des haplotypes mixtes, la divergence présentée vient d'une région non-recombinante.

urartu et *T. turgidum* évaluée précédemment (Huang *et al.* 2002). Par contre, les larges différences observées entre les deux hexaploïdes sont étonnantes. En particulier, les deux larges délétions par recombinaison illégitime dans CSA ne sont pas présentes dans RenanA. Une analyse précise des comparaisons entre les trois polyploïdes nous a permis de mettre en évidence une recombinaison génétique, près du gène *Unknown-2a* (Figure 1A), ayant introgressée une partie de la séquence de DurumA dans RenanA.

Au final, nous avons utilisé les divergences des séquences inter-géniques pour calculer la divergence entre les différents haplotypes. Dans le cas d'haplotypes mixtes, nous avons utilisé les régions de l'haplotype non réarrangées (Figure II-6).

Conclusion

Nous avons donc observé, à travers l'étude d'haplotypes des génomes A, B, S et D une variabilité extensive au niveau du locus *Ha*, due à la fois à des divergences importantes des séquences, des insertions différentielles d'éléments transposables, des recombinaisons génétiques entre différents haplotypes et des réarrangements (notamment délétions) par recombinaisons illégitimes.

Nous n'avons pas trouvé d'effet génome ou d'effet polyploïdie en relation avec l'activité insertionnelle des éléments transposables. En effet, on trouve une ou quelques insertions différentielles de TEs dans tous les haplotypes, sans apparente distinction, à l'exception de l'haplotype 137E09 du génome D qui est particulièrement riche en TEs. Ceci confirme nos travaux sur des séquences représentatives des génomes A et B, qui montraient que la majorité des insertions d'éléments transposables avait eu lieu avant les évènements de polyploïdisation (Charles *et al.* 2008).

Les comparaisons des différents haplotypes diploïdes des génomes D, appartenant à la même espèce (*Ae. tauschii*), a permis de mettre en évidence une importante variabilité intra-spécifique essentiellement due à de multiples recombinaisons génétiques. L'étude du locus *Lr34* (Wicker *et al.* 2009a) avait précédemment mis en évidence des recombinaisons génétique dans *Ae. tauschii*. C'est donc probablement un phénomène global dans cette espèce. Nous avons également identifié une recombinaison génétique entre des espèces différentes pour le génome A. La barrière d'espèces induite par la polyploïdie est faible au niveau tétraploïde/hexaploïde dans le génome A, confirmant des résultats de transfert de gènes entre tétraploïde et hexaploïde observés dans ce génome au locus *Xpsr920* (Dvorak *et al.* 2006).

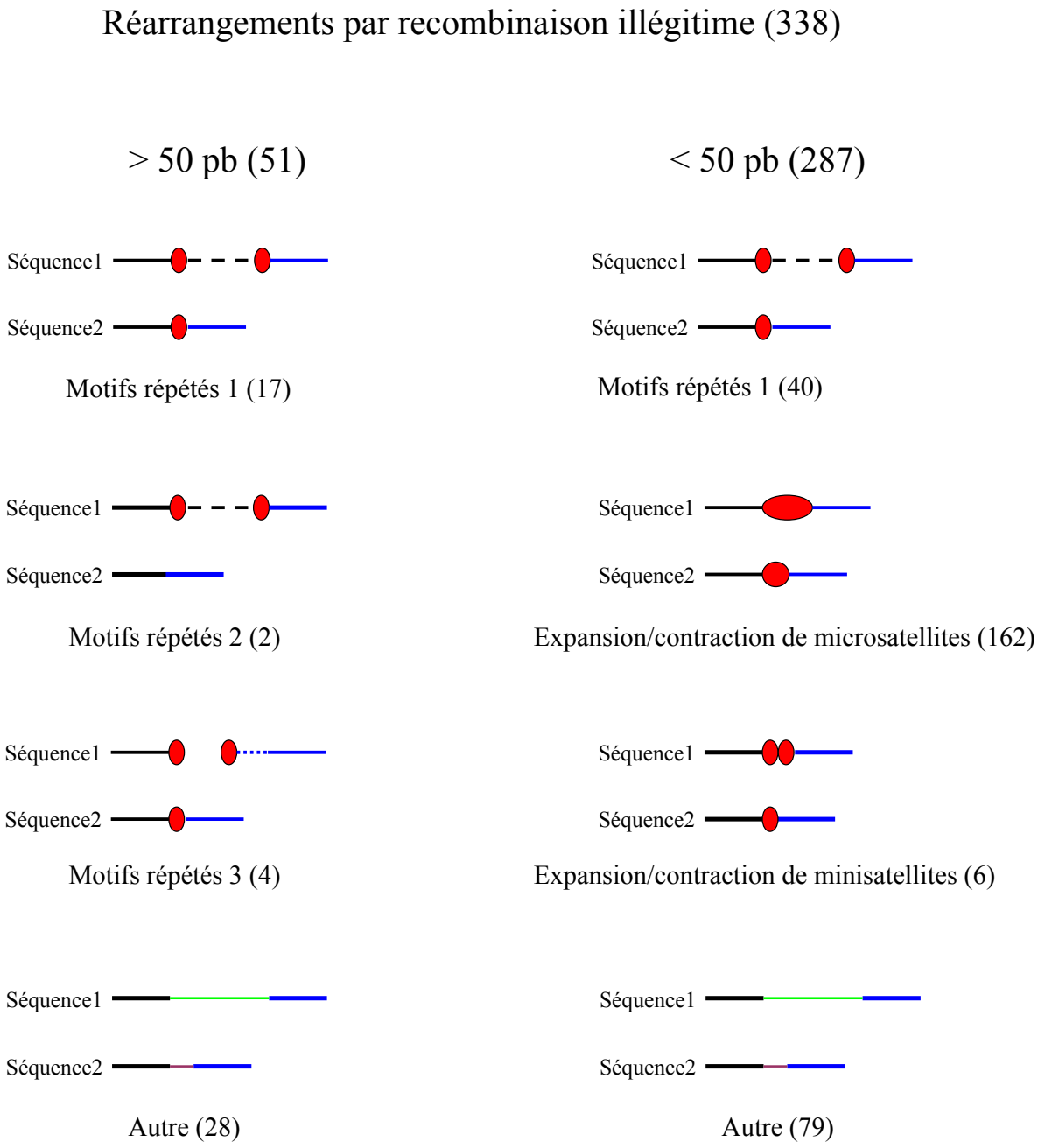


Figure II-7. Bases moléculaires des différentes recombinaisons illégitimes observées. Les 338 recombinaisons illégitimes (non-homologues) polarisées (insertion ou délétion) identifiées dans les séquences des génomes A, B, et D se répartisent en 51 grandes recombinaisons (>50 pb) et 287 petites recombinaisons (<50pb).
Les recombinaisons de grande taille sont très majoritairement des délétions (49/51). Des petits motifs répétés (cercle rouge, 3-20 pb) ont été détectés pour la moitié d'entre-elles (23/51).
Les recombinaisons de petite taille sont plus equilibrées entre les insertions (120/287) et les délétions (167/287). Les expansions/contractions de micro/mini-satellites sont les mécanismes les plus fréquents (168/287).

Deux autres recombinaisons potentielles ont été détectées dans le génome A sur le tétraploïde et un diploïde (8K2) mais le manque de séquences de référence et couvrantes ne nous a pas permis de les confirmer. Les haplotypes mixtes trouvés dans le génome D concernent uniquement les diploïdes. Notre étude n'a donc pas montré une perméabilité pour le génome D entre les niveaux hexaploïde et diploïde.

Mon étude a révélé et confirmé la généralisation des réarrangements génomiques, parfois sur plusieurs dizaines de kb par recombinaisons illégitimes. L'étude comparative présentée ici révèle que les réarrangements induisant des éliminations ont eu lieu, pour la plupart, dans des séquences d'éléments transposables (Figure II-1, II-4). Comme nous l'avons constaté dans l'étude des haplotypes diploïdes du génome D, les recombinaisons génétiques peuvent amplifier l'importance des recombinaisons illégitimes en apportant des fragments d'haplotypes plus divergents, et donc des délétions par recombinaisons supplémentaires. Un dernier type de recombinaison, les recombinaisons homologues inégales, semblent rares comme prédit dans mon étude précédente (Article 1) et ont une importance mineure dans les réarrangements. Cette comparaison haplotypique sur différents niveaux a montré que les recombinaisons illégitimes d'ADN représentent le principal mécanisme à l'origine de la dynamique de l'espace TEs.

La variabilité haplotypique importante trouvée dans les génomes du blé ressemble ou dépasse celle décrite pour le maïs (Tikhonov *et al.* 1999, Fu et Dooner 2002, Song et Messing 2003, Brunner *et al.* 2005, Lai *et al.* 2005, Morgante *et al.* 2005, Wang et Dooner 2006, Xu et Messing 2006). Contrairement au blé, la prolifération différentielle des éléments transposables et surtout le mouvement des gènes par les TEs de type *Helitron*, très actifs dans le maïs sont les principaux responsables de la variabilité importante du maïs (Lai *et al.* 2005, Morgante *et al.* 2005, Wang et Dooner 2006, Xu et Messing 2006). Nous n'avons trouvé aucune évidence de réarrangements médiés par des helitrons dans les séquences du blé. La bonne conservation de l'espace inter-génique entre les séquences des haplotypes d'un même génome nous a permis d'examiner les bordures des régions éliminées pour essayer de leur associer un mécanisme (Figure II-7). Les petites recombinaisons, largement majoritaires en nombre, sont principalement des expansions/contractions de mini/micro-satellites. Les grandes recombinaisons sont quasiment toutes des délétions (49/51). Dans la moitié d'entre-elles, et en particulier pour les grandes délétions du génome A, on distingue des motifs répétés probablement impliqués dans la délétion (Figure II-7).

La faible variabilité haplotypique observée dans les haplotypes polyploïdes du génome B contraste avec les importantes différences trouvées entre les haplotypes polyploïdes du génome A. Pourtant, ces deux génomes cohabitent dans ces polyploïdes depuis la tétraploidisation. Il serait intéressant de confirmer ou d'infirmer cette surprenante différence de comportement évolutif entre deux génomes co-résidents dans le même noyau.

Partie III : Evolution du caractère ‘grain tendre’ dans les *Poaceae* au cours des 60 derniers Ma : Emergence des gènes *Ha* dans l’ancêtre commun des *Erhrartoideae* et des *Pooideae*, après leur divergence avec les *Panicoideae*

I Introduction

Dans les analyses que j'ai présentées en Parties I et II, j'ai pu apprécier les effets des forces majeures de l'évolution des TEs (insertions et éliminations) sur la dynamique des génomes du blé à court terme (< 4 Ma). L'analyse des dates d'insertion des rétrotransposons d'un échantillon de séquences représentatif des génomes A et B a montré que 87% des insertions de ces éléments ont eu lieu avant les événements récents de polyploïdisation (0,5 Ma et 0,01 Ma). Elle a aussi montré que l'activité insertionnelle était différente entre ces deux génomes, en particulier au niveau des périodes et des intensités de prolifération des principales super-familles et familles de TEs (Résultats, Partie I). Sur une échelle d'évolution encore plus courte (0-1,2 Ma), la comparaison de régions orthologues entre des haplotypes des même génomes du blé a mis en évidence des différences de séquences impliquant des dizaines voire des centaines de kb. Les recombinaisons illégitimes, principales responsables de ces larges différences, contribuent à cette dynamique des génomes du blé (Résultats, Partie II). Elles sont également associées avec des recombinaisons génétiques qui amplifient les différences de séquences.

Cette dynamique importante de l'espace TEs a permis la divergence rapide des différents génomes du blé et a donc joué un effet stabilisateur, en défavorisant la recombinaison homéologue, dans les blés polyploïdes qui les ont réunis.

La divergence rapide de l'espace TEs nous permet d'apprécier leurs effets sur l'organisation et la dynamique des génomes du blé uniquement sur une courte échelle de temps (< 3 Ma). Au-delà de cette échelle, l'espace TEs n'est plus comparable (entièrement différent). Cependant, les comparaisons de l'espace gènes entre le blé et les différentes espèces des *Poaceae* permet d'apprécier la dynamique des génomes sur une échelle d'évolution plus longue (50-60 Ma).

Le locus *Ha*, connu pour son importante dynamique dans le blé (Chantret *et al.* 2005, 2008), représente un excellent modèle pour des études comparatives entre les différentes espèces de *Poaceae*. Ce locus porte les gènes *Gsp-1*, *Pina* et *Pinb* codant pour des protéines du grain ['Grain Softness Protein' (GSP), 'Puroindoline A' (PinA) et 'Puroindoline B' (PinB)], dont la présence dans la graine confère le caractère 'grain tendre' alors que leur absence (ou mutation) confère le caractère 'grain dur'. Une précédente analyse par génomique

comparée du locus *Ha* a expliqué l'absence du locus dans le blé dur tétraploïde par sa délétion indépendante dans les génomes A et B. Des recombinaisons illégitimes impliquant des TEs sont à l'origine de ces délétions (Chantret *et al.* 2005). Le blé tendre a retrouvé son caractère grain tendre en incorporant le génome d'*Ae. tauschii* (DD). Les analyses présentées en Partie II ont bien confirmé que ces événements de recombinaisons illégitimes sont fréquents dans cette région et sont responsables de la délétion récurrente des gènes *Pina* et *Pinb* dans différents génomes des blés polyploïdes (Chantret *et al.* 2005, Li *et al.* 2008).

Des orthologues des gènes *Ha* ont été trouvés dans l'orge (*Gsp-1*, *HindA*, *HindB1* et *HindB2*) (Caldwell *et al.* 2004). De plus, des homologues aux gènes *Ha* ont été détectées dans l'avoine (Avenoindoline) et le seigle (Secaloindoline), appartenant à la même tribu que le blé (*Triticeae*), mais aucun homologue aux gènes *Ha* n'a été détecté dans les *Panicoideae* (maïs et sorgho) et les *Ehrhartoideae* (riz), qui ont des grains durs (Fabijanski *et al.* 1988; Gautier *et al.* 2000; Darlington *et al.* 2001; Morris 2002). Cependant, des comparaisons des séquences ont permis l'identification, dans la région orthologue du génome du riz, d'un petit segment de 105pb, appelé *Ha-relic,* qui montre 67% de similarité en acides aminés avec le gène *Gsp-1* (Chantret *et al.* 2004). La situation n'était pas claire chez les *Panicoideae* (sorgho et maïs) qui ont divergé plus tôt des *Pooideae* (blé, orge) et des *Ehrhartoideae* (riz) et l'émergence du locus *Ha* et/ou sa disparition récurrente n'étaient donc pas bien précisées au niveau des *Poaceae*.

Pour répondre à ces questions et obtenir plus de précisions sur l'histoire évolutive du locus *Ha* dans différentes espèces, j'ai donc augmenté le champ de la comparaison génomique en réalisant du séquençage génomique complémentaire dans une espèce intermédiaire de *Poaceae* (*Brachypodium sylvaticum*) appartenant à la même sous-famille que le blé (les *Pooideae*, Figure Intro-2), mais aussi en analysant les régions orthologues d'autres espèces qui étaient en cours de séquençage (*Brachypodium distachyon* et sorgho). J'ai essayé d'élucider la difficile question de l'émergence de ce locus au cours de l'évolution des *Poaceae* en étudiant l'ensemble de ces séquences ainsi que l'évolution des gènes de la famille des prolamines qui codent pour des protéines de réserve proches des protéines Ha.

Les travaux présentés dans ce chapitre ont été publiés sous la forme d'article dans la revue Molecular Biology and Evolution (2009). Je présente aussi sous forme de résultats complémentaires les 'Supplemental data' de l'article, disponibles en ligne, ainsi que l'annotation des gènes *Ha-like* et des *prolamines* dans le génome complet de *B. distachyon*. Cette analyse a été réalisée sur une version plus avancée de la séquence de ce génome

(couverture 8x, http://mips.helmholtz-muenchen.de/plant/brachypodium/index.jsp) que celle utilisée dans l'Article 2 (couverture 4x). J'ai eu accès à cette version grâce à la contribution de mon équipe au consortium international de séquençage de *B. distachyon*. Ce séquençage et l'analyse du génome de *B. distachyon* fait également l'objet d'une publication à laquelle j'ai modestement contribué (The International Brachypodium Initiative 2010) (Annexe 5).

II Article 2

Cet article a été présenté dans la revue Molecular Biology and Evolution

Sixty Million Years in Evolution of Soft Grain Trait in Grasses: Emergence of the Softness Locus in the Common Ancestor of *Pooideae* and *Ehrhartoideae*, after their Divergence from *Panicoideae*

Mathieu Charles, Haibao Tang, Harry Belcram, Andrew Paterson, Piotr Gornicki, and Boulos Chalhoub.

Molecular Biology and Evolution **26**(7):1651–1661, juillet 2009

Sixty Million Years in Evolution of Soft Grain Trait in Grasses: Emergence of the Softness Locus in the Common Ancestor of *Pooideae* and *Ehrhartoideae*, after their Divergence from *Panicoideae*

Mathieu Charles, Haibao Tang,† Harry Belcram,* Andrew Paterson,† Piotr Gornicki,‡ and Boulos Chalhoub**

*Unité de Recherches en Génomique Végétale (UMR INRA 1165–CNRS 8114UEVE), Organization and evolution of Plant Genomes, Evry, France; †Plant Genome Mapping Laboratory, University of Georgia; and ‡Department of Molecular Genetics and Cell Biology, University of Chicago

Together maize, Sorghum, rice, and wheat grass (*Poaceae*) species are the most important cereal crops in the world and exhibit different "grain endosperm texture." This trait has been studied extensively in wheat because of its pivotal role in determining quality of products obtained from wheat grain. Grain softness protein-1 and Puroindolines A and B (grain storage proteins), encoded by *Ha-like* genes: *Gsp-1*, *Pina*, and *Pinb*, of the *Hardness* (*Ha*) locus, are the main determinants of the grain softness/hardness trait in wheat. The origin and evolution of grain endosperm texture in grasses was addressed by comparing genomic sequences of the *Ha* orthologous region of wheat, *Brachypodium*, rice, and Sorghum. Results show that the *Ha-like* genes are present in wheat and *Brachypodium* but are absent from *Sorghum bicolor*. A truncated remnant of an *Ha-like* gene is present in rice. Synteny analysis of the genomes of these grass species shows that only one of the paralogous *Ha* regions, created 70 My by whole-genome duplication, contained *Ha-like* genes. The comparative genome analysis and evolutionary comparison with genes encoding grain reserve proteins of grasses suggest that an ancestral *Ha-like* gene emerged, as a new member of the prolamin gene family, in a common ancestor of the *Pooideae* (*Triticeae* and *Brachypoidieae* tribes) and *Ehrhartoideae* (rice), between 60 and 50 My, after their divergence from *Panicoideae* (Sorghum). It was subsequently lost in *Ehrhartoideae*. Recurring duplications, deletions, and/or truncations occurred independently and appear to characterize *Ha-like* gene evolution in the grass species. The *Ha-like* genes gained a new function in *Triticeae*, such as wheat, underlying the soft grain phenotype. Loss of these genes in some wheat species leads, in turn, to hard endosperm seeds.

Introduction

GRASSES (*Poaceae*), with 10,000 species growing under diverse climates and latitudes, exceed all other plant families in ecological dominance and economic importance. Analysis of fossil records and phylogenetic data established that the grass subfamilies diverged from a common ancestor 50–80 My (for review, see Kellogg 2001; Gaut 2002; Prasad et al. 2005; Chalupska et al. 2008). Divergence time of several important grass lineages (*Triticum*, *Hordeum*, *Brachypodium*, *Oryza*, *Sorghum*, and *Zea*) has been recently reexamined based on sequence comparison of *Acc* and other genes, using 60 My for the divergence time of the *Panicoideae* (*Sorghum*, *Zea*) and *Ehrhartoideae* (rice) to calibrate the molecular clock (Chalupska et al. 2008). This and several earlier studies (Paterson et al. 2004; Bossolini et al. 2007; Faris et al. 2008) concluded that *Pooideae* (wheat, barley, and *Brachypodium*) and *Ehrhartoideae* (rice) diverged from each other 50 My, early after their divergence from *Panicoideae* (maize, Sorghum). Among the *Pooideae*, *Brachypoidieae* (*Brachypodium*), and *Triticeae* (wheat, barley), tribes diverged about 35 My. *Brachypodium*, with its small diploid genome, has become a model *Pooideae* grass with a potential to aid analysis of the large genomes of the *Triticeae* (Draper et al. 2001; Foote et al. 2004; Faris et al. 2008). A 4× draft version of the *Brachypodium distachyon* genomic sequence is already publicly available (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008), and the *Brachypodium* consortium is assembling an 8× genome sequence coverage (Vogel J, personal communication).

The *Hardness* (*Ha*) locus in wheat spans *Pina*, *Pinb*, and *Gsp-1* genes (called in this study *Ha-like* genes) and encodes Puroindolines A and B (PinA and PinB) and grain softness protein-1 (GSP-1) that determine the wheat grain hardness/softness or endosperm texture (for review, see Morris 2002). Because of the pivotal role of grain texture in determining quality of products obtained from wheat grain, this trait has been studied by geneticists (Law et al. 1978), chemists (Schofield 1986; Blochet et al. 1991, 1993), and molecular biologists (Gautier et al. 1994, 2000; Chantret et al. 2004, 2005, 2008; Li et al. 2008). At the genome organization level, the *Ha* locus is about 65 kb in the D genome of hexaploid wheat *Triticum aestivum* and contains three functional *Ha-like* genes: *Gsp-1*, *Pina*, and *Pinb*, as well as a *PseudoPinb*, a *Pinb-relic*, two other predicted genes (*Gene3* and *Gene5*), and several transposable elements (Chantret et al. 2005; fig. 1*A*). Upstream of *Gsp-1* gene, the *BGGP* (*Gene1*), encoding β-1-3-galactosyl-O-glycosyl-glycoprotein, delimits the 5′ boundary of the *Ha* locus. A *Nodulin* gene (*Gene8*) and a cluster of *ATPase* genes (*Genes7-1, 7-2, 7-3, 7-2′, and 7-3′*), located 20 kb downstream of *PseudoPinb*, delimit the 3′ boundary of the *Ha* locus (fig. 1*A*; Chantret et al. 2005, 2008). *Pina* and *Pinb* genes were also found in *Triticeae* species, in which soft endosperm is a dominant trait: in diploid and hexaploid wheat (*Triticum* and *Aegilops* species), barley (*Hordeum vulgare*), rye (*Secale cereale*), and oats (*Avena sativa*). Surprisingly, *Pina* and *Pinb* genes are absent from the A and B genomes of the tetraploid (*Triticum turgidum*) and hexaploid (*T. aestivum*) wheat species, although present in their progenitor species (Gautier et al. 2000). Comparative genomic analysis showed that *Pina*

Key words: *Poaceae*, evolution, comparative genomics, grain endosperm softness.

E-mail: chalhoub@evry.inra.fr.

Mol. Biol. Evol. 26(7):1651–1661. 2009
doi:10.1093/molbev/msp076
Advance Access publication April 24, 2009

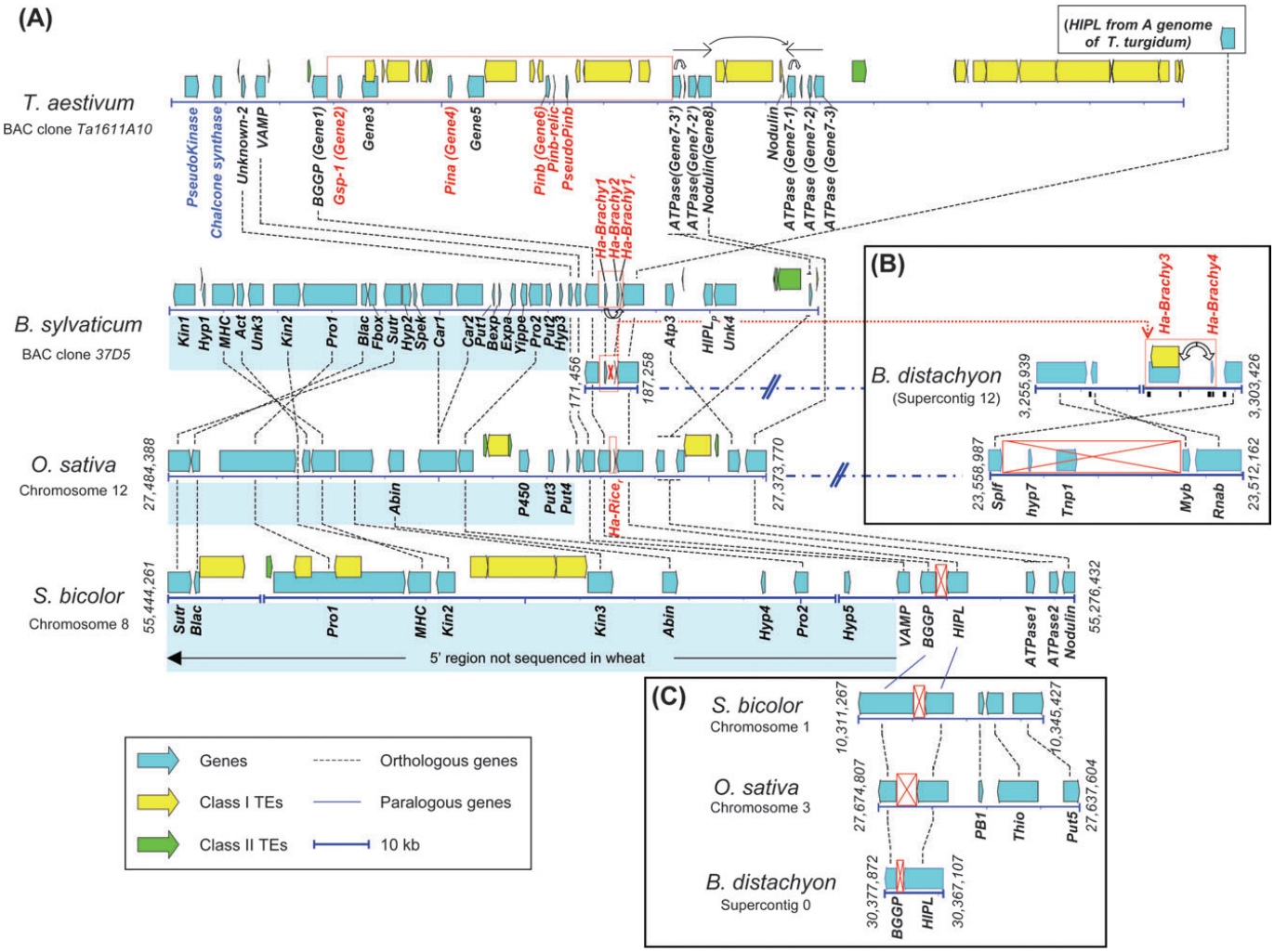


FIG. 1.—Comparison of orthologous and paralogous genomic regions including the *Ha* locus of wheat, *Brachypodium sylvaticum*, *Brachypodium distachyon*, rice, and *Sorghum bicolor*. (*A*) Comparison of orthologous regions between the five species. An overview of the 187,340 bp sequence (BAC clone Ta1611A10) of the D genome of hexaploid wheat *Triticum aestivum* (from Chantret et al. 2008). Relative position of *HIPL* gene as found in the sequence of the A genome of *Triticum turgidum* species (Chantret et al. 2008) is also shown. *Brachypodium sylvaticum* BAC clone (BAC37D5) of 120,033 bp was sequenced in this study. *Oryza sativa* and *S. bicolor* orthologous region sequences and annotations were, respectively, retrieved from the Michigan State University site (http://rice.plantbiology.msu.edu, release 6 January 2009) and from the Joint Genome Institute Web site (http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html, release March 2008; Paterson et al. 2009). The *B. distachyon* 4X genome sequence is from http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/ (release October 2008). Sequence gaps at *Ha-like* genes were supplied by the international *Brachypodium* initiative (Vogel J, USDA, Albany, USA). The wheat genes were named in the same way as in previous studies (Chantret et al. 2005, 2008). *Ha-like* genes and related sequences are shown in red. *Ha-Brachy2* gene is deleted in *B. distachyon* (marked by a red cross). Genes present in wheat but not in other species are shown in blue. Genes conserved in multiple species are connected by dashed lines. Light blue boxes represent region, 5′ to the *Ha* locus, compared between *B. sylvaticum*, rice, and Sorghum but not sequenced in wheat. (*B*) Additional duplications of *Ha-like* genes (*Ha-Brachy3* and *Ha-Brachy4*) observed in *B. distachyon* at 3 Mb of the *Ha* locus (red dashed arrow). Flanked genes are also shown. Corresponding rice orthologous region is also presented and shows no *Ha-like* genes (red cross). Double arrow on *Ha-Brachy3* and *Ha-Brachy4* indicates that they are derived from recent tandem duplication (from each other's). Black bars, below the *B. distachyon* presented region, indicate location of sequences successfully used to derive PCR markers and confirm the presence of *Ha-Brachy3*, *Ha-Brachy4* as well as flanked *Myb* and *Splf* genes on same BAC clones of *B. sylvaticum*. Thickness of these bars is proportional to the length of the sequence used. (*C*) Synteny and collinearity of paralogous *Ha* regions (derived from last-shared ancestral whole-genome duplication) of rice, Sorghum, and *B. distachyon*. The predicted location of the *Ha-like* genes is between *BGGP* and *HIPL* genes. Absence of any *Ha-like* genes or related sequences is indicated by a red cross. Abbreviations of predicted gene names are detailed in supplementary figure 2 (Supplementary Material online). Nucleotide positions of analyzed regions of the *B. distachyon*, rice, and Sorghum are indicated.

and *Pinb* genes were deleted from the A and B genomes of polyploid wheat species (Chantret et al. 2005). A large deletion at this locus occurred independently not only in the A and B genomes but also in the G genome of another wheat allotetraploid (*Triticum timopheevii*; Li et al. 2008).

Homologs of the *Ha-like* genes have not been found in *Panicoideae* (maize and Sorghum) and *Ehrhartoideae* (rice), all with hard endosperm (Fabijanski et al. 1988; Gautier et al. 2000; Darlington et al. 2001; Morris 2002). Nevertheless, comparative genome analysis shows that a short genomic sequence of 105 bp, with 67% amino acids similarity to *Gsp-1* gene, is present in an otherwise orthologous rice locus (called *Ha-rice-relic*; Caldwell et al. 2004; Chantret et al. 2004, 2005). *Ha-rice-relic* is located between *BGGP* gene, orthologous to wheat *BGGP* and a gene called *HIPL*, encoding a Hedgehog-interacting–like

protein, followed by a *Nodulin* gene and a cluster of *ATPase* genes (Chantret et al. 2004, 2008; fig. 1*A*). The situation is not clear for the *Panicoideae* (Sorghum, maize), which diverged earlier from *Pooideae* (wheat, barley) and *Ehrhartoideae* (rice).

In the present study, we used comparative genome analysis of orthologous *Ha* regions from wheat, *Brachypodium*, rice, and recently sequenced *Sorghum bicolor* (Paterson et al. 2009) to analyze the evolutionary origin and trace the relative time of emergence of *Ha-like* genes in grasses.

Materials and Methods

Brachypodium sylvaticum Bacterial Artificial Chromosome Library Screening

A six genome coverage bacterial artificial chromosome (BAC) library of *B. sylvaticum* (Foote et al. 2004) arrayed on high density filters was initially screened with probes prepared from separate, as well as mixture of, polymerase chain reaction (PCR) products, amplified from the hexaploid wheat *Gsp-1*, *Pina*, and *Pinb* genes using primers described in Chantret et al. (2005). Eighteen BAC clones were initially detected, indicating that homologs of these three genes are probably present in *Brachypodium*. Six of these BAC clones were retained in a second step after screening based on hybridization signal intensity, fingerprinting, and PCR confirmations. BAC clone (BAC37D5) was sequenced as described by Chantret et al. (2005).

Another round of PCR screening was also made to check the presence in *B. sylvaticum* of additional *Ha-like* gene duplicates, revealed from the analysis of the *B. distachyon* 4× genome sequence (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008; see Results). Primers were designed based on *B. distachyon* genome sequence, and the *B. sylvaticum* BAC library, organized into pools, was PCR screened.

Ha Genomic Regions from Grass Species Sequenced Genomes

Ha region from the *B. distachyon* was extracted from the available 4× coverage genome sequence (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/; release October 2008), that of rice from http://rice.plantbiology.msu.edu (release 6 January 2009), and that of *S. bicolor* from http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html (release March 2008; Paterson et al. 2009).

Sequence Annotation

Genomic sequences were annotated as described by Chantret et al. (2005). The first step of our annotation method is to detect transposable elements (TEs). Primarily, TEs were detected by a BlastN search against two databases of repetitive elements: TREP (Wicker et al. 2002, http://wheat.pw.usda.gov/ITMI/Repeats/index.shtml) and Repbase (Jurka 2000, http://www.girinst.org/Repbase_Update.html). Core domains (nucleic coordinates of known elements) were identified through BlastN alignments against TREPn. Long terminal repeats (LTRs) and limits were identified through BlastN and CENSOR (Jurka et al. 1996) alignments against Repbase and TREP databases. Putative polyproteins were identified by BlastX alignments against TREPprot. No a priori cutoff was imposed for BlastX and BlastN. We also used structural detection method using LTR_STRUC (McCarthy and McDonald 2003) and DOTTER program (Sonnhammer and Durbin 1995) for de novo identification of TEs. TE prediction and classification were performed as essentially suggested by the unified classification system for eukaryotic TEs, based on the 80–80–80 rule (Wicker et al. 2007). Retrotransposon insertion dates were estimated when necessary based on their LTR divergence as described (Charles et al. 2008).

The next step is the gene annotation. We used the gene prediction given by the program FGENESH (http://www.softberry.com; with the Monocot matrix) as well as BlastN and BlastX and TBlastX alignments against dbEST (http://www.ncbi.nlm.nih.gov/), SwissProt (http://expasy.org/sprot/), and synteny with characterized rice gene to precise gene structure and potential functions.

Finally, we systematically proceeded to a comparative annotation of genes common to several species, checking the coding sequence, and introns/exons transitions.

Gene Classification

Genes of known and unknown functions or putative genes were defined based on FGENESH predictions and the existence of rice or other *Triticeae* homologs. Hypothetical genes were identified based on FGENESH prediction only. Pseudogenes were not well predicted by FGENESH program, and frameshifts need to be introduced within the coding sequences (CDS) structure to better fit a putative function based on BlastX (mainly with rice). Large part of genes, truncated at one end (by TE insertion or unassigned DNA), potentially conversing coding capacity were qualified as "truncated." Truncated pseudogenes (genes disrupted by large insertion or deletion) and highly degenerated CDS sequences were considered as gene relics.

Identification of Duplicated Paralogous Regions in *B. distachyon*, *Oryza*, and *Sorghum*

We used the gene annotation of *Oryza sativa* (http://rice.plantbiology.msu.edu, MSU rice genome annotation release 6 January 2009) and *S. bicolor* (http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html, Sbi version 1.4, release March 2008, Paterson et al. 2009), along with analysis of the *B. distachyon* 4× genome sequence coverage (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008) to identify the syntenic blocks both within and among the three genomes. Identified syntenic blocks from *B. distachyon* were also annotated in the study (FGENESH predictions and Blast against the National Center for Biotechnology Information nonredundant databases). BlastP results ($E < 1 \times 10^{-5}$) among the predicted genes were used as input to feed the collinearity detection program MCscan with the default parameters (score >300, $E < 0.01$; Tang, Bowers, et al. 2008). MCscan generates a

number of syntenic blocks, among which we selected the set of regions that are collinear to the identified *Ha* region.

Nucleotide and Protein (amino acid) Sequence Comparisons

We used MEGA3 (Kumar et al. 2004) to make all the nucleic/proteic multiple alignments. We manually enhance these alignments taking into account special feature conservations (such as cysteine skeleton and tryptophan-rich domain [TRD]) or other domain conservation. The pairwise similarity comparisons are based on multiple alignments.

Results

Sequence analysis of the *Ha* locus in the D genome of hexaploid wheat *T. aestivum* and the orthologous region of rice were previously described (Chantret et al. 2005, 2008). We isolated and sequenced in the present study the *Ha* orthologous region from *B. sylvaticum* and conduct comparative genome analysis between all three grass species, along with that from the *B. distachyon* genome (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008) as well as the recently sequenced *S. bicolor* genome (Paterson et al. 2009; fig. 1).

Isolation and Sequencing of the *Ha* Locus in *B. sylvaticum*

Six BAC clones were retained after screening of a six genome coverage BAC library of *B. sylvaticum* (Foote et al. 2004) arrayed on high density filters, using probes prepared from wheat *Gsp-1*, *Pina*, and *Pinb* genes and further characterization, based on hybridization signal intensity, fingerprinting, and PCR. The longest BAC clone (BAC37D5) of 120,033 bp was sequenced.

Two *Ha-like* genes, *Ha-Brachy1* and *Ha-Brachy2*, were found in a 120-kb fragment of the *B. sylvaticum* genome, flanked by a *BGGP* gene on one side and by an *HIPL* and an *ATPase* gene on the other (fig. 1*A*). The tandemly duplicated *Ha-Brachy1* and *Ha-Brachy2* genes contain a single exon each and show 62% amino acid similarity to each other (fig. 2; supplementary table 1, Supplementary Material online). Predicted products of these two genes show 48–54% sequence similarity to wheat GSP-1, PinA, and PinB proteins throughout their entire length (151 and 146 amino acids; fig. 2; supplementary table 1, Supplementary Material online), indicating that an *Ha-like* gene was present in a common ancestor of the *Triticeae* and the *Brachypoidieae* tribes.

Comparison with *Ha* Locus Region from *B. distachyon*

A sequence similarity search on the available 4× shotgun sequences of the *B. distachyon* genome (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008), completed with additional 751 bp sequence, kindly provided by Dr John Vogel (USDA, Albany, USA) to fill the sequence gap, identified only the *Ha-Brachy1* gene and the *Ha-Brachy1-relic* at the *Ha* locus region of this species. The *B. distachyon Ha* locus is located on a region (super_12 contig), which is orthologous to rice chromosome12 and Sorghum chromosome8 (fig. 1*A*). The *Ha-Brachy1* gene and the *Ha-Brachy1-relic* are both very close to their *B. sylvaticum* counterparts, showing 97% and 93% amino acid similarity, respectively (fig. 2; supplementary table 1, Supplementary Material online). Thus, *Ha-Brachy1-relic* was present prior to the two *Brachypodium* species divergence, estimated to 4.2 ± 0.78 My in this study (data not shown). Surprisingly, the *Ha-Brachy2* gene is absent from the *Ha* locus region of *B. distachyon* (fig. 1*A*). The relatively old tandem duplication of *Ha-Brachy1* and *Ha-Brachy2* genes, indicated by the low level of observed amino acid similarities in *B. sylvaticum* (fig. 2; supplementary table 1, Supplementary Material online), from one side, and precise sequence comparisons between the two *Brachypodium* species from the other side, suggest that *Ha-Brachy2* has been deleted from *B. distachyon*. This occurred apparently by an illegitimate DNA recombination, driven by 62–65 bp direct repeats that flank the 842 bp deleted segment (data not shown).

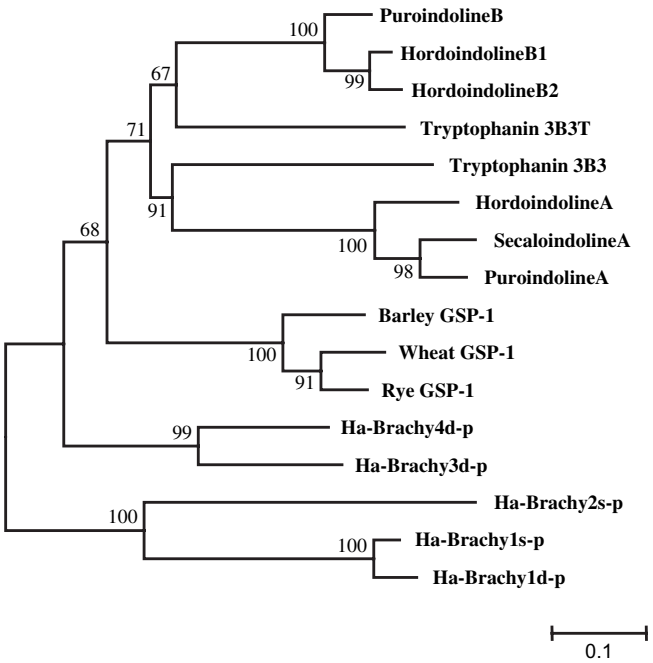


FIG. 2.—Neighbor-Joining tree, illustrating relationships between Ha-like proteins of the *Pooideae* family. Amino acid sequence alignment is shown in supplementary figure 1 (Supplementary Material online). The Ha-like proteins from *Brachypodium distachyon* are ended by (d), those of *Brachypodium sylvaticum* by (s).
Protein reference sequences:

Wheat GSP-1 (CAH10195.1), PuroindolineA (CAH10197.1), and PuroindolineB (CAH10199.1) from Chantret et al. (2005).
Rye GSP-1 (AAT76525.1) from Simeone and Lafiandra (2005).
SecaloindolineA (ABB88759.1) from Massa and Morris (2006).
Barley GSP-1 (AAV49992.1), hordoindolineA (AAV49987.1), hordoindolineB1 (AAV49986.1), and hordoindolineB2 (AAV49985.1) from Caldwell et al. (2004).
Tryptophanin 3B3 (ABU39829.1) and 3B3T (ABU39832.1) from Tanchak et al. (1998).
Ha-like proteins from *B. distachyon* from http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/ (release October 2008).
Ha-like proteins from *B. sylvaticum* were determined in this study.

Additional Duplications of *Ha-Like* Gene in the *Brachypoidieae*

Blast similarity searches of *Ha-like* genes against *B. distachyon* genome sequence (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008) allow identification of two other *Ha-like* genes that we called *Ha-Brachy3* and *Ha-Brachy4*. They are located on the same region (super_12 contig), separated by approximately 3 Mb (3,015,111 bp) from *Ha-Brachy1* gene of the *Ha* locus (fig. 1*B*). *Ha-Brachy3* and *Ha-Brachy4* genes are separated by 5.8 kb and show 83% amino acid similarity (fig. 2; supplementary table 1, Supplementary Material online) and 81% nucleotide sequence identity, indicating that they are more likely derived from recent tandem duplication between each other. *Ha-Brachy3* is inserted in this turn by an LTR retrotransposon for which we estimate insertion date to 1.2 ± 0.36 My. These additional *Ha-Brachy* gene copies show between 50% and 60% amino acid similarity to the other *Ha-like* genes (fig. 2; supplementary table 1, Supplementary Material online).

PCR-derived markers (fig. 1*B*) and BAC library screening confirm the presence of both *Ha-Brachy3 and Ha-Brachy4* as well as flanked *Myb* and *Splf* genes (fig. 1*B*) on common BAC clones of *B. sylvaticum* (data not shown). As expected, PCR analysis confirms that the retrotransposon insertion in the *Ha-Brachy3* gene of *B. distachyon* (fig. 1*B*) is not common to that of *B. sylvaticum* (data not shown). Thus, the *Ha-Brachy3* gene is not interrupted in *B. sylvaticum*.

Comparison of *B. distachyon Ha-Brachy3* and *Ha-Brachy4* genomic region (super_12 contig) with corresponding orthologous regions from rice chromosome12 (fig. 1*B*) and Sorghum chromosome8 (data not shown), identified based on flanking conserved genes, did not show any traces of *Ha-like* genes in these two later grass species. These comparisons suggest that *Ha-Brachy3* and *Ha-Brachy4* genes were generated in the *Brachypoidieae* through duplication from an *Ha-like* gene of the *Ha* locus, after divergence from *Ehrhartoideae* (rice).

The situation is not clear for *Triticeae* (wheat and barley) as their genomes are not entirely sequenced yet. Nevertheless, no *Ha-like* genes, other than *puroindolines* or *Gsp-1* genes, were so far described in these *Triticeae* species (reviewed by Morris 2002). Moreover, physical characterization of BAC clones from these species, identified as harboring *Ha-like* genes, revealed one single *Ha-like* region (Caldwell et al. 2004; Chantret et al. 2004, 2005). Further characterizations would better confirm whether this additional *Ha-like* gene duplication is specific to *Brachypoidieae*.

Thus, recurring gene duplications and/or deletions occurred independently at different stages of the grass species evolution, as indicated by the number of *Ha-like* gene copies as well as related gene fragments (partially deleted or incompletely duplicated genes) and pseudogenes found in the *Triticeae* and *Brachypoidieae Ha* locus (fig. 1*A* and *B*; supplementary table 1, Supplementary Material online; Caldwell et al. 2004; Chantret et al. 2005, 2008; discussed also hereafter).

The Orthologous *Ha* Locus Region in *S. bicolor*

A 168-kb fragment of *S. bicolor* genome (coordinates 55,276,432–55,444,261 on chromosome8; Paterson et al. 2009) was identified as containing a region orthologous to that spanning the *Ha* locus sequenced from *B. sylvaticum* (fig. 1*A*). The arguments supporting the orthologous relationship are presented hereafter. DNA sequence of this fragment is available in three contigs and includes 18 genes and putative genes (33% of the sequence), class I TEs (23%), and class II TEs (0.6%). All three numbers are substantially lower than the corresponding genome-wide averages (Paterson et al. 2009).

We found no evidence of any *Ha-like* genes or their relics, such as those found in *Pooideae* and rice, in the *Ha* orthologous region or anywhere in the sequenced Sorghum genome.

Collinearity of the Orthologous *Ha* Region in Wheat and Three Other Grasses

We compared the gene order of the 187-kb region including the D genome *Ha* locus of hexaploid wheat and amino acid sequences they encode to those of the orthologous region of rice, Sorghum, and *B. sylvaticum* (fig. 1*A*). The wheat region is larger because of the expansion of repetitive elements (fig. 1*A*)—it has a higher TE content (47%) and a lower gene content (12%). The corresponding orthologous regions of the other three species are of similar sizes and have comparable gene content (40%). Nevertheless, gene content is higher in *B. sylvaticum* than in sorghum when we extend comparison to the entire sequenced region (detailed hereafter). We detected fewer TEs in *Brachypodium* than in rice (and wheat). Although some *Brachypodium* TEs may have escaped detection because a comprehensive library of TE sequences for this species is not yet available, there is limited remaining space to detect an important proportion of TEs because of the high gene content. However, wheat, rice, and Sorghum also contain fewer TEs in this region than predicted from the genome-wide averages (Charles et al. 2008; Charles H, unpublished data).

Six single-copy genes and a cluster of *ATPase* genes are found in at least three of the four species (fig. 1*A*). The *HIPL* gene is not present in the sequenced fragment of the D genome of hexaploid wheat but instead we used the *HIPL* gene from the *Ha* region of the A genome of *T. turgidum* (fig. 1*A*; Chantret et al. 2008) for comparisons. Different levels of conservation at the amino acid level are observed for the genes when the four species are considered (fig. 3; supplementary table 2, Supplementary Material online). In Sorghum, we have not found any sequences related to the gene *Unknown-2* (fig. 1*A*).

The level of amino acid sequence similarity is consistent with closer evolutionary relationship between *Brachypodium* and wheat (*Triticeae*) than between these two species, rice and Sorghum (fig. 3; supplementary table 2, Supplementary Material online), with the exception of the *ATPase* genes. These genes are often found in clusters of complete and truncated genes, as well as pseudogenes

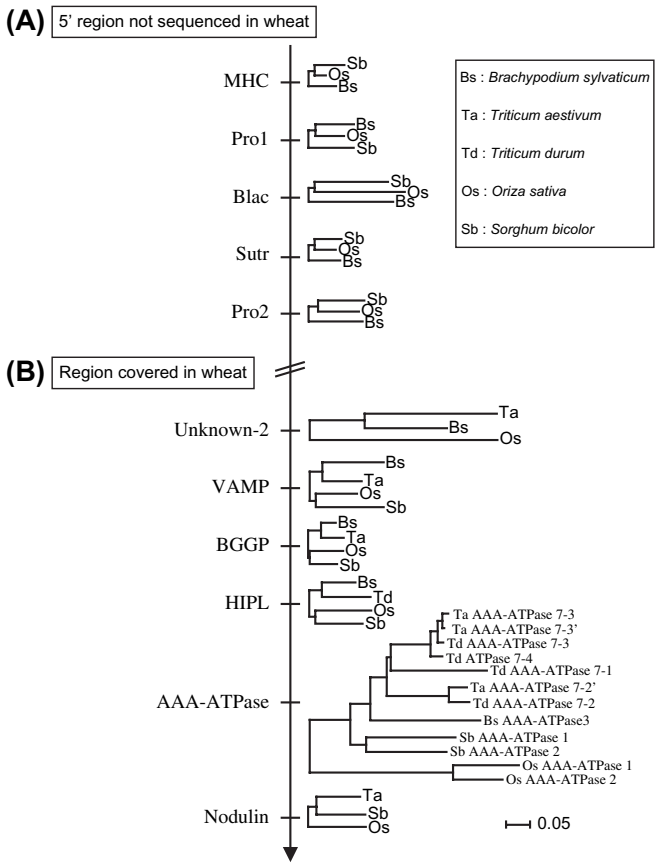


FIG. 3.—Amino acid sequence comparisons of proteins encoded by the predicted genes at the *Ha* locus regions of wheat, *Brachypodium sylvaticum*, rice, and Sorghum, shown in the order, the genes are found in *B. sylvaticum*. (*A*) Genes not present in the sequenced region of the wheat genome. (*B*) Genes present in the sequenced region of the wheat genome. Neighbor-Joining trees for genes, present in at least three of the four grass species, are shown. Pairwise sequence identities for all predicted proteins are listed in supplementary table 2 (Supplementary Material online).

(fig. 1*A*) making orthology assignments difficult (fig. 3), suggesting possible gene conversion as previously reported for rice genes (Wang et al. 2007) and further complicates the analysis.

Synteny and Collinearity Perturbation

The *chalcone synthase* gene and a *kinase* pseudogene (names shown in blue in fig. 1*A*) are not present at a corresponding site in the other grass species. Duplication of the *ATPase* gene and *Nodulin* gene cluster in inverse orientation is also specific to wheat (fig. 1*A*). In previous papers (Chantret et al. 2005, 2008), we suggested that the second cluster was ancestral, based on orientation of the *ATPase* genes in wheat, rice, and barley. However, the order and orientation of the *ATPase* and *Nodulin* genes in *B. sylvaticum* and *S. bicolor* are the same as the order and orientation of the first cluster of the genes in wheat (fig. 1*A*), suggesting that it is ancestral. *ATPase3* gene, conserved between rice and *B. sylvaticum*, has no orthologs in sorghum and was probably not covered in the wheat sequenced region.

In *B. sylvaticum*, Sorghum, and rice, the *HIPL* gene is located between the *Ha* locus and the *ATPase* genes, but in the A genome of tetraploid, wheat is located at a noncollinear position separated from the *Ha* locus by 50 kb (fig. 1*A*; Chantret et al. 2008).

Collinearity between *Brachypodium*, Rice, and Sorghum outside of the Sequenced Wheat Region

We extended comparison between the 120-kb BAC clone of *B. sylvaticum* and the corresponding regions in Sorghum and rice (fig. 1*A*). Comparisons of the additional sequence in *Brachypodium*, rice, and Sorghum confirmed a high level of collinearity and similarity between the three grass species: 10 genes (of known or unknown functions, putative genes, pseudogenes, and gene relics) out of 21 in *B. sylvaticum*, 12 in rice, and 10 in Sorghum are orthologous in at least two of the species (figs. 1*A* and 3; supplementary table 2, Supplementary Material online). A 34-kb large inversion, including eight of the genes, differentiates *B. sylvaticum* from rice and Sorghum (fig. 1*A*).

Time of Emergence and Evolutionary Origin of the *Ha-Like* Genes

We searched the available genomic sequences of *B. sylvaticum*, rice, and Sorghum to determine whether ancestral *Ha-like* genes existed before the whole-genome duplication (ancient polyploidy) of the cereal genome, which occurred ~70 My, before the radiation of the major subfamilies (Paterson et al. 2004; Salse et al. 2008; Tang, Wang, et al. 2008). The paralogous regions resulting from the ancestral duplication and collinear to the *Ha* region, based on the overall content of conserved genes, were identified for rice, Sorghum, and *B. sylvaticum* (fig. 1*C*) using MCscan search (see Materials and Methods). The two genes flanking the *Ha* locus, *BGGP* and *HIPL*, are preserved in the three collinear paralogous genomic segments from *B. distachyon*, rice, and Sorghum separated by less then 10 kb (fig. 1*C*). These intergenic sequences were searched extensively, and no *Ha-like* genes or related sequences were identified. We concluded that the *Ha-like* genes emerged after the whole-genome duplication and after the divergence of *Pooideae* and *Ehrhartoideae* from *Panicoideae*.

Homologs of *Ha-like* genes (*Gsp-1*, *Pina* and *Pinb*) encoding GSP-1 and Puroindolines were previously identified in the *Triticeae* (wheat [*Triticum* and *Aegilops* species], barley [*H. vulgare*], and rye [*S. cereale*]) and *Aveneae* (oats: *A. sativa*) tribes (Tanchak et al. 1998; Gautier et al. 2000; Kan et al. 2006; Gollan et al. 2007; Mohammadi et al. 2007; reviewed by Bhave and Morris 2008).

Puroindoline-like proteins (products of *Ha-like* genes) from wheat endosperm and several other grasses (Blochet et al. 1993; Tanchak et al. 1998; Gautier et al. 2000; Kan et al. 2006; Gollan et al. 2007; Mohammadi et al. 2007; reviewed by Bhave and Morris 2008) are characterized by a cysteine skeleton and a unique TRD. Products of *Ha-Brachy1*, *Ha-Brachy2*, *Ha-Brachy3*, and *Ha-Brachy4* genes have the cysteine skeleton and one and two conserved residues of the TRD (fig. 4*A* and *B*; supplementary fig. 1, Supplementary Material online). The N-terminal 19-amino acid signal peptide and 100-amino acid domain also found in alpha amylase inhibitor and seed storage proteins

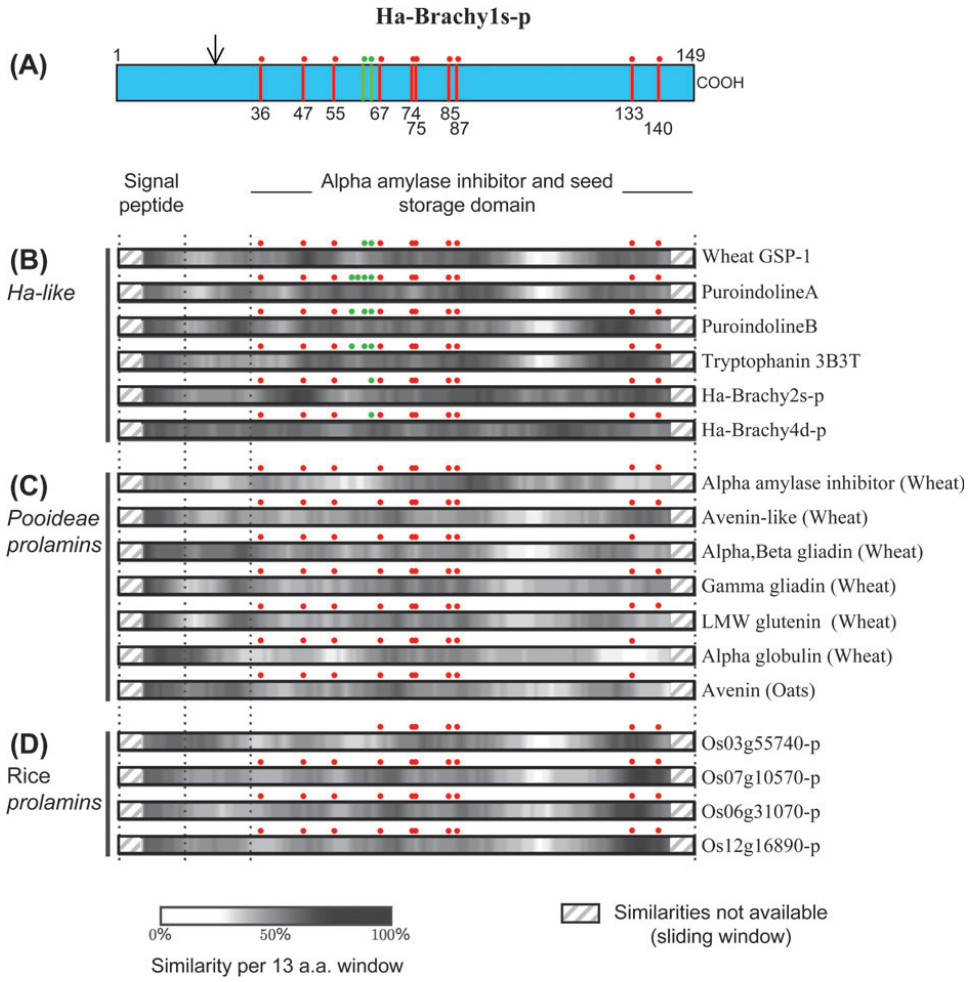


FIG. 4.—Similarity comparisons (Heatmap) between a putative protein encoded by *Ha-Brachy1s* gene of *Brachypodium sylvaticum* used as reference, GSP-1, Puroindolines, other Ha-like and Prolamins proteins. (*A*) An overview of the primary structure of the Ha-Brachy1s-p protein with its characteristic features: A 10-residue cysteine skeleton (red dots and vertical lines) and tryptophan residues (green dots and vertical lines) of the TRD. The N-terminal position of the processed protein is indicated by black arrow. (*B*) Comparison with representative Ha-like proteins from *Pooideae*. The *Brachypodium distachyon* Ha-like proteins are ended by (d), those of *B. sylvaticum* by (s). (*C*) Comparison with representative known *Pooideae* prolamins. (*D*) Comparison with representative cysteine-rich prolamins from rice (for an overview of all the rice prolamins, see supplementary fig. 2, Supplementary Material online). The 13-amino acid window sliding from the left to the right boundaries shown in gray was used for Heatmap comparisons. For the clarity of the illustration, comparisons of only representative sequences per each group of proteins showing a very high level of similarity are shown. The signal peptide is strongly conserved in all proteins, the cysteine skeleton is found in all proteins, whereas tryptophan residues of the TRD are found only in proteins encoded by *Ha-like* genes.

Uniprot reference of wheat and oats prolamin representative sequences:
Alpha/beta gliadin (P04721, P04722, P04723, P04724, and P04725) are represented by P04721.
Gamma gliadin (P04729, P04730, and P08453) are represented by P04729.
LMW glutenin (P10385 and P10386) are represented by P10386.
Avenin-like: A5A4L4.
Alpha amylase inhibitor (A4ZIZ0, Q4U195, and P01085) are represented by A4ZIZ0.
Alpha globulin: Q0Q5D4.
Avenin: DQ370180.

Reference of rice prolamin representative sequences:
Os03g55740-p represents Os11g0535100-p/Os03g55740-p/Os03g55734-p.
Os07g10570-p represents Os07g10570-p/Os07g10580-p.
Os06g31070-p represents Os06g31060-p/Os06g31070-p.
Os12g16890-p represents Os12g16880-p/Os12g16890-p/Os12g17010-p.

domain (IPR006106; http://www.ebi.ac.uk/interpro/IEntry?ac=IPR006106) are highly conserved in Ha-like proteins (fig. 4*B* and *C*; supplementary fig. 1, Supplementary Material online).

The cysteine skeleton of Puroindolines and GSP-1 proteins is also present in seed storage proteins of the prolamin superfamily (Kan et al. 2006; Bhave and Morris 2008). Prolamins encoded by *Alpha*, *Beta*, and *Gamma gliadin* and *low molecular weight (LMW)-glutenin* genes (Gao et al. 2007) as well as the *avenin* and *avenin-like* genes from oats and wheat show significant sequence similarities with Ha-like proteins (38–49%, depending on the domain, detailed in fig. 4). One gene from each of these prolamins was chosen as a reference for the subsequent sequence comparisons with *Ha-Brachy1* gene (fig. 4*C*). Although the cysteine skeleton is also generally well conserved (at least 7 out

of 10 cysteine residues found in orthologous position), no tryptophan residues of the TRD found in Ha-like proteins are found in *Pooideae* prolamins. The peptide signal domain is still strongly conserved with wheat prolamins and avenins, whereas lower conservations were observed between IPR006106 domain of the *Ha-like* encoded proteins and the corresponding *Pooideae* prolamins (fig. 4*C*).

None of the 31 prolamin genes (annotated as prolamin or putative prolamin genes) found in the rice genome (http://rice.plantbiology.msu.edu/cgi-bin/putative_function_search.pl) contains the TRD characteristic of puroindolines. The 29 complete copies of these genes group in six clades (supplementary fig. 2, Supplementary Material online), four of which encode proteins with the cysteine skeleton and the IPR001954 domain (a "child" domain of IPR006106 found in gliadins and LMW glutenins). The coding sequence of the *Ha-rice-relic* is most similar to prolamin encoding genes belonging to these four groups. Interestingly, Ha-like proteins show higher amino acid sequence similarity to these rice prolamins than to *Triticeae* prolamins: gliadins and LMW glutenins (fig. 4*C* and *D*).

Finally, our analysis shows that several prolamins of *Panicoideae*, such as beta and gamma zeins (Woo et al. 2001), exhibit cysteine skeleton. None of these could be compared (aligned) with prolamins of *Ehrhartoideae* and *Triticeae* analyzed above (data not shown) because sequences are too divergent.

Discussion

Our study shows that *Ha-like* genes are present in *Brachypoidieae* (*B. sylvaticum and B. distachyon*), tribe sister to the *Triticeae,* and *Aveneae* tribes of the *Pooideae* subfamily of grasses. Although *Ha-like* genes were not initially found in *Ehrhartoideae* (rice: *O. sativa*) and *Panicoideae* (maize: *Zea mays*, and sorghum: *S. bicolor*; Gautier et al. 2000), genome sequence analysis of the *Ha* orthologous region from rice showed a short sequence related to *Ha-like* genes (Caldwell et al. 2004; Chantret et al. 2004, 2005) that is probably a nonfunctional truncated gene remnant (*Ha-rice-relic*). Similarly, *Ha-like* genes, with specific deletions, duplications and/or truncations, were identified at the *Ha* locus region of the *Brachypoidieae* tribe and additional *Ha-like* gene duplications (*Ha-Brachy3* and *Ha-Brachy4* genes) also occurred at 3 Mb from the *Ha* locus region. Thus, it was important to analyze and confirm the absence of *Ha-like*–related sequences at the *Ha* orthologous region of recently sequenced *S. bicolor* (*Panicoideae* subfamily of grasses; Paterson et al. 2009), which diverged earlier from *Pooideae* (wheat, barley, *Brachypodium*) and *Ehrhartoideae* (rice). Overall, comparative genome analysis of orthologous *Ha* regions as well as comparison with sequences of genes encoding prolamins from wheat, *Brachypodium*, rice, and Sorghum allow elucidation of evolutionary origin and time of emergence of *Ha-like* genes in grasses.

Evolutionary Origin of *Ha-Like* Genes

As the Ha-like proteins of *Triticeae* and *Aveneae*, the *Brachypoidieae* Ha-like proteins contain one and two conserved tryptophan residues of the TRD and a conserved cysteine skeleton (fig. 4*A*; Blochet et al. 1993; Gautier et al. 2000). These conserved features suggest that the *Brachypodium* Ha-like proteins may also play a role in determining endosperm hardness/softness, although this trait has not yet been investigated in this model species.

Our sequence comparisons confirmed previous observations of a common evolutionary origin of GSP-1 and Puroindolines encoded by the *Ha-like* genes and proteins of the prolamin superfamily (Kan et al. 2006; Bhave and Morris 2008). The prolamin superfamily was defined by Kreis et al. (1985) and initially comprised three groups of seed proteins rich in prolines and glutamines: the major prolamin storage proteins of *Triticeae* (alpha, beta and gamma gliadins; LMW glutenins), the alpha amylase/trypsin inhibitors of cereal seeds, and the 2S storage albumins from oilseed rape and other dicotyledonous plants. An expanded family includes also, among others, the major prolamins of *Panicoideae* and the alpha globulins of cereals (Shewry et al. 2004; Kan et al. 2006). All these prolamins are seed-specific proteins found only in the Plant Kingdom. It has been postulated that addition of a repetitive domain in grass *prolamin* genes accelerated their divergence and drastically limited their sequence homology with prolamins from other species (Shewry et al. 2002; Nagy et al. 2005). The TRD motif is specific to GSP-1 and Puroindolines encoded by *Ha-like* genes and is not shared with other prolamins (see Results and fig. 4).

None of the rice prolamin genes are conserved at orthologous position in Sorghum, confirming previously reported highly divergent and dynamic evolution of grass prolamins, similar to other seed storage proteins, not syntenic, often clustered and known to generate recombinant copies by gene fusion, duplication, or other types of genomic rearrangements (recombination, frameshifts; Kreis et al. 1985; Shewry et al. 2002; Nagy et al. 2005; Gao et al. 2007).

Evolution of *Ha-Like* Genes by Recurring and Independent Duplications and/or Deletions

The present study supplies further insights about dynamic evolution of the *Ha-like* genes through independent duplications and/or deletions, which appear to occur recurrently at different stages of the grass species evolution. *Gsp-1*, *Pina*/*Hina*, and *Pinb*/*Hinb* genes of *Triticeae* (wheat/barley) were most likely formed by duplication of an ancestral *Ha-like* gene (Caldwell et al. 2004; Chantret et al. 2005, 2008), closely after the divergence of the three tribes (*Triticeae*, *Aveneae*, and *Brachypoidieae*). Our study also shows that independent duplications and deletions of *Ha-like* genes (*Ha-Brachy1*, *Ha-Brachy2*, *Ha-Brachy1-relic*, *Ha-Brachy3*, and *Ha-Brachy4*) have also occurred in the *Brachypoidieae* lineage (figs. 1*A*, 1*B*, and 2). Another recent duplication occurred independently in barley (*Hinb-1 and Hinb-2*) (Caldwell et al. 2004). Deletions of *Ha-like* gene occurred also independently in the A and B genomes of *T. turgidum* (*Pina and Pinb*; Chantret et al. 2005), the G genome of in *T. timopheevi* (*Pinb*; Li et al. 2008), and in *B. distachyon* (*Ha-Brachy2*) as revealed in the present study.

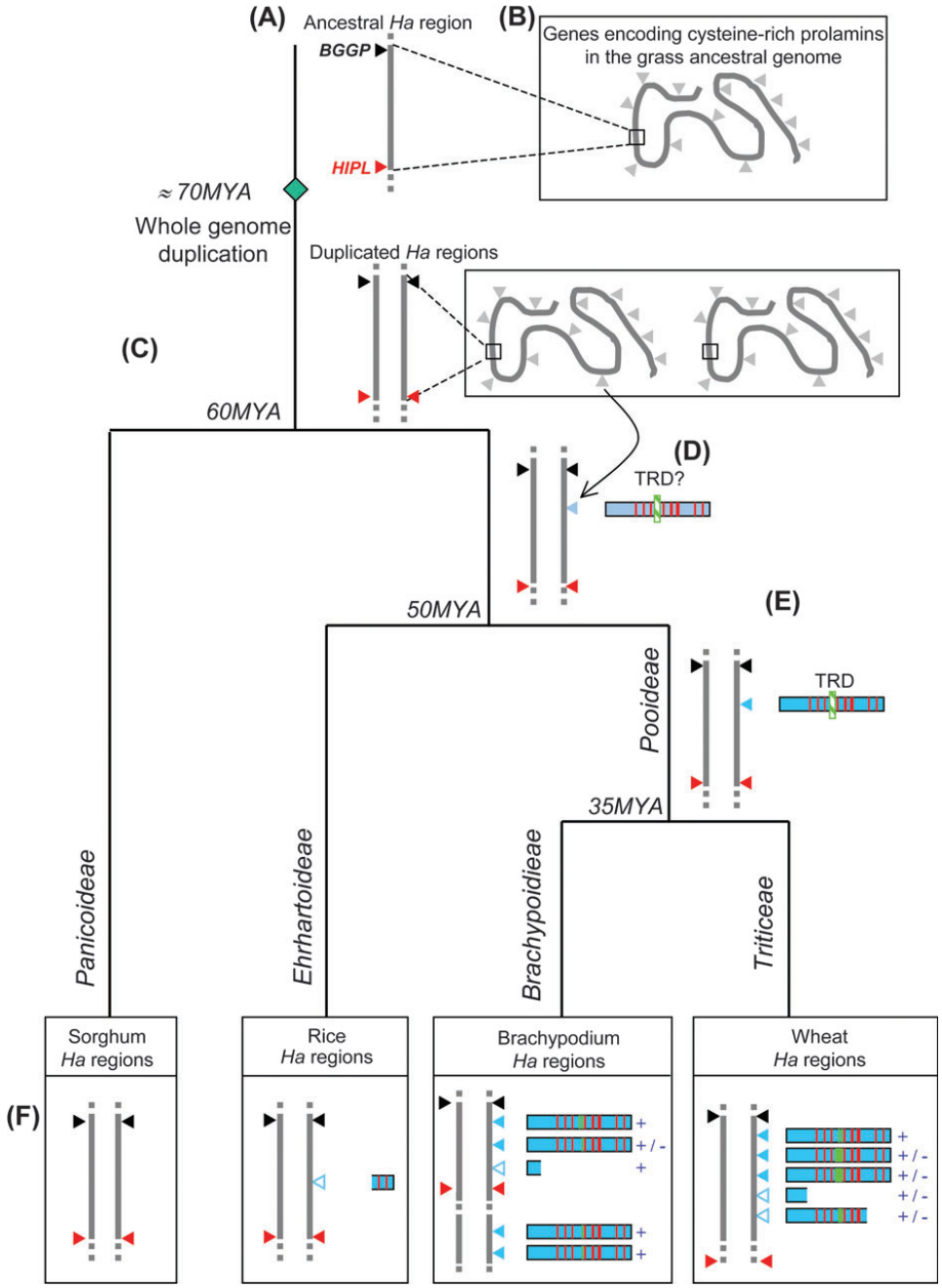


FIG. 5.—Suggested model for origin, time of emergence, and evolution of the *Ha-like* genes and locus in grasses. (*A*) Ancestral *Ha* region from the *BGGP* gene (black arrow) to the *HIPL* gene (red arrow) before the whole-genome duplication (polyploidization) predating divergence of grass subfamilies, with no *Ha-like* genes or related sequences. (*B*) Genes of the cysteine-rich prolamin superfamily were present in the ancestral grass genome. (*C*) The whole-genome duplication occurred in grasses 70 My (Paterson et al. 2004; Salse et al. 2008). (*D*) Emergence of an *Ha-like* gene in an ancestor of *Ehrhartoideae* and *Pooideae*, at one paralogous *Ha* region, after their divergence from *Panicoideae* by gene duplication, translocation, and divergence of a member of the prolamin superfamily. The cysteine residues are shown as red vertical lines in a rectangle representing the protein. (*E*) The TRD characteristic of GSP-1 and Puroindolines encoded by *Ha-like* genes (green vertical bar) appeared in an *Ha-like* gene ancestor either before or shortly after *Pooideae* and *Ehrhartoideae* diverged. (*F*) Evolution of *Ha* locus and genes by duplications, deletions, and/or truncations, occurring independently in each of the *Pooideae* and *Ehrhartoideae* families and tribes. "+" indicates *Ha-like* gene copies observed in all studies species, "+/−" those that were found deleted in specific species of the indicated grass family or tribe. *Ha-like* genes and prolamin encoding genes positions are marked with small blue triangles (filled for complete; empty for truncated or "pseudoized" copies). Rectangles represent primary structures of putative encoded proteins where the cysteine residues and tryptophan residues of the TRD are represented by, respectively, red and green vertical bars.

Time of Emergence of the *Ha* Locus in Grasses

There are two possible explanations of the presence of *Ha-like* genes on only one duplicated region in wheat, *Brachypodium*, and rice and their absence from both duplicated regions of Sorghum (the whole-genome duplication predating radiation of the major grass subfamilies is considered here) 1) The *Ha* genes emerged in this locus in a common ancestor of *Pooideae* and *Ehrhartoideae* after the duplication and after their divergence from *Panicoideae* or 2) an *Ha-like* gene was present in the ancestral grass genome but survived in only one of the two paralogous regions and only survived in some lineages, *Pooideae* and *Ehrhartoideae*,

but not *Panicoideae*. Current evidence on the evolutionary origin of *Ha-like* genes—their closer relatedness to genes encoding prolamins of *Pooideae* and *Ehrhartoideae* than to those of Sorghum—favors the first explanation.

Concluding Remarks

As summarized in figure 5, the present study allows retracing of emergence, origin, and specific evolution of the *Ha-like* genes and locus. This locus emerged, in the ancestor of the *Pooideae* and *Ehrhartoideae*, between 60 and 50 My, as a new member of the prolamin gene family. The genes were subsequently lost in *Ehrhartoideae*. After independent duplications and divergent evolution, illustrating their rapid dynamic, *Ha-like* genes gained a new function in *Pooideae*, such as wheat, underlying the soft grain phenotype. Loss of these genes in some wheats leads, in turn, to hard endosperm seeds.

Supplementary Material

Supplementary figures 1 and 2 and tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). Sequence of *B. sylvaticum* BAC clone 37D5 was deposited at EMBL/GenBank under the accession number FJ234838.

Acknowledgments

We sincerely thank Dr Graham Moore and Dr Simon Griffiths (John Innes Center, Norwich, UK) for the *B. sylvaticum* BAC library (Foote et al. 2004) and for screening BAC clones with *Gsp-1*, *Pina*, and *Pinb* gene-specific probes; Dr Nathalie Boudet (URGV) for advices in gene annotation, Mrs Cécile Huneau (URGV) for technical assistance, Dr John Vogel (USDA, Albany, USA), and PI of the international *Brachypodium* initiative (http://www.brachypodium.org) for authorizing using the 4× available genome sequence of *Brachypodium distachyon* and supplying complementary sequences for gaps filling at *Ha-like* genes.

Literature Cited

Bhave M, Morris CF. 2008. Molecular genetics of puroindolines and related genes: allelic diversity in wheat and other grasses. Plant Mol Biol. 66:205–219.

Blochet JE, Chevalier C, Forest E, Pebay-Peyroula E, Gautier MF, Joudrier P, Pezolet M, Marion D. 1993. Complete amino acid sequence of puroindoline, a new basic and cystine-rich protein with a unique tryptophan-rich domain, isolated from wheat endosperm by Triton X-114 phase partitioning. FEBS Lett. 329:336–340.

Blochet JE, Kaboulou A, Compoint JP, Marion D. 1991. Gluten proteins. In: Bushuk W, Tkachuk R, editors. Gluten Proteins 1990. St Paul (MN): American Association of Cereal Chemists. p. 314–325.

Bossolini E, Wicker T, Knobel PA, Keller B. 2007. Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. Plant J. 49:704–717.

Caldwell KS, Langridge P, Powell W. 2004. Comparative sequence analysis of the region harboring the hardness locus in barley and its colinear region in rice. Plant Physiol. 136: 3177–3190.

Chalupska D, Lee HY, Faris JD, Evrard A, Chalhoub B, Haselkorn R, Gornicki P. 2008. Acc homoeoloci and the evolution of wheat genomes. Proc Natl Acad Sci USA. 105: 9691–9696.

Chantret N, Cenci A, Sabot F, Anderson O, Dubcovsky J. 2004. Sequencing of the Triticum monococcum hardness locus reveals good microcolinearity with rice. Mol Genet Genomics. 271:377–386.

Chantret N, Salse J, Sabot F, et al. (19 co-authors). 2005. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). Plant Cell. 17:1033–1045.

Chantret N, Salse J, Sabot F, et al. (17 co-authors). 2008. Contrasted microcolinearity and gene evolution within a homoeologous region of wheat and barley species. J Mol Evol. 66:138–150.

Charles M, Belcram H, Just J, et al. (13 co-authors). 2008. Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics. 180:1071–1086.

Darlington HF, Rouster J, Hoffmann L, Halford NG, Shewry PR, Simpson DJ. 2001. Identification and molecular characterisation of hordoindolines from barley grain. Plant Mol Biol. 47:785–794.

Draper J, Mur LA, Jenkins G, Ghosh-Biswas GC, Bablak P, Hasterok R, Routledge AP. 2001. *Brachypodium distachyon*. A new model system for functional genomics in grasses. Plant Physiol. 127:1539–1555.

Fabijanski S, Chang S-C, Dukiandjiev S, Bahramian MB, Ferrara P. 1988. The nucleotide sequence of a cDNA for a major prolamin (avenin) in oat (*Avena sativa* L. cultivar Hinoat) which reveals homology with oat globulin. Biochem Physiol Pflanzen. 183:143–152.

Faris JD, Zhang Z, Fellers JP, Gill BS. 2008. Micro-colinearity between rice, Brachypodium, and Triticum monococcum at the wheat domestication locus Q. Funct Integr Genomics. 8: 149–164.

Foote TN, Griffiths S, Allouis S, Moore G. 2004. Construction and analysis of a BAC library in the grass Brachypodium sylvaticum: its use as a tool to bridge the gap between rice and wheat in elucidating gene content. Funct Integr Genomics. 4: 26–33.

Gao S, Gu YQ, Wu J, et al. (11 co-authors). 2007. Rapid evolution and complex structural organization in genomic regions harboring multiple prolamin genes in the polyploid wheat genome. Plant Mol Biol. 65:189–203.

Gaut BS. 2002. Evolutionary dynamics of grass genomes. New Phytol. 154:15–28.

Gautier MF, Aleman ME, Guirao A, Marion D, Joudrier P. 1994. *Triticum aestivum* puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. Plant Mol Biol. 25:43–57.

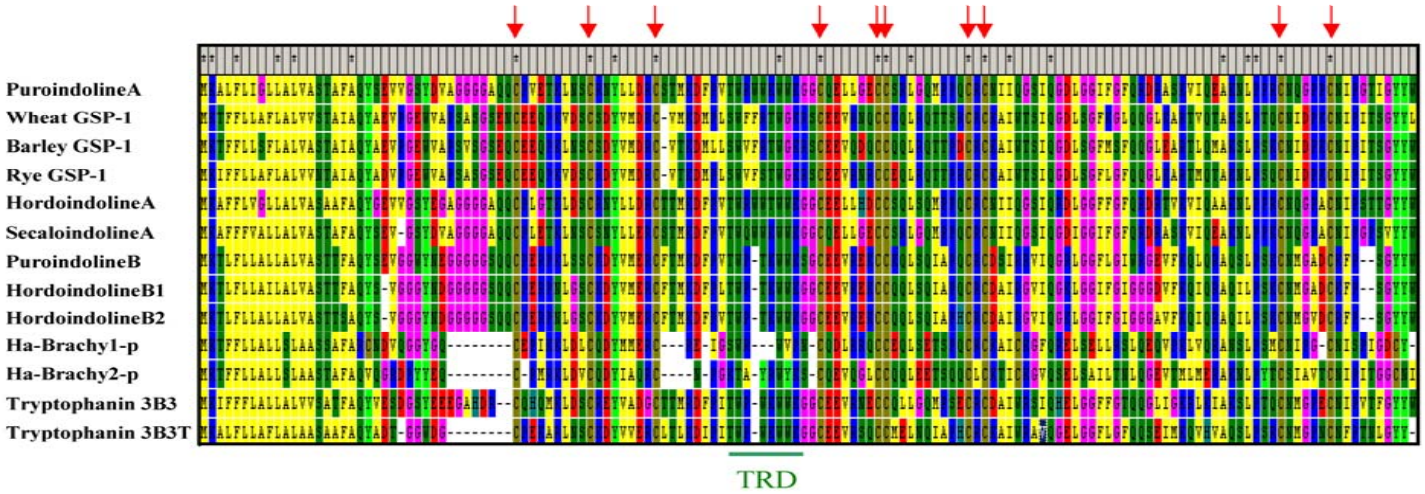
Gautier MF, Cosson P, Guirao A, Alary M, Joudrier P. 2000. Puroindoline genes are highly conserved in diploid ancestor wheats and related species but absent in tetraploid Triticum species. Plant Sci. 153:81–91.

Gollan P, Smith K, Bhave M. 2007. Gsp-1 genes comprise a multigene family in wheat that exhibits a unique combination of sequence diversity yet conservation. J Cereal Sci. 45: 184–198.

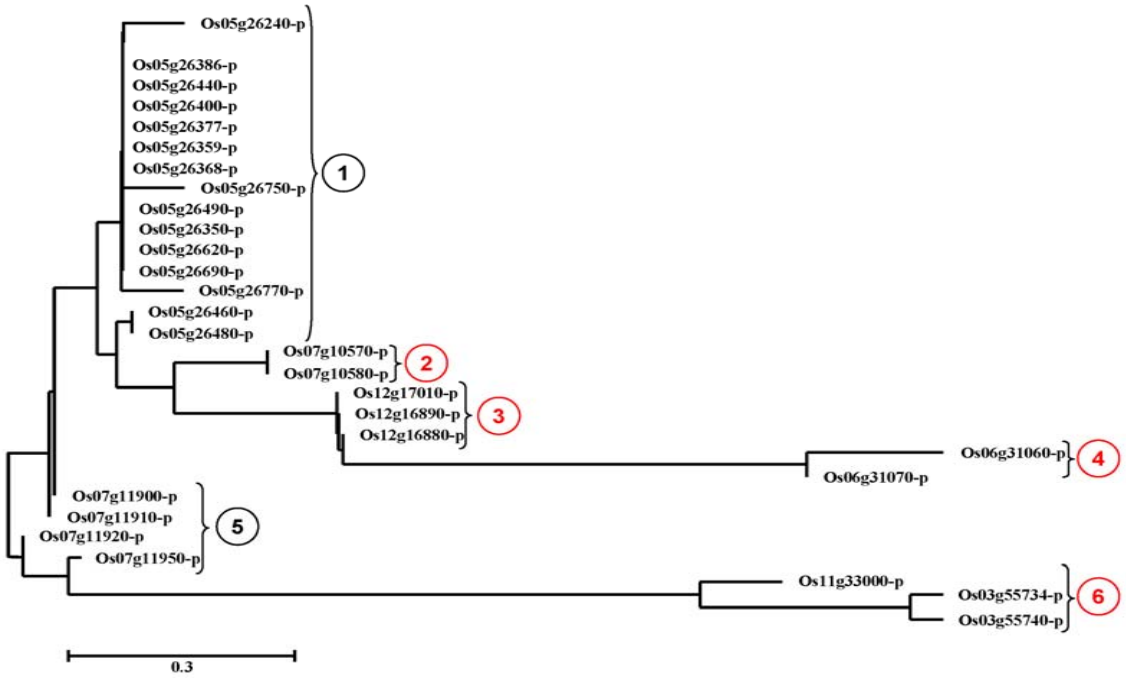
Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 16: 418–420.
Jurka J, Klonowski P, Dagman V, Pelton P. 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem. 20:119–121.
Kan Y, Wan Y, Beaudoin F, Leader DJ, Edwards K, Poole R, Wang D, Mitchell RAC, Shewry PR. 2006. Transcriptome analysis reveals differentially expressed storage protein transcripts in seeds of Aegilops and wheat. J Cereal Sci. 44:75–85.
Kellogg EA. 2001. Evolutionary history of the grasses. Plant Physiol. 125:1198–1205.
Kreis M, Forde BG, Rahman S, Miflin BJ, Shewry PR. 1985. Molecular evolution of the seed storage proteins of barley, rye and wheat. J Mol Biol. 183:499–502.
Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.
Law CN, Young CF, Brown JWS, Snape JW, Worland AJ. 1978. The study of grain protein control in wheat using whole chromosomes substitution lines. In: I.A.E. Agency. 1978. Seed protein improvement by nuclear techniques. Vienna (Austria): I.A.E. Agency. p. 483–502.
Li W, Huang L, Gill BS. 2008. Recurrent deletions of puroindoline genes at the grain hardness locus in four independent lineages of polyploid wheat. Plant Physiol. 146:200–212.
Massa AN, Morris CF. 2006. Molecular evolution of the puroindoline-a, puroindoline-b, and grain softness protein-1 genes in the tribe Triticeae. J Mol Evol. 63:526–536.
McCarthy EM, McDonald JF. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 19:362–367.
Mohammadi M, Zaidi MA, Ochalski A, Tanchak MA, Altosaar I. 2007. Immunodetection and immunolocalization of tryptophanins in oat (*Avena sativa* L.) seeds. Plant Sci. 172:579–587.
Morris CF. 2002. Puroindolines: the molecular genetic basis of wheat grain hardness. Plant Mol Biol. 48:633–647.
Nagy IJ, Takacs I, Juhasz A, Tamas L, Bedo Z. 2005. Identification of a new class of recombinant prolamin genes in wheat. Genome. 48:840–847.
Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA. 101:9903–9908.
Paterson AH, Bowers JE, Bruggmann R, et al. (45 co-authors). 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature. 457:551–556.
Prasad V, Stromberg CA, Alimohammadian H, Sahni A. 2005. Dinosaur coprolites and the early evolution of grasses and grazers. Science. 310:1177–1180.
Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C. 2008. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell. 20:11–24.
Schofield JD. 1986. Flour proteins: structure and functionality in baked products. In: Blanshard JMV, Frazier PJ, Galliard T, editors. Chemistry and physics of baking. London: The Royal Society of Chemistry. p. 14–29.
Shewry PR, Beaudoin F, Jenkins J, Griffiths-Jones S, Mills EN. 2002. Plant protein families and their relationships to food allergy. Biochem Soc Trans. 30:906–910.
Shewry PR, Jenkins J, Beaudoin F, Mills ENC. 2004. The classification, functions and evolutionary relationships of plant proteins in relation to food allergens. In: Mills ENC, Shewry PR, editors. Plant food allergens. Oxford (UK): Blackwell Science. p. 24–41.
Simeone MC, Lafiandra D. 2005. Isolation and characterisation of friabilin genes in rye. J Cereal Sci. 41:115–122.
Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene. 167:GC1–GC10.
Tanchak MA, Schernthaner JP, Giband M, Altosaar I. 1998. Tryptophanins: isolation and molecular characterization of oat cDNA clones encoding proteins structurally related to puroindoline and wheat grain softness proteins. Plant Sci. 137:173–184.
Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. Science. 320:486–488.
Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. 18:1944–1954.
Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. Genetics. 177:1753–1763.
Wicker T, Matthews D, Keller B. 2002. TREP: a database for Triticeae repetitive elements. Trends Plant Sci. 7: 561–562.
Wicker T, Sabot F, Hua-Van A, et al. (13 co-authors). 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Woo YM, Hu DW, Larkins BA, Jung R. 2001. Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. Plant Cell. 13:2297–2317.

Neelima Sinha, Associate Editor

Accepted April 13, 2009



Supplementary Figure 1: Amino acid alignement of different *Pooideae Ha-like* genes (*Pina* gene as reference). The cysteine skeleton, formed by ten cysteine residues (red arrows), and the tryptophan rich domain (TRD) are indicated.



Supplementary Figure 2: Phylogenetic relationship (Neighboor Joining tree) between the rice prolamins. Six clades could be identified, only four of which (in red) show cysteine skeleton and were compared to *Ha-like* gene (figs. 3, 4).

III Résultats complémentaires

III.1 ‘Supplemental Data’ en ligne de l’Article 2

Ces résultats complémentaires présentent les analyses de similarités entre les différents gènes *Ha-like* des *Pooideae* ainsi que les différents gènes de *prolamine* des *Poaceae*. Ces analyses présentent l’alignement des protéines des gènes *Ha-like* (Supplementary Figure 1) et l’arbre phylogénétique des gènes de *prolamine* trouvés dans le génome du riz (Supplementary Figure 2). Elles présentent aussi les similarités entre les gènes *Ha-like* (Supplementary Table 1) et entre l’ensemble des gènes adjacents trouvés dans les séquences des différentes espèces (Supplementary Table 2).

	Wheat			*B. sylvaticum*	
	Gsp-1	*Pina*	*Pinb*	*Ha-Brachy1*	*Ha-Brachy2*
Pina	56.7	-	70.9	50.3	48.8
Pinb	58.8	70.9	-	53.9	51.2
Hordo Gsp-1	90.8	59.4	60.2	49.4	46.4
Hordoa	56.7	88.6	66.2	52.0	49.4
Hordob1	55.8	69.3	91.9	51.3	51.0
Hordob2	55.2	68.0	89.9	50.7	49.0
Ha-Brachy1	48.3	50.3	53.9	-	62.3
Ha-Brachy2	51.4	48.8	51.2	62.3	-
Ha-Brachy1-relic (a)	77.8	88.9	77.8	94.4	66.7
Ha-Rice-relic (b)	68.4	63.2	68.4	84.2	70.0

(a) Comparison only with the 5' part of the *Ha-like* genes

(b) Comparison only with the 3' part of the *Ha-like* genes

Supplementary Table 1: Amino acid similarities between wheat, *Brachypodium sylvaticum,* barley (*Hordeum-Gsp-1, Hinda*, *Hindb1, Hindb2*) and rice *Ha-like* genes found at orthologous *Ha* regions.

Gene abbreviation	*Gene name*	*B. sylvaticum*		Rice		Sorghum		Wheat
Kin1	*Kinase-1*	+						
Hyp1	*Hypothetical-1*	+						
MHC	*Myosin heavy chain*	+	94%	+	94%	+		
		+		93%		+		
Act	*Actin*	+	98%	+				
Kin3	*Kinase-3*			+	78%	+		
Unk3	*Unknown-3*	+						
Kin2	*Kinase-2*	+		93%		+		
Pro1	*Sodium/proton antiporter*	+	89%	+	89%	+		
				86%				
Blac	*beta-lactamase*	+	72%	+	77%	+		
		+		71%		+		
Fbox	*F-box*	+						
Sutr	*Sucrose transport*	+	93%	+	94%	+		
		+		91%		+		
Hyp2	*Hypothetical-2*	+						
Spek	*Speckle*	+						
Abin	*ATP binding*			+	76%	+		
Car1	*carotenoid cleavage dioxygenase-1*	+	94%	+				
Car2	*carotenoid cleavage dioxygenase-2*	+						
Put1	*Putative-1*	+						
Bexp	*Beta-expansin*	+						
Expa	*Expansin*	+						
Yippe	*Yippee*	+						
Pro2	*Monovalent cation/proton exchanger*	+	85%	+	86%	+		
		+		84%		+		
Put2	*Putative-2*	+						
P450	*Cytochrome P450*			+				
Put3	*Putative-3*			+				
Put4	*Putative-4*			+				
Hyp3	*Hypothetical-3*	+						
Hyp4	*Hypothetical-4*					+		
Hyp5	*Hypothetical-5*					+		
	Pseudo-Kinase							+
	Chalcone synthase							+
	Unknown-2	+	42%	+		39%		+
		+			64%			+
	VAMP	+	91%	+	81%	+	83%	+
		+		82%		+		+
		+		+		87%		+
		+			91%			+
	BGGP	+	92%	+	90%	+	93%	+
		+		93%		+		+
		+		+		91%		+
		+			96%			+
	Gsp-1							+
	Gene3							+
	Pina							+
	Gene5							+
	Pinb							+
	Pinb-relic							+
	Pseudo-Pinb							+
	Ha-Rice-relic			+				
	Ha-Brachy1	+						
	Ha-Brachy2	+						
	Ha-Brachy1-relic	+						
	HIPL	+	88%	+	87%	+	87%	+
		+		88%		+		+
		+		+		87%		+
		+			91%			+
Atp3	*ATPase3*	+	83%	+				
HIPLp	*Pseudo-HIPL*	+						
Unk4	*Unknown-4*	+						
	ATPase1	+	76%	+	71%	+	83%	+
		+		87%		+		+
		+		+		72%		+
		+			88%			+
	ATPase-c							+
	ATPase2	+	69%	+	71%	+	80%	+
		+		80%		+		+
		+		+		70%		+
		+			86%			+
	Nodulin			+	89%	+	90%	+
				+		89%		+
	Pseudo-Nodulin							+
	Pseudo-ATPase							+
	Pseudo-ATPase							+
	ATPase							+
PB1	*PB1 domain protein*			+	85%	+		
Put5	*Putative-5*			+	88%	+		
Thio	*Thioredoxine*			+	86%	+		

Supplementary Table 2. Pairwise amino acid similarity comparison between genes of B. sylvaticum, rice, Sorghum and wheat, found at orthologous *Ha* regions (genes are presented according to their order in *B. sylvaticum*, fig. 1).

	Longueur (a.a.)	chr	Début - Fin	# cluster	BLASTP sur les gènes *Ha-like*	Résultats BLASTP sur NR	Orthologues trouvés dans le riz	Orthologues trouvés dans le sorgho
Bradi1g07520.2	128	1	5,278,138 - 5,278,521	1	-	2-S albumin	Os03g0766000	
Bradi1g50200.1	167	1	48,850,006 - 48,850,506	2	49%	Gliadin/LMW glutenin		
Bradi1g50290.1	153	1	48,943,896 - 48,944,354	2	50%	Gliadin/LMW glutenin		
Bradi1g50300.1	171	1	48,946,755 - 48,947,267	2	48%	Gliadin/LMW glutenin		
Bradi1g53990.1	150	1	52,359,847 - 52,359,398	3	49% (81 a.a.)	Alpha amylase / Trypsin inhibitor	Os07g0216600	
Bradi2g20860.1	233	2	18,283,530 - 18,284,228	4	-	Alpha globulin	Os05g0499100	Sb09g024570.1
Bradi2g20870.1	100*	2	18,299,900 - 18,300,199	4	-	HMW glutenin		
Bradi2g20910.1	295	2	18,325,829 - 18,326,713	4	-	HMW glutenin		
Bradi2g33280.1	140	2	33,350,098 - 33,350,517	5	54% (86 a.a.)	Gliadin/LMW glutenin		
Bradi2g38530.1	182	2	38,816,240 - 38,815,695	5	57% (90 a.a.)	Gliadin/LMW glutenin		
Bradi2g39910.1	105	2	39,926,289 - 39,925,975	5	55% (72 a.a.)	Gliadin/LMW glutenin		
Bradi2g39920.1	126	2	39,927,955 - 39,927,578	5	52% (91 a.a.)	Gliadin/LMW glutenin		
Bradi2g39940.1	186	2	39,934,955 - 39,934,398	5	45%	Gliadin/LMW glutenin		
Bradi3g17070.1	335	3	15,198,520 - 15,197,516	6	51%	Gliadin/LMW glutenin		
Bradi4g00420.1	151	4	171,653 - 172,105	7	-	*Ha-Brachy1*	Ha-rice-relic	
Bradi4g03950.1	160	4	3,198,715 - 3,198,236	8	-	*Ha-Brachy4*		
Bradi4g03960.1	133	4	3,210,075 - 3,209,677	8	-	*Ha-Brachy3*		

Tableau III-1. Résultats des BLASTX sur le génome de *B. distachyon* avec les protéines Ha-like comme références. Les gènes en rouge sont les *Ha-like* de *B. distachyon*, ceux en noir sont les *prolamines* proches des *Ha-like* et en bleu des *prolamines* plus lointaines.

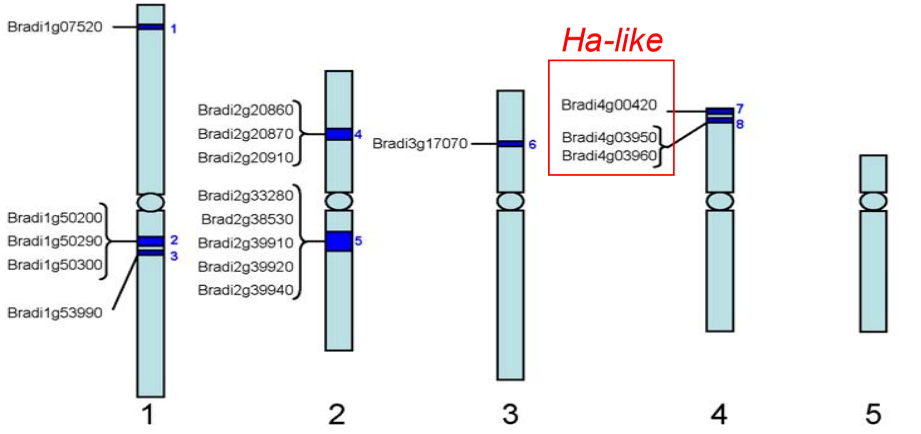


Figure III-1. Répartition en clusters de gènes *Ha-like* et de *prolamine* sur les chromosomes de *B. distachyon*.

III.2 Organisation et évolution des gènes *Ha-like* dans le génome de *Brachypodium distachyon*

La séquence du génome de *Brachypodium* s'est révélée primordiale dans l'étude de génomique comparée des *Poaceae* présentée dans l'Article 2, pour son rôle d'intermédiaire évolutif entre le riz et le blé. L'analyse présentée dans cet article suggère que les gènes '*Ha-like*' ont évolué à partir d'un gène de la famille des prolamines dans l'ancêtre commun des *Ehrhartoideae* et des *Pooideae*. La disponibilité de la séquence complète du génome de *B. distachyon* (couverture 4x) a permis d'élargir l'étude des gènes *Ha-like* à cette espèce. Nous avons donc analysé, dans le cadre de notre collaboration au consortium international de séquencage et d'annotation des gènes de *B. distachyon* (The International Brachypodium Initiative 2010), les gènes *Ha-like* et ainsi que les *prolamines*, en collaboration avec Yong-Qiang Gu (USDA, Albany, Etats-Unis) apportant ainsi ma modeste contribution aux efforts internationaux portant sur l'annotation des gènes de cette espèce.

J'ai donc eu accès à une version plus fine de la séquence du génome de *B. distachyon* me permettant de préciser l'organisation et les relations entre les gènes *Ha-like* et les *prolamines* en commençant par leur identification avec une recherche de similarité sur tout le génome. Une fois les gènes identifiés et annotés, j'ai pu faire une comparaison entre ceux-ci et ceux trouvés dans d'autres espèces de *Poaceae* (Article 2) dans une analyse phylogénétique et étudier les relations d'orthologie entre ces différents gènes.

III.2.1 Identification des gènes *Ha-like* et des *prolamines* dans le génome entier de *B. distachyon* et comparaison à ceux trouvés dans le riz

J'ai recherché, par similarité (BLASTX) des gènes ressemblant aux *Ha-like* dans le génome entier de *B. distachyon*, en utilisant des séquences de copies complètes des gènes *Ha* (*Ha-Brachy1*, *Ha-Brachy3* et *Ha-Brachy4*) identifiés dans l'Article 2. La recherche a permis de détecter 13 gènes (prédictions) montrant des niveaux de similarité variables avec les gènes de référence (http://mips.gsf.de/proj/plant/jsf/Brachypodium/index.jsp,Tableau III-1, en noir et rouge). Parmi ceux-ci, il y a évidemment trois gènes *Ha-like* (Tableau III-1, en rouge) mais également 10 autres gènes de la famille des prolamines (Tableau III-1, en noir). Nous avons aussi détecté quatre gènes plus éloignés de la famille des prolamines (1 gène de *2-S albumin*, 1 gène d'*alpha-globulin* et 2 gènes d'*HMW-glutenin*), ayant des homologues dans le blé et ne

Gène	# cluster	Annotation des gènes d'après *MSU Rice Genome Annotation (version 6)*
Os03g55730	1	SSA2 - 2S albumin seed storage family protein precursor, putative, expressed
Os03g55734	1	SSA3 - 2S albumin seed storage family protein precursor
Os03g55740	1	SSA4 - 2S albumin seed storage family protein precursor, expressed
Os05g26240	2	PROLM1 - Prolamin precursor, expressed
Os05g26250	2	PROLM3 - Prolamin precursor, putative, expressed
Os05g26260	2	retrotransposon protein, putative, Ty3-gypsy subclass
Os05g26350	2	PROLM4 - Prolamin precursor, expressed
Os05g26359	2	PROLM6 - Prolamin precursor, putative, expressed
Os05g26368	2	prolamin precursor, putative, expressed
Os05g26377	2	PROLM9 - Prolamin precursor, expressed
Os05g26386	2	prolamin precursor, putative, expressed
Os05g26400	2	prolamin precursor, putative, expressed
Os05g26440	2	PROLM10 - Prolamin precursor, putative
Os05g26460	2	PROLM11 - Prolamin precursor, expressed
Os05g26480	2	PROLM12 - Prolamin precursor, putative, expressed
Os05g26490	2	PROLM13 - Prolamin precursor, expressed
Os05g26620	2	PROLM14 - Prolamin precursor, putative, expressed
Os05g26690	2	PROLM15 - Prolamin precursor, putative, expressed
Os05g26720	2	PROLM16 - Prolamin precursor, expressed
Os05g26750	2	PROLM17 - Prolamin precursor, expressed
Os05g26770	2	PROLM18 - Prolamin precursor, expressed
Os05g41970	3	SSA1 - 2S albumin seed storage family protein precursor, expressed
Os06g31060	4	PROLM23 - Prolamin precursor, expressed
Os06g31070	4	PROLM24 - Prolamin precursor, expressed
Os07g10570	5	PROLM25 - Prolamin precursor, expressed
Os07g10580	5	PROLM26 - Prolamin precursor, expressed
Os07g11310	6	LTPL166 - Protease inhibitor/seed storage/LTP family protein precursor, expressed
Os07g11320	6	RAL1 - Seed allergenic protein RA5/RA14/RA17 precursor
Os07g11330	6	RAL2 - Seed allergenic protein RA5/RA14/RA17 precursor, expressed
Os07g11360	6	RAL3 - Seed allergenic protein RA5/RA14/RA17 precursor, expressed
Os07g11380	6	RAL4 - Seed allergenic protein RA5/RA14/RA17 precursor, expressed
Os07g11410	6	RAL5 - Seed allergenic protein RA5/RA14/RA17 precursor, expressed
Os07g11510	6	RAL6 - Seed allergenic protein RA5/RA14/RA17 precursor, expressed
Os07g11630	6	LTPL163 - Protease inhibitor/seed storage/LTP family protein precursor, expressed
Os07g11650	6	LTPL164 - Protease inhibitor/seed storage/LTP family protein precursor, expressed
Os07g11900	7	PROLM19 - Prolamin precursor, putative, expressed
Os07g11910	7	PROLM20 - Prolamin precursor, expressed
Os07g11920	7	PROLM22 - Prolamin precursor, expressed
Os07g11950	7	No gene annotation
Os07g12080	8	LTPL169 - Protease inhibitor/seed storage/LTP family protein precursor
Os07g12090	8	LTPL165 - Protease inhibitor/seed storage/LTP family protein precursor
Os11g33000	9	SSA5 - 2S albumin seed storage family protein precursor, expressed
Os12g16880	10	PROLM27 - Prolamin precursor, expressed
Os12g16890	10	PROLM28 - Prolamin precursor, expressed
Os12g17010	10	PROLM29 - Prolamin precursor, expressed
Os12g17030	10	PROLM30 - Prolamin precursor, expressed

Tableau III-2. Résultats des BLASTX sur le génome du riz avec les protèines Ha-like comme références. Ce sont des gènes de *prolamines* regroupés en 10 clusters et déjà majoritairement annotées en temps que telles.

montrant pas une similarité détectable avec les gènes *Ha*, bien que possédant des similarités structurales communes (squelette de cystéines) (Tableau III-1, en bleu). Au total, ces 17 gènes, ressemblant plus ou moins aux gènes *Ha*, sont regroupés dans 8 régions du génome (Figure III-1). Ils sont tous constitués d'un seul exon qui code pour une protéine de 128 à 335 a.a. possédant un squelette de cystéines et le domaine IPR003612 (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR003612, Plant Lipid Transfert Protein, seed sotrage, alpha amylase and trypsin inhibitor), confirmant leur appartenance à la famille des prolamines. De tous ces gènes, seuls les 3 gènes *Ha-like* présentent le domaine TRD (domaine riche en tryptophane).

En comparaison, 46 gènes montrant des similarités avec les gènes *Ha* ont été identifiés dans le riz (Tableau III-2). Comme pour *B. distachyon*, ils sont composés d'un seul exon (140-190 a.a.), contiennent un squelette de cystéines et possèdent le domaine protéine IPR003612. Ils sont également organisés en clusters (10 principaux) avec globalement plus de copies en tandem que dans *B. distachyon*.

III.2.2 Comparaison phylogénétique entre les gènes *Ha-like* et de *prolamines* des *Poaceae*

Nous avons fait un arbre phylogénétique ('bootstrap Neighbour Joining') avec les gènes détectés dans *B. distachyon* et les gènes *Ha-like* ou de *prolamines* trouvés dans le blé et le riz (Figure III-2). Pour alléger la figure, deux des clades formées par des gènes du riz ont été symbolisées par une simple flèche sur l'arbre à leur point d'embranchement. Les gènes *HMW-glutenin* sont trop distants pour être représentés. L'ensemble des gènes forme cinq groupes bien distincts (Figure III-2) :

- Gliadin/LMW-glutenin
- Ha-like
- 2-S albumin
- Alpha-globulin
- Alpha amylase/trypsin inhibitor

Les noms de ces cinq groupes sont basés sur la fonction putative des protéines encodées. La valeur du bootstrap soutenant les clades *Gliadin/LMW-glutenin* et *2-S albumin* est trop faible (33) pour déterminer avec confiance laquelle est plus proche des *Ha-like*. Par contre, les gènes des clades *Alpha-globulin* et *Alpha amylase/trypsin inhibitor* sont clairement plus éloignés des gènes *Ha-like* que ceux des deux clades évoquées précédemment.

Figure III-2. Arbre phylogénétique (‘Neighbour joining’) des différent gènes *Ha-like* et de *prolamine* trouvés dans le blé, le riz, le sorgho et *B. distachyon*. On distingue clairement 5 clades dont celle des gènes *Ha-like*.

III.2.3 Relations d'orthologie

Nous avons recherché d'éventuelles relations d'orthologie entre les gènes *Ha-like* et de *prolamines* des génomes du riz, de *Brachypodium*, du blé et du sorgho. En plus des gènes *Ha-like* trouvés en position orthologue (*Ha-Brachy1*, *Ha-rice-relic* et *Gsp-1/Pina/Pinb*), nous avons constaté que les gènes d'*alpha-globulin* (cluster 4 pour *B. distachyon*, cluster 3 pour le riz) sont présents en une seule copie conservée en position orthologue dans les génomes du riz, de *Brachypodium* et du sorgho (Tableau III-1, Tableau III-2, Figure III-2). Les gènes de *2-S albumin* (cluster 1 pour *B. distachyon*, cluster 1 pour le riz) sont conservés en positions orthologues même si le nombre variable de copies (dupliquées en tandem) empêche de préciser des relations d'orthologie un pour un (Tableau III-1, Tableau III-2, Figure III-2). Finalement, les gènes Bradi2g53990.1 de *B. sylvaticum* et Os07g11630/Os07g114650 du riz sont situés dans des régions orthologues, même si leur position n'est pas précisément conservée (Tableau III-1, Tableau III-2, Figure III-2). Tous les autres gènes de *prolamines* (36 et 11 gènes pour le riz et *B. distachyon*) ne sont pas conservés en position orthologue, soulignant l'importante dynamique de cette famille de gènes.

IV Discussion

J'ai étudié le rôle important des TEs dans l'évolution et l'organisation des génomes du blé sur une courte échelle de temps (< 4 Ma) dans les parties I et II. Il était aussi important d'apprécier l'évolution des différentes espèces de *Poaceae* en analysant la conservation des séquences géniques sur une échelle d'évolution plus longue (50-60 Ma).

A cette échelle de l'évolution, les différentes espèces de *Poaceae* ont conservé une bonne colinéarité au niveau génique, contrastant avec l'évolution très rapide de l'espace TEs. Comme attendu, ce dernier n'est pas conservé entre les différentes espèces de *Poaceae*, laissant les gènes comme seuls éléments comparables pour l'étude de l'évolution à l'échelle de cette famille. En regroupant l'analyse comparative de 12 locus, The International Brachypodium Initiative (2010) ont trouvé 62,5% des gènes conservés dans le même ordre entre *Pooideae* (*Brachypodium*, orge et blé soit 35 Ma de divergence) et 55% entre les *Poaceae* (sorgho, riz, *Brachypodium*, orge et blé soit 60 Ma de divergence).

L'étude présentée dans ce chapitre m'a permis de tracer pour la première fois l'émergence du locus *Ha* en analysant, par génomique comparée, l'ensemble des régions *Ha* (orthologues et paralogues) trouvées dans les différentes espèces de *Poaceae*. Des traces des gènes du locus *Ha* ne sont trouvées que dans une seule des deux régions *Ha* paralogues, provenant de la duplications ancestrale commune aux *Poaceae*, dans les *Ehrhartoideae* (riz) et les *Pooideae* (*Brachypodium*, orge et blé). Nous avons confirmé que les *Panicoideae* (sorgho) étaient dépourvus d'homologues aux gènes *Ha* en analysant précisément les deux régions.

D'autre part, les relations évolutives que j'ai pu établir avec les gènes codant pour des protéines de réserve du grain montrent que les gènes *Ha* sont plus proches des gènes de la famille des prolamines du riz et du *Brachypodium* que du sorgho (Article 2). Ces résultats suggèrent que l'ancêtre des gènes *Ha* a émergé il y a 50-60 Ma, comme un nouveau membre de la famille des prolamines dans l'ancêtre commun aux *Pooideae* (tribus des *Triticeae* et *Brachypoidieae*) et *Ehrhartoideae* (riz), après la divergence des *Panicoideae* (sorgho) et donc après la polyploïdisation ancestrale commune aux *Poaceae*. A partir de ce gène ancestral, les gènes *Ha* ont eu une évolution fonctionnelle dans les *Triticeae*, comme l'émergence du domaine riche en tryptophane (TRD), conférant ainsi le caractère grain tendre.

La perte de ces gènes dans certains blés polyploïdes a conduit au caractère grain dur (Chantret *et al.* 2005). La caractérisation des gènes *Ha* dans la séquence de *B. distachyon* (génome complet) et de *B. sylvaticum* (clone BAC de la région *Ha*) m'a permis de mettre en évidence une variabilité du nombre de copies de ces gènes par duplication et délétion. Cette variabilité avait déjà été observée dans les espèces du blé, généralisant l'évolution dynamique des gènes *Ha*. Une duplication supplémentaire et plus récente de ces gènes (*Ha-Brachy3* et *Ha-Brachy4*) a même été observée dans les *Brachypodieae*, confirmant la récurrence de cette dynamique, indépendamment dans les espèces du blé et de *Brachypodium*.

Conclusion générale

Dans la famille des *Poaceae*, les espèces du blé (genres *Triticum* et *Aegilops*) ont particulièrement évolué par des évènements de polyploïdisation récurrents et relativement récents ainsi que la prolifération en éléments transposables (>80% du génome), contribuant a leurs larges génomes (17 Gb pour le blé tendre hexaploïde). L'évolution et l'organisation des génomes du blé, dans ce contexte, est le point central de ma thèse. Pour mes travaux, j'ai comparé les séquences de différentes espèces de blé entre-elles ainsi qu'avec d'autres espèces de *Poaceae* pour couvrir des échelles d'évolution courtes (< 4 Ma, Parties I et II) et longues (60 Ma, Partie III).

La longue échelle d'évolution, a été analysé en appréciant la conservation des gènes au niveau du locus *Ha* et en retraçant ainsi son histoire évolutive pendant les 60 derniers Ma. J'ai ainsi élucidé l'émergence des gènes *Ha*, conférant le caractère grain tendre, à partir d'un gène de la très dynamique famille des prolamines. Contrairement à la plupart des autres membres de cette famille, il s'est fixé à un locus (*Ha*) dans l'ancêtre commun des *Pooideae* et des *Ehrhartoideae*, après leur divergence des *Panicoideae* il y a 50-60 Ma (Partie III, Article 2) et donc après la polyploïdisation ancestrale commune aux *Poaceae* (65-73 Ma). Il a aussi subi des changements fonctionnels particuliers comme l'acquisition du domaine riche en tryptophane (TRD). Son évolution dynamique est à l'origine de duplications en tandem des gènes *Ha-like*, de façon indépendante dans différentes espèces de la sous-famille des *Pooideae* possédant le locus *Ha*.

La relative conservation de synténie entre les *Poaceae*, observée dans cette étude comparative, n'est pas spécifique du locus *Ha*. Elle a déjà été observée à l'échelle de carte génétique (Moore *et al.* 1995) ainsi que récemment à l'échelle des séquences génomiques et des ESTs (Bolot *et al.* 2009). Une étude récente couvrant l'analyse de 12 locus différents a montré que 62,5% des gènes sont conservés dans le même ordre en comparant les *Pooideae* (35 Ma de divergence) contre 55% des gènes en comparant les *Poaceae* (60 Ma de divergence) (The International Brachypodium Initiative 2010).

Si les niveaux de conservation des gènes, qui représentent moins de 1% du génome, sont relativement élevés, l'espace TEs, représentant plus de 80%, est fortement dynamique. Ainsi, la dynamique des insertions et des éliminations des éléments transposables, conduisant à une prolifération différentielle dans les différents génomes du blé sur une courte échelle de temps d'évolution, ont été évaluées en analysant pour la première fois plusieurs haplotypes des principaux génomes du blé, dans plusieurs espèces et à différents niveaux de ploïdie. Les

taux de remplacement de l'espace TEs sont très variables (de 0,1% à 90% sur des périodes de 0,02 à 1,2 Ma) mais donnent une moyenne de 86% par millions d'années. Un taux similaire a été évalué par Dubcovsky et Dvorak (2007) utilisant des données moins représentatives. Avec ce taux élevé, il n'est pas étonnant de voir que l'espace TEs est complètement différent entre les principaux génomes des blés qui ont divergé il y a 2,5-4 Ma.

Les deux forces majeures impliquées dans cette importante dynamique de l'espace TEs sont donc les insertions des éléments transposables et leur élimination active par recombinaisons illégitimes. Ces dernières peuvent être associées à des recombinaisons génétiques. Mon étude montre que les TEs s'insèrent en continu dans le génome du blé, à un taux relativement stable (environ 1-3 insertions / 100 kb / Ma dans les haplotypes) sans aucun effet notable de la polyploïdie. Par contre, l'étendue des délétions par recombinaisons illégitimes, affectant des segments génomiques de tailles allant de quelques paires de bases à plusieurs dizaines de kb, est très variable. Dans les haplotypes diploïdes du génome A, les délétions sont moins importantes que les insertions de TEs en quantité d'ADN remplacée, alors que leur importance relative semble plus équilibrée pour le génome D.

La situation est encore plus accentuée pour les polyploïdes où des larges délétions par recombinaisons illégitimes sont plus importantes que les insertions observées dans le génome A. Les haplotypes des polyploïdes des génomes B et D sont très semblables et ont divergé récemment (0,02 Ma). Cette différence de variabilité haplotypique entre les génomes A et B des polyploïdes est d'autant plus surprenante que les génomes co-existent dans le même noyau depuis l'allotétraploïdisation, qui les a réunis dans *T. turgidum*.

La recombinaison génétique découverte entre haplotypes tétraploïdes et hexaploïdes ne peut s'expliquer que par la perméabilité de la barrière de polyploïdie entre ces espèces. Nous avons également mis en évidence de probables recombinaisons génétiques entre des haplotypes très divergents dans les génomes D.

L'abondance des TEs varie largement entre les différents organismes. Mes travaux de thèse ont montré qu'elle pouvait aussi varier dans des espèces proches et même entre individus de la même espèce (parties I et II). La variabilité haplotypique importante chez le blé, caractérisée pour la première fois au cours de ma thèse, confirme ainsi les hypothèses de 'Pan-Genome' (Morgante *et al.* 2007). Ces hypothèses, dérivées de l'analyse des génomes des micro-organismes et essentiellement basées chez les plantes sur l'analyse de la variabilité du maïs, suggèrent que les génomes se composent d'une partie indispensable, comme l'espace

génique, soumise à une forte pression de sélection, et d'une partie dispensable, comme l'espace TEs, où la pression de sélection est beaucoup moins forte (Morgante *et al.* 2007).

Si mon étude et celles antérieures sur le blé (Isidore *et al.* 2005, Scherrer *et al.* 2005) confirment les ressemblances avec le maïs sur la variabilité importante de la fraction dispensable de leur génomes, les mécanismes ne sont pas similaires. En effet, mon étude a révélé et confirmé la généralisation des réarrangements génomiques de plusieurs dizaines de kb, essentiellement par recombinaisons illégitimes. La prolifération différentielle des éléments transposables et surtout le mouvement des gènes par les TEs de type *Helitron* sont les principaux responsables de la variabilité haplotypique importante du maïs (Lai *et al.* 2005, Morgante *et al.* 2005, Wang et Dooner 2006, Xu et Messing 2006). Dans les espèces du blé, nous n'avons trouvé aucune évidence de réarrangements médiés par les *Helitrons*, similaires à ceux du maïs. L'étude comparative réalisée au cours de ma thèse révèle relativement peu de recombinaisons homologues inégales alors que différents types de recombinaisons illégitimes d'ADN sont les responsables majeurs de ces réarrangements. La recombinaison génétique peut également amplifier la divergence haplotypique. Les insertions de TEs représentent pour leur part un taux plus faible et relativement stable de cette variabilité haplotypique. Il n'est cependant pas étonnant que l'espace TEs des blés évolue de façon différente de celle du maïs. Les deux espèces ont divergé il y a 60 Ma (Chaluspka *et al.* 2008) et les TEs ont fortement proliférés dans ces deux groupes d'espèces de *Poaceae* les 10 derniers millions d'années de façon indépendante.

Par ailleurs, mon étude révèle que 7 des 10 régions du génome B, séquencées sans a priori d'ancrage par des gènes, portent des gènes complets (5) ou tronqués (mutation ou relique) (2). Ceci confirme des conclusions antérieures montrant que les gènes sont distribués de façon plus 'randomisée' sur les chromosomes du blé (Devos *et al.* 2005). Il est intéressant de noter que trois des six gènes tronqués sont dus à des insertions d'éléments transposables (Charles *et al.* 2008) illustrant le rôle de ces derniers dans l'évolution des génomes du blé.

Les nouvelles initiatives de séquençage actuelles permettent d'envisager la disponibilité ou une couverture plus importante ou complète d'un génome ou d'un chromosome du blé dans un avenir plus ou moins proche. Les possibilités qui seront alors offertes offriront de nombreux axes de recherche prometteurs. Cependant, les travaux de ma thèse ont bien montré l'importance de comparer différentes espèces et génotypes pour espérer avoir une vue d'ensemble dans les études de génomique comparée. La disponibilité de la

séquence de différents génotypes et génomes du blé reste nécessaire pour explorer complètement la richesse et la diversité importante des génomes du blé.

Pour finir ce rapport de thèse, et avant de passer aux annexes, je conclurai sur les apports de la thèse sur un plan plus personnel. Ma transition d'informaticien à bioinformaticien a été grandement facilitée par ma présence dans une équipe composée de biologistes et de bioanalystes. J'ai progressivement rattrapé mon retard dans les connaissances biologiques pour les mettre à profit dans mes analyses de bioinformatique en les enrichissant. J'ai acquis un savoir-faire et une expertise dans l'annotation de séquences génomiques de plantes qui, j'en suis sûr, me servira dans la suite de mon parcours. J'ai aussi appris qu'être un chercheur, ce n'était pas seulement obtenir des résultats mais également savoir les transmettre à la communauté. Je me suis donc initié à l'art délicat de la rédaction d'articles scientifiques encadré par Boulos. Arrivée en fin de thèse, mon expérience englobe maintenant l'ensemble des aspects du travail de recherche.

Références bibliographiques

Adams, K. L., R. Cronn, R. Percifield, and J. F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. USA* **100**:4649-4654.
Adams, K. L., and J. F. Wendel. 2005. Polyploidy and genome evolution in plants. *Curr. Opin. Plant. Biol.* **8**:135-141.
Agrawal, A., Q. M. Eastman, and D. G. Schatz. 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394**:744-751.
Akhunov, E. D., A. R. Akhunova, and J. Dvorak. 2005. BAC libraries of Triticum urartu, Aegilops speltoides and Ae. tauschii, the diploid ancestors of polyploid wheat. *Theor. Appl. Genet.* **111**:1617-1622.
Allouis, S., G. Moore, A. Bellec, R. Sharp, P. Faivre-Rampant, K. Mortimer, S. Pateyron, T. N. Foote, S. Griffiths, M. Caboche, and B. Chalhoub. 2003. Construction and characterisation of a hexaploid wheat (Triticum aestivum L.) BAC library from the reference germplasm 'Chinese Spring'. *Cereal research Communications* **31**:331-338.
Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic Local Alignement Search Tool. *J. Mol. Biol.* **215**:403-410.
Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389-3402.
Bao, Z., and S. R. Eddy. 2002. Automated *de novo* Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Res.earch* **12**:1269-1276.
Bartos, J., E. Paux, R. Kofler, M. Havrankova, D. Kopecky, P. Suchankova, J. Safar, H. Simkova, C. D. Town, T. Lelley, C. Feuillet, and J. Dolezel. 2008. A first survey of the rye (Secale cereale) genome composition through BAC end sequencing of the short arm of chromosome 1R. *BMC Plant Biol.* **8**:95.
Bennett, M. D., and J. B. Smith. 1976. Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **274**:227-274.
Bennett, M. D., and J. B. Smith. 1991. Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **334**:309-345.
Bennett, M. D., and I. J. Leitch. 1997. Nuclear DNA Amounts in Angiosperms—583 New Estimates. *Ann. Bot. (Lond.)* **80**:169-196.
Bennett, M. D., and I. J. Leitch. 2005. Plant Genome Size Research: A Field In Focus. *Ann. Bot. (Lond.)* **95**:1-6.
Bennetzen, J. L., and E. A. Kellogg. 1997. Do Plants Have a One-Way Ticket to Genomic Obesity? *Plant Cell* **9**:1509-1514.
Bennetzen, J. L. 2000a. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* **12**:1021-1029.
Bennetzen, J. L. 2000b. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**:251-269.
Bennetzen, J. L. 2002a. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* **115**:29-36.
Bennetzen, J. L. 2002b. The rice genome. Opening the door to comparative plant biology. *Science* **296**:60-63.
Bennetzen, J. L., J. Ma, and K. M. Devos. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot. (Lond.)* **95**:127-132.
Bergman, C. M., and H. Quesneville. 2007. Discovering and detectiong transposable elements in genome sequences. *Brief Bioinform.* **8**:382-392.
Bhave, M., and C. F. Morris. 2008. Molecular genetics of puroindolines and related genes: allelic diversity in wheat and other grasses. *Plant Mol. Biol.* **66**:205-219.

Blake, N. K., B. R. Lehfeldt, M. Lavin, and L. E. Talbert. 1999. Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat. *Genome* **42**:351-360.

Blanc, G., A. Barakat, R. Guyot, R. Cooke, and M. Delseny. 2000. Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell* **12**:1093-1101.

Blanc, G., and K. H. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**:1667-1678.

Blochet, J. E., C. Chevalier, E. Forest, E. Pebay-Peyroula, M. F. Gautier, P. Joudrier, M. Pezolet, and D. Marion. 1993. Complete amino acid sequence of puroindoline, a new basic and cystine-rich protein with a unique tryptophan-rich domain, isolated from wheat endosperm by Triton X-114 phase partitioning. *FEBS Lett.* **329**:336-340.

Blochet, J. E., A. Kaboulou, J. P. Compoint, and D. Marion. 1991. *Gluten Proteins* (Eds. Bushuk, W. and Tkachuk, R.) pp. 314-325, American Association of Cereal Chemists, Minnesota, St-Paul, USA.

Bolot, S., M. Abrouk, U. Masood-Quraishi, N. Stein, J. Messing, C. Feuillet, and J. Salse. 2009. The 'inner circle' of the cereal genomes. *Curr. Opin. Plant. Biol.* **12**:119-125.

Bossolini, E., T. Wicker, P. A. Knobel, and B. Keller. 2007. Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. *Plant J.* **49**:704-717.

Brenner, S., G. Elgar, R. Sandford, A. Macrae, B. Venkatesh, and S. Aparicio. 1993. Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* **366**:265-268.

Bretagnolle, F., and J. D. Thompson. 1995. Tansley Review No-78 - Gametes With The Somatic Chromosome-Number - Mechanisms Of Their Formation And Role In The Evolution Of Autopolyploid Plants. *New Phytologist* **129**:1-22.

Brunner, S., K. Fengler, M. Morgante, S. Tingey, and A. Rafalski. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17**:343-360.

Caldwell, K. S., P. Langridge, and W. Powell. 2004. Comparative sequence analysis of the region harboring the hardness locus in barley and its colinear region in rice. *Plant Physiol.* **136**:3177-3190.

Capy, P., C. Bazin, D. Higuet, and T. Langin. 1998. Dynamics and Evolution of Transposable Elements. *Springer, Landes Bio-sciences, Library of Congress, Austin, Texas*.

Capy, P., G. Gasperi, C. Biemont, and C. Bazin. 2000. Stress and transposable elements: co-evolution or useful parasites? *Heredity* **85 (Pt 2)**:101-106.

Cenci, A., N. Chantret, X. Kong, Y. Gu, O. D. Anderson, T. Fahima, A. Distelfeld, and J. Dubcovsky. 2003. Construction and characterization of a half million clone BAC library of durum wheat (Triticum turgidum ssp. durum). *Theor. Appl. Genet.* **107**:931-939.

Chalupska, D., H. Y. Lee, J. D. Faris, A. Evrard, B. Chalhoub, R. Haselkorn, and P. Gornicki. 2008. Acc homoeoloci and the evolution of wheat genomes. *Proc. Natl. Acad. Sci. USA* **105**:9691-9696.

Chantret, N., A. Cenci, F. Sabot, O. Anderson, and J. Dubcovsky. 2004. Sequencing of the Triticum monococcum hardness locus reveals good microcolinearity with rice. *Mol. Genet. Genomics* **271**:377-386.

Chantret, N., J. Salse, F. Sabot, S. Rahman, A. Bellec, B. Laubin, I. Dubois, C. Dossat, P. Sourdille, P. Joudrier, M. F. Gautier, L. Cattolico, M. Beckert, S. Aubourg, J. Weissenbach, M. Caboche, M. Bernard, P. Leroy, and B. Chalhoub. 2005. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *Plant Cell* **17**:1033-1045.

Chantret, N., J. Salse, F. Sabot, A. Bellec, B. Laubin, I. Dubois, C. Dossat, P. Sourdille, P. Joudrier, M. F. Gautier, L. Cattolico, M. Beckert, S. Aubourg, J. Weissenbach, M. Caboche, P. Leroy, M. Bernard, and B. Chalhoub. 2008. Contrasted microcolinearity and gene evolution within a homoeologous region of wheat and barley species. *J. Mol. Evol.* **66**:138-150.

Charles, M., H. Belcram, J. Just, C. Huneau, A. Viollet, A. Couloux, B. Segurens, M. Carter, V. Huteau, O. Coriton, R. Appels, S. Samain, and B. Chalhoub. 2008. Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. *Genetics* **180**:1071-1086.

Charles, M., H. Tang, H. Belcram, A. Paterson, P. Gornicki, and B. Chalhoub. 2009. Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of Pooideae and Ehrhartoideae, after their divergence from Panicoideae. *Mol. Biol. Evol.* **26**:1651-1661.

Chooi, W. Y. 1971. Variation in nuclear DNA content in the genus *Vicia*. *Genetics* **68**:195-211.

Comai, L., A. P. Tyagi, and M. A. Lysak. 2003. FISH analysis of meiosis in Arabidopsis allopolyploids. *Chromosome Res.* **11**:217-226.

Comai, L. 2005. The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6**:836-846.

Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, J. E. Carlson, K. Arumuganathan, A. Barakat, V. A. Albert, H. Ma, and C. W. dePamphilis. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**:738-749.

Darlington, H. F., J. Rouster, L. Hoffmann, N. G. Halford, P. R. Shewry, and D. J. Simpson. 2001. Identification and molecular characterisation of hordoindolines from barley grain. *Plant Mol. Biol.* **47**:785-794.

Devos, K. M., J. K. Brown, and J. L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* **12**:1075-1079.

Devos, K. M., J. Ma, A. C. Pontaroli, L. H. Pratt, and J. L. Bennetzen. 2005. Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proc. Natl. Acad. Sci. USA* **102**:19243-19248.

Dolezel, J., M. Kubalakova, J. Bartos, and J. Macas. 2004. Flow cytogenetics and plant genome mapping. *Chromosome Res.* **12**:77-91.

Doolittle, W. F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**:601-603.

Draper, J., L. A. Mur, G. Jenkins, G. C. Ghosh-Biswas, P. Bablak, R. Hasterok, and A. P. Routledge. 2001. Brachypodium distachyon. A new model system for functional genomics in grasses. *Plant Physiol.* **127**:1539-1555.

Du, C., Z. Swigonova, and J. Messing. 2006. Retrotranspositions in orthologous regions of closely related grass species. *BMC Evol. Biol.* **6**:62.

Dubcovsky, J., and J. Dvorak. 2007. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**:1862-1866.

Dvorak, J., P. DiTerlizzi, H.-B. Zhang, and R. P. 1993. The evolution of polyploid wheats: identification of the A genome donor species. *Genome* **36**:21-31.

Dvorak, J., M. C. Luo, and Z. L. Yang. 1998. Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing aegilops species. *Genetics* **148**:423-434.

Dvorak, J., E. D. Akhunov, A. R. Akhunov, K. R. Deal, and M. C. Luo. 2006. Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat.

Mol. Biol. Evol. **23**:1386-1396.
Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**:1-26.
Fabijanski, S., S.-C. Chang, S. Dukiandjiev, M. B. Bahramian, and P. Ferrara. 1988. The nucleotide sequence of a cDNA for a major prolamin (avenin) in oat (*Avena sativa* L. cultivar Hinoat) which reveals homology with oat globulin. *Biochem. Physiol. Pflanzen* **183**:143-152.
Faris, J. D., Z. Zhang, J. P. Fellers, and B. S. Gill. 2008. Micro-colinearity between rice, Brachypodium, and Triticum monococcum at the wheat domestication locus Q. *Funct. Integr. Genomics* **8**:149-164.
Feldman, M., F. G. H. Lupton, and T. E. Miller. 1995. *Wheats*. In: evolution of crops, 2nd ed., J. Smartt and N. W. Simmonds, eds (London : Longman Scientific).
Feldman, M., B. Liu, G. Segal, S. Abbo, A. A. Levy, and J. M. Vega. 1997. Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics* **147**:1381-1387.
Feldman, M. 2001. The origin of cultivated wheat. *Paris, France: Lavousier Tech & Doc.*
Feldman, M., and A. A. Levy. 2005. Allopolyploidy--a shaping force in the evolution of wheat genomes. *Cytogenet. Genome Res.* **109**:250-258.
Férignac, P. 1962. Test de Kolmogorov-Smirnov sur la validité d'une fonction de distribution. *Revue de statistique appliquée* **10**:13-32.
Finnegan, D. J. 1990. Transposable elements and DNA transposition in eukaryotes. *Curr Opin. Cell. Biol.* **2**:471-477.
Flagel, L., J. Udall, D. Nettleton, and J. Wendel. 2008. Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. *BMC Biol.* **6**:16.
Flavell, A. J., M. R. Knox, S. R. Pearce, and T. H. Ellis. 1998. Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. *Plant J.* **16**:643-650.
Foote, T. N., S. Griffiths, S. Allouis, and G. Moore. 2004. Construction and analysis of a BAC library in the grass Brachypodium sylvaticum: its use as a tool to bridge the gap between rice and wheat in elucidating gene content. *Funct. Integr. Genomics* **4**:26-33.
Fu, H., and H. K. Dooner. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. USA* **99**:9573-9578.
Gaeta, R. T., J. C. Pires, F. Iniguez-Luy, E. Leon, and T. C. Osborn. 2007. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. *Plant Cell* **19**:3403-3417.
Gao, X., E. R. Havecker, P. V. Baranov, J. F. Atkins, and D. F. Voytas. 2003. Translational recoding signals between gag and pol in diverse LTR retrotransposons. *Rna* **9**:1422-1430.
Gao, L., E. M. McCarthy, E. W. Ganko, and J. F. McDonald. 2004. Evolutionary history of Oryza sativa LTR retrotransposons: a preliminary survey of the rice genome sequences. *BMC Genomics* **5**:18.
Gao, S., Y. Q. Gu, J. Wu, D. Coleman-Derr, N. Huo, C. Crossman, J. Jia, Q. Zuo, Z. Ren, O. D. Anderson, and X. Kong. 2007. Rapid evolution and complex structural organization in genomic regions harboring multiple prolamin genes in the polyploid wheat genome. *Plant Mol. Biol.* **65**:189-203.
Gaut, B. S. 2002. Evolutionary dynamics of grass genomes. *New Phytologist* **154**:15-28.
Gautier, M. F., M. E. Aleman, A. Guirao, D. Marion, and P. Joudrier. 1994. Triticum aestivum puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. *Plant Mol. Biol.* **25**:43-57.

Gautier, M. F., P. Cosson, A. Guirao, M. Alary, and P. Joudrier. 2000. Puroindoline genes are highly conserved in diploid ancestor wheats and related species but absent in tetraploid Triticum species. *Plant Science* **153**:81-91.
Gollan, P., K. Smith, and M. Bhave. 2007. *Gsp-1* genes comprise a multigene family in wheat that exhibits a unique combination of sequence diversity yet conservation. *Journal of Cereal Science* **45**:184-198.
Graner, A., H. Siedler, A. Jahoor, R. G. Herrman, and G. Wenzal. 1990. Assessment of the degree and the type of restriction fragment length polymorphism in barley (*Hordeum vulgare*). *Theor. Appl. Genet.* **80**:826-832.
Griffiths, S., R. Sharp, T. N. Foote, I. Bertin, M. Wanous, S. Reader, I. Colas, and G. Moore. 2006. Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature* **439**:749-752.
Grover, C. E., Y. Yu, R. A. Wing, A. H. Paterson, and J. F. Wendel. 2008. A phylogenetic analysis of indel dynamics in the cotton genus. *Mol. Biol. Evol.* **25**:1415-1428.
Gu, Y. Q., D. Coleman-Derr, X. Kong, and O. D. Anderson. 2004. Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four *Triticeae* genomes. *Plant Physiol.* **135**:459–470.
Gu, Y. Q., J. Salse, D. Coleman-Derr, A. Dupin, C. Crossman, G. R. Lazo, N. Huo, H. Belcram, C. Ravel, G. Charmet, M. Charles, O. D. Anderson, and B. Chalhoub. 2006. Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. *Genetics* **174**:1493-1504.
Guyot, R., and B. Keller. 2004. Ancestral genome duplication in rice. *Genome* **47**:610-614.
Haberer, G., S. Young, A. K. Bharti, H. Gundlach, C. Raymond, G. Fuks, E. Butler, R. A. Wing, S. Rounsley, B. Birren, C. Nusbaum, K. F. Mayer, and J. Messing. 2005. Structure and architecture of the maize genome. *Plant Physiol.* **139**:1612-1624.
Han, B., and Y. Xue. 2003. Genome-wide intraspecific DNA-sequence variations in rice. *Curr. Opin. Plant. Biol.* **6**:134-138.
Harlan, J. R., and J. M. De Wet. 1975. On O Winge and a prayer: the origins of polyploidy. *Botanical Review* **41**:361-390.
Hawkins, J. S., H. Kim, J. D. Nason, R. A. Wing, and J. F. Wendel. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res.* **16**:1252-1261.
He, Y. D., H. Dai, E. E. Schadt, G. Cavet, S. W. Edwards, S. B. Stepaniants, S. Duenwald, R. Kleinhanz, A. R. Jones, D. D. Shoemaker, and R. B. Stoughton. 2003. Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. *Bioinformatics* **19**:956-965.
Hegarty, M. J., J. M. Jones, I. D. Wilson, G. L. Barker, J. A. Coghill, P. Sanchez-Baracaldo, G. Liu, R. J. Buggs, R. J. Abbott, K. J. Edwards, and S. J. Hiscock. 2005. Development of anonymous cDNA microarrays to study changes to the Senecio floral transcriptome during hybrid speciation. *Mol. Ecol.* **14**:2493-2510.
Hoskins, R. A., J. W. Carlson, C. Kennedy, D. Acevedo, M. Evans-Holm, E. Frise, K. H. Wan, S. Park, M. Mendez-Lago, F. Rossi, A. Villasante, P. Dimitri, G. H. Karpen, and S. E. Celniker. 2007. Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science* **316**:1625-1628.
Hovav, R., J. A. Udall, B. Chaudhary, R. Rapp, L. Flagel, and J. F. Wendel. 2008. Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proc. Natl. Acad. Sci. USA* **105**:6191-6195.
Huang, S., A. Sirikhachornkit, X. Su, J. Faris, B. Gill, R. Haselkorn, and P. Gornicki. 2002. Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate

kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc. Natl. Acad. Sci. USA* **99**:8133-8138.
Hughes, A. L. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* **48**:565-576.
Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**:796-815.
Isidore, E., B. Scherrer, B. Chalhoub, C. Feuillet, and B. Keller. 2005. Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels. *Genome Res.* **15**:526-536.
Jaillon, O., J. M. Aury, F. Brunet, J. L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot, S. Nicaud, D. Jaffe, S. Fisher, G. Lutfalla, C. Dossat, B. Segurens, C. Dasilva, M. Salanoubat, M. Levy, N. Boudet, S. Castellano, V. Anthouard, C. Jubin, V. Castelli, M. Katinka, B. Vacherie, C. Biemont, Z. Skalli, L. Cattolico, J. Poulain, V. De Berardinis, C. Cruaud, S. Duprat, P. Brottier, J. P. Coutanceau, J. Gouzy, G. Parra, G. Lardier, C. Chapple, K. J. McKernan, P. McEwan, S. Bosak, M. Kellis, J. N. Volff, R. Guigo, M. C. Zody, J. Mesirov, K. Lindblad-Toh, B. Birren, C. Nusbaum, D. Kahn, M. Robinson-Rechavi, V. Laudet, V. Schachter, F. Quetier, W. Saurin, C. Scarpelli, P. Wincker, E. S. Lander, J. Weissenbach, and H. Roest Crollius. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**:946-957.
Jaillon, O., J. M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyere, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pe, G. Valle, M. Morgante, M. Caboche, A. F. Adam-Blondon, J. Weissenbach, F. Quetier, and P. Wincker. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**:463-467.
Jiang, N., Z. Bao, X. Zhang, S. R. Eddy, and S. R. Wessler. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**:569-573.
Jones, R. N., and L. M. Brown. 1976. Chromosome evolution and DNA variation in *Crepis*. *Heredity* **36**:91-104.
Jurka, J., P. Klonowski, V. Dagman, and P. Pelton. 1996. CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20**:119-121.
Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**:418-420.
Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**:462-467.
Kaminker, J. S., C. M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas, S. Patel, E. Frise, D. A. Wheeler, S. E. Lewis, G. M. Rubin, M. Ashburner, and S. E. Celniker. 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol.* **3**:RESEARCH0084.
Kan, Y., Y. Wan, F. Beaudoin, D. J. Leader, K. Edwards, R. Poole, D. Wang, R. A.

C. Mitchell, and P. R. Shewry. 2006. Transcriptome analysis reveals differentially expressed storage protein transcripts in seeds of *Aegilops* and wheat. *Journal of Cereal Science* **44**:75-85.
Kapitonov, V. V., and J. Jurka. 2001. Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA* **98**:8714-8719.
Kapitonov, V. V., and J. Jurka. 2004. Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol.* **23**:311-324.
Kapitonov, V. V., and J. Jurka. 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol*. **3**:e181.
Kapitonov, V. V., and J. Jurka. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet.* **23**:521-529.
Kashkush, K., M. Feldman, and A. A. Levy. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* **33**:102-106.
Kellogg, E. A. 2001. Evolutionary history of the grasses. *Plant Physiol.* **125**:1198-1205.
Kidwell, M. G. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**:49-63.
Kihara, H. 1944. Discovery of the DD analyser, one of the ancestors of Triticum vulgare. *Agricultural Horticulture* **143**:253-255.
Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111-120.
Kong, X. Y., Y. Q. Gu, F. M. You, J. Dubcovsky, and O. D. Anderson. 2004. Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat. *Plant. Mol. Biol.* **54**:55-69.
Kreis, M., B. G. Forde, S. Rahman, B. J. Miflin, and P. R. Shewry. 1985. Molecular evolution of the seed storage proteins of barley, rye and wheat. *J. Mol. Biol.* **183**:499-502.
Kronmiller, B. A., and R. P. Wise. 2008. TE nest: Automated chronological annotation and visualization of nested plant transposable elements. *Plant Physiol.*
Kumar, S., K. Tamura, and M. Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* **5**:150-163.
Lai, J., Y. Li, J. Messing, and H. K. Dooner. 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci. USA* **102**:9068-9073.
Lander, E. S.L. M. LintonB. BirrenC. NusbaumM. C. ZodyJ. BaldwinK. DevonK. DewarM. DoyleW. FitzHughR. FunkeD. GageK. HarrisA. HeafordJ. HowlandL. KannJ. LehoczkyR. LeVineP. McEwanK. McKernanJ. MeldrimJ. P. MesirovC. MirandaW. MorrisJ. NaylorC. RaymondM. RosettiR. SantosA. SheridanC. SougnezN. Stange-ThomannN. StojanovicA. SubramanianD. WymanJ. RogersJ. SulstonR. AinscoughS. BeckD. BentleyJ. BurtonC. CleeN. CarterA. CoulsonR. DeadmanP. DeloukasA. DunhamI. DunhamR. DurbinL. FrenchD. GrafhamS. GregoryT. HubbardS. HumphrayA. HuntM. JonesC. LloydA. McMurrayL. MatthewsS. MercerS. MilneJ. C. MullikinA. MungallR. PlumbM. RossR. ShownkeenS. SimsR. H. WaterstonR. K. WilsonL. W. HillierJ. D. McPhersonM. A. MarraE. R. MardisL. A. FultonA. T. ChinwallaK. H. PepinW. R. GishS. L. ChissoeM. C. WendlK. D. DelehauntyT. L. MinerA. DelehauntyJ. B. KramerL. L. CookR. S. FultonD. L. JohnsonP. J. MinxS. W. CliftonT. HawkinsE. BranscombP. PredkiP. RichardsonS. WenningT. SlezakN. DoggettJ. F. ChengA. OlsenS. LucasC. ElkinE. UberbacherM. FrazierR. A. GibbsD. M. MuznyS. E. SchererJ. B. BouckE. J. SodergrenK. C. WorleyC. M. RivesJ. H. GorrellM. L. MetzkerS. L. NaylorR. S. KucherlapatiD. L. NelsonG. M. WeinstockY. SakakiA. FujiyamaM. HattoriT.

YadaA. ToyodaT. ItohC. KawagoeH. WatanabeY. TotokiT. TaylorJ. WeissenbachR. HeiligW. SaurinF. ArtiguenaveP. BrottierT. BrulsE. PelletierC. RobertP. WinckerD. R. SmithL. Doucette-StammM. RubenfieldK. WeinstockH. M. LeeJ. DuboisA. RosenthalM. PlatzerG. NyakaturaS. TaudienA. RumpH. YangJ. YuJ. WangG. HuangJ. GuL. HoodL. RowenA. MadanS. QinR. W. DavisN. A. FederspielA. P. AbolaM. J. ProctorR. M. MyersJ. SchmutzM. DicksonJ. GrimwoodD. R. CoxM. V. OlsonR. KaulN. ShimizuK. KawasakiS. MinoshimaG. A. EvansM. AthanasiouR. SchultzB. A. RoeF. ChenH. PanJ. RamserH. LehrachR. ReinhardtW. R. McCombieM. de la BastideN. DedhiaH. BlockerK. HornischerG. NordsiekR. AgarwalaL. AravindJ. A. BaileyA. BatemanS. BatzoglouE. BirneyP. BorkD. G. BrownC. B. BurgeL. CeruttiH. C. ChenD. ChurchM. ClampR. R. CopleyT. DoerksS. R. EddyE. E. EichlerT. S. FureyJ. GalaganJ. G. GilbertC. HarmonY. HayashizakiD. HausslerH. HermjakobK. HokampW. JangL. S. JohnsonT. A. JonesS. KasifA. KaspryzkS. KennedyW. J. KentP. KittsE. V. KooninI. KorfD. KulpD. LancetT. M. LoweA. McLysaghtT. MikkelsenJ. V. MoranN. MulderV. J. PollaraC. P. PontingG. SchulerJ. SchultzG. SlaterA. F. SmitE. StupkaJ. SzustakowskiD. Thierry-MiegJ. Thierry-MiegL. WagnerJ. WallisR. WheelerA. WilliamsY. I. WolfK. H. WolfeS. P. YangR. F. YehF. CollinsM. S. GuyerJ. PetersonA. FelsenfeldK. A. WetterstrandA. PatrinosM. J. MorganP. de JongJ. J. CataneseK. OsoegawaH. ShizuyaS. Choi, and Y. J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860-921.

Law, C. N., C. F. Young, J. W. S. Brown, J. W. Snape, and A. J. Worland. 1978. The study of grain protein control in wheat using whole chromosomes substitution lines. In: I.A.E. Agency, editor. *Seed Protein Improvement by Nuclear Techniques*. Austria, I.A.E. Agency. p. 483-502.

Leitch, I. J., M. W. Chase, and M. D. Bennett. 1998. Phylogenetic analysis of DNAC-values provides evidence for a small ancestral genome size in flowering plants. *Ann. Bot. (Suppl. A)* **82**:85-94.

Levis, R. W., R. Ganesan, K. Houtchens, L. A. Tolar, and F. M. Sheen. 1993. Transposons in place of telomeric repeats at a Drosophila telomere. *Cell* **75**:1083-1093.

Levy, A. A., and M. Feldman. 2004. Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. *Biological Journal of the Linnean Society* **82**:607-613.

Li, W., P. Zhang, J. P. Fellers, B. Friebe, and B. S. Gill. 2004. Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J.* **40**:500-511.

Li, W., L. Huang, and B. S. Gill. 2008. Recurrent deletions of puroindoline genes at the grain hardness locus in four independent lineages of polyploid wheat. *Plant Physiol.* **146**:200-212.

Lijavetzky, D., G. Muzzi, T. Wicker, B. Keller, R. Wing, and J. Dubcovsky. 1999. Construction and characterization of a bacterial artificial chromosome (BAC) library for the A genome of wheat. *Genome* **42**:1176-1182.

Liu, R., C. Vitte, J. Ma, A. A. Mahama, T. Dhliwayo, M. Lee, and J. L. Bennetzen. 2007. A GeneTrek analysis of the maize genome. *Proc. Natl. Acad. Sci. USA* **104**:11844-11849.

Lynch, C., and M. Tristem. 2003. A co-opted gypsy-type LTR-retrotransposon is conserved in the genomes of humans, sheep, mice, and rats. *Curr Biol* **13**:1518-1523.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**:1151-1155.

Lynch, M., and A. Force. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**:459-473.

Ma, J., and J. L. Bennetzen. 2004. Rapid recent growth and divergence of rice nuclear

genomes. *Proc. Natl. Acad. Sci. USA* **101**:12404-12410.
Ma, J., K. M. Devos, and J. L. Bennetzen. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**:860-869.
Ma, J., and J. L. Bennetzen. 2006. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc. Natl. Acad. Sci. USA* **103**:383-388.
Madlung, A., A. P. Tyagi, B. Watson, H. Jiang, T. Kagochi, R. W. Doerge, R. Martienssen, and L. Comai. 2005. Genomic changes in synthetic Arabidopsis polyploids. *Plant J.* **41**:221-230.
Maizel, J. V., Jr., and R. P. Lenk. 1981. Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc. Natl. Acad. Sci. USA* **78**:7665-7669.
Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376-380.
Massa, A. N., and C. F. Morris. 2006. Molecular evolution of the puroindoline-a, puroindoline-b, and grain softness protein-1 genes in the tribe Triticeae. *J. Mol. Evol.* **63**:526-536.
McCarthy, E. M., and J. F. McDonald. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**:362-367.
McClintock, B. 1950. Mutable loci in maize. *In: Carnegie Institute of Washington Year Book,* pp. 174–181, Washington.
McFadden, E. S., and E. R. Sears. 1946. The origin of Triticum spelta and its free-threshing hexaploid relatives. *Journal of Heredity* **37**:81-89.
Meyers, B. C., S. V. Tingey, and M. Morgante. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11**:1660-1676.
Miller, A. K., G. Galiba, and J. Dubcovsky. 2006. A cluster of 11 CBF transcription factors is located at the frost tolerance locus Fr-Am2 in Triticum monococcum. *Mol. Genet. Genomics* **275**:193-203.
Mohammadi, M., M. A. Zaidi, A. Ochalski, M. A. Tanchak, and I. Altosaar. 2007. Immunodetection and immunolocalization of tryptophanins in oat (*Avena sativa* L.) seeds. *Plant Science* **172**:579-587.
Moore, G., K. M. Devos, Z. Wang, and M. D. Gale. 1995. Cereal genome evolution. Grasses, line up and form a circle. *Curr Biol* **5**:737-739.
Morgante, M., S. Brunner, G. Pea, K. Fengler, A. Zuccolo, and A. Rafalski. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**:997-1002.
Morgante, M., E. De Paoli, and S. Radovic. 2007. Transposable elements and the plant pan-genomes. *Curr. Opin. Plant. Biol.* **10**:149-155.
Morris, C. F. 2002. Puroindolines: the molecular genetic basis of wheat grain hardness. *Plant Mol. Biol.* **48**:633-647.
Moullet, O., H. B. Zhang, and E. S. Lagudah. 1999. Construction and characterisation of a large DNA insert library from the D genome of wheat. *Theor. Appl. Genet.* **99**:305-313.
Muehlbauer, G. J., B. S. Bhau, N. H. Syed, S. Heinen, S. Cho, D. Marshall, S.

Pateyron, N. Buisine, B. Chalhoub, and A. J. Flavell. 2006. A hAT superfamily transposase recruited by the cereal grass genome. *Mol. Genet. Genomics* **275**:553-563.
Nagy, I. J., I. Takacs, A. Juhasz, L. Tamas, and Z. Bedo. 2005. Identification of a new class of recombinant prolamin genes in wheat. *Genome* **48**:840-847.
Nesbitt, M., and D. Samuel. 1996. From staple crop to extinction? The archaeology and history of the hulled wheats. In. *First International Workshop on Hulled Wheats. Promoting the conservation and use of underutilized and neglected crops*. Rome : International Plant Genetic Resources Institute.41-100.
Ohno, S. 1970. Evolution by gene duplication. *Springer Verlag, New York.*
Orgel, L. E., and F. H. Crick. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**:604-607.
Ouyang, S., and C. R. Buell. 2004. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.* **32**:D360-363.
Ozkan, H., A. A. Levy, and M. Feldman. 2001. Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops-Triticum) group. *Plant Cell* **13**:1735-1747.
DePaoli, E. 2009. Evergreen with a fossil genome: a novel paradigm for higher plant genome evolution revealed by the analysis of gymnosperm species. *Proceeding of the 9th international plant molecular biology congress*, St Louis, USA, October 25-30, 2009, editor Perry Gustafson.
Paterson, A. H., J. E. Bowers, and B. A. Chapman. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. USA* **101**:9903-9908.
Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V. Grigoriev, E. Lyons, C. A. Maher, M. Martis, A. Narechania, R. P. Otillar, B. W. Penning, A. A. Salamov, Y. Wang, L. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C. McCann, R. Ming, D. G. Peterson, R. Mehboob ur, D. Ware, P. Westhoff, K. F. Mayer, J. Messing, and D. S. Rokhsar. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**:551-556.
Paux, E., D. Roger, E. Badaeva, G. Gay, M. Bernard, P. Sourdille, and C. Feuillet. 2006. Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J.* **48**:463-474.
Pereira, V. 2004. Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. *Genome Biol.* **5**:R79.
Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl, and K. L. Shaw. 2000. Evidence for DNA loss as a determinant of genome size. *Science* **287**:1060-1062.
Petrov, D. A. 2002a. Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* **61**:531-544.
Petrov, D. A. 2002b. DNA loss and evolution of genome size in Drosophila. *Genetica* **115**:81-91.
Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Saniyal, H. Kim, K. Collura, D. S. Brar, S. Jackson, R. A. Wing, and O. Panaud. 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Res.* **16**:1262-1269.
Prasad, V., C. A. Stromberg, H. Alimohammadian, and A. Sahni. 2005. Dinosaur coprolites and the early evolution of grasses and grazers. *Science* **310**:1177-1180.
Pritham, E. J. 2009. Transposable elements and factors influencing their success in eukaryotes. *J. Hered.* **100**:648-655.

International Rice Genome Sequencing Project 2005. The map-based sequence of the rice genome. *Nature* **436**:793-800.
Qi, L., B. Echalier, B. Friebe, and B. S. Gill. 2003. Molecular characterization of a set of wheat deletion stocks for use in chromosome bin mapping of ESTs. *Funct. Integr. Genomics* **3**:39-55.
Ramsey, J., and D. W. Schemske. 1998. Pathways, Mechanisms, an Rates of polyploid formation in flowering plants. *Annual Review Of Ecology And Systematics* **29**:467-501.
Rapp, R. A., J. A. Udall, and J. F. Wendel. 2009. Genomic expression dominance in allopolyploids. *BMC Biol.* **7**:18.
Rieseberg, L. H. 2001. Polyploid evolution: keeping the peace at genomic reunions. *Curr Biol* **11**:R925-928.
Riley, R., and V. Chapman. 1958. Genetic control of cytologically diploid behaviour of hexaploid wheat. *Nature* **182**:713-715.
Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**:365-386.
Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16**:944-945.
Sabot, F., D. Simon, and M. Bernard. 2004. Plant transposable elements, with an emphasis on grass species. *Euphytica* **139**:227–247.
Sabot, F., R. Guyot, T. Wicker, N. Chantret, B. Laubin, B. Chalhoub, P. Leroy, P. Sourdille, and M. Bernard. 2005. Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. *Mol. Genet. Genomics* **274**:119-130.
Safar, J., J. Bartos, J. Janda, A. Bellec, M. Kubalakova, M. Valarik, S. Pateyron, J. Weiserova, R. Tuskova, J. Cihalikova, J. Vrana, H. Simkova, P. Faivre-Rampant, P. Sourdille, M. Caboche, M. Bernard, J. Dolezel, and B. Chalhoub. 2004. Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.* **39**:960-968.
Salse, J., S. Bolot, M. Throude, V. Jouffe, B. Piegu, U. Masood Quraishi, T. Calcagno, R. Cooke, M. Delseny, and C. Feuillet. 2008a. Identification and Characterization of Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution. *Plant Cell*.
Salse, J., V. Chague, S. Bolot, G. Magdelenat, C. Huneau, C. Pont, H. Belcram, A. Couloux, S. Gardais, A. Evrard, B. Segurens, M. Charles, C. Ravel, S. Samain, G. Charmet, N. Boudet, and B. Chalhoub. 2008b. New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the *SPA* genomic region with the S genome of the diploid relative *Aegilops speltoides*. *BMC Genomics* 9:555.
SanMiguel, P., A. Tikhonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P. S. Springer, K. J. Edwards, M. Lee, Z. Avramova, and J. L. Bennetzen. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**:765-768.
SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima, and J. L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**:43-45.
SanMiguel, P. J., W. Ramakrishna, J. L. Bennetzen, C. S. Busso, and J. Dubcovsky. 2002. Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). *Funct. Integr. Genomics* **2**:70-80.
Sasaki, T., W. Jianzhong, T. Itoh, and T. Matsumoto. 2005. [Complete rice genome sequence information: the key for elucidation of Rosetta stones of other cereal genome]. *Tanpakushitsu Kakusan Koso* **50**:2167-2173.
Scherrer, B., E. Isidore, P. Klein, J. S. Kim, A. Bellec, B. Chalhoub, B. Keller, and

C. Feuillet. 2005. Large intraspecific haplotype variability at the Rph7 locus results from rapid and recent divergence in the barley genome. *Plant Cell* **17**:361-374.
Schiex, T., A. Moisan, and P. Rouzé. 2001. EuGene: An Eucaryotic Gene Finder that combines several sources of evidence. *Computational Biology, Eds. O. Gascuel and M-F. Sagot, LNCS 2066, pp. 111-125*.
Schlueter, J. A., P. Dixon, C. Granger, D. Grant, L. Clark, J. J. Doyle, and R. C. Shoemaker. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**:868-876.
Schofield, J. D. 1986. Flour proteins: structure and functionality in baked products. In : Blanshard JMV, Frazier PJ, Galliard T, editors. *Chemistry and Physics of Baking.* London, The Royal Society of Chemistry. p. 14-29.
Semon, M., and K. H. Wolfe. 2007. Consequences of genome duplication. *Curr Opin Genet Dev* **17**:505-512.
Shaked, H., K. Kashkush, H. Ozkan, M. Feldman, and A. A. Levy. 2001. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* **13**:1749-1759.
Shewry, P. R., F. Beaudoin, J. Jenkins, S. Griffiths-Jones, and E. N. Mills. 2002. Plant protein families and their relationships to food allergy. *Biochem Soc. Trans.* **30**:906-910.
Shewry, P. R., J. Jenkins, F. Beaudoin, and E. N. C. Mills. 2004. The classification, functions and evolutionary relationships of plant proteins in relation to food allergens. In: Mills ENC, Shewry PR, editors. *Plant Food Allergens.* Oxford, U.K., Blackwell Science. p. 24-41.
Silverman, B. W. 1986. Density Estimation for Statistics and Data Analysis, Chapman & Hall, eds Hardcover.
Simeone, M. C., and D. Lafiandra. 2005. Isolation and characterisation of friabilin genes in rye. *Journal of Cereal Science* **41**:115-122.
Slotkin, R. K., and R. Martienssen. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**:272-285.
Smith, D., and R. Flavell. 1975. Characterization of the wheat genome by renaturation kinetics. *Chromosoma* **50**:223-242.
Soltis, D. E., V. A. Albert, J. Leebens-Mack, C. D. Bell, A. H. Paterson, C. Zheng, D. Sankoff, C. W. DePamphilis, P. Kerr Wall, and P. S. Soltis. 2009. Polyploidiy and angiosperm diversification. *American Journal of Botany* **96**:336-348.
Song, K., P. Lu, K. Tang, and T. C. Osborn. 1995. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proc. Natl. Acad. Sci. USA* **92**:7719-7723.
Song, R., and J. Messing. 2003. Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc. Natl. Acad. Sci. USA* **100**:9055-9060.
Sonnhammer, E. L., and R. Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**:GC1-10.
Stein, N. 2007. Triticeae genomics: advances in sequence analysis of large genome cereal crops. *Chromosome Res.* **15**:21-31.
Swanson-Wagner, R. A., Y. Jia, R. DeCook, L. A. Borsuk, D. Nettleton, and P. S. Schnable. 2006. All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proc. Natl. Acad. Sci. USA* **103**:6805-6810.
Tanchak, M. A., J. P. Schernthaner, M. Giband, and I. Altosaar. 1998. Tryptophanins: isolation and molecular characterization of oat cDNA clones encoding proteins structurally related to puroindoline and wheat grain softness proteins. *Plant*

Science **137**:173-184.
Tang, H., J. E. Bowers, X. Wang, R. Ming, M. Alam, and A. H. Paterson. 2008a. Synteny and collinearity in plant genomes. *Science* **320**:486-488.
Tang, H., X. Wang, J. E. Bowers, R. Ming, M. Alam, and A. H. Paterson. 2008b. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**:1944-1954.
Tate, J. A., Z. Ni, A. C. Scheen, J. Koh, C. A. Gilbert, D. Lefkowitz, Z. J. Chen, P. S. Soltis, and D. E. Soltis. 2006. Evolution and expression of homeologous loci in Tragopogon miscellus (*Asteraceae*), a recent and reciprocally formed allopolyploid. *Genetics* **173**:1599-1611.
The International Brachypodium Initiative 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**:763-768.
Tikhonov, A. P., P. J. SanMiguel, Y. Nakajima, N. M. Gorenstein, J. L. Bennetzen, and Z. Avramova. 1999. Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc. Natl. Acad. Sci. USA* **96**:7409-7414.
Vedel, F., and M. Delseny. 1987. Repetitivity and variability of higher plant genomes. *Plant Physiol. biochem.* **25**:191-210.
Velasco, R., A. Zharkikh, M. Troggio, D. A. Cartwright, A. Cestaro, D. Pruss, M. Pindo, L. M. Fitzgerald, S. Vezzulli, J. Reid, G. Malacarne, D. Iliev, G. Coppola, B. Wardell, D. Micheletti, T. Macalma, M. Facci, J. T. Mitchell, M. Perazzolli, G. Eldredge, P. Gatto, R. Oyzerski, M. Moretto, N. Gutin, M. Stefanini, Y. Chen, C. Segala, C. Davenport, L. Dematte, A. Mraz, J. Battilana, K. Stormo, F. Costa, Q. Tao, A. Si-Ammour, T. Harkins, A. Lackey, C. Perbost, B. Taillon, A. Stella, V. Solovyev, J. A. Fawcett, L. Sterck, K. Vandepoele, S. M. Grando, S. Toppo, C. Moser, J. Lanchbury, R. Bogden, M. Skolnick, V. Sgaramella, S. K. Bhatnagar, P. Fontana, A. Gutin, Y. Van de Peer, F. Salamini, and R. Viola. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* **2**:e1326.
Vicient, C. M., A. Suoniemi, K. Anamthawat-Jonsson, J. Tanskanen, A. Beharav, E. Nevo, and A. H. Schulman. 1999. Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum. *Plant Cell* **11**:1769-1784.
Vitte, C., and J. L. Bennetzen. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc. Natl. Acad. Sci. USA* **103**:17638-17643.
Wang, J., L. Tian, H. S. Lee, N. E. Wei, H. Jiang, B. Watson, A. Madlung, T. C. Osborn, R. W. Doerge, L. Comai, and Z. J. Chen. 2006. Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics* **172**:507-517.
Wang, Q., and H. K. Dooner. 2006. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proc. Natl. Acad. Sci. USA* **103**:17644-17649.
Wang, X., H. Tang, J. E. Bowers, F. A. Feltus, and A. H. Paterson. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* **177**:1753-1763.
Wendel, J. F., R. C. Cronn, J. S. Johnston, and H. J. Price. 2002. Feast and famine in plant genomes. *Genetica* **115**:37-47.
Wicker, T., D. Matthews, and B. Keller. 2002. TREP: a database for Triticeae repetitive elements. *Trends Plant. Sci.* **7**:561-562.
Wicker, T., R. Guyot, N. Yahiaoui, and B. Keller. 2003a. CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiol.* **132**:52-63.
Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z. D. Liu, J. Dubcovsky, and B. Keller. 2003b. Rapid genome divergence at orthologous low molecular weight glutenin

loci of the A and Am genomes of wheat. *Plant Cell* **15**:1186-1197.
Wicker, T., and B. Keller. 2007. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res.*
Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman. 2007a. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**:973-982.
Wicker, T., N. Yahiaoui, and B. Keller. 2007b. Contrasting rates of evolution in *Pm3* loci from three wheat species and rice. *Genetics* **177**:1207-1216.
Wicker, T., S. G. Krattinger, E. S. Lagudah, T. Komatsuda, M. Pourkheirandish, T. Matsumoto, S. Cloutier, L. Reiser, H. Kanamori, K. Sato, D. Perovic, N. Stein, and B. Keller. 2009. Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight into the structure of diploid and hexaploid triticeae gene pools. *Plant Physiol.* **149**:258-270.
Wicker, T., W. Zimmermann, D. Perovic, A. H. Paterson, M. Ganal, A. Graner, and N. Stein. 2005. A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley Hv-eIF4E locus: recombination, rearrangements and repeats. *Plant J.* **41**:184-194.
Willcox, G. 1996. Evidence for plant exploitation and vegetation history from three Early Neolithic pre-pottery sites on the Euphrate (Syria). *Vegetation History and Archaeobotany* **5**:143–152.
Wolfe, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2**:333-341.
Woo, Y. M., D. W. Hu, B. A. Larkins, and R. Jung. 2001. Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. *Plant Cell* **13**:2297-2317.
Xu, J. H., and J. Messing. 2006. Maize haplotype with a helitron-amplified cytidine deaminase gene copy. *BMC Genet.* **7**:52.
Yan, L., V. Echenique, C. Busso, P. SanMiguel, W. Ramakrishna, J. L. Bennetzen, S. Harrington, and J. Dubcovsky. 2002. Cereal genes similar to Snf2 define a new subfamily that includes human and mouse genes. *Mol. Genet. Genomics* **268**:488-499.
Yan, L., A. Loukoianov, G. Tranquilli, M. Helguera, T. Fahima, and J. Dubcovsky. 2003. Positional cloning of the wheat vernalization gene VRN1. *Proc. Natl. Acad. Sci. USA* **100**:6263-6268.
Yan, L., M. Helguera, K. Kato, S. Fukuyama, J. Sherman, and J. Dubcovsky. 2004. Allelic variation at the VRN-1 promoter region in polyploid wheat. *Theor. Appl. Genet.* **109**:1677-1686.
Zabala, G., and L. O. Vodkin. 2005. The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. *Plant Cell* **17**:2619-2632.
Zuccolo, A., A. Sebastian, J. Talag, Y. Yu, H. Kim, K. Collura, D. Kudrna, and R. A. Wing. 2007. Transposable element distribution, abundance and role in genome size variation in the genus Oryza. *BMC Evol. Biol.* **7**:152.

Annexes

Annexe 1: Liste des figures

Introduction

Matériels et méthode d'annotation

Résultats - Partie I

Résultats - Partie II

Résultats - Partie III

Annexe 2 : Liste des tableaux

Introduction

Résultats-Partie I

Résultats-Partie II

Résultats-Partie III

Annexe 3 : Article 3

DOI: 10.1534/genetics.106.060756

Types and Rates of Sequence Evolution at the High-Molecular-Weight Glutenin Locus in Hexaploid Wheat and Its Ancestral Genomes

Yong Qiang Gu,*[,1] Jérôme Salse,[†,‡] Devin Coleman-Derr,* Adeline Dupin,[†] Curt Crossman,* Gerard R. Lazo,* Naxin Huo,* Harry Belcram,[†] Catherine Ravel,[‡] Gilles Charmet,[‡] Mathieu Charles,[†] Olin D. Anderson* and Boulos Chalhoub[†]

*United States Department of Agriculture-Agricultural Research Service, Western Regional Research Center, Albany, California 94710, [†]Laboratory of Genome Organization, Unité de Recherches en Génomique Végétale (URGV-INRA), 91057 Evry Cedex, France and [‡]UMR INRA-UBP ASP Amélioration et Santé des Plantes, 63039 Clermont Ferrand, France

Manuscript received May 15, 2006
Accepted for publication August 29, 2006

ABSTRACT

The *Glu-1* locus, encoding the high-molecular-weight glutenin protein subunits, controls bread-making quality in hexaploid wheat (*Triticum aestivum*) and represents a recently evolved region unique to Triticeae genomes. To understand the molecular evolution of this locus region, three orthologous *Glu-1* regions from the three subgenomes of a single hexaploid wheat species were sequenced, totaling 729 kb of sequence. Comparing each *Glu-1* region with its corresponding homologous region from the D genome of diploid wheat, *Aegilops tauschii*, and the A and B genomes of tetraploid wheat, *Triticum turgidum*, revealed that, in addition to the conservation of microsynteny in the genic regions, sequences in the intergenic regions, composed of blocks of nested retroelements, are also generally conserved, although a few nonshared retroelements that differentiate the homologous *Glu-1* regions were detected in each pair of the A and D genomes. Analysis of the indel frequency and the rate of nucleotide substitution, which represent the most frequent types of sequence changes in the *Glu-1* regions, demonstrated that the two A genomes are significantly more divergent than the two B genomes, further supporting the hypothesis that hexaploid wheat may have more than one tetraploid ancestor.

POLYPLOIDIZATION, an evolutionary process resulting in more than one genome per cell, has played a significant role in the evolutionary history of plants, particularly in agriculturally important crops (Masterson 1994; Soltis and Soltis 1999; Wendel 2000). Bread wheat is an allohexaploid species (*Triticum aestivum* L. $2n = 6x = 42$), consisting of three sets of highly related genomes (A, B, and D). Hexaploid wheat originated from two independent polyploidization events. The first event involved the hybridization of two diploid progenitors: an ancestor of *Triticum urartu* ($2n = 2x = 14$, genome AA) and an unconfirmed species (BB genome) related to *Aegilops speltoides* ($2n = 2x = 14$, genome SS), which resulted in cultivated allotetraploid emmer wheat (*T. turgidum* ssp. *dicoccum*, $2n = 4x = 28$, genomes AABB) (Dvorak *et al.* 1992; Blake *et al.* 1999). In the second event, which occurred 8000–10,000 years ago, an ancestor of the diploid *Aegilops tauschii* (DD genome) hybridized with the allotetraploid to form a hexaploid wheat ($2n = 6x = 42$) (Feldman *et al.* 1995).

Because of its relatively recent speciation, wheat represents an excellent system for studying evolutionary events that occur in genomes shortly after polyploidization. Different types of genetic and epigenetic changes, including DNA removal, changes in gene expression, reactivation of transposable elements, and functional diversification of duplicate genes, are all known to be significant to the evolutionary process in polyploid species (Ozkan *et al.* 2001; Shake *et al.* 2001; Comai *et al.* 2002; He *et al.* 2003; Kashkush *et al.* 2003; Blanc and Wolfe 2004). However, the particular molecular mechanisms underlying the rapid genome evolution observed in polyploid genomes are not well understood. Sequence comparisons of polyploid genomes with their ancestral genomes may represent the strategy of choice to identify sequence changes caused by recent genome evolution. Recently, this strategy has been successfully used to elucidate the molecular basis of the polyploidy-related deletion of the *Hardness* locus from the tetraploid wheat (*T. turgidum*) (Chantret *et al.* 2005)

One of the great challenges for wheat genome research is the size of Triticeae genomes. Compared with model plant species, such as Arabidopsis (130 Mb) and rice (430 Mb), the bread wheat genome (~16,000 Mb) is extremely large. We now know that although the sudden increase in chromosome numbers caused by polyploidization contributed to the overall increase in the wheat

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. DQ537335, DQ537336, and DQ537337.

[1]*Corresponding author:* USDA-ARS, Western Regional Research Center, 800 Buchanan St., Albany, CA 94710. E-mail: ygu@pw.usda.gov

genome size, the replication and insertion of repetitive DNA, particularly long terminal repeat (LTR) retrotransposons (SanMiguel *et al.* 1996), is another major cause for genome expansion (Bennetzen *et al.* 2005). In wheat, repetitive DNA accounts for ~90% of the genome, of which retrotransposons constitute 60–80% (SanMiguel *et al.* 2002; Wicker *et al.* 2003; Gu *et al.* 2004).

In addition to being responsible for genome size variation, transposable elements often stimulate other types of genomic rearrangements, including unequal homologous and illegitimate recombination to remove nuclear DNA (Devos *et al.* 2002; Ma *et al.* 2004; Chantret *et al.* 2005). Their transpositions can cause gene inactivation or changing expression of adjacent genes (Kashkush *et al.* 2003; Gu *et al.* 2004). Repetitive DNA elements also make useful markers for revealing recent genomic changes because they are abundant and believed to be neutral under natural selection. Repetitive DNA, including retrotransposons, has a relative short turnover period (Ma *et al.* 2004) so that colinear retrotransposons are usually not found in distantly related genomes. Our previous studies of comparative analyses of orthologous high-molecular-weight (HMW) glutenin regions indicated that only a few remnants of elements are colinear in the closely related wheat A, B, and D genomes that diverged from each other within the last 4–5 million years (MY) (Huang *et al.* 2002; Gu *et al.* 2004; Kong *et al.* 2004). Because of this, the intergenic regions are largely not conserved between the A, B, and D genomes of wheat, although gene colinearity is retained (Gu *et al.* 2004). Furthermore, several recent studies suggested that retrotransposable elements may be one of the major causes for intraspecific sequence variations in the intergenic regions (Brunner *et al.* 2005; Scherrer *et al.* 2005). A comparison of homologous regions from different inbred lines in maize indicated that >59% of the compared sequence is noncolinear largely due to the insertion of novel or allele-specific LTR retrotransposons (Brunner *et al.* 2005).

In addition to retroelements, other genomic rearrangements, particularly sequence insertions and deletions (indels), play significant roles in genome evolution (Petrov *et al.* 2000; Gregory 2004). In rice, an estimated minimum of 119 Mb of sequence has been removed from the genome in the last 5 MY (Ma and Bennetzen 2004). A comparison of the *Hardness* locus in diploid and polyploid wheat species has indicated that multiple genomic deletions occurred independently in different genomes and that indels are one of the primary evolutionary mechanisms involved in reconstructing this domestication gene region (Chantret *et al.* 2005). Gene disruption caused by deletions and large inversions has resulted in various haplotypes in the wheat A genome (Isidore *et al.* 2005). Despite rapid DNA rearrangements and intraspecific violation of genetic colinearity observed in recent studies (Brunner *et al.* 2005; Chantret *et al.* 2005; Isidore *et al.* 2005; Scherrer *et al.* 2005), it remains to be shown whether such rapid sequence diversification has occurred throughout the whole genome or only in certain genetic loci. Future sequence comparisons of regions from homologous genomes, such as wheat genomes at different ploidy levels, will help answer this question.

In wheat, the *Glu-1* locus encodes HMW glutenin protein subunits, which are the major determinants of bread-making quality for wheat flour, making *Glu-1* one of the most important genetic loci in wheat and frequently the target of genetic engineering efforts for the improvement of grain quality (Blechl and Anderson 1996; Rooke *et al.* 1999). The wheat HMW glutenin locus and the orthologous loci of barley and rye are unique to Triticeae species (Shewry and Tatham 1990), which suggests that these loci evolved relatively recently within the Triticeae tribe. Studies focusing on these specific orthologous HMW glutenin regions from diploid (Anderson *et al.* 2003) and tetraploid wheat genomes (Kong *et al.* 2004), as well as from barley (Gu *et al.* 2003), have provided the first view of genome evolution in three homeologous regions from the wheat A, B, and D genomes (Gu *et al.* 2004). In this article, we report sequencing the three HMW glutenin locus regions located on the long arms of homeologous group 1 chromosomes from a single hexaploid wheat species, which has allowed us to compare the homologous *Glu-1* regions from the hexaploid wheat with its diploid and tetraploid ancestors. The sequence conservation and divergence detected could shed light on the evolutionary history of the wheat genomes.

MATERIALS AND METHODS

Isolation and sequencing of hexaploid wheat BACs: Hexaploid wheat BAC clones were obtained from the *T. aestivum cv.* Renan BAC library by screening with PCR primers specific to HMW glutenin genes. Assignment of BAC clones to the A, B, and D genomes was based on their characterization by restriction fragment length polymorphisms and contig assembly by BAC fingerprinting on agarose gels as described previously (Kong *et al.* 2004; Chantret *et al.* 2005). Selection of BACs for sequencing was based on contig maps constructed for each orthologous *Glu-1* locus and Southern hybridization data using various probes flanking the x-type and y-type HMW glutenin genes. BAC clones that covered the largest regions of each *Glu-l* locus were selected. The shotgun-sequencing libraries for hexaploid wheat BAC clones were constructed by the method described by either Gu *et al.* (2003) or Chantret *et al.* (2005). Plasmid DNAs from single colonies were purified and inserts were sequenced from both directions with T7 and T3 primers using BigDye terminator chemistry (Applied Biosystems, Foster City, CA) on ABI3730 capillary sequencers. Gaps between sequence contigs were filled and sequenced by primer-walking and transposition reactions (Finnzymes, Espoo, Finland). Gaps caused by GC-rich regions were usually filled by resequencing using the dGTP BigDye terminator chemistry (Applied Biosystems).

Sequence analysis: For sequence assembly, a target of 10-fold coverage was chosen. Base calling and quality of the shotgun sequences were processed using Phred (Ewing and Green 1998). The sequence data generated for each BAC clone was used to assemble continuous contigs using both the Lasergene SeqMan module (DNAStar) (http://www.DNAStar.com) and Phrap assembly engine (http://www.phrap.org). In some cases, assemblies with two different programs helped resolve gap regions and the order of contigs. The consensus sequences were obtained by analyzing at least three sequence reads (on both strands) or using sequencing methods based on two different labeling procedures applied on one strand. To validate the accuracy of the sequence assembly, digestion patterns of BAC DNAs with *Hin*dIII, *Eco*RI, and *Not*I were compared with the predicted restriction patterns of the computer-assembled sequences.

For annotation, the assembled sequences of the A, B, and D genome BACs from the hexaploid wheat were compared with the previous annotations for the *Glu-1* regions from the barley BAC (AY268139), diploid D-genome BAC (AF497474), tetraploid A-genome BAC (AY494981), and B genome BAC (AY368673). In addition, a homology search was performed against NCBI nonredundant and dbEST databases using BLASTN, BLASTX, and TBLASTX algorithms. FGENESH (http://www.softberry.com/nucleo.html) and GENESCAN (http://genemark.mit.edu/GENESCAN.htm) were used for gene prediction. DNA repetitive elements were identified with NCBI BLAST searches, with DNAStar MegAlign dot-plot analysis, and by comparison with the Triticeae Repeat Sequence Database at the GrainGenes website at http://wheat.pw.usda.gov/ITMI/Repeats/. The definition and naming of new retrotransposons were done according to the method described by SanMiguel *et al.* (1998).

The rate of nonsynonymous (K_a) *vs.* synonymous (K_s) substitutions were calculated for four genes (x-type, y-type HMW glutenin genes, globulin, and protein kinase genes) with MEGA3 (Kumar *et al.* 2004). Dating of retrotransposon insertions was performed on the basis of the method described in SanMiguel *et al.* (1998). The number of transition and transversion mutations was calculated using the MEGA3 software (Kumar *et al.* 2004). The average substitution rates of colinear retrotransposons in the two homologous genomes were also calculated in the same way and used to estimate the divergence times on the basis of the method described by Wicker *et al.* (2003).

The statistical significance regarding differences in rates of nucleotide substitution and indel frequency among each pair of the homologous A, B, and D genomes was measured by randomly sampling ~10-kb intervals in the *Glu-1* regions. The Student's *t*-test was employed for statistical analyses.

RESULTS

Isolation and sequencing of *Glu-1* regions from hexaploid wheat: Previously, we have reported sequences of the orthologous HMW glutenin regions from the D genome of the diploid *A. tauschii* and the A and B genomes of the tetraploid *T. turgidum*, representing the ancestral genomes of hexaploid wheat. To further examine the evolution of these genomic regions in hexaploid wheat, we screened a large insert BAC library constructed from the hexaploid wheat species *cv.* Renan with PCR primers specific to the different copies of HMW glutenin genes. The screening of this BAC library identified 12 HMW glutenin BAC clones from the A genome, 10 from the B genome, and 18 from the D genome. Further characterization of these BAC clones indicated that none of the BAC clones belonging to the A and B genomes contains both the y-type and x-type HMW glutenin genes (data not shown). BAC DNA was fingerprinted using the *Hin*dIII restriction enzyme and, using FPC software, contigs were built that span the *Glu-1* loci from each wheat genome (Soderlund *et al.* 2000). The assembled contig maps and the BAC Southern data were used to guide the selection of BAC clones for sequencing. BAC clone 706G08 represents the *Glu-1* region from the D genome. The *Glu-1* region from the B genome was covered by the sequences from two BAC clones, 1289J04 and 2001P20. Three BACs were sequenced for the A-genome HMW glutenin region (BAC clones 1344C16, 754K10, and 1031P08). A 152,010-bp region derived from the D-genome BAC, a 285,506-bp region derived from the combined sequences of two B-genome BAC clones, and a 292,044-bp region from three A-genome BAC clones were annotated using a combination of bioinformatics tools and BLAST searches against publically available databases.

Gene colinearity between HMW glutenin genomic sequences across the three wheat ploidy levels: Sequence comparison of orthologous *Glu-1* regions between three homeologous wheat genomes revealed microcolinearity in the genic regions, but large-scale sequence divergence in the intergenic regions (Gu *et al.* 2004). Sequencing the corresponding *Glu-1* regions from the three subgenomes of hexaploid wheat (A, B, and D) enables us to analyze sequence changes compared to homologous genomes from the diploid and tetraploid wheats. In all the sequenced *Glu-1* regions, there are a total of six genes (Gu *et al.* 2004). The order of these genes is as follows: a leucine-rich-repeat receptor-like protein kinase, a globulin, a y-type HMW glutenin, a duplicate globulin, an x-type HMW glutenin, and a serine/threonine protein kinase. In the three genomes of hexaploid wheat, these six genes exist in the same order and orientation as those in the ancestor's genomes (Figure 1). Previously, a tandem duplication of an ancestral region containing the globulin and HMW glutenin genes was characterized at the *Glu-1* locus regions (Kong *et al.* 2004). This duplication, which led to the presence of the x-type and y-type copies of HMW glutenin genes in the wheat genome, occurred after the divergence of wheat and barley (Gu *et al.* 2003; Figure 1). Extension of sequences in the 5′ HMW glutenin regions identified a second tandem duplication, two copies of the receptor kinase genes present near the *Glu-1* locus. This duplication exists in at least the A and D genomes of the hexaploid wheat, suggesting the occurrence of the event prior to the divergence of the A and D genomes. The paralogous receptor kinase 1a and 1b in the hexaploid D genome share 90% sequence identity, whereas the orthologous receptor kinase 1a from the hexaploid D genome and the hexaploid A

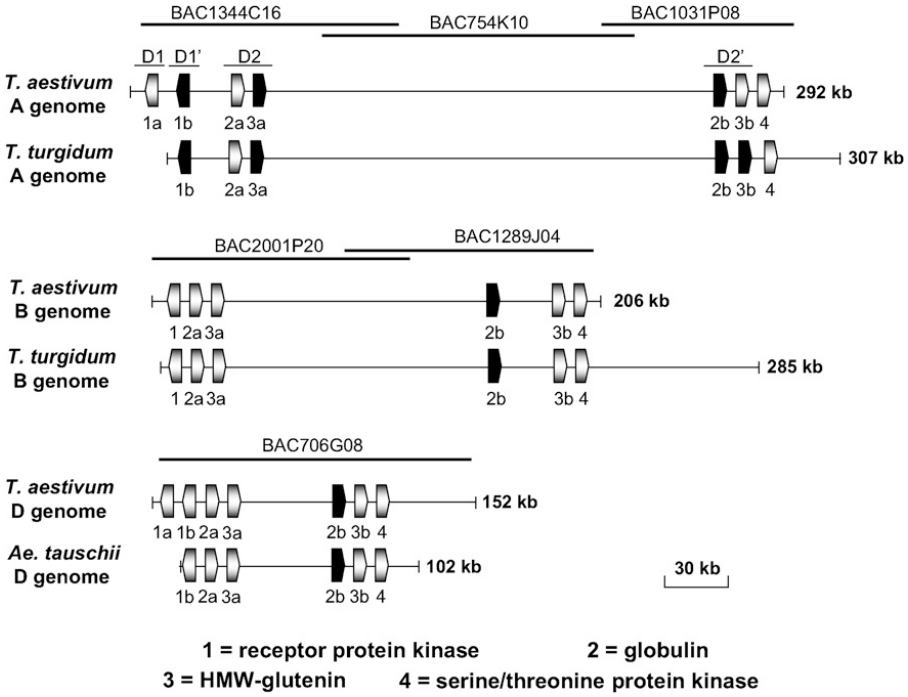


FIGURE 1.—Gene organization and comparison of orthologous and homologous *Glu-1* regions from ancestral wheat and hexaploid wheat genomes. Genes are represented by numbered pentagons, with their corresponding names listed below. The letters "a" and "b" denote distinct copies of the duplicated genes. The arrows of the pentagons indicate the direction of the potential transcription of these genes. Solid pentagons represent inactive genes or pseudogenes caused by sequence rearrangements identified in the analysis. Positions and addresses of BAC clones selected for complete sequencing of the regions are indicated. Regions involved in sequence duplications resulting in two copies of receptor protein kinase (D1, D1′) and two copies of HMW glutenin (D2, D2′) are labeled only in the A genome. A 30-kb segment is provided as a scale reference.

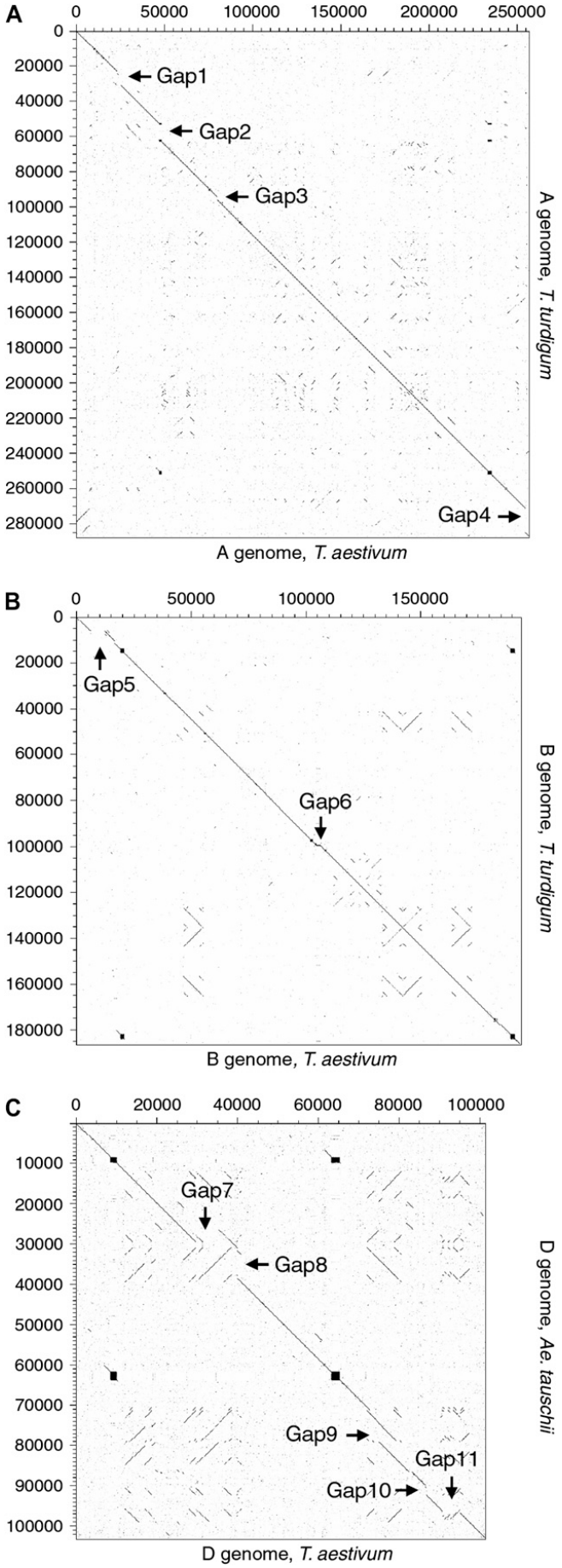
genome share 93% sequence identity. The receptor kinase 1b in both the hexaploid A genome and the tetraploid A genome is disrupted by the same transposable elements with identical insertion patterns (data not shown). Whether or not the duplication of the receptor kinase gene is present in the B genome is not determinable from the current data; the 5′-end of the B-genome BAC is downstream of where the receptor kinase 1a would be putatively located (Figure 1).

In an earlier study, we reported that although gene colinearity was not violated between homeologous wheat genomes, several genes were subjected to differential disruptions by various mechanisms, including repetitive DNA insertion, sequence deletion, and nucleotide substitutions causing an in-frame stop codon (GU *et al.* 2004). Five genes found to be disrupted at the *Glu-1* locus regions in the hexaploid wheat appear to have the same patterns of gene disruption as those identified in their ancestor genomes (Table 1). For example, the second globulin gene in the hexaploid B genome (2b) has the same four retrotransposon insertion events as those identified in the tetraploid B genome (KONG *et al.* 2004). Previously, we reported that the second globulin genes in the diploid D and tetraploid A genomes were disrupted by an identical deletion event, suggesting that this gene disruption occurred before the divergence of the wheat A and D genomes (ANDERSON *et al.* 2003; GU *et al.* 2004). In the hexaploid wheat, the second globulin gene (2b) in both the A and D genomes share this same deletion. The receptor kinase 1b in the hexaploid A genome contains the same multiple insertions of retroelements and miniature inverted repeat sequences (MITES)as the receptor kinase gene in the durum A genome (GU *et al.* 2004). On the basis of this evidence, it

TABLE 1

Comparison of gene disruptions observed in hexaploid wheat and its ancestral genomes

	Receptor kinase 1b (A genome)	*γ-type HMW-glutenin 3a* (A genome)	*Globulin 2b* (A genome)	*Globulin 2b* (B genome)	*Globulin 2b* (D genome)
T. turgidum	Insertions of multiple retrotransposons	Point mutation resulting in a premature stop codon	Deletion of a coding sequence region	Insertions of multiple retrotransposons	
T. aestivum	Insertions of multiple retrotransposons	Point mutation resulting in a premature stop codon	Deletion of a coding sequence region	Insertions of multiple retrotransposons	Deletion of a coding sequence region
Ae. tauschii					Deletion of a coding sequence region



appears that the disruption of these genes in the hexaploid wheat had already occurred in its diploid or tetraploid progenitors.

Genes disrupted in only one of the homologous genomes were also identified. The Ax HMW glutenin gene falls within this category. A point mutation in the tetraploid Ax HMW glutenin gene has resulted in a premature stop codon, while the Ax HMW glutenin gene in the hexaploid wheat is intact.

Sequence comparison of the *Glu-1* regions across the three wheat ploidy levels: To further analyze the sequence variation present in homologous wheat genomes, we performed dot matrix analyses between pairs of corresponding genomes from different ploidy wheats. Sequence divergences are seen as disruptions in the main matrix diagonal line. Figure 2 suggests that the sequences between the homologous *Glu-1* regions are generally conserved. The gaps along the diagonal lines represent types of sequence rearrangements, such as deletions/insertions, duplications, and inversions, that differentiate the two homologous genomes. LTR retrotransposable elements are usually 6–8 kb and insertions of such elements that occur in only one of two compared regions will result in large gaps in the dot matrix analysis. The HMW glutenin region of the A genome of *T. turgidum* contains three nonshared retroelement insertions when compared with the corresponding region from the A genome of *T. aestivum*. In the region between the receptor kinase 1b and globulin 2a, there is a WIS-type element (*Wis-3*) insertion present in the coding region of a gypsy-class element *Boba-1* (Gap1). Gap2 was caused by the insertion of *Wis-4* in the coding region of the Ay HMW glutenin gene in tetraploid wheat. In the region between the Ax HMW glutenin and the serine/threonine protein kinase genes, there is a gypsy-class element, *Erika-2*, that is not present in the corresponding area of our *T. aestivum* sequence (Gap4). These results provide evidence that these elements have inserted since the two genomes last shared a common ancestor. In addition to these retroelement insertions, there are also two significant indel events in the *T. aestivum* sequence. The first of these removed the entire coding region and a fragment of each LTR from the element *Madil-1*, located in the intergenic region between the receptor kinase and globulin genes, leaving behind a sequence similar to a solo LTR in size and structure (Gap3). The second indel,

FIGURE 2.—Pairwise comparisons of the *Glu-1* regions between the two homologous genomes. The dot plot was performed using the DOTTER program (SONNHAMMER and DURBIN 1995) with default parameters between two A-genome sequences from *T. turgidum* and *T. aestivum* (A), two B-genome sequences from *T. turgidum* and *T. aestivum* (B), and two D-genome sequences from *Ae. tauschii* and *T. aestivum* (C). Each distinct gap in the dot plot is assigned a number, and sequence rearrangements causing the gaps are described in the text.

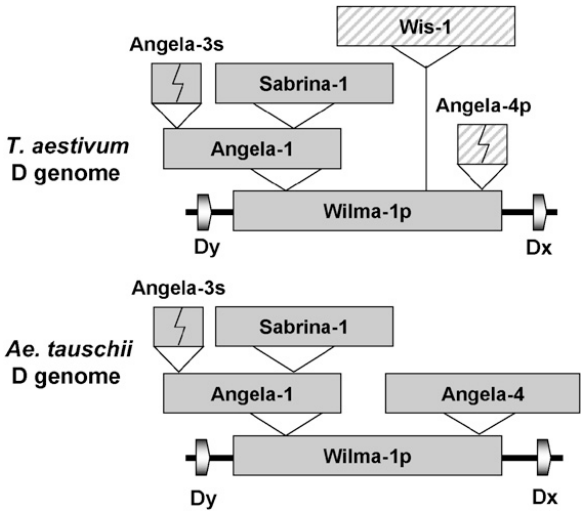


FIGURE 3.—Insertion and deletion of retrotransposons that differentiate two homologous D genomes. Retrotransposons that are colinear between the hexaploid D and diploid D genomes are boxed in gray. Retrotransposons that have inserted or rearranged by deletion in only one of the genomes are represented by hatched boxes. Genes are represented by pentagons. Dy and Dx are short for the y-type and x-type HMW-glutenin genes from the D genome. The letters "s" and "p" following the names of the retroelements indicate solo or partial, respectively.

which occurred in a relatively old CACTA element, *Jorge-1*, removed a portion of the coding sequence and possibly a MITE, *Eos-1*, as well.

By contrast, the HMW glutenin regions of the two B genomes revealed only a few major sequence differences. A 6-kb indel (Gap5) is present in the intergenic region between the receptor protein kinase 1b and globulin 2a coding regions. While the sequence of this 6-kb indel showed no sequence similarity to known transposable element sequences or structures, it is flanked by an 8-bp (TAAGAATT) perfect repeat, suggesting that illegitimate recombination resulted in this indel (WICKER *et al.* 2003; CHANTRET *et al.* 2005). In addition to the 6-kb indel, there is a smaller indel, ~1.7 kb (Gap6), present within the class I transposable element *Sogi-1*. Sequence comparisons revealed no novel retroelement insertions that could serve to differentiate the two B genomes.

The HMW glutenin regions of the two D genomes are the most divergent at the large-scale sequence rearrangement and insertion/deletion level. Since the time when the D-genome donors of sequenced accessions of *Ae. tauschii* and *T. aestivum* shared a common ancestor, three novel retroelement insertions have occurred in this area. In *Ae. tauschii*, two copia class retroelements, *Angela-5* (Gap9) and *Wis-1s* (Gap10), have inserted into a large block of retroelements in the region downstream of the serine/threonine protein kinase. In *T. aestivum*, a complete copy of the element *Wis-1* (Gap8) is present in the nested retroelement structure between the two HMW glutenin genes (Figure 3). In addition to these differential insertions, there is also a deletion of the coding region and portions of both LTRs from *Angela-4* (Gap7), resulting in a partial LTR retroelement, *Angla-4p* in *Ae. tauschii* (Figure 3). A series of sequence rearrangements/insertions in the 5′ LTR sequence of *Angela-7p* in *T. aestivum* generated Gap11 (Figure 2).

Shared and nonshared transposable elements between *Glu-1* homologous regions: Despite indel events shown in Figure 2, the dot matrix analyses revealed general sequence conservation of the *Glu-1* regions between two homologous genomes, suggesting that many transposable elements in the intergenic regions are shared or colinear. To test this, we calculated the number of shared retroelements *vs.* nonshared retroelements in the *Glu-1* regions from pairs of homologous genomes. On the basis of the activity of retroelement insertions, the two B genomes are the most conserved, with all 20 of the retroelements identified present in both genomes (Figure 4A). In the ~300-kb *Glu-1* regions from the two A genomes, we identified 35 retroelements that are present in both genomes. The two nonshared retroelements are both present in the A genome of durum wheat. In the ~100-kb homologous *Glu-1* regions of the two D genomes, 14 retroelements are found to be shared, while 3 are nonshared. Among these 3 new retroelements, there is 1 intact retroelement, one solo LTR, and 1 partial retroelement missing one LTR and a portion of the coding region. The solo LTR and the partial LTR retroelement that differentiate the two D genomes are somewhat surprising since newly inserted retroelements would be expected to be intact (MA *et al.* 2004; BRUNNER *et al.* 2005). However, this may suggest that insertions of new transposable elements can be followed by DNA recombination.

Although nonshared retroelements were observed in the homologous *Glu-1* regions between two A and D genomes, the frequency of nonshared retroelements is much lower than those found in homologous regions from different maize inbred lines, the *Rph7* locus regions from two different barley accessions, the *Lr10* locus from homologous A genomes in different ploidy levels, and the *Ha* locus from homologous A, B, and D genomes in different ploidy levels (BRUNNER *et al.* 2005; CHANTRET *et al.* 2005; ISIDORE *et al.* 2005; SCHERRER *et al.* 2005). In this study, only 5.7% of the retroelements are nonshared in the *Glu-1* regions from the two A genomes, 0% from the two B genomes, and 21% from the two D genomes. The nonshared retroelements account for 18.1% of the sequence divergence between the two D genomes and 8.2% between the two A genomes.

Sequence insertions and deletions in the homologous *Glu-1* regions: Indels play a fundamental role in genome evolution through the loss or gain of DNA in specific regions (GREGORY 2004; TAYLOR *et al.* 2004). There is still little known about the distribution and frequency of indels in plant genomes. In addition to the large indels that cause visible gaps in the dot matrix analysis (Figure 2), we also analyzed smaller indels

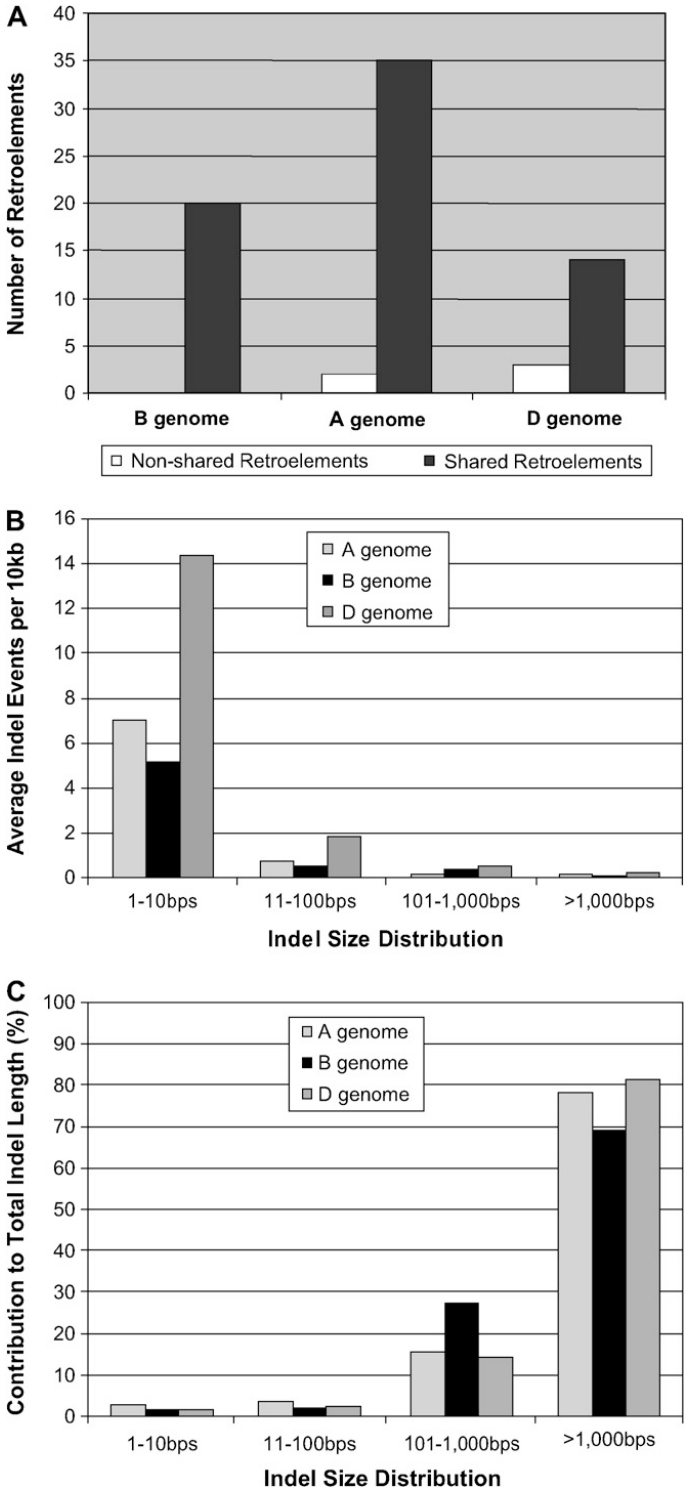


FIGURE 4.—A. Number of shared *vs.* nonshared retroelements between the two homologous wheat genomes. Shared and non-shared retroelements between the homologous genomes are represented by the black and gray bars, respectively. B. Indel frequency and size distribution in three wheat genomes. Indels identified in each pair of homologous genomes were categorized into four groups. The indel frequency for each group was normalized to the number of indels per 10 kb in the graph. C. Percentage of total indel length contributed by indels of different sizes. Indel length represented by different indel sizes were calculated and then divided by the total indel length for each of the A, B and D genomes. The total indel lengths are 10,738, 11,382, and 13,612 bp for the A, B, and D genomes, respectively.

present in the *Glu-1* regions of each pair of the homologous wheat genomes (Figure 4B). Clearly, smaller indels (≤10 bp) greatly outnumber large indels in terms of frequency in all three genomes. They account for 88.2% of the indel frequency for the A genome, 83.3% for the B genome, and 85.0% for the D genome.

When the indel frequencies were compared in the A, B, and D genomes, the frequency of indels differed significantly between any two genomes in the *Glu-1* regions ($P < 0.05$, Student's *t*-test). The D genome had the highest frequency of indels (16.9 indels/10 kb). The indel frequency for the A and B genomes was 8.0 and 6.2, respectively. Indels are a type of sequence rearrangement event and the frequency reflects the evolutionary distance of two related genomes in comparison (OGURTSOV *et al.* 2004). Therefore, our data suggest that the two D genomes are more distantly related as compared to the two A and B genomes.

Indels directly contribute to the sequence divergence between two compared sequences. We calculated the total length of sequences represented by the indels. The total length of the indel sequence is 10,738 bp for the A genome, 11,382 bp for the B genome, and 13,612 bp for the D genome. They account for 4.0, 6.2, and 13.3% of sequence divergence in the compared regions for the A, B, and D genomes, respectively. When the length of indels represented by different indel sizes was calculated, indels >100 bp contributed much more to the total indel length (Figure 4C). They account for 93.5, 96.5, and 95.8% of the total indel length for the A, B, and D genomes, respectively. These results suggest that large indels are primarily responsible for the indel-based sequence divergence between the two homologous wheat genomes.

Nucleotide substitutions in the *Glu-1* region of the homologous wheat genomes: Another type of common sequence variation, occurring in even the most conserved regions, is the single nucleotide polymorphism, caused by nucleotide substitutions (BRUMFIELD *et al.* 2003). The high level of sequence conservation observed in the *Glu-1* regions of the homologous genomes allows us to examine nucleotide substitutions in large contiguous segments spanning both the genic and intergenic regions. We calculated that the average number of nucleotide substitutions *per site* is 0.0093, 0.0056, and 0.0137 for the A, B, and D genomes, respectively. The variations in average nucleotide substitutions are all statistically significant ($P < 0.001$). The data indicate that the nucleotide substitution rates for the A and B genomes are different despite the fact that the two genomes have been co-evolving in the nuclei of the same species.

The rate of nucleotide substitutions in intergenic regions is most often considerably higher than those found in genic regions and protein-coding sequences (MA and BENNETZEN 2004). We also compared the

TABLE 2

K_s and K_a values and standard errors for pairwise comparisons of four genes in the *Glu-1* region

	A genome		B genome		D genome	
	K_s	K_a	K_s	K_a	K_s	K_a
Globulin 2a	0	0.0019 (±0.0019)	0.0059 (±0.0064)	0.0020 (±0.0020)	0.0059 (±0.0058)	0.0020 (±0.0019)
HMW-glutenin 3a	0.0162 (±0.0068)	0.0078 (±0.0026)	0.0055 (±0.0031)	0.0055 (±0.0017)	0.0395 (±0.0101)	0.0108 (±0.0029)
HMW-glutenin 3b	0.0081 (±0.0035)	0.0069 (±0.0019)	0.0142 (±0.0048)	0.0165 (±0.0030)	0.0061 (±0.0031)	0.0050 (±0.0017)
Protein kinase 4	0	0.0014 (±0.0010)	0	0.0012 (±0.0011)	0.0049 (±0.0034)	0.0020 (±0.0012)

number of nucleotide substitutions *per site* between the HMW glutenin genes with those of the intact globulin and protein kinase genes that flank the *Glu-1* locus in all three wheat genomes. The number of synonymous substitutions (K_s) and nonsynonymous substitutions (K_a) *per site* between genes from the homologous genomes is given in Table 2. Considerable variations in nucleotide substitutions are observed among these genes, suggesting that different genes evolve at different rates. The globulin 2a in the two A genomes and protein kinase genes in both the A and B genomes appear to be more conserved. The number of nucleotide substitutions *per site* in the HMW glutenin regions is at least similar to or greater than that found in the neighboring intergenic regions. The high nucleotide substitution rate coincides with the previous observation that HMW glutenin genes are highly polymorphic in wild wheat and display significant variation in modern cultivars (Allaby *et al.* 1999; Shewry *et al.* 2003).

Time of divergence of the homologous genomes: In the *Glu-1* regions, the identification of a number of LTR retrotransposons that are intact and shared by two homologous genomes allows us to further examine the sequence changes and divergence times of the homologous genomes in wheat. Because of the particular mechanism of reverse transposition, the two LTR sequences from a single retrotransposon are identical at the time of its insertion (Boeke and Corces 1989). On the basis of this property, the nucleotide substitutions identified in two LTRs of a retrotransposon can be used to estimate the date of its insertion (SanMiguel *et al.* 1998). Previously, we estimated insertion times for 22 LTR retrotransposons in the wheat *Glu-1* region using a mutation rate of 6.5×10^{-9} substitutions/synonymous site/year (Guat *et al.* 1996; SanMiguel *et al.* 1998). It appears that these datable LTR retrotransposons have all inserted into their current positions within the last 4–5 MY (Gu *et al.* 2004). A similar methodology can be employed to estimate divergence times for two homologous sequences on the basis of the number of nucleotide substitutions accumulated after the ancestor genome started to split into two distinct descendants. However, considerable variation in estimated divergence time has been noted when single gene sequences are used (Huang *et al.* 2002; Devos *et al.* 2005). One possible explanation is that rates of nucleotide substitution are different for genes that are under different selective forces. A better representation of divergence time for two closely related sequences may be obtained by comparing multiple regions, ideally intergenic sequences that have few selective forces acting on them (Wakeley and Hey 1997). In this study, we used the same colinear LTR retrotransposons previously identified and dated in the *Glu-1* regions to estimate the divergence time of the two homologous genomes. A total of 8–9 intact colinear LTR retrotransposons for each B and each A genome pair, representing ~40–50 kb of sequence, were used to estimate the divergence times. For the two homologous D genomes, only two full-length, datable LTR retrotransposons are colinear; we identified additional partial yet colinear retroelements to obtain ~40 kb of total sequence for comparison. Using the same molecular clock for dating the LTR retrotransposon insertions, we estimated the divergence time by calculating rates of nucleotide substitution between each pair of colinear retroelements to examine the variation in different sequences (Table 3). The divergence times estimated, using common retroelements, for the two homologous A genomes range from 0.38 to 1.2 MY, with an average of 0.81 MY. The two B genomes are estimated to have diverged in the last 0.24–0.78 MY, with an average of 0.48 million years ago (MYA). The two D genomes are the most divergent, ranging from 1.12 to 1.86 MYA with an average of 1.39 MYA. The longer divergence time between the two D genomes shows that the diploid D-genome sequence is divergent from the D-genome donor(s) of the hexaploid wheat by ~1.39 MY (Table 3). It has been noted that LTR retroelements evolve at least two times faster than genes and UTR regions (SanMiguel *et al.* 1998; Wicker *et al.* 2003; Ma and Bennetzen 2004). If this is the case, the divergence times estimated for each pair of the homologous genomes should be divided by a factor of 2.

DISCUSSION

In this study, a detailed sequence comparison was performed to study the organization and evolution of the hexaploid wheat genome with its ancestral diploid and tetraploid wheat genomes. On the basis of the

TABLE 3

Estimates of divergence time of the homologous wheat genomes

LTR retrotransposons	Divergence time (MYA)	SD
Shared in A genome		
Pivu-1	0.38	0.11
Gujog-1	0.80	0.35
Nusif-1	0.97	0.29
Apiip-1	1.20	0.27
Sabrina-2	0.60	0.17
Fatimah-1	1.13	0.42
Sabrina-3	0.71	0.19
Ames-1	0.75	0.12
Fatimah-2	0.74	0.35
Average	0.81	0.24
Shared in B genome		
Ifis-1	0.37	0.22
Wis-1	0.55	0.16
Fatimah-1	0.57	0.15
Sabrina-1	0.24	0.10
Derami-1	0.38	0.16
Wis-2	0.37	0.13
Wis-3	0.57	0.16
Wis-4	0.78	0.18
Average	0.48	0.16
Shared in D genome		
Wilma-1p	1.18	0.25
Angela-2	1.12	0.23
Sabrina-2	1.34	0.25
Angela-3s	1.59	0.27
Angela-7p	1.30	0.23
Latitude-1s	1.74	0.20
Wham-1p	1.86	0.32
Sabrina-4p	1.13	0.23
Average	1.39	0.28

Full-length LTR retrotransposons that are colinear in the *Glu-1* regions were used to calculate the divergence time of two homologous genomes. Divergence time was estimated as described in SanMiguel *et al.* (1998).

sequence analysis of *Glu-1* regions, our results reveal that although sequence rearrangements differentiating each pair of two homologous genomes are easily visible, a considerable portion of the sequences are highly conserved between the two homologous genomes, including large segments of intergenic regions composed of nested retroelements. Although retroelements provide the majority of sequence divergence between two homologous genomes, indels and nucleotide substitutions are the most frequent events in the compared *Glu-1* regions.

Sequence conservation in the homologous *Glu-1* region: Our previous study indicated that microcolinearity is maintained in the orthologous *Glu-1* regions from homeologous wheat genomes, but intergenic regions were not conserved due to rapid amplification/deletion of retroelements (Gu *et al.* 2004). This study revealed general sequence conservation between two homologous genomes from the different wheat ploidy levels. In addition to the conservation of the genic regions, sequences in the intergenic regions are also highly colinear (Figure 4A). This suggests that a vast majority of retroelements in the hexaploid wheat genome were inherited from their diploid and tetraploid ancestral genomes. In the *Glu-1* region of the tetraploid A genome, an intergenic region between the y-type and x-type HMW glutenin genes contains large blocks of nested retrotransposon insertions, with as many as 19 members spanning a region of 140 kb (Gu *et al.* 2004). The same nested retroelement structure is also present in the hexaploid A genome, suggesting that the intergenic region has not been drastically changed due to major sequence rearrangements. The presence of a high number of colinear retroelements in the *Glu-1* regions from different ploidy wheat genomes is surprising, considering the results from similar comparative sequence studies on other homologous locus regions. In the genomes of different maize inbred lines, it was found that 70% of LTR retrotransposons are allele specific (Brunner *et al.* 2005). When homologous regions containing the barley *Rph7* locus from resistance and susceptible lines were compared, the number and type of repetitive elements were completely different in 65% of the sequenced regions (Scherrer *et al.* 2005). In the wheat *Lr10* region, transposon insertions resulted in >70% sequence divergence among three A genomes from diploid, tetraploid, and hexaploid wheats (Isidore *et al.* 2005). In the *Glu-1* regions, retroelements resulted in only 8.2% of nonshared sequence between the two A genomes, 18.1% between the two D genomes, and 0% between the two B genomes.

In the *Glu-1* regions, violations of gene colinearity were not detected between two homologous genomes. This is also significantly different from the other genetic loci studied in maize (Brunner *et al.* 2005), barley (Scherrer *et al.* 2005), and wheat (Wicker *et al.* 2003; Chantret *et al.* 2005; Isidore *et al.* 2005). One possible explanation for the sequence conservation observed in the *Glu-1* region could be that the hexaploid line (Renan) selected in this study has the same or a very similar haplotype as the tetraploid wheat *cv.* Langdon. However, our haplotype analyses suggest that the two homologous A and B genomes belong to different haplotypes (see supplemental data at http://www.genetics.org/supplemental). Therefore, our results suggest that the wheat *Glu-1* locus is a conserved genetic region, likely with a low recombination rate since recombination rates are positively correlated with the frequency of locus duplication and deletions that often cause synteny perturbation between wheat homeologous chromosomes (Akhunov *et al.* 2003). Furthermore, the conservation of the *Glu-1* locus regions is further supported by the lack of recombination events observed between x- and y-type HMW glutenin genes, despite the wide

interest in breaking the linkage for wheat quality improvement in breeding (Shewry *et al.* 2003). In addition, our data indicate that at the HMW glutenin loci there is no evidence for polyploidization-induced genomic changes that resulted in significant sequence rearrangements.

Sequence rearrangements by indels in the wheat *Glu-1* regions: Despite the general conservation in the *Glu-1* regions, indels, particularly small indels, are frequently observed between the homologous genomes. In the *Glu-1* regions, small indels (<10 bp, ~85%) greatly outnumber larger indels (>10 bp, ~15%) in all three pairs of homologous genomes. Small indels often occur in coding regions, making them a major player in gene evolution (Taylor *et al.* 2004). In addition, small indel events might be associated with nucleotide substitutions, resulting in an increased rate of single nucleotide polymorphisms (Ma and Bennetzen 2004). Indel occurrences have been used to estimate evolutionary distance between related genomes (Ogurtsov *et al.* 2004). In this study, we also detect a positive correlation between the frequency of indels and the rate of nucleotide substitution. The two homologous D genomes that have the longest divergence times show the highest indel frequency and greatest rate of nucleotide substitution.

In the homologous *Glu-1* regions, although small indels are more ubiquitous, they contribute less to the genome size difference. However, large indels contribute more heavily to overall sequence divergence in closely related genomes (Figure 4C). In addition, it is often difficult to distinguish if an indel was caused by an insertion or a deletion. Deletions generally outnumber insertions (Blumenstiel *et al.* 2002; Gregory 2004; Ma and Bennetzen 2004). In the *Glu-1* regions, most large indels were identified to be deletions since they occur in the sequences of well-characterized retroelements. It is likely that nonintact or repetitive DNA fragments observed in the intergenic regions are caused mainly by deletion events.

In general, small indels are caused by replication slippages (Gregory 2004), whereas large indels seems to involve different mechanisms. One mechanism is the unequal homologous recombination that often acts on LTR transposable elements, resulting in solo LTRs (Devos *et al.* 2002; Ma *et al.* 2004). In contrast, indels caused by illegitimate recombination appear to occur in any region of nucleotide sequence with as little as a few base pairs of sequence identity (Wicker *et al.* 2003; Chantret *et al.* 2005). It has been reported that the deletion events attributable to illegitimate recombination are much more frequent than deletion events caused by unequal homologous recombination (Wicker *et al.* 2003). Illegitimate recombination is primarily responsible for the removal of nonessential DNA in Arabidopsis (Devos *et al.* 2002). In wheat, indels caused by illegitimate DNA recombination are one of the major molecular mechanisms responsible for reshaping the *Ha* locus during wheat genome evolution (Chantret *et al.* 2005). Our results indicate that indels are dynamic evolutionary processes that contribute to sequence divergence between homologous wheat genomes.

Wheat genome evolution: Hexaploid wheat contains three homeologous genomes. The origin and evolution of these evolutionarily closely related genomes has been analyzed phylogenetically using sequences from HMW glutenin genes (Allaby *et al.* 1999; Blatter *et al.* 2004). The results indicated that wheat homeologous genomes diverged ~5.0–6.9 MYA (Allaby *et al.* 1999).

In this study, we compared sequence changes among the homologous wheat genomes. Analysis of the *Glu-1* regions from tetraploid and hexaploid wheat revealed that the A genome exhibits higher sequence variation than the B genome. This is supported by at least four lines of evidence. First, the rate of nucleotide substitution in the A genome is significantly higher than that detected in the B genome ($P < 0.01$). Second, the B genomes are devoid of nonshared retroelements, whereas two nonshared retrotransposons were identified in the A genome (Figure 4A). Third, the indel frequency in the A genome is also significantly higher than that found in the B genome ($P < 0.01$) (Figure 4). Fourth, the divergence times estimated for A and B genomes are significantly different (Table 3). The greater sequence conservation was also reported for the same two B genomes at the *Ha* locus on chromosome 5, where there are no nonshared retroelements and the two sequences have 99% identity (Chantret *et al.* 2005), suggesting that such sequence conservation is present at multiple genetic regions and is unlikely caused by introgression recombination. Although the possibility of introgression recombination cannot be completely excluded, the variation of sequence divergence between the A and B homologous pairs could be explained by a hypothesis that the A and B genomes in the tetraploid and hexaploid wheats evolve at different rates. This explanation would have to be based on the assumption that significant sequence changes have occurred since the tetraploidization event, which was estimated to have occurred in the last 0.36–0.5 MY (Huang *et al.* 2002; Dvorak and Akhunov 2005). Different rates of genome evolution have been noted for *japonica* and *indica* rice, which have experienced independent variation for ~0.44 MY (Ma and Bennetzen 2004). However, in this study, both the A and B genomes are the subgenomes in a single polyploid wheat. As far as we know, no reports have shown that different subgenomes in a single species are subject to different evolutionary rates.

Another likely explanation for the observed difference in sequence variation is that hexaploid wheats have more than one tetraploid ancestor, resulting from independent tetraploidization events, in which two divergent A genomes hybridized with two less divergent B genomes, resulting in the different rates of sequence variation observed in our study. Our estimates of the

divergence times for the two A genomes and the two B genomes, calculated by comparing colinear retroelements, is in accordance with this hypothesis. Our results indicated that the two B genomes have diverged for ~0.48 MY, while the two A genomes have been separated for the last ~0.81 MY. Another piece of supporting evidence includes the finding from another study that the A genome from the hexaploid wheat Renan has a different haplotype than the A genome from durum wheat at the *Lr10* locus and that this haplotype originated from ancient DNA rearrangements at the diploid level (Isidore *et al.* 2005). Furthermore, we previously reported that the sequence in the *Glu-1* locus from the hexaploid wheat *cv.* Chinese Spring is more closely related to that of the A genome from durum wheat, primarily because of shared mechanisms of gene disruption in the two orthologous Ay HMW glutenin genes and because of higher sequence identity compared to allelic HMW glutenin genes in other bread wheats, such as Cheyenne (Gu *et al.* 2004). It is likely that the Chinese Spring A genome has the same lineage as the A genome from the durum wheat and that the A genome in Renan and Cheyenne represents a different A genome lineage. Wheat exists at three ploidy levels, and multiple origins of hexaploid wheat have been suggested (Dvorak *et al.* 1998; Talbert *et al.* 1998; Blatter *et al.* 2004). Such multiple independent polyploidizations would serve to greatly increase the genetic diversity in the wheat gene pool. Our results provide further evidence based on the analysis of sequences of large wheat genomic regions spanning the *Glu-1* locus across three ploidy levels. This study allows us to examine the various events of sequence changes that have occurred in the evolutionary history of wheat genomes. A genomewide analysis using various tools such as haplotype genotyping in different ploidy wheats will promote us to elucidate the occurrence of wheat speciation and the molecular evolution of the wheat genome.

We thank Mingcheng Luo for providing wheat materials for the haplotype analysis, Frank You for assistance in statistical and bioinformatics analysis, and Roger Thilmony for critical reading of the manuscript. We sincerely thank the Genoplante consortium (http://www.genoplate.com) for making available the BAC library from the *T. aestivum* cultivar Renan and the *Glu-1* BAC clones. Sequencing of the BAC clones 1289J4 and 1001P20 covering the *Glu-1* regions from the B genome was supported by the Genoplante consortium. Sequencing of the other BAC clones was supported by U.S. Department of Agriculture-Agriculture Research Service grant CRIS 5325022100-011. This work is also supported in part by National Science Foundation Plant Genome grant no. DBI-0321757.

LITERATURE CITED

Akhunov, E. D., A. R. Akhunov, A. M. Linkiewicz, J. Dubcovsky, D. Hummel *et al.*, 2003 Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. Proc. Natl. Acad. Sci. USA **100:** 10836–10841.

Allaby, R. G., M. Banerjee and T. A. Brown, 1999 Evolution of the high molecular weight glutenin loci of the A, B, D, and G genomes of wheat. Genome **42:** 296–307.

Anderson, O. D., C. Rausch, O. Moullet and E. S. Lagudah, 2003 The wheat D-genome HMW-glutenin locus: BAC sequencing, gene distribution, and retrotransposon cluster. Funct. Integr. Genomics **3:** 56–68.

Bennetzen, J. L., J. Ma and K. M. Devos, 2005 Mechanism of recent genome size variation in flowering plants. Ann. Bot. **95:** 127–132.

Blake, N. K., B. R. Lehfeldt, M. Lavin and L. E. Talbert, 1999 Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat. Genome **42:** 351–360.

Blanc, G., and K. H. Wolfe, 2004 Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell **16:** 1679–1691.

Blatter, R. H. E., S. Jacomet and A. Schlumbaum, 2004 About the origin of European spelt (*Triticum spelta* L): allelic differentiation of the HMW glutenin B1–1 and A1–2 subunit genes. Theor. Appl. Genet. **108:** 360–367.

Blechl, A. E., and O. D. Anderson, 1996 Expression of a novel high-molecular-weight glutenin subunit gene in transgenic wheat. Nat. Biotechnol. **14:** 875–879.

Blumenstiel, J. P., D. L. Hartl and E. R. Lozovsky, 2002 Pattern of insertion and deletion in contrasting chromatin domains. Mol. Biol. Evol. **19:** 2211–2225.

Boeke, J. D., and V. G. Corces, 1989 Transcription and reverse transcription of retrotransposons. Annu. Rev. Microbiol. **43:** 403–434.

Brumfield, R. T., P. Beerli, D. A. Nickerson and S. T. Edwards, 2003 The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol. Evol. **18:** 249–256.

Brunner, S., K. Fengler, M. Morgante, S. Tingey and A. Rafalski, 2005 Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell **17:** 343–360.

Chantret, N., J. Salse, F. Sabot, S. Rahman, A. Bellac *et al.*, 2005 Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploidy wheat species (*Triticum* and *Aegilops*). Plant Cell **17:** 1033–1045.

Comai, L., A. P. Tyagi, K. Winter, R. Holmes-Davis, S. H. Reynolds *et al.*, 2002 Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. Plant Cell **12:** 1551–1568.

Devos, K. M., J. K. Brown and J. L. Bennetzen, 2002 Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. **12:** 1075–1079.

Devos, K. M., J. Beales, Y. Ogihara and A. N. Doust, 2005 Comparative sequence analysis of the phytochrome C gene and its upstream region in allohexaploid wheat reveals new data on the evolution of its three constituent genomes. Plant Mol. Biol. **58:** 625–641.

Dvorak, J., and E. D. Akhunov, 2005 Tempos of gene locus deletion and duplications and their relationship to recombination rate during diploid and polyploidy evolution in the Aegilops-Triticum alliance. Genetics **171:** 323–332.

Dvorak, J., P. Di Terlizzi, H-B. Zhang and P. Resta, 1992 The evolution of polyploid wheats: identification of the A genome donor species. Genome **36:** 21–31.

Dvorak, J., M. C. Luo, Z. L. Yang and H. B. Zhang, 1998 The structure of the *Aegilops tauschii* genepool and the evolution of hexaploid wheat. Theor. Appl. Genet. **97:** 657–670.

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using *Phred II*. Error probabilities. Genome Res. **8:** 186–194.

Feldman, M., F. G. H. Lupton and T. E. Miller, 1995 Wheats, pp. 184–192 in *Evolution of Crops*, Ed. 2, edited by J. Smartt and N. W. Simmonds. Longman Scientific, London.

Gregory, T. R., 2004 Insertion-deletion biases and the evolution of genome size. Gene **423:** 15–34.

Gu, Y. Q., O. D. Anderson, C. Londeore, X. Kong, R. N. Chibbar *et al.*, 2003 Structural organization of the barley D-hordein locus in comparison with its orthologous regions of wheat genomes. Genome **46:** 1084–1097.

Gu, Y. Q., D. Coleman-Derr, X. Kong and O. D. Anderson, 2004 Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four Triticeae genomes. Plant Physiol. **135:** 459–470.

Guat, B. S., B. R. Morton, B. C. McCaig and M. T. Clegg, 1996 Substitution rate comparisons between grasses and palm: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc. Natl. Acad. Sci. USA **93:** 10274–10279.

He, P., B. R. Friebe, B. S. Gill and J. M. Zhou, 2003 Allopolyploidy alters gene expression in the highly stable hexaploid wheat. Plant Mol. Biol. **52:** 401–414.

Huang, S., A. Sirikhachornkit, X. Su, J. Faris, B. Gill *et al.*, 2002 Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. Proc. Natl. Acad. Sci. USA **99:** 8133–8138.

Isidore, E., B. Scherrer, B. Chalhoub, C. Feuillet and B. Keller, 2005 Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels. Genome Res. **15:** 526–536.

Kashkush, K., M. Feldman and A. A. Levy, 2003 Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat. Genet. **33:** 102–106.

Kong, X., Y. Q. Gu, F. M. You, J. Dubcovsky and O. D. Anderson, 2004 Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat. Plant Mol. Biol. **54:** 56–69.

Kumar, S., K. Tamura and M. Nei, 2004 MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinformatics **5:** 150–163.

Ma, L., and J. L. Bennetzen, 2004 Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl. Acad. Sci. USA **101:** 12404–12410.

Ma, J., K. M. Devos and J. L. Bennetzen, 2004 Analyses of LTR-retrotranposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. **14:** 860–869.

Masterson, J., 1994 Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science **264:** 421–423.

Ogurtsov, A. Y., S. Sunyaev and A. S. Kondrashov, 2004 Indel-based evolutionary distance and mouse-human divergence. Genome Res. **14:** 1610–1616.

Ozkan, H., A. A. Levy and M. Feldman, 2001 Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. Plant Cell **13:** 1735–1747.

Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl and K. L. Shaw, 2000 Evidence for DNA loss as a determinant of genome size. Science **287:** 1060–1062.

Rooke, L., F. Bekes, R. Fido, F. Barro, P. Gras *et al.*, 1999 Overexpresssion of a gluten protein in transgenic wheat results in greatly increased dough strength. J. Cereal Sci. **30:** 115–120.

SanMiguel, P., A. Tiknonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768.

SanMiguel, P., B. S. Gau, A. Tikhonov, Y. Nakajima and J. L. Bennetzen, 1998 The paleontology of intergene retrotransposons of maize. Nat. Genet. **20:** 43–45.

SanMiguel, P. J., W. Ramakrishna, J. L. Bennetzen, C. S. Busso and J. Dubcovsky, 2002 Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). Funct. Integr. Genomics **2:** 70–80.

Scherrer, B., E. Isidore, P. Klein, J. S. Kim, A. Bellac *et al.*, 2005 Large intraspecific haplotype variability at the *Rph7* locus results from rapid and recent divergence in the barley genome. Plant Cell **17:** 361–374.

Shake, H., K. Kashkush, H. Ozkan, M. Feldman and A. A. Levy, 2001 Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. Plant Cell **13:** 1749–1759.

Shewry, P. R., and A. S. Tatham, 1990 The prolamin storage proteins of cereals: structure and evolution. Biochem. J. **267:** 1–12.

Shewry, P. R., N. G. Halford and A. S. Tatham, 2003 The high molecular weight subunits of wheat glutenin and their role in determining wheat processing properties. Adv. Food Nutri. Res. **45:** 221–303.

Soderlund, C., S. Humphry, A. Dunham and L. French, 2000 Contigs built with fingerprints, markers, and PC v4.7. Genome Res. **10:** 1817–1825.

Soltis, D. E., and P. S. Soltis, 1999 Polyploidy: recurrent formation and genome evolution. Trends Ecol. Evol. **14:** 348–352.

Sonnhammer, E. L., and R. Durbin, 1995 A dot matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene **167:** GC1–10.

Talbert, L. E., L. Y. Smith and N. K. Blake, 1998 More than one origin of hexaploid wheat is indicated by sequence comparison of low-copy DNA. Genome **41:** 402–407.

Taylor, M. S., C. P. Ponting and R. R. Copley, 2004 Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. Genome Res. **14:** 555–566.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

Wendel, J. F., 2000 Genome evolution in polyploids. Plant Mol. Biol. **42:** 225–249.

Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z. D. Liu *et al.*, 2003 Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. Plant Cell **15:** 1187–1197.

Communicating editor: S. R. Wessler

Annexe 4 : Article 4

BMC Genomics

BioMed Central

Research article

Open Access

New insights into the origin of the B genome of hexaploid wheat: Evolutionary relationships at the *SPA* genomic region with the S genome of the diploid relative *Aegilops speltoides*

Jérome Salse[1,2], Véronique Chagué[1], Stéphanie Bolot[2], Ghislaine Magdelenat[3], Cécile Huneau[1], Caroline Pont[2], Harry Belcram[1], Arnaud Couloux[3], Soazic Gardais[1], Aurélie Evrard[1], Béatrice Segurens[3], Mathieu Charles[1], Catherine Ravel[2], Sylvie Samain[3], Gilles Charmet[2], Nathalie Boudet[1] and Boulos Chalhoub*[1]

Address: [1]UMR INRA 1165 – CNRS 8114 UEVE – Unité de Recherche en Génomique Végétale (URGV), 2, rue Gaston Crémieux, CP5708, 91057 Evry cedex, France, [2]UMR 1095 INRA – Université Blaise Pascal – Génétique Diversité Ecophysiologie de Céréales (GDEC), Domaine de Crouelle, 234, avenue du Brézet, F-63100, Clermont-Ferrand, France and [3]CEA: Institut de génomique – GENOSCOPE, 2, rue Gaston Crémieux, CP 5706, 91057, EVRY Cedex, France

Email: Jérome Salse - jsalse@clermont.inra.fr; Véronique Chagué - chague@evry.inra.fr; Stéphanie Bolot - sbolot@clermont.inra.fr; Ghislaine Magdelenat - gmagdele@geosope.cns.fr; Cécile Huneau - huneau@evry.inra.fr; Caroline Pont - cpont@clermont.inra.fr; Harry Belcram - belcram@evry.inra.fr; Arnaud Couloux - acouloux@genoscope.cns.fr; Soazic Gardais - soazicgardais@hotmail.com; Aurélie Evrard - aurelie.evrard@acpfg.com.au; Béatrice Segurens - segurens@genoscope.cns.fr; Mathieu Charles - charles@evry.inra.fr; Catherine Ravel - Catherine.Ravel@clermont.inra.fr; Sylvie Samain - samain@genoscope.cns.fr; Gilles Charmet - gilles.charmet@clermont.inra.fr; Nathalie Boudet - boudet@evry.inra.fr; Boulos Chalhoub* - chalhoub@evry.inra.fr

* Corresponding author

Published: 25 November 2008

BMC Genomics 2008, **9**:555 doi:10.1186/1471-2164-9-555

Received: 16 June 2008
Accepted: 25 November 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/555

Abstract

Background: Several studies suggested that the diploid ancestor of the B genome of tetraploid and hexaploid wheat species belongs to the *Sitopsis* section, having *Aegilops speltoides* (SS, 2n = 14) as the closest identified relative. However molecular relationships based on genomic sequence comparison, including both coding and non-coding DNA, have never been investigated. In an attempt to clarify these relationships, we compared, in this study, sequences of the Storage Protein Activator (SPA) locus region of the S genome of *Ae. speltoides* (2n = 14) to that of the A, B and D genomes co-resident in the hexaploid wheat species (*Triticum aestivum, AABBDD*, 2n = 42).

Results: Four BAC clones, spanning the SPA locus of respectively the A, B, D and S genomes, were isolated and sequenced. Orthologous genomic regions were identified as delimited by shared non-transposable elements and non-coding sequences surrounding the SPA gene and correspond to 35 268, 22 739, 43 397 and 53 919 bp for the A, B, D and S genomes, respectively. Sequence length discrepancies within and outside the SPA orthologous regions are the result of non-shared transposable elements (TE) insertions, all of which inserted after the progenitors of the four genomes divergence.

Conclusion: On the basis of conserved sequence length as well as identity of the shared non-TE regions and the SPA coding sequence, *Ae speltoides* appears to be more evolutionary related to the B genome of *T. aestivum* than the A and D genomes. However, the differential insertions of TEs, none of which are conserved between the two genomes led to the conclusion that the S genome of *Ae. speltoides* has diverged very early from the progenitor of the B genome which remains to be identified.

Background

All cereal crop species are members of the grass (*Poaceae*) family that is the fourth largest family of flowering plants. With about 10 000 species growing under nearly all climates and latitudes, grasses exceed all other plant families in ecological dominance and economic importance. In terms of genome organisation they represent a very diverse family with basic chromosome numbers ranging from 4 to 50 and genome sizes ranging from 350 Mb to 17 Gb [1]. Fossil data and phylogenetic studies have estimated that the grasses have diverged from a common ancestor 50 to 70 million years ago (MYA) [2,3]. Archaeological records suggest that farming started concomitantly in at least three widely separated regions between 10 000-5 000 years ago during the late Neolithic period. The three most important cereals were independently domesticated in three centres: wheat in south western Asia in the 'Fertile Crescent' region, maize in Mexico and rice in both south east Asia and west Africa [4-6].

Within the *Poaceae*, the genera *Aegilops* and *Triticum* include several diploid species (2n = 14) that, via allopolyploidization, produced several tetraploid and hexaploid wheat species, most of which have been domesticated [7-9]. *T. turgidum* (2n = 28, AABB) was derived from a hybridization event that happened (< 0.5 MYA) between *T. urartu*, (2n = 14, AA), the diploid donor of the A genome (here after gA), and another unknown species of the *Sitopsis* section, donor of the B genome (here after gB), for which the closest known relative is *Ae. speltoides* [7,9,10]. The hexaploid wheat (*T. aestivum*, 2n = 42, AABBDD) originated from an additional polyploidization event between the early-domesticated tetraploid *T. turgidum ssp dicoccum* and the diploid donor of the D genome (here after gD), *Ae. tauschii* (2n = 14, DD), 7 000 to 12 000 years ago (for review [11]). Several wheat phylogeny studies have tried to identify the progenitor of the B genome of polyploid wheat based on cytology [12], nuclear and mitochondrial DNA sequences [13-15] as well as chromosome rearrangement studies (*i.e.* common translocation events) [16-24]. It remains controversial from those studies whether the progenitor of the B genome is a unique *Aegilops* species (*i.e.* monophyletic) or whether this genome resulted from an introgression of several parental *Aegilops* species (*i.e.* polyphyletic origin). More recent and representative molecular comparisons using germplasm collections have shown that the B genome could be related to several *Ae. speltoides* lines but not to other species of the *Sitopsis* section [25,26].

Transposable elements (TEs) have been shown since the seventies to be well represented in the wheat genome, ~80% [27,28]. Comparative studies have shown that beside the general conservation in coding sequences, no TE insertions are conserved between the A, B and D genomes of wheat whereas important proportion of TE insertions are shared between the A or D genomes of polyploid wheat and their respective progenitors *T. urartu* and *Ae. tauschii* [29-33]. No such studies have been yet reported comparing the B genome of these polyploid wheat species to that of its closest known diploid relative, *i.e. Ae. speltoides*. In the present study, we compared for the first time coding and non-coding sequences as well as dynamics of TE insertions between the S genome of *Ae. speltoïdes* and that of the A, B and D co-resident in the hexaploid wheat (*T. Aestivum*). The SPA (for Storage Protein Activator [34]) locus region, belonging to BZIP (Basic Leucine Zipper), located on chromosome 1BL [35], has been chosen because of its importance as trans-acting elements of seed storage protein and its conservation in several other cereals such as maize (Opaque 2 [36-38]), rice (RISBZ1-5 [39]), and barley (BLZ1-2 [40,41]). Updating phylogeny relationships and insights onto the origin of the B genome are discussed.

Results

Organization of SPA locus region in the A, B, D and S genomes

Three BAC clones spanning the SPA gene of the A, B and D genomes of *T. aestivum* were screened from cv Renan BAC library with PCR markers specific for each of the three SPA genes [42]. Sequencing resulted in 113 460, 94 732 and 120 879 bp for, respectively, the A, B and D genomes. Screening of an *Ae. speltoides* pooled BAC library with the same SPA-specific PCR markers allowed us to identify and sequence a BAC clone of 80 493 bp sequence spanning the SPA locus gene. Annotation has been performed to identify and compare gene and repeat contents of the four available genome sequences, graphically presented in Figure 1A. More details are also presented in Additional File 1. As expected for wheat, the four genomic sequences are very rich in TEs.

Overall, the 113 460 bp A genome sequence is structured as 56 830 bp (50.1% of the sequence) of class I TE, 3 934 bp (3.5% of the sequence) of class II elements and 4.9% of unclassified TE. Fourteen class I TEs are identified as one incompletely sequenced (at the BAC sequence extremity), five truncated (with a 5' or 3' truncated region due to nested TE insertion), 4 relics (only visible through alignment remnants), one fragmented (inserted by other TEs, *i.e.* nested insertion) and three complete elements. The class II TEs is represented as a complete CACTA element (CACTA_1_comp, *cf* Additional File 2) and three MITEs (Miniature Inverted-repeat Transposable Element). Besides the identification of TEs a pseudo tubulin gene separated by 55 614 bp from the SPA gene was also identified, both genes covering 4.7% of the sequence.

The 94 732 bp B genome sequence is structured as 38 126 bp (40.2% of the sequence) of class I TEs, 22 602 bp (23.9% of the sequence) of class II elements and 0.6% of

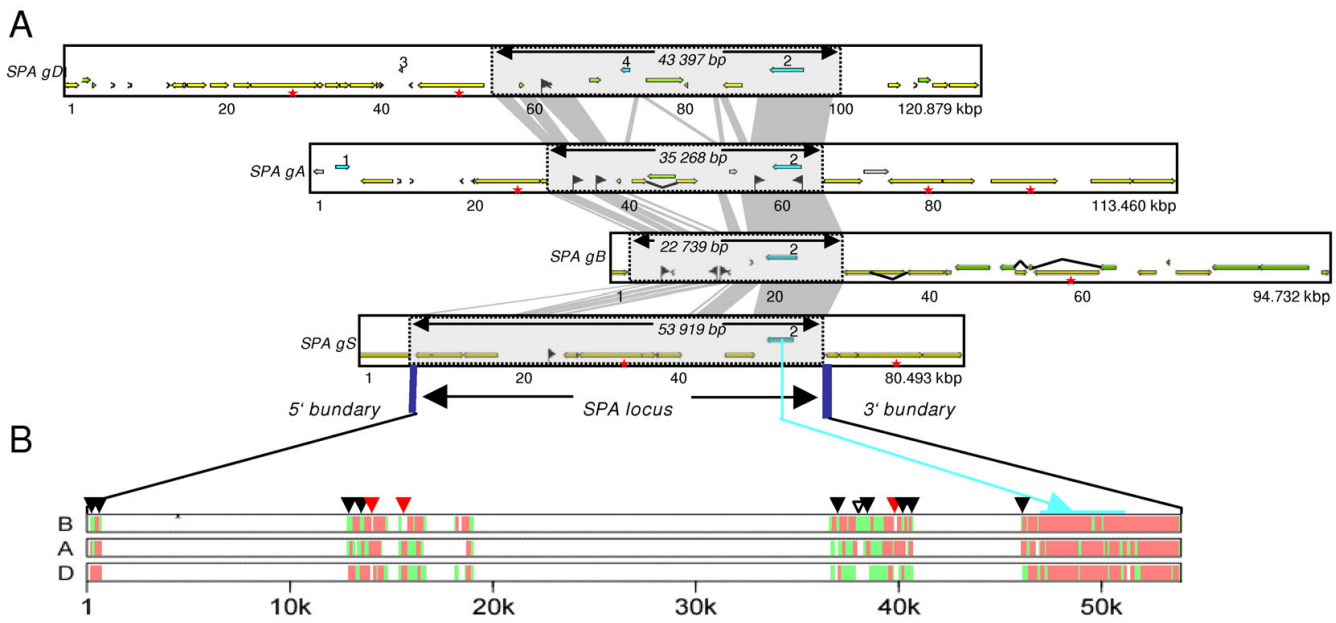


Figure 1

Identification of the 'SPA orthologous region' and comparative annotation of the homoeologous A, B, D and S sequences. (A) Scaled diagram of annotation results of the SPA locus region in which (CDS) (light blue), class I TEs (yellow blocks), class II TEs (green blocks), unclassified elements (grey), MITEs (vertical black flags) are shown. The remaining white spaces correspond to unassigned DNA (no features of annotation). Grey blocks represent sequence conservation between the different genomes defining the 'SPA orthologous region'. Genes are numbered as follow: 1: Pseudo tubulin gene; 2: SPA; 3: Putative cortical cell-delineating gene; 4: Putative kinesin gene. Eight class I TE displaying complete LTR and TSD suitable for the estimation of the insertion dates are highlighted with red stars. (B) Multipipmaker alignment using the sequence of the SPA orthologous region of the S genome of *Ae. speltoides* as a matrix compared with the 3 other sequences available, *i.e. T. aestivum* gB (top), gA (center), gD (bottom). Coloured blocks show the percentage of sequence identity (> 90 in red; between 50 to 90% in green). The SPA gene is indicated as a blue box.

unclassified elements. Twelve Class I elements are identified as two incompletes, six truncated, two relics, one fragmented and one complete element. The class II TEs consists of two complete, one fragmented and one truncated CACTA (CACTA_1 to _4, *cf* Additional File 2) as well as three MITEs. The SPA gene is the only gene identified on the B genome sequence, representing 4.4% of the sequence.

The 120 879 bp D genome sequence is structured as 50 540 bp (41.8% of the sequence) of class I TEs, 9 446 bp (7.8% of the sequence) of class II elements. Twenty-two class I TEs are identified as two incomplete, eight truncated, eight relics, two fragmented and two complete elements. Class II TEs are represented as three truncated CACTA elements (CACTA_1 to 3, *cf* Additional File 2), one mutator relic and one MITE. Three genes have been annotated on the D genome sequence, the SPA gene, a putative kinesin and a putative cortical cell-delineating gene, covering 5.2% of a 48 440 bp interval.

The 80 493 bp S genome sequence is structured as 54 965 bp (68.3% of the sequence) of class I TEs, and a single MITE class II TE. Thirteen class I TEs are identified as one incomplete, six truncated, four fragmented and two complete TEs (*cf* Additional File 2). As in the B genome sequence, only the SPA gene, covering 4.3% of the annotated sequence, has been identified on the S genome sequence.

Identification and characterization of conserved sequences

Alignment of the four genomic regions allows the identification of the 'SPA orthologous region', which we have defined as the shared common regions delimitated by conserved non-coding sequence (CNS) stretches (5' and 3' locus boundaries) that do not correspond to TEs. The 'SPA orthologous region' spans respectively 35 268 bp, 22 739 bp, 43 397 bp and 53 919 bp for the A, B, D and the S genomes (*cf* grey boxes in the Figure 1A, 1B).

Dot plot analysis performed between *Ae. speltoides* gS (horizontal) and the *T. aestivum* gA, -gB, -gD genome (ver-

tical) sequences, allows the identification of four conserved sequence stretches, highlighted by blue dotted circles in the Figure 2. The majority of the remaining DNA within the 'SPA orthologous region' (as well as outside the flanking boundaries) is composed of class I and class II TEs that were differentially inserted and/or deleted in each of the four genomes (*i.e.* shown by diagonal breaks on the dot plot in the Figure 2). The cumulative length of the conserved sequence stretches, within the 'SPA orthologous region' of the four genomes are approximately similar between the genomes pairs gA/gB (15 118 bp), gA/gD (14 677 bp), gA/gS (14 504 bp), gB/gD (14 628 bp), gB/gS(15 877 bp), gD/gS (13 985 bp). These could be considered as the *Aegilops-Triticum* 'ancestral SPA Locus' covering 16 598 bp of cumulative length considering sequences stretches conserved between at least two of the compared sequence. Other stretches of sequence conservation were observed outside the 'SPA orthologous region' when comparing pairs of genomes but these sequences were not determined in the available BAC clone sequences of the other genomes (*data not shown*). As we cannot rule out whether these sequences were not covered in the sequenced BAC clones or were not really conserved across the four genomes, they were not considered in the evolutionary relationship analysis.

No genes, other than SPA can be predicted from these four conserved sequence stretches. As coding and non-coding sequences can evolve at different rates, we perform evolutionary analysis separately for the SPA CDS (CoDing Sequence) and the remaining conserved non-coding sequences (CNS).

Conserved non-coding sequences (CNS) analysis

The conserved non-coding sequences consist of the four shared sequence stretches, excluding the SPA gene itself (from methionine start to the stop codon). The gB/gS genome comparison shows the highest sequence identity and cumulative length (89.9% over 11 976 bp) compared to the other sequence comparisons, *i.e.* gA/gB (85.9% over 11 152 bp), gA/gD (87.9% over 10 838 bp), gA/gS (86.8% over 10 597 bp), gB/gD (85.8% over 10 666 bp), and gD/gS (85.3% over 10 039 bp) (*cf* Table 1). Nevertheless, only a 824 bp sequence was shown to be conserved between gS/gB (within the 11 976 bp of aligned sequence) and absent from other genomes (highlighted with white arrows in the Figure 1B). On the contrary, three sequence stretches (respectively 168, 340 and 218 bp) are conserved between the S, the A and/or D genomes and absent from the B genome (*cf* Figure 1B, red arrows). Moreover, although it represents the majority of the CNS comparisons, sequence conservation was not always the highest between the S and B genomes across the CNS as 9 small stretches (representing a total of 726 bp) of sequences were more conserved between the S and the A and/or D

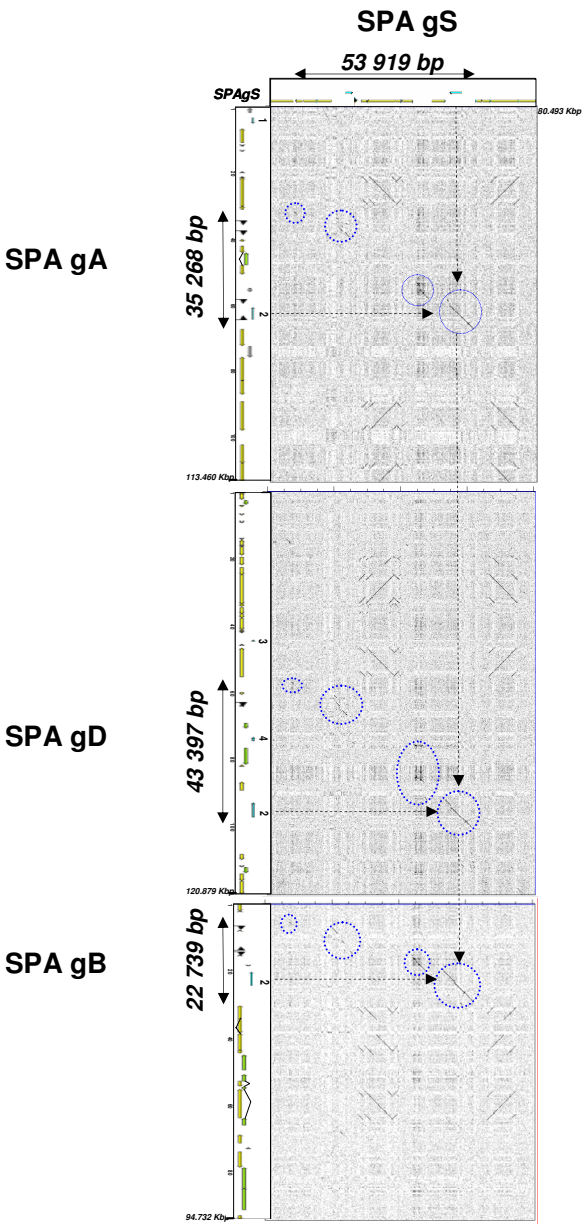


Figure 2
Comparison of the Ae. speltoides sequence with the A, B D genome sequence of T. aestivum. The dot plot was performed using the DOTTER program with default parameters between *Ae. speltoides* gS (horizontal) and *the T. aestivum* gA, -gB, -gD genome (vertical) sequences. Annotation features identified for these sequences are reported on the corresponding axes. Gene numbers and names as well as color codes for TEs and other DNA sequence classes are as in figure 1. Diagonals on the dot plot output that represent nucleotide conservation between the two analyzed sequences are highlighted with dotted blue circles. The loss of micro-colinearity corresponds to diagonal breaks. 'SPA orthologous region' defined as conserved sequences between *Ae. speltoides* gS and *T. aestivum* -gA, -gB, -gD sequences are mentioned with plain arrows on the four annotation features. SPA gene is shown with dotted arrows on the dot plot out put.

Table 1: Conserved Coding (SPA gene) and Non-coding Sequences (CNS) identified between SPA-gA-gB-gD-gS at the 'SPA orthologous region'

Non coding 'SPA orthologous loci' sequences				Coding 'SPA orthologous loci' sequences			
	B	**D**	**S**		**B**	**D**	**S**
A				**A**			
CNS size (bp)	11 152	10 838	10 597	**Nb of transitions**	33	20	38
% Identity	85,9	87,9	86,8	**Nb of transversions**	22	10	23
Ks	0,874+- 0,036	1,037+-0,036	0,716+-0,024	**Ratio**	1,5	2	1,65
Ka	0,664+- 0,014	0,848+-0,015	0,57+-0,01	**Ks**	0,055+-0,015	0,042+-0,013	0,065+-0,016
Ks/Ka	1,3	1,2	1,3	**Ka**	0,042+-0,007	0,021+-0,005	0,049+-0,007
B				**MYA**	6,2–10,8	4,5–8,5	7,5–12,5
CNS size (bp)		10 666	11 976	**B**			
% Identity		85,8	89,9	**Nb de transitions**		35	25
Ks		0,991+-0,034	0,617+-0,026	**Nb de transversions**		20	19
Ka		0,797+-0,015	0,492+-0,012	**Ratio**		1,75	1,32
Ks/Ka		1,2	1,3	**Ks**		0,071+-0,017	0,035+-0,012
D				**Ka**		0,039+-0,007	0,035+-0,006
CNS size (bp)			10 039	**MYA**		8,3–13,5	3,5–7,2
% Identity			85,3	**D**			
Ks			0,902+-0,031	**Nb de transitions**			39
Ka			0,791+-0,014	**Nb de transversions**			23
Ks/Ka			1,1	**Ratio**			1,7
				Ks			0,089+-0,019
				Ka			0,043+-0,007
				MYA			10,8–16,6

Non-coding 'SPA orthologous loci' sequences. Detailed features obtained for the 6 pairwised alignments of the 4 SPA orthologous regions excluding the SPA gene itself are mentioned with the CNS length, percentage of sequence identity, Ks and Ka values and rate. As an example, alignment of the gA and gB SPA orthologous regions of respectively 31 498 bp and 18 589 bp correspond to a cumulative CNS length of 11 152 bp with 85.9% of sequence identity and 0.874, 0.664, 1.3 values for respectively Ks, Ka and Ks/Ka.

Coding 'SPA orthologous loci' sequences. 6 pairwised comparisons of the SPA gene between A/B B/D D/A A/S B/S D/S sequences are associated with the number of substitutions, the number of transition, the number of transversion, the transition/transversion ratio, the Ks value, the Ka value and speciation date (MYA).

genomes than with the B genome (Figure 1B, black arrows).

We also estimated divergence times on the basis of the number of base substitutions (Ks) accumulated after the split-time from the ancestor genome. Ks values were obtained for the 6 pairwise alignment combinations (Table 1). The lowest and highest Ks values correspond respectively to the gB/gS (0.617, *i.e.* identifying the closest related sequences), and gB/gD (1.037, *i.e.* the more divergent sequences).

Conserved coding sequences analysis: SPA gene structure and evolution

SPA genes are structured as six exons (*cf* Additional File 2). In wheat, SPA gene (and CDS) are respectively 3 426(1 218) bp, 3 486(230) bp, 3 796(1 212) bp, 4 080(1 233) bp, long for A, B, D and S genes (hereafter designated SPAgA, -gB, -gD and -gS genes). These SPA genes are composed of six exons ranging in size from 76 (SPAgA, -gB, -gD, -gS exon 4) to 432 bp (SPAgA-gB-gS exon 1) and five introns ranging in size from 92 (SPAgA, -gB, -gD intron 4) to 1297 bp (SPAgD intron 5). All of exon-intron junction sites obey the GT/AG rule as identified in other eukaryotic genes. The relative organization of the exons and introns is the same for the others SPA-like bZIP protein genes characterized to date in cereal, *i.e.* the number of exons and introns is conserved and individual introns occur at relatively the same sites for the maize *O2* [36-43], sorghum *O2* [44], and barley *Blz1* genes [40]. It is interesting to note that the first and fifth introns of the homoeologous SPA genes are respectively much shorter and larger, compare to the other cereal SPA-like bZIP protein genes (*cf* Additional File 2).

We conducted a phylogenic analysis based on SPA CDS of the four wheat genomes as well as that available from other cereals. A graphical representation of these data is shown in the Figure 3 with a classical phylogenic tree including SPA homologs available from other cereals (*cf* parameters in material and method) and illustrates that wheat SPA and barley BLZ2 consists in the same *Triticeae* subfamily in which *Ae. speltoides* and *T. aestivum*-gB SPA sequences are linked on the same branch. Such phylogenic analysis shows that the lowest synonymous (Ks) and non-synonymous (Ka) substitution rates were obtained between *Ae. speltoides* and *T. aestivum* -gB, with Ks (0.035+/-0.012) and Ka (0.035+/-0.006) values corresponding to a 3.5 to 7.2 MYA divergence time, while rates obtained when *Ae. speltoides* is compared to -gA and -gD are respectively Ks (0.065+/-0.016) and Ka (0.049+/-0.007) values corresponding to 7.5 to 12.5 MYA divergence time; and Ks (0.089+/-0.019) and Ka (0.043+/-0.007) values corresponding to 10.8 to 16.6 MYA divergence time (*cf* Table 1). This result strongly suggests that,

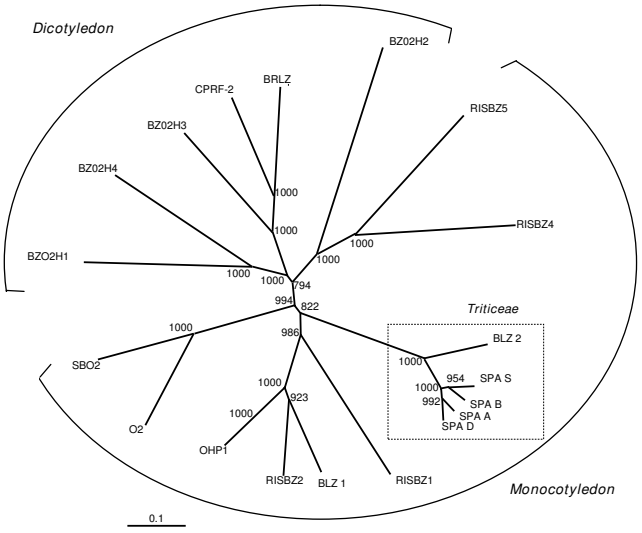


Figure 3
Phylogenic analysis of the SPA protein among plant species. 4 rice (RISBZ1-2-4-5 respectively AB053475, AB021736, AB053473, AB053474), 2 barley (BLZ1-2 respectively BLZ2, Y10834), 2 maize (O2-OHP1 respectively AJ491297, L00623), 1 sorghum (SBO2, X71636), 4 *Arabidopsis thaliana* (BZO2H1-4 respectively NM178959, NM122389, NM122760, NM115319), 1 *Nicotiana tabacum* (BRLZ, AY061648), 1 *Petroselinum crispum* (CPRF-2, X58577) and 4 wheat (SPA-gA, -gB, -gD, -gS, present analysis) sequences are involved in the tree. Parameters used to construct the tree are mentioned in the material and method section.

despite the strong nucleotide conservation between the 3 homoeologous copies of the SPA CDS in *T. aestivum*, *Ae. speltoides* CDS is closest to the *T. aestivum* SPA-gB than the two other homoeologous -gA and -gD sequences.

As reported by Guillaumie et al. [35], a stop codon TGA (+19 bp from the ATG transcription initiation) site had been identified in the SPA-gB sequence suggesting that it might be no more functional. No proof of expression could be also provided for the SPA gB haplotype presenting this stop codon as we were unable to find any corresponding ESTs. In order to clarify the apparition of the TGA stop codon in the B genome, the stop codon allele distribution was analyzed using 18 wheat genotypes which cover, 1 diploid genome S (*Ae. longissima*), 11 tetraploid (3 *T. turgidum* durum, 3 *T. turgidum* dicoccoïdes, 2 *T. turgidum* dicoccum, 2 *T. timophevii*, 1 *T. turgidum turgidum*) and 6 hexaploid (*T. aestivum cv* soisson, arminda, vilmorin, chinese spring, renan, recital) genotypes. Genotyping data demonstrate that the TGA allele is present at 50% in hexaploid wheat (*T. cv soisson*, *vilmorin*, *renan*) and for the first time in one tetraploid (*T. turgidum durum*) genotype over 11 tested and absent in *Ae. longissima* (*cf* Additional file 3).

Differential transposable elements insertions and evolution

Size discrepancies of the 'SPA orthologous regions' can be attributed to differential TE insertions or eliminations (*cf* Additional File 2 and Figures 1A and 2), which occurred after the four genomes divergence. Hence, the size increase observed for the 'SPA orthologous region' in *Ae. speltoides* (35 268 bp) when compared to *T. aestivum*-gB (22 739 bp) is due to 7 class I elements, *i.e*. 2 truncated Angela solo-LTRs (soloLTR_Angela_1 and _3), one complete Angela (Angela_2), one truncated Rada (Rada_1), 2 fragmented LINEs (LINE_1 and _2) and one MITE (*cf* Figure 2 and Additional File 2). These TEs may correspond to insertions, which occurred in the *Ae. speltoides* genome after its divergence from the ancestor of the B genome as they are dispersed between CNS stretches and not present in the B genome of *T. aestivum*. Occurrence of eight class I TEs displaying complete LTR and TSD (Target Site Duplication), identified in the four annotated genomes (highlighted with red stars in the Figure 1A) allows to estimate the insertion dates, based on nucleotide substitution pattern analysis (*cf* material and method; Additional File 4). Thus, the complete Angela_2 identified in *Ae. speltoides* (gS) located in the 'SPA orthologous region' exhibits a transition and tranversion value of 0.02 +/- 0.004 respectively associated with an estimated insertion time of 1.3 to 1.9 MYA. The youngest insertion time was observed for the Angela_5 element annotated outside the 'SPA orthologous region' in the *Ae. speltoides* sequence, *i.e*. 0.6 to 1.1 MYA.

Discussion

We sequenced for the first time an *Ae. speltoides* genomic region (SPA locus region) and compared it to orthologous regions of the A, B and D genomes coresident in the hexaploid wheat *T. aestivum* at the SPA CDS, the CNS and the TE insertion dynamics levels.

SPA gene structure comparison and haplotype variability

The SPA gene is the only gene conserved across the four genomes. A phylogenic analysis involving SPA protein sequences from *T. aestivum*, *Ae. speltoides*, rice, barley, maize, sorghum, *Arabidopsis thaliana*, *Nicotiana tabacum*, *Petroselinum crispum*, clearly identified a *Triticeae* outgroup in which *Ae. speltoides* SPA sequence is more closely related to *T. aestivum*-gB SPA than any other sequence involved in the tree. Interestingly, in this study we showed that the stop codon TGA allele, 19 bases downstream the ATG transcription initiation site, previously identified in the B genome of hexaploid wheat [42], is also present in the tetraploid *T. turgidum*. This indicates that the stop-codon TGA SPA allele has been generated before the allo-hexaploidization event. The presence of both stop TGA and TCA SPA alleles in tetraploid and hexaploid wheat accessions provides further evidences for the hypothesis of (i) recurrent hexapolyploidization events or (ii) gene flow through introgression between the different wheat species with different ploidy levels [30-33].

Differential pattern of CNS conservation

Our results reveal that, a large proportion of the remaining non-genes and non-transposable elements sequences are highly conserved between the four genomes (CNS). At the 'SPA orthologous region', excluding the SPA gene itself, the gB/gS genome comparison shows the highest sequence identity and cumulative length as well as the lowest Ks value (89.9% over 11 976 bp with Ks = 0.617) compared to the other sequences (*cf* Table 1). Thus, the S genome was confirmed to be the closest to the B genome in term of cumulative conserved sequence length as well as identity as compared to any other pairwise genome combinations. Small stretches of sequences, which were more conserved between the S and/or the A and D genomes (*cf* Figure 1B), do not contradict with the general pattern of an overall higher CNS conservation between the S and B genomes. This is the first time that we precisely report close relationships between the S and B genomes based on both coding and non-coding sequence comparisons. CNS (within introns or upstream regulatory sequences), have been recently surveyed in cereals (maize *vs* rice) and mammals (human *vs* mouse) [45,46]. It has been shown that CNSs are more abundant in loci embedding regulatory genes such as transcription factors (as SPA gene described in our study) and that despite divergence from a common ancestors, grass genes have dramatically fewer (5- to 20-fold) and smaller CNSs than mammalian genes. One possible explanation is that, in contrast to vertebrate genomes, plant genomes have been subjected to more rounds of whole genome duplications (polyploidization) events that have profoundly affected their organisation, the subfunctionalisation of duplicated genes leading to a greater per gene loss of CNS [47].

Differential TE insertion dynamics

No class I or class II TE annotated within or outside the 'SPA orthologous region' is common when comparing any two-genome combinations. The two *WIS* retrotransposons, displaying similar apparent insertion positions in the 5' SPA locus boundaries of the A and D genomes correspond to independent insertions as Target Site Duplication (TSD) signature-motifs are distinct (respectively TATTG and TGTGA). This is also confirmed by estimation of their insertion dates with a transition and transversion ratio of 0.0029+/-0.004 (*i.e*. insertion date of 1.9–2.6 MYA) and 0.012+/-0.003 (*i.e*. insertion date of 0.7–1.2 MYA) for respectively the A and D genome sequences (*cf* Additional File 4). The differential insertion of TEs is surprisingly the case of the B and S genomes. Overall, we count six (two class II TEs, one unclassified TE and three MITEs) and eight (five class I TEs, two class II TEs and one

MITE) TEs differentially inserted in the B and S genomes respectively (*cf* Figure 1A). The 'SPA orthologous region' of the S genome has been invaded by retrotransposons, whereas outside the 'SPA orthologous region' the B genome seems to have a specific site for the insertion of class II TEs (mainly CACTA elements representing 23.9% of the sequence). Overall, we were able to estimate insertion dates for 8 retrotransposons. Out of them, only one (Angela_2) has been inserted into the 'SPA orthologous region' of *Ae. speltoides*, (estimated insertion date 1.3 to 1.9 MYA). Thus, the differential insertions of TEs in the S genome might be posterior to the S and B genome progenitors divergence from a common ancestor 2.5 MYA, 3.5 in the present study. Figure 4 retraces the process of TE differential insertion-deletions from a suggested *Triticum-Aegilops* 'ancestral SPA Locus' sequence of 16 598 bp that has been subjected to intensive TE insertions in the A, D and S genomes as compared to the B genome analysed in the present study.

The progenitor enigma of the B genome of polyploid wheat species

According to the two allopolyploidization events that gave rise to *T. aestivum*, the D genomes of the hexaploid wheat have diverged relatively recently from that of its donnor *Ae. tauschii* (0.08–0.12 MYA) whereas divergence of the A and B genomes from their respective progenitors occurred much more earlier (< 0.5 MYA) [7,9,10]. For almost 50 years, it remained controversial whether the source of the B genome is unique (*i.e.* monophyletic origin) related to *Ae. speltoides* or whether this genome resulted from an introgression of several parental *Aegilops* species (*i.e.* polyphyletic origin) [9,12-24,48]. Recent data on molecular comparisons using germplasm collections clearly show that the B genome could be related to several *Ae. speltoides* lines but not to other species of the Sitopsis section [25,49].

Comparison between the A genome of polyploid wheat species to that of its progenitor *T.urartu* at the PSR920 region [32] has shown a very high CDS conservation (99.5% of sequence identity at the third base of codons and 99.6% for introns). Moreover, Dvorak et al. [32] found in the 103 kb intergenic sequences four conserved TEs (inserted prior to their divergence) whereas four and one other TEs were respectively inserted in the A genome of *T. urartu* and that of *T. durum*, after their divergence from a common ancestor. Our present comparison based on CDS and CNS confirms that the B genome is closer to the S genome of *Ae. speltoides* than the A and D genomes. However, SPA sequence divergence and the differential insertions/deletions of TEs, none of which is conserved between the two genomes, indicate that *Ae. speltoids* have diverged very early (> 3MYA, in our study) from the B genome progenitor.

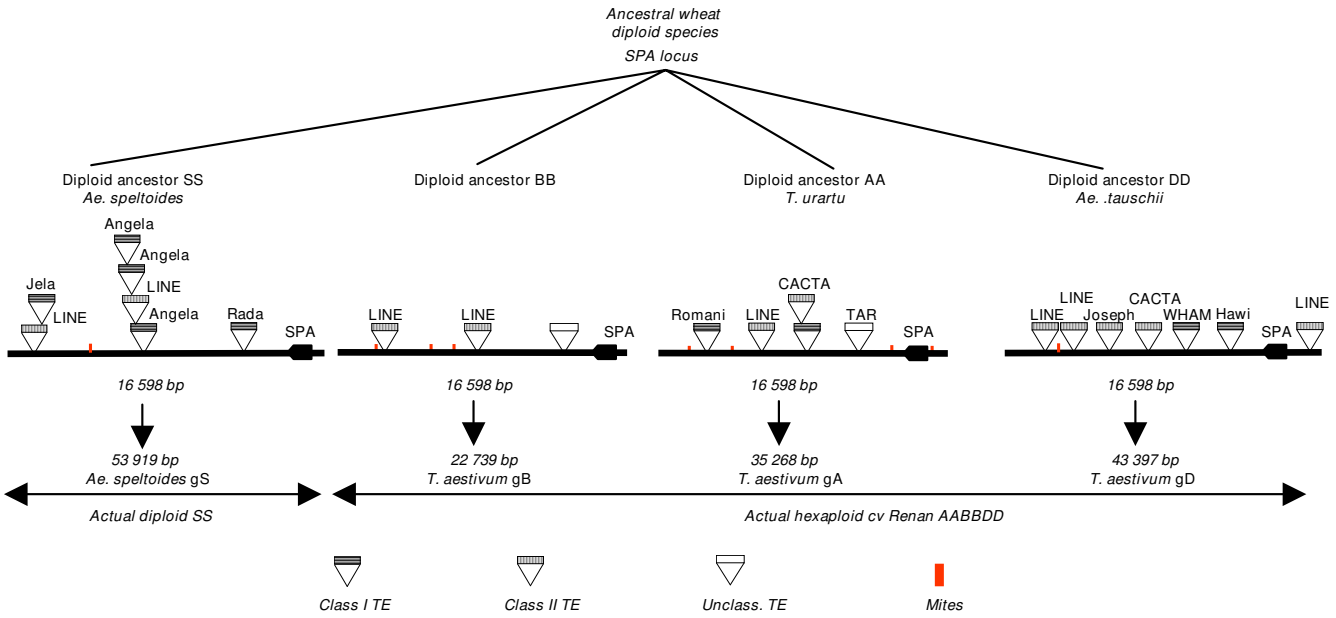


Figure 4

Evolutionary structure of the 'Ancestral SPA Locus'. From the 'Ancestral SPA Locus' of 16 598 bp, nested insertions of identified TE are shown for the four sequences (gA, gB, gD, gS). Graphical motifs used to materialize class I, class II unclassified TE as well as MITE are mentioned on the figure.

Conclusion

The present study based on detailed CDS, CNS and TE dynamics comparisons, clearly shows that evolutionary relationship between the B genome and the S genome of *Ae. speltoides* is not as close as it has been reported in the literature for the A genome of polyploid wheat species compared to its identified progenitor, *T. urartu*. Thus, a B genome progenitor remains to be identified.

Methods

BAC Clone Isolation

A BAC (Bacterial Artificial Chromosomes) library from *T. aestivum* cv renan [50] and *Ae. speltoides* BAC library (Chalhoub et al., unpublished) were screened with SPA PCR markers [34,42]. Assignment to the A, B, or D genomes of the BAC clones from the hexaploid species was based on their further characterization by Hind*III* restriction fragment length polymorphisms and specific PCR primers [42]. To ensure maximum coverage of the SPA locus, the longest BAC clones for the A (Ren1424A05, Accession#: FM242575), B (Ren0871J20, Accession#: FM242576), D (Ren2409K09, Accession#: FM242578) and S (Sho42-9K3, Accession#: FM242577) genomes were sequenced.

BAC sequencing and annotation

BAC shotgun sequencing was performed at the Centre National de Sequencage (Evry, France). Genes and repeated elements (TEs and short repeats) were identified by computing and integrating results based on BLAST algorithms [51,52], predictor programs, and different software detailed as follows.

Gene structure analysis

Gene structures and putative functions were identified by combining results of BLASTN and BLASTX alignments against dbEST http://www.ncbi.nlm.nih.gov/ and SwissProt databases http://expasy.org/sprot/, with results of 2 gene predictor programs, Eugene [53] with rice (*Oryza sativa*) training version and FgeneSH [54] (with default parameters http://linux1.softberry.com/berry.phtml. To incorporate heterologous information, we only recovered potential gene coding sequences. The CDS (CoDing Sequence) structures correspond to a consensus derived from the three preceding information sources. The gene content parameter represents the sum of known genes, hypothetical genes, unknown genes, and pseudogenes. Known genes were named based on BLASTX results against proteins with known functions (SwissProt). CDSs were considered as (i) hypothetical genes if their identification was only based on the predictors (as a consensus of the structures suggested by both predictors), without any evidence of putative function based on BLASTX results; (ii) unknown genes if the identification was only based on matching ESTs, without any evidence of putative function based on BLASTX results; (iii) pseudogenes if frame shifts need to be introduced within the CDS structure to better fit a putative function based on BLASTX results. Truncated pseudogenes, (genes disrupted by large insertion or deletion) and highly degenerated CDS sequences were considered as gene relics.

Transposable elements (TE)

TEs were detected by comparison with two databases of repetitive elements: TREP ([55]; http://wheat.pw.usda.gov/ITMI/Repeats/), and Repbase ([56]; http://www.girinst.org/Repbase_Update.html). Core domains (nucleic coordinates of known elements) were identified through BLASTN alignments against TREPnr. LTRs (Long Terminal Repeats) and TE boundaries were identified through BLASTN alignments against Repbase. Putative polyproteins were identified by BLASTX alignments against TREPprot. We used $1e^{-04}$ as a cutoff for BLASTN alignment results (either on TREPnr or Repbase). No cut-off was imposed for BLASTX results on TREPprot. Nested insertions of TEs were considered only when complete reconstruction of the split element was possible with no ambiguity. Other TE structures (either novel or highly degenerated TEs) were identified within the remaining unassigned DNA either by LTR_STRUC [57] or by BLASTX against the NCBI nr database http://www.ncbi.nlm.nih.gov/. When it was possible (*i.e.* for complete TEs), target-site duplications were indicated in the commentary of the element.

Pairwise comparisons of the four BAC clones, including the analysis of each BAC sequence against itself, were performed using the program Dotter [58] in order to identify or confirm direct repeats, LTRs, local duplications, and deletion events as well as MITEs. Multiple sequences comparisons were performed with PIPMAKER software [59]. As a final screening, unassigned DNA (free of annotated genes or TEs) was aligned using BLASTX against the NCBI nonredundant database http://www.ncbi.nlm.nih.gov. This BLASTX analysis allows the extension of several TE features already identified. TEs were classified and named based on the unified classification from Wicker et al. [60] according to referred nomenclature (*i.e.*, element name, BAC name, appearance rank) and designed as complete, truncated, and degenerated sequences as suggested by TREP or Repbase databases.

Short repeated motifs

Short repeated motifs were identified either as inverted repeats (by using EINVERTED with default parameters; http://emboss.bioinformatics.nl/cgi-bin/emboss/einverted) or tandem repeats (Tandem Repeat Finder, with default parameters; http://tandem.bu.edu/trf/trf.advanced.submit.html). Only repeated domains (*i.e.* tandem or inverted) longer than 100 bp were kept in our annotation results.

Unassigned DNA sequences

Unassigned DNA corresponds to sequences in which neither CDS nor TE was identified. Such unassigned DNA may contain short repetitive units (tandem repeats or inverted repeats).

Integration of annotation results

Cross-analysis of the information obtained for genes and TEs as short repeats was integrated into ARTEMIS [61].

Sequence analysis

Multiple alignments

Identification of conserved domains was performed based on multiple alignments (clustalw, [62]) on translated SPA CDS (identified from the sequence annotation procedure).

Phylogeny analysis

The phylogenetic analysis was performed using Neighbor-joining method with clustalx alignment of protein sequences with 1 000 repetition bootstraps. The BLOSUM 62 matrix was chosen for substitution identification. The sequence divergence datation was performed based on the rate of nonsynonymous (*Ka*) vs. synonymous (*Ks*) substitutions calculated with MEGA-3 [63]. The average substitution rate (r) of 6.5×10^{-9} substitutions per synonymous site per year for grasses was used to calibrate the ages of the considered gene ([64,65]. The time (*T*) since gene insertion was estimated using the formula $T = Ks/r$.

Determination retrotransposons insertion dates

Full-length retrotransposons were analysed by comparing their 5' and 3' LTR sequences in order to date their insertion time [65] based on the assumption that the two LTRs of a single element are identical at the time of insertion. The two LTRs were aligned and the number of transition and transversion mutation were counted. The insertion times were dated using the Kimura parameter method (K2P, [66]) and a mutation rate of 6.5×10^{-9} substitutions per synonymous site per year [64]. The time (*T*) since element insertion was estimated using the formula $T = K2P/2r$.

Authors' contributions

JSperformed the BAC sequence annotation and analysis and comparative annotation and wrote the manuscript. VC, SB, CP, MC and NB contributed in sequence analysis and annotation as well as transposable elements evolution. CH, HB, SG and AEwere implicated into *(i)* the construction of the *Ae. speltoides* BAC library, *(ii)* the screening of the BAC libraries, *(iii)* the identification and verification of the positive BAC clones as well as PCR genotyping. GM, AC, BSand SSwere implicated in BAC clone sequencing, sequence assembly and verification of assembled sequences. CRand GCwere involved in the interpretation of SPA gene sequence comparisons. BC, coordinator of the project, set up the project and followed analysis and interpretation of the results as well as wrote and edited the manuscript.

Additional material

Additional file 1

***BAC clones annotation**. detailed annotation features for T. aestivum -gA, -gB, -gD and Ae. speltoides gS sequences as GenBank format files.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-9-555-S1.doc]

Additional file 2

***BAC clones gene and TE content**. Detailed features (genes, TE) of the 4 annotated BAC clone T. aestivum -gA, -gB, -gD and Ae. speltoides gS sequences.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-9-555-S2.xls]

Additional file 3

***SPA genotyping data**. SPA genotyping data among 18 wheat genotypes.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-9-555-S3.xls]

Additional file 4

***TE divergence analysis**. Divergence time for 8 complete Class I transposable elements.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-9-555-S4.xls]

Acknowledgements

The Renan BAC library and sequencing of the BAC clones from the A, B and D genomes of hexaploid wheat were supported by the Genoplante consortium http://www.genoplante.com. Sequencing of the *Aegilops tauschii* BAC clone was supported by the APCNS2003 project 'Comparative genome sequencing in wheat' http://www.cns.fr/externe/English/Projets/Projet_LE/LE.html.

References

1. Feuillet C, Keller B: **Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution.** *Ann Bot (Lond)* 2002, **89:**3-10.
2. Kellogg EA: **Evolutionary history of the grasses.** *Plant Physiol* 2001, **125:**1198-1205.
3. Gaut BS: **Evolutionary dynamics of grass geno.** *New phytologist* 2002, **154:**15-28.
4. Harlan JR: **Crops and Man.** Madison, Wisconsin: American Society of Agronomy, Inc; 1992.
5. Zohary D, Hopf M: **Domestication of plants in the Old World.** 3rd edition. New York: Oxford University Press; 2000.
6. Piperno DR, Flannery KV: **The earliest archaeological maize (*Zea mays L*) from highland Mexico: new accelerator mass spectrometry dates and their implications.** *Proc Natl Acad Sci USA* 2001, **13:**2101-2103.

7. Feldman M, Lupton FGH, Miller TE: **Wheats.** In *Evolution of Crops* 2nd edition. Edited by: Smartt J, Simmonds NW. London: Longman Scientific; 1995:184-192.
8. Eckardt NA: **A sense of self: The role of DNA sequence elimination in allopolyploidization.** *Plant Cell* 2001, **13:**1699-1704.
9. Huang S, Sirikhachornkit A, Su XJ, Faris J, Gill B, Haselkorn R, Gornicki P: **Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat.** *Proc Natl Acad Sci USA* 2002, **99:**8133-8138.
10. Blake NK, Lehfeldt BR, Lavin M, Talbert LE: **Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: The B genome of wheat.** *Genome* 1999, **42:**351-360.
11. Nesbitt M, Samuel D: **From staple crop to extinction? The archaeology and history of the hulled wheats.** *Hulled wheats. Proceedings of the First International Workshop on Hulled Wheats. Promoting the conservation and use of underutilized and neglected crops 4* 1996:41-100.
12. Zohary D, Feldman M: **Hybridization between amphidiploids and the evolution of polyploids in the wheat (Aegilops-Triticum) group.** *Evolution* 1962, **16:**44-61.
13. Dvorák J, Zhang HB, Kota RS, Lassner M: **Organization and evolution of the 5S ribosomal RNA gene family in wheat and related species.** *Genome* 1989, **32:**1003-1016.
14. Dvorák J, Zhang HB: **Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes.** *Proc Natl Acad Sci USA* 1990, **87:**9640-9644.
15. Terachi T, Ogihara Y, Tsunewaki K: **The molecular basis of genetic diversity among cytoplasms of Triticum and Aegilops. 7. Restriction endonuclease analysis of mitochondrial DNA from polyploid wheats and their ancestral species.** *Theor Appl Genet* 1990, **80:**366-373.
16. Feldman M: **Identification of unpaired chromosomes in F1 hybrids involving Triticum aestivum and T. timopheevii.** *Can J Genet Cytol* 1966, **8:**144-151.
17. Feldman M: **The mechanism regulating pairing in *Triticum timopheevii*.** *Wheat Inf Serv* 1966, **21:**1-2.
18. Hutchinson J, Miller TE, Jahier J, Shepherd KW: **Comparison of the chromosomes of *Triticum timopheevii* with related wheats using the techniques of C-banding and in situ hybridization.** *Theor Appl Genet* 1982, **64:**31-40.
19. Gill BS, Chen PD: **Role of cytoplasm specific introgression in the evolution of the polyploid wheats.** *Proc Natl Acad Sci USA* 1987, **84:**6800-6804.
20. Naranjo T, Roca A, Goicoechea PG, Giráldez R: **Arm homoeology of wheat and rye chromosomes.** *Genome* 1987, **29:**873-882.
21. Naranjo T: **Chromosome structure of durum wheat.** *Theor Appl Genet* 1990, **79:**397-400.
22. Jiang J, Gill BS: **Different species-specific chromosome translocations in *Triticum timopheevii* and *T. turgidum* support the diphyletic origin of polyploid wheats.** *Chromosome Res* 1994, **2:**59-64.
23. Devos KM, Dubcovsky J, Dvorák J, Chinoy CN, Gale MD: **Structural evolution of wheat chromosomes 4A, 5A and 7B and its impact on recombination.** *Theor Appl Genet* 1995, **91:**282-288.
24. Maestra B, Naranjo T: **Structural chromosome differentiation between *Triticum timopheevii* and *T. turgidum* and *T. aestivum*.** *Theor Appl Genet* 1999, **98:**744-750.
25. Kilian B, Ozkan H, Deusch O, Effgen S, Brandolini A, Kohl J, Martin W, Salamini F: **Independent wheat B and G genome origins in outcrossing Aegilops progenitor haplotypes.** *Mol Biol Evol* 2007, **24:**217-227.
26. Salina EA, Lim KY, Badaeva ED, Shcherban AB, Adonina IG, Amosova AV, Samatadze TE, Vatolina TY, Zoshchuk SA, Leitch AR: **Phylogenetic reconstruction of Aegilops section Sitopsis and the evolution of tandem repeats in the diploids and derived wheat polyploids.** *Genome* 2006, **49(8):**1023-35.
27. Smith DB, Flavell RB: **Characterisation of the wheat genome by renaturation kinetics.** *Chromosoma (Berl)* 1975, **50:**223242.
28. Vedel E, Delseny M: **Repetitivity and variability of higher plant genomes.** *Pl Physiol Biochem* 1987, **25:**191-210.
29. Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu ZD, Dubcovsky J, Keller B: **Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat.** *Plant Cell* 2003, **15:**1186-1197.
30. Isidore E, Scherrer B, Chalhoub B, Feuillet C, Keller B: **Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels.** *Genome Res* 2005, **15(4):**526-36.
31. Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, Gautier MF, Cattolico L, Beckert M, Aubourg S, Weissenbach J, Caboche M, Bernard M, Leroy P, Chalhoub B: **Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops).** *Plant Cell* 2005, **17(4):**1033-45.
32. Dvorak J, Akhunov ED, Akhunov AR, Deal KR, Luo MC: **Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat.** *Mol Biol Evol* 2006, **23(7):**1386-96.
33. Gu YQ, Salse J, Coleman-Derr D, Dupin A, Crossman C, Lazo GR, Huo N, Belcram H, Ravel C, Charmet G, Charles M, Anderson OD, Chalhoub B: **Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes.** *Genetics* 2006, **174(3):**1493-504.
34. Albani D, Hammond-Kosack MC, Smith C, Conlan S, Colot V, Holdsworth M, Bevan MW: **The wheat transcriptional activator SPA: a seed-specific bZIP protein that recognizes the GCN4-like motif in the bifactorial endosperm box of prolamin genes.** *Plant Cell* 1997, **9:**171-184.
35. Guillaumie S, Charmet G, Linossier L, Torney V, Robert N, Ravel C: **Colocation between a gene encoding the bZip factor SPA and an eQTL for a high-molecular-weight glutenin subunit in wheat (*Triticum aestivum*).** *Genome* 2004, **47(4):**705-13.
36. Schmidt RJ, Ketudat M, Aukerman MJ, Hoschek G: **Opaque-2 is a transcriptional activator that recognizes a pecific target site in 22-kD zein genes.** *Plant Cell* 1992, **4:**689-700.
37. Schmidt RJ: **Opaque-2 and zein gene expression.** In *Control of Plant Gene Expression* Edited by: Verma DPS. Boca Raton, FL: CRC Press; 1993:337-355.
38. Vicente-Carbajosa J, Moose SP, Parsons RL, Schmidt RJ: **A maize zinc-finger protein binds the prolamin box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator Opaque2.** *Proc Natl Acad Sci USA* 1997, **94(14):**7685-90.
39. Onodera Y, Suzuki A, Wu CY, Washida H, Takaiwa F: **A rice functional transcriptional activator, RISBZ1, responsible for endosperm-specific expression of storage protein genes through GCN4 motif.** *J Biol Chem* 2001, **276(17):**14139-52.
40. Vicente-Carbajosa J, Onate L, Lara P, Diaz I, Carbonero P: **Barley BLZ1: a bZIP transcriptional activator that interacts with endosperm-specific gene promoters.** *Plant J* 1998, **13:**629-640.
41. Onate L, Vicente-Carbajosa J, Lara P, Diaz I, Carbonero P: **Barley BLZ2, a seed-specific bZIP protein that interacts with BLZ1 in vivo and activates transcription from the GCN4-like motif of B-hordein promoters in barley endosperm.** *J Biol Chem* 1999, **274(14):**9175-82.
42. Ravel C, Praud S, Murigneux A, Canaguier A, Sapet F, Samson D, Balfourier F, Dufour P, Chalhoub B, Brunel D, Beckert M, Charmet G: **Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.).** *Genome* 2006, **49(9):**1131-9.
43. Hartings H, Maddaloni M, Lazzaroni N, Di Fonzo N, Motto M, Salamini F, Thompson R: **The O2 gene which regulates zein deposition in maize endosperm encodes a protein with structural homologies to transcriptional activators.** *EMBO J* 1989, **8:**2795-2801.
44. Pirovano L, Lanzini S, Hartings H, Lazzaroni N, Rossi V, Joshi R, Thompson RD, Salamini F, Motto M: **Structural and functional analysis of an Opaque-2-related gene from sorghum.** *Plant Mol Biol* 1994, **24(3):**515-23.
45. Kaplinsky NJ, Braun DM, Penterman J, Gof SA, Freeling M: **Utility and distribution of conserved noncoding sequences in the grasses.** *Proc Natl Acad Sci USA* 2002, **99:**6147-6151.
46. Inada DC, Bashir A, Lee C, Thomas BC Ko C, Goff SA, Freeling M: **Conserved noncoding sequences in the grasses.** *Genome Res* 2003, **13:**2030-2041.
47. Lockton S, Gaut BS: **Plant conserved non-coding sequences and paralogue evolution.** *Trends Gene* 2005, **21:**60-65.

48. Buchner P, Prosser IM, Hawkesford MJ: **Phylogeny and expression of paralogous and orthologous sulphate transporter genes in diploid and hexaploid wheats.** *Genome* 2004, **47(3):**526-34.
49. Wang JR, Zhang L, Wei YM, Yan ZH, Baum BR, Nevo E, Zheng YL: **Sequence polymorphisms and relationships of dimeric α-amylase inhibitor genes in the B genomes of Triticum and S genomes of Aegilops.** *Plant Science* 2007, **173:**1-11.
50. Chalhoub B, Belcram H, Caboche M: **Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size.** *Plant Biotechnol J* 2004, **2(3):**181-8.
51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Bi* 1990, **215:**403-410.
52. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **A new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
53. Mathe C, Sagot MF, Schiex T, Rouze P: **Current methods of gene prediction, their strengths and weaknesses.** *Nucleic Acids Res* 2002, **30(19):**4103-17.
54. Salamov A, Solovyev V: **Ab initio gene finding in Drosophila genomic DNA.** *Genome Res* 2000, **10:**516-522.
55. Wicker T, Matthews DE, Keller B: **TREP: A database for Triticeae repetitive elements.** *Trends Plant Sci* 2002, **7:**561-562.
56. Jurka J: **Repbase update: A database and an electronic journal of repetitive elements.** *Trends Genet* 2000, **9:**418-420.
57. McCarthy E, McDonald J: **LTR_STRUC: A novel search and identification program for LTR retrotransposons.** *Bioinformatics* 2003, **19:**362-367.
58. Sonnhammer EL, Durbin R: **A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis.** *Gene* 1995, **167:**1-10.
59. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker-a web server for aligning two genomic DNA sequences.** *Genome Res* 2000, **10(4):**577-86.
60. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH: **A unified classification system for eukaryotic transposable elements.** *Nat Rev Genet* 2007, **8(12):**973-82.
61. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: Sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945.
62. Aiyar A: **The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment.** *Methods Mol Biol* 2000, **132:**221-41.
63. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Brief Bioinform* 2004, **5:**150-163.
64. Gaut BS, Morton BR, McCaig BC, Clegg MT: **Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL.** *Proc Natl Acad Sci USA* 1996, **93(19):**10274-9.
65. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize.** *Nat Genet* 1998, **20(1):**43-5.
66. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16(2):**111-20.

Annexe 5 : Article 5

ARTICLES

Genome sequencing and analysis of the model grass *Brachypodium distachyon*

The International Brachypodium Initiative*

Three subfamilies of grasses, the Ehrhartoideae, Panicoideae and Pooideae, provide the bulk of human nutrition and are poised to become major sources of renewable energy. Here we describe the genome sequence of the wild grass *Brachypodium distachyon* (*Brachypodium*), which is, to our knowledge, the first member of the Pooideae subfamily to be sequenced. Comparison of the *Brachypodium*, rice and sorghum genomes shows a precise history of genome evolution across a broad diversity of the grasses, and establishes a template for analysis of the large genomes of economically important pooid grasses such as wheat. The high-quality genome sequence, coupled with ease of cultivation and transformation, small size and rapid life cycle, will help *Brachypodium* reach its potential as an important model system for developing new energy and food crops.

Grasses provide the bulk of human nutrition, and highly productive grasses are promising sources of sustainable energy[1]. The grass family (Poaceae) comprises over 600 genera and more than 10,000 species that dominate many ecological and agricultural systems[2,3]. So far, genomic efforts have largely focused on two economically important grass subfamilies, the Ehrhartoideae (rice) and the Panicoideae (maize, sorghum, sugarcane and millets). The rice[4] and sorghum[5] genome sequences and a detailed physical map of maize[6] showed extensive conservation of gene order[5,7] and both ancient and relatively recent polyploidization.

Most cool season cereal, forage and turf grasses belong to the Pooideae subfamily, which is also the largest grass subfamily. The genomes of many pooids are characterized by daunting size and complexity. For example, the bread wheat genome is approximately 17,000 megabases (Mb) and contains three independent genomes[8]. This has prohibited genome-scale comparisons spanning the three most economically important grass subfamilies.

Brachypodium, a member of the Pooideae subfamily, is a wild annual grass endemic to the Mediterranean and Middle East[9] that has promise as a model system. This has led to the development of highly efficient transformation[10,11], germplasm collections[12–14], genetic markers[14], a genetic linkage map[15], bacterial artificial chromosome (BAC) libraries[16,17], physical maps[18] (M.F., unpublished observations), mutant collections (http://brachypodium.pw.usda.gov, http://www.brachytag.org), microarrays and databases (http://www.brachybase.org, http://www.phytozome.net, http://www.modelcrop.org, http://mips.helmholtz-muenchen.de/plant/index.jsp) that are facilitating the use of *Brachypodium* by the research community. The genome sequence described here will allow *Brachypodium* to act as a powerful functional genomics resource for the grasses. It is also an important advance in grass structural genomics, permitting, for the first time, whole-genome comparisons between members of the three most economically important grass subfamilies.

Genome sequence assembly and annotation

The diploid inbred line Bd21 (ref. 19) was sequenced using whole-genome shotgun sequencing (Supplementary Table 1). The ten largest scaffolds contained 99.6% of all sequenced nucleotides (Supplementary Table 2). Comparison of these ten scaffolds with a genetic map (Supplementary Fig. 1) detected two false joins and created a further seven joins to produce five pseudomolecules that spanned 272 Mb (Supplementary Table 3), within the range measured by flow cytometry[20,21]. The assembly was confirmed by cytogenetic analysis (Supplementary Fig. 2) and alignment with two physical maps and sequenced BACs (Supplementary Data). More than 98% of expressed sequence tags (ESTs) mapped to the sequence assembly, consistent with a near-complete genome (Supplementary Table 4 and Supplementary Fig. 3). Compared to other grasses, the *Brachypodium* genome is very compact, with retrotransposons concentrated at the centromeres and syntenic breakpoints (Fig. 1). DNA transposons and derivatives are broadly distributed and primarily associated with gene-rich regions.

We analysed small RNA populations from inflorescence tissues with deep Illumina sequencing, and mapped them onto the genome sequence (Fig. 2a, Supplementary Fig. 4 and Supplementary Table 5). Small RNA reads were most dense in regions of high repeat density, similar to the distribution reported in *Arabidopsis*[22]. We identified 413 and 198 21- and 24-nucleotide phased short interfering RNA (siRNA) loci, respectively. Using the same algorithm, the only phased loci identified in *Arabidopsis* were five of the eight *trans*-acting siRNA loci, and none was 24-nucelotide phased. The biological functions of these clusters of *Brachypodium* phased siRNAs, which account for a significant number of small RNAs that map outside repeat regions, are not known at present.

A total of 25,532 protein-coding gene loci was predicted in the v1.0 annotation (Supplementary Information and Supplementary Table 6). This is in the same range as rice (RAP2, 28,236)[23] and sorghum (v1.4, 27,640)[5], suggesting similar gene numbers across a broad diversity of grasses. Gene models were evaluated using ~10.2 gigabases (Gb) of Illumina RNA-seq data (Supplementary Fig. 5)[24]. Overall, 92.7% of predicted coding sequences (CDS) were supported by Illumina data (Fig. 2b), demonstrating the high accuracy of the *Brachypodium* gene predictions. These gene models are available from several databases (such as http://www.brachybase.org, http://www.phytozome.net, http://www.modelcrop.org and http://mips.org).

Between 77 and 84% of gene families (defined according to Supplementary Fig. 6) are shared among the three grass subfamilies represented by *Brachypodium*, rice and sorghum, reflecting a relatively

*A list of participants and their affiliations appears at the end of the paper.

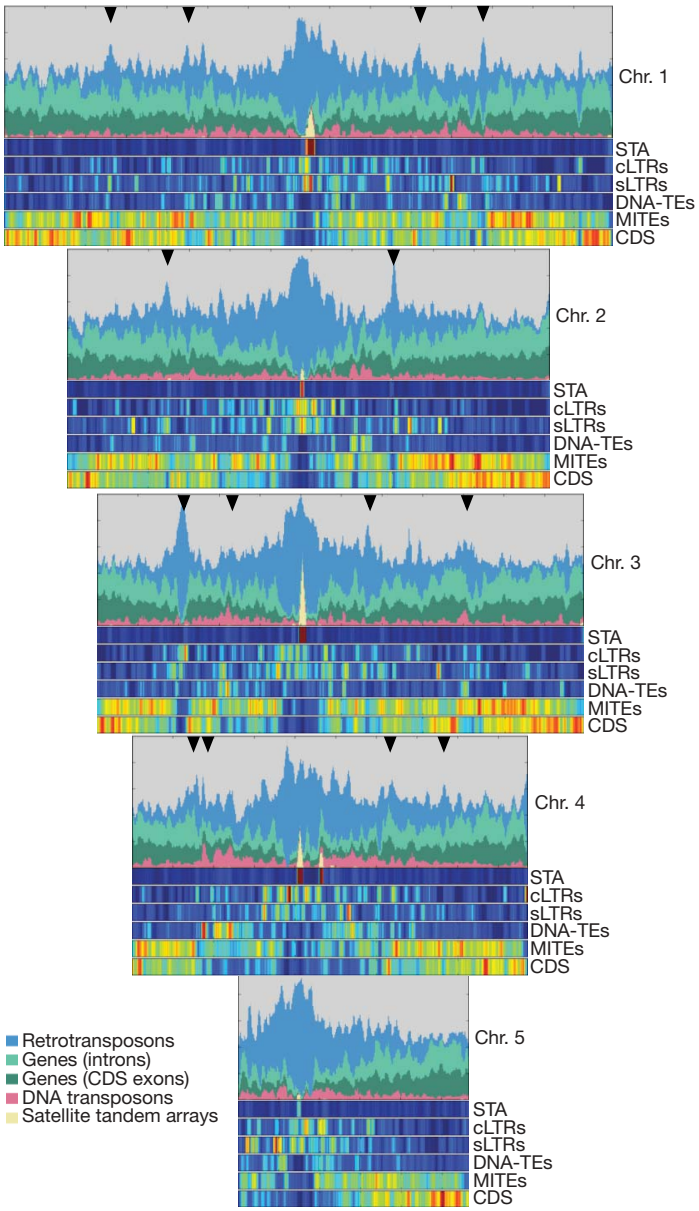


Figure 1 | Chromosomal distribution of the main *Brachypodium* genome features. The abundance and distribution of the following genome elements are shown: complete LTR retroelements (cLTRs); solo-LTRs (sLTRs); potentially autonomous DNA transposons that are not miniature inverted-repeat transposable elements (MITEs) (DNA-TEs); MITEs; gene exons (CDS); gene introns and satellite tandem arrays (STA). Graphs are from 0 to 100 per cent base-pair (%bp) coverage of the respective window. The heat map tracks have different ranges and different maximum (max) pseudocolour levels: STA (0–55, scaled to max 10) %bp; cLTRs (0–36, scaled to max 20) %bp; sLTRs (0–4) %bp; DNA-TEs (0–20) %bp; MITEs (0–22) %bp; CDS (exons) (0–22.3) %bp. The triangles identify syntenic breakpoints.

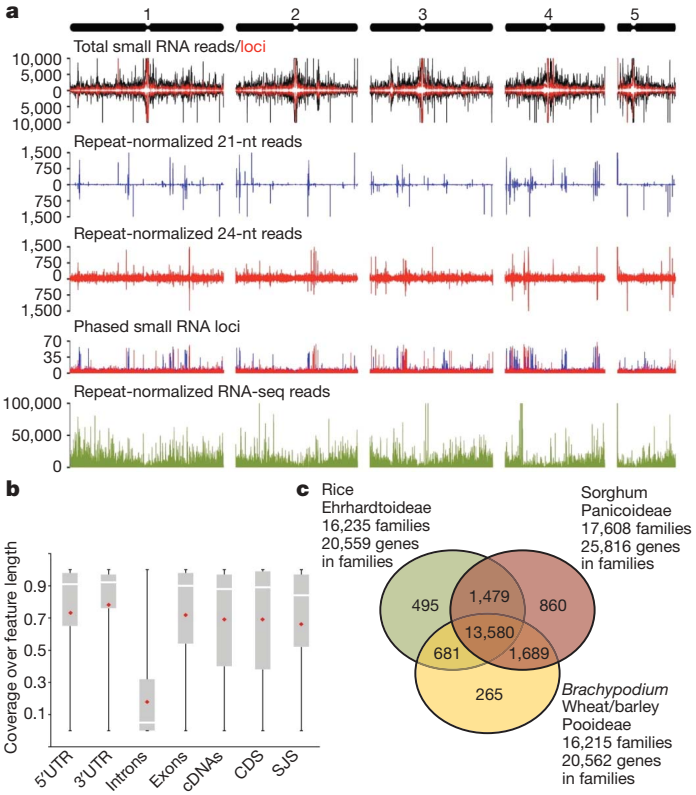


Figure 2 | Transcript and gene identification and distribution among three grass subfamilies. **a**, Genome-wide distribution of small RNA loci and transcripts in the *Brachypodium* genome. *Brachypodium* chromosomes (1–5) are shown at the top. Total small RNA reads (black lines) and total small RNA loci (red lines) are shown on the top panel. Histograms plot 21-nucleotide (nt) (blue) or 24-nucleotide (red) small RNA reads normalized for repeated matches to the genome. The phased loci histograms plot the position and phase-score of 21-nucleotide (blue) and 24-nucleotide (red) phased small RNA loci. Repeat-normalized RNA-seq read histograms plot the abundance of reads matching RNA transcripts (green), normalized for ambiguous matches to the genome. **b**, Transcript coverage over gene features. Perfect match 32-base oligonucleotide Illumina reads were mapped to the *Brachypodium* v1.0 annotation features using HashMatch (http://mocklerlab-tools.cgrb.oregonstate.edu/). Plots of Illumina coverage were calculated as the percentage of bases along the length of the sequence feature supported by Illumina reads for the indicated gene model features. The bottom and top of the box represent the 25th and 75th quartiles, respectively. The white line is the median and the red diamonds denote the mean. SJS, splice junction site. **c**, Venn diagram showing the distribution of shared gene families between representatives of Ehrhartoideae (rice RAP2), Panicoideae (sorghum v1.4) and Pooideae (*Brachypodium* v1.0, and *Triticum aestivum* and *Hordeum vulgare* TCs (transcript consensus)/EST sequences). Paralogous gene families were collapsed in these data sets.

recent common origin (Fig. 2c). Grass-specific genes include transmembrane receptor protein kinases, glycosyltransferases, peroxidases and P450 proteins (Supplementary Table 7B). The Pooideae-specific gene set contains only 265 gene families (Supplementary Table 7C) comprising 811 genes (1,400 including singletons). Genes enriched in grasses were significantly more likely to be contained in tandem arrays than random genes, demonstrating a prominent role for tandem gene expansion in the evolution of grass-specific genes (Supplementary Fig. 7 and Supplementary Table 8).

To validate and improve the v1.0 gene models, we manually annotated 2,755 gene models from 97 diverse gene families (Supplementary Tables 9–11) relevant to bioenergy and food crop improvement. We annotated 866 genes involved in cell wall biosynthesis/modification and 948 transcription factors from 16 families[25]. Only 13% of the gene models required modification and very few pseudogenes were identified, demonstrating the accuracy of the v1.0 annotation. Phylogenetic trees for 62 gene families were constructed using genes from rice, *Arabidopsis*, sorghum and poplar. In nearly all cases, *Brachypodium* genes had a similar distribution to rice and sorghum, demonstrating that *Brachypodium* is suitably generic for grass functional genomics research (Supplementary Figs 8 and 9). Analysis of the predicted secretome identified substantial differences in the distribution of cell wall metabolism genes between dicots and grasses (Supplementary Tables 12, 13 and Supplementary Fig. 10), consistent with their different cell walls[26]. Signal peptide probability curves also suggested that start codons were accurately predicted (Supplementary Fig. 11).

Maintaining a small grass genome size

Exhaustive analysis of transposable elements (Supplementary Information and Supplementary Table 14) showed retrotransposon sequences comprise 21.4% of the genome, compared to 26% in rice,

54% in sorghum, and more than 80% in wheat[27]. Thirteen retroelement sets were younger than 20,000 years, showing a recent activation compared to rice[28] (Supplementary Fig. 12), and a further 53 retroelement sets were less than 0.1 million years (Myr) old. A minimum of 17.4 Mb has been lost by long terminal repeat (LTR)–LTR recombination, demonstrating that retroelement expansion is countered by removal through recombination. In contrast, retroelements persist for very long periods of time in the closely related Triticeae[28].

DNA transposons comprise 4.77% of the *Brachypodium* genome, within the range found in other grass genomes[5,29]. Transcriptome data and structural analysis suggest that many non-autonomous *Mariner DTT* and *Harbinger* elements recruit transposases from other families. Two *CACTA DTC* families (M and N) carried five non-element genes, and the *Harbinger U* family has amplified a NBS-LRR gene family (Supplementary Figs 13 and 14), adding it to the group of transposable elements implicated in gene mobility[30,31]. Centromeric regions were characterized by low gene density, characteristic repeats and retroelement clusters (Supplementary Fig. 15). Other repeat classes are described in Supplementary Table 15. Conserved non-coding sequences are described in Supplementary Fig. 16.

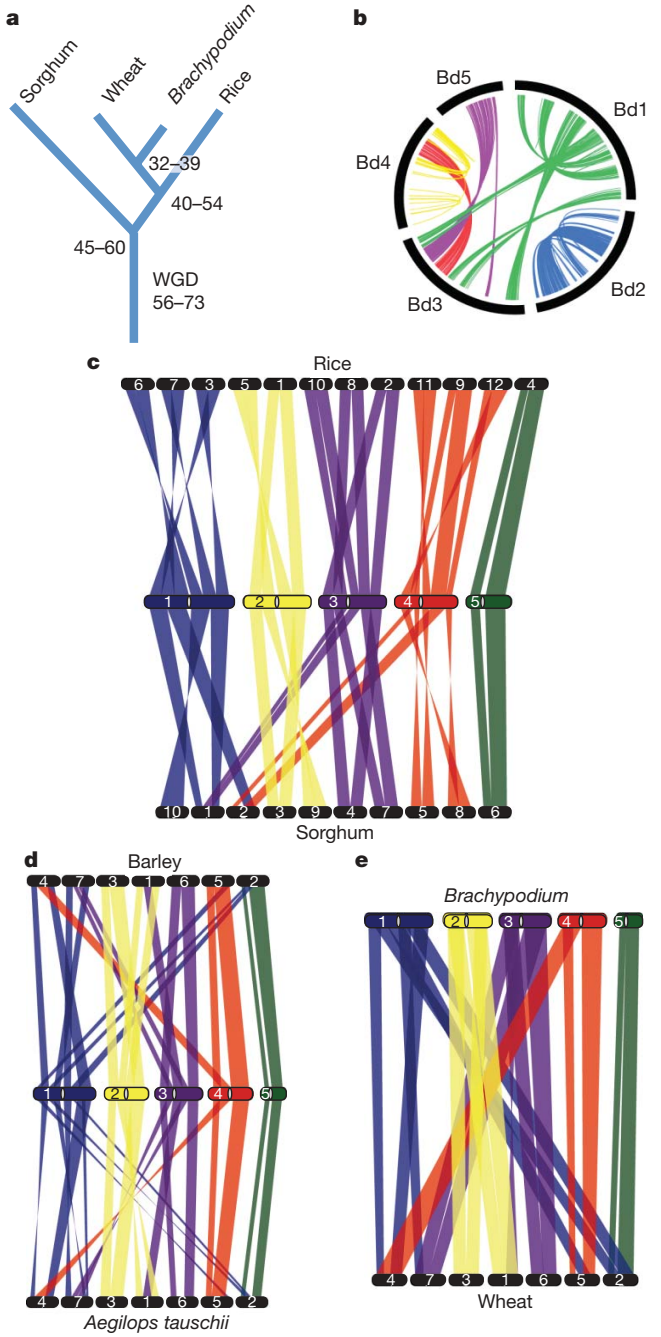


Figure 3 | *Brachypodium* genome evolution and synteny between grass subfamilies. a, The distribution maxima of mean synonymous substitution rates (K_s) of *Brachypodium*, rice, sorghum and wheat orthologous gene pairs (Supplementary Table 16) were used to define the divergence times of these species and the age of interchromosomal duplications in *Brachypodium*. WGD, whole-genome duplication. The numbers refer to the predicted divergence times measured as Myr ago by the NG or ML methods. **b**, Diagram showing the six major interchromosomal *Brachypodium* duplications, defined by 723 paralogous relationships, as coloured bands linking the five chromosomes. **c**, Identification of chromosome relationships between the *Brachypodium*, rice and sorghum genomes. Orthologous relationships between the 25,532 protein-coding *Brachypodium* genes, 7,216 sorghum orthologues (12 syntenic blocks), and 8,533 rice orthologues (12 syntenic blocks) were defined. Sets of collinear orthologous relationships are represented by a coloured band according to each *Brachypodium* chromosome (blue, chromosome (chr.) 1; yellow, chr. 2; violet, chr. 3; red, chr. 4; green, chr. 5). The white region in each *Brachypodium* chromosome represents the centromeric region. **d**, Orthologous gene relationships between *Brachypodium* and barley and *Ae. tauschii* were identified using genetically mapped ESTs. 2,516 orthologous relationships defined 12 syntenic blocks. These are shown as coloured bands. **e**, Orthologous gene relationships between *Brachypodium* and hexaploid bread wheat defined by 5,003 ESTs mapped to wheat deletion bins. Each set of orthologous relationships is represented by a band that is evenly spread across each deletion interval on the wheat chromosomes.

Whole-genome comparison of three diverse grass genomes

The evolutionary relationships between *Brachypodium*, sorghum, rice and wheat were assessed by measuring the mean synonymous substitution rates (K_s) of orthologous gene pairs (Supplementary Information, Supplementary Fig. 17 and Supplementary Table 16), from which divergence times of *Brachypodium* from wheat 32–39 Myr ago, rice 40–53 Myr ago, and sorghum 45–60 Myr ago (Fig. 3a) were estimated. The K_s of orthologous gene pairs in the intragenomic *Brachypodium* duplications (Fig. 3b) suggests duplication 56–72 Myr ago, before the diversification of the grasses. This is consistent with previous evolutionary histories inferred from a small number of genes[3,32–34].

Paralogous relationships among *Brachypodium* chromosomes showed six major chromosomal duplications covering 92.1% of the genome (Fig. 3b), representing ancestral whole-genome duplication[35]. Using the rice and sorghum genome sequences, genetic maps of barley[36] and *Aegilops tauschii* (the D genome donor of hexaploid wheat)[37], and bin-mapped wheat ESTs[38,39], 21,045 orthologous relationships between *Brachypodium*, rice, sorghum and Triticeae were identified (Supplementary Information). These identified 59 blocks of collinear genes covering 99.2% of the *Brachypodium* genome (Fig. 3c–e). The orthologous relationships are consistent with an evolutionary model that shaped five *Brachypodium* chromosomes from a five-chromosome ancestral genome by a 12-chromosome intermediate involving seven major chromosome fusions[39] (Supplementary Fig. 18). These collinear blocks of orthologous genes provide a robust and precise sequence framework for understanding grass genome evolution and aiding the assembly of sequences from other pooid grasses. We identified 14 major syntenic disruptions between *Brachypodium* and rice/sorghum that can be explained by nested insertions of entire chromosomes into centromeric regions (Fig. 4a, b)[2,37,40]. Similar nested insertions in sorghum[37] and barley (Fig. 4c, d) were also identified. Centromeric repeats and peaks in retroelements at the junctions of chromosome insertions are footprints of these insertion events (Supplementary Fig. 15C and Fig. 1), as is higher gene density at the former distal regions of the inserted chromosomes (Fig. 1). Notably, the reduction in chromosome number in *Brachypodium* and wheat occurred independently because none of the chromosome fusions are shared by *Brachypodium* and the Triticeae[37] (Supplementary Fig. 18).

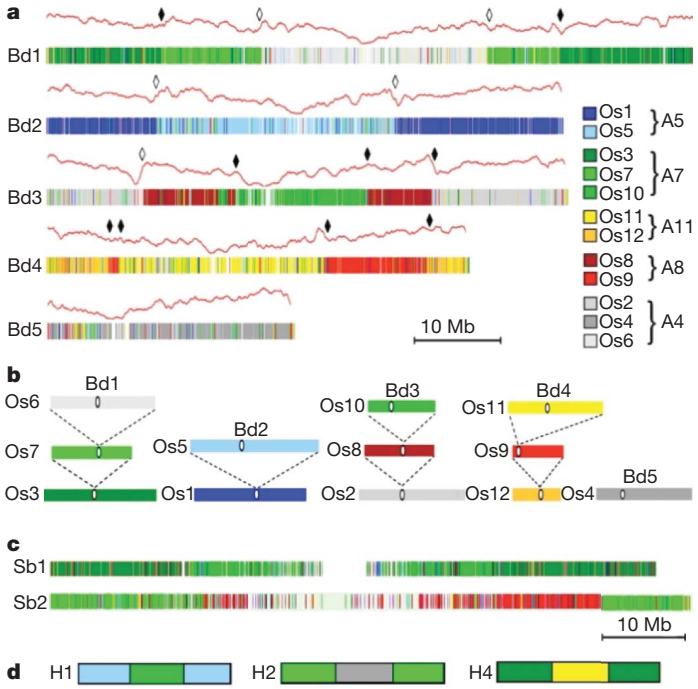


Figure 4 | A recurring pattern of nested chromosome fusions in grasses. **a**, The five *Brachypodium* chromosomes are coloured according to homology with rice chromosomes (Os1–Os12). Chromosomes descended from an ancestral chromosome (A4–A11) through whole-genome duplication are shown in shades of the same colour. Gene density is indicated as a red line above the chromosome maps. Major discontinuities in gene density identify syntenic breakpoints, which are marked by a diamond. White diamonds identify fusion points containing remnant centromeric repeats. **b**, A pattern of nested insertions of whole chromosomes into centromeric regions explains the observed syntenic break points. Bd5 has not undergone chromosome fusion. **c**, Examples of nested chromosome insertions in sorghum (Sb) chromosomes 1 and 2. **d**, Examples of nested chromosome insertions in barley (H chromosomes) inferred from genetic maps. Nested insertions were not identified in other chromosomes, possibly owing to the low resolution of genetic maps.

Comparisons of evolutionary rates between *Brachypodium*, sorghum, rice and *Ae. tauschii* demonstrated a substantially higher rate of genome change in *Ae. tauschii* (Supplementary Table 17). This may be due to retroelement activity that increases syntenic disruptions, as proposed for chromosome 5S later[41]. Among seven relatively large gene families, four were highly syntenic and two (NBS-LRR and F-box) were almost never found in syntenic order when compared to rice and sorghum (Supplementary Table 18), consistent with the rapid diversification of the NBS-LRR and F-box gene families[42].

The short arm of chromosome 5 (Bd5S) has a gene density roughly half of the rest of the genome, high LTR retrotransposon density, the youngest intact *Gypsy* elements and the lowest solo LTR density. Thus, unlike the rest of the *Brachypodium* genome, Bd5S is gaining retrotransposons by replication and losing fewer by recombination. Syntenic regions of rice (Os4S) and sorghum (Sb6S) demonstrate maintenance of this high repeat content for ~50–70 Myr (Supplementary Fig. 19)[43]. Bd5S, Os4S and Sb6S also have the lowest proportion of collinear genes (Fig. 4a and Supplementary Fig. 19). We propose that the chromosome ancestral to Bd5S reached a tipping point in which high retrotransposon density had deleterious effects on genes.

Discussion

As the first genome sequence of a pooid grass, the *Brachypodium* genome aids genome analysis and gene identification in the large and complex genomes of wheat and barley, two other pooid grasses that are among the world's most important crops. The very high quality of the *Brachypodium* genome sequence, in combination with those from two other grass subfamilies, enabled reconstruction of chromosome evolution across a broad diversity of grasses. This analysis contributes to our understanding of grass diversification by explaining how the varying chromosome numbers found in the major grass subfamilies derive from an ancestral set of five chromosomes by nested insertions of whole chromosomes into centromeres. The relatively small genome of *Brachypodium* contains many active retroelement families, but recombination between these keeps genome expansion in check. The short arm of chromosome 5 deviates from the rest of the genome by exhibiting a trend towards genome expansion through increased retroelement numbers and disruption of gene order more typical of the larger genomes of closely related grasses.

Grass crop improvement for sustainable fuel[44] and food[45] production requires a substantial increase in research in species such as *Miscanthus*, switchgrass, wheat and cool season forage grasses. These considerations have led to the rapid adoption of *Brachypodium* as an experimental system for grass research. The similarities in gene content and gene family structure between *Brachypodium*, rice and sorghum support the value of *Brachypodium* as a functional genomics model for all grasses. The *Brachypodium* genome sequence analysis reported here is therefore an important advance towards securing sustainable supplies of food, feed and fuel from new generations of grass crops.

METHODS SUMMARY

Genome sequencing and assembly. Sanger sequencing was used to generate paired-end reads from 3 kb, 8 kb, fosmid (35 kb) and BAC (100 kb) clones to generate 9.4× coverage (Supplementary Table 1). The final assembly of 83 scaffolds covers 271.9 Mb (Supplementary Table 3). Sequence scaffolds were aligned to a genetic map to create pseudomolecules covering each chromosome (Supplementary Figs 1 and 2).

Protein-coding gene annotation. Gene models were derived from weighted consensus prediction from several *ab initio* gene finders, optimal spliced alignments of ESTs and transcript assemblies, and protein homology. Illumina transcriptome sequence was aligned to predicted genome features to validate exons, splice sites and alternatively spliced transcripts.

Repeats analysis. The MIPS ANGELA pipeline was used to integrate analyses from expert groups. LTR-STRUCT and LTR-HARVEST[46] were used for *de novo* retroelement searches.

Received 29 August; accepted 9 December 2009.

1. Somerville, C. The billion-ton biofuels vision. *Science* **312,** 1277 (2006).
2. Kellogg, E. A. Evolutionary history of the grasses. *Plant Physiol.* **125,** 1198–1205 (2001).
3. Gaut, B. S. Evolutionary dynamics of grass genomes. *New Phytol.* **154,** 15–28 (2002).
4. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436,** 793–800 (2005).
5. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457,** 551–556 (2009).
6. Wei, F. *et al.* Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet.* **3,** e123 (2007).
7. Moore, G., Devos, K. M., Wang, Z. & Gale, M. D. Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.* **5,** 737–739 (1995).
8. Salamini, F., Ozkan, H., Brandolini, A., Schafer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nature Rev. Genet.* **3,** 429–441 (2002).
9. Draper, J. *et al. Brachypodium distachyon.* A new model system for functional genomics in grasses. *Plant Physiol.* **127,** 1539–1555 (2001).
10. Vain, P. *et al.* Agrobacterium-mediated transformation of the temperate grass *Brachypodium distachyon* (genotype Bd21) for T-DNA insertional mutagenesis. *Plant Biotechnol. J.* **6,** 236–245 (2008).
11. Vogel, J. & Hill, T. High-efficiency *Agrobacterium*-mediated transformation of *Brachypodium distachyon* inbred line Bd21–3. *Plant Cell Rep.* **27,** 471–478 (2008).
12. Vogel, J. P., Garvin, D. F., Leong, O. M. & Hayden, D. M. *Agrobacterium*-mediated transformation and inbred line development in the model grass *Brachypodium distachyon*. *Plant Cell Tissue Organ Cult.* **84,** 100179–100191 (2006).
13. Filiz, E. *et al.* Molecular, morphological and cytological analysis of diverse *Brachypodium distachyon* inbred lines. *Genome* **52,** 876–890 (2009).
14. Vogel, J. P. *et al.* Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. *BMC Plant Biol.* **9,** 88 (2009).

15. Garvin, D. F. *et al.* An SSR-based genetic linkage map of the model grass *Brachypodium distachyon*. *Genome* **53,** 1–13 (2009).
16. Huo, N. *et al.* Construction and characterization of two BAC libraries from *Brachypodium distachyon*, a new model for grass genomics. *Genome* **49,** 1099–1108 (2006).
17. Huo, N. *et al.* The nuclear genome of *Brachypodium distachyon*: analysis of BAC end sequences. *Funct. Integr. Genomics* **8,** 135–147 (2008).
18. Gu, Y. Q. *et al.* A BAC-based physical map of *Brachypodium distachyon* and its comparative analysis with rice and wheat. *BMC Genomics* **10,** 496 (2009).
19. Garvin, D. F. *et al.* Development of genetic and genomic research resources for *Brachypodium distachyon*, a new model system for grass crop research. *Crop Sci.* **48,** S-69–S-84 (2008).
20. Bennett, M. D. & Leitch, I. J. Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Ann. Bot. (Lond.)* **95,** 45–90 (2005).
21. Vogel, J. P. *et al.* EST sequencing and phylogenetic analysis of the model grass *Brachypodium distachyon*. *Theor. Appl. Genet.* **113,** 186–195 (2006).
22. Rajagopalan, R., Vaucheret, H., Trejo, J. & Bartel, D. P. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis* thaliana. *Genes Dev.* **20,** 3407–3425 (2006).
23. Tanaka, T. *et al.* The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res.* **36,** D1028–D1033 (2008).
24. Fox, S., Filichkin, S. & Mockler, T. Applications of ultra-high-throughput sequencing. *Methods Mol. Biol.* **553,** 79–108 (2009).
25. Gray, J. *et al.* A recommendation for naming transcription factor proteins in the grasses. *Plant Physiol.* **149,** 4–6 (2009).
26. Vogel, J. Unique aspects of the grass cell wall. *Curr. Opin. Plant Biol.* **11,** 301–307 (2008).
27. Bennetzen, J. L. & Kellogg, E. A. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9,** 1509–1514 (1997).
28. Wicker, T. & Keller, B. Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Res.* **17,** 1072–1081 (2007).
29. Wicker, T. *et al.* Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight into the structure of diploid and hexaploid triticeae gene pools. *Plant Physiol.* **149,** 258–270 (2009).
30. Jiang, N., Bao, Z., Zhang, X., Eddy, S. R. & Wessler, S. R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431,** 569–573 (2004).
31. Morgante, M. *et al.* Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genet.* **37,** 997–1002 (2005).
32. Grass Phylogeny Working Group. Phylogeny and subfamilial classification of the grasses (Poaceae). *Ann. Mo. Bot. Gard.* **88,** 373–457 (2001).
33. Bossolini, E., Wicker, T., Knobel, P. A. & Keller, B. Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: implications for wheat genomics and grass genome annotation. *Plant J.* **49,** 704–717 (2007).
34. Charles, M. *et al.* Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of *Pooideae* and *Ehrhartoideae*, after their divergence from *Panicoideae*. *Mol. Biol. Evol.* **26,** 1651–1661 (2009).
35. Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA* **101,** 9903–9908 (2004).
36. Stein, N. *et al.* A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor. Appl. Genet.* **114,** 823–839 (2007).
37. Luo, M. C. *et al.* Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc. Natl Acad. Sci. USA* **106,** 15780–15785 (2009).
38. Qi, L. L. *et al.* A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168,** 701–712 (2004).
39. Salse, J. *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20,** 11–24 (2008).
40. Srinivasachary, Dida M. M., Gale, M. D. & Devos, K. M. Comparative analyses reveal high levels of conserved colinearity between the finger millet and rice genomes. *Theor. Appl. Genet.* **115,** 489–499 (2007).
41. Vicient, C. M., Kalendar, R. & Schulman, A. H. Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. *J. Mol. Evol.* **61,** 275–291 (2005).
42. Meyers, B. C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R. W. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* **15,** 809–834 (2003).
43. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA* **101,** 12404–12410 (2004).
44. U.S. Department of Energy Office of Science. *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda* ⟨ http://genomicscience.energy.gov/biofuels/b2bworkshop.shtml⟩ (2006).
45. Food and Agriculture Organization of the United Nations. *World Agriculture: Towards 2030/2050 Interim Report*. ⟨ http://www.fao.org/ES/esd/AT2050web.pdf⟩ (2006).
46. McCarthy, E. M. & McDonald, J. F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19,** 362–367 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the contributions of the late M. Gale, who identified the importance of conserved gene order in grass genomes. This work was mainly supported by the US Department of Energy Joint Genome Institute Community Sequencing Program project with J.P.V., D.F.G., T.C.M. and M.W.B., a BBSRC grant to M.W.B., an EU Contract Agronomics grant to M.W.B. and K.F.X.M., and GABI Barlex grant to K.F.X.M. Illumina transcriptome sequencing was supported by a DOE Plant Feedstock Genomics for Bioenergy grant and an Oregon State Agricultural Research Foundation grant to T.C.M.; small RNA research was supported by the DOE Plant Feedstock Genomics for Bioenergy grants to P.J.G. and T.C.M.; annotation was supported by a DOE Plant Feedstocks for Genomics Bioenergy grant to J.P.V. A full list of support and acknowledgements is in the Supplementary Information.

Author Information The whole-genome shotgun sequence of *Brachypodium distachyon* has been deposited at DDBJ/EMBL/GenBank under the accession ADDN00000000. (The version described in this manuscript is the first version, accession ADDN01000000). EST sequences have been deposited with dbEST (accessions 67946317–68053959) and GenBank (accessions GT758162–GT865804). The short read archive accession for RNA-seq data is SRA010177. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.P.V. (john.vogel@ars.usda.gov) or D.F.G. (david.garvin@ars.usda.gov) or T.C.M. (tmockler@cgrb.oregonstate.edu) or M.W.B. (michael.bevan@bbsrc.ac.uk).

Author Contributions See list of consortium authors below.

The International Brachypodium Initiative

Principal investigators John P. Vogel[1], David F. Garvin[2], Todd C. Mockler[3], Jeremy Schmutz[4], Dan Rokhsar[5,6], Michael W. Bevan[7]; **DNA sequencing and assembly** Kerrie Barry[5], Susan Lucas[5], Miranda Harmon-Smith[5], Kathleen Lail[5], Hope Tice[5], Jeremy Schmutz[4] (Leader), Jane Grimwood[4], Neil McKenzie[7], Michael W. Bevan[7]; **Pseudomolecule assembly and BAC end sequencing** Naxin Huo[1], Yong Q. Gu[1], Gerard R. Lazo[1], Olin D. Anderson[1], John P. Vogel[1] (Leader), Frank M. You[8], Ming-Cheng Luo[8], Jan Dvorak[8], Jonathan Wright[7], Melanie Febrer[7], Michael W. Bevan[7], Dominika Idziak[9], Robert Hasterok[9], David F. Garvin[2]; **Transcriptome sequencing and analysis** Erika Lindquist[5], Mei Wang[5], Samuel E. Fox[3], Henry D. Priest[3], Sergei A. Filichkin[3], Scott A. Givan[3], Douglas W. Bryant[3], Jeff H. Chang[3], Todd C. Mockler[3] (Leader), Haiyan Wu[10,24], Wei Wu[10], An-Ping Hsia[10], Patrick S. Schnable[10,24], Anantharaman Kalyanaraman[11], Brad Barbazuk[12], Todd P. Michael[13], Samuel P. Hazen[14], Jennifer N. Bragg[1], Debbie Laudencia-Chingcuanco[1], John P. Vogel[1], David F. Garvin[2], Yiqun Weng[15], Neil McKenzie[7], Michael W. Bevan[7]; **Gene analysis and annotation** Georg Haberer[16], Manuel Spannagl[16], Klaus Mayer[16] (Leader), Thomas Rattei[17], Therese Mitros[6], Dan Rokhsar[6], Sang-Jik Lee[18], Jocelyn K. C. Rose[18], Lukas A. Mueller[19], Thomas L. York[19]; **Repeats analysis** Thomas Wicker[20] (Leader), Jan P. Buchmann[20], Jaakko Tanskanen[21], Alan H. Schulman[21] (Leader), Heidrun Gundlach[16], Jonathan Wright[7], Michael Bevan[7], Antonio Costa de Oliveira[22], Luciano da C. Maia[22], William Belknap[1], Yong Q. Gu[1], Ning Jiang[23], Jinsheng Lai[24], Liucun Zhu[25], Jianxin Ma[25], Cheng Sun[26], Ellen Pritham[26]; **Comparative genomics** Jerome Salse[27] (Leader), Florent Murat[27], Michael Abrouk[27], Georg Haberer[16], Manuel Spannagl[16], Klaus Mayer[16], Remy Bruggmann[13], Joachim Messing[13], Frank M. You[8], Ming-Cheng Luo[8], Jan Dvorak[8]; **Small RNA analysis** Noah Fahlgren[3], Samuel E. Fox[3], Christopher M. Sullivan[3], Todd C. Mockler[3], James C. Carrington[3], Elisabeth J. Chapman[3,28], Greg D. May[29], Jixian Zhai[30], Matthias Ganssmann[30], Sai Guna Ranjan Gurazada[30], Marcelo German[30], Blake C. Meyers[30], Pamela J. Green[30] (Leader); **Manual annotation and gene family analysis** Jennifer N. Bragg[1], Ludmila Tyler[1,6], Jiajie Wu[1,8], Yong Q. Gu[1], Gerard R. Lazo[1], Debbie Laudencia-Chingcuanco[1], James Thomson[1], John P. Vogel[1] (Leader), Samuel P. Hazen[14], Shan Chen[14], Henrik V. Scheller[31], Jesper Harholt[32], Peter Ulvskov[32], Samuel E. Fox[3], Sergei A. Filichkin[3], Noah Fahlgren[3], Jeffrey A. Kimbrel[3], Jeff H. Chang[3], Christopher M. Sullivan[3], Elisabeth J. Chapman[3,27], James C. Carrington[3], Todd C. Mockler[3], Laura E. Bartley[8,31], Peijian Cao[8,31], Ki-Hong Jung[8,31]†, Manoj K Sharma[8,31], Miguel Vega-Sanchez[8,31], Pamela Ronald[8,31], Christopher D. Dardick[33], Stefanie De Bodt[34], Wim Verelst[34], Dirk Inzé[34], Maren Heese[35], Arp Schnittger[35], Xiaohan Yang[36], Udaya C. Kalluri[36], Gerald A. Tuskan[36], Zhihua Hua[37], Richard D. Vierstra[37], David F. Garvin[3], Yu Cui[24], Shuhong Ouyang[24], Qixin Sun[24], Zhiyong Liu[24], Alper Yilmaz[38], Erich Grotewold[38], Richard Sibout[39], Kian Hematy[39], Gregory Mouille[39], Herman Höfte[39], Todd Michael[13], Jérome Pelloux[40], Devin O'Connor[41], James Schnable[41], Scott Rowe[41], Frank Harmon[41], Cynthia L. Cass[42], John C. Sedbrook[42], Mary E. Byrne[7], Sean Walsh[7], Janet Higgins[7], Michael Bevan[7], Pinghua Li[19], Thomas Brutnell[19], Turgay Unver[43], Hikmet Budak[43], Harry Belcram[44], Mathieu Charles[44], Boulos Chalhoub[44], Ivan Baxter[45]

[1]USDA-ARS Western Regional Research Center, Albany, California 94710, USA. [2]USDA-ARS Plant Science Research Unit and University of Minnesota, St Paul, Minnesota 55108, USA. [3]Oregon State University, Corvallis, Oregon 97331-4501, USA. [4]HudsonAlpha Institute, Huntsville, Alabama 35806, USA. [5]US DOE Joint Genome Institute, Walnut Creek, California 94598, USA. [6]University of California Berkeley, Berkeley, California 94720, USA. [7]John Innes Centre, Norwich NR4 7UJ, UK. [8]University of California Davis, Davis, California 95616, USA. [9]University of Silesia, 40-032 Katowice, Poland. [10]Iowa State University, Ames, Iowa 50011, USA. [11]Washington State University, Pullman, Washington 99163, USA. [12]University of Florida, Gainsville, Florida 32611, USA. [13]Rutgers University, Piscataway, New Jersey 08855-0759, USA. [14]University of Massachusetts, Amherst, Massachusetts 01003-9292, USA. [15]USDA-ARS Vegetable Crops Research Unit, Horticulture Department, University of Wisconsin, Madison, Wisconsin 53706, USA. [16]Helmholtz Zentrum München, D-85764 Neuherberg, Germany. [17]Technical University München, 80333 München, Germany. [18]Cornell University, Ithaca, New York 14853, USA. [19]Boyce Thompson Institute for Plant Research, Ithaca, New York 14853-1801, USA. [20]University of Zurich, 8008 Zurich, Switzerland. [21]MTT Agrifood Research and University of Helsinki, FIN-00014 Helsinki, Finland. [22]Federal University of Pelotas, Pelotas, 96001-970, RS, Brazil. [23]Michigan State University, East Lansing, Michigan 48824, USA. [24]China Agricultural University, Beijing 10094, China. [25]Purdue University, West Lafayette, Indiana 47907, USA. [26]The University of Texas, Arlington, Arlington, Texas 76019, USA. [27]Institut National de la Recherché Agronomique UMR 1095, 63100 Clermont-Ferrand, France. [28]University of California San Diego, La Jolla, California 92093, USA. [29]National Centre for Genome Resources, Santa Fe, New Mexico 87505, USA. [30]University of Delaware, Newark, Delaware 19716, USA. [31]Joint Bioenergy Institute, Emeryville, California 94720, USA. [32]University of Copenhagen, Frederiksberg DK-1871, Denmark. [33]USDA-ARS Appalachian Fruit Research Station, Kearneysville, West Virginia 25430, USA. [34]VIB Department of Plant Systems Biology, VIB and Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, 9052 Gent, Belgium. [35]Institut de Biologie Moléculaire des Plantes du CNRS, Strasbourg 67084, France. [36]BioEnergy Science Center and Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6422, USA. [37]University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [38]The Ohio State University, Columbus, Ohio 43210, USA. [39]Institut Jean-Pierre Bourgin, UMR1318, Institut National de la Recherche Agronomique, 78026 Versailles cedex, France. [40]Université de Picardie, Amiens 80039, France. [41]Plant Gene Expression Center, University of California Berkeley, Albany, California 94710, USA. [42]Illinois State University and DOE Great Lakes Bioenergy Research Center, Normal, Illinois 61790, USA. [43]Sabanci University, Istanbul 34956, Turkey. [44]Unité de Recherche en Génomique Végétale: URGV (INRA-CNRS-UEVE), Evry 91057, France. [45]USDA-ARS/Donald Danforth Plant Science Center, St Louis, Missouri 63130, USA. †Present address: The School of Plant Molecular Systems Biotechnology, Kyung Hee University, Yongin 446-701, Korea.

Abstract

Together maize, Sorghum, rice, and wheat grass (*Poaceae*) species are the most important cereal crops cultivated worldwide.

My PhD aims to characterize dynamic evolution and organization of wheat genomes from différent species (*Triticum* and *Aegilops* genera) in relation to transposable element (TE) proliferation in their genomes (>80%), polyploidizations and synteny with other *Poaceae* species.

Little was known about the dynamics of TEs in wheat genomes. By constituting and comparing representative genomic sequences, I have characterized dynamics and differential proliferation of various TE superfamilies and families in A and B genomes of wheat, as resulting from the combinations of their insertions and deletions. Differential proliferation of TEs as well as extents of resulting rearrangements and consequences on wheat genome evolution were more precisely appreciated by analyzing for the first time haplotype variability of the A, B, S and D genomes (a total of 16 haplotypes). Thus, we compared several genotypes of the diploid and polyploid wheat species of the *Hardness* (*Ha*) locus. Mean replacement rate of the TE space, which measures sequence differences due to insertion and removal of TEs between two haplotypes, was estimated to 86% per one million year (My). This is more important than the well-documented haplotype variability found in maize. Thus, TE space is completely different between the A, B, S and D genomes that have diverged about three MYA. It was observed that TE insertions and DNA elimination by illegitimate recombination (implicating several 'tens' of kb) as well as homologous recombination between divergent haplotypes represent the main molecular basis for rapid change of the TE space.

At a longer evolutionary scale (60 My), I have compared gene conservation at the *Ha* locus region between different *Poaceae* species. The comparative genome analysis and evolutionary comparison with genes encoding grain reserve proteins of grasses suggest that an ancestral *Ha-like* gene emerged, as a new member of the *Prolamin* gene family, in a common ancestor of the *Pooideae* (wheat and *Brachypodium* from the *Triticeae* and *Brachypodieae* tribes) and *Ehrhartoideae* (rice), between 60 and 50 My, after their divergence from *Panicoideae* (Sorghum).

My results suggest important dynamic and plasticity of TE spaces in the wheat genome as compared to a relatively high conservation of genes.

Mots clés : Wheat, Evolution, Transposable elements, Comparative genomics, Haplotype variability

Résumé

Le blé (famille des *Poaceae*) constitue, avec le riz, le sorgho et le maïs, les céréales les plus cultivées au monde.

Ma thèse vise à caractériser l'évolution dynamique et l'organisation des génomes des différentes espèces du blé (genres *Triticum* et *Aegilops*) en relation avec la prolifération des éléments transposables (TEs) dans leur génome (>80%), les polyploïdisations récurrentes ainsi que la syntenie avec d'autres espèces de *Poaceae*.

Très peu de choses étaient connues sur la dynamique des TEs dans les génomes du blé. En constituant des sets de séquences génomiques représentatives, j'ai caractérisé la dynamique et la prolifération différentielle des différentes superfamilles et familles de TEs, dans les génomes A et B du blé. Elle est la résultante de l'équilibre entre leurs insertions et aussi leurs éliminations actives. L'étendue et les conséquences de cette prolifération différentielle des TEs sur le génome ont été plus précisément appréciées en analysant la variabilité haplotypique des génomes A, B, S et D (16 haplotypes au total). Nous avons ainsi comparé plusieurs génotypes des espèces diploïdes et polyploïdes du blé au niveau du locus de la dureté de la graine (*Ha : Hardness*). Le taux moyen de remplacement de l'espace TEs, mesurant les différences de séquences dues aux insertions et aux délétions entre deux haplotypes, a été ainsi estimé à 86% par million d'années (Ma) et dépasse celles bien documentées du maïs. Ainsi, l'espace TE est complètement différent entre les régions orthologues des génomes A, B, S et D qui ont divergé il y a moins de trois Ma. Les insertions des TEs mais aussi leurs éliminations par recombinaisons illégitimes de l'ADN (pouvant atteindre plusieurs dizaines de kb) ainsi que les recombinaisons génétiques entre haplotypes divergents représentent les principaux mécanismes à la base des changements rapides de l'espace TEs.

Sur une échelle d'évolution plus longue (60 Ma), j'ai analysé la conservation des gènes et l'évolution du locus (*Ha*) entre différentes espèces des *Poaceae*. J'ai pu ainsi préciser l'émergence du caractère grain tendre et des gènes *Ha,* comme nouveaux membres de la famille des gènes de *Prolamine*, dans l'ancêtre commun des *Pooideae* (blé et *Brachypodium*, de la tribu des *Triticeae* et des *Brachypodieae)* et des *Ehrhartoideae* (riz), après leur divergence des *Panicoideae* (maïs, sorgho).
Mes travaux de thèse suggèrent une importante dynamique et plasticité de l'espace TEs du blé, comparées à une relative conservation des gènes.

Mots clés : Blé, Évolution, Élément transposables, Génomique comparée, Variabilité haplotypique