université
PARIS-SACLAY

Étude des Mutations Microsatellites par des Approches Basées sur l'Utilisation de Séquenceurs de Première, Deuxième et Troisième Génération

Study of Microsatellite Mutations Using Approaches Based on First, Second, and Third-Generation Sequencers

Thèse de doctorat de l'université Paris-Saclay

École doctorale 577, Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Sciences de la vie et de la santé
Graduate School : Life Science and Health. Référent : Université d'Évry Val d'Essonne

Thèse préparée dans le Laboratoire de génomique de la Fondation Jean Dausset - Centre d'Etude du Polymorphisme Humain, sous la direction d'**Alexandre HOW KIT**, Directeur de recherche Fondation Jean Dausset - CEPH, et l'encadrement de **Jean-François DELEUZE**, Directeur Scientifique de la Fondation Jean Dausset - CEPH.

Thèse soutenue à Paris, le 27 février 2025, par

Sophie JEANJEAN

Composition du Jury

Membres du jury avec voix délibérative

Florence GONNET Professeure, Université d'Evry Paris-Saclay	Présidente
Benoit ARCANGIOLI Directeur de recherche, Institut Pasteur, Université Paris Cité	Rapporteur
Julien BARC Directeur de recherche, INSERM, Nantes Université	Rapporteur
Judith LOPES Chargée de recherche, INSERM, Sorbonne Université	Examinatrice

THESE DE DOCTORAT

NNT : 2025UPASL010

université PARIS-SACLAY | ÉCOLE DOCTORALE Structure et dynamique des systèmes vivants (SDSV)

Résumé

Titre : Etude des mutations microsatellites par des approches basées sur l'utilisation de séquenceurs de première, deuxième et troisième génération.

Mots clés : Microsatellites, Modélisation, NGS, Mutations, Système MMR

Les microsatellites sont des séquences d'ADN répétées en tandem avec un motif de 1 à 6 nucléotides, présents dans presque tous les génomes. Leur taux de mutation élevé, responsable d'une évolution rapide, leur confère un haut niveau de polymorphisme basé sur le nombre de répétitions. Ils sont donc utilisés comme marqueurs génétiques dans de nombreuses applications biomédicales. Le principal mécanisme de mutation est le glissement de la polymérase lors de la réplication, provoquant des contractions (délétions) ou des expansions (insertions) des microsatellites. Ce processus présente un biais directionnel, avec une préférence pour les délétions chez les procaryotes et les insertions chez les eucaryotes. Le système de réparation des mésappariements (MMR) corrige ces erreurs et réduit fortement le taux de mutation.

Plusieurs facteurs influencent les mutations des microsatellites. Parmi les facteurs endogènes, on trouve la composition nucléotidique, la taille et le nombre de répétitions. Les facteurs exogènes incluent la taille du microsatellite et l'efficacité du système MMR. Mon projet de thèse visait à étudier et caractériser ces mutations à l'aide de séquenceurs de première, deuxième et troisième génération. Il se divisait en trois parties principales.

Dans la première partie, j'ai contribué au développement d'une méthode d'inférence des taux de mutation, une mesure cruciale en génétique. Cette méthode permet d'identifier le modèle génétique à l'origine du glissement de la polymérase parmi 32 modèles considérés. Elle intègre des taux de délétion et d'insertion allant jusqu'à quatre sauts d'unités de répétition, avec une relation linéaire ou exponentielle entre le taux de mutation et le nombre d'unités. La méthode a été appliquée à des données expérimentales obtenues par amplification *in vitro* de microsatellites via PCR et méthode isotherme (RPA), combinées à l'analyse de fragments. Cela a révélé des différences mutationnelles significatives. Dans la deuxième partie, j'ai comparé différentes technologies de séquençage haut débit (NGS) pour déterminer la meilleure approche pour identifier la taille des allèles microsatellites. J'ai évalué diverses méthodes de préparation de librairies pour le séquençage à lecture courte avec la technologie Illumina, incluant des protocoles avec ou sans PCR (« PCR-free »), utilisant de 1 à 20 cycles PCR et incorporant ou non des codes-barres moléculaires. Un protocole sans PCR a été utilisé pour le séquençage à lecture longue avec la technologie PacBio. Mes résultats ont montré que le séquençage à lecture courte sans étape PCR est l'approche la plus fidèle. Enfin, dans la troisième partie, j'ai caractérisé les mutations des microsatellites en fonction de divers facteurs endogènes (unité répétée, nombre de répétitions) et exogènes (polymérase, température). J'ai développé un système rapporteur basé sur l'utilisation de plasmides contenant des microsatellites, combiné au séquençage capillaire et/ou NGS. Les études *in vitro* ont impliqué trois méthodes d'amplification ADN avec des températures allant de 15 à 65 °C. Les études *in vivo* ont été réalisées chez *Escherichia coli* et *Saccharomyces cerevisiae*, utilisant des souches sauvages et déficientes en MMR, cultivées à des températures de 20 à 45 °C. Ces travaux ont révélé un effet modulateur de la température sur les mutations des microsatellites.

En conclusion, ce travail a permis d'améliorer la caractérisation des mutations microsatellites grâce à des approches innovantes, mathématiques et technologiques, et a mis en évidence un nouveau facteur exogène influençant leur formation *in vitro* et *in vivo*.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Abstract

Titre : Study of microsatellite mutations using approaches based on first, second, and third-generation sequencers.

Mots clés : Microsatellites, Modelisation, NGS, Mutations, MMR System

Microsatellites are tandem repeat DNA sequences with motifs of 1 to 6 nucleotides, present in nearly all genomes. Their high mutation rate, driving rapid evolution, results in significant polymorphism based on the number of repeats. They are widely used as genetic markers in various biomedical applications. The primary mutation mechanism involves polymerase slippage during replication, leading to contractions (deletions) or expansions (insertions) of microsatellites. This process exhibits directional bias, favoring deletions in prokaryotes and insertions in eukaryotes. The mismatch repair system (MMR) corrects these errors, significantly reducing the mutation rate.

Several factors influence microsatellite mutations. Endogenous factors include nucleotide composition, motif size, and repeat number. Exogenous factors include microsatellite size and MMR efficiency. My thesis project aimed to study and characterize these mutations using first-, second-, and third-generation sequencing technologies. It was divided into three main parts.

In the first part, I contributed to the development of a method to infer mutation rates, a crucial measure in genetics. This method identifies the genetic model underlying polymerase slippage among 32 considered models. It accounts for deletion and insertion rates up to four repeat unit jumps, with linear or exponential relationships between mutation rates and repeat units. The method was applied to experimental data from *in vitro* microsatellite amplification via PCR and isothermal methods (RPA), combined with fragment analysis. This revealed significant mutational differences.

In the second part, I compared various next-generation sequencing (NGS) technologies to identify the best approach for determining the length of microsatellite alleles. I evaluated library preparation methods for short-read sequencing with Illumina technology, including protocols with or without PCR ("PCR-free"), using 1 to 20 PCR cycles and incorporating molecular barcodes. A PCR-free protocol was used for long-read sequencing with PacBio technology. My results showed that short-read sequencing without PCR is the most reliable approach.

Finally, in the third part, I characterized microsatellite mutations based on various endogenous (repeat unit, repeat number) and exogenous (polymerase, temperature) factors. I developed a reporter system using plasmids containing microsatellites combined with capillary and/or NGS sequencing. *In vitro* studies involved three DNA amplification methods across temperatures from 15 to 65 °C. *In vivo* studies were conducted in *Escherichia coli* and *Saccharomyces cerevisiae*, using wild-type and MMR-deficient strains grown at temperatures from 20 to 45 °C. These studies revealed a modulatory effect of temperature on microsatellite mutations.

In conclusion, this work improved microsatellite mutation characterization through innovative mathematical and technological approaches, highlighting a novel exogenous factor influencing their formation *in vitro* and *in vivo*.

Remerciements

Je tiens à exprimer mes remerciements au Professeure Florence GONNET qui m'a fait l'honneur d'accepter de présider mon jury de thèse. J'adresse mes profonds remerciements au Docteur Julien BARC ainsi qu'au Professeur Benoit ARCANGIOLI pour l'intérêt qu'ils ont porté à ce travail en acceptant de le rapporter. Je remercie également la Docteure Judith LOPES pour avoir accepté d'examiner mes travaux de recherche.

Je tiens également à remercier le Docteur Alexandre HOW-KIT et le Docteur Jean-François DELEUZE de m'avoir encadré et accueilli au sein de leur laboratoire. Merci de m'avoir permis de découvrir le monde captivant de la génétique des microsatellites, de m'avoir fait confiance ainsi que pour le temps et les moyens que vous m'avez accordés à la bonne réalisation de mes projets. Merci également à Alexandre de m'avoir guidé tout au long de cette thèse. Merci de m'avoir transmis tes connaissances pratiques et scientifiques, et pour ton aide et ta patience.

Je remercie la Fondation ARC et la Fondation Jean Dausset-CEPH pour avoir financé mon doctorat.

Je remercie Antoine DAUNAY et la Docteure Lise HARDY pour leur aide technique, leur soutien et leur compagnie depuis le début de cette aventure.

Je remercie particulièrement Nicolas TESSIER et Ilef HCHAICHI, mes "co-doctorants", et Julie DE CHEFDEBIEN pour votre soutien pendant les heures passées à la paillasse ou à l'écriture, le partage de sucreries et d'aneries, et votre amitié.

Je remercie Emmanuel TUBACHER, Docteur Yimin SHEN et Mourad SAHBATOU pour leur expertise et leur aide technique pour les analyses bioinformatiques.

Je remercie Schayma BEN MARZOUGUI, Rojo NOFIDIANTSOA et Nasreddine RAJOUA pour leur expertise et leur aide technique sur le développement de MSmrc32.

Je remercie l'ensemble des membres de l'équipe du CRB : Docteure Hélène BLANCHÉ, Laetitia GRESSIN, Jean-Christophe BEAUDOIN, Nandhinie ZANEGUY ainsi que tous les autres.

Je remercie en particulier la Docteure Hélène LE BUANEC, David BERGERAT et la Docteure Marine MERANDET pour leur soutien et leur aide lorsque notre matériel nous a abandonné. Vive les allers-retours pour faire autoclaver des milieux !

Je remercie l'équipe de l'iGenSeq du Docteur Yannick MARIE de l'Institut du Cerveau pour nous avoir permis d'utiliser leur plateforme de ddPCR et leur aide technique.

Je remercie le Docteur Antonio ALBERDI pour son aide technique et son soutien avec l'utilisation du NextSeq 500.

Je remercie l'équipe de la plateforme de séquençage et de génotypage du CNRGH et tout particulièrement Anne BOLAND et Christian DAVIAUD pour leur expertise et leur aide technique.

Je remercie la Docteure Shufang RENAULT et le Docteur Guy-Franck RICHARD qui m'ont fait l'honneur de faire partie de mon comité de suivi individuel de thèse.

Un immense merci à mes amis pour m'avoir soutenu du début à la fin malgré les frontières des disciplines, pour les multiples soirées, vacances, jeux, discussions sur Discord et détentes en tout genre. Vous êtes malheureusement bien trop nombreux pour pouvoir tous vous citer ici, mais mon attention

se porte particulièrement sur le noyau dur que sont Jonathan, John, Romane, Camille, Axel (et Sasha), Victorien, Bastien, Guillaume, Aku, Mathieu, Amandine et Christophe.

Enfin, un grand merci à mes parents Solange et Michel et à mon frère Teiki, pour leur soutien indéfectible.

Table des matières

Résumé 3
Abstract 4
Remerciements 5
Table des matières 7
Tables des Figures 15
Tables des Tableaux 22
Abréviations 23
Synthèse Bibliographique 25
GENERALITES SUR L'ADN ET LE GENOME 25
1. L'ADN REPETE EN TANDEMS 27
1.1. Satellites 28
1.2. Minisatellites 29
1.3. Microsatellites 29
2. LES MICROSATELLITES ET LEURS PARTICULARITES 31
2.1. Distributions dans les génomes 31
2.1.1. Régions non-codantes 33
2.1.2. Régions codantes : cibles 33
2.2. Mutations dans les microsatellites 34
2.2.1. Taux de mutation 34
2.2.2. Type de mutations : Mutations ponctuelles 34
2.2.3. Mécanisme de mutations 35
2.2.3.1. Déamination des cytosines méthylées 35
2.2.3.2. Recombinaison homologue 36
2.2.3.3. Glissement de Polymérase 37
2.2.3.4. Cas particulier 39
2.3. Cycle de vie du microsatellite 39
2.3.1. Conception 41
2.3.2. Naissance 42
2.3.3. Croissance 42
2.3.4. Sénescence / Mort / Renaissance 42
3. LE SYSTEME MMR 43

3.1. Son rôle et son fonctionnement 43
3.1.1. Chez les procaryotes 44
3.1.2. Chez les eucaryotes 44
3.2. Conséquences de la déficience du MMR et des mutations microsatellites 45
3.2.1. Instabilité des microsatellites (MSI) 45
3.2.2. Cancers MMRd - Déficience constitutionnelle du MMR 47
3.2.3. Maladies neuro-dégénératives 48
4. FACTEURS INFLUENÇANT LES MUTATIONS DES MICROSATELLITES 49
4.1. Intrinsèques 49
4.1.1. Nombre de répétitions 49
4.1.2. Taille du motif 50
4.1.3. Composition du motif 51
4.2. Extrinsèques 51
4.2.1. Systèmes de corrections 51
4.2.2. Organismes vivants (procaryotes/eucaryotes) 51
4.3. Directionnalité des mutations 52
5. LES METHODES D'ANALYSES DES MICROSATELLITES 53
5.1. Génotypage & mutations microsatellites 53
5.2. Méthodes bas débit 54
5.2.1. PCR et électrophorèse en gel d'agarose 54
5.2.2. RFLP 54
5.2.3. AFLP 54
5.2.4. PCR avec amorces fluorescentes et électrophorèse capillaire 55
5.2.5. LT-RPA et électrophorèse capillaire 55
5.3. Méthodes de séquençage 56
5.3.1. Séquençage de première génération - Sanger 56
5.3.2. Séquençage de deuxième génération - Illumina 57
5.3.3. Séquençage de troisième génération 59
5.3.3.1. PacBio - SMRT 59
5.3.3.2. Nanopore 60
5.4. Analyse et algorithmes bio-informatiques 61
5.4.1. Identification de microsatellites dans le génome 62
5.4.2. Détection d'indels 62
5.4.3. Génotypage 62
5.4.4. Génotypage quantitatif 63
5.4.5. Alignement 64
6. METHODES D'ANALYSE DES TAUX DE MUTATIONS 64
6.1. *In vitro* 64

6.1.1. HSV-tk mutation assay *64*
6.1.2. Fréquences alléliques après PCR *65*
6.2. *In vivo / Ex vivo* 65
6.3. *In populo* 66
7. LES APPLICATIONS DES MICROSATELLITES 67
7.1. La cartographie génétique 67
7.2. Génétique évolutive 67
7.3. Génétique des populations et études de la biodiversité 68
7.4. Empreinte génétique 69
7.4.1. Application non humaine *69*
7.4.1.1. Animaux : certification de races (pedigree) et Parentalité *69*
7.4.1.2. Végétaux : Sélection assistée par marqueur, Plantes, certification de graines *69*
7.4.2. Sciences médico-légales *70*
7.5. Maladies humaines 70
8. INTRODUCTION DU SUJET DE THESE 71
8.1. Organisme d'accueil 71
8.1.1. Histoire de la Fondation Jean Dausset - CEPH *71*
8.1.2. Laboratoire de Génomique *72*
8.2. Objectifs 73
8.2.1. Projet 1 : Inférence des taux de mutation des microsatellites à partir des fréquences alléliques obtenues à partir d'approches d'amplification d'ADN (non) isothermes en utilisant trente-deux modèles génétiques *74*
8.2.2. Projet 2 : Analyse détaillée des approches de séquençage de deuxième et troisième génération pour la détermination précise de la longueur des répétitions courtes en tandem et des homopolymères *75*
8.2.3. Projet 3 : Influence de la température sur les mutations des microsatellites *75*

Travaux de recherche 77

1. PROJET 1 : INFERENCE DES TAUX DE MUTATION DES MICROSATELLITES A PARTIR DES FREQUENCES ALLELIQUES OBTENUES A PARTIR D'APPROCHES D'AMPLIFICATION D'ADN (NON) ISOTHERMES EN UTILISANT TRENTE-DEUX MODELES GENETIQUES 77
1.1. Résultats 77
1.1.1. Modélisation mathématique de l'amplification de l'ADN et du processus de mutation des microsatellites au cours de la PCR *77*

1.1.2. Estimation des taux de mutation, des paramètres α/β et γ par le critère de minimisation de l'erreur quadratique moyenne à l'aide de l'algorithme de recuit simulé couplé à une recherche par grille 80
1.1.3. Evaluation des performances d'inférence des taux de mutation des microsatellites 81
1.1.4. Évaluation de MSmrc32 pour l'identification du modèle génétique sous-jacent au mécanisme de mutation des microsatellites 84
1.1.5. Mise en œuvre de trois modes pour l'inférence des taux de mutation des microsatellites compatibles avec les méthodes d'amplification d'ADN basées sur la PCR et les méthodes isothermes 85
1.1.6. Evaluation de l'algorithme sur des données expérimentales de PCR et RPA 86
1.2. Discussion 97
2. PROJET 2 : ANALYSE DETAILLEE DES APPROCHES DE SEQUENÇAGE DE DEUXIEME ET TROISIEME GENERATION POUR LA DETERMINATION PRECISE DE LA LONGUEUR DES REPETITIONS COURTES EN TANDEM ET DES HOMOPOLYMERES. 101
2.1. Résultats 101
2.1.1. Description des approches NGS utilisées dans notre étude comparative 101
2.1.2. Impact des cycles de PCR dans la préparation de librairies short-read Illumina sur l'estimation de la longueur des allèles microsatellites 103
2.1.3. Détermination de la longueur des allèles des microsatellites par la correction d'erreur à partir d'UMI simple 107
2.1.4. Amélioration de l'estimation de la longueur des allèles des microsatellites par correction d'erreur combinée à partir des données de séquençage en duplex 111
2.1.5. Précision et amélioration de l'identification de la longueur des allèles des microsatellites par séquençage en long-read PacBio 114
2.2. Discussion 116
3. PROJET 3 : INFLUENCE DE LA TEMPERATURE SUR LES MUTATIONS DES MICROSATELLITES 119
3.1. Résultats 119
3.1.1. Expérimentations en contexte in vitro 119
3.1.1.1. WGA 121
3.1.1.2. RPA 125
3.1.1.3. PCR 137
3.1.2. Expérimentations en contexte in vivo 149
3.1.2.1. Dans un organisme procaryote : Escherichia coli 149
3.1.2.1.1. Optimisations du protocole de transformation et de culture 149
3.1.2.1.2. Doublement cellulaire et doublement de molécule plasmidique 151

3.1.2.1.3. Impact de la température sur les mutations microsatellites 153
3.1.2.2. Dans un organisme eucaryote : Saccharomyces cerevisiae *154*
3.1.2.2.1. Etudes plasmidiques 154
3.1.2.2.1.1. Optimisation du protocole de transformation et de culture 155
3.1.2.2.1.2. Doublement cellulaire et doublement de molécule plasmidique 157
3.1.2.2.1.3. Impact de la température sur les mutations microsatellites 159
3.1.2.2.2. Etude des mutations microsatellites dans le génome de S. cerevisiae 161
3.1.2.2.2.1. Optimisation du protocole de culture 161
3.1.2.2.2.2. Doublement de cellule et de molécule 162
3.1.2.2.2.3. Impact de la température sur les microsatellites du génome de la levure 162
3.2. Discussion 165
4. DISCUSSION GENERALE, CONCLUSION ET PERSPECTIVES 169

Matériels et Méthodes 177

1. MATERIEL BIOLOGIQUE 177
1.1. Souches Bactériennes 177
1.2. Souches de Levures 178
1.3. Vecteurs et inserts 179
1.3.1. Stratégie *179*
1.3.2. Vecteurs *180*
1.3.2.1. Bactérien *180*
1.3.2.2. Levures *180*
1.3.3. Insert microsatellites *181*
1.3.4. Oligonucléotides *181*
2. MILIEUX DE CULTURE 184
2.1. Milieux de culture bactériens 184
2.2. Milieux de culture pour la levure 184
2.3. Quantification des bactéries et levures en culture liquide par mesure de la densité optique 185
2.4. Sélection de clones 186
2.4.1. Isolement *186*
2.4.2. Strike *187*
2.5. Culture liquide de bactéries et levures non transformées en vue de la compétence 188
2.6. Compétence 188
2.6.1. Bactéries E. coli *189*
2.6.2. Levures S. cerevisiae *189*
2.7. Transformation 189
2.7.1. Bactéries (Choc Thermique) *189*
2.7.2. Levures *190*

2.8. Détermination du taux de transformation et du nombre de doublement cellulaire 191
2.8.1. Taux de transformation 191
2.8.2. Nombre de doublement cellulaire 191
3. MANIPULATION DES ACIDES NUCLEIQUES 192
3.1. Extraction d'ADN 192
3.1.1. ADN plasmidique de bactérie 192
3.1.2. ADN génomique des levures 193
3.1.3. ADN génomique et plasmidique de levures 194
3.1.4. ADN plasmidique des levures 195
3.2. Méthode d'amplification 196
3.2.1. PCR 196
3.2.1.1. PCR classique 196
3.2.1.2. PCR en temps réel (rtPCR) 197
3.2.1.2.1. *Le mix réactionnel* 197
3.2.1.2.2. *Le cyclage PCR* 197
3.2.1.2.3. *La courbe de fusion* 198
3.2.1.2.4. *Efficacité* 198
3.2.1.3. PCR quantitative 199
3.2.1.4. PCR digitale par gouttelettes (ddPCR) 200
3.2.2. RPA 200
3.2.2.1. TwistAmp® Basic Kit 201
3.2.2.2. TwistAmp® Liquid Basic Kit 202
3.2.2.3. Digestion contre polyadénilation 202
3.2.3. WGA 202
3.3. Méthode de dosage 203
3.3.1. NanoDrop 203
3.3.2. Qubit 203
3.4. Méthodes de purification de l'ADN 204
3.4.1. Billes magnétiques 204
3.4.2. QIAquick PCR purification kit (Qiagen) 205
3.4.3. Précipitation à l'éthanol avec de l'acétate de sodium 206
3.5. Méthode d'analyse de l'ADN basée sur l'électrophorèse 206
3.5.1. Electrophorèse Horizontale 206
3.5.1.1. Gel d'agarose 206
3.5.1.2. BluePippin 207
3.5.1.3. TapeStation 207
3.5.2. Electrophorèse Capillaire 208
3.5.2.1. Bioanalyzer 208
3.5.2.2. Fragment Analyzer 5200 209
3.5.3. SeqStudio 210

4. METHODES DE SEQUENÇAGE....211
4.1. Séquençage de seconde génération....211
4.1.1. Préparation des librairies....211
4.1.1.1. Digestion / Fragmentation....211
4.1.1.1.1. Site – spécifique (SSDP)....211
4.1.1.1.2. Fragmentation aléatoire (RFP)....212
4.1.1.1.2.1. Enzymatique....212
4.1.1.1.2.2. Mécanique (xGen PRISM)....212
4.1.1.2. Ligation des adaptateurs / indexes....213
4.1.1.2.1. QIAseq 1-step Amplicon Library (QIAGEN®)....213
4.1.1.2.2. QIAseq FX DNA Library Kit (Qiagen)....213
4.1.1.2.3. xGen™ cfDNA & FFPE DNA Library Prep v2 MC Kit (Integrated DNA Technologies)....214
4.1.2. Technologie Illumina en lecture courte – short read....215
4.1.2.1. Iseq 100....216
4.1.2.2. NextSeq 500....216
4.1.2.3. NovaSeq 6000....217
4.2. Séquençage de troisième génération....218
4.2.1. Préparation des libraires....218
4.2.1.1. Linéarisation des plasmides....218
4.2.1.2. SMRTbell prep kit 3.0 (Pacific Biosciences)....218
4.2.2. Technologie PacBio Sequel II....218
5. BIOINFORMATIQUE....219
5.1. Projet 2....219
5.1.1. Analyses des short-reads Illumina....219
5.1.1.1. Analyse standard de la longueur des allèles de microsatellites....219
5.1.1.2. Correction des erreurs de PCR et de séquençage des longueurs d'allèles de microsatellites avec un UMI unique....219
5.1.1.3. Correction des erreurs de PCR et de séquençage des longueurs d'allèles de microsatellites avec un UMI double....220
5.1.2. Analyses des long-reads PacBio....222
5.1.2.1. Analyse standard de la distribution des longueurs d'allèles de microsatellites 222
5.1.2.2. Analyse interne des données de subreads brutes pour la distribution des longueurs d'allèles de microsatellites....222
5.2. Projet 3....223
5.2.1. R1R2....223
5.2.2. BASIC....224
5.2.3. Whole Genome Sequencing de Levure....224
6. ANALYSES STATISTIQUES ET REPRESENTATIONS GRAPHIQUES....225
6.1. Schématisation....225
6.2. Représentation graphique....225

6.3. Statistiques 225
7. INTELLIGENCE ARTIFICIELLE 225
Liste des communications scientifiques 227
1. CHAPITRE PUBLIE 227
2. ARTICLE SOUMIS 227
3. ARTICLE EN FINALISATION 227
4. POSTERS 227
Bibliographie 229

Tables des Figures

Figure 1 : Frise chronologique représentant les grandes étapes de l'histoire de la génétique. 25

Figure 2 : Représentation de l'organisation dans le génome humain 26

Figure 3 : L'histoire des microsatellites 30

Figure 4 : Les microsatellites chez Homo sapiens, Escherichia coli et Saccharomyces cerevisiae. 32

Figure 5 : Mécanismes par lesquels les microsatellites peuvent modifier l'expression de gènes 33

Figure 6 : Mécanisme des mutations ponctuelles 35

Figure 7 : Processus de déamination des cytosines méthylées 36

Figure 8 : Crossing-over inégal 37

Figure 9 : Le glissement de la polymérase lors de la réplication d'une séquence répétée 37

Figure 10 : Représentation schématique sous forme de cycle de vie dynamique d'un locus microsatellite (par exemple (CA)n) 40

Figure 11 : Représentation schématique du système de réparation des mésappariements de bases de l'ADN ou système MMR (mismatch repair) 44

Figure 12 : Répartition des séquences ainsi que des mutations répétées et non-répétées dans les tumeurs MSI chez l'humain 46

Figure 13 : Les taux de mutation des souches germinales pour les STR de longueur et de taille différentes 50

Figure 14 : Taux de mutation pour 2 747 répétitions courtes en tandem uniques génotypées, avec des longueurs de motif allant de 1 pb (mononucléotide) à 6 pb (hexanucléotide) 50

Figure 15 : Efficacité de la correction des mésappariements (pourcentage) en fonction de la taille de l'unité de répétition du microsatellite (en paires de bases) 51

Figure 16 : Représentation des différents types de microsatellites présents à travers trois organismes vivants 52

Figure 17 : Exemples de génotypage de microsatellites et de détection de l'instabilité des microsatellites (MSI) utilisant LT-RPA et PCR combinés avec électrophorèse capillaire 55

Figure 18 : Représentation schématique du principe de séquençage par la méthode de Sanger 56

Figure 19 : Représentation des étapes du développement du séquençage tout au long de l'histoire de la biologie moléculaire 57

Figure 20 : Représentation du principe de séquençage Illumina 58

Figure 21 : Représentation du principe de séquençage de la technologie PacBio 60

Figure 22 : Représentation schématique du principe du séquençage par nanopore 61

Figure 23 : Représentation schématique du principe du HSV-tk mutation assay 65

Figure 24 : Organisation et relations entre les différentes objectifs et les projets de la thèse. 74

Figure 25 : Représentation schématique des modèles génétiques sous-jacents au glissement de la polymérase 78

Figure 26 : Représentation graphique des 32 types différents de modèles génétiques utilisés pour les mécanismes de mutation des microsatellites basés sur le glissement de la polymérase 79

Figure 27 : Effet de l'allèle de référence et des variations du taux de mutation des microsatellites sur la performance d'inférence de notre algorithme avec quatre modèles génétiques..... 82

Figure 28 : Effet des coefficients linéaires (α) et exponentiels (β) sur la performance d'inférence de notre algorithme avec quatre modèles génétiques..... 83

Figure 29 : Effet du nombre de cycles PCR et des ensembles de données sur les performances d'inférence de notre algorithme avec quatre modèles génétiques..... 84

Figure 30 : Évaluation de la capacité de notre algorithme à identifier le modèle génétique sous-jacent parmi les 32 disponibles 85

Figure 31 : Evaluation des différents modes d'inférence des taux de mutations des microsatellites, incluant un mode d'efficacité variable, un mode d'efficacité fixe et un mode de doublement de molécule 86

Figure 32 : Séquences de microsatellites utilisées dans notre étude. Séquençage Sanger des huit séquences de microsatellites insérées dans le vecteur pUC57-Mini. 87

Figure 33 : Conditions d'amplification de PCR en temps réel réalisées dans notre étude utilisant différents types de microsatellites à des quantités variables d'ADN matrice, également avec différents nombres de cycles PCR 87

Figure 34 : Fréquences alléliques de microsatellites obtenues par PCR en utilisant des microsatellites avec des répétitions de mono- et di-nucléotides comme matrice 88

Figure 35 : Les valeurs de Ct, calculées à partir du maximum de la seconde dérivée des courbes d'amplification PCR, indiquent approximativement le cycle auquel une diminution de l'efficacité, ainsi que la transition de la phase exponentielle à la phase linéaire, sont observées au cours de l'amplification PCR 89

Figure 36 : Fréquences alléliques de microsatellites obtenues à partir des expériences RPA utilisant des microsatellites mononucléotidiques (A) et dinucléotidiques (B) comme matrices (105 copies d'ADN). 89

Figure 37 : Précision de la quantification de l'ADN à l'aide du Qubit dsDNA HS Kit. 90

Figure 38 : Taux de mutation inférés à partir des données expérimentales de PCR et RPA..... 93

Figure 39 : Résultat du séquençage SANGER des 9 microsatellites clonés en plasmide pUC57-Mini pour le séquençage NGS..... 101

Figure 40 : Workflow des expériences NGS réalisées dans notre étude..... 102

Figure 41 : Taux d'erreur de séquençage obtenus à partir des bibliothèques sans PCR de type I, en utilisant le séquençage *short reads* avec les instruments iSeq100 et NextSeq 500 104

Figure 42 : Effet des cycles de PCR inclus dans la préparation des bibliothèques sur la longueur des allèles de microsatellites obtenue après séquençage en short-read Illumina utilisant la digestion SSDP 105

Figure 43 : Comparaison des performances des instruments Illumina iSeq 100 et NextSeq 500 pour la détermination précise de la longueur des neuf microsatellites étudiés en utilisant la digestion SSDP comme modèle 106

Figure 44 : Comparaison des performances de la détermination précise de la longueur des neuf microsatellites étudiés à partir de bibliothèques sans PCR de type I et du séquençage short-read Illumina NextSeq 500, en utilisant les digestions SSDP et RFP comme modèles. 106

Figure 45 : Comparaison des différentes stratégies utilisées pour la correction d'erreur par UMI afin de déterminer avec précision la longueur des microsatellites 108

Figure 46 : Impact de la correction d'erreur par UMI sur les données de séquençage en short-read Illumina pour une détermination précise de la longueur des microsatellites..... 109

Figure 47 : Influence du nombre de reads par groupe d'UMI sur la précision de l'identification de la longueur des allèles microsatellites après correction d'erreur par UMI 110

Figure 48 : Fréquences des allèles originaux obtenues sur les données de la bibliothèque de type III après correction d'erreur par UMI standard incluant les readss consensuelles αβ+βα, αβ et βα 111

Figure 49 : Impact de la correction d'erreur dual-UMI du séquençage en duplex sur les données de séquençage short-read Illumina pour une détermination précise de la longueur des microsatellites . 112

Figure 50 : Comptages des *reads*, groupes UMI, et consensus combinés obtenus par séquençage en «short-read» Illumina des bibliothèques de type III 113

Figure 51 : Analyse 'brute' des données à partir du séquençage long-Read PacBio (Bibliothèques de Type IV) en utilisant différentes stratégies..... 114

Figure 52 : Taux d'erreur de séquençage obtenus à partir de bibliothèques sans PCR de type IV, en utilisant le séquençage «long-read» avec le Sequel II 115

Figure 53 : Détermination précise de la longueur des microsatellites à partir du séquençage long-Read PacBio (Bibliothèques de Type IV) en utilisant différentes stratégies 116

Figure 54 : Echelle des températures couvertes par les différentes méthodes d'amplification utilisées ainsi que les méthodes d'analyse utilisées. 120

Figure 55 : Représentation schématique de l'effet de la soustraction du bruit de fond sur la répartition des fréquences alléliques..... 120

Figure 56 : Représentation schématique du processus d'amplification d'un plasmide par WGA 121

Figure 57 : Nombre de doublement de molécule après amplification d'un pool de 6 microsatellites par WGA en quadruplicats à différentes températures allant de 15 à 37,5 °C. 122

Figure 58 : FMDM par rapport à la température obtenue par NGS sur WGA après soustraction du bruit de fond..... 123

Figure 59 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et di-nucléotidiques $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$ de 15°C à 37,5°C obtenu par NGS sur WGA après soustraction du bruit de fond..... 124

Figure 60 : Représentation schématique du processus d'amplification par LT-RPA 125

Figure 61 : Représentation graphique du nombre de doublement de molécule en fonction de la température selon la méthode d'analyse utilisée (Electrophorèse capillaire ou NGS) pour tous les microsatellites et marqueurs étudiés. 126

Figure 62 : FMDM par rapport à la température obtenues par NGS sur RPA après soustraction du bruit de fond..... 127

Figure 63 : FMDM par rapport à la température obtenues par électrophorèse capillaire sur RPA 128

Figure 64 : FMDM par rapport à la température obtenues par NGS sur RPA des marqueurs mono-nucléotidiques NR24, CAT25 et BAT26..... 129

Figure 65 : FMDM par rapport à la température obtenues par électrophorèse capillaire sur RPA des marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 130

Figure 66 : FMDM par rapport à la température obtenues par NGS sur RPA des marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ 131

Figure 67 : FMDM par rapport à la température obtenues par électrophorèse capillaire sur RPA des marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ 132

Figure 68 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et les marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 de 25°C à 45°C obtenu par NGS sur RPA avec soustraction du bruit de fond pour les plasmides 133

Figure 69 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et les marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 de 25°C à 45°C obtenu par électrophorèse capillaire sur RPA sans soustraction du bruit de fond pour les plasmides afin d'uniformisé les données comparées 134

Figure 70 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites di-nucléotidiques AC15, AC20 et AC24 et les marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ de 25°C à 45°C obtenues par NGS sur RPA avec soustraction du bruit de fond pour les plasmides 135

Figure 71 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites di-nucléotidiques AC15, AC20 et AC24 et les marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ de 25°C à 45°C obtenues par électrophorèse capillaire sur RPA sans soustraction du bruit de fond pour les plasmides afin d'uniformisé les données comparées 136

Figure 72 : Représentation schématique du processus d'amplification par PCR en 2 étape 137

Figure 73 : Représentation graphique du nombre de doublement de molécule ou de la Ct (Crossing threshold) en fonction de la température selon la méthode d'analyse utilisée (Electrophorèse capillaire ou NGS) pour tous les microsatellites et marqueurs étudiés 138

Figure 74 : FMDM par rapport à la température obtenue par NGS sur PCR avec soustraction du bruit de fond 139

Figure 75 : FMDM par rapport à la température obtenue par électrophorèse capillaire sur PCR 140

Figure 76 : FMDM par rapport à la température obtenue par NGS sur PCR des marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 141

Figure 77 : FMDM par rapport à la température obtenue par électrophorèse capillaire sur PCR des marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 142

Figure 78 : FMDM par rapport à la température obtenue par NGS sur PCR des marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ 143

Figure 79 : FMDM par rapport à la température obtenue par électrophorèse capillaire sur PCR des marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ 144

Figure 80 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et les marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 de 45°C à 65°C obtenu par NGS sur PCR avec soustraction du bruit de fond pour les plasmides 145

Figure 81 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et les marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 de 45°C à 65°C obtenues par électrophorèse capillaire sur PCR sans soustraction du bruit de fond pour les plasmides afin d'uniformisé les données comparées 146

Figure 82 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites di-nucléotidiques $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$ et les marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ de 45°C à 65°C obtenues par NGS sur PCR avec soustraction du bruit de fond pour les plasmides..... 147

Figure 83 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites di-nucléotidiques $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$ et les marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ de 45°C à 65°C obtenues par électrophorèse capillaire sur PCR sans soustraction du bruit de fond pour les plasmides afin d'uniformisé les données comparées 148

Figure 84 : Workflow Production Bactéries 149

Figure 85 : Cinétique de détermination de la D.O. d'arrêt des cultures de la souche WT et du mutant mutS de E.coli pour une production optimale en restant dans la phase exponentielle..... 151

Figure 86 : Evolution de la quantité des bactéries transformées durant le processus de production... 152

Figure 87 : Evolution de la représentation des plasmides au sein des bactéries transformées..... 153

Figure 88 : FMDM par rapport à la température de culture des différentes souches étudiées..... 154

Figure 89 : Workflow Production Levures Plasmides 155

Figure 90 : Cinétique de détermination de la D.O. d'arrêt des cultures de la souches WT, et des mutant mlh1 et msh2 de S. cerevisiae pour une production optimale en restant dans la phase exponentielle 156

Figure 91 : Evolution de la quantité des levures transformées durant le processus de production 157

Figure 92 : Evolution de la représentation des plasmides au sein des levures transformées 158

Figure 93 : Evolution de la quantité des bactéries transformées durant le processus de production... 159

Figure 94 : FMDM par rapport à la température de culture des différentes souches étudiées..... 160

Figure 95 : Workflow Production Levures Génomiques..... 161

Figure 96 : Cinétique de détermination de la D.O. d'arrêt des cultures de la souches WT, et des mutant mlh1 et msh2 de S. cerevisiae pour une production optimale en restant dans la phase exponentielle 162

Figure 97 : Evolution de la quantité des levures durant le processus de production 162

Figure 98 : Contrôle Qualité du séquence WGS de la levure 163

Figure 99 : Représentation en diagramme de la représentation des différentes catégories de microsatellites présent dans le génome de la levure a travers les trois souches étudié (WT, mlh1 et msh2) 164

Figure 100 : Evaluation de la proportionnalité de représentation des différentes catégories de microsatellites pour chaque souche utilisée 165

Figure 101 : Conception des amorces et construction des mutants à délétion d'un seul gène 177

Figure 102 : Stratégie de délétion des souches. Schéma du fournisseur : SnapGene 179

Figure 103 : Représentation schématique de la construction complète (plasmide et insert) pour les expérimentations in vitro et in vivo bactérienne 180

Figure 104 : Plan du vecteur pUC57-Mini 180

Figure 105 : Plan du vecteur pESC-URA avant modification 181

Figure 106 : Courbe de perte de linéarité entre densité optique et concentration bactérienne 186

Figure 107 : Représentation schématique d'un isolement par la méthode des cadrans. 187

Figure 108 : Représentation schématique de strike de 8 colonies sur une boîte de petri 187

Figure 109 : Schéma représentatif de la procédure du Kit QIAGEN® Plasmid Plus Midi Kit™ 193

Figure 110 : Schéma représentatif de la procédure du Kit Quick-DNA™ Fungal/Bacterial Miniprep Kit 194

Figure 111 : Représentation schématique du processus d'amplification par PCR 196

Figure 112 : Représentation schématique du processus de fluorescence par dye de la rtPCR donnant la qPCR 198

Figure 113 : Représentation graphique d'une courbe de fusion d'une amplification spécifique. 198

Figure 114 : Principe du kit de quantification de librairies KAPA (tiré de leur brochure) 199

Figure 115 : Représentation schématique du Workflow de la ddPCR 200

Figure 116 : Représentation schématique du processus d'amplification par LT-RPA 201

Figure 117 : Représentation des ratios de sélection de tailles par les billes AMPure™ 204

Figure 118 : Représentation des ratios de sélection de tailles par les billes SPRIselect™ 205

Figure 119 : Représentation des résultats sur TapeStation par gel et électrophorégramme virtuel 208

Figure 120 : Représentation des résultats sur Bioanalyzer par électrophorégramme et gel 209

Figure 121 : Protocole de préparation des Gels et solution de conditionnement 210

Figure 122 : Principe de double Workflow du SeqStudio Genetic Analyzer réalisant un séquençage SANGER et une analyse de fragment, le tout par électrophorèse capillaire. 211

Figure 123 : Représentation schématique de la méthode de construction de librairie du xGen cfDNA & FFPE DNA Library Prep v2 MC Kit 215

Figure 124 : Représentation schématique de la chimie du système monocolore 216

Figure 125 : Représentation schématique de la chimie du système bicolore 217

Figure 126 : Workflow des différentes analyses bioinformatiques réalisées à partir des données de séquençage en *short reads* Illumina 221

Figure 127 : Workflow des différentes analyses bioinformatiques réalisées à partir des données de séquençage en long-reads de PacBio (librairies de type IV). 223

Tables des Tableaux

Tableau 1 : Classification et paramètres des ADN répétés en tandems 28

Tableau 2 : Les différents types de motifs microsatellites 30

Tableau 3 : Répartition proportionnelle de la présence des microsatellites à travers les différentes régions du génome humain 32

Tableau 4 : Tableau récapitulatif des caractéristiques des différents mécanismes de mutations chez les microsatellites 35

Tableau 5 : Caractéristiques des différents modèles de glissement de la polymérase 38

Tableau 6 : Mécanisme de mutation majoritaire pour chaque phase du cycle de vie d'un microsatellite 41

Tableau 7 : Paramètres du système de modélisation des 32 modèles. 80

Tableau 8 : Identification des modèles génétiques les mieux ajustés pour chaque microsatellite mononucléotidique selon les trois modes d'inférence des taux de mutation des microsatellites et les types d'expériences de PCR 91

Tableau 9 : Identification des modèles génétiques les mieux ajustés pour chaque microsatellite mononucléotidique selon les trois modes d'inférence des taux de mutation des microsatellites et les types d'expériences de PCR. 92

Tableau 10 : Statistiques de régression linéaire et exponentielle calculées pour les taux de mutation des microsatellites à répétitions mononucléotidiques obtenues à partir des expériences de PCR. 94

Tableau 11 : Statistiques de régression linéaire et exponentielle calculées pour les taux de mutation des microsatellites à répétitions di-nucléotidiques obtenues à partir des expériences de PCR. 95

Tableau 12 : Identification du meilleur modèle génétique pour chaque microsatellite mono- et di-nucléotidiques par RPA utilisant l'inférence du taux de mutations basé sur le mode de doublement de molécule. 96

Tableau 13 : Statistiques des régressions linéaires et exponentielles calculées pour les taux de mutations microsatellites mono- et di-nucléotidiques obtenus par PCR. 96

Tableau 14 : Les microsatellites au sein du génome de Saccharomyces cerevisiae 163

Tableau 15 : Catégorisation et répartition des microsatellites conservés pour l'étude 164

Tableau 16 : Séquences des insert microsatellites 182

Tableau 17 : Informations concernant les oligonucléotides. 183

Tableau 18 : Détails de la composition des différents milieux de culture pour bactérie. 184

Tableau 19 : Détails de la composition des différents milieux de culture pour levure. 185

Tableau 20 : Composition des Tampons Buffer SCE et Lysis Buffer. 195

Abréviations

A	**A**dénosine
ADN	**A**cide **D**ésoxiribo**N**ucléique
AFLP	***A**mplified **F**ragment **L**ength **P**olymorphims*
Amp	**Amp**icilline
ARN	**A**cide **R**ibo**N**ucléique
AS	**A**nalyses de **S**équence
ATM	***A**taxia **T**elangiectasia **M**utated*
AV	**A**nalyses de **V**ariabilité
C	**C**ytosine
CCS	***C**hildhood **C**ancer **S**yndrome*
CDB	**C**assures **D**ouble-**B**rin
CEPH	**C**entre d'**E**tude du **P**olymorphisme **H**umain
CMMRD	***C**onstitutional **M**is**M**atch **R**epair **D**eficiency*
CoLoN	***C**olon tumours or/and **L**eukaemia/**L**ymphoma or/and **N**eurofibromatosis*
ddNTPs	**d**i**d**ésoxyribo**N**ucléosides **T**ri**P**hosphates
ddPCR	***d**roplet **d**igital **PCR*** (PCR digitale en gouttelette)
dNTPs	**d**ésoxyribo**N**ucléosides **T**ri**P**hosphates
E. coli	***E**scherichia **coli***
G	**G**uanine
Gen	**Gén**éticine
GSM	***G**eneral **S**tepwise **M**odel*
HRM	***H**igh **R**esolution **M**elting* (Fusion à haute résolution)
HTT	**H**un**T**ing**T**in
IAM	***I**nfinite **A**llele **M**odel*
KAM	***K**-**A**llele **M**odel*
Kana	**Kana**mycine
LB	***L**ysogenic **B**roth* (milieu Lysogenique)
LINE	***L**ong **I**nterspersed **N**uclear **E**lements* (Eléments nucléaires dispersés longs)
LT-RPA	***L**ow **T**emperature **RPA*** (RPA à basse température)
Lys	**Lys**ine
MMR	***M**is**M**atch **R**epair* (Réparation de mésappariement)
MSDB	***M**icro**S**atellite **D**ata**B**ase*
MSI	***M**icro**S**attelite **I**nstability* (Instabilité des microsatellites)
MSI-H	**MSI-H**igh (MSI élevé)
NaHR	***N**on-**A**llelique **H**omologus **R**ecombinaison* (Recombinaison homologue non allélique)
NF1	**N**euro**F**ibromatose **1**
NGS	***N**ext-**G**eneration **S**equencing* (Séquençage de nouvelle génération)
ORF	***O**pen **R**eading **F**rame* (Cadre de lecture ouvert)
PCNA	***P**roliferating **C**ell **N**uclear **A**ntigen*
PCR	***P**olymerase **C**hain **R**eaction* (Polymérisation en chaîne)
PSMD	***P**an-**S**pecies **M**icrosatellite **D**atabase*
QC	***Q**uality **C**heck* (Contrôle qualité)
qPCR	**q**uantitative **PCR**
RFC	***R**eplication **F**actor **C***
RFLP	***R**estriction **F**ragment **L**ength **P**olymorphism* (Polymorphisme de longueur des fragments de restriction)
RPA	***R**ecombinase **P**olymerase **A**mplification* (Amplification par polymérase à recombinase)
rtPCR	***r**eal-**t**ime **PCR*** (PCR en temps réel)

S. cerevisiae	***S**accharomyces **cerevisiae***
S. Do	***S**ynthetic **D**rop-**o**ut* (Milieu synthétique sans complément)
SDSA	***S**ynthesis-**D**ependant **S**trand **A**nnealing*
SINE	***S**hort **I**nterspersed **N**uclear **E**lements* (Eléments nucléaires dispersés courts)
SMM	***S**tepwise **M**utation **M**odel*
SNP	***S**ingle **N**ucleotide **P**olymorphism* (Polymorphisme d'un seul nucléotide)
SSM	***S**lipped **S**trand **M**ispairing*
SSR	***S**imple **S**equence **R**epeat*
STR	***S**hort **T**andem **R**epeat* (Répétition en tandems courts)
T	**T**hymine
TPM	***T**wo-**P**hase **M**odel*
Ura	**U**racile
UTR	***U**n**T**ranslated **R**egions*
VNTR	***V**ariable **N**umber of **T**andem **R**epeats* (Répétition en tandems à nombre variable)
WGA	***W**hole **G**enome **A**mplification* (Amplification pangénomique)
YNB	***Y**east **N**itrogen **B**ase* (Base azotée pour levure)
YPD	***Y**east extract **P**eptone **D**extrose* (Extrait de levure, peptone, dextrose)

Unités et facteurs de multiplications

°C	degré **C**elsius
µ	micro (10^{-6})
Da	**Da**lton
DO	**D**ensité **O**ptique
g	**g**ramme
h	**h**eure
k	**k**ilo (10^{3})
L	**L**itre
m	**m**ètre
m	**m**illi (10^{-3})
M	**M**olaire (mole/litre)
min	**min**ute
n	**n**ano (10^{-9})
p	**p**ico (10^{-12})
p/v	**p**oids**/v**olume
pb	**p**aires de **b**ases
pH	**p**otentiel **H**ydrogène
sec	**sec**onde
U	**U**nité enzymatique
v/v	**v**olume**/v**olume

Généralités sur l'ADN et le génome

La volonté de compréhension de l'hérédité et de la variabilité des êtres vivants a poussé l'humanité à créer une nouvelle discipline fondamentale qui a transformé notre compréhension de la vie depuis le milieu du XIXe siècle : la génétique. Cette révolution scientifique commença avec la définition de ses lois par Gregor Mendel en 1865, suivie par la découverte des liens entre gènes et chromosomes par Thomas Hunt en 1910. Un grand pas fut franchi avec la découverte de l'ADN comme support de l'information génétique par Avery, MacLeod et McCarty en 1944 (Avery et al., 1944) suivie, grâce au cliché 51 de Rosalind Franklin, par la découverte de sa structure en double hélice par Watson et Crick en 1953 (Watson and Crick, 1953), permettant ainsi de révéler la nature chimique et, par la suite, le mécanisme de réplication du matériel génétique. À partir de ce moment, la génétique connut des avancées spectaculaires, en grande partie grâce à l'essor des technologies de séquençage de l'ADN et à la capacité croissante de décoder et de manipuler l'information génétique (Figure 1).

Avec l'arrivée de la génomique dans les années 1990, la recherche a franchi une nouvelle étape (A. J. Jeffreys et al., 1985). Là où la génétique classique se concentrait principalement sur les gènes individuels et leur transmission, la génomique vise à une compréhension globale de l'ensemble du matériel génétique d'un organisme, son génome. Le séquençage du génome humain, achevé en 2003 après 13 ans (International Human Genome Sequencing Consortium, 2004), a marqué un tournant décisif, offrant pour la première fois, un aperçu complet des 3 milliards de paires de bases qui composent notre ADN et des milliers de gènes qui codent pour les protéines, les régulateurs et autres éléments essentiels à la vie (Figure 1).

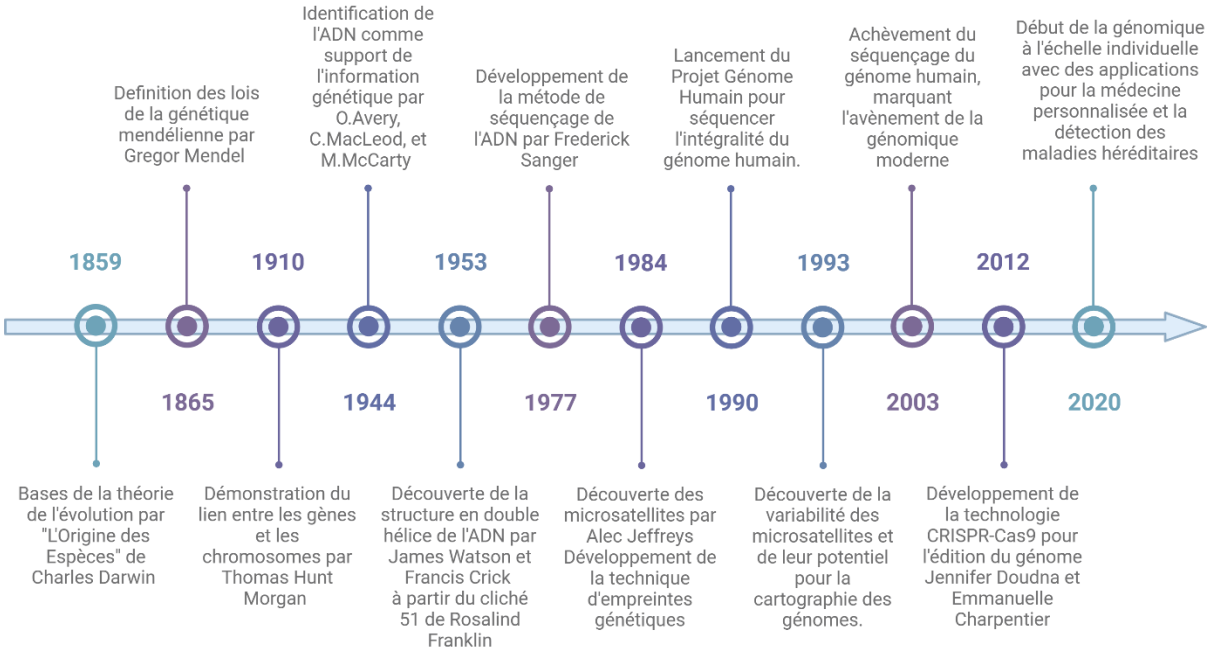

Figure 1 : Frise chronologique représentant les grandes étapes de l'histoire de la génétique (Avery et al., 1944; Darwins, 1859; International Human Genome Sequencing Consortium, 2004; A. J. Jeffreys et al., 1985; Mendel, 1866; Sanger et al., 1977; Watson and Crick, 1953).

Chez les organismes eucaryotes, comme les plantes, les animaux et les humains, le génome peut être divisé en deux : le génome nucléaire et le génome extranucléaire provenant d'autres organites (mitochondrie et chloroplaste). Chez l'homme, on trouve d'une part le génome mitochondrial composé d'un seul chromosome circulaire appelé Chromosome M, présent en de nombreuses copies au sein de la mitochondrie, et contenant 37 gènes codant pour autant de protéines spécifiquement utilisées dans les processus du métabolisme énergétique de la cellule. D'autre part, on trouve le génome nucléaire organisé en chromosomes linéaires situés dans le noyau de la cellule. Toutefois, chez les procaryotes, comme les bactéries, il est généralement constitué d'une molécule d'ADN circulaire. L'exploration de la structure et de l'organisation du génome humain a permis de révéler sa complexité étonnante. L'organisation du génome peut être subdivisée en trois grandes catégories : l'ADN codant, l'ADN non-codant et l'ADN extra-génique (Figure 2).

L'ADN codant comprend les séquences qui sont transcrites en ARN messager (ARNm) et traduites en protéines. Ces séquences représentent environ 1 à 2 % du génome humain. L'ADN non-codant, quant à lui, est constitué des séquences qui ne codent pas pour des protéines, mais qui remplissent des rôles variés dans la régulation de l'expression génique, la structuration du génome et la protection contre l'instabilité génétique (Figure 2) (Aganezov et al., 2022; International Human Genome Sequencing Consortium, 2004; Nurk et al., 2022).

Enfin, l'ADN extra-génique inclut les séquences situées en dehors des régions géniques, qui ne participent pas directement à la transcription des gènes mais sont néanmoins essentielles pour la structure et la fonction du génome. Une part importante de l'ADN extra-génique est constituée de séquences répétées, qui peuvent être classées en séquences répétées intercalées et séquences répétées en tandems (Figure 2). Les séquences répétées intercalées, telles que les éléments transposables (comme les SINEs et LINEs), sont des fragments d'ADN capables de se déplacer dans le génome par un mécanisme de retro-transposition. Ces éléments transposables représentent une proportion non négligeable du génome (environ 15%) et contribuent à l'évolution génétique tout en jouant un rôle potentiel dans certaines mutations et maladies génétiques (Aganezov et al., 2022; Nurk et al., 2022).

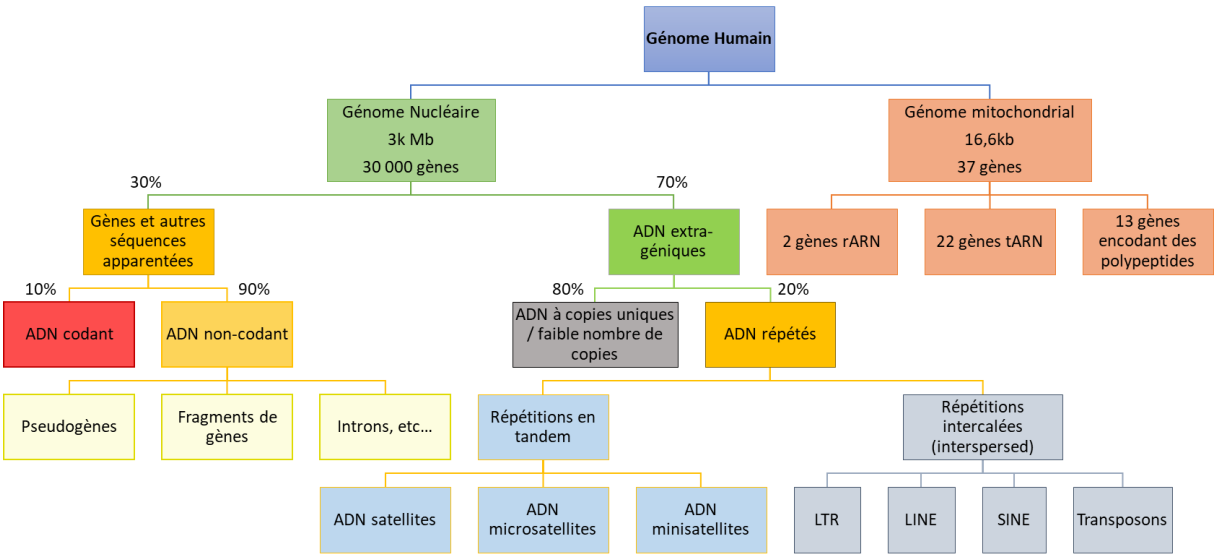

Figure 2 : Représentation de l'organisation dans le génome humain (Aganezov et al., 2022; International Human Genome Sequencing Consortium, 2004; Nurk et al., 2022).

En parallèle, dans ce vaste paysage génomique, progressivement l'attention s'est portée vers les séquences répétées en tandems qui, bien que ne participant pas directement à la production de protéines, jouent un rôle crucial dans la régulation et la structure du génome. Parmi ces éléments, les microsatellites se sont avérés être des outils de choix pour la recherche génétique et génomique. Ces séquences courtes et répétitives, composées de 1 à 6 nucléotides, sont réparties au sein des génomes, dans les parties codantes comme non-codantes de l'ADN, mais majoritairement dans les parties extra-géniques et présentent une grande variabilité d'un individu à l'autre, ce qui leur confère une valeur inestimable pour l'étude de la diversité génétique.

Les microsatellites sont devenus des marqueurs génétiques privilégiés pour diverses applications, notamment en génétique des populations, en identification individuelle, en cartographie génétique, et en étude de maladies, héréditaires ou non. Leur fort polymorphisme, une facilité de génotypage par des méthodes basées sur la PCR (*Polymerase Chain Reaction*) et la capacité à fournir des informations sur la variabilité génétique les rendent particulièrement utiles, notamment le suivi de populations animales et végétales, la détection de maladies génétiques, ainsi que pour les études de paternité et d'empreintes génétiques dans le domaine médico-légal.

Cependant, les microsatellites offrent également une fenêtre unique sur les processus évolutifs et les mécanismes de mutation qui sous-tendent la dynamique du génome. La compréhension de leur origine, de leur maintien et de leur évolution dans les génomes demeure un domaine de recherche d'intérêt, posant des questions fondamentales sur les forces sélectives ou neutres qui façonnent la structure et la diversité du matériel génétique.

Ainsi, cette synthèse bibliographique se propose d'explorer en profondeur le rôle des microsatellites dans la génomique moderne, en examinant les mécanismes moléculaires qui gouvernent leur dynamique et leur évolution, et en soulignant leur importance en tant qu'outils d'analyse génétique. Elle se basera principalement sur des connaissances concernant le génome humain, sauf lorsque précisé.

1. L'ADN répété en tandems

Les séquences répétées en tandems, souvent appelées « séquences satellites », ont été mises en lumière grâce aux travaux de Britten et Kohne en 1968 (Britten and Kohne, 1968). Ils ont démontré que ces séquences représentaient une proportion significative du génome et se caractérisaient par une vitesse de réassociation rapide distincte, indicative de leur nature répétitive. Les séquences étant constituées de motifs répétés en tandems de taille variée, un classement en trois catégories a été établi en fonction de la longueur du motif de répétition : satellites, minisatellites, microsatellites. Les satellites présentent des motifs longs, souvent de plusieurs centaines de paires de bases (parfois appelé macrosatellites), les minisatellites ont des motifs de 10 à 100 paires de bases (aussi appelés VNTR, *Variable Number of Tandem Repeats*), et les microsatellites se caractérisent par des motifs courts de 1 à 6 paires de bases (aussi appelé STR, *Short Tandem Repeat*, ou même SSR, *Simple Sequence Repeat*) (Tableau 1). Toutefois, la délimitation de tailles entre chaque catégorie peut varier en fonction des publications et des standards choisis (Dantas et al., 2024; Richard et al., 2008).

Dans cette étude, nous avons choisi de considérer les paramètres présentés précédemment (Tableau 1).

Tableau 1 : Classification et paramètres des ADN répétés en tandems (Dantas et al., 2024; Richard et al., 2008)

Nom	Taille du motif	Longueur totale	Distribution dans le génome	Caractéristique principale
Satellites (Macrosatellites / Mégasatellites)	Supérieur à 100 pb	Plusieurs kilobases à mégabases	Régions centromériques et péri-centromériques Régions spécifiques, moins courant	Impliqués dans la structure chromosomique, ségrégation lors de la division cellulaire, potentiellement impliqué dans la régulation des gènes
Minisatellites (VNTRs)	10 à 100 pb	De 500 pb à plusieurs kilobases	Principalement dans les régions (sub-) télomériques	Polymorphiques Utilisés pour les empreintes génétiques
Microsatellites	1 à 6 pb (allant jusqu'à 10 pb)	Quelques dizaines à centaines de pb	De manière dispersée dans le génome	Polymorphiques Utilisés comme marqueurs génétiques dans la plupart des applications bio-médicales et médico-légales Variation interindividuelle fréquente

1.1. Satellites

La dénomination « satellite » provient de la première mise en évidence de bandes mineures (ou satellites) lors de centrifugation en gradient de densité d'ADN génomique (John and Gabor Miklos, 1979). Lors de ces centrifugations, trois bandes peuvent être identifiées, représentant différentes familles de séquences satellites. Les satellites I, riches en AT, possédant des motifs allant de 25 à 50 pb et un nombre de répétions allant jusqu'à plus de 1000, sont essentiellement présents dans l'hétérochromatine de la plupart des chromosomes. Les satellites II et III, quant à eux, possèdent une unité de base de 5 à 25 pb et sont présents dans la plupart des chromosomes. Enfin, l'ADN satellite alpha, constituant principal de l'hétérochromatine centromérique, possède un motif de 171 pb. Ces blocs évoluent au cours du temps, où le nombre de répétitions peut augmenter ou diminuer au sein d'une même lignée d'individus. Ce biais de séquences serait dû à la présence ou non de protéines de liaisons de l'ADN. En effet, un chromosome portant une répétition élargie se maintiendrait seulement si le bloc satellite était en présence d'une protéine de liaison de l'ADN permettant une stabilisation pendant la mitose (Csink and Henikoff, 1998).

Initialement considérées comme de l'ADN « poubelle » (« *junk* » DNA) en raison de leur non-transcription apparente, il a été révélé ces dernières années qu'elles possédaient des fonctions essentielles dans le génome. Les séquences satellites jouent un rôle clé dans la structure chromosomique, notamment en contribuant à la formation et à la fonction des centromères, nécessaires à la ségrégation des chromosomes pendant la division cellulaire (Shatskikh et al., 2020). De plus, elles participent à la réponse au stress cellulaire, influencent la signalisation immunitaire, et leur évolution rapide pourrait faciliter l'adaptation aux environnements changeants et même la spéciation (Shatskikh et al., 2020).

Selon les conventions choisies, le terme « satellite » peut être utilisé comme nom générique d'une famille de répétitions en tandems contenant les minisatellites et les microsatellites ainsi que les macrosatellites ou/et les mégasatellites décrivant alors des sous-catégories de ce que nous avons décrit comme les satellites. Les macrosatellites semblent alors définir des satellites spéciaux, pouvant posséder un cadre de lecture ouvert (ORF) et ainsi produire de l'ARN codant pour des protéines. Ils sont souvent spécifiques à un ou deux chromosomes seulement (Dantas et al., 2024; Rich et al., 2014). Quant à eux,

les mégasatellites semblent être des satellites avec des motifs supérieurs à 100 pb, répétés au moins 3 fois et insérés dans des gènes codants. Toutefois, leurs différences avec les macrosatellites restent floues car ils semblent également posséder une ORF (Descorps-Declère and Richard, 2022; Tekaia et al., 2013; Thierry et al., 2008).

1.2. Minisatellites

Les minisatellites sont des régions du génome possédant des séquences en tandems d'un motif répété allant de 6 jusqu'à 100 pb (Alec J. Jeffreys et al., 1985). Le nombre de loci de minisatellites au sein du génome humain est estimé entre 1000 et 2000 par génome haploïde (Jeffreys, 1987; Vergnaud and Denoeud, 2000). Ils ne sont pas répartis uniformément à travers le génome mais sont principalement localisés aux extrémités des chromosomes, impliquant de ce fait une limitation de leurs utilisations dans certaines analyses de lignage (Armour and Jeffreys, 1992). De plus, les minisatellites semblent être des points chauds pour les mutations de la lignée germinale, présentant des taux de mutation bien plus élevés que dans la lignée somatique. Ces mutations résultent principalement de mécanismes de conversion génique, souvent accompagnés de recombinaisons inégales, qui altèrent la longueur et la structure des motifs répétés. Les mutations seraient influencées par des facteurs tels que la longueur des répétitions, la séquence spécifique des motifs et les contextes génomiques environnants (Jeffreys et al., 1994). Par exemple, dans le cas de cinq loci hautement instables, le taux de mutation était lié à leur hétérozygotie observée, ce qui semblait indiquer que les changements étaient sélectivement neutres (Ramel, 1997). Enfin, ces séquences répétitives contribuent à la diversité génétique en générant des variations rapides, ce qui peut influencer l'évolution et l'adaptation des populations liées à des réarrangements chromosomiques. L'étude des minisatellites est appliquée dans les domaines de la génétique des populations, la médecine légale et la compréhension des mécanismes fondamentaux de l'évolution du génome et de certaines maladies humaines (Vergnaud and Denoeud, 2000).

1.3. Microsatellites

Les microsatellites (aussi appelé STRs ou SSRs), sont des satellites courts et simples avec un haut taux de mutations (Ananda et al., 2013). Leur motif est composé généralement de 1 à 6 nucléotides répétés en tandem (Baptiste et al., 2013; Tautz, 1989). Ils sont classés dans différentes familles désignées par la composition de leurs motifs allant de mono-nucléotide (ex : $(A)_n$) à hexa-nucléotide (ex : $(ACTCGA)_n$). De plus, les répétitions des motifs peuvent être d'un seul tenant, donnant des microsatellites parfaits, ou interrompus ou composés, donnant des microsatellites imparfaits. Lorsque le nombre de répétions varie selon les individus et au cours des générations, on parle de polymorphisme de taille pour les microsatellites simples et composés (Tableau 2). On peut également parler de « polymorphisme de structure » qui différencie les microsatellites dits « parfaits » comparés à ceux ayant subi un changement d'une base en une autre au sein d'un motif, nommé alors « imparfait » (Tableau 2) (Balaresque, 2007; Brazda et al., 2020).

Comptant pour environ 3% du génome humain, les microsatellites sont à la fois présents dans les régions intra-géniques (introns, exons, non traduites) et inter-géniques. Il y a environ 17% des cadres de lecture qui contiennent des microsatellites. Ceux dans les régions intra-géniques peuvent jouer un rôle de régulation de l'expression de gène et la fonction de protéine (Gemayel et al., 2010; Kelkar et al., 2010). Toutefois, les microsatellites présents dans les exons peuvent influencer l'épissage alternatif, notamment par des mécanismes tels que *l'exon skipping*. Ce phénomène, induit par des variations dans les microsatellites, peut altérer la reconnaissance des sites d'épissage par les complexes spliceosomes,

conduisant à l'exclusion d'un exon donné. Ces modifications de l'épissage peuvent aboutir à des protéines tronquées ou modifiées, influençant ainsi divers processus cellulaires et pathologiques, y compris les maladies génétiques et les troubles cardiaques héréditaires (Collura et al., 2014; Dorard et al., 2011; Jonchère et al., 2024).

Tableau 2 : Les différents types de motifs microsatellites (Balaresque, 2007; Brazda et al., 2020)

Microsatellites Simples	Parfaits	CCCCCCCCCCCCC	$(C)_{13}$
		CTCTCTCTCTCTCTCTCT	$(CT)_9$
	Imparfaits Interrompus	CCCCGCCCCCCCC	$(C)_4$ G $(C)_8$
		CTCTCTCTCTCACTCTCT	$(CT)_5$ CA $(CT)_3$
Microsatellites Composés	Parfaits	CCCCCCAAAAAAAAA	$(C)_6$ $(A)_9$
		GTGTGTGTGTGTGTACACACAC	$(GT)_7$ $(AC)_4$
	Imparfaits Interrompus	CCCCCCAAAAATAAA	$(C)_6$ $(A)_5$ T $(A)_3$
		GTGTGTATGTGTGTACACACAC	$(GT)_3$ AT $(GT)_3$ $(AC)_4$

Bien que les microsatellites aient initialement été identifiés comme des séquences d'ADN répétées dans les années 1970 et nommés par le terme "microsatellite" en 1981, ce n'est que dans les années 1980 que leur rôle comme marqueurs génétiques a été exploré. Cela a été suivi par une utilisation croissante de ces séquences dans de nombreuses applications en biologie, en médecine et en science forensique, qui sont toujours d'actualité aujourd'hui. En 1989, Litt et Luty (Litt and Luty, 1989), ainsi que Weber et May (Weber and May, 1989), démontrent leur utilité pour la cartographie génétique et l'identification individuelle, marquant ainsi le début de leur utilisation en génétique des populations, en médecine légale et en recherche biomédicale. Dans les années 1990, les mécanismes mutationnels des microsatellites, principalement le glissement de l'ADN polymérase, sont élucidés, soulignant leur instabilité génétique (Levinson and Gutman, 1987a) (Figure 3).

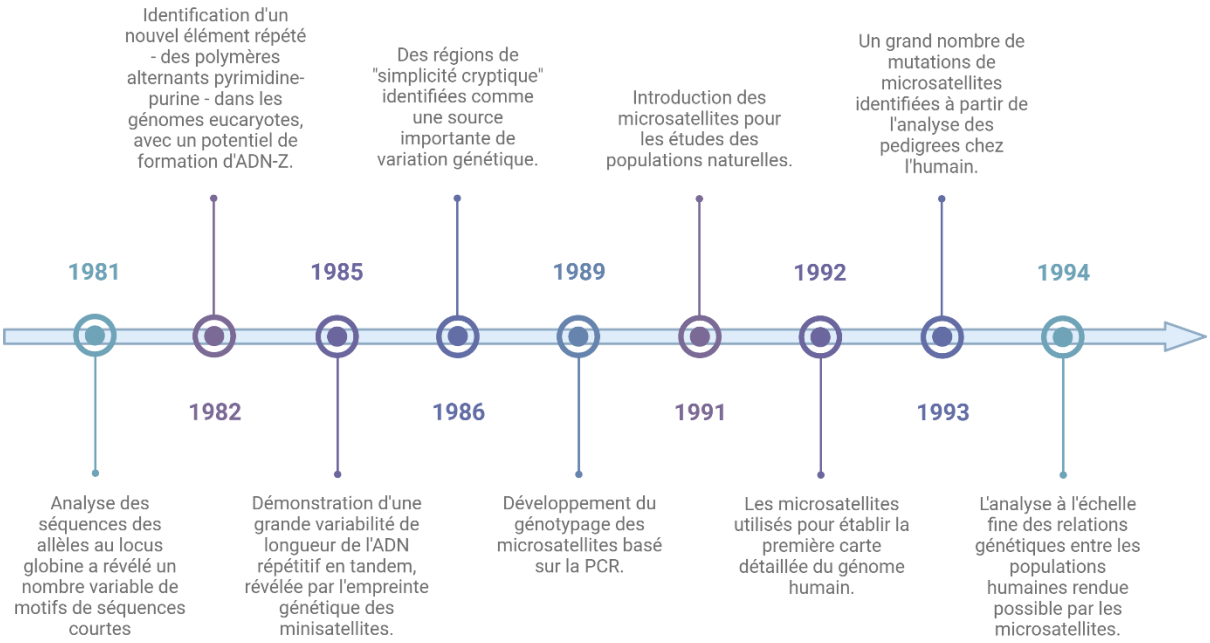

Figure 3 : L'histoire des microsatellites (Levinson and Gutman, 1987a; Litt and Luty, 1989; Weber and May, 1989)

Les années 2000 voient l'émergence des technologies de séquençage à haut débit, permettant une exploration plus approfondie de leur distribution et de leur évolution dans les génomes (Cao et al., 2015; Dean et al., 2002). Dans les années 2010, des études révèlent leurs rôles potentiels dans la régulation de l'expression génique et leur implication dans des maladies génétiques et le cancer (Hause et al., 2016; Lyon et al., 2010; Shah et al., 2010). Aujourd'hui, les microsatellites continuent d'être importants dans la recherche biomédicale, notamment pour l'édition génétique et comme biomarqueurs, montrant leur pertinence continue dans la compréhension des mécanismes évolutifs et pathologiques du génome.

2. Les microsatellites et leurs particularités

Du fait que les microsatellites sont localisés dans tout le génome et extrêmement polymorphes, ils sont utilisés en tant que marqueurs dans de nombreuses applications en génétique. En fonction de leurs positions, de leurs motifs et des interactions avec les séquences environnantes, des différences ont été observées dans l'action des mécanismes mutationnels sur ces séquences.

2.1. Distributions dans les génomes

Les microsatellites sont présents chez tous les organismes vivants. Toutefois, leur densité est variable selon les règnes (Coenye and Vandamme, 2005; Ellegren, 2004; Trivedi, 2006). Au fil des années, l'analyse de la distribution des microsatellites a été facilitée par le séquençage de génomes entiers. Aujourd'hui, même s'ils ne sont pas tous assemblés ou avec une couverture optimale, plus de 42 000 génomes eucaryotes et plus de 2 millions de génomes bactériens sont accessibles sur le site du NCBI (http://www.ncbi.nlm.nih.gov/Genomes/index.html). De plus, des banques de données se sont développées, par exemple le site de la PSMD (*Pan-Species Microsatellite Database*) (https://big.cdu.edu.cn/psmd/) regroupant 678 millions de microsatellites parfaits et 43 millions de microsatellites composés à partir de 18 408 organismes, ou le site de MSDB (*MicroSatellite DataBase*) (https://data.ccmb.res.in/msdb/) regroupant 5,29 milliards de microsatellites composés à partir de 46 122 génomes provenant de 40 204 espèces recouvrant pratiquement toutes les espèces avec des génomes disponibles.

De manière générale, le génome des procaryotes semble présenter peu de microsatellites comparé au génome des eucaryotes (Li et al., 2004). Comme on peut le voir chez *Escherichia coli* (Figure 4), la densité en microsatellites est faible mais il y a aussi une plus forte proportion de microsatellites mononucléotidiques. Ceci est possiblement lié à la proportion importante de régions codantes dans leur génome, restreignant ainsi le développement des microsatellites (Coenye and Vandamme, 2005).

Concernant les eucaryotes, il y a une différence entre les eucaryotes unicellulaires et les eucaryotes supérieurs (Bagshaw, 2017; Li et al., 2004). Le génome de *Saccharomyces cerevisiae* maintient une densité et fréquence faibles en STRs par rapport au génome humain. De plus, il est notable qu'il y a en proportion une plus faible diversité de distribution des STRs chez l'homme que chez la levure, probablement liée à une fonctionnalisation potentielle de nombreux microsatellites, puisque 52 % de ses microsatellites sont dans ses exons, 47 % sont intergéniques et moins de 1% sont dans des introns (Avvaru et al., 2020).

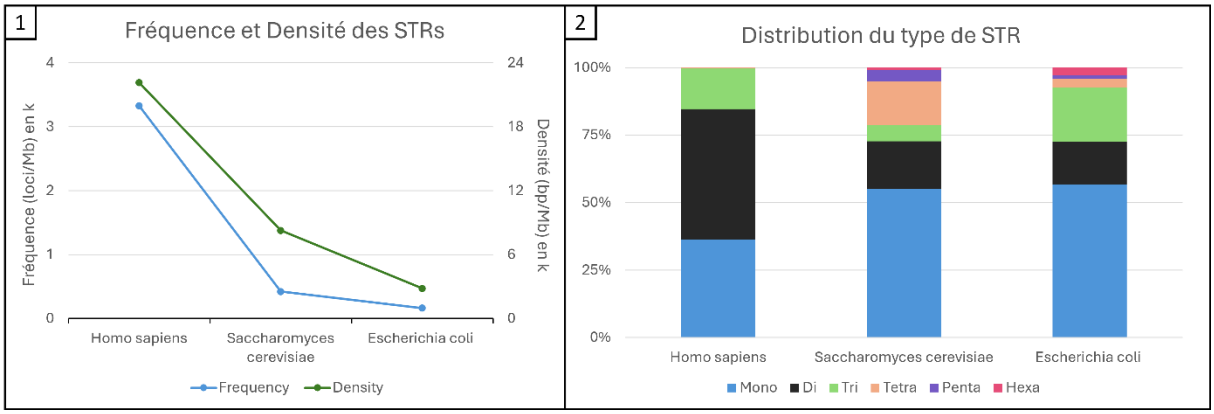

Figure 4 : Les microsatellites chez Homo sapiens, Escherichia coli et Saccharomyces cerevisiae. 1. Représentation graphique de la fréquence et de la densité des STR en fonction de l'organisme hôte. 2. Représentation graphique de la distribution des types de STR en fonction de l'organisme hôte. Données provenant de PSMD.

Toutefois, il est à noter que les données issues d'analyses et de publications différentes peuvent varier substantiellement selon la méthode de détection utilisée et les microsatellites étudiés. Certaines études se basent uniquement sur des microsatellites parfaits pour la distribution, d'autres uniquement sur des di-nucléotides pour calculer les densités. Le choix du chromosome est également important, car certaines études se focalisent sur tous les microsatellites d'un chromosome là où d'autres se basent sur plusieurs chromosomes mais restreignent les microsatellites analysés (Katti et al., 2001; Subramanian et al., 2003). (Pour la Figure 4, toutes les données disponibles ont été utilisées).

Dans le génome humain, les microsatellites sont distribués de manière relativement aléatoire, mais leur densité et leur localisation peuvent varier en fonction du type de microsatellite et des régions génomiques spécifiques. Ils sont présents dans l'ensemble du génome, aussi bien dans les régions codantes (exons) que non-codantes (introns, régions intergéniques, régions promotrices, etc…), bien qu'ils ont une plus grande propension à se trouver dans ces dernières. Ils se retrouvent le plus fréquemment dans les introns, les régions régulatrices, riches en hétérochromatine ou intergéniques (Tableau 3) (Ellegren, 2004; Gelfand et al., 2007; Gemayel et al., 2010; Kozlowski et al., 2010; Li et al., 2002; Subramanian et al., 2003).

Tableau 3 : Répartition proportionnelle de la présence des microsatellites à travers les différentes régions du génome humain (Ellegren, 2004; Gelfand et al., 2007; Gemayel et al., 2010; Kozlowski et al., 2010; Li et al., 2002; Subramanian et al., 2003).

Région du génome	Proportion de microsatellites (%)	Nombre de microsatellites par Mb
Exons	2 - 5 %	10 - 20
Introns	40 - 50 %	500 – 700
Régions intergéniques	30 - 40 %	300 – 500
Régions promotrices	5 - 10 %	50 – 150
Régions d'enhancer	5 - 10 %	50 – 100
Régions UTR (5' et 3')	3 – 7 %	15 – 45
Régions riches en GC	1 - 5 %	5 - 50
Régions riches en hétérochromatine	< 1 %	< 5

2.1.1. Régions non-codantes

La majorité des microsatellites sont présents dans l'ADN non codant, principalement dans les séquences intergéniques et les introns. Bien que l'évolution de nombreux microsatellites soit considérée comme neutre, signifiant que le degré de polymorphisme est proportionnel au taux de mutation sous-jacent, certaines études suggèrent que leur fréquence et leur distribution dans le génome peuvent refléter non seulement le processus de mutation, mais aussi des pressions sélectives spécifiques liées à des fonctions régulatrices (Ellegren, 2004; Li et al., 2004). Ces séquences n'influencent pas directement la fonction des protéines, mais elles peuvent jouer un rôle important dans la régulation de l'expression génique et l'organisation de la chromatine. Elles sont connues pour moduler la transcription, l'épissage des ARN et la traduction, en agissant comme des éléments cis-régulateurs qui peuvent affecter l'activité des gènes voisins (Gemayel et al., 2010; Sawaya et al., 2013). Par exemple, des microsatellites situés dans les régions promotrices ou « *enhancers* » peuvent modifier la liaison des facteurs de transcription, tandis que ceux présents dans les régions 5' ou 3' UTR peuvent influencer la stabilité et la traduction des ARN messagers (Gymrek et al., 2016; Kozlowski et al., 2010).

2.1.2. Régions codantes : cibles

Dans les régions codantes de l'ADN, la sélection négative, ou purifiante, des mutations décale le cadre de lecture, entravant efficacement l'expansion des microsatellites. Cependant, les répétitions tri- et hexa-nucléotidiques sont relativement fréquentes dans les régions codantes, car elles ne modifient pas le cadre, et donc sont moins susceptibles d'entraîner des mutations nuisibles. Toutefois, des contraintes de longueur, influant sur la stabilité et la fonction des protéines, peuvent impacter la distribution et la taille de ces répétitions dans les régions codantes (Gymrek et al., 2016; Marshall et al., 2021; Usdin et al., 2015) (Figure 5). Par exemple, les répétitions tri-nucléotidiques CAG, codant pour la glutamine, sont courantes dans certains gènes et jouent un rôle dans les maladies neurodégénératives telles que la maladie de Huntington lorsqu'elles s'étendent au-delà d'un certain seuil (Pearson et al., 2005) (voir partie 3.2.3.).

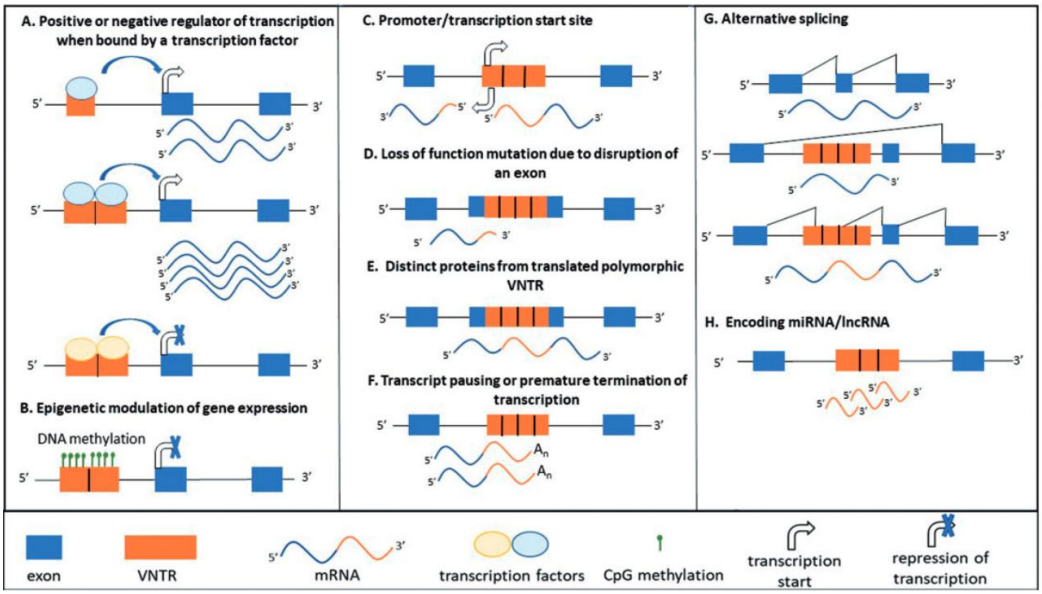

Figure 5 : Mécanismes par lesquels les microsatellites peuvent modifier l'expression de gènes (Marshall et al., 2021).

2.2. Mutations dans les microsatellites

2.2.1. Taux de mutation

La notion de taux de mutation est une métrique majeure en génétique et d'une importance fondamentale pour la compréhension de l'évolution des génomes. Elle est un marqueur indispensable comme source de la diversité génétique, de même que pour différentes applications. Plusieurs mécanismes sont responsables de l'augmentation du taux de mutations, et ainsi du polymorphisme. Ce taux de mutations peut varier en fonction du microsatellite, mais également en fonction de l'organisme hôte et de la localisation du microsatellite dans le contexte nucléotidique.

Les microsatellites, en raison de leur nature répétée, sont particulièrement vulnérables aux erreurs de réplication commises par l'ADN polymérase. Leur stabilité pendant la réplication dépend notamment de certaines caractéristiques structurelles spécifiques aux microsatellites. Il a été constaté que la fréquence des erreurs de l'ADN polymérase augmente avec la longueur de l'unité répétée (Kroutil et al., 1996; Strauss et al., 1997). La composition en bases joue également un rôle important dans la mutabilité : les microsatellites composés de C ou G sont plus instables que ceux composés de A ou T, tant chez les procaryotes que chez les eucaryotes (Alhopuro et al., 2012; Harfe and Jinks-Robertson, 2000). De plus, la nature des répétitions semble influencer le taux de mutation, puisqu'il a été observé que pour un même nombre de répétitions, la stabilité augmente avec la taille du motif répété ; en d'autres termes, les répétitions mononucléotidiques sont plus instables que les dinucléotidiques, qui sont elles-mêmes plus instables que les trinucléotidiques, et ainsi de suite (Boyer et al., 2002). La composition des séquences adjacentes joue également un rôle, rendant certains sites plus susceptibles au glissement et, par conséquent, aux erreurs de la polymérase (Duval and Hamelin, 2002). Enfin, il a été démontré que le taux de mutation des microsatellites est positivement corrélé à l'état de compaction de la chromatine et à l'activité transcriptionnelle de ces régions de l'ADN : les domaines d'euchromatine et les introns présentent une plus grande fréquence de mutations par rapport aux régions d'hétérochromatine et intergéniques, respectivement (Kim et al., 2013).

Dans le cadre des microsatellites, certains mécanismes de mutations induisant des mutations de type ponctuel sont favorisés (Lynch, 2010), comme les recombinaisons inégales de chromosomes et/ou de chromatides sœurs et le glissement de la polymérase. Également, les déaminations des cytosines méthylées peuvent avoir un impact, notamment lors de la création des microsatellites (Poulos et al., 2017).

2.2.2. Type de mutations : Mutations ponctuelles

Le type de mutation le plus couramment observé dans le génome est la mutation dite ponctuelle. Elles se produisent lorsqu'une base azotée dans la séquence d'ADN est remplacée par une autre (Figure 6.1). Ce phénomène inclut principalement les transitions, pour lesquelles une purine (A ou G) est remplacée par une autre purine ou une pyrimidine (C ou T) par une autre pyrimidine, et les transversions, qui impliquent un échange entre une purine et une pyrimidine (Lynch, 2010) (Figure 6.2). En plus des substitutions, les mutations ponctuelles peuvent également comprendre des insertions ou des délétions de quelques bases, appelées indels (Stenson et al., 2014). Ces mutations surviennent principalement lors de la réplication de l'ADN, en raison des erreurs de la polymérase, bien que d'autres processus, tels que la réparation des cassures double-brin ou l'intégration d'éléments exogènes, puissent également en être la source. Parmi les différents types de mutations ponctuelles, les transitions sont généralement plus fréquentes que les transversions, tandis que les indels restent relativement rares (Nachman and Crowell, 2000). Contrairement aux mutations dites « larges », qui concernent des

mutations à l'échelle du chromosome, ici, une mutation ponctuelle peut concerner des tailles allant d'une unique base à des séquences d'une centaine de bases.

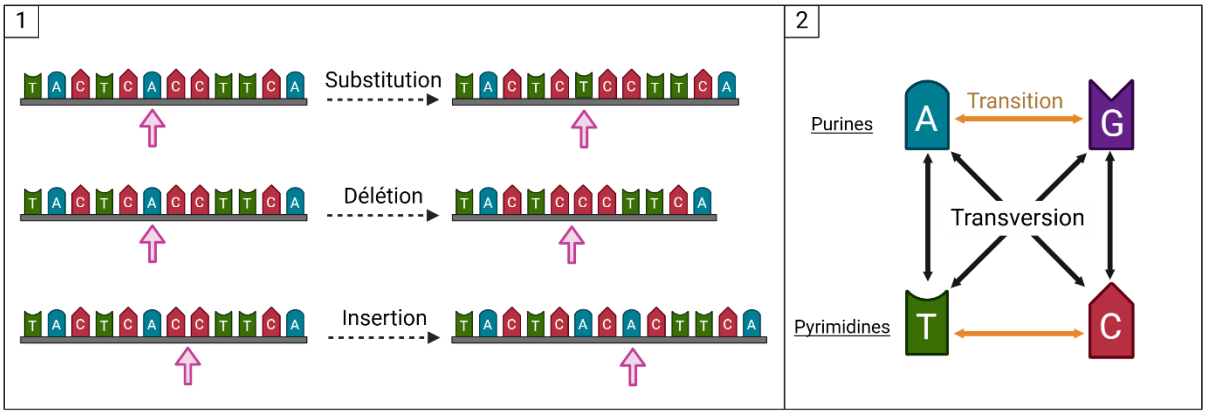

Figure 6 : Mécanisme des mutations ponctuelles. 1. Les trois cas possibles de mutation ponctuelle. Les sites fléchés sont les sites de la mutation. 2. Les différentes substitutions possibles.

2.2.3. Mécanisme de mutations

Pour obtenir des mutations de types ponctuelles, de nombreux mécanismes rentrent en jeu, mais ici ne seront présentés que ceux concernant les microsatellites (Tableau 4).

Tableau 4 : Tableau récapitulatif des caractéristiques des différents mécanismes de mutations chez les microsatellites.

Mécanisme	Taille de mutation	Fréquence	Référence
Désamination des cytosines méthylées	1 base (C→T) dans les CpG	Moins de 5 %	(Walser and Furano, 2010)
Recombinaison homologue	Séquences longues jusqu'à segments chromosomiques	5 à 10 %	(Kelkar et al., 2008)
Glissement de polymérase	1 à plusieurs répétitions	90 à 95 %	(Ellegren, 2004)

2.2.3.1. Déamination des cytosines méthylées

La déamination des cytosines est une forme courante de dégradation de l'ADN, souvent déclenchée par divers types de stress (Poulos et al., 2017; Wang et al., 1980). Ce processus transforme une base cytosine (C) en uracile (U). En général, cette erreur est détectée et corrigée par le système de réparation de l'ADN, cependant, lorsque la cytosine a été méthylée (m5C), la déamination génère une thymine (T) à la place. Le système de réparation peut alors corriger soit le T en C, soit la guanine complémentaire (G) en adénine (A) (Figure 7) ; seule cette dernière correction mène à une mutation dans 20% des cas de cytosines méthylées. La méthylation des cytosines est un phénomène courant chez les eucaryotes, notamment chez les vertébrés et les plantes. Cette méthylation concerne 90 % des cytosines des dinucléotides CG (Ehrlich and Wang, 1981), appelés CpG pour indiquer que la guanine suit la cytosine dans le sens 5'-3'. L'action combinée de la méthylation et de la déamination entraîne une sur-représentation des transitions de C vers T, ou de G vers A, dans les sites CpG, avec un taux de mutation 2 à 4 fois plus élevé que dans les autres sites (Podlutsky et al., 1998). Ces mutations ne peuvent pas être

pleinement considérées comme des mutations ponctuelles, car elles dépendent du processus de méthylation qui n'est pas aléatoire.

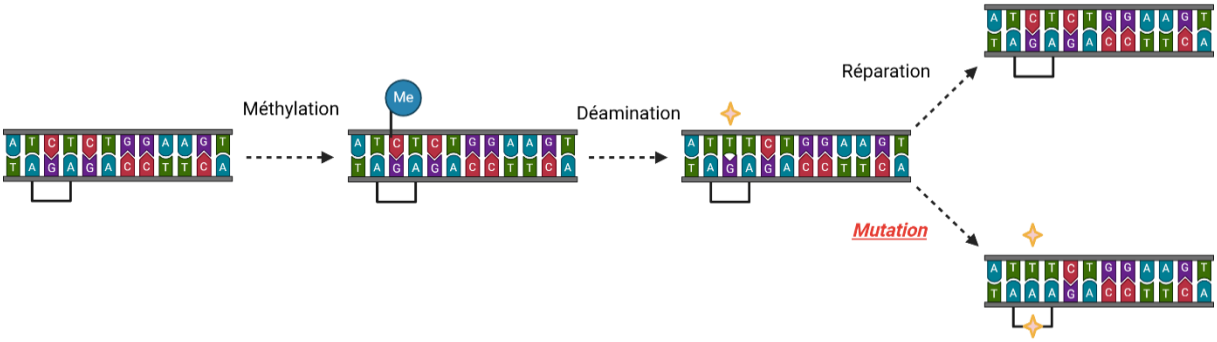

Figure 7 : Processus de déamination des cytosines méthylées. Le site encadré est le site de mutation. La bulle «Me» indique la cytosine méthylée.

2.2.3.2. Recombinaison homologue

La recombinaison homologue est un mécanisme clé dans la réparation des cassures double-brin (CDB) de l'ADN et joue un rôle crucial dans la diversité génétique ainsi que dans le maintien de la stabilité du génome. En temps normal, les CDB sont accidentelles et causées par des stress cellulaires (radiations UV par exemple), mais elles sont aussi provoquées lors de la phase de division méiotique des cellules. Plus de dix gènes sont, par exemple, responsables des CDB méiotiques chez *Saccharomyces cerevisiae*, augmentant ainsi le nombre de recombinaisons de 100 à 1000 fois par rapport aux cellules végétatives (Smith and Nicolas, 1998).

Ce processus utilise une séquence d'ADN homologue comme matrice pour réparer les cassures, garantissant ainsi la fidélité de la réparation. Cependant, la présence de séquences répétées, telles que les microsatellites, peut influencer ce processus de manière significative. Les microsatellites, constitués de répétitions courtes de motifs nucléotidiques, sont particulièrement sujets à des erreurs de réplication et à des glissements de brins, qui peuvent provoquer des expansions ou des contractions de ces répétitions lors de la recombinaison homologue.

Ces séquences répétées peuvent favoriser une recombinaison homologue non allélique (NAHR), où des segments de l'ADN non homologues mais similaires en séquence sont mal alignés et échangent des segments, ce qui peut entraîner des délétions, des duplications, ou d'autres réarrangements chromosomiques (Stankiewicz and Lupski, 2002). De plus, la stabilité des microsatellites pendant la recombinaison homologue dépend de divers facteurs, notamment la longueur et la composition des répétitions (Bayliss et al., 2004).

La recombinaison homologue est également influencée par l'état de la chromatine. Les régions de chromatine ouverte (euchromatine), qui sont plus transcriptionnellement actives, montrent une fréquence accrue de mutations et de recombinaisons comparativement aux régions de chromatine condensée (hétérochromatine) (Kim et al., 2013). Cela suggère que l'organisation de la chromatine et les activités transcriptionnelles associées jouent un rôle non négligeable dans la modulation des événements de recombinaison dans les microsatellites, et que ces éléments peuvent contribuer à la diversité génétique et à l'évolution des génomes eucaryotes.

Enfin, il peut aussi y avoir l'implication de crossing-over inégaux. En effet, lors de ces crossing-over, la région située entre les répétitions d'un des brins est alors dupliquée sur l'une des molécules

d'ADN, tandis que la région homologue de l'autre brin est délétée (Lynch, 2007) (Figure 8). Les événements de duplication et délétion ne peuvent cependant se produire que lorsque les régions d'homologies sont situées sur deux chromatides sœurs ou chromosomes homologues. Ainsi, les crossing-over sont responsables de mutations chez les microsatellites avec des tailles variables et donc transmissibles à la descendance (Kelkar et al., 2008).

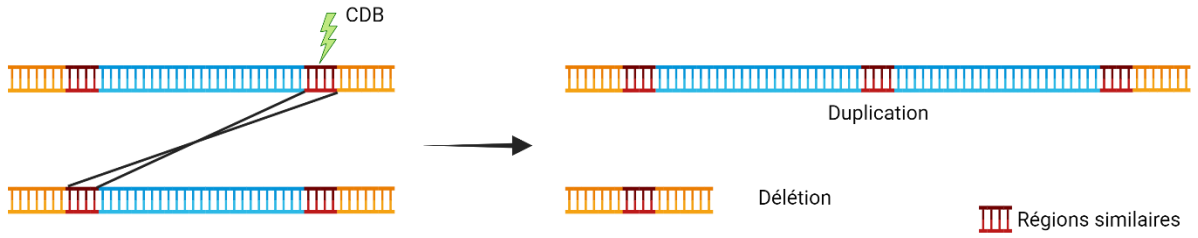

Figure 8 : Crossing-over inégal. Le crossing-over inégal ne peut se produire que lors d'une recombinaison ectopique et provoque la duplication d'un fragment d'ADN dans une séquence, et sa délétion dans l'autre séquence.

2.2.3.3. Glissement de Polymérase

La mutation par glissement (SSM, *Slipped Strand Mispairing*) (Levinson and Gutman, 1987a) se produit lors de la réplication de l'ADN. Pendant la synthèse du nouveau brin par les polymérases, il peut arriver que l'extrémité du brin en cours de synthèse se détache temporairement du brin matrice. Ce détachement peut être accompagné d'une pause de la polymérase, notamment dans les régions à séquences répétées en tandem. Ces pauses s'expliquent par la difficulté pour la polymérase de maintenir un alignement stable et une progression régulière sur ces motifs. Bien que cela puisse ralentir le processus, ce n'est généralement pas problématique, car le brin peut se réapparier grâce à la complémentarité des bases avec le brin matrice.

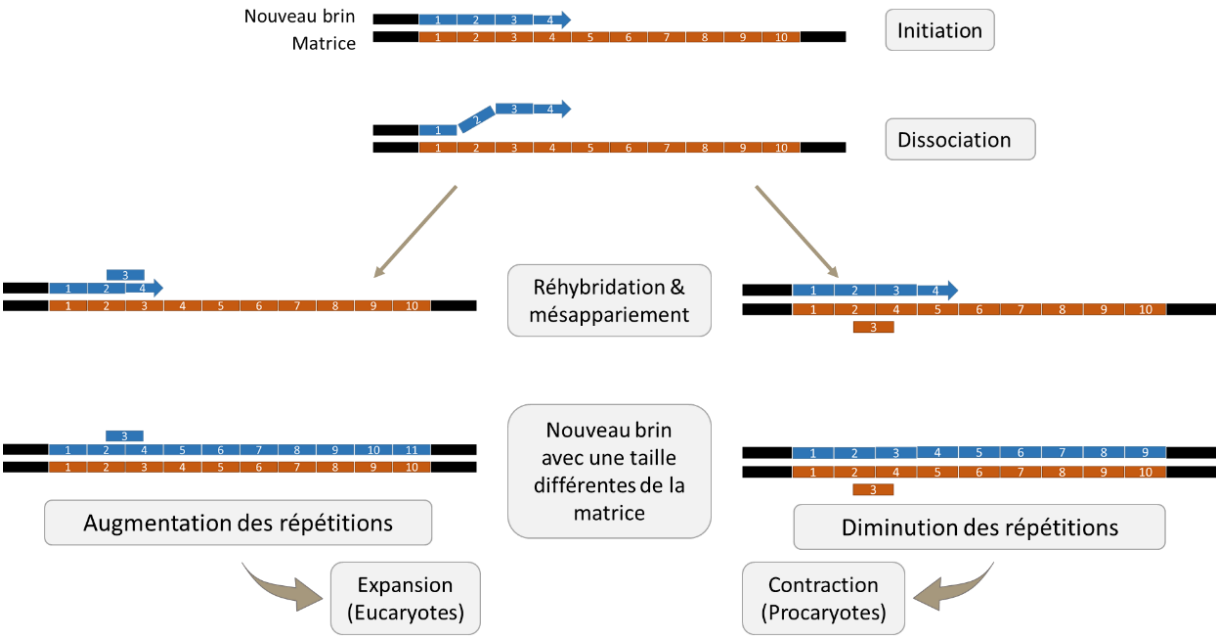

Figure 9 : Le glissement de la polymérase lors de la réplication d'une séquence répétée. Les répétitions sont caractérisées par leur numéro.

Cependant, lorsqu'un mésappariement se produit dans une région où les séquences sont répétées en tandem, cela peut provoquer un décalage pendant la réplication. En effet, le réappariement peut se faire sur l'une des répétitions adjacentes ou former des structures secondaires car la complémentarité est toujours respectée par les motifs. Dans ce cas, on dit que la polymérase a "glissé" d'une ou plusieurs répétitions (Tanudisastro et al., 2024; Viguera et al., 2001). Si le réappariement se fait en amont, la répétition sera synthétisée deux fois, entraînant ainsi l'ajout d'une répétition dans le nouveau brin d'ADN. À l'inverse, si le réappariement se fait en aval, l'une des répétitions ne sera pas du tout synthétisée, ce qui conduit à la perte d'une répétition dans l'ADN (Figure 9).

Le glissement peut suivre plusieurs modèles qui se répartissent en deux grandes catégories théoriques. Les modèles IAM (*Infinite Allele Model*)(Kimura and Crow, 1964) et KAM (*K-Allele Model*)(Kimura, 1980) considèrent que chaque mutation crée un nouvel allèle, soit unique pour l'IAM, soit tiré aléatoirement parmi K allèles possibles pour le KAM. Ces modèles, initialement développés pour les mutations ponctuelles, peuvent être appliqués aux microsatellites en supposant que les mutations produisent des allèles de tailles variées (Estoup et al., 2002). En parallèle, le modèle SMM (*Stepwise Mutation Model*)(Kimura and Ohta, 1978) stipule que les mutations modifient progressivement la taille des allèles, de manière incrémentale (+1 ou -1 répétition). Ce modèle s'adapte particulièrement bien aux microsatellites, en considérant des mutations symétriques et de petites variations de taille (Kimura and Ohta, 1978). Ces modèles ont été enrichis pour intégrer des mutations à pas multiple, comme dans le TPM (*Two-Phase Model*)(Di Rienzo et al., 1994), où la probabilité de mutations à pas simple ou multiple est définie, ou le GSM (*General Stepwise Model*)(Fu and Chakraborty, 1998), où la taille des mutations suit une loi géométrique. Des mécanismes comme le biais de contraction vers une taille focale ou un taux de mutation proportionnel à la taille du locus ont également été introduits (Kruglyak et al., 1998). Ces avancées incluent des mécanismes d'apparition des microsatellites pour éviter leur disparition complète par contractions excessives (Tableau 5).

Tableau 5 : Caractéristiques des différents modèles de glissement de la polymérase (Di Rienzo et al., 1994; Estoup et al., 2002; Goldstein and Schlotterer, 1999; Kimmel et al., 1998; Kimura and Crow, 1964; Levinson and Gutman, 1987a; Shriver et al., 1995).

Modèle	Type de mutation	Variabilité allélique	Contexte d'utilisation
Simple Mutation Model (SMM)	Insertion ou délétion d'une unité à la fois	Modérée	Microsatellites, mutations progressives
Two-Phase Model (TPM)	Mutation simple et multiple	Plus élevée que SMM	Régions répétées où mutations complexes possibles
Stepwise Mutation Model (SMM étendu)	Mutation simple (fréquente) et double (rare)	Modérée à élevée	Séquences répétées avec variation limitée
Generalized Stepwise Model (GSM)	Mutation simple et multiple	Haute	Séquences répétées avec changements de longueur plus importants
Infinite Alleles Model (IAM)	Un nouvel allèle pour chaque mutation	Très élevée	Études de diversité génétique dans populations
K-Allele Model (KAM)	Nouvelle mutation dans les limites de K allèles	Limitée	Modèles d'évolution et diversité génétique dans les populations

Les modèles sont ajustés à deux types de données : les analyses de variabilité (AV), qui comparent la distribution d'allèles à quelques locus, et les analyses de séquence (AS), qui considèrent un seul allèle sur plusieurs locus (Shriver et al., 1995). Des approches comme les processus markoviens

permettent de simuler l'évolution des tailles de locus, bien que des phénomènes plus complexes que les simples mutations ponctuelles soient nécessaires pour expliquer l'apparition des microsatellites (Kruglyak et al., 1998). Enfin, des tests par maximum de vraisemblance ont été introduits pour comparer l'ajustement des modèles aux données en fonction de leurs paramètres. La comparaison la plus aboutie conclut qu'un modèle avec un taux de mutation proportionnel à la taille du locus, biaisé vers une taille focale, offre le meilleur ajustement, tandis que les mutations à pas multiple n'apportent pas d'amélioration significative par rapport à celles à pas simple (Sainudiin et al., 2004).

Cependant, un événement de mésappariement ne conduit pas systématiquement à une mutation en taille, car la plupart de ces erreurs sont corrigées par le système de réparation MMR (Levinson and Gutman, 1987b; Strand et al., 1993) ou l'activité correctrice (3'->5' exonucléase) de la polymérase, qui interviennent immédiatement après la synthèse pour vérifier que le nouveau brin est conforme au brin matrice. Toutefois, ces réparations ne sont pas systématiques (Tautz and Schlötterer, 1994), car elles sont sensibles à de nombreux facteurs comme la nature du microsatellite, sa composition, sa localisation, son pourcentage de GC ou ses conditions environnementales, par exemple (Sia et al., 1997a). De plus, il est possible que le système MMR se retrouve défaillant, induisant une instabilité responsable de certains cancers (Ellegren, 2004; Goldstein and Schlotterer, 1999). Ainsi, le glissement de la polymérase est considéré comme le principal facteur d'évolution des microsatellites.

Il est également important de noter que ce phénomène de glissement ne se limite pas à la synthèse lors de la réplication de l'ADN, mais peut aussi se produire pendant les événements de réparation de l'ADN tels que la SDSA (*Synthesis-Dependant Strand Annealing*) (Richard and Paques, 2000). Le principe reste le même : le brin en cours de synthèse peut se désapparier de la séquence matrice et se réapparier de manière décalée. Si l'erreur n'est pas corrigée à ce moment-là, la mutation devient définitive, car le brin erroné servira ensuite de matrice pour la réparation du brin complémentaire.

2.2.3.4. Cas particulier

Il existe toutefois un cas particulier : les expansions de triplets. Les mécanismes de mutation des microsatellites et des expansions de triplets diffèrent en ce qui concerne leur dynamique et leurs conséquences biologiques. Alors que les microsatellites mutent en suivant majoritairement le modèle du glissement de la polymérase, les triplets, composés de motifs de trois nucléotides tels que $(CAG)_n$ ou $(CGG)_n$, présentent un seuil critique d'instabilité. Au-delà d'un certain nombre de répétitions, ces séquences deviennent sujettes à des expansions massives, favorisées par la formation de structures secondaires telles que des épingles à cheveux ou des triplex. Ces mécanismes complexes, souvent multi-pas, perturbent la réplication et la réparation de l'ADN, entraînant des expansions progressives (Iyer et al., 2015; Lieberman and Fischbeck, 2000; Otten and Tapscott, 1995). Contrairement aux microsatellites, les triplets sont souvent localisés dans des gènes essentiels, et leurs expansions sont directement associées à des maladies graves telles que la maladie de Huntington ou le syndrome de l'X fragile (Dragileva et al., 2009; Rajan-Babu et al., 2017).

2.3. Cycle de vie du microsatellite

Certaines recherches se sont intéressées au cycle de vie des microsatellites, c'est-à-dire comprendre et expliquer comment ces séquences se forment, mutent et évoluent dans les génomes d'une population au cours des générations. Les microsatellites ont en effet un potentiel évolutif, voire

adaptatif, phénoménal. Chaque locus peut changer de taille au fil des générations, et cela aussi bien dans les régions codantes que non-codantes, ce qui leur confère une très grande plasticité. Toutefois, ce mécanisme repose majoritairement sur deux facteurs, la taille du motif répété et le nombre de répétitions au sein du microsatellite. De plus, un microsatellite se définit également par sa capacité à perdre ou gagner des motifs. En effet, si on prend un motif CA avec une répétition, on ne pourra pas le considérer comme le microsatellite $(CA)_2$ car il ne présente pas de capacité de contraction ou d'élongation contrairement à $(CA)_8$. Une taille minimale a donc été proposée afin de pouvoir qualifier une séquence de microsatellite et de pouvoir correctement étudier leurs apparitions et disparitions (Balaresque, 2007; Kelkar et al., 2011).

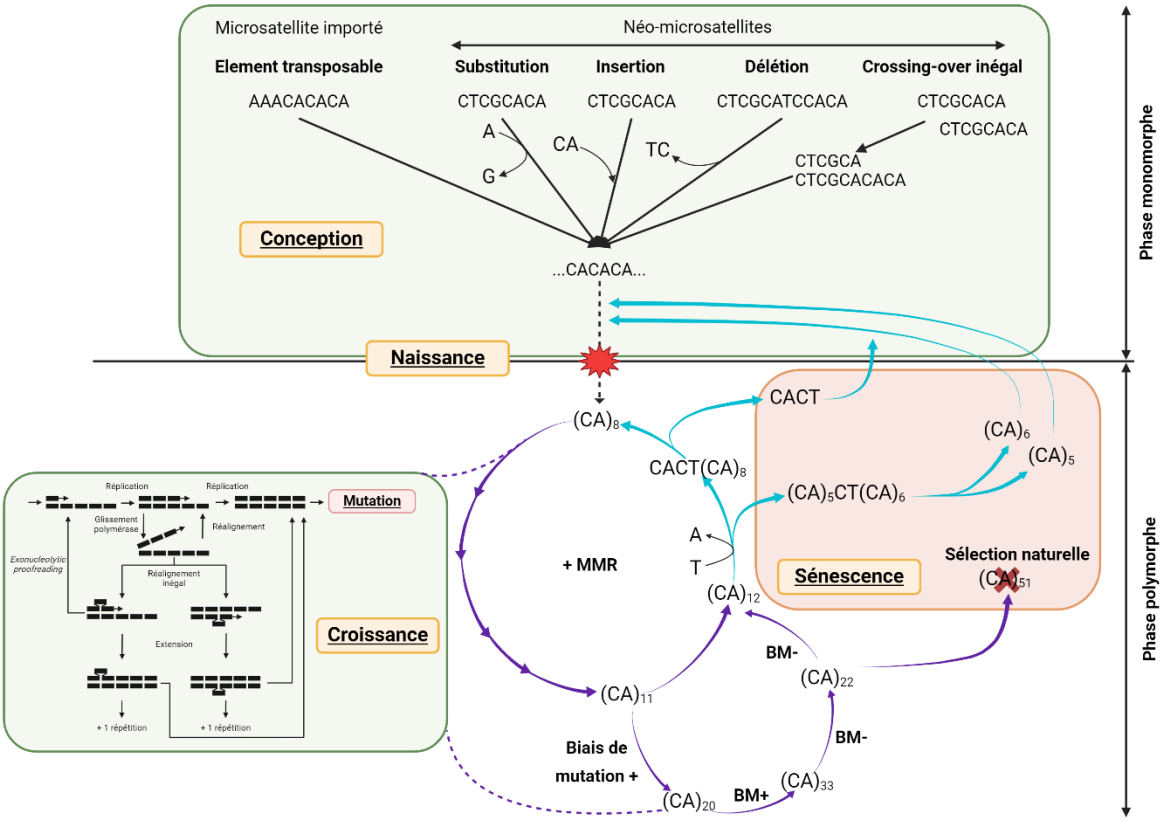

Figure 10 : Représentation schématique sous forme de cycle de vie dynamique d'un locus microsatellite (par exemple $(CA)_n$) inspiré de (Balaresque, 2007)

La capacité d'expansion et de contraction est transitoire car cyclique, et sa variabilité est dépendante du taux de mutations d'un locus. Ainsi, le cycle de vie des microsatellites peut être divisé en quatre phases (Figure 10) :

- Conception : la mise en place de la matrice minimale (monomorphe, par exemple $(CA)_3$)
- Naissance : le passage de la matrice initiale au microsatellite possédant sa capacité de variations dans le nombre de répétitions.
- Croissance : la phase de maturité caractérisée par le maintien d'une taille ou de sa capacité de mutations stables pendant plusieurs générations
- Sénescence / renaissance : la disparition du locus ou le redémarrage d'un cycle.

De plus, chaque étape est régie par de nombreux mécanismes de mutations, bien que certains soient majoritaires (Tableau 6) (Baptiste et al., 2013; Buschiazzo and Gemmell, 2006; Ellegren, 2004; Kelkar et al., 2011; Taylor et al., 1999).

Tableau 6 : Mécanisme de mutation majoritaire pour chaque phase du cycle de vie d'un microsatellite (Baptiste et al., 2013; Buschiazzo and Gemmell, 2006; Ellegren, 2004; Kelkar et al., 2011; Taylor et al., 1999)

Phase du cycle de vie	Mécanismes de mutation majoritaire
Conception	Glissement de la polymérase (insertion)
Naissance	Glissement de la polymérase (insertion) Recombinaison Homologue
Croissance	Glissement de la polymérase (expansion / contraction) Recombinaison Homologue
Sénescence	Glissement de la polymérase (contraction) Désamination Recombinaison Homologue
Mort	Glissement de la polymérase (délétion) Désamination Recombinaison Homologue

2.3.1. Conception

Malgré les années passées par un grand nombre de scientifiques à étudier les microsatellites, un débat important règne encore concernant leur conception. Existe-il un seuil en dessous duquel un microsatellite ne peut pas être polymorphe ? Pour certains, ce « seuil » serait entre 7 et 9 (Kelkar et al., 2010; Lai and Sun, 2003; Rose and Falush, 1998; James L. Weber, 1990), pour d'autres de 4 à 8 selon que ce soit pour les microsatellites di- à tétranucléotidiques, mais pour d'autres ce « seuil » n'existe pas (Leclercq et al., 2010; Pupko and Graur, 1999). La naissance du polymorphisme dépendrait alors dans tous les cas de la conception du microsatellite. Deux mécanismes proposés semblent possibles pour expliquer l'origine de la création d'un nouveau microsatellite (Figure 10).

Le premier consiste en une matrice initiale qui naît spontanément par le biais de substitutions ponctuelles (Messier et al., 1996), ou d'insertions / délétions. Cependant, les substitutions ponctuelles apportent quelques répétitions mais ne produisent pas une matrice assez longue, à l'inverse des insertions/délétions, plus prometteuses (Zhu et al., 2000). Ainsi, une combinaison des deux est fortement envisageable. Ici on parlera de néo-proto-microsatellites (Figure 10).

Le second consiste en une séquence anonyme qui peut être copiée *via* transposition d'éléments ou duplications, notamment auprès de séquences abondantes comme des «*short or long interspersed elements*» (SINE et LINE) ou des rétro-pseudogènes (Wilder and Hollocher, 2001). La présence de leurs queues poly-A expliquerait la genèse de microsatellites de type poly-A (Nadir et al., 1996). On parlera ici de proto-microsatellites (Figure 10).

2.3.2. Naissance

Après la conception, vient la naissance, même si cette phase est la plus trouble du cycle. Il faut, dans un premier temps, qu'un nombre minimal d'unités répétées, le «seuil», nécessaire à la mise en place des facteurs de croissance, soit atteint. Ce seuil semblerait se situer entre 2 et 10 répétitions en fonction du motif et de la séquence environnante (Figure 10) (Balaresque, 2007; Kelkar et al., 2011).

Cependant, pour certains ce « seuil » n'existe pas et la naissance du microsatellite ne serait pas directement liée au contexte génomique et aux fonctions du locus (Rose and Falush, 1998; J. L. Weber, 1990). Toutefois, il a été montré qu'un polymorphisme de 6 répétitions était observable sur des microsatellites liés aux duplications segmentaires, ce qui semble appuyer l'hypothèse d'une augmentation du polymorphisme des microsatellites dans un environnement répété (Balaresque et al., 2003).

2.3.3. Croissance

La phase de croissance d'un microsatellite représente sa phase de maturité, où il est le plus polymorphe et ainsi informatif et utile selon les domaines d'applications. À partir de là, on utilise les termes d'expansion et de contraction plus que d'insertion et de délétion pour définir les mouvements de croissance. On parle alors de microsatellite mature à partir généralement de 10, 12, 15, 20 ou 30 répétitions pour les mono- et di-, tri-, tétra- et penta-nucléotides ou plus, respectivement. Cette fluctuation du nombre de répétitions peut être expliquée par cinq « groupes » de facteurs (Figure 10) (Baptiste et al., 2013; Buschiazzo and Gemmell, 2006; Kelkar et al., 2011).

Le premier et sans doute le plus important, comprend les mécanismes mutationnels tels que les mutations ponctuelles, les recombinaisons homologues, les crossing-over inégaux mais principalement le glissement de la polymérase (Tableau 6).

Le second comprend la nature du microsatellite, tels que sa pureté, la composition de son motif, la longueur de son motif, la longueur de ses allèles ou encore son hétérozygotie.

Le troisième correspond au contexte génomique, tel que la proportion en GC, le taux de recombinaison et/ou de mutation locale, la composition des séquences flanquantes, les liaisons possibles avec des éléments transposables ou encore le type de régions (transcrite, non-transcrite ou régulatrice) où se trouve le microsatellite (Figure 2).

Le quatrième représente le contexte du porteur (ou individu), tels que le sexe, l'âge, l'ethnicité, l'hétérozygotie, la structure chromosomique associée ou les caractéristiques environnementales.

Pour finir, le cinquième concerne les influences sélectives comme les systèmes de réparation, le temps de génération ou encore le taux de métabolisme.

2.3.4. Sénescence / Mort / Renaissance

Même s'il semble se créer un équilibre entre expansion et contraction et la réparation des erreurs, l'augmentation du nombre de répétitions entraîne généralement encore plus de répétitions – les microsatellites sont considérés en expansion dans les génomes eucaryotes - la croissance devrait tendre vers de très grandes tailles. Or, il est très rare de voir des microsatellites de plus de 200 répétitions (Schlötterer, 2004).

Trois hypothèses ont été proposées. Il y aurait un biais de mutations envers les contractions à partir d'une certaine taille ou nombre de répétitions, ce qui permettrait de maintenir les microsatellites dans un intervalle préférentiel de taille permettant un équilibre (Kelkar et al., 2011; Xu et al., 2000). Par ailleurs, la sélection naturelle jouerait un rôle régulateur possible afin de garantir une préservation structurelle de l'ADN. Enfin, certaines substitutions, ou insertions/délétions pourraient interrompre le motif (Goldstein and Schlötterer, 1999), provoquant une perte de séquence microsatellite et du potentiel polymorphique (Angers and Bernatchez, 1997), et dégénérer le microsatellite en séquence classique (Taylor et al., 1999) entraînant ainsi la mort du microsatellite (Kelkar et al., 2011). Toutefois, si l'interruption ne se déroule pas au centre du microsatellite mais de manière dissymétrique, cela raccourcira le motif principal et induira la continuation du cycle sous forme de renaissance (Figure 10).

3. Le système MMR

Le système de réparation des mésappariements de l'ADN, MMR (*MisMatch Repair*), est un mécanisme essentiel pour maintenir l'intégrité et la stabilité du génome. Il joue un rôle crucial dans la correction des erreurs survenues lors de la réplication de l'ADN, prévenant ainsi l'accumulation de mutations qui peuvent conduire à des maladies génétiques et à des cancers.

3.1. Son rôle et son fonctionnement

Le système MMR joue un rôle crucial dans la correction des erreurs de réplication de l'ADN, garantissant ainsi la fidélité de l'information génétique. Il détecte et répare les bases mal appariées ainsi que les petites insertions et délétions qui surviennent lors de la réplication de l'ADN. Mais ce mécanisme diffère selon l'organisme qui le contient.

Lors d'un glissement de polymérase, le mésappariement crée une boucle, soit sur le brin matrice, soit sur le brin en cours de synthèse. Si cette boucle mesure moins d'une dizaine de bases, elle est reconnue par le système MMR (Iyer et al., 2006). Selon le brin sur lequel se trouve la boucle, deux scénarios sont à envisager : si elle se trouve sur le brin en cours de synthèse, elle sera excisée, mais si elle se trouve sur le brin matrice, le brin complémentaire sera clivé pour permettre la resynthèse des bases manquantes (Figure 11). Cette reconnaissance est brin spécifique. Bien que pour les eucaryotes et les procaryotes le mécanisme ne soit pas encore pleinement défini, chez les bactéries cette reconnaissance semble être couplée à la réplication où le brin néosynthétisé est repéré par un système de marquage.

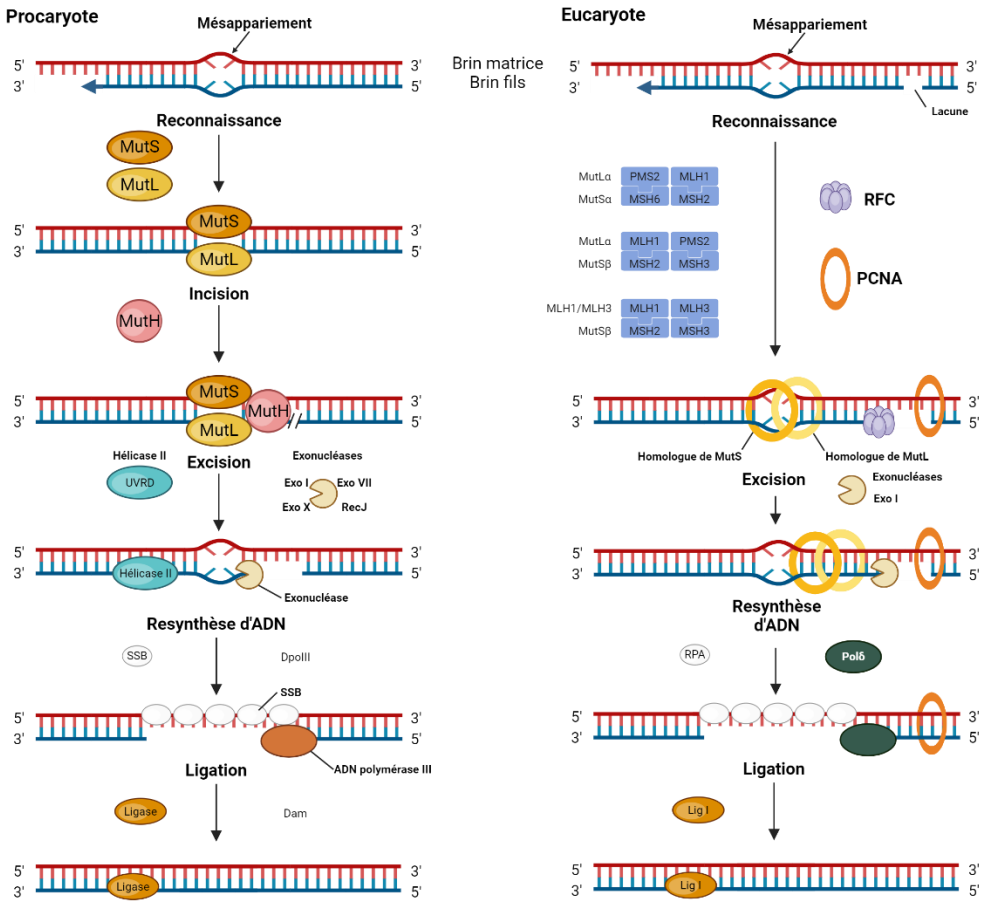

Figure 11 : Représentation schématique du système de réparation des mésappariements de bases de l'ADN ou système MMR (MisMatch Repair) inspiré du schéma issu de la base de données KEGG (« Kyoto Encyclopedia of Genes and Genomes »)

3.1.1. Chez les procaryotes

Chez les procaryotes, et notamment chez la bactérie *Escherichia coli* qui a été la plus étudiée, le système MMR est initié lorsque l'homodimère d'enzyme MutS se fixe comme une pince asymétrique à l'ADN possédant des mésappariements. L'homodimère MutL se couple alors avec MutS et distingue le brin matrice du nouveau brin. Chez *E. coli*, l'absence de méthylation sur l'adénine catalysée par l'adénine méthyltransférase (Dam), au niveau des séquences GATC nouvellement synthétisées, permet à MutH d'entailler le nouveau brin. Cette coupure sert alors de guide pour le retrait du mésappariement impliquant l'hélicase UvrD et pour l'excision par l'exonucléase d'ADN simple brin dans les deux sens selon la polarité de l'entaille par rapport au mésappariement. L'ADN polymérase III vient alors resynthétiser correctement l'ADN avec l'aide de la protéine de liaison d'ADN simple brin SSB (*Single-Strand dna-Binding protein*) puis la ligase finalise la réparation (Larrea et al., 2010) (Figure 11).

3.1.2. Chez les eucaryotes

Le système MMR des eucaryotes est similaire aux procaryotes, toutefois il comporte plusieurs homologues différents de MutS et MutL. MutSα (MSH2/MSH6) reconnaît les mésappariements d'une seule base et les insertions/délétions de 1 à 2 bases, tandis que MutSβ (MSH2/MSH3) reconnaît les

insertions/délétions de 2 bases ou plus (Li, 2008). Il y a également 3 hétérodimères de MutL : MutLα (MLH1/PMS1 chez la levure, MLH1/PMS2 chez l'homme), MutLβ (MLH1/MLH3) et MutLγ (MLH1/MLH2 chez la levure, MLH1/PMS1 chez l'homme). De plus, les hétérodimères MutS et MutL des eucaryotes ont un chevauchement partiel de spécificité de substrat. MutLα et MutLβ ont une activité endonucléase, avec le site actif présent sur PMS1 pour la levure et sur PMS2 et MLH3 chez l'homme.

Le processus de reconnaissance du brin matrice est controversé à l'heure actuelle car il n'est pas élucidé. L'entaille réalisée est utilisée pour l'excision par l'exonucléase à ADN double brin, Exo1. Certaines études montrent que l'activité endonucléase permettant l'entaille en 3' chez MutLα est dépendante de PCNA (*Proliferating Cell Nuclear Antigen*) / RFC (*Replication Factor C*) (Kadyrov et al., 2006). PCNA interagit également avec MSH2 et MLH1 et jouerait un rôle dans l'initiation et les étapes de resynthèse de l'ADN (Umar et al., 1996). Il semblerait également que PCNA interagit avec MSH6 et MSH3 via la PIP box, un motif d'interaction de PCNA, afin d'aider à la localisation des mésappariements (Kleczkowska et al., 2001; Shell et al., 2007). Enfin, PCNA est nécessaire pour les entailles dirigées en 3', et non essentielle pour celles en 5', car Exo1 est une 5'→3' exonucléase et elle interagit également avec MSH2 et MLH1 (Tishkoff et al., 1997). Aucune hélicase n'a été impliquée dans le système MMR eucaryote. Enfin, l'ADN est resynthétisé par une ADN polymérase (principalement l'ADN polymérase δ) aidée de la RPA (*Replication Protein A*) et ligué par l'ADN ligase 1 (Larrea et al., 2010; Qiu et al., 2015) (Figure 11).

3.2. Conséquences de la déficience du MMR et des mutations microsatellites

Bien que le rôle du système MMR soit de maintenir la fidélité de la réplication de l'ADN en corrigeant les mésappariements et les boucles d'insertion/délétion apparaissant lors de la réplication de l'ADN, il joue aussi d'autres rôles. Il a une contribution importante lors des contrôles des phases G1 et G2 du cycle cellulaire, lors de la réponse apoptotique initiée par certains dommages à l'ADN, et concernant la fidélité de la recombinaison génétique (Chao and Lipkin, 2006). De plus, il contribue au traitement des mésappariements lors de dommages sur l'ADN liés à l'O^6-méthylguanine, la 8-oxoguanine, les adduits de cisplatine et les photo-produits d'ultraviolets (Iyer et al., 2006).

Toutefois, bien que la fidélité de la réplication de l'ADN soit élevée dans les cellules possédant un système MMR fonctionnel, des erreurs ponctuelles subsistent tout de même. Le système MMR corrige efficacement la plupart des erreurs dues aux glissements de polymérase, par exemple. Cependant, certaines échappent au processus de réparation en raison de facteurs tels que la saturation du système, la présence de structures secondaires dans l'ADN ou la proximité de régions difficiles à répliquer (Chung et al., 2021; Cilona et al., 2020). Ces erreurs non corrigées peuvent alors s'accumuler sur de nombreuses générations, générant du polymorphisme dans les séquences microsatellites (Li et al., 2004). Ce polymorphisme joue notamment un rôle clé dans la variabilité génétique au sein des populations, offrant des marqueurs utiles pour les études de génétique des populations, les diagnostics médicaux et l'analyse de traits adaptatifs (Anmarkrud et al., 2008).

Ainsi, au vu de l'impact d'un système opérationnel, une déficience du système MMR favorise fortement le développement de malignité par le biais de nombreux mécanismes différents.

3.2.1. Instabilité des microsatellites (MSI)

La déficience du système MMR conduit à l'accumulation d'erreurs lors de la réplication de l'ADN, qui se traduit dans des séquences de microsatellites par des insertions et délétions en grand nombre

(Hause et al., 2016). Ces mutations induisent un phénomène génétique connu sous le terme d'instabilité des microsatellites (MSI), initialement décrit dans des tumeurs familiales associées au syndrome de Lynch, une forme héréditaire de cancer colorectal, distingué par une mutation germinale dans les gènes du système MMR (Boland and Goel, 2010).

Outre son association avec le syndrome de Lynch, le phénotype MSI a également été identifié dans des cancers sporadiques incluant ceux du colon, de l'estomac, de l'endomètre et plus rarement dans d'autres types de cancers tels que ceux des voies biliaires, du pancréas et des ovaires (Cortes-Ciriano et al., 2017). Dans ces contextes sporadiques, l'instabilité des microsatellites est souvent liée à l'inactivation épigénétique du gène *MLH1*, conduisant à une déficience acquise du système MMR (Imai and Yamamoto, 2008; The Cancer Genome Atlas Network et al., 2012). L'instabilité qui en résulte est un marqueur important pour la prédiction de la réponse à certains traitements, notamment l'immunothérapie. En effet, les tumeurs dites MSI-H (MSI élevé) présentent une forte charge mutationnelle, ce qui les rend plus susceptibles de répondre aux inhibiteurs de points de contrôles immunitaires (Le et al., 2017).

Sur le plan cytogénétique, les tumeurs MSI se distinguent par un caryotype généralement normal ou sub-normal, on parle de phénotype « chromosome-stable ». Contrairement à d'autres types de cancers qui présentent fréquemment des remaniements chromosomiques complexes et une aneuploïdie marquée, les tumeurs MSI montrent peu ou pas de changements similaires (Boland and Goel, 2010). Dans environ 75% des cas, les tumeurs MSI ne présentent pas de remaniements chromosomiques majeurs, ce qui contraste avec d'autres formes de cancers où l'instabilité chromosomique est un trait distinctif (Duval and Hamelin, 2002). Cette particularité cytogénétique souligne l'importance de l'instabilité microsatellitaire dans la genèse des tumeurs MSI, où l'accumulation progressive de mutations des microsatellites joue un rôle crucial dans la progression tumorale (Figure 12) (Cortes-Ciriano et al., 2017; Jonchère et al., 2024).

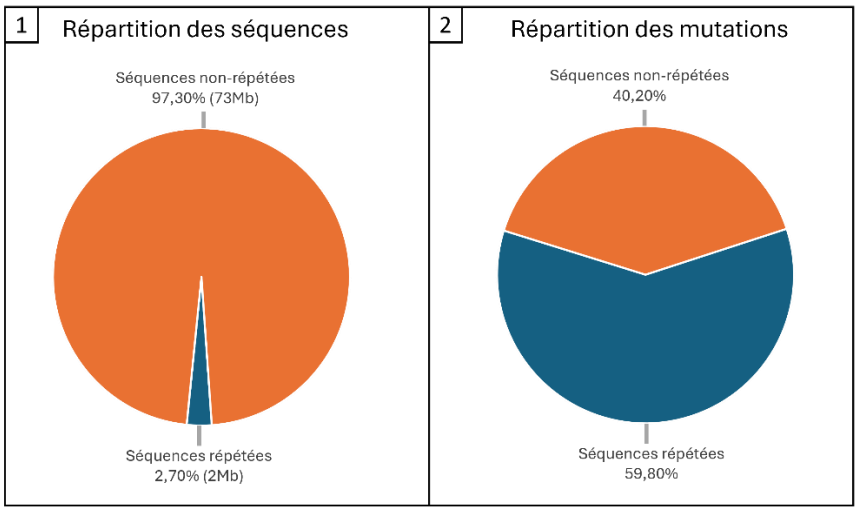

Figure 12 : Répartition des séquences ainsi que des mutations répétées et non-répétées dans les tumeurs MSI chez l'humain. 1. Représentation de la proportion des séquences répétées (bleu) et non-répétées (orange). 2. Représentation de la proportion des mutations affectant les séquences répétées (bleu) et non-répétées (orange) dans le cancer du côlon MSI. Adapté de (Jonchere et al., 2018).

L'épidémiologie des tumeurs MSI révèle des variations notables en fonction de la localisation du cancer et de son stade. Par exemple, dans le cancer colorectal, environ 15% des cas sont caractérisés par une instabilité des microsatellites, avec une proportion encore plus élevée observée dans les cancers

de l'endomètre (Boland and Goel, 2010). Cette variabilité épidémiologique reflète non seulement les différences biologiques sous-jacentes entre les types de tumeurs, mais aussi l'impact de facteurs environnementaux et génétiques sur la survenue de ces altérations (Hampel et al., 2005). En outre, la présence d'une instabilité des microsatellites a des implications pronostiques significatives. Dans les cancers colorectaux, par exemple, les tumeurs MSI sont souvent associées à un meilleur pronostic par rapport aux tumeurs sans instabilité, en raison de leur comportement moins agressif et de leur réponse accrue à l'immunothérapie (Jin and Sinicrope, 2021).

La description de ces phénotypes et de ses implications cliniques souligne l'importance de l'instabilité des microsatellites dans la compréhension et la prise en charge des cancers. Les recherches se poursuivent sur les mécanismes sous-jacents de cette instabilité, ainsi que leur rôle dans la progression tumorale et la réponse au traitement, car essentielles pour développer des stratégies thérapeutiques plus ciblées et efficaces (Boland and Goel, 2010; How-Kit and Ye, 2023).

3.2.2. Cancers MMRd - Déficience constitutionnelle du MMR

La déficience du système MMR pouvant être constitutionnelle, elle a été nommée différemment au fil du temps, tels que « *Childhood Cancer Syndrome* » (CCS) (Krüger et al., 2008), syndrome de Lynch III (Felton et al., 2007), ou encore CoLoN (*Colon tumours or/and Leukaemia/Lymphoma or/and Neurofibromatosis*) (Bandipalliam, 2005). Aujourd'hui, cette pathologie est plus connue sous le nom de syndrome de déficience constitutionnelle du système MMR (CMMRD - *Constitutional MisMatch Repair Deficiency*) depuis la mise en évidence des mécanismes sous-jacents du syndrome (Scott et al., 2007).

Le syndrome de Lynch est l'un des syndromes héréditaires les plus fréquents prédisposant au cancer, principalement colorectal, mais aussi à d'autres cancers tels que ceux de l'endomètre, de l'estomac et des voies urinaires. Il est causé par des mutations monoalléliques de type ponctuel dans les gènes du système MMR, principalement *MLH1* et *MSH2*, qui sont responsables de la majorité des cas. Des mutations dans *MSH6* et *PMS2* sont également observées, mais elles ont une pénétrance généralement plus faible (Senter et al., 2008; Truninger et al., 2005). Toutefois, les mutations du gène *PMS2* sont sous-représentées dans le syndrome de Lynch. Cela s'explique par la pénétrance réduite des mutations monoalléliques dans ce gène, ce qui rend leur contribution au risque de cancer moins évidente comparée aux mutations des gènes *MLH1* et *MSH2*. En revanche, le CMMRD est une maladie pédiatrique beaucoup plus sévère, causée par des mutations germinales bialléliques des gènes du système MMR. Dans environ 60 % des cas, des mutations bialléliques du gène *PMS2* sont identifiées, tandis que les 40% restants concernent les gènes *MSH6*, *MLH1* et *MSH2* (Wimmer et al., 2014).

Une différence importante entre ces deux syndromes est la viabilité embryonnaire. Alors que des mutations bialléliques dans les gènes *MLH1* et *MSH2* pourraient être létales à l'état embryonnaire (Wimmer and Etzler, 2008), les mutations bialléliques de *PMS2* ou *MSH6* permettent la survie et se manifestent par une forte prédisposition à des cancers pédiatriques. De plus, une corrélation entre le gène MMR impliqué et le spectre tumoral a été observée, notamment les tumeurs cérébrales sont plus fréquentes chez les porteurs de mutations bialléliques de *PMS2*, tandis que les hémopathies sont plus fréquentes chez les porteurs de mutations bialléliques de *MLH1/MSH2*. Le syndrome de Lynch est dominé par les mutations des gènes *MLH1* et *MSH2*, avec une apparition de cancers principalement à l'âge adulte, tandis que le CMMRD est associé à des mutations bialléliques, notamment dans *PMS2*, entraînant des cancers précoces et graves dès l'enfance (Peltomäki et al., 2023; Wimmer et al., 2014).

Enfin, le dépistage du syndrome de Lynch chez les patients atteints de cancers colorectaux avec MSI est crucial, car ce syndrome héréditaire prédispose à de multiples cancers et peut être transmis aux

descendants. Identifier ce syndrome permet de mettre en place un suivi précoce et ciblé pour les patients et leurs familles, d'adapter les traitements (comme l'immunothérapie pour les tumeurs MSI) et de prévenir d'autres cancers (Biller et al., 2019; Stoffel et al., 2015).

3.2.3. Maladies neuro-dégénératives

Bien que souvent moins étudiée que dans le contexte du cancer, la déficience du système MMR et les mutations microsatellites sont également liées à quelques maladies neurodégénératives.

Il existe notamment les maladies débilitantes dites à expansions de triplets qui sont un groupe de pathologies génétiques causées par une augmentation anormale du nombre de répétitions de séquences tri-nucléotidiques (triplets) dans certains gènes, qui ne causent pas d'intéruption du cadre de lecture (Richard et al., 2008). Ces expansions provoquent des dysfonctionnements cellulaires et des maladies, telles que le syndrome de Huntington, le syndrome du X fragile, l'ataxie spinocérébelleuse de type 1,2 et 3, la myotonie dystrophique, l'ataxie de Friedreich, ou encore la démence frontotemporale avec parkinsonisme, dont deux exemples sont développés ci-dessous (Mirkin, 2007).

Le syndrome de Huntington est une maladie neurodégénérative qui se traduit par d'importants troubles moteurs, cognitifs et psychiatriques évoluant jusqu'à la perte de l'autonomie, puis le décès. Cela est causé par une expansion somatique répétée du triplet CAG dans le gène *HTT*. Chez les individus sains, le gène *HTT* contient entre 6 et 35 répétitions CAG. Dans la maladie de Huntington, le nombre de répétitions dépasse 36, et plus le nombre de répétitions augmente, plus les symptômes apparaissent tôt et sont sévères (MacDonald et al., 1993). Des recherches récentes ont montré que le système MMR pourrait influencer la dynamique de cette expansion répétée. Plus spécifiquement, il a été démontré que MLH1 est essentiel pour les expansions CAG somatiques et qu'il favorise l'accumulation nucléaire de protéines de huntingtine dans les neurones striataux (Arning and Nguyen, 2021; Dragileva et al., 2009; Kovalenko et al., 2012). Il a également été démontré que MLH3, faisant partie du complexe MutLγ, est un facteur clé de l'expansion somatique (Kunkel and Erie, 2005; Polosina and Cupples, 2010). Cela conduit à une anticipation où les générations successives développent des symptômes plus précoces et plus sévères de la maladie de Huntington. Une déficience du système MMR pourrait donc accélérer cette expansion, exacerbant la progression des symptômes neurodégénératifs chez les patients atteints de cette maladie (Pinto et al., 2013).

Un autre exemple est le syndrome du X fragile, une maladie génétique causée par une expansion anormale des répétitions CGG dans le gène *FMR1*, situé sur le chromosome X. Lorsque le nombre de répétitions dépasse 200 (contre 5 à 44 chez un patient sain), le gène est méthylé et sa protéine FMRP n'est plus produite, ce qui perturbe le développement neuronal et entraîne des symptômes cognitifs et comportementaux. Transmise de manière récessive liée à l'X, la maladie touche principalement les hommes, n'en possédant qu'un seul. Les femmes peuvent être moins affectées, car elles possèdent un deuxième chromosome X non muté. Le phénomène d'anticipation génétique provoque une augmentation du nombre de répétitions à chaque génération, aggravant souvent les symptômes. Les signes cliniques incluent un retard intellectuel, des troubles comportementaux (hyperactivité, anxiété, stéréotypies), des caractéristiques physiques (visage allongé, grandes oreilles) et des troubles moteurs. Le diagnostic repose sur un test génétique pour détecter l'expansion de CGG et la méthylation du gène (Rajan-Babu et al., 2017).

Il existe toutefois d'autres pathologies qui ne font pas partie de la famille des maladies à expansion de triplet liées à une déficience du système MMR. Par exemple, la neurofibromatose de type 1 (NF1) est un trouble génétique autosomique dominant causé par des mutations dans le gène *NF1*, qui

entraîne des anomalies osseuses, des problèmes de coordination et sensoriels, ainsi que des lésions cutanées. Ce gène, riche en microsatellites, pourrait être sensible à l'instabilité génétique liée à une déficience du système MMR. Cette instabilité génétique pourrait augmenter le risque de tumeurs et contribuer à des processus neurodégénératifs neuronaux (Alotaibi et al., 2008; Gallinger et al., 2004; Raevaara et al., 2004). Le syndrome de Cockayne, une maladie neurodégénérative rare, résulte de mutations dans les gènes *ERCC6* et *ERCC8* impliqués dans la réparation de l'ADN. Ces mutations augmentent la sensibilité aux dommages de l'ADN, associée à une déficience du système MMR, entraînant une accumulation de mutations et aggravant l'instabilité génétique, particulièrement dans les cellules neuronales, ce qui contribue à la dégénérescence (Bertola et al., 2006; Laugel, 2013; Wei et al., 2015). Enfin, l'ataxie télangiectasie (A-T), une maladie neurodégénérative autosomique récessive, est causée par des mutations dans le gène *ATM*, crucial pour la réparation des cassures double brin de l'ADN. Cette défaillance, combinée à une déficience du système MMR, aggrave l'instabilité génétique et l'accumulation de mutations dans les cellules neuronales, contribuant ainsi à la dégénérescence progressive observée dans cette maladie (Derheimer and Kastan, 2010; Rothblum-Oviatt et al., 2016).

4. Facteurs influençant les mutations des microsatellites

Bien qu'il soit tentant de penser que tous les loci de microsatellites suivent les mêmes dynamiques évolutives, il est maintenant bien établi que les taux de mutations des microsatellites varient d'un locus à l'autre (Di Rienzo et al., 1998). Plusieurs facteurs peuvent expliquer ces différences dans leurs dynamiques évolutives : le nombre de répétitions, la séquence du motif, la longueur de l'unité de répétition, les séquences environnantes, les interruptions au sein du microsatellite, le taux de recombinaison, de transcription, etc… Certains de ces facteurs, les plus représentatifs, seront abordés ici.

4.1. Intrinsèques

4.1.1. Nombre de répétitions

Depuis les années 1990, des analyses de microsatellites clonés (Wierdl et al., 1997), des observations directes à partir de pédigrés connues (Brinkmann et al., 1998; Schlötterer, 1998; Schug et al., 1998a; Weber and Wong, 1993) ainsi que des études de population (Goldstein and Clark, 1995) indiquent avec confiance que le taux de mutation des microsatellites augmente en fonction du nombre de répétitions avec une tendance de type exponentiel (Baptiste et al., 2013; Eckert et al., 2002; Kelkar et al., 2010; Whittaker et al., 2003) (Figure 13).

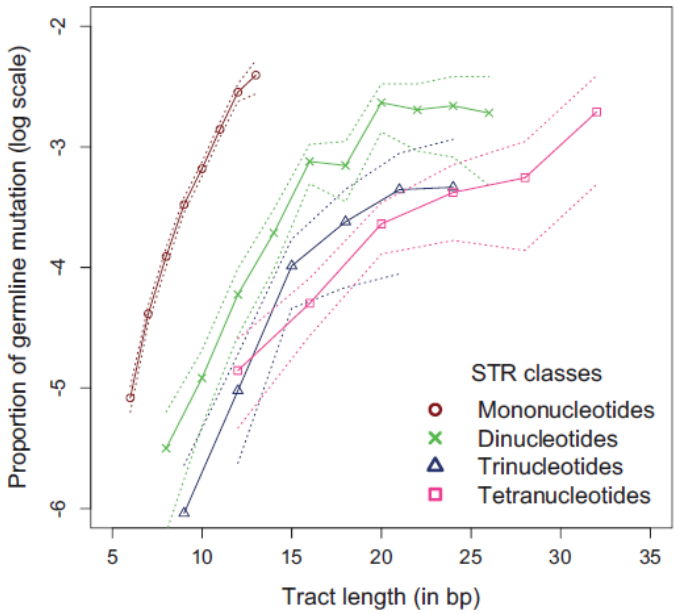

Figure 13 : Les taux de mutation des souches germinales pour les STR de longueur et de taille différentes. Les lignes pointillées représentent l'intervalle de confiance à 95 % de l'échantillonnage multinomial. Seules les longueurs de répétition avec un support de 2000 loci ont été tracées (Fungtammasan et al., 2015).

4.1.2. Taille du motif

De nombreuses études ayant examiné les taux de mutations de différents motifs de microsatellites, semblerait montrer que les motifs courts muteraient plus rapidement que les motifs longs (Chakraborty et al., 1997; Kruglyak et al., 1998). Cependant, certaines études montrent des motifs tetra-nucléotidiques mutant plus rapidement que les courts (Sun et al., 2012; Weber and Wong, 1993). Toutefois, la vérité semble être plutôt en la faveur de la première hypothèse, avec une diminution du taux de mutations, avec l'augmentation de la longueur du motif (Figure 14)(Baptiste et al., 2013; Kelkar et al., 2010; Schlötterer, 2000; Steely et al., 2022).

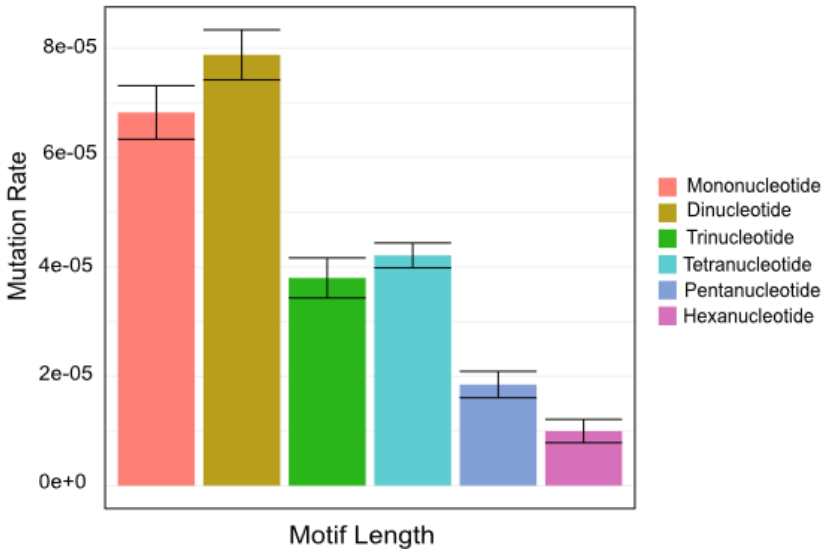

Figure 14 : Taux de mutation pour 2 747 répétitions courtes en tandem uniques génotypées, avec des longueurs de motif allant de 1 pb (mononucléotide) à 6 pb (hexanucléotide). Les taux de mutation diminuent généralement à mesure que la longueur du motif augmente, avec quelques exceptions (Steely et al., 2022).

4.1.3. Composition du motif

Dans différentes études réalisées, le pourcentage de GC ne semblait pas avoir de lien direct avec le taux de mutations des microsatellites, outre la stabilité pure de l'ADN (Balloux et al., 1998; Glenn et al., 1996) mais des études plus récentes montrent, à l'inverse, que ces zones sont plus sujettes aux mutations de par leur facilité à créer des structures secondaires favorisant les erreurs lors de la réplication (Kelkar et al., 2008). Par ailleurs, il a été montré que des motifs parfaits mutaient environ deux fois plus que les motifs imparfaits, ces derniers permettant par leurs inégalités un moindre glissement (Steely et al., 2022).

4.2. Extrinsèques

4.2.1. Systèmes de corrections

Le système MMR fonctionnel est lui-même un facteur de mutation des microsatellites puisqu'il ne corrige pas 100% des erreurs. En effet, même si lors d'une correction de mésappariement sur le brin matrice, induisant une potentielle délétion, le système MMR est très efficace, ce n'est pas le cas lors d'une correction sur le brin néo-synthétisé induisant une insertion (Figure 15). De plus, il semble que ce système corrige préférentiellement les boucles de mésappariement sur les motifs courts (Sia et al., 1997b).

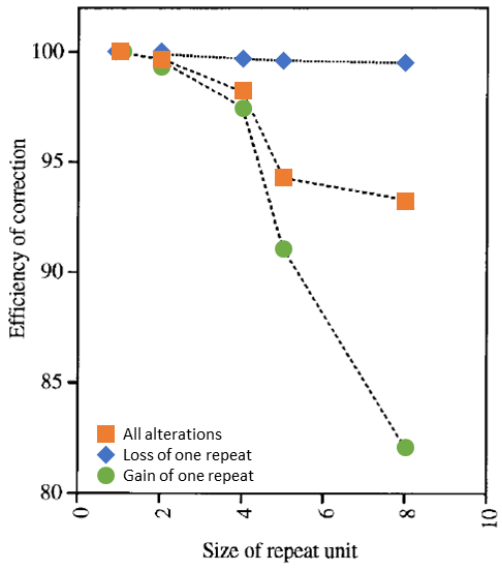

Figure 15 : Efficacité de la correction des mésappariements (pourcentage) en fonction de la taille de l'unité de répétition du microsatellite (en paires de bases). L'efficacité de la réparation a été calculée pour toutes les altérations (carré), les altérations conduisant à un gain d'une répétition (cercle) et les altérations menant à une perte d'une répétition (losange). Adaptée de (Sia et al., 1997b).

4.2.2. Organismes vivants (procaryotes/eucaryotes)

Chaque organisme possède et présente une proportion très différente de microsatellites et de manière globale, la prépondérance des microsatellites va du plus petit motif répété au plus grand pour un nombre de répétitions similaire. Dans le cas de *Escherichia coli*, son génome possède particulièrement peu de microsatellites par rapport à la levure et même à l'humain (Figure 16). Cela vient appuyer le fait que les microsatellites sont plutôt en expansion chez les eucaryotes et en contraction chez les

procaryotes, allant jusqu'à leur disparition (Ellegren, 2004). De plus, il est remarquable de noter que le seuil à partir duquel on définit un microsatellite diffère selon le type de microsatellite et l'organisme (Du et al., 2020) (Figure 16).

Ainsi, l'organisme au sein duquel se trouvent les microsatellites va également impacter les mutations microsatellites s'ajoutant aux autres facteurs de mutations présentés précédemment, créant ainsi un ensemble de facteurs responsables des mutations des microsatellites.

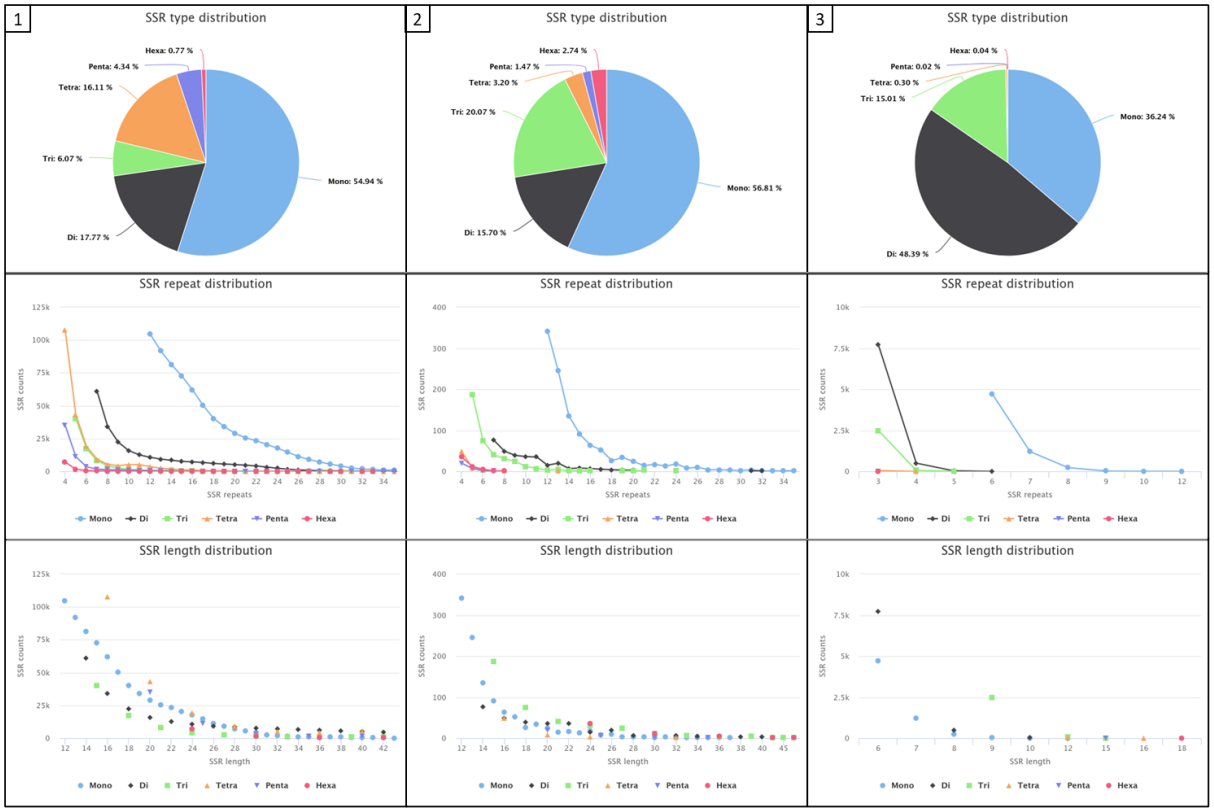

Figure 16 : Représentation des différents types de microsatellites présents à travers trois organismes vivants. De gauche à droite : Homo Sapiens, Saccharomyces cerevisiae, Escherichia coli. Données de PSMD.

4.3. Directionnalité des mutations

Enfin, il est important de prendre en compte le concept de directionnalité des mutations. Les mutations qui affectent les microsatellites, bien qu'aléatoires, présentent aussi une directionnalité, qui favorise préférentiellement soit les insertions, soit les délétions. Tout comme pour le taux de mutations, ce phénomène résulte de nombreux facteurs, incluant également la nature des séquences microsatellites, le mécanisme de mutation, la réparation, ainsi que les pressions évolutives spécifiques aux espèces et aux contextes génomiques. Par exemple, comme précédemment mentionné, dans de nombreux génomes eucaryotes, les mutations tendent à favoriser les insertions dans les séquences courtes, augmentant ainsi leur taille, alors que dans les microsatellites déjà longs, les délétions sont souvent prédominantes pour limiter leur instabilité (Ellegren, 2004; Tóth et al., 2000). Ces biais directionnels varient également selon les espèces : certains organismes présentent une forte tendance aux insertions, tandis que d'autres montrent une prédominance des délétions, reflet d'adaptations évolutives ou de différences dans les mécanismes de réparation de l'ADN (Sia et al., 1997b). De plus, des facteurs tels que la composition en nucléotides du microsatellite, la localisation génomique (intra- ou

intergénique) et l'influence des protéines de liaison à l'ADN peuvent moduler la directionnalité des mutations (Bhargava and Fuentes, 2009).

5. Les méthodes d'analyses des microsatellites

5.1. Génotypage & mutations microsatellites

La plupart des applications liées aux microsatellites nécessitent un génotypage, qui consiste en l'analyse de la variabilité des séquences des microsatellites par l'identification des différentes tailles de répétitions qui caractérisent les allèles. Ce processus permet de détecter des variations génétiques, donnant aux microsatellites une caractéristique de marqueur génétique dans diverses applications biomédicales, en raison de leur polymorphisme élevé. Ces applications concernent différents domaines tels que la conservation et la génétique des populations (Abdul-Muneer, 2014; Putman and Carbone, 2014), la cartographie génétique (Gulcher, 2012), la sélection assistée par marqueurs (Miah et al., 2013), la génétique évolutive (Kelkar et al., 2008; Sainudiin et al., 2004), l'analyse de parenté (Städele and Vigilant, 2016), l'identification en médecine légale (Gettings et al., 2015; Zhang et al., 2023), ainsi que le diagnostic de maladies, soit liées à l'expansion de répétitions (Lyon et al., 2010; Mirkin, 2007) soit dans des cancers MSI (Baudrin et al., 2018).

Un génotypage peut être complexe en raison des limitations des technologies de génotypage actuelles. Ces technologies reposent principalement sur des méthodes de mesure de taille, notamment par l'analyse des fragments par électrophorèse capillaire ou par des méthodes de séquençage où, à partir d'une séquence, il est possible d'en déduire un allèle (Baudrin et al., 2018; Vieira et al., 2016). L'avantage de ce dernier étant que, possédant l'intégralité de la séquence du microsatellite, il est possible de distinguer des microsatellites de même taille mais avec des variations dans leur séquence. L'objectif final est, dans tous les cas, de déterminer la taille des allèles.

On peut considérer deux types de génotypage ; d'une part, le génotypage classique, qui consiste à l'unique acquisition de l'information des deux allèles, et d'autre part, un génotypage quantitatif demandant une quantification de chaque allèle présent dans l'échantillon obtenu par séquençage ou par analyse de fragment. Ce dernier est grandement utilisé pour la détection de MSI dans le cancer (Boland and Goel, 2010; Willis et al., 2019) mais également dans des études plus fondamentales (Lee et al., 2012) (voir Travaux de recherche). Pour certaines applications qui ne requièrent qu'un génotype, les informations seront alors converties en génotype par des algorithmes bio-informatiques, notamment pour des données provenant de NGS, ou même issus de l'analyse de fragment (Tae et al., 2014).

Toutefois, le génotypage des microsatellites est rendu difficile par les artefacts issus du glissement de polymérase *in vitro*, connus comme les artefacts « stutter », ce qui a conduit au développement d'algorithmes de correction, particulièrement pour les données NGS (Ganschow et al., 2018; Gymrek et al., 2012; Highnam et al., 2013). En outre, les données NGS présentent d'autres défis spécifiques à la technologie de séquençage utilisée, tels que les erreurs de séquençage induites par les répétitions courtes (STR) et les homopolymères, ainsi que des limitations dans la longueur des microsatellites analysés avec des lectures courtes, ce qui complique l'identification correcte des STR et des indels (Cao et al., 2015; Fungtammasan et al., 2015; Ross et al., 2013; Tae et al., 2014; Tang et al., 2017; Zavodna et al., 2014).

5.2. Méthodes bas débit

5.2.1. PCR et électrophorèse en gel d'agarose

La PCR, inventée par le Dr. Kary Mullis en 1986 et pour laquelle il obtiendra un prix Nobel de chimie en 1993, est une méthode de biologie moléculaire primordiale. Elle permet, par l'alternance de température de manière cyclique et l'ajout d'amorces spécifiques, une amplification par une ADN polymérase d'une séquence d'ADN de manière exponentielle (Mullis et al., 1986). Elle peut ensuite être couplée à une méthode de visualisation telle qu'un gel d'électrophorèse, consistant en la migration des brins d'ADN à travers un gel d'agarose contenant un intercalant par l'application d'un courant électrique, permettant ainsi une séparation selon la taille des brins d'ADN. La visualisation est alors réalisée par exposition aux rayons UV qui activent la fluorescence de l'intercalant et ainsi des brins d'ADN (Aaij and Borst, 1972).

Cependant, dans le cadre d'études sur les microsatellites, cette méthode d'analyse n'est pas préconisée. En effet, la phase de PCR induit un polymorphisme d'amplification à cause des cycles répétés de réplication par la polymérase. De plus, la visualisation par gel d'agarose n'est pas suffisamment précise car elle permet une estimation de la taille d'une séquence mais pas sa détermination exacte ; il est alors impossible de voir et comptabiliser la perte ou le gain d'un ou plusieurs motifs, quelles que soient la nature ou la longueur de ceux-ci.

5.2.2. RFLP

La RFLP (*Restriction Fragment Length Polymorphism*), inventée par Sir Alec Jeffreys en 1985, est une méthode utilisée pour la détection de variations dans les séquences d'ADN pouvant être utilisées pour créer des « empreintes génétiques ». Elle repose sur la digestion par des enzymes de restriction de l'ADN en des fragments de tailles différentes selon les polymorphismes. Ces fragments sont ensuite séparés par migration sur gel d'agarose, dénaturés par de la soude, et transférés sur une membrane pour enfin les hybrider avec des sondes permettant leur visualisation (A. J. Jeffreys et al., 1985).

Cette technique n'impliquant pas d'étape de PCR est intéressante pour l'étude des microsatellites, cependant, la présence de tous les fragments et donc de toutes les variations de polymorphismes induit une difficulté d'identification précise des mouvements de mutations.

5.2.3. AFLP

L'AFLP (*Amplified Fragment Length Polymorphims*), inventée par les Dr. Pieter Vos et Marc Zabeau en 1995, est également une méthode de détection des polymorphismes de l'ADN comme la RFLP. Toutefois, cette technique, après digestion de l'ADN par des enzymes de restrictions, va hybrider des adaptateurs sur les brins, et les amplifier de manière sélective par PCR afin de détecter les polymorphismes sans nécessité de connaissance préalable des séquences (Vos et al., 1995).

Bien que contrairement à la RFLP, les polymorphismes soient plus identifiables à travers l'étendue du génome analysable, la présence de la phase de PCR vient rajouter des polymorphismes d'amplification, faussant ainsi les résultats lors de l'étude de microsatellites.

5.2.4. PCR avec amorces fluorescentes et électrophorèse capillaire

Lors d'une PCR, il est possible de coupler un marqueur fluorescent à une des deux amorces utilisées. Ainsi, il est possible d'utiliser des méthodes de visualisation plus précises telles que l'électrophorèse capillaire. Cette technique est basée sur le même principe que l'électrophorèse sur gel d'agarose, mais celui-ci est désormais intégré à des capillaires. Lors de la migration, plus lente et sur une plus longue distance, les brins d'ADN peuvent être séparés à la base près. Par l'excitation des marqueurs fluorescents, incorporés lors de la PCR, il est donc possible d'obtenir un électrophorégramme très précis.

Cependant, dans le cadre d'études sur les microsatellites, l'utilisation de cette méthode ne s'affranchit pas des polymorphismes d'amplification. De plus, même si la méthode de visualisation est beaucoup plus précise, un phénomène de *stutters* est visible. Ce phénomène consiste en l'apparition de bandes parasites présentes à un multiple de la taille du motif du microsatellite (si le microsatellite est un mono-nucléotide, les *stutters* apparaîtront à chaque multiple de 1 en amont et en aval du pic réel, si c'est un dinucléotide, les *stutters* seront présents à chaque multiple de 2, etc…)(Murray et al., 1993).

5.2.5. LT-RPA et électrophorèse capillaire

La RPA (*Recombinase Polymerase Amplification*), inventée par les Dr. Olaf Piepenburg, Colin Williams, Derek Stemple et Niall Armes en 2006, est une technique d'amplification isotherme permettant une amplification rapide et sensible de séquences cibles sans passage par les cycles thermiques de la PCR traditionnelle (Piepenburg et al., 2006).

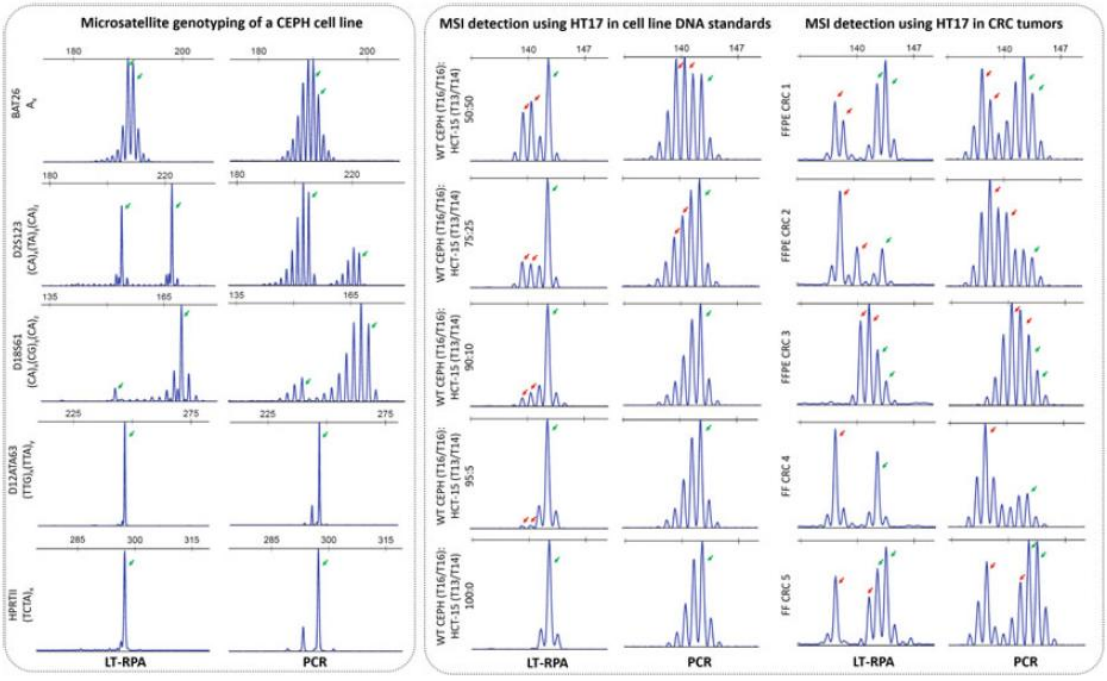

Figure 17 : Exemples de génotypage de microsatellites et de détection de l'instabilité des microsatellites (MSI) utilisant LT-RPA et PCR combinés avec électrophorèse capillaire. À gauche, un microsatellite mono-(BAT26), deux di-(D2S123 et D18D61), un tri-(D12ATA63) et un tétranucléotide (HPRTII) ont été génotypés en utilisant une lignée cellulaire lymphoblastoïde CEPH (NA12762). À droite, le microsatellite HT17 a été utilisé pour la détection de MSI dans l'ADN de cellules mélangées WT (lignée cellulaire CEPH, NA12812) et MSI (lignée cellulaire HCT-15), identifiant la limite de détection de MSI avec LT-RPA (5% de MSI) et PCR (25% de MSI) ainsi que pour la détection de MSI dans les tumeurs colorectales déficientes en MMR (trois échantillons FFPE et deux échantillons congelés frais (FF)). Les flèches vertes indiquent les allèles de microsatellites WT. Les flèches rouges indiquent les allèles mutants instables HT17. (Jeanjean et al., 2023)

Par la suite, la LT-RPA (*Low-Temperature RPA*) a été développée par notre laboratoire en 2019. Cette méthode a pour principe une RPA à faible température permettant ainsi la réduction drastique de la présence de *stutter* comparé aux amplifications par PCR de microsatellites (Daunay et al., 2019). En effet, en couplant les primers avec des marqueurs fluorescents, lors de la visualisation par électrophorèse capillaire, le génotypage des microsatellites est beaucoup plus précis et simple (Figure 17) (Jeanjean et al., 2023).

5.3. Méthodes de séquençage

5.3.1. Séquençage de première génération - Sanger

Dans les années 1970, alors que la biologie moléculaire et la génétique deviennent de plus en plus importantes, le Dr. Walter Gilbert aux États-Unis et le Dr. Frederick Sanger au Royaume-Uni ainsi que leurs équipes travaillent sur des méthodes pour permettre le séquençage de l'ADN. C'est en 1977 que les deux équipes mettent au jour deux méthodes différentes. La méthode des Dr. Gilbert et Maxam est fondée sur le principe de dégradation chimique sélective (Maxam and Gilbert, 1977), là où celle du Dr. Sanger est tournée vers la synthèse enzymatique sélective (Sanger et al., 1977). Ils seront récompensés pour leurs découvertes par le Prix Nobel de Chimie en 1980. Toutefois, la méthode de Sanger deviendra la méthode de référence pour les années à venir.

Le principe d'un séquençage Sanger ou à terminaison de chaîne se base sur le principe d'une PCR. Le brin d'ADN cible est mis en présence d'une amorce, des nucléotides libres (dNTPs), une ADN polymérase, mais également des didésoxyribonucléotides (ddNTPs) marqués par des fluorochromes différents pour chaque base (Figure 18). Ces ddNTPs ont la particularité d'être dépourvus de groupe 3' hydroxyle, empêchant alors l'incorporation de n'importe quel nucléotide à la suite. Ainsi, lors de la phase d'amplification, l'ADN polymérase incorpore les dNTPs mais aussi les ddNTPs, ce qui termine la chaîne d'amplification à des positions différentes. Ensuite, les brins amplifiés sont séparés par électrophorèse sur gel ou capillaire et lors de la lecture, grâce aux fluorochromes de couleurs différentes et aux fragments de différentes tailles, la séquence peut être déduite (Sanger et al., 1977).

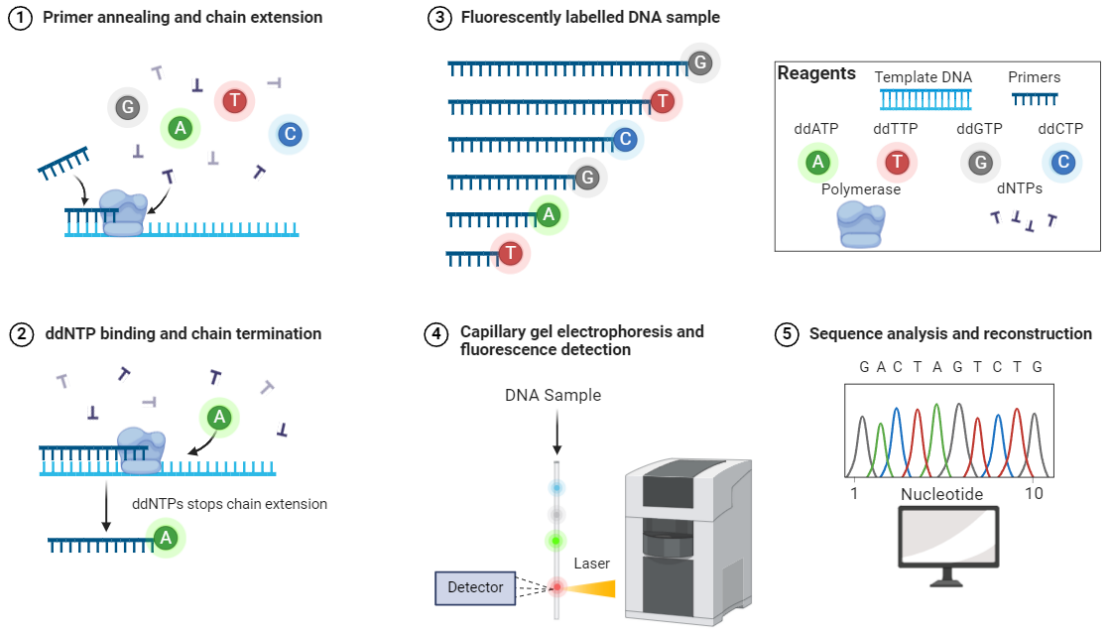

Figure 18 : Représentation schématique du principe de séquençage par la méthode de Sanger.

Toutefois, dans le cadre des microsatellites, et plus particulièrement des homopolymères, les séquences sont complexes à lire. En effet, la séquence en elle-même et la séquence suivant le microsatellite sont difficiles à analyser de par la répétition d'un même motif. Ainsi, réaliser un génotypage même quantitatif est ardu à cause de la superposition des séquences, contrairement au SNPs. C'est pourquoi la méthode d'analyse de fragment réalisée sur le même instrument, un séquenceur capillaire, a été préférée.

5.3.2. Séquençage de deuxième génération - Illumina

De nouvelles technologies permettant le séquençage de masse en parallèle, où des millions de fragments d'ADN sont séquencés simultanément, ont été développées à partir des années 2000, permettant une révolution dans le domaine de la génomique. La taille des fragments reste généralement courte (entre 50 et 300 bases), mais le débit est beaucoup plus rapide et moins coûteux. C'est ce que l'on appelle le séquençage de deuxième génération ou *short-reads*.

De nombreuses technologies telles que le pyroséquençage avec le Roche 454® ou l'ion Torrent™ de PGM® ont pu faire leurs preuves (Figure 19), mais dans le cadre de l'étude des microsatellites, nous nous sommes concentrés ici sur les technologies Illumina.

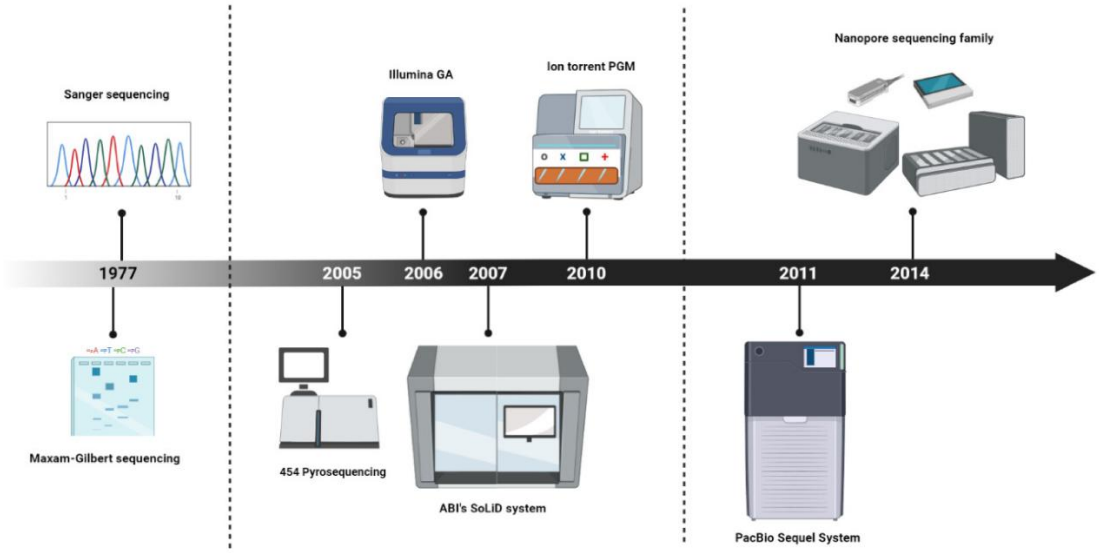

Figure 19 : Représentation des étapes du développement du séquençage tout au long de l'histoire de la biologie moléculaire (Athanasopoulou et al., 2021).

En 1998, une société américaine se lance dans le développement d'une nouvelle technologie de séquençage de deuxième génération. Cette technologie va permettre en 2013 de réduire les coûts d'un séquençage du génome humain entier à environ 7 000 dollars contre de 25 million en 2007, tandis que le prix du premier génome humain terminé en 2003 avait couté à près de 3 millards d'euros.

Le séquençage Illumina est une technologie de séquençage de l'ADN basée sur la synthèse par incorporation de nucléotides fluorescents (SBS, Séquençage par synthèse). Le processus classique commence par la fragmentation de l'ADN en petits morceaux, qui sont ensuite ligués à des adaptateurs spécifiques afin d'obtenir une librairie. Ces fragments sont fixés sur une surface solide (*flow-cell*) et

amplifiés pour créer des *clusters* d'ADN identiques par la méthode de la transition en pont (Figure 20). Cela signifie qu'après fixation sur la *flow-cell* (Figure 20.1), le brin complémentaire est synthétisé, le brin original éliminé (Figure 20.2) et le brin fixé à la *flow-cell* va alors créer un pont en hybridant son extrémité libre sur l'amorce adjacente (Figure 20.3). Un nouveau brin complémentaire (identique à l'original) est alors synthétisé pour créer un pont double-brin (Figure 20.4). Ce pont est ensuite dénaturé (Figure 20.5) et une nouvelle amplification est alors réalisée en suivant les mêmes étapes jusqu'à remplissage complet de la *flow-cell* (Figure 20.6). Enfin, seuls les brins identiques à l'original sont conservés et leurs extrémités 3' restantes sont bloquées (Figure 20.7).

Lors du séquençage, des nucléotides marqués avec des fluorochromes sont ajoutés (un seul pour l'iSeq, deux avec un codage à deux canaux pour le NextSeq et deux à quatre distincts pour le NovaSeq) (Figure 20.8). Une caméra haute résolution capture les signaux fluorescents émis par chaque nucléotide incorporé (Figure 20.9-11), permettant ainsi de déterminer la séquence des bases une à une (Figure 20.12). De plus, une lecture *paired-end* permet de séquencer les deux extrémités d'un fragment d'ADN. Après séquençage, le brin d'ADN est retourné par une transition en pont afin de commencer le séquençage à partir de l'autre extrémité. Ce type de séquençage permet, en alignant les deux extrémités, de positionner l'intégralité des inserts dans la librairie. Il est actuellement utilisé dans toutes les applications car il permet un meilleur positionnement du fragment séquencé.

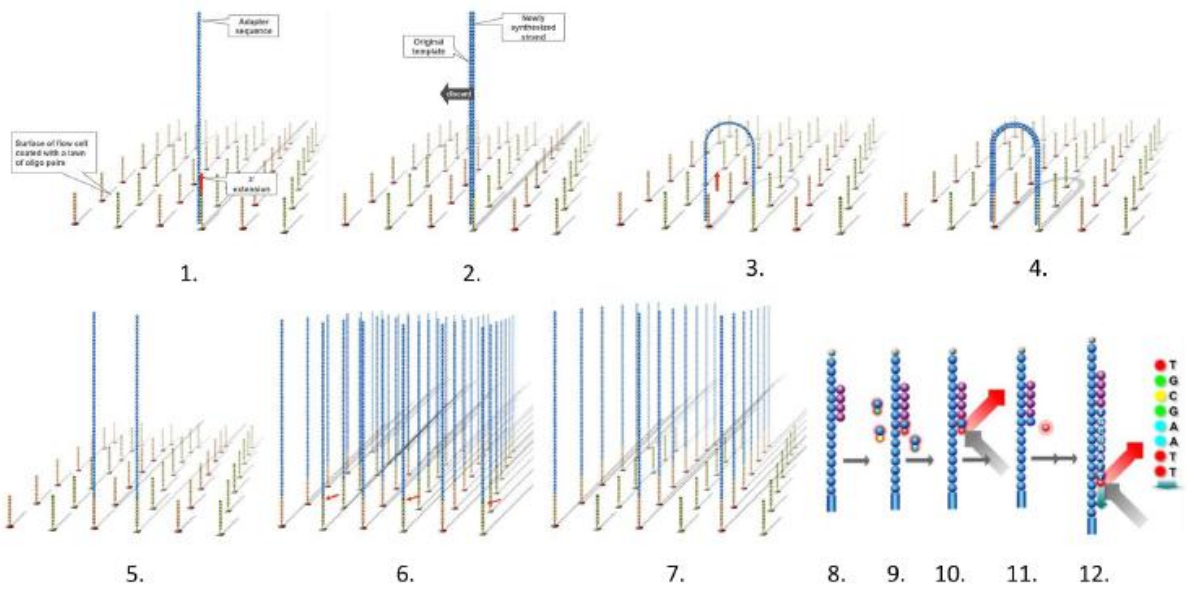

Figure 20 : Représentation du principe de séquençage Illumina (Adapté du site Illumina.com). 1. Ligation aux amorces de la flow-cell. 2. Synthèse du brin complémentaire et elimination du brin originel . 3. Hybridation à l'amorce adjacente. 4. Synthèse du brin complémentaire. 5. Séparation des brins. 6. Création du cluster par répétition des phases d'hybridation, de synthèse et de séparation. 7. Clivage des brins antisens afin de garder les brins des copies originels, et les extrémités 3' restantes sont bloquées. 8. Ajout des nucléotides complémentaires au brin d'ADN à séquencer. 9. Incorporation du nucléotide et lavage du milieu réactionnel. 10. Excitation des fluorochrome et émission du signal. 11. Clivage du fluorochromes. 12. Début d'un nouveau cycle.

L'utilisation de cycles de lecture réversible, isothermes et de fluorochromes détachables permet de lire de longues séquences d'ADN de manière précise et efficace, avec une haute couverture et une faible marge d'erreur. Dernièrement, Illumina a développé la technologie XLEAP-SBS permettant d'optimiser la méthode SBS classique. Elle s'associe aux derniers séquenceurs tels que le MiSeq i100 et les NextSeq 1000 et 2000 et le NovaSeq 6000 / 6000 plus pour obtenir des qualités de séquençage plus

proches d'une qualité Q40, et notamment une meilleure fidélité annoncée dans les zones répétitives telles que les microsatellites.

5.3.3. Séquençage de troisième génération

Après le développement des technologies permettant de séquencer beaucoup en peu de temps et à bas prix, un inconvénient reste présent : la nécessité de réassembler les séquences. En effet, les séquençages en *short-reads* présentent un defi concernant les variations structurales et les régions complexes, ou encore les homopolymères et les séquences répétées très longues, comme les microsatellites, dans les maladies à expansion de triplets. C'est pourquoi, des nouvelles technologies ont été développées, permettant la lecture de très longues molécules d'ADN (plusieurs milliers de bases) en une expérience sans nécessité d'amplification. C'est ce que l'on appelle le séquençage de troisième génération ou à longue lecture (*long-reads*) (Figure 19).

5.3.3.1. PacBio - SMRT

Le Dr. Stephen Turner créa en 2004 la société Pacific Biosiences® et y développa en 2009 la technologie SMRT (*Single Molecule Real-Time*) aussi appelée PacBio. C'est la première technologie de séquençage qui a permis la lecture de séquences supérieures à 10 000 bases.

Le séquençage PacBio obtient la séquence de la matrice cible durant son processus de réplication (Figure 21.1). Cette matrice double brin a été fermée en une boucle, la SMRTbell, par la ligation d'adapteurs bouclant des deux côtés de la séquence (Travers et al., 2010) (Figure 21.2). La boucle est ensuite chargée dans le puits d'une puce de la plus petite taille possible permettant la détection d'émission de lumière (*zero-mode waveguide*, ZMW). L'un des deux adapteurs, sans préférence, va venir s'hybrider à une polymérase immobilisée au fond de la ZMW pour démarrer la réplication (Eid et al., 2009). Le mélange réactionnel contenant des dNTPs marqués par quatre fluorochromes spécifiques des bases, à chaque intégration d'un dNTP par la polymérase, un flash lumineux est produit et détecté par les capteurs (Figure 21.1). La séquence cible étant une boucle, la réplication se fait donc en continu (*Continuous Long Read*, CLR), permettant d'acquérir plusieurs fois et dans les deux sens l'information de la séquence. Le CLR peut ensuite être séparé en multiple *reads*, appelés *subreads*, par la reconnaissance des adaptateurs bornant la séquence. En alignant tous les *subreads* obtenus, un consensus est créé permettant de corriger les erreurs de réplication de la polymérase ou de détection de la fluorescence (Koren and Phillippy, 2015) (Figure 21.2).

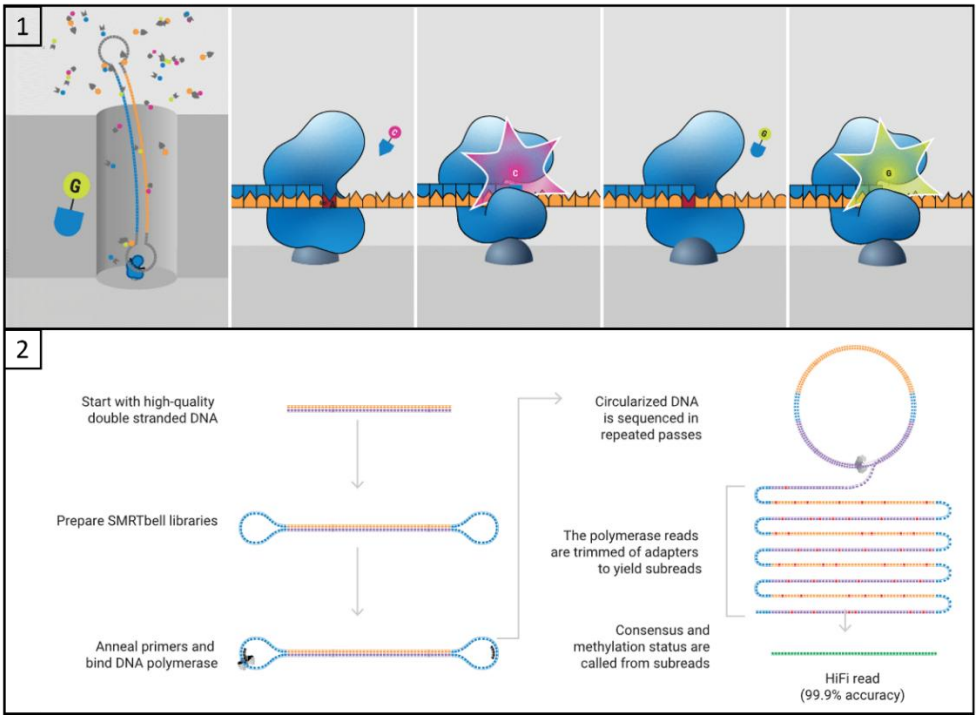

Figure 21 : Représentation du principe de séquençage de la technologie PacBio. 1. Processus de réplication et d'émission de fluorescence dans la ZMW par la polymérase immobilisée. 2. Workflow du séquençage jusqu'à obtention de la séquence consensus.

Cette technologie serait particulièrement utile pour caractériser les régions riches en microsatellites, notamment des longues expansions répétées impliquées dans des maladies génétiques comme la maladie de Huntington ou le syndrome de l'X fragile, souvent difficiles à lire avec d'autres technologies de séquençage en raison de leur taille et de leur nature répétitive. En effet, grâce à ses lectures longues et continues, elle permettrait de capturer intégralement des régions microsatellitaires complexes sans reconstruction, offrant une mesure précise du nombre de répétitions. De plus, sa capacité à produire plusieurs *subreads* pour chaque molécule permettrait de corriger les erreurs et d'obtenir un consensus fiable, essentiel dans les régions riches en répétitions. Cette technologie permet aussi de détecter les mosaïcismes somatiques, d'identifier d'éventuelles interruptions dans les séquences répétées, et d'explorer les mécanismes de mutation ou de stabilité des microsatellites (Dolzhenko et al., 2024; Liu et al., 2017; Logsdon et al., 2020; Ummat and Bashir, 2014).

5.3.3.2. Nanopore

Le concept du séquençage par nanopores a été exploré dès les années 1990 par des chercheurs tels que le Dr. David Deamer de l'Université de Californie à Santa Cruz ou encore le Dr. Daniel Branton de l'Université de Harvard. Mais cette technologie basée sur le changement de potentiel électrique lors du passage de l'ADN dans des pores extrêmement petits ne s'est concrétisée qu'en 2005 lors de la fondation d'Oxford Nanopore Technologies par le Dr. Hagan Bayley, puis en 2014 lors de la commercialisation du MinION, le premier séquenceur à nanopore possédant plus de 2 000 nanopores et séquençant à environ 450 bases par seconde.

La technologie du séquençage par nanopore repose sur l'utilisation d'un pore protéique d'une taille nanométrique (nanopore) servant de biocapteur en étant intégré dans une membrane polymère

électriquement résistante (Deamer et al., 2016) (Figure 22). Une tension constante est alors appliquée grâce à la présence d'une solution électrolytique afin de produire un courant ionique à travers le nanopore. Ainsi les molécules d'ADN (ou d'ARN) simple brin, naturellement chargées négativement, sont happées au travers du nanopore passant du côté « cis » (négatif) vers le côté « trans » (positif) (Jain et al., 2016). Le brin d'ADN traverse le nanopore de manière progressive grâce à une protéine motrice qui contrôle la vitesse de translocation. En passant dans le nanopore, la molécule d'ADN induit des variations dans le courant ionique correspondant à sa séquence (Wang et al., 2021). Grâce à la membrane servant de puce de détection, ses fluctuations sont récupérées et traduites par le biais d'un algorithme informatique permettant ainsi l'obtention de la séquence en temps réel (Leggett and Clark, 2017) (Figure 22). De plus, en même temps que sa fonction de contrôle de vitesse de translocation, la protéine motrice possède également une activité d'hélicase, permettant de dérouler des ADN double-brin et molécule simple brin afin de traverser le nanopore (Montel, 2018).

Le séquençage par nanopore est particulièrement utile pour l'analyse des microsatellites, comme par exemple les expansions de triplets impliquées dans la maladie de Huntington et la myotonie dystrophique. Grâce à sa capacité à lire de longues séquences en temps réel, cette technologie surmonte les limitations des technologies classiques, qui peinent à résoudre les répétitions longues et variables. Cette technologie permet de capturer intégralement les régions répétées, d'identifier les variations subtiles et de détecter les interruptions dans les répétitions (Miyatake et al., 2022).

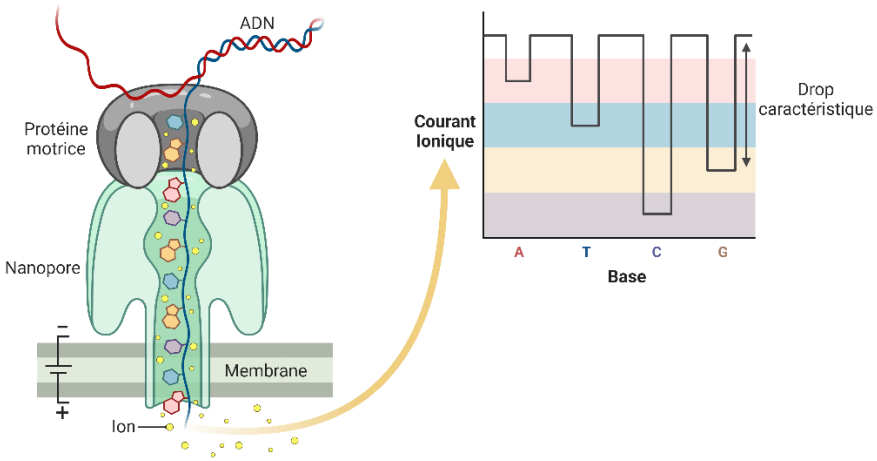

Figure 22 : Représentation schématique du principe du séquençage par nanopore.

5.4. Analyse et algorithmes bio-informatiques

Après la génération des séquences, l'étape cruciale suivante dans l'analyse génomique est l'alignement des *reads* sur une séquence de génome de référence, pour lequel plus de 80 outils spécialisés sont disponibles (Fonseca et al., 2012). Une fois cet alignement effectué, différents outils bio-informatiques sont utilisés pour détecter et caractériser les variations génétiques et les séquences répétitives.

5.4.1. Identification de microsatellites dans le génome

Sputnik, *RepeatMasker* et *Tandem Repeat Finder* sont des logiciels conçus pour identifier et analyser les séquences répétitives, telles que les STRs ou leurs homologues longues (Ellegren, 2004). *Sputnik* (Rudd et al., 2003) permet de localiser les microsatellites et autres STRs, tandis que *RepeatMasker* (Smit et al., 2013) identifie et masque les séquences répétitives connues pour améliorer la précision de l'alignement des reads et la détection des variations génétiques. *Tandem repeat Finder* (Benson, 1999) détecte spécifiquement les répétions en tandem, offrant une analyse précise même lorsque les répétions contiennent des mutations ou des indels.

5.4.2. Détection d'indels

Pour la détection des indels (insertion/délétion), des outils comme Dindel, *SAMtools mpileup* et *GATK* sont largement utilisés (Cao et al., 2015). Dindel (Albers et al., 2011) modélise les insertions et les délétions potentielles pour estimer les génotypes avec précision, tandis que *SAMtools mpileup* (Li et al., 2009) génère des fichiers de variantes en analysant la profondeur de couverture des séquences alignées. *GATK* (DePristo et al., 2011) est une suite d'outils polyvalente pour l'appel de variantes, offrant des fonctions avancées pour le recalibrage de la qualité des bases et la détection d'indels avec une grande précision. Ils peuvent également être employés pour analyser les variations STR.

5.4.3. Génotypage

Pour le génotypage des STRs à partir des données de séquençage haut débit, des outils spécialisés tels que *LobSTR*, *RepeatSeq* et *STRViper* sont utilisés. *LobSTR* (Gymrek et al., 2012) offre un alignement rapide et efficace pour identifier et génotyper les STRs, tandis que *RepeatSeq* (Highnam et al., 2013) corrige les erreurs de séquençage pour une estimation plus précise des allèles STR. *STRViper* (Cao et al., 2014), quant à lui, est optimisé pour détecter les variations dans les STRs, y compris les insertions et les délétions, en comparant les séquences avec des références génomiques. Dans les applications forensiques, des logiciels spécialisés comme *STRmix*™ et *TrueAllele*® sont utilisés (Orozco, 2023). *STRmix*™ repose sur un modèle probabiliste qui permet de résoudre des mélanges complexes d'ADN en déterminant les contributeurs multiples potentiels, même lorsque les échantillons sont dégradés ou contiennent une faible quantité d'ADN (Bright et al., 2016). *TrueAllele*®, quant à lui, utilise des modèles mathématiques avancés pour analyser les mélanges d'ADN, une méthode très utile dans les enquêtes criminelles où l'ADN est souvent contaminé (Bauer et al., 2020). Le logiciel *GeneMarker*™ *HID* permet également d'analyser les *loci* STR dans des mélanges complexes et est couramment utilisé dans les laboratoires forensiques pour l'identification humaine (Holland and Parson, 2011). Dans le domaine du génotypage de « *single-cell* », essentiel pour des applications comme la fertilisation *in vitro* (FIV), des outils comme Cell Ranger (10x Genomics) et Seurat sont employés pour analyser l'ADN ou l'ARN des « *single-cells* » (Niu et al., 2023; Raz et al., 2019).

PacmonSTR (Ummat and Bashir, 2014) est spécifiquement conçu pour identifier et génotyper les STRs à partir de données de séquençage de *long-reads*, telles que celles produites par les technologies PacBio et Oxford Nanopore Technologies (ONT). Grâce à l'utilisation de *long-reads*, *PacmonSTR* peut caractériser avec précision les STRs dans des régions génomiques complexes et détecter de grandes expansions de répétitions, particulièrement utiles pour l'étude de maladies génétiques liées aux STRs. *RepeatHMM* (Liu et al., 2017), quant à lui, utilise un modèle caché de Markov (HMM) pour génotyper les STRs en utilisant des données de séquençage à *short-* et *long-reads*. Cette approche probabiliste permet de corriger les erreurs de séquençage et d'améliorer la précision des appels d'allèles, même dans des zones à forte densité de répétitions (Mitsuhashi et al., 2019).

Toutes ces méthodes reposent sur trois approches principales : la profondeur de couverture, les *reads* fractionnées (*split reads*) et le mappage par paires d'extrémités (*paired-end mapping*). Les approches basées sur la profondeur de couverture sont efficaces pour détecter de grands indels (>50 pb) mais sont limitées par les biais de séquençage et les difficultés d'alignement dans les régions répétitives (Medvedev et al., 2009). Les méthodes utilisant les *reads* fractionnées offrent une sensibilité accrue pour la détection de nouveaux indels, mais leur efficacité est restreinte aux STR de courte longueur et dépend de la qualité d'alignement dans les zones répétitives. Enfin, l'alignement par paires d'extrémités permet une détection plus fiable sans nécessiter des *long-reads*, mais exige une couverture de séquençage élevée et une distribution précise des tailles de fragments pour identifier efficacement les petits indels (Cao et al., 2015).

5.4.4. Génotypage quantitatif

Plusieurs logiciels se distinguent pour l'analyse du génotypage quantitatif et particulièrement pour l'instabilité des microsatellites. *MSIsensor* et *MANTIS* sont des outils dédiés à l'évaluation de l'instabilité des microsatellites, avec des capacités quantitatives utiles pour des études de cancers et d'autres pathologies génétiques. Ils se démarquent notamment par leur capacité à analyser des échantillons tumoraux même en l'absence de données normales (Baudrin et al., 2018; Bonneville et al., 2017; Kautto et al., 2016; Niu et al., 2014; Renault et al., 2022). *TrueMark MSI Analysis Software*, développé par Thermo Fisher®, est spécifiquement conçu pour l'analyse des marqueurs MSI dans des échantillons fixés, avec des fonctionnalités d'automatisation adaptées à la clinique.

D'autres outils, comme *HipSTR*, *ExpansionHunter*, et *GangSTR*, se concentrent sur l'analyse des STRs à partir de données de séquençage en *short-reads*. *HipSTR* (Willems et al., 2017) utilise un algorithme d'alignement local flexible pour minimiser les erreurs d'appel d'allèles dans les régions répétitives, le rendant particulièrement efficace pour les études de population à grande échelle et les analyses cliniques. *ExpansionHunter* (Dolzhenko et al., 2024) est spécialisé dans la détection d'expansions de répétitions STR associées à des maladies génétiques, telles que la maladie de Huntington, en utilisant une base de motifs répétitifs pathogènes connus. *GangSTR* (Mousavi et al., 2019) combine des modèles probabilistes pour améliorer la précision du génotypage des STRs et est optimisé pour les grandes études de cohorte et les recherches sur les populations. Pour détecter les expansions de STRs rares ou pathogéniques, *STRetch* (Dashnow et al., 2018) utilise une approche de modélisation de la profondeur de couverture pour identifier des variations qui ne sont pas facilement détectables par d'autres outils, rendant cet outil particulièrement pertinent dans les études de maladies rares.

En complément, des outils comme *TRGT* et *adVNTR* apportent des méthodes spécialisées pour l'analyse des répétitions. *TRGT* (Dolzhenko et al., 2024) est conçu pour génotyper les « *tandem repeats* » avec une grande précision à partir de données de séquençage en *long-reads*, en intégrant des modèles de variation qui prennent en compte les caractéristiques spécifiques des STRs dans des contextes génomiques complexes. *adVNTR* (Bakhtiari et al., 2021) est particulièrement efficace pour détecter et génotyper les VNTRs dans les données de séquençage en *short-reads*, utilisant des algorithmes d'apprentissage automatique pour distinguer les variations des VNTRs avec une précision accrue, même dans des régions à forte densité de répétitions. *Straglr* (Chiu et al., 2021) est un autre outil focalisé sur l'identification des variations dans les séquences répétitives, mais il est spécialement conçu pour analyser les STRs en utilisant des *long-reads* afin de surmonter les défis posés par les erreurs de séquençage fréquentes dans ces données.

De plus, *LongTR* (Jam et al., 2024) est un logiciel émergent conçu pour analyser les STRs et les répétitions plus longues à partir de données de séquençage en *long-reads*, exploitant pleinement les

capacités des technologies de troisième génération pour identifier et caractériser les variations répétitives complexes à travers l'ensemble du génome.

Enfin, *PrancSTR* (Sehgal et al., 2024)est également utilisé pour le génotypage précis des STRs à partir de données de séquençage en *short-reads*. En intégrant des algorithmes de filtrage des lectures et une méthode de recalibration des scores d'alignement, *PrancSTR* réduit les erreurs de génotypage dans les régions répétitives et améliore la fiabilité des appels d'allèles, ce qui en fait un outil précieux pour les études génétiques de grande envergure nécessitant une analyse précise des STRs.

5.4.5. Alignement

En complément de ces outils, *LAST* (LocAL Sequence Alignment Tool) (Frith and Khan, 2018) est un outil performant pour l'alignement de séquences de grande taille ou fortement divergentes, comme les génomes entiers ou les *long-reads* issues des technologies PacBio ou ONT. Grâce à sa capacité à gérer efficacement des alignements avec des différences importantes, telles que des variations structurelles et des séquences à forte densité d'erreurs, *LAST* est particulièrement adapté pour des analyses génomiques complexes nécessitant une précision élevée et une réduction des faux positifs.

6. Méthodes d'analyse des taux de mutations

6.1. *In vitro*

6.1.1. HSV-tk mutation assay

Lorsque l'on veut réaliser une étude *in vitro* des mutations des microsatellites, la méthode de référence est la *HSV-tk mutation assay*, un *GAP-filling assay* développée par Kristin Eckert et Guang Yan (Eckert and Yan, 2000). Elle repose sur l'utilisation du gène *HSV-tk* (thymidine kinase du virus Herpès Simplex), qui convertit la floxuridine (FUdR) en une molécule toxique pour les bactéries. Dans cette technique, des microsatellites d'intérêt sont insérés dans une région codante ou régulatrice du gène *HSV-tk* (Figure 23.1). Toute mutation dans les microsatellites ou dans le gène entraîne une perte de fonction de HSV-tk, ce qui rend les bactéries résistantes au FUdR et permet ainsi de détecter les mutations.

La méthode débute par la préparation d'un vecteur contenant le gène *HSV-tk*, modifié pour inclure les séquences microsatellites étudiées. Après digestion enzymatique et hybridation, un hétéroduplex est formé. Cet hétéroduplex est ensuite transformé dans des bactéries *E. coli*, où il est répliqué. Une première phase de sélection par antibiotique permet d'isoler les bactéries ayant intégré le plasmide. Puis, une seconde sélection est réalisée avec la floxuridine : seules les bactéries ayant des mutations abolissant l'activité du gène *HSV-tk* survivent (Figure 23.2). Enfin, les plasmides sont extraits des bactéries résistantes pour être séquencés et analyser les mutations. Ce cycle d'étapes est répété plusieurs fois pour obtenir un volume de données suffisant (Baptiste et al., 2013). Cette méthode permet de mesurer directement les mutations survenant pendant une simple élongation et de distinguer les mutations affectant le brin matrice du brin néo-synthétisé.

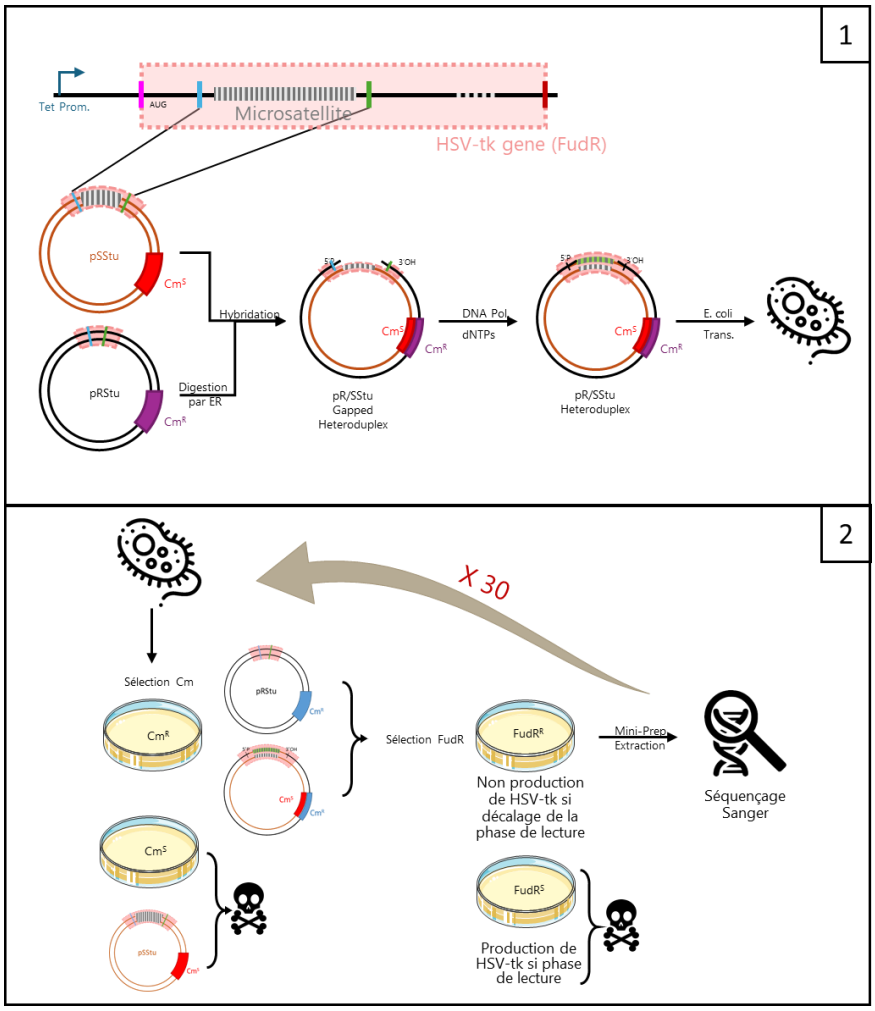

Figure 23 : Représentation schématique du principe du HSV-tk mutation assay. 1. Phase 1. 2. Phase 2.

6.1.2. Fréquences alléliques après PCR

Le calcul du taux de mutation par la fréquence allélique après PCR consiste à analyser les variations des allèles d'un locus spécifique avant et après amplification. D'abord, une séquence cible est amplifiée par PCR, et les produits sont analysés pour détecter les mutations, souvent par séquençage ou génotypage. La fréquence allélique avant PCR est mesurée dans un échantillon témoin, puis la fréquence des allèles mutants est déterminée après amplification. Le taux de mutation par dérapage est alors estimé en utilisant une approche de quasi-vraisemblance en comparant ces fréquences alléliques avant et après la PCR, en utilisant la formule du taux de mutation qui est généralement le rapport du nombre de mutations observées sur le nombre total d'individus ou d'événements mesurés. Les résultats doivent être interprétés en prenant en compte des facteurs comme les biais liés à la PCR (efficacité d'amplification et fidélité de la polymérase) ainsi que les erreurs de séquençage, qui peuvent fausser les conclusions (Schlötterer, 2000).

6.2. *In vivo / Ex vivo*

La *HSV-tk mutation assay* peut être appliquée à des expérimentations *in vitro* (comme montré précédemment) mais également en *in vivo* (bactéries, levures) et *ex vivo* (cellules humaines). Dans le cas

des expériences *ex vivo*, les cassettes géniques HSV-tk contenant des microsatellites sont alors clonées dans des vecteurs de navettes dérivés d'EBV-oriP notamment pour être electroporées dans des lignées cellulaires lymphoblastoïdes humaines (LCL) et ainsi étudier leurs mutations (Baptiste et al., 2013). Toutefois, cela reste une méthode lourde et laborieuse, nécessitant transformation bactérienne et séquençage de clone individuel. De plus, peu de molécules sont analysées par conditions expérimentales et les microsatellites long dits « matures » ne peuvent être analysés. Enfin, les mutations ayant lieu lors de la phase de lecture ne sont pas prises en compte.

Il existe d'autres méthodes comme par exemple, dans *E. coli*, des mutations spontanées dans des séquences de microsatellites peuvent être surveillées par des tests de survie de colonies bactériennes, et l'impact de mutations spécifiques sur la stabilité du génome est évalué par la fréquence de l'apparition de mutants résistants à un traitement donné. De plus, des tests génétiques sont utilisés pour identifier les mutations dans des cassettes contenant des gènes sensibles aux mésappariements, comme le gène *lacZ*. Pour les levures, comme *S. cerevisiae*, des stratégies similaires sont employées en utilisant des vecteurs de plasmides contenant des microsatellites et en analysant les populations clonales pour déterminer le taux de mutation (Hashem et al., 2002; Spampinato et al., 2009).

6.3. *In populo*

Les méthodes les plus couramment utilisées pour étudier les taux de mutation des microsatellites dans des contextes *in populo*, comme à partir de données familiales, populationnelles ou comparatives entre génomes, exploitent des approches bioinformatiques et statistiques adaptées à ces échelles. L'analyse des mutations *de novo* peut être réalisée grâce à des modèles familiaux (pédigrées), où les profils microsatellites des membres d'une famille sont comparés pour estimer les taux de mutation intergénérationnels (Antão-Sousa et al., 2022; Sainudiin et al., 2004). Ces études permettent de détecter des facteurs influençant les mutations (Schlötterer et al., 1998; Schug et al., 1998b), comme l'âge parental, mais nécessitent des données génétiques familiales exhaustives. Les études de population, basées sur des bases de données comme le *1000 Genomes Project* (Devuyst, 2015), permettent d'analyser la diversité génétique, d'évaluer l'instabilité des microsatellites et de détecter des signatures d'instabilité liées à des pathologies telles que le cancer. Ces analyses à grande échelle sont souvent réalisées à l'aide de logiciels comme *Arlequin* (Excoffier et al., 2007) ou *STRUCTURE* (Porras-Hurtado et al., 2013), mais dépendent de la taille et de la représentativité des échantillons. Les comparaisons entre génomes, qu'elles soient intra-populationnelles ou interspécifiques, sont également largement utilisées pour identifier les variations alléliques et estimer les taux de mutation ou les pressions évolutives. Ces études, souvent appuyées par des outils d'alignement comme *BLAST* (*Basic Local Alignment Search Tool*) (McGinnis and Madden, 2004; Ye et al., 2006), permettent de mettre en lumière les mécanismes évolutifs des microsatellites dans différents contextes biologiques. Enfin, les approches basées sur des simulations et modèles statistiques, comme celles utilisant des simulateurs génétiques (*msprime*, *SLiM*)(Baumdicker et al., 2022; Haller and Messer, 2017) ou des méthodes d'inférence bayésienne (*BEAST*, *MSVAR*) (Bouckaert et al., 2019; Girod et al., 2011), permettent de tester divers scénarios mutationnels et d'estimer les taux de mutation sur la base de données théoriques ou empiriques. Ces méthodes, bien que puissantes, dépendent fortement des hypothèses formulées dans les modèles.

7. Les applications des microsatellites

Les microsatellites, jouant un rôle essentiel dans de nombreuses disciplines de la biologie et de la génétique, sont de très bons marqueurs génétiques. Grâce à leur polymorphisme et leur présence dans de nombreux emplacements des génomes, ils constituent des outils puissants pour étudier la diversité génétique et les mécanismes évolutifs. Ces marqueurs trouvent des applications variées, notamment en cartographie génétique, en génétique évolutive et des populations, dans les études de biodiversité, les empreintes génétiques, ainsi que dans l'analyse des maladies humaines.

7.1. La cartographie génétique

En raison de leur forte variabilité et de leur distribution abondante dans le génome, les microsatellites sont idéaux pour la cartographie génétique. Ayant pour but de déterminer la position relative des gènes sur un chromosome et pour établir la distance génétique entre eux, les microsatellites sont utilisés pour suivre la transmission de segments chromosomiques au sein de familles ou de populations (Gulcher, 2012). En comparant les génotypes des microsatellites entre les individus, il est alors possible d'identifier des associations entre certains marqueurs et de traits phénotypiques spécifiques, tels que des maladies héréditaires ou des caractéristiques agronomiques (Solignac et al., 2004). Cette approche permet de localiser approximativement les gènes responsables de ces traits sur le génome.

La cartographie génétique basée sur les microsatellites est particulièrement précise pour les espèces où les ressources génétiques sont limitées, car elle ne nécessite pas de séquençage complet du génome. Elle est couramment utilisée dans la recherche sur les plantes (Song et al., 2005; Zhang et al., 2005) et les animaux pour améliorer les programmes de sélection et pour comprendre la base génétique de traits complexes (McComish et al., 2024).

La carte physique peut également être basée sur les microsatellites car ils forment des balises idéales pour jalonner le génome. Dans le cadre de la construction d'une carte physique, les microsatellites servent de points de repère stables pour localiser des régions spécifiques du génome, permettant d'établir des correspondances précises entre des fragments d'ADN et leur position sur les chromosomes (Cann, 1992; Cohen et al., 1993; Weissenbach, 1993a). En outre, les microsatellites facilitent le liage génétique entre les cartes physiques et les cartes génétiques (basées sur les taux de recombinaison), permettant une intégration des données génomiques.

En résumé, les microsatellites facilitent la détection de liaisons génétiques et la construction de cartes génétiques, ce qui contribue à une meilleure compréhension de l'organisation du génome et à la découverte de gênes d'intérêt.

7.2. Génétique évolutive

En génétique évolutive, les microsatellites fournissent des outils puissants pour étudier la diversité génétique, les relations phylogénétiques et les processus évolutifs au sein et entre populations. Leur haute variabilité génétique et leur large distribution dans le génome en font des marqueurs idéaux pour comprendre comment les populations évoluent au fil du temps.

Les microsatellites permettent d'évaluer la diversité génétique au sein des populations en mesurant la variation des allèles. Cette diversité est un indicateur clé de la santé évolutive d'une

population, car une haute diversité génétique est souvent associée à une meilleure adaptabilité aux changements environnementaux. En comparant les profils des microsatellites entre différentes populations, il est possible d'identifier des goulots d'étranglement génétiques, des événements de dérive génétique ou des flux génétiques, qui influencent l'évolution de la population (Carreras et al., 2007; Gugala et al., 2016).

Ils peuvent aussi aider à reconstruire les relations phylogénétiques entre des espèces ou populations. En analysant les similitudes et les différences dans les microsatellites (motifs, tailles, longueur…) entre différentes espèces, il est possible de déterminer les relations de parenté et de retracer les trajectoires évolutives. Par exemple, ils permettent d'identifier des lignées distinctes au sein d'une même espèce ou de clarifier les relations entre espèces proches, contribuant ainsi à la compréhension de l'arbre de la vie (Karssene et al., 2019; Tumendemberel et al., 2019).

Les microsatellites peuvent également être utilisés pour identifier des loci impliqués dans l'adaptation à des environnements spécifiques. En associant la variabilité des microsatellites avec des traits phénotypiques adaptatifs, il est possible de localiser des régions du génome soumises à une sélection naturelle. Cela aide à comprendre comment les populations s'adaptent à des conditions écologiques variées, telles que la température, l'altitude, ou la disponibilité des ressources (MANEL et al., 2012; Müller et al., 2024; Wijfjes et al., 2024).

Également, en analysant la divergence génétique entre populations qui ont été géographiquement isolées, les microsatellites peuvent être utiles pour étudier la spéciation, processus d'émergence de nouvelles espèces. Il est en effet possible de détecter les signes précoces de spéciations et ainsi déterminer si des populations sont entrain de diverger génétiquement à un tel point qu'elles ne pourront plus se reproduire entre elles (Maletzky et al., 2009; Takuno et al., 2019; Teeter et al., 2008).

Enfin, les microsatellites permettent de reconstituer l'histoire démographique des populations, notamment en identifiant des événements passés de migration, de colonisation ou de déclin démographique. Par exemple, ils peuvent être utilisés pour retracer les routes migratoires des espèces humaines anciennes ou pour comprendre l'expansion des populations animales dans de nouveaux territoires (Perrier et al., 2013; Rossi et al., 2020; Szövényi et al., 2009).

7.3. Génétique des populations et études de la biodiversité

En génétique des populations, les microsatellites jouent un rôle prépondérant en tant que marqueurs génétiques hautement informatifs pour étudier les structures génétiques, la dynamique des populations et le processus évolutif au sein des populations (Putman and Carbone, 2014).

Ils permettent d'analyser la structure génétique des populations, c'est-à-dire la répartition de la variation génétique à travers l'espace et les groupes sociaux. Ils sont utilisés pour identifier les sous-populations (ou groupe génétiquement distincts) et pour quantifier le degré de différenciation épigénétique entre elles. Par exemple, des valeurs de FST (mesure de différentiation génétique) peuvent être calculées à partir de données de microsatellites pour déterminer dans quelle mesure les populations sont isolées les unes des autres, ce qui est essentiel pour comprendre les flux géniques et les barrières reproductives (Piotti et al., 2013; Urzi et al., 2021).

Les microsatellites peuvent aussi être utilisés pour détecter des événements historiques qui ont affecté la taille effective des populations où une population subit une réduction drastique de sa taille. Ces événements laissent des signatures génétiques détectables par la réduction de la diversité des

microsatellites. En les étudiant on peut ainsi comprendre comment les populations ont répondu à des pressions environnementales ou anthropiques (Schmidt-Küntzel et al., 2018; Weinberger et al., 2021).

Les microsatellites peuvent être utilisés pour mesurer la diversité génétique au sein et entre les populations. La diversité génétique, souvent évaluée par le nombre d'allèles et les fréquences alléliques, est un indicateur clé de la santé génétique d'une population. Les populations avec une diversité génétique élevée sont généralement plus résistantes aux maladies et plus capables de s'adapter aux changements environnementaux. Les microsatellites permettent donc de surveiller la variabilité génétique, d'identifier les membres d'une population, de définir les populations à faible diversité et ainsi prioriser les efforts de conservation (Dou et al., 2016; Fernando et al., 2006).

Dans cet objectif, les microsatellites permettent l'identification de populations en danger en évaluant la réussite ou non des programmes de réintroduction, en élaborant des stratégies de gestion génétique. Par exemple, en surveillant la diversité génétique des populations réintroduites, les microsatellites peuvent aider à évaluer si ces populations maintiennent un niveau de diversité suffisant pour survivre à long terme ou s'il y a eu croisement avec une population déjà présente. De plus, ils sont utilisés pour définir des unités de gestion distinctes (ESU, *Evolutionarily Significant Units*) dans les espèces, en identifiant des populations ou des sous-populations génétiquement distinctes qui nécessitent une gestion séparée. Ces unités sont cruciales pour élaborer des plans de conservation spécifiques à chaque population, surtout lorsque des populations géographiquement proches présentent des différences génétiques importantes dues à l'isolement ou à la sélection locale (Abdul-Muneer, 2014; Rutledge et al., 2017).

7.4. Empreinte génétique

7.4.1. Application non humaine

7.4.1.1. Animaux : certification de races (pedigree) et Parentalité

Dans le domaine très prisé des animaux de race (canines, félines, équine…), les microsatellites sont très utilisés pour la certification de pedigree (Radko and Podbielska, 2021). En raison de leur haute résolution, ils peuvent distinguer les relations de parenté et d'ascendance d'individus au sein d'une population (Städele and Vigilant, 2016). De plus, ces informations sont importantes pour la gestion des populations en captivité et pour les programmes de conservation, où il est essentiel de minimiser la consanguinité et de maintenir la diversité génétique (Buono et al., 2022; Pei et al., 2018).

7.4.1.2. Végétaux : Sélection assistée par marqueur, Plantes, certification de graines

Les microsatellites sont également utilisés dans le domaine végétal, notamment pour la sélection assistée par marqueur (SAM), un outil précieux dans l'amélioration génétique des cultures. Cette approche permet de certifier des lots de graines, la qualité des semences en s'assurant qu'elles portent les bons caractères ou variations génétiques souhaitées (Gupta and Varshney, 2000). Grâce à leur haute variabilité et leur large distribution dans le génome, les microsatellites sont également employés pour étudier la diversité génétique des plantes, analyser les populations naturelles et menacées, et pour la gestion des ressources génétiques dans les programmes de conservation (Jiménez et al., 2010; Wu et al., 2020).

7.4.2. Sciences médico-légales

Les microsatellites possédant un nombre de répétitions variant considérablement d'un individu à l'autre, même entre parents proches, il est possible de les utiliser comme marqueurs de différentiation. Cette variabilité élevée est utilisée en médecine légale pour identifier un individu de manière unique (Gettings et al., 2015). Cette propriété exploitée pour générer un profil génétique unique est appelée «empreinte génétique » et est pratiquement unique à chaque personne (Keerti and Ninave, 2022).

Ainsi, un panel de plusieurs *loci* (généralement entre 13 et 20) est analysé, pour chaque *locus*, le nombre de répétitions est déterminé, créant ainsi un profil génétique composé de plusieurs valeurs numériques. Ce profil est unique pour chaque individu (exception des vrais jumeaux) (Zhang et al., 2023), permettant ainsi d'associer de manière fiable un échantillon biologique à une personne spécifique. Une fois ces profils générés, ils peuvent être comparés à d'autres profils dans une base de données ADN ou directement à des échantillons prélevés sur des suspects, des victimes ou des scènes de crime. Si les profils correspondent, cela peut indiquer que l'échantillon biologique provient de la même personne ou de personnes très étroitement apparentées. En médecine légale, cette correspondance peut être utilisée pour établir des liens entre un suspect et une scène de crime, ou pour identifier des victimes inconnues (Wang, 2023). Les systèmes utilisant des microsatellites, comme le système CODIS (*Combined DNA Index System*) aux États-Unis, sont devenus des standards mondiaux.

Au-delà même de l'identification de personnes recherchées, en comparant le profil microsatellites d'un enfant et de ses parents présumés, ou inversement, il est possible de déterminer les relations de parenté avec une précision très élevée (Chen, 2012). L'un des avantages des microsatellites est leur capacité à fournir des probabilités statistiques très élevées en cas de correspondance. Par exemple, la probabilité que deux personnes non apparentées partagent le même profil microsatellite sur 13 *loci* est extrêmement faible, souvent de l'ordre de 1 sur plusieurs millions ou plus. Cette spécificité permet non seulement d'identifier des suspects, mais aussi d'exonérer des innocents en cas de non-correspondance (Butler, 2006).

7.5. Maladies humaines

Les microsatellites jouent un rôle important dans l'étude et la détection de certaines maladies chez l'homme. Ces études permettent de déterminer si un microsatellite particulier est co-transmis avec une maladie spécifique au sein d'une famille, indiquant ainsi que le gène responsable de la maladie est situé à proximité de ce microsatellite sur le même chromosome (Iossifov et al., 2008; Weissenbach, 1993b). Par exemple, dans des maladies monogéniques, comme la chorée de Huntington, des microsatellites situés près du gène *HTT* sont utilisés pour suivre la transmission de l'allèle muté au sein des familles à risque (Gusella, 1984).

Près de 30 maladies héréditaires chez l'homme résultent d'une augmentation du nombre de copies de microsatellites dans l'ADN génomique (Mirkin, 2007). Par exemple, la dystrophie myotonique de type 1 (DM1) est liée à l'augmentation du nombre de répétitions de CTG dans le gène *DMPK* (Cho and Tapscott, 2007), ou encore pour le syndrome du X fragile, une expansion du triplet CGG dans le gène *FMR1* conduit à la maladie (Lyon et al., 2010). Leur détection est donc primordiale.

Les microsatellites sont également utilisés dans le diagnostic prénatal et le diagnostic génétique préimplantatoire (DPI). En étudiant leur présence dans les cellules embryonnaires ou fœtales, il est possible de détecter des anomalies génétiques ou chromosomiques avant la naissance. Cette technique est utilisée pour des maladies génétiques connues dans les familles, où les microsatellites peuvent être

analysés pour déterminer si l'embryon ou le fœtus a hérité de l'allèle mutant (Hu et al., 1996; Lim et al., 2023). Par exemple, pour des maladies comme la fibrose kystique, le DPI peut être utilisé pour sélectionner des embryons non porteurs de la mutation *CFTR*, si on fait partie d'une famille à risque, en utilisant des microsatellites pour confirmer l'absence du gène muté (Morral et al., 1993).

Enfin comme présenté précédemment, l'instabilité des microsatellites (MSI) est un phénomène où le nombre de répétitions microsatellitaires change au cours de la division cellulaire en raison de défauts dans le système de réparation des mésappariements d'ADN. Ce phénomène est couramment observé dans certains cancers, notamment les cancers colorectaux associés au syndrome de Lynch (Hampel et al., 2005). La détection de MSI est utilisée à la fois pour diagnostiquer ces cancers et pour prédire la réponse au traitement (Baudrin et al., 2018).

Une fois qu'un patient est diagnostiqué avec une maladie liée à un microsatellite, ces marqueurs peuvent être utilisés pour surveiller l'évolution de la maladie et l'efficacité des traitements. Par exemple, dans le cadre du suivi du cancer, les changements dans le profil MSI peuvent indiquer une réponse ou une résistance à la thérapie (Baudrin et al., 2018; Evrard et al., 2019).

8. Introduction du sujet de thèse

8.1. Organisme d'accueil

8.1.1. Histoire de la Fondation Jean Dausset - CEPH

Le Centre d'Étude du Polymorphisme Humain (CEPH) a été fondé en 1984 par le Pr. Jean Dausset (Prix Nobel de Médecine 1980), en collaboration avec les Pr. Daniel Cohen et Howard Cann, et le Dr. Mark Lathrop. Leur objectif était de coordonner la première collaboration internationale pour cartographier le génome humain. Une collection d'ADN de 40 familles de référence a été mise à disposition de la communauté scientifique, permettant de créer plus de dix cartes génétiques dans les années 1990.

En collaboration avec l'Association Française contre les Myopathies (AFM), le CEPH a aussi contribué à la création du Généthon, qui par la suite a permis au CEPH de réaliser la cartographie physique du génome via des chromosomes artificiels de levure (YAC), sous la direction du Pr. Daniel Cohen. Ces cartes ont permis de localiser des gènes liés à des maladies génétiques, comme le gène de la glucokinase associé au diabète MODY.

Le modèle de collaboration du CEPH a inspiré des projets tels que le séquençage du génome humain, HapMap et 1000 Genomes. Le CEPH a aussi initié des études sur les phénotypes complexes, comme la recherche de gènes liés à la longévité, en 1991. Ces études sont devenues une partie importante de domaine de recherche avec tout particulièrement l'identification de d'*APOE*, premier gène de la longévité identifié chez l'humain, grâce à une étude de centenaires, au CEPH (Schächter et al., 1994).

Devenu Fondation Reconnue d'Utilité Publique en 1993, le CEPH a poursuivi des recherches majeures, notamment l'identification du gène *CARD15/NOD2* pour la maladie de Crohn en 2001 sous la Direction du Pr. Gilles Thomas, et la mise en place du panel HGDP-CEPH pour l'étude de la diversité génétique des populations humaines lors d'une collaboration entre le Pr. Howard Cann et le Pr. Cavalli-Sforza (Université de Stanford, USA).

Sous la direction du Pr. Mark Lathrop, la Fondation Jean Dausset-CEPH a été un partenaire clé dans le Plan National de Génomique du Cancer, aboutissant à l'identification de gènes liés au cancer du

poumon et au développement d'une collection de prélèvements sur le cancer du sein. Elle a aussi contribué au Plan Alzheimer, identifiant en 2009 les gènes *CLU* et *CR1* associés à cette maladie.

Actuellement, la Fondation est impliquée dans la recherche sur des maladies complexes et divers types de cancer, exploitant les technologies de séquençage de nouvelle génération (NGS). Dirigée depuis 2012 par le Dr. Jean-François Deleuze, le CEPH a coordonné le LABoratoire d'EXCellence en GENomique MEDicale (GenMed) pour identifier des gènes de maladies, personnaliser les traitements, développer des réseaux scientifiques et contribuer aux réflexions éthiques et à la formation médicale jusqu'en 2022. Elle coordonne actuellement le projet Agenomics (2023-2026) qui vise à exploiter des données génomiques de centenaires et de cohortes longitudinales pour le développement de stratégies ciblant le vieillissement. Elle accueille également l'infrastructure BioCF (Biobanque Cohortes Française, 2021-2029), qui regroupe l'ensemble des échantillons issus des cohortes nationales en population générale. Cette cohorte regroupe un ensemble d'échantillons biologiques et de données associées issues de cinq grandes cohortes françaises, totalisant près de 400 000 volontaires suivis sur plusieurs décennies. Cette infrastructure intègre des analyses génétiques avancées (650 000 SNPs pour 230 000 individus), favorisant des recherches comme les GWAS et l'étude des interactions gène-environnement. Elle vise à renforcer la compétitivité de la France dans un contexte international en rattrapant des initiatives telles que le UK Biobank, tout en couvrant toutes les tranches d'âge, de la naissance à la vieillesse.

8.1.2. Laboratoire de Génomique

Le laboratoire de Génomique, dirigé par le Dr. Alexandre How-Kit, mène des projets de recherche fondamentale, translationnelle et technologique en utilisant des approches (épi)génétiques et (épi)génomiques. Ces travaux s'appuient sur des technologies de séquençage de première, deuxième et troisième génération, ainsi que sur des méthodes de génotypage de bas, moyen et haut débit.

Le programme de recherche du laboratoire s'articule autour de deux axes principaux : l'étude du vieillissement et de la longévité humaine, ainsi que l'étude des microsatellites et de leur instabilité.

Le vieillissement de la population représente un défi majeur sur les plans sociétal, économique et de santé publique en France et en Europe. Face à cette réalité, il est crucial de trouver des solutions pour faire face à l'augmentation des coûts du système de santé liée à cette tendance démographique, en particulier en favorisant un vieillissement en bonne santé. Cela implique une compréhension approfondie des mécanismes du vieillissement et de la longévité, qui varient d'une personne à l'autre. Dans ce contexte, le laboratoire se consacre à l'étude des mécanismes génétiques et épigénétiques sous-jacents à la longévité. Ils analysent les variations du génome mitochondrial, les modifications épigénétiques et les horloges épigénétiques, en particulier chez des centenaires, supercentenaires et leurs descendants.

En parallèle, le laboratoire se consacre à l'étude des microsatellites et de l'instabilité des microsatellites dans le cancer. Les travaux menés portent sur la compréhension fondamentale des mécanismes de formation et de mutation des microsatellites dans des modèles *in vitro* et *in vivo*, procaryotes et eucaryotes. Ils visent également à améliorer les technologies de séquençage des régions répétées, notamment via des approches haut débit (NGS). Depuis 2017, plusieurs méthodes de pointe ont été également développées pour améliorer la détection des instabilités microsatellitaires (MSI) dans le cancer. Ces technologies incluent l'Enhanced-ice-COLD-PCR, qui amplifie préférentiellement les mutants en bloquant l'ADN sauvage grâce à une sonde spécifique (Mauger et al., 2016), et NaMSIE, une méthode enzymatique permettant d'enrichir les fragments mutés (Baudrin et al., 2018). Une autre avancée, la RPA, reposant sur une PCR isotherme qui réduit les artefacts thermiques, offre une meilleure

précision pour le génotypage et la détection des MSI (Jeanjean et al., 2023). Cependant, ces méthodes actuelles rencontrent encore des limitations pour détecter avec une grande sensibilité les mutations microsatellitaires précoces, notamment les délétions et/ou insertions d'un seul nucléotide. C'est pourquoi le laboratoire continue de rechercher des approches encore plus performantes, avec des applications cliniques prometteuses, notamment pour le dépistage des syndromes de Lynch et CMMRD à partir d'échantillons de sang ou de plasma.

8.2. Objectifs

Ma thèse s'est inscrite dans ce contexte de recherche, sur la deuxième thématique de recherche du laboratoire, où des problématiques fondamentales de compréhension des mécanismes de mutations des microsatellites avec une volonté de vouloir caractériser des modèles génétiques sous-jacents, associés à des problématiques technologiques de séquençage des microsatellites. Ma thèse s'est articulée autour de deux grands objectifs, regroupant de la recherche fondamentale, technologique et méthodologique (Figure 24).

D'une part, nous avons eu la volonté de développer un système rapporteur permettant l'étude des mutations microsatellites *in vitro* et *in vivo* compatibles au NGS. En effet comme présenté précédemment, les microsatellites sont compliqués à étudier à cause des mutations induites par la moindre phase d'amplification. Ces mutations se produisent à chaque étape de la manipulation des microsatellites, de la PCR jusque dans les résultats de séquençage. L'objectif était de pouvoir développer un système rapporteur avec différents microsatellites mono- et di-nucléotidiques afin de pouvoir étudier leurs mutations en conditions *in vitro* et *in vivo* avec une compatibilité pour les technologies NGS permettant le séquençage simultané d'une multitude de séquences microsatellites. La réalisation de cet objectif s'est confrontée à la problématique de l'analyse des données, car si la méthode d'amplification génère des mutations, la méthode de séquençage également. C'est pourquoi, nous avons réalisé une comparaison de différentes technologies NGS afin de définir si une méthode était plus optimale pour l'étude des microsatellites et de leurs mutations. Cet objectif nous a alors permis de regrouper des connaissances d'un point de vue technologique et méthodologique pour de futures études sur les microsatellites. En parallèle nous avons également souhaité développer une méthode d'inférence des taux de mutation permettant aussi d'identifier le modèle génétique à l'origine des mutations à partir d'un profil de distribution de taille de microsatellite obtenu après plusieurs cycles de réplication de l'ADN. Nous avons ainsi pu comparer et valider notre méthodologie avec des résultats expérimentaux. Cette méthode d'inférence avait pour but de permettre son application sur l'ensemble des résultats généré dans ma thèse.

D'un autre côté, j'avais également un objectif plus fondamental dans ma thèse. Nous avons voulu étudier l'effet de différents facteurs, notamment exogènes, sur le processus mutationnel des séquences microsatellites. Désormais, il est connu que les microsatellites possèdent plusieurs facteurs qui influent leur mutabilité, notamment des facteurs intrinsèques comme leur taille, leur composition, le nombre de répétitions etc… Mais ici, je me suis posé la question : est-ce que d'autres facteurs, exogènes cette fois, tels que la température, les ADN polymérases, les organismes hôtes, etc. pouvaient impacter le processus mutationnel des séquences microsatellites? C'est pourquoi nous avons étudié l'effet de ces facteurs sur les mutations de différents types de microsatellites, afin d'en déterminer les effets sur les mutations microsatellites.

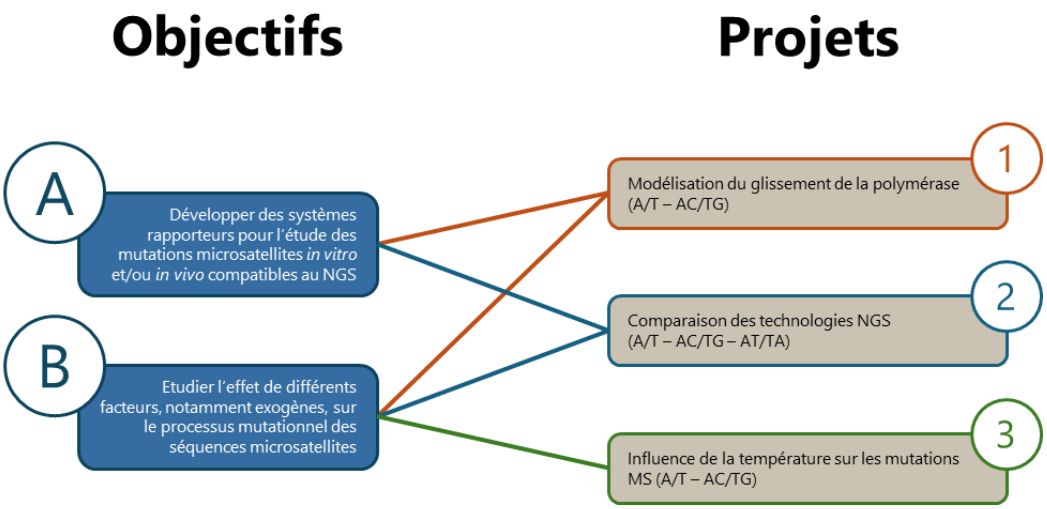

Figure 24 : Organisation et relations entre les différentes objectifs et les projets de la thèse.

Afin de réaliser de manière optimale ces deux objectifs, nous avons reparti la thèse en trois projets (Figure 24).

8.2.1. Projet 1 : Inférence des taux de mutation des microsatellites à partir des fréquences alléliques obtenues à partir d'approches d'amplification d'ADN (non) isothermes en utilisant trente-deux modèles génétiques

Dans ce premier projet, le but était de développer une méthode d'inférence des taux de mutation, qui est l'une des mesures utilisées en génétique pour décrire et caractériser les mécanismes mutationnels. Cette méthode permet également d'identifier le modèle génétique à l'origine du glissement de la polymérase parmi trente-deux modèles développés. La méthode est basée sur le critère de minimisation de l'erreur quadratique moyenne (RMSE) avec un algorithme d'optimisation stochastique pour tester les paramètres — ici, nous avons utilisé l'algorithme de recuit simulé (SA) couplé à une méthode de recherche par grille pour affiner les valeurs à trouver — afin de déterminer les taux de mutation des microsatellites et d'identifier le modèle génétique qui correspond le mieux à la distribution observée des allèles microsatellites.

Cette méthode a d'abord été développée sur des données simulées provenant des différents modèles génétiques utilisés, et les taux de mutation ont été comparés aux taux inférés pour évaluer les performances de notre algorithme en fonction de plusieurs paramètres donnés et caractéristiques génétiques afin d'identifier ses limites. Notre algorithme développé, que nous avons appelé MSmrc32 (*MicroSatellite mutation rate calculator 32*), comprend trois modes d'inférence du taux de mutation à partir des données de PCR basés soit sur (i) le nombre de cycles et les efficacités cycle par cycle, (ii) le nombre de cycles et une efficacité unique, ou (iii) le nombre de doublements de molécules. Ce dernier étant applicable aux méthodes d'amplification isotherme. Ainsi, les taux de délétion et d'insertion allant d'un jusqu'à quatre sauts d'unité de répétitions ont été considérés, avec une relation croissante linéaire ou exponentielle entre le taux de mutation et le nombre d'unité de répétition.

Enfin, l'algorithme a été appliqué à des données expérimentales de PCR et RPA générées à partir de microsatellites à répétitions mono- (A/T) et di-nucléotidiques (AC/TG) et a permis de calculer les taux

de mutation des microsatellites et d'identifier les modèles de glissement de la polymérase les plus susceptibles de se produire dans ces expériences *in vitro*.

8.2.2. Projet 2 : Analyse détaillée des approches de séquençage de deuxième et troisième génération pour la détermination précise de la longueur des répétitions courtes en tandem et des homopolymères.

Dans ce second projet, nous avons évalué et comparé différentes méthodes et instruments de séquençage de nouvelle génération (NGS) pour estimer avec précision la longueur de différents microsatellites, afin d'identifier l'approche la plus efficace pour minimiser les erreurs de longueur de répétition et refléter fidèlement la distribution réelle des allèles présents dans un échantillon. En utilisant plusieurs plasmides contenant des microsatellites mono-(A/T) et di-(AC/TG et AT/TA) nucléotidiques de 15 à 25 répétitions, nous avons généré plusieurs bibliothèques NGS basées sur les technologies de séquençage de 2ème génération d'Illumina et de 3ème génération de PacBio, y compris certaines spécifiquement dédiées à la réduction des erreurs de séquençage et à la détection des mutations rares. Ainsi, avec les instruments Illumina (iSeq100 et/ou NextSeq 500), nous avons évalué un protocole sans PCR, des protocoles contenant des cycles PCR (1 à 20 cycles), et des protocoles de PCR avec UMI (*Unique Molecular Identifier*) unique ou double (séquençage duplex) (12 à 20 cycles) qui permettent la correction des erreurs de PCR et de séquençage, tandis que nous avons testé une approche sans PCR avec le séquenceur PacBio Sequel II.

En raison des spécificités des microsatellites et des homopolymères, plusieurs approches bioinformatiques ont été développées pour l'analyse des données de séquençage des microsatellites. Ainsi, nous avons une comparaison exhaustive des différentes plateformes et protocoles NGS, ainsi que des stratégies bioinformatiques pour la détermination correcte de la distribution de la longueur des allèles de différents types de microsatellites.

8.2.3. Projet 3 : Influence de la température sur les mutations des microsatellites

Enfin, dans ce dernier projet, notre intérêt s'est porté sur la caractérisation des mutations microsatellites en fonction de différents facteurs endogènes (unité répétée et nombre de répétitions) et exogènes (polymérase et température) pouvant influencer le mécanisme mutationnel. Nous avons développé un système rapporteur en se basant sur l'utilisation de plasmides contenant des séquences microsatellites combinées à l'électrophorèse capillaire et/ou NGS ce qui a permis l'étude des mutations microsatellites *in vitro* et *in vivo*.

Les études *in vitro* ont été réalisées sur des plasmides de synthèse ainsi que sur des marqueurs biologiques connus, à la fois mono- (A/T) et di-nucléotidiques (AC/TG). Trois méthodes d'amplification de l'ADN ont permis des amplitudes de températures distinctes allant de 15 à 65°C ; le WGA, la RPA et la PCR. La double analyse par électrophorèse capillaire et NGS nous a permis d'évaluer et de confirmer l'importance de l'outil d'analyse lors de l'étude des fréquences de mutations chez les microsatellites.

Les études *in vivo* ont quant à elles été réalisées au sein de deux organismes différents. Dans un premier temps, chez un organisme procaryote : la bactérie *Escherichia coli* et dans un deuxième temps chez un organisme eucaryote : la levure *Saccharomyces cerevisiae*. L'impact de la présence d'un système de réparation des mésappariements fonctionnels chez ces deux organismes a pu être quantifié à l'aide de souches sauvages (WT) et déficientes pour différentes sous-unités du système MMR. Afin de pouvoir également étudier l'effet de la température sur les taux de mutations des microsatellites, les cultures des

différentes souches transformées en parallèle par un plasmide mono- (A_{24}) et di-nucléotidiques ($(AC)_{24}$) ont été réalisées à des températures allant de 20 à 45°C. *Saccharomyces cerevisiae* possédant suffisamment de microsatellites dans son génome, ils ont également été étudiés.

1. Projet 1 : Inférence des taux de mutation des microsatellites à partir des fréquences alléliques obtenues à partir d'approches d'amplification d'ADN (non) isothermes en utilisant trente-deux modèles génétiques

1.1. Résultats

Dans cette partie sera présenté l'intégralité des résultats obtenus pour ce projet dont l'objectif était de pouvoir inférer les taux de mutations à partir de fréquences alléliques de microsatellites, obtenus après amplification *in vitro* par PCR et RPA. La réalisation, le développement et l'optimisation des algorithmes mathématiques, informatiques et statistiques ont été réalisés par Schayma Ben Marzougui (ingénieure de recherche) pour les prémices et Rojo Nofidiantsoa (développeur) pour la finalisation. Cette partie ne sera pas abordée en détail ici, et sera décrite dans la partie méthode.

1.1.1. Modélisation mathématique de l'amplification de l'ADN et du processus de mutation des microsatellites au cours de la PCR

Dans un premier temps, afin de permettre le développement de notre approche d'inférence des taux de mutations des microsatellites à partir des fréquences de microsatellites obtenues après amplification PCR, les mécanismes d'amplification et de mutations des microsatellites ayant lieu lors d'une PCR ont été modélisés. Les formules utilisées sont basées principalement sur certaines formules décrites dans une précédente étude (Shinde et al., 2003), auxquelles nous avons apporté des modifications. Au cours de la PCR, le nombre total de molécules d'ADN (n_i) pendant chaque cycle (i) de PCR augmente selon la formule :

$$n_{i+1} = (1 + \lambda_i) \times n_i$$

où λ_i correspond à l'efficacité de la réplication de l'ADN au cycle i ($0 \leq \lambda \leq 1$).

Toutefois, après chaque cycle de réplication, un microsatellite possédant x unités de répétition a pu avoir été copié à l'identique ou a pu subir une mutation (une insertion ou une délétion) en suivant le mécanisme de glissement de polymérase (Synthèse Bibliographique 2.2.3.3.). De ce fait, le taux de mutation du microsatellite (μ_x) peut être définit par un taux de délétion (μ_{x_d}) et un taux d'insertion (μ_{x_u}), décrit selon la formule :

$$\mu_x = \boxed{\mu_{x_d}} + \boxed{\mu_{x_u}}$$

Ainsi, si l'on considère uniquement la possibilité pour un microsatellite de muter d'un motif (un pas), le dénombrement du microsatellite avec x répétitions accumulées au fil des cycles PCR se fait suivant la formule :

$$n(x)_{i+1} = n(x)_i + \boxed{n(x+1)_i \times \lambda_i \times \mu_{x+1_d}} + \boxed{n(x-1)_i . \lambda_i . \mu_{x-1_u}} + n(x)_i \times \lambda_i \times \left(1 - \mu_{x_d} - \mu_{x_u}\right)$$

où $n(x)_i$ correspond au nombre de molécules précédentes, $\boxed{n(x+1)_i \times \lambda_i \times \mu_{x+1_d}}$ correspond au nombre de molécules $n(x+1)_i$ ayant subi une délétion, $\boxed{n(x-1)_i . \lambda_i . \mu_{x-1_u}}$ correspond au nombre de molécules $n(x-1)_i$ ayant subi une insertion et $n(x)_i \times \lambda_i \times \left(1 - \mu_{x_d} - \mu_{x_u}\right)$ correspond au nombre de molécules microsatellites ayant été copié à l'identique à partir de $n(x)_i$.

La modélisation du mécanisme de mutation des microsatellites survenant durant la PCR à cause du glissement de la polymérase a été par la suite développée à l'aide de 32 modèles génétiques en se basant sur des modèles suggérés ou utilisés dans de précédentes études (Baptiste et al., 2015; Eckert et al., 2002; Huang et al., 2002; Kelkar et al., 2010, 2008; Raz et al., 2019; Rose and Falush, 1998; Sainudiin et al., 2004; Shinde et al., 2003; Sia et al., 1997b; Whittaker et al., 2003; Wierdl et al., 1997; Xu et al., 2000). Ces modèles mutationnels varient en fonction de la relation entre taux de mutation et nombre d'unités de répétition et de la possibilité de réaliser des mutations de 1 (modèle à un pas, Figure 25.A) à 4 motifs (modèle à plusieurs pas, Figure 25.B).

Ainsi, une dépendance entre les taux de mutations et le nombre de répétition du motif microsatellite a été introduite, qui était soit linéaire ou exponentielle, et donnée par les équations récursives suivantes (exemple donné pour le modèle à 1 pas, Figure 25.A) :

- Relation linéaire croissante : $\mu_{x+1_d} = \mu_{x_d} + \alpha_d$ pour le taux de délétion et $\mu_{x+1_u} = \mu_{x_u} + \alpha_u$ pour le taux d'insertion, avec $0 < \alpha_{d/u} < 1$, Figure 25.A)
- Relation exponentielle croissante : $\mu_{x+1_d} = \beta_d \times \mu_{x_d}$ pour le taux de délétion et $\mu_{x+1_u} = \beta_u \times \mu_{x_u}$ pour le taux d'insertion, avec $\beta_{d/u} > 1$, Figure 25.A).

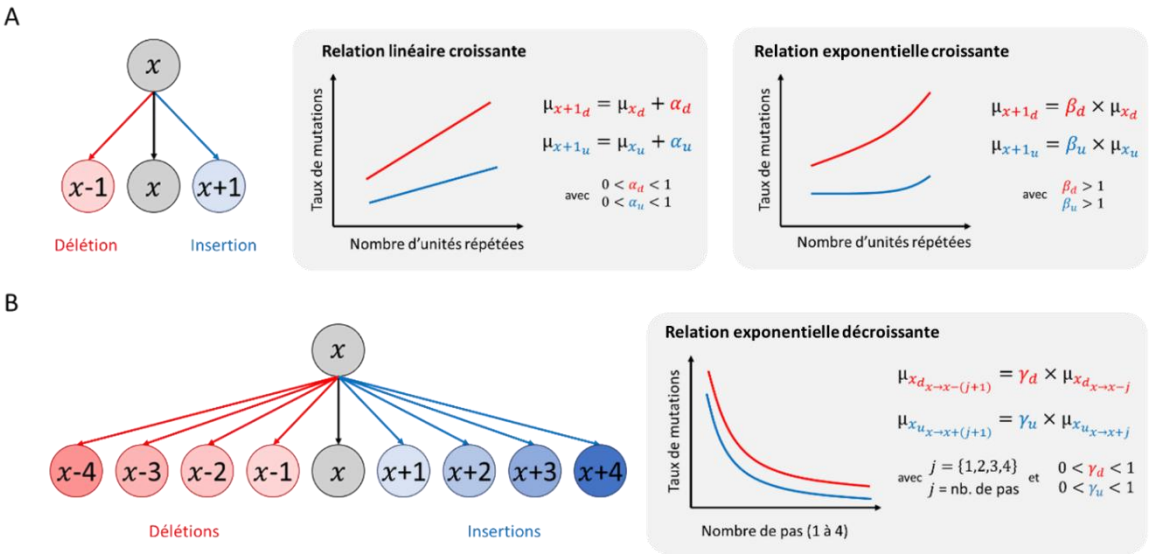

Figure 25 : Représentation schématique des modèles génétiques sous-jacents au glissement de la polymérase. A. Modèle à un seul pas. Dans ce modèle, le microsatellite de x unités de répétition ne peut glisser que d'une unité de répétition. La relation entre les taux de mutation (μ) et le nombre d'unités de répétition (x) augmente de manière linéaire ou exponentielle selon les coefficients α et β. d = délétion, u = insertion. B. Modèle à plusieurs pas avec jusqu'à quatre étapes pour les insertions et les délétions. La relation entre les taux de mutation (μ) et le nombre d'unités de répétition (x) est similaire au modèle de glissement à un seul pas. Dans ce modèle, les taux de mutation (μ) diminuent de manière géométrique à mesure que le nombre de pas (j) augmente selon les coefficients γ.

Pour les modèles à plusieurs pas, une relation exponentielle décroissante a été introduite, mais cette fois entre les taux de mutation des microsatellites de pas ou saut croissant, donnée par les équations : $\mu_{x_{d_{x\to x-(j+1)}}} = \gamma_d \times \mu_{x_{d_{x\to x-j}}}$ pour les taux de délétions et $\mu_{x_{u_{x\to x+(j+1)}}} = \gamma_u \times \mu_{x_{u_{x\to x+j}}}$ pour les taux d'insertions, avec $0 < \gamma_{d/u} < 1$ et j = nombre de pas/saut autorisé, Figure 25.B).

En combinant ces deux relations, un total de 32 modèles génétiques a pu être considéré pour le mécanisme mutationnel du glissement de la polymérase au cours de la réplication des microsatellites par PCR (Figure 26).

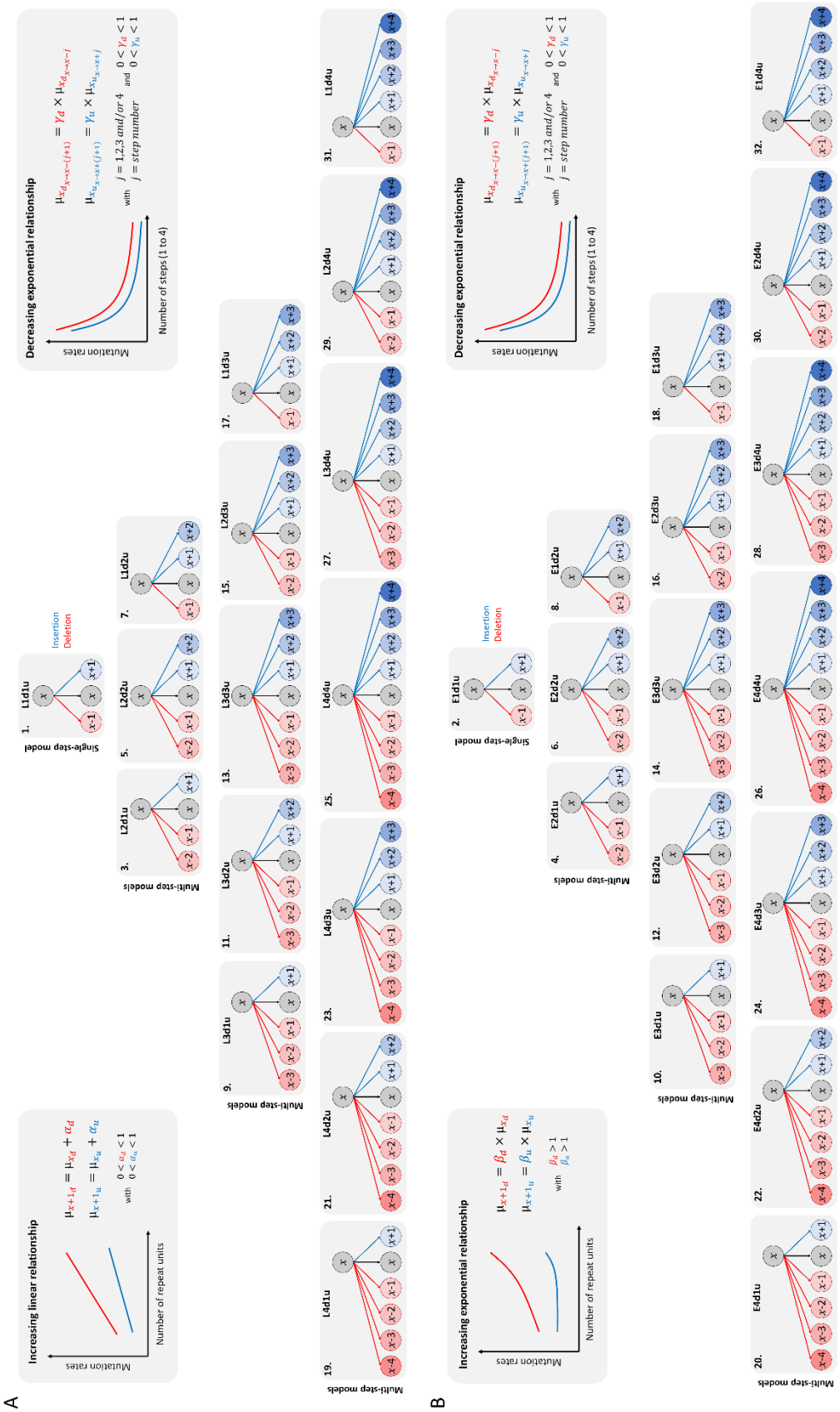

Figure 26 : Représentation graphique des 32 types différents de modèles génétiques utilisés pour les mécanismes de mutation des microsatellites basés sur le glissement de la polymérase. Chaque modèle permet une combinaison spécifique d'un nombre de délétions et d'insertions de l'unité répétée (pas) variant de 1 à 4. Les modèles sont liés par une relation (A) linéaire ou (B) exponentielle croissante, ce qui donne un total de 32 modèles différents.

1.1.2. Estimation des taux de mutation, des paramètres α/β et γ par le critère de minimisation de l'erreur quadratique moyenne à l'aide de l'algorithme de recuit simulé couplé à une recherche par grille

Nous avons utilisé un générateur infinitésimal de la chaîne de Markov X pour prédire les fréquences des allèles des microsatellites obtenus après les expériences de PCR à chaque cycle, en utilisant un nombre défini de copies de matrices d'ADN avec les trente-deux modèles génétiques. Les trente-deux modèles comprenaient quatre à six paramètres à estimer en fonction du nombre de pas/sauts considéré(s) dans chaque modèle (Tableau 7).

Tableau 7 : Paramètres du système de modélisation des 32 modèles.

Modèle	**Paramètres des modèles**
1. L1d1u	$\mu_u, \mu_d, \alpha_u, \alpha_d$
2. E1d1u	$\mu_u, \mu_d, \beta_u, \beta_d$
7. L1d2u 17. L1d3u 31. L1d4u	$\mu_u, \mu_d, \alpha_u, \alpha_d, \gamma_u$
3. L2d1u 9. L3d1u 19. L4d1u	$\mu_u, \mu_d, \alpha_u, \alpha_d, \gamma_d$
8. E1d2u 18. E1d3u 32. E1d4u	$\mu_u, \mu_d, \beta_u, \beta_d, \gamma_u$
4. E2d1u 10. E3d1u 20. E4d1u	$\mu_u, \mu_d, \beta_u, \beta_d, \gamma_d$
5. L2d2u 23. L4d3u 11. L3d2u 25. L4d4u 13. L3d3u 27. L3d4u 15. L2d3u 29. L2d4u 21. L4d2u	$\mu_u, \mu_d, \alpha_u, \alpha_d, \gamma_u, \gamma_d$
6. E2d2u 24. E4d3u 12. E3d2u 26. E4d4u 14. E3d3u 28. E3d4u 16. E2d3u 30. E2d4u 22. E4d2u	$\mu_u, \mu_d, \beta_u, \beta_d, \gamma_u, \gamma_d$

Les paramètres de chaque modèle ont été inférés en utilisant l'algorithme d'optimisation stochastique du recuit simulé (*Simulated Annealing*, SA), combiné à une recherche par grille (*Grid Search*, GS), en utilisant le critère de minimisation du RMSE (*Root Mean Square Error*) obtenu à partir de la distribution des allèles des microsatellites ayant une fréquence supérieure à 0,01%, selon la formule suivante :

$$RMSE = \sqrt{\sum_{allèle=1}^{K} \frac{(fréquence\ observée_{allèle} - fréquence\ simulée_{allèle})^2}{K}}$$

Ainsi, pour chacun des 32 modèles, les paramètres optimaux ont été déduits à partir du RMSE le plus bas obtenu permettant de reproduire une distribution allélique au plus proche de celle obtenue dans l'expérience., De plus, le modèle présentant le RMSE le plus bas parmi les 32, qui permet le mieux d'expliquer la distribution de taille des allèles microsatellites observée, pouvait ainsi correspondre au meilleur modèle génétique expliquant le mécanisme mutationnel sous-jacent au glissement de la polymérase de l'expérience observée.

1.1.3. Evaluation des performances d'inférence des taux de mutation des microsatellites

Pour évaluer la performance et la précision de notre algorithme pour déterminer les taux de mutation des microsatellites, des données artificielles de fréquence d'allèles de microsatellites ont été générées avec les modèles linéaires et exponentiels à un seul pas/sauts exponentiel (1. L1d1u et 2. E1d1u) et les modèles à quatre pas/sauts (25. L4d4u et 26. E4d4u), en modifiant plusieurs paramètres, y compris différentes conditions de PCR. La performance et la précision de l'algorithme ont été évaluées à l'aide du RMSE obtenu avec les données artificielles et de l'erreur relative ($RR = \frac{|\mu_{simulated} - \mu_{real}|}{\mu_{real}}$) des taux de mutation des microsatellites inférés.

Sans détailler les manipulations informatiques qui ont été effectuées par Schayma Ben Marzougui et Rojo Nofidiantsoa, les résultats ont montré que l'algorithme était capable d'inférer correctement les taux de mutation de 10 à 90 unités de répétition, indiqués par un RR inférieur à 0,1 dans la plupart des cas, sauf pour les taux de mutation avec un nombre de sauts/pas supérieur à 1 et de plus en plus grand (Figure 27).

De plus, dans les modèles à quatre pas, la performance d'inférence des taux de mutation a diminué à mesure que le nombre de pas des taux de mutation augmentait. Cependant, ces performances décroissantes devraient avoir peu d'impact sur la détermination des taux globaux d'insertion et de délétion, en raison de la contribution respective des taux de mutation au nombre de sauts de plus en plus grand.

Pour les modèles linéaires, l'augmentation des coefficients α_u et α_d a diminué les performances d'inférence dans le modèle à quatre pas, alors que pour les modèles exponentiels, l'augmentation des coefficients β_u et β_d a légèrement diminué la performance de notre algorithme dans les modèles à un et quatre pas. Ainsi, l'algorithme d'inférence a montré des performances comparables pour les plages de valeurs utilisées pour les coefficients α et β (Figure 28).

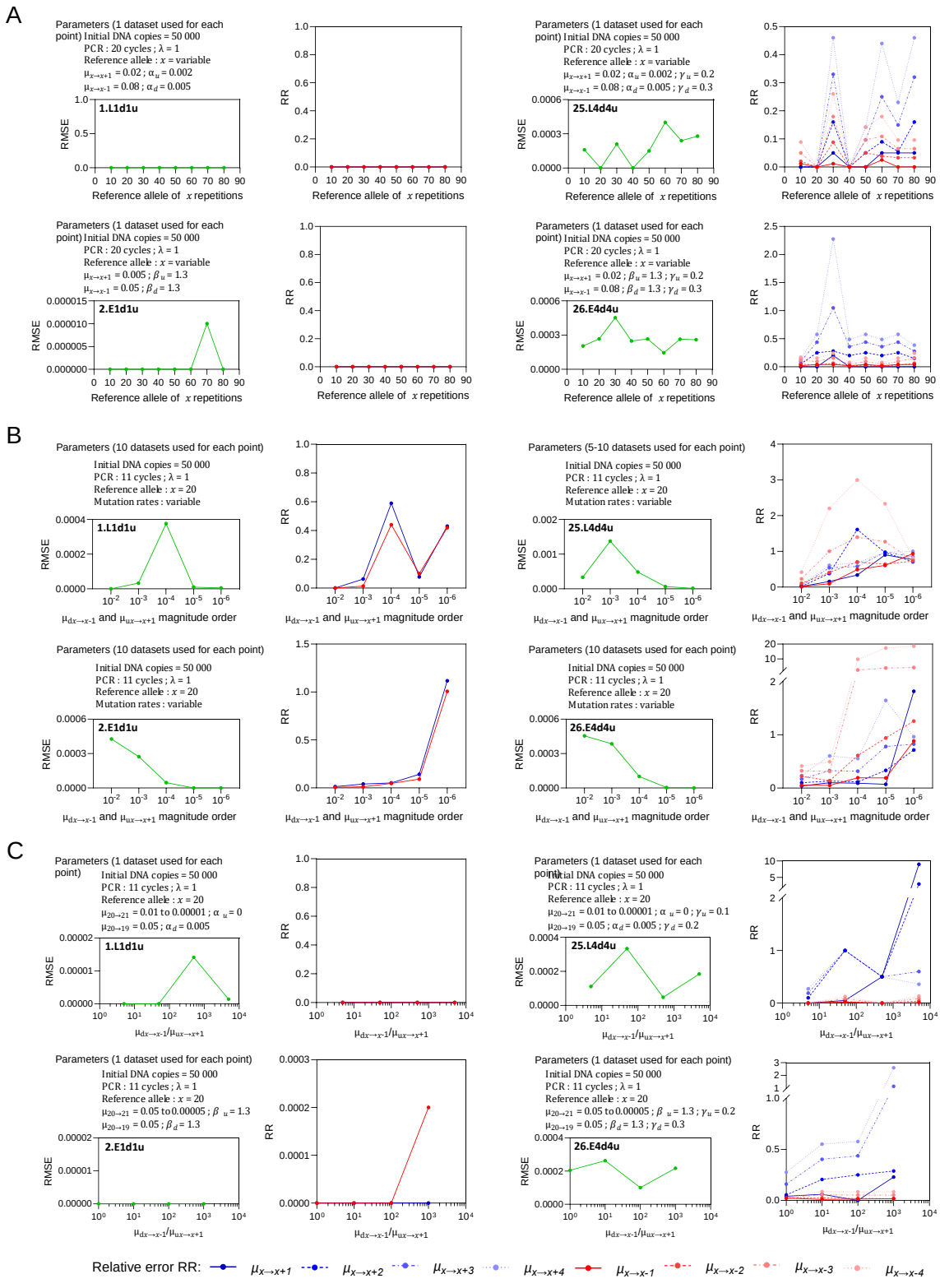

Figure 27 : Effet de l'allèle de référence et des variations du taux de mutation des microsatellites sur la performance d'inférence de notre algorithme avec quatre modèles génétiques (L1d1u, L4d4u, E1d1u et E4d4u). Le plus petit RMSE obtenu avec notre algorithme et le RR correspondant des taux de mutation inférés en utilisant (A) une longueur croissante d'allèles de référence, (B) un ordre décroissant de grandeur des deux taux de mutation ou (C) une différence log-augmentant entre les taux d'insertion et de délétion. Les paramètres utilisés pour générer les fréquences d'allèles artificielles (données d'entrée) ont été donnés pour chaque condition.

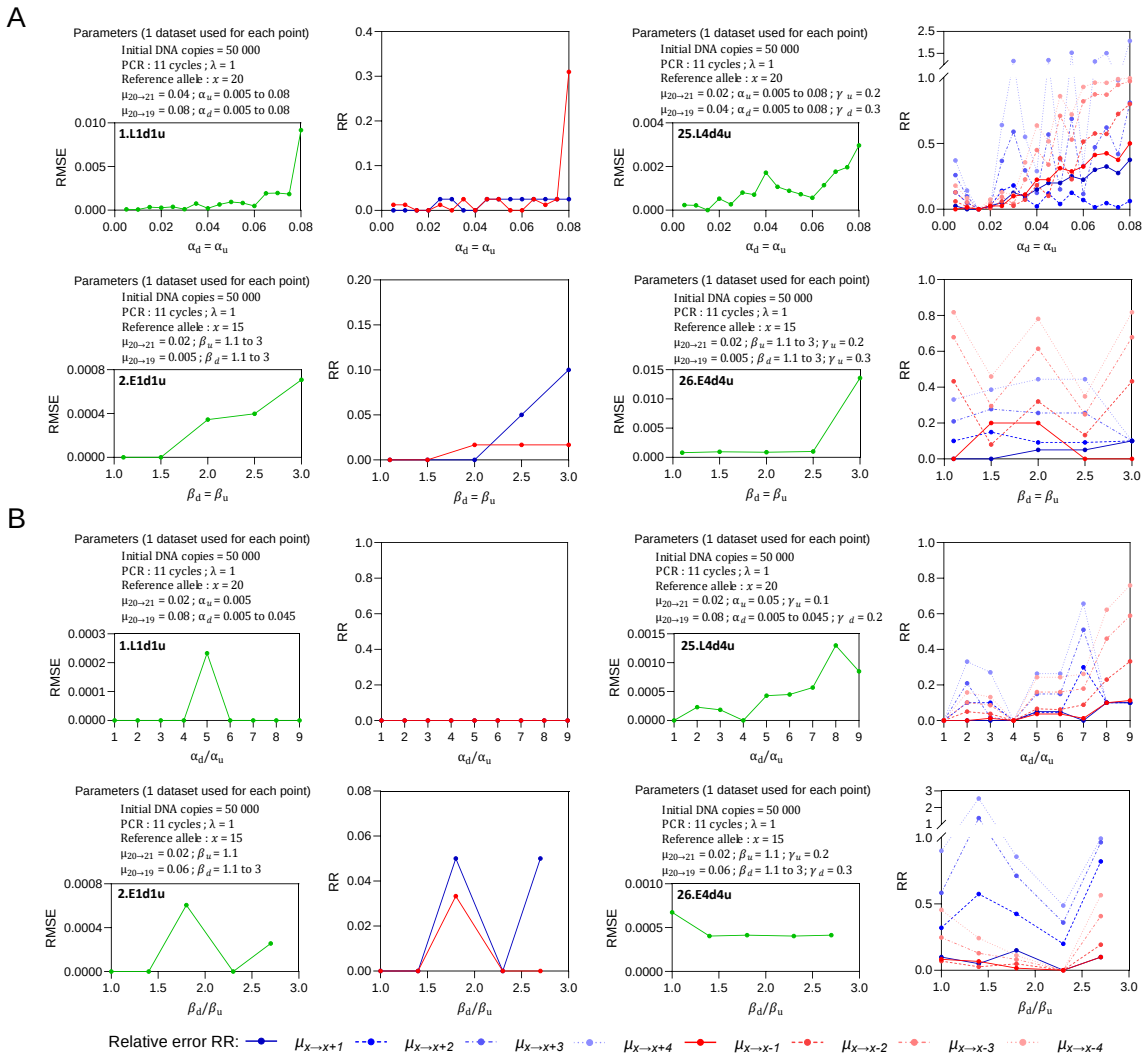

Figure 28 : Effet des coefficients linéaires (α) et exponentiels (β) sur la performance d'inférence de notre algorithme avec quatre modèles génétiques (L1d1u, L4d4u, E1d1u et E4d4u). Le plus petit RMSE obtenu avec notre algorithme et le RR correspondant des taux de mutation inférés en utilisant (A) un coéfficient égal ou croissant (α_d=α_u et β_d=β_u) pour les insertions et délétions (B) ou des coefficients de délétions croissants (α_d or β_d) et d'insertions fixes (α_u or β_u). Les paramètres utilisés pour générer les fréquences d'allèles artificielles (données d'entrée) ont été donnés pour chaque condition.

Enfin, nos résultats ont montré que le nombre de cycles de PCR n'affectait pas les performances d'inférence dans les modèles à un seul pas, tandis qu'elles diminuaient progressivement dans les modèles à quatre pas à mesure que le nombre de cycles de PCR augmentait (Figure 29.A). La performance de l'algorithme était similaire pour les différents nombres de jeux de données utilisés, indiquant qu'un seul jeu de données pouvait donner de bons résultats (Figure 29.B).

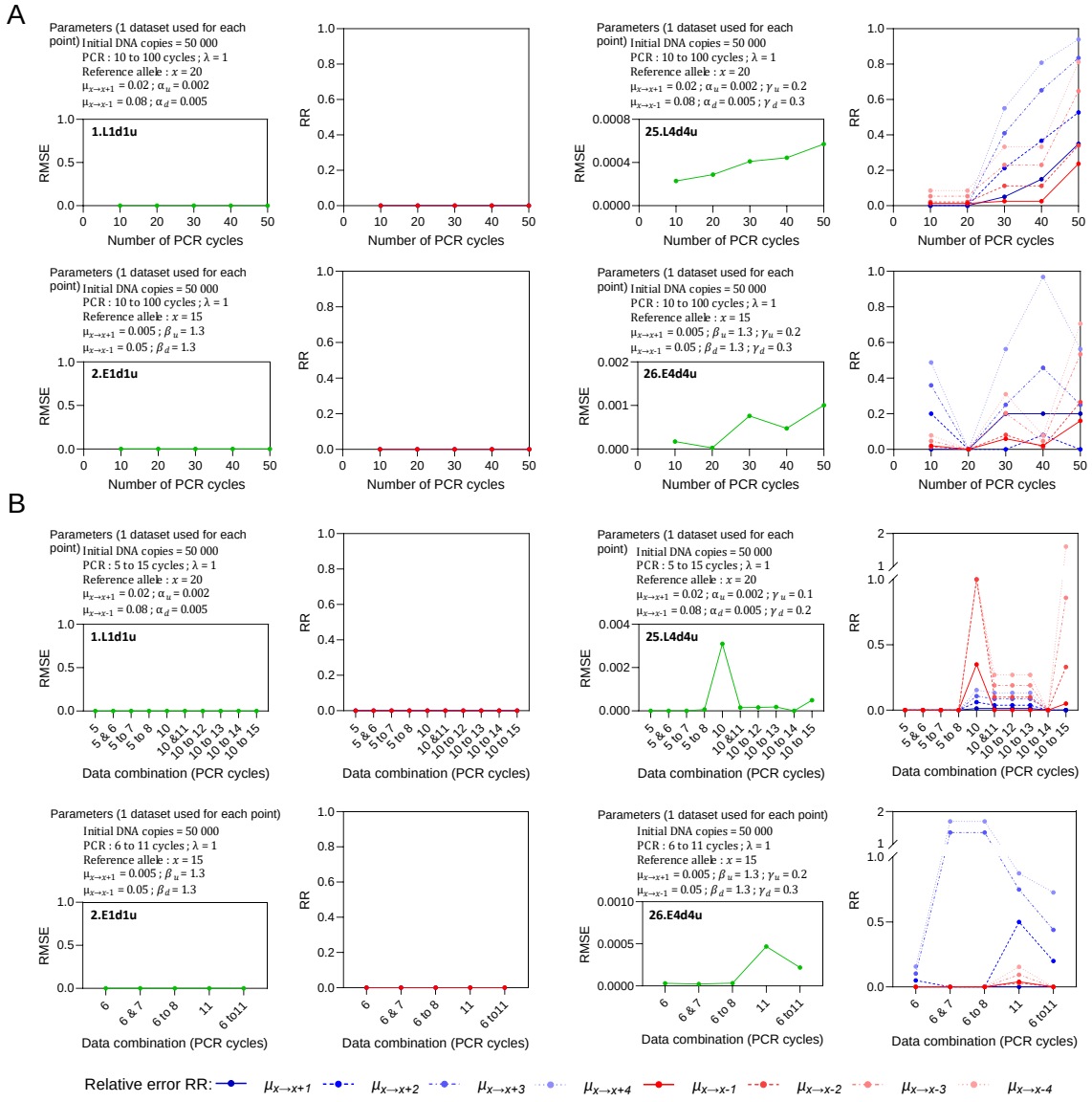

Figure 29 : Effet du nombre de cycles PCR et des ensembles de données sur les performances d'inférence de notre algorithme avec quatre modèles génétiques (L1d1u, L4d4u, E1d1u et E4d4u). Le plus petit RMSE obtenu avec notre algorithme et le RR correspondant des taux de mutation inférés en utilisant (A) un nombre croissant de cycles de PCR ou (B) un nombre croissant d'ensembles de données disponibles. Les paramètres utilisés pour générer les fréquences d'allèles artificielles (données d'entrée) ont été donnés pour chaque condition.

1.1.4. Évaluation de MSmrc32 pour l'identification du modèle génétique sous-jacent au mécanisme de mutation des microsatellites

Afin de pouvoir évaluer la capacité de *MSmrc32* (notre algorithme) à identifier le modèle génétique optimal, nous avons généré des fréquences alléliques de microsatellites basées sur les 32 modèles génétiques en utilisant les mêmes valeurs pour chaque paramètre lorsque cela était possible. En comparant les 32 jeux de données artificielles entre eux après passage dans l'algorithme, la matrice RMSE a montré des différences entre toutes les fréquences alléliques, avec une valeur nulle obtenue uniquement lorsqu'un jeu de données est comparé à lui-même (Figure 30.A). De même, le plus faible RMSE obtenu a également été comparé aux données et a montré un schéma similaire, bien que légèrement plus flou, à la matrice précédente (Figure 30.B). La diagonale, présentant certaines des

valeurs de RMSE les plus basses, suggère que le plus faible RMSE pourrait être utilisé pour déterminer le modèle génétique sous-jacent au mécanisme de glissement de la polymérase lors d'expériences *in vitro*. Enfin, quel que soit le modèle utilisé pour l'inférence, les taux de mutation inférés étaient proches des valeurs réelles, ce qui pourrait indiquer que le modèle sélectionné pour l'inférence a peu d'influence sur l'estimation des taux de mutation (Figure 30.C).

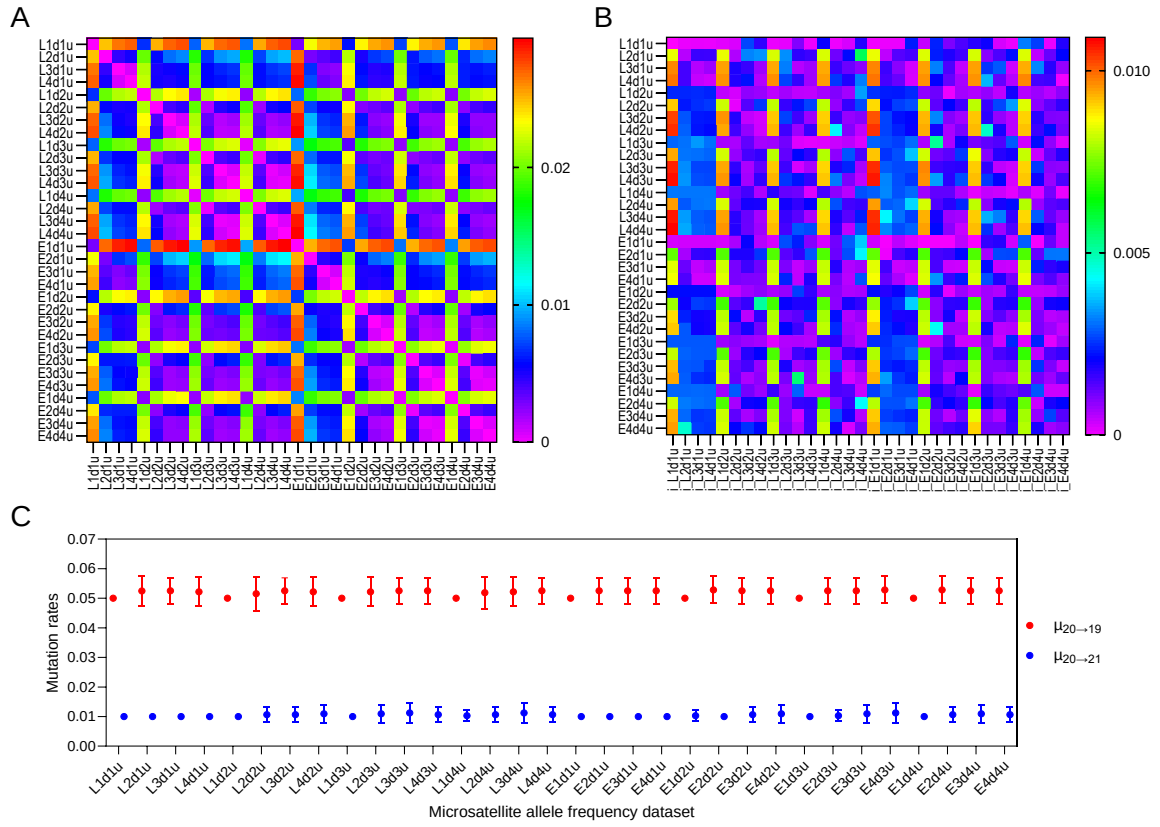

Figure 30 : Évaluation de la capacité de notre algorithme à identifier le modèle génétique sous-jacent parmi les 32 disponibles. Cartes thermiques des valeurs de RMSE obtenues avec les 32 jeux de données artificielles générés avec chacun des 32 modèles, comparés soit (A) à eux-mêmes, soit (B) aux données "i_" inférées à partir des 32 modèles. C. Taux de mutation moyens inférés avec les 32 modèles en utilisant les 32 jeux de données de fréquences alléliques de microsatellites. Paramètres utilisés pour la génération des données artificielles : copies d'ADN initiales = 5000 ; allèle de référence = 20 ; cycles de PCR = 15 ; Modèles linéaires ; Modèles exponentiels.

1.1.5. Mise en œuvre de trois modes pour l'inférence des taux de mutation des microsatellites compatibles avec les méthodes d'amplification d'ADN basées sur la PCR et les méthodes isothermes

L'algorithme développé permet d'inférer les taux de mutation des microsatellites à partir des fréquences d'allèles de microsatellites obtenues après la réaction de PCR, avec la connaissance de l'allèle microsatellite original (considéré comme pur), du nombre de molécules d'ADN de départ utilisées comme matrice, du nombre total de cycles de PCR et de l'efficacité cycle par cycle de la PCR (mode I d'efficacité variable, e.v.). Alors qu'il est facile d'obtenir les efficacités cycle par cycle de la PCR avec la PCR en temps réel, cela n'est pas possible avec d'autres méthodes utilisées pour l'amplification de l'ADN. Ainsi, les méthodes standard de PCR et d'amplification d'ADN isotherme, y compris la RPA, ne peuvent pas fournir ces données, en raison de l'absence de données en temps réel (pour la PCR standard) et/ou

de cycles de PCR (pour les méthodes isothermes dont la RPA). Pour adapter l'algorithme à ces méthodes, deux autres modes d'inférence ont été mis en place pour l'inférence des taux de mutation soit à partir d'une efficacité fixe de PCR fixée à 1 (efficacité de réplication de l'ADN de 100%, mode II d'efficacité fixe, e.f.) ou d'un nombre d'événements dits de doublement des molécules (mode III de doublement des molécules, d.m.) qui peut être calculé simplement comme suit : $log_2(\frac{DNA\ molecules\ après\ amplification}{DNA\ molecules\ avant\ amplification})$.

Des données PCR artificielles correspondant à une PCR expérimentale de 45 cycles ont été générées, où l'amplification de l'ADN suivait une courbe sigmoïde, incluant des phases initiales exponentielles (cycles 1 à 21), log-linéaire (cycles 22 à 35) et plateau (cycles 36 à 45), causées par la diminution des efficacités de PCR de 1 à 0 du cycle 13 au cycle 35 (Figure 31.A). Pour divers cycles de PCR, nous avons utilisé les fréquences alléliques des microsatellites pour inférer les taux de mutation avec le modèle linéaire à un seul pas, basés sur le RMSE le plus bas obtenu avec les trois modes (Figure 31.B).

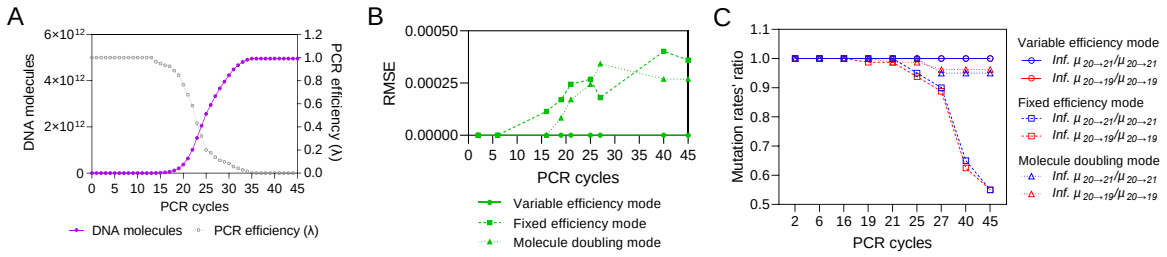

Figure 31 : Evaluation des différents modes d'inférence des taux de mutations des microsatellites, incluant un mode d'efficacité variable, un mode d'efficacité fixe et un mode de doublement de molécule. A. Graphiques d'amplification par PCR et d'efficacité des données artificielles générées en utilisant le modèle 1. L1d1u (copies initiales d'ADN = 5.10^5 ; allèle de référence x = 20 ; $\mu_{(20\rightarrow21)}$ = 0,02 α_u = 0,002 ; $\mu_{(20\rightarrow19)}$ = 0,08 et α_d = 0,01). B. Plus petite RMSE obtenue avec les trois modes différents. C. Ratio des taux d'insertion ($\mu_{(20\rightarrow21)}$) / délétion ($\mu_{(20\rightarrow19)}$) des microsatellites inférés sur les taux d'insertion ($\mu_{(20\rightarrow21)}$) / délétion ($\mu_{(20\rightarrow19)}$) des microsatellites réels.

Les résultats ont montré que le mode I permettait une détermination parfaite des taux de mutation des microsatellites durant chaque cycle de PCR, ce qui était également le cas avec les deux autres modes (II et III) pendant les cycles où l'efficacité était égale ou très proche de 1 (Figure 31.C). Avec la diminution de l'efficacité de la PCR, les modes II et III sous-estimaient initialement légèrement les taux de mutation durant la phase log-linéaire. Ensuite, pendant la phase plateau, le mode III atteignait une limite de sous-estimation de 5 %, tandis que le mode II continuait à sous-estimer de plus en plus les taux de mutation (Figure 31.C). Bien que le mode I soit le meilleur pour inférer les taux de mutation quel que soit le cycle de PCR, le mode III, qui peut exploiter les données des PCR standard et de l'amplification d'ADN isotherme, a également donné une bonne approximation des taux de mutation malgré une légère sous-estimation dans les phases/cycles tardifs de la PCR. D'un autre côté, l'utilisation du mode II devrait être limitée aux cycles avec des efficacités de PCR proches de 1, c'est-à-dire la phase exponentielle et les premiers cycles de la phase linéaire (Figure 31).

1.1.6. Evaluation de l'algorithme sur des données expérimentales de PCR et RPA

En utilisant comme matrices des microsatellites mononucléotidiques A/T et dinucléotidiques AC/TG avec des longueurs de 15 à 24 répétitions, contenus dans des plasmides pUC57-Mini (Figure 32 et Tableau 16), nous avons généré plusieurs données réelles de fréquences d'allèles de microsatellites à partir d'expériences de PCR en temps réel et RPA.

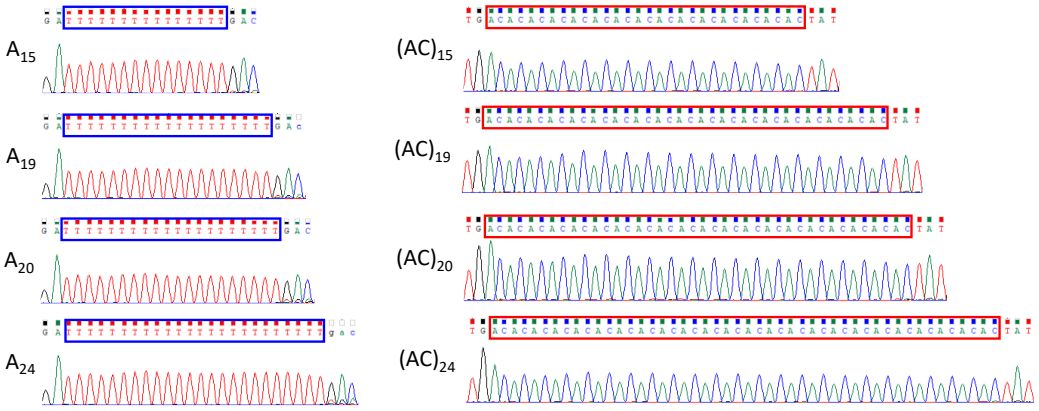

Figure 32 : Séquences de microsatellites utilisées dans notre étude. Séquençage Sanger des huit séquences de microsatellites insérées dans le vecteur pUC57-Mini.

Les amorces *forward* étaient marquées avec un flurochrome FAM, qui permettait la détection des tailles d'amplicons par analyse de fragment au séquenceur capillaire SeqStudio. Les fréquences alléliques ont été calculées à partir des pics obtenus sur *GeneMapper*.

Les expériences de PCR étaient de deux types : (i) des réactions de 50 cycles, utilisant de 10^3 à 10^8 copies de plasmides comme matrice d'ADN, réalisées jusqu'à la phase de « plateau » (Figure 33 & Figure 34.A), et (ii) des réactions de PCR cinétiques, utilisant 10^8 et 10^5 copies de plasmides comme matrices, arrêtées à 10, 12, 14 et 16 cycles et 21, 23, 25 et 27 cycles, respectivement (Figure 33 & Figure 34.B). Chaque expérience a été effectuée en quadruplicat.

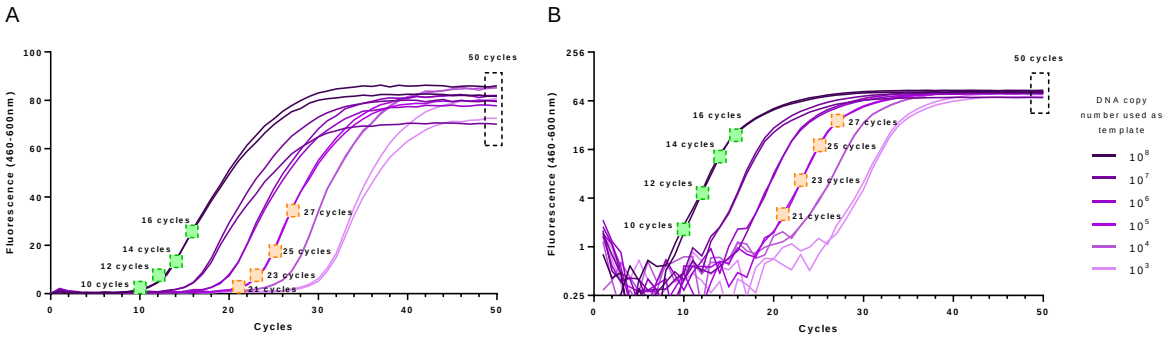

Figure 33 : Conditions d'amplification de PCR en temps réel réalisées dans notre étude utilisant différents types de microsatellites (mononucléotides : A_{15}, A_{19}, A_{20} et A_{24} et dinucléotides : $(AC)_{15}$, $(AC)_{19}$, $(AC)_{20}$ et $(AC)_{24}$) à des quantités variables d'ADN matrice (10^3, 10^4, 10^5, 10^6, 10^7 et 10^8 copies), également avec différents nombres de cycles PCR (10, 12, 14, 16, 21, 23, 25, 27 et 50) A. Fluorescence exprimée en échelle linéaire. B. Fluorescence exprimée en échelle logarithmique binaire (log2). Les courbes d'amplification présentées ont été obtenues avec le plasmide A20 sur 50 cycles et incluaient deux répétitions par condition. La fluorescence de base a été soustraite des données brutes de fluorescence avant le traitement.

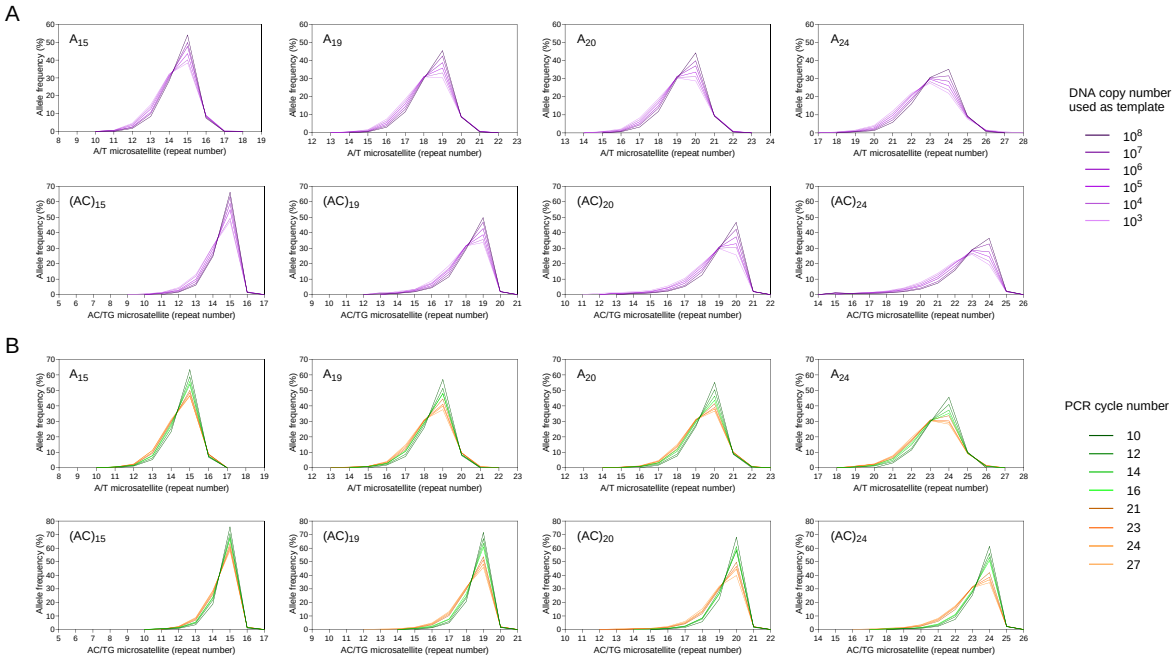

Figure 34 : Fréquences alléliques de microsatellites obtenues par PCR en utilisant des microsatellites avec des répétitions de mono- et di-nucléotides comme matrice. A. PCR de 50 cycles avec différentes quantités de matrice d'ADN, allant de 10^3 à 10^8 copies. B. PCR cinétique avec 10^8 copies d'ADN en entrée (10, 12, 14 et 16 cycles de PCR) et 10^5 copies d'ADN en entrée (21, 23, 25 et 27 cycles de PCR). Chaque point représente des PCR effectuées en quadruplicat. Les moyennes des fréquences alléliques ont été utilisées pour la représentation graphique.

Les valeurs de Ct ont été calculées à partir des expériences de PCR de 50 cycles et ont indiqué le moment approximatif où l'efficacité a commencé à diminuer, correspondant à la transition de la phase exponentielle à la phase linéaire lors de l'amplification par PCR (Figure 33 & Figure 35). Ainsi, la plupart des expériences de PCR cinétiques pourraient être considérées comme arrêtées pendant la phase exponentielle ou au début de la phase linéaire, sans jamais atteindre la phase de « plateau » (Figure 33 & Figure 35).

Pour l'acquisition des données d'efficacité «cycle à cycle», elles ont été calculées comme suit :

- Du cycle 1 au cycle arrivant après le Ct (le Ct est souvent une valeur non entière), l'efficacité correspond à l'efficacité standard, calculée pour chaque plasmide avec le couple d'amorce d'HT17-145 grâce aux gammes de dilution, avec la formule :

$$Efficacité = 10^{-\frac{1}{pente\ de\ la\ gamme\ standard}} - 1$$

- Pour les cycles suivants, l'efficacité a été calculée en fonction des données d'intensité de fluorescence mesurées au LC480, avec la formule :

$$Efficacité\ cycle_i = log_2\left(\frac{intensité\ de\ fluorescence\ cycle_{i+1}}{intensité\ de\ fluorescence\ cycle_i}\right)$$

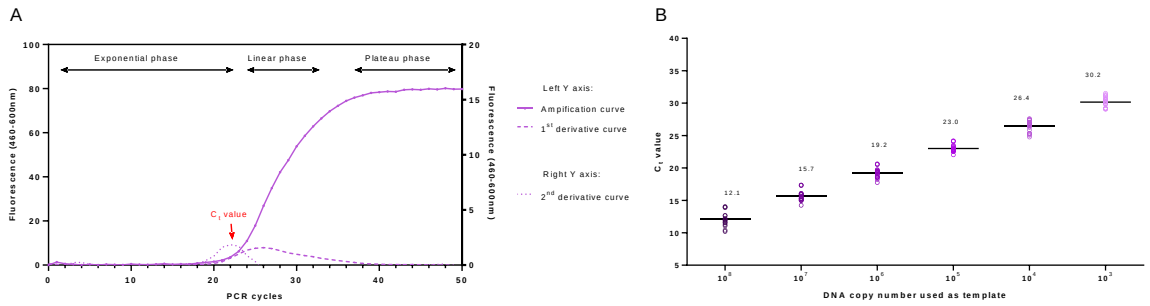

Figure 35 : Les valeurs de Ct, calculées à partir du maximum de la seconde dérivée des courbes d'amplification PCR, indiquent approximativement le cycle auquel une diminution de l'efficacité, ainsi que la transition de la phase exponentielle à la phase linéaire, sont observées au cours de l'amplification PCR. A. Exemple d'une courbe d'amplification PCR sur 50 cycles et de ses premières et secondes dérivées. La valeur de Ct (le maximum de la 2e dérivée), ainsi que les différentes phases de l'amplification PCR, sont indiquées sur le graphique. Cet exemple correspond à une PCR réalisée sur 10^5 copies du plasmide A_{20} sur 50 cycles. La fluorescence de base a été soustraite des données brutes de fluorescence avant le traitement. B. Valeurs de Ct obtenues pour toutes les expériences PCR sur 50 cycles utilisant différentes quantités d'ADN comme matrice. Les cercles correspondent aux valeurs de Ct obtenues à partir de PCR en quadruplicat utilisant des microsatellites mononucléotidiques (A_{15}, A_{19}, A_{20} et A_{24}) et dinucléotidiques ($(AC)_{15}$, $(AC)_{19}$, $(AC)_{20}$ et $(AC)_{24}$) comme matrice. Les barres horizontales et les nombres au-dessus représentent les valeurs moyennes de Ct. Les valeurs de Ct ont été calculées selon la méthode du «maximum de la seconde dérivée» du logiciel LightCycler® 480.

Concernant les expériences de RPA, elles ont été réalisées à 40°C pendant 30 min à partir de 10^5 copies de plasmides comme matrice d'ADN (Figure 36). Chaque expérience a été effectuée en quadruplicat.

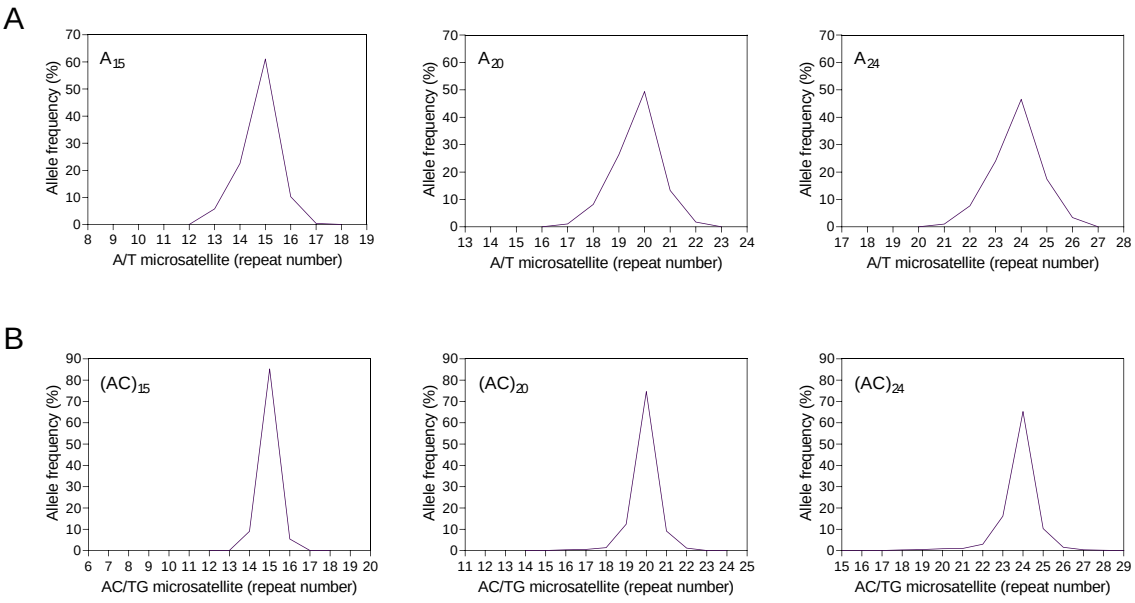

Figure 36 : Fréquences alléliques de microsatellites obtenues à partir des expériences RPA utilisant des microsatellites mononucléotidiques (A) et dinucléotidiques (B) comme matrices (10^5 copies d'ADN). Pour la représentation graphique, les moyennes des fréquences ont été utilisées à partir des expériences RPA réalisées en quadruplicats.

Pour le calcul du nombre de doublement de molécules, elle est faite comme décrite précédemment par la formule :

$$Nb.doublement\ molécule = log_2(\frac{molécule\ d'ADN\ après\ amplification}{molécules\ d'ADN\ avant\ amplification}).$$

Pour ce faire il était nécessaire de connaître la quantité de molécules d'ADN avant et après amplification dans nos expériences de PCR et de RPA, ce qui a été effectué par dosage par fluorométrie à l'aide d'un Qubit dsDNA HS Kit™ (voir Matériels et Méthodes.3.3.2.).

Afin de vérifier que les autres fluorochromes utilisés dans la réaction de PCR en temps réel et RPA n'interféraient pas avec notre dosage post-amplification, nous avons évalué l'effet de leur présence dans différents mélanges d'ADN dosés par fluorimérie. 20 ng/µl d'ADN (ADN commercial Promega) avec des amorces non marquées, des amorces marquées FAM, des amorces marquées HEX et/ou SYTO 9 ont été quantifiées à l'aide du Qubit dsDNA HS Kit™ en appliquant une dilution 1:1 et 1:10. La présence des amorces marquées FAM et/ou de SYTO 9 semble avoir peu d'influence sur la quantification de l'ADN, contrairement aux amorces marquées HEX. Sans le colorant fluorescent du Qubit dsDNA HS Kit™, l'ADN ne peut pas être quantifié. Les quantifications de l'ADN ont été réalisées en triplicat (Figure 37.A).

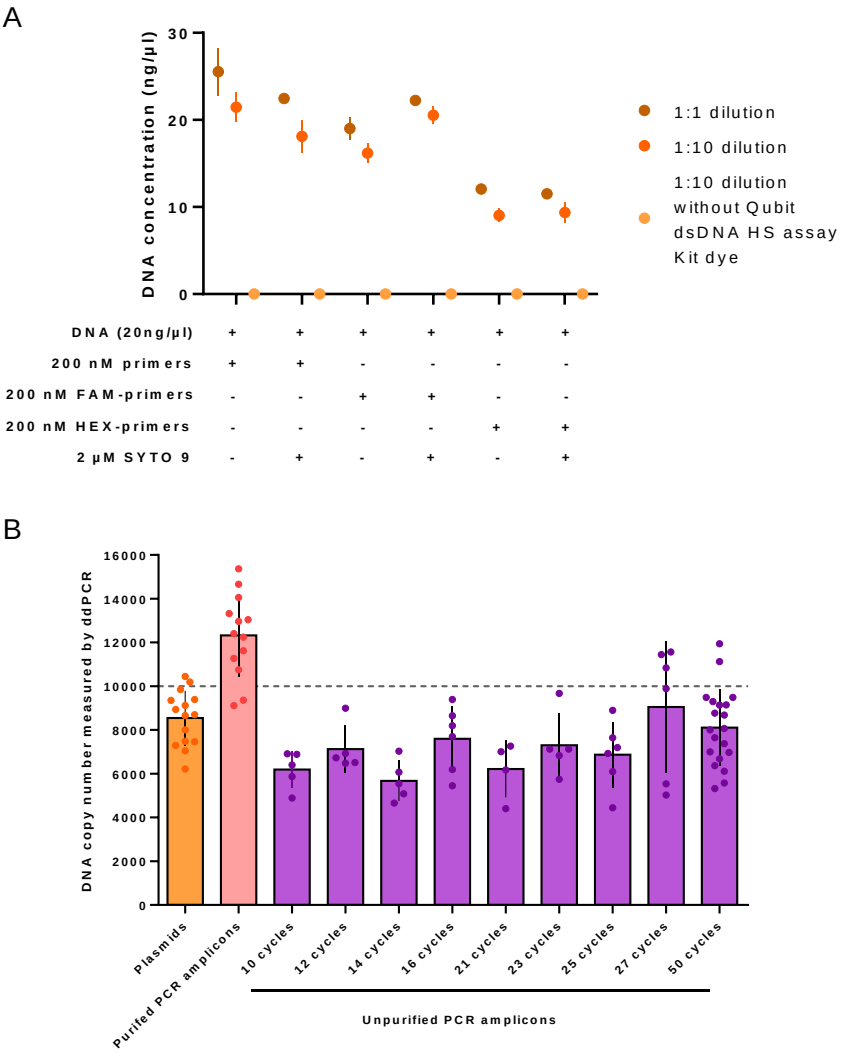

Figure 37 : Précision de la quantification de l'ADN à l'aide du Qubit dsDNA HS Kit™. A. Influence des amorces marquées par FAM, HEX et du colorant SYTO 9 sur la quantification de l'ADN avec le Qubit dsDNA HS Kit™. B. Comparaison des quantifications d'ADN par Qubit dsDNA HS Kit™ et ddPCR.

De plus, afin de vérifier que le dosage par le dsDNA HS Kit™ quantifie correctement les échantillons, une comparaison avec une quantification par ddPCR a été réalisée. Les plasmides modèles, les amplicons PCR purifiés et non purifiés (amplifiés avec des amorces marquées FAM en présence de 2 µM SYTO9 à différents cycles PCR) ont été quantifiés par le Qubit dsDNA HS Kit™. Les échantillons d'ADN ont été dilués à une concentration de 5000 copies d'ADN par µl, puis 10 000 molécules ont été utilisées comme modèles dans les expériences ddPCR réalisées sur un QX200™ (Bio-Rad®) (voir Matériels et Méthodes 3.2.1.4.). Les résultats ddPCR ont montré une légère surestimation de la quantification de l'ADN par le Qubit dsDNA HS Kit™ pour les plasmides modèles et les amplicons non purifiés (18 % et 35 % en moyenne pour les plasmides et les amplicons, respectivement), tandis que la quantification des amplicons PCR purifiés a été légèrement sous-estimée par le Qubit dsDNA HS Kit™ (19 % en moyenne) (Figure 37.B). Ces biais de quantification n'affectent que marginalement le nombre de doubles moléculaires calculés dans nos expériences.

Les taux de mutation ont été inférés en utilisant les 32 modèles génétiques et les trois modes de calcul décrits précédemment (e.v., e.f. et d.m.), en utilisant les données expérimentales moyennes calculées à partir de quadruplicats pour chaque condition. Pour chaque type de microsatellites et conditions, les RMSE obtenus avec chaque modèle ont été classés du plus petit au plus grand (de 1 à 32), puis la médiane des RMSE a été calculée pour un même microsatellite, soit en considérant l'ensemble des conditions, soit en séparant les PCR en phase plateau (50 cycles) et les PCR cinétiques (Tableau 8 & Tableau 9). Les modèles obtenant la valeur la plus basse des modes e.v. et d.m. ont été considérés comme les meilleurs modèles génétiques sous-jacents responsables du glissement de la polymérase et des fréquences d'allèles microsatellites observées dans mes expériences de PCR. Ainsi, alors que les deux types de microsatellites permettent des insertions à un seul pas et des suppressions à plusieurs pas, les taux de mutation augmentent préférentiellement de manière exponentielle (10. E3d1u) pour les microsatellites mononucléotidiques et linéairement (19. L4d1u) pour les microsatellites dinucléotidiques (Tableau 8 & Tableau 9).

Tableau 8 : Identification des modèles génétiques les mieux ajustés pour chaque microsatellite mononucléotidique selon les trois modes d'inférence des taux de mutation des microsatellites et les types d'expériences de PCR.

Mode	Rang	Modèles génétiques les mieux ajustés basés sur le classement médian obtenu à partir des expériences de PCR[1]											
		Toutes données PCR				Données PCR 50-cycles				Données de PCR cinétique (10-16 et 21-27 cycles)			
		A_{15}	A_{19}	A_{20}	A_{24}	A_{15}	A_{19}	A_{20}	A_{24}	A_{15}	A_{19}	A_{20}	A_{24}
Efficacité Variable	1	E2d1 **E3d1u**[2]	L2d3u **E3d1u**	**E3d1u**	**E3d1u** E3d4u	**E3d1u**	**E3d1u**	**E3d1u**	E3d2u E3d4u	**E3d1u**	E4d1u	E4d1u	**E3d1u**
	2	L3d1u	E4d1u	E4d1u	E2d1u E3d2u	E2d1u	L2d1u	L2d1u E3d4u	**E3d1u**	L3d1u E2d1u	L3d2u E3d3u	L3d2u **E3d1u**	E2d1u
	3	L4d1u E4d1u	E2d1u E3d3u	L2d1u L2d4u	L2d1u L2d2u	L2d1u E2d2u	E3d2u	L2d3u	E3d3u	L4d1u	**E3d1u**	E3d2u	L2d1u
Efficacité Fixée	1	L2d1u L3d1u	L2d2u	L2d1u	L2d4u	E3d3u	**E3d1u**	L2d1u	L2d1u	L3d1u E4d1u	L2d2u	E4d3u	L2d4u **E3d1u**
	2	L4d1u	L2d1u	E3d3u	**E3d1u**	L2d1u	L2d1u	L2d4u	L2d4u	L4d1u	L2d1u	E3d3u	L3d1u E3d3u
	3	L4d3u E4d1u E4d2u	L2d4u	L3d1u L4d1u L2d4u	L3d1u	L2d3u E2d3u	L2d4u	L3d1u	L3d1u E3d2u	L4d3u	E4d1u	E4d1u	L2d1u
Doublement de molécules	1	**E3d1u**	**E3d1u**	L2d4u **E3d1u** E4d1u	**E3d1u**	E2d1u	**E3d1u**	E3d3u	**E3d1u**	L3d1u L4d1u E4d1u	L2d1u	E4d1u	L2d1u
	2	L3d1u L4d1	L2d1u	L4d1u	E2d1u E2d4u	E2d2u	L2d2u	**E3d1u**	E2d1u	L3d2u	L2d4u	L2d4u	L3d1u L4d1u
	3	E2d1u E2d3u	L2d3u L2d4u	L2d1u	L2d1u	**E3d1u**	L2d1u E2d1u	L4d1u L2d2u	E2d4u	L4d2u L3d3u L4d3u	**E3d1u**	L3d1u **E3d1u**	E2d4u

[1] Pour chaque point expérimental (6 pour la PCR de 50 cycles et 8 pour la PCR cinétique), les fréquences alléliques des microsatellites ont été moyennées à partir d'expériences en quadruplicate pour déduire les taux de mutation des microsatellites en utilisant les 32 modèles génétiques et les trois modes d'inférence. Les RMSE obtenues ont été classées de 1 à 32 pour chaque point expérimental, et la médiane des rangs a été utilisée pour identifier le modèle génétique le mieux ajusté.

[2] Le modèle classé premier dans la plupart des cas, à la fois dans les modes «efficacité variable» et «doublement des molécules», a été considéré comme le modèle génétique le mieux ajusté et a été mis en gras et coloré en rouge.

Tableau 9 : Identification des modèles génétiques les mieux ajustés pour chaque microsatellite mononucléotidique selon les trois modes d'inférence des taux de mutation des microsatellites et les types d'expériences de PCR.

Mode	Rang	Modèles génétiques les mieux ajustés basés sur le classement médian obtenu à partir des expériences de PCR[1]											
		Toutes données PCR				Données PCR 50-cycles				Données de PCR cinétique (10-16 et 21-27 cycles)			
		$(AC)_{15}$	$(AC)_{19}$	$(AC)_{20}$	$(AC)_{24}$	$(AC)_{15}$	$(AC)_{19}$	$(AC)_{20}$	$(AC)_{24}$	$(AC)_{15}$	$(AC)_{19}$	$(AC)_{20}$	$(AC)_{24}$
Efficacité Variable	1	**L4d1u**[2]	**L4d1u**	**L4d1u**	**L4d1u** L4d2u	**L4d1u**	**L4d1u**	**L4d1u**	L4d2u L4d4u	**L4d1u**	**L4d1u**	**L4d1u**	**L4d1u** L4d2u
	2	L3d1u E4d1u	L4d2u	L3d1u L4d2u L4d3u	L4d3u L4d4u	L3d2u	L4d2u L4d3u L4d4u	L4d2u	**L4d1u** L4d3u	E4d1u	L4d2u	L3d1u	L4d4u
	3	L3d2u L4d3u E4d2u	L3d1u L4d4u	L4d4u	L3d1u L3d4u	L3d1u	L3d1u	L4d3u L4d4u	L3d4u	L3d1u	L3d1u L4d3u	L3d4u E4d1u	L3d1u L3d4u
Efficacité Fixée	1	E4d1u	**L4d1u**	**L4d1u**	**L4d1u**	**L4d1u**	**L4d1u**	L4d2u	**L4d1u** L4d3u	E4d1u	L4d4u	**L4d1u**	**L4d1u** L4d4u
	2	**L4d1u**	L4d4u	L4d2u L4d3u	L4d4u	L3d3u L4d3u	L4d3u L4d4u	**L4d1u**	L4d2u L4d4u	L3d1u **L4d1u**	**L4d1u**	L4d3u	L3d1u
	3	L3d1u	L3d1u L4d2u L4d3u	E4d2u	L4d2u L4d3u	L3d1u E4d1u	L3d1u L4d2u	L4d3u	E4d1u E4d2u E4d3u	E3d1u	L3d1u	L3d1u	L4d3u
Doublement de molécules	1	L3d1u **L4d1u**	**L4d1u**	**L4d1u**	**L4d1u**	L3d1u	**L4d1u**	**L4d1u**	**L4d1u** L4d2u L4d3u	L3d1u **L4d1u**	**L4d1u**	**L4d1u**	**L4d1u** L4d4u
	2	E4d1u	L4d3u	L4d2u L4d3u L4d4u	L4d2u L4d3u L4d4u	**L4d1u** L3d3u	L4d2u L4d3u	L4d2u L4d3u L4d4u	L4d4u	E4d1u	L3d1u L4d3u	L3d1u L4d3u	L3d1u
	3	L3d3u L4d3u	L3d1u L4d4u	L3d1u	L3d1u E4d2u	L4d3u	L4d4u	E4d4u	L3d4u	E3d1u	L4d4u	L4d2u E4d2u	L3d3u

[1] Pour chaque point expérimental (6 pour la PCR de 50 cycles et 8 pour la PCR cinétique), les fréquences alléliques des microsatellites ont été moyennées à partir d'expériences en quadruplicate pour déduire les taux de mutation des microsatellites en utilisant les 32 modèles génétiques et les trois modes d'inférence. Les RMSE obtenues ont été classées de 1 à 32 pour chaque point expérimental, et la médiane des rangs a été utilisée pour identifier le modèle génétique le mieux ajusté.

[2] Le modèle classé premier dans la plupart des cas, à la fois dans les modes «efficacité variable» et «doublement des molécules», a été considéré comme le modèle génétique le mieux ajusté et a été mis en gras et coloré en rouge.

Les taux de mutation des microsatellites obtenus à partir des trois modes ont montré un fort biais pour les délétions, plus prononcé pour les microsatellites dinucléotidiques que pour les mononucléotidiques (Figure 38). En raison de l'utilisation de l'efficacité cycle par cycle, les résultats du mode e.v. peuvent être considérés comme les plus précis et ont montré une bonne concordance entre les expériences de PCR de 50 cycles et de PCR cinétiques (Figure 38). Les résultats étaient également concordants pour le mode d.m. qui a donné une bonne approximation des taux de mutation pour tous les types d'expériences de PCR, similaire au mode e.v. (Figure 38). En revanche, malgré des résultats similaires pour les données de PCR cinétiques, le mode e.f. a montré une variabilité accrue pour les différents points (allant de 10^3 à 10^8 copies d'ADN comme matrice) ainsi qu'une légère sous-estimation moyenne des taux de mutation inférés des expériences de PCR de 50 cycles en raison de la phase de « plateau » atteinte à différents cycles (Figure 38 & Figure 33).

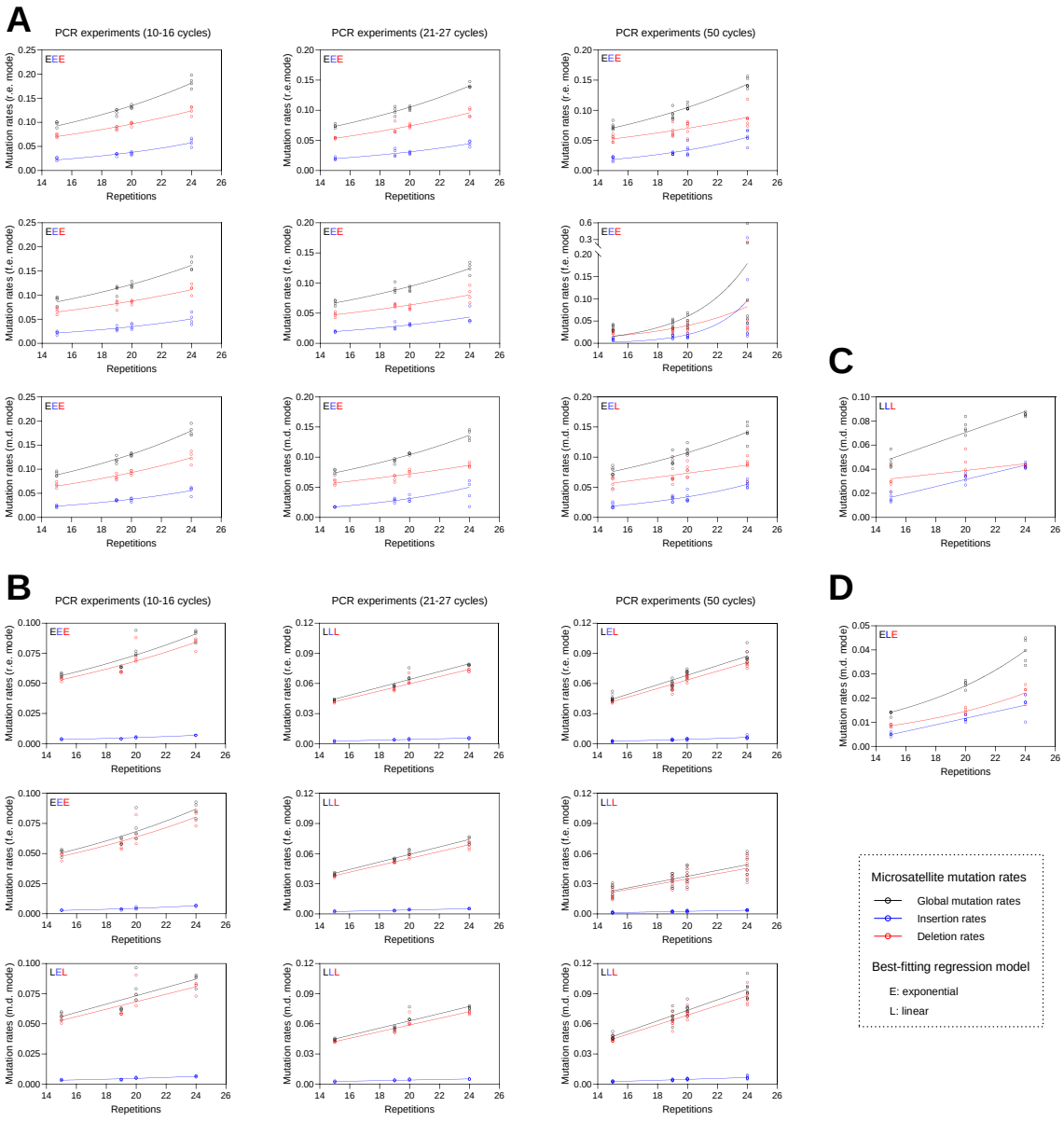

Figure 38 : Taux de mutation inférés à partir des données expérimentales de PCR et RPA. Les taux de mutation ont été calculés en utilisant les trois modes d'inférence (e.v., e.f. et d.m.) pour : (A) les microsatellites mononucléotidiques (A_{15}, A_{19}, A_{20} et A_{24}) (modèle E3d1u), (B) les microsatellites dinucléotidiques ($(AC)_{15}$, $(AC)_{19}$, $(AC)_{20}$ et $(AC)_{24}$) (modèle L4d1u) en utilisant soit des expériences de PCR de 10 à 16 cycles ou de 21 à 27 cycles dans la phase de «cinétique», soit des expériences de PCR de 50 cycles dans la phase de «plateau». Pour les expériences de RPA, les taux de mutation ont été inférés avec le mode d.m. pour (C) les microsatellites mononucléotidiques (A_{15}, A_{19}, A_{20} et A_{24}), (D) les microsatellites dinucléotidiques ($(AC)_{15}$, $(AC)_{19}$, $(AC)_{20}$ et $(AC)_{24}$). Pour la PCR, les taux de mutation ont été inférés à partir de la moyenne des fréquences alléliques des microsatellites des réplicats expérimentaux pour chaque point. Pour la RPA, les taux de mutation ont été inférés à partir des fréquences alléliques des microsatellites des réplicats expérimentaux ainsi que de leur moyenne pour le même point. La meilleure régression ajustée entre des courbes de croissance linéaire (L) et exponentielle (E) est montrée dans chaque graphique et colorée selon les taux de mutation.

Il est intéressant de noter que lorsque l'on régresse les taux de mutations en fonction du nombre de répétitions, les courbes de régression les mieux adaptées montraient principalement une courbe exponentielle croissante pour les microsatellites mononucléotidiques et une ligne croissante pour les microsatellites dinucléotidiques (Figure 38 & Tableau 10 & Tableau 11). Ces résultats semblent soutenir

notre approche basée sur le RMSE pour l'identification du modèle génétique sous-jacent au glissement de la polymérase.

Tableau 10 : Statistiques de régression linéaire et exponentielle calculées pour les taux de mutation des microsatellites à répétitions mononucléotidiques obtenues à partir des expériences de PCR.

Type de données	Mode	Taux de mutation	Regression linéaire			Regression exponentielle			Meilleur modèle de régression[1]
			Equation	R^2	Somme des carrés	Equation	R^2	Somme des carrés	
50 cycles PCR	Efficacité variable	Global	Y=0.007938X-0.05077	0.907	0.001595	$Y=0.0218e^{(0.07851X)}$	0.936	0.001092	Exponentielle
		Insertion	Y=0.003982X-0.04302	0.739	0.001378	$Y=0.002969e^{(0.122X)}$	0.790	0.001111	Exponentielle
		Deletion	Y=0.003957X-0.00775	0.615	0.002414	$Y=0.0222e^{(0.05757X)}$	0.622	0.002367	Exponentielle
10-16 cycles PCR		Global	Y=0.009643X-0.05463	0.914	0.001440	$Y=0.03017e^{(0.07475X)}$	0.948	0.000871	Exponentielle
		Insertion	Y=0.003787X-0.03584	0.823	0.000507	$Y=0.004511e^{(0.1062X)}$	0.881	0.000339	Exponentielle
		Deletion	Y=0.005857X-0.01879	0.908	0.000571	$Y=0.02756e^{(0.06264X)}$	0.929	0.000442	Exponentielle
21-27 cycles PCR		Global	Y=0.007426X-0.0405	0.952	0.000452	$Y=0.0248e^{(0.07229X)}$	0.970	0.000285	Exponentielle
		Insertion	Y=0.002738X-0.02239	0.823	0.000265	$Y=0.005089e^{(0.0905X)}$	0.847	0.000229	Exponentielle
		Deletion	Y=0.004689X-0.01811	0.883	0.000477	$Y=0.02034e^{(0.06466X)}$	0.895	0.000428	Exponentielle
50 cycles PCR	Efficacité fixée	Global	Y=0.0161X-0.2359	0.204	0.248900	$Y=0.0002699e^{(0.2708X)}$	0.266	0.229500	Exponentielle
		Insertion	Y=0.009596X-0.1528	0.207	0.086690	$Y=0.000006985e^{(0.3972X)}$	0.283	0.078420	Exponentielle
		Deletion	Y=0.006501X-0.08305	0.197	0.042480	$Y=0.001048e^{(0.182X)}$	0.241	0.040120	Exponentielle
10-16 cycles PCR		Global	Y=0.008279X-0.04019	0.880	0.001531	$Y=0.03006e^{(0.07021X)}$	0.908	0.001182	Exponentielle
		Insertion	Y=0.003241X-0.02863	0.732	0.000631	$Y=0.005105e^{(0.09568X)}$	0.751	0.000587	Exponentielle
		Deletion	Y=0.005038X-0.01156	0.819	0.000920	$Y=0.02646e^{(0.05988X)}$	0.846	0.000785	Exponentielle
21-27 cycles PCR		Global	Y=0.006298X-0.02932	0.894	0.000775	$Y=0.02406e^{(0.06838X)}$	0.909	0.000661	Exponentielle
		Insertion	Y=0.002677X-0.0217	0.668	0.000585	$Y=0.005221e^{(0.08848X)}$	0.676	0.000570	Exponentielle
		Deletion	Y=0.003621X-0.007622	0.740	0.000755	$Y=0.01979e^{(0.05848X)}$	0.754	0.000714	Exponentielle
50 cycles PCR	Doublement de molécule	Global	Y=0.007265X-0.03524	0.856	0.002185	$Y=0.02701e^{(0.06906X)}$	0.869	0.001993	Exponentielle
		Insertion	Y=0.003957X-0.04254	0.822	0.000833	$Y=0.003168e^{(0.1189X)}$	0.861	0.000650	Exponentielle
		Deletion	Y=0.003308X+0.0073	0.507	0.002614	$Y=0.02908e^{(0.0458X)}$	0.505	0.002626	Linéaire
10-16 cycles PCR		Global	Y=0.01005X-0.06585	0.939	0.001077	$Y=0.02701e^{(0.07895X)}$	0.963	0.000651	Exponentielle
		Insertion	Y=0.003607X-0.03252	0.862	0.000341	$Y=0.005476e^{(0.09666X)}$	0.878	0.000301	Exponentielle
		Deletion	Y=0.006439X-0.03333	0.873	0.000987	$Y=0.02224e^{(0.07157X)}$	0.898	0.000795	Exponentielle
21-27 cycles PCR		Global	Y=0.006864X-0.0309	0.934	0.000548	$Y=0.02677e^{(0.06786X)}$	0.953	0.000388	Exponentielle
		Insertion	Y=0.003622X-0.03888	0.801	0.000536	$Y=0.003018e^{(0.1171X)}$	0.828	0.000463	Exponentielle
		Deletion	Y=0.003242X+0.007981	0.824	0.000369	$Y=0.02873e^{(0.04598X)}$	0.833	0.000349	Exponentielle

[1] Le modèle de régression le mieux ajusté a été déterminé comme étant celui ayant le plus faible R^2 et la somme des carrés la plus faible entre les régressions linéaire et exponentielle.

Tableau 11 : Statistiques de régression linéaire et exponentielle calculées pour les taux de mutation des microsatellites à répétitions di-nucléotidiques obtenues à partir des expériences de PCR.

Type de données	Mode	Taux de mutation	Regression linéaire			Regression exponentielle			Meilleur modèle de régression[1]
			Equation	R^2	Sum of Squares	Equation	R^2	Equation	
50 cycles PCR	Efficacité variable	Global	Y=0.004752X-0.02687	0.909	0.000651	$Y=0.01576e^{(0.07195X)}$	0.909	0.000653	Linéaire
		Insertion	Y=0.0004303X-0.003998	0.752	0.000018	$Y=0.000627e^{(0.09741X)}$	0.752	0.000018	Exponentielle
		Deletion	Y=0.004322X-0.02287	0.913	0.000512	$Y=0.01526e^{(0.07011X)}$	0.913	0.000514	Linéaire
10-16 cycles PCR		Global	Y=0.003843X-0.002193	0.780	0.000682	$Y=0.02547e^{(0.05309X)}$	0.786	0.000665	Exponentielle
		Insertion	Y=0.000372X-0.002253	0.810	0.000005	$Y=0.001045e^{(0.07862X)}$	0.860	0.000004	Exponentielle
		Deletion	Y=0.003471X+0.0000597	0.767	0.000601	$Y=0.02458e^{(0.05129X)}$	0.769	0.000595	Exponentielle
21-27 cycles PCR		Global	Y=0.003916X-0.0144	0.935	0.000175	$Y=0.01828e^{(0.06164X)}$	0.915	0.000229	Linéaire
		Insertion	Y=0.0003354X-0.002415	0.848	0.000003	$Y=0.0008846e^{(0.07748X)}$	0.812	0.000004	Linéaire
		Deletion	Y=0.003581X-0.01198	0.929	0.000161	$Y=0.01747e^{(0.06048X)}$	0.910	0.000204	Linéaire
50 cycles PCR	Efficacité fixée	Global	Y=0.002898X-0.02051	0.615	0.001510	$Y=0.007562e^{(0.07847X)}$	0.602	0.001561	Linéaire
		Insertion	Y=0.0002641X-0.002636	0.719	0.000008	$Y=0.000324e^{(0.1025X)}$	0.706	0.000008	Linéaire
		Deletion	Y=0.002634X-0.01787	0.596	0.001352	$Y=0.007297e^{(0.07666X)}$	0.583	0.001394	Linéaire
10-16 cycles PCR		Global	Y=0.004036X-0.01113	0.790	0.000708	$Y=0.02049e^{(0.06023X)}$	0.799	0.000679	Exponentielle
		Insertion	Y=0.0004238X-0.003701	0.820	0.000006	$Y=0.0006611e^{(0.09658X)}$	0.857	0.000005	Exponentielle
		Deletion	Y=0.003612X-0.007432	0.781	0.000600	$Y=0.02012e^{(0.05767X)}$	0.788	0.000582	Exponentielle
21-27 cycles PCR		Global	Y=0.003736X-0.01553	0.956	0.000107	$Y=0.01633e^{(0.0634X)}$	0.932	0.000162	Linéaire
		Insertion	Y=0.000314X-0.002311	0.791	0.000004	$Y=0.0007474e^{(0.08184X)}$	0.790	0.000004	Linéaire
		Deletion	Y=0.003422X-0.01322	0.957	0.000087	$Y=0.01564e^{(0.06211X)}$	0.932	0.000136	Linéaire
50 cycles PCR	Doublement de molécule	Global	Y=0.005226X-0.03107	0.891	0.000957	$Y=0.01686e^{(0.07228X)}$	0.876	0.001087	Linéaire
		Insertion	Y=0.0004512X-0.004263	0.759	0.000019	$Y=0.0006505e^{(0.09722X)}$	0.747	0.000019	Linéaire
		Deletion	Y=0.004775X-0.0268	0.895	0.000768	$Y=0.01634e^{(0.07056X)}$	0.880	0.000878	Linéaire
10-16 cycles PCR		Global	Y=0.003452X+0.004243	0.677	0.000934	$Y=0.02767e^{(0.04812X)}$	0.676	0.000935	Linéaire
		Insertion	Y=0.0003476X-0.002027	0.734	0.000007	$Y=0.001055e^{(0.07561X)}$	0.761	0.000006	Exponentielle
		Deletion	Y=0.003104X+0.00627	0.659	0.000819	$Y=0.02683e^{(0.04623X)}$	0.657	0.000824	Linéaire
21-27 cycles PCR		Global	Y=0.003599X-0.009049	0.891	0.000260	$Y=0.01956e^{(0.05761X)}$	0.874	0.000301	Linéaire
		Insertion	Y=0.0002835X-0.001591	0.778	0.000004	$Y=0.001028e^{(0.06777X)}$	0.736	0.000004	Linéaire
		Deletion	Y=0.003316X-0.007457	0.888	0.000228	$Y=0.01857e^{(0.05688X)}$	0.873	0.000257	Linéaire

[1] Le modèle de régression le mieux ajusté a été déterminé comme étant celui ayant le plus faible R^2 et la somme des carrés la plus faible entre les régressions linéaire et exponentielle.

Notre algorithme a été ensuite évalué en utilisant les fréquences alléliques de microsatellites obtenues à partir des expériences de RPA utilisant 10^5 copies de plasmides comme matrices d'ADN. Comme la RPA est une approche isotherme réalisée en une seule étape d'amplification de l'ADN, le seul mode d'inférence utilisable avec les données de RPA était le mode d.m. En raison de la faible quantité de données RPA, chaque fréquence d'allèle de microsatellite répliquée a été considérée ainsi que leur moyenne comme données d'entrée pour l'inférence des taux de mutation des microsatellites. Bien que les résultats soient moins clairs qu'avec la PCR, nous avons pu identifier les modèles génétiques sous-jacents 4. E2d1u et 12. E3d2u pour les microsatellites mono- et di-nucléotidiques, respectivement, sur la base de la médiane des rangs de RMSE (Tableau 12).

Tableau 12 : Identification du meilleur modèle génétique pour chaque microsatellite mono- et di-nucléotidiques par RPA utilisant l'inférence du taux de mutations basé sur le mode de doublement de molécule.

Mode	Rang	Modèles génétiques les mieux ajustés basés sur le classement médian obtenu à partir des expériences de RPA[1]							
		A_{15}	A_{20}	A_{24}	A_{15-24}	$(AC)_{15}$	$(AC)_{20}$	$(AC)_{24}$	$(AC)_{15-24}$
Doublement de moélcule	1	E2d4u	E2d2u	**E2d1u**	**E2d1u**	L1d2u	**E3d2u**	**E3d2u**	L4d2u L4d3u
	2	**E2d1u**	**E2d1u**	E2d4u	E2d4u	L1d1u E2d1u	L3d4u	L4d3u	L3d3u
	3	L2d1u	L2d1u	E2d2u	E2d2u	L1d4u	L4d2u	L4d2u, L3d4u	L3d4u **E3d2u**
	4	L2d4u	L2d4u	L2d2u	L2d1u	L4d1u E2d1u	E3d1u	L3d3u	L2d2u E4d1u

[1] Pour chaque point expérimental, chaque fréquence allèlique de microsatellite répliquée ainsi que les fréquences alléliques moyennes des microsatellites ont été utilisées pour déduire les taux de mutation des microsatellites en utilisant les 32 modèles génétiques et le mode de doublement des molécules. Chaque microsatellite a été considéré individuellement ou regroupé par type de microsatellite (mono- ou dinucléotidique). Les RMSE obtenues ont été classées de 1 à 32 pour chaque point expérimental, et la médiane des rangs a été utilisée pour identifier le modèle génétique le mieux ajusté.

[2] Le modèle classé premier dans la plupart des cas a été considéré comme le modèle génétique le mieux ajusté, et il a été mis en gras et coloré en rouge.

Les taux de mutation des mononucléotides ont montré des taux croissants avec le nombre de répétitions, plus prononcés pour les insertions que pour les délétions, ce qui a conduit à des taux de mutation équivalents pour les insertions et les délétions à 24 répétitions (Figure 38.C). En ce qui concerne les microsatellites dinucléotidiques, les taux d'insertion et de délétion avaient tendance à augmenter de manière similaire, les taux de délétion étant légèrement supérieurs aux taux d'insertion à chaque répétition (Figure 38.D). La régression des taux de mutation en fonction du nombre de répétitions était mieux ajustée pour les microsatellites dinucléotidiques par une courbe exponentielle pour les taux de mutation globaux et de délétion et par une droite pour les taux d'insertion, tandis que pour les microsatellites mononucléotidiques seules des droites ont été obtenues (Figure 38.C & D & Tableau 13) ce qui est partiellement en désaccord avec les modèles génétiques sous-jacents identifiés par notre algorithme basé sur le RMSE.

Tableau 13 : Statistiques des régressions linéaires et exponentielles calculées pour les taux de mutations microsatellites mono- et di-nucléotidiques obtenus par PCR.

Type de Microsatellite	Taux de mutation	Regression Linéaire			Regression Exponentielle			Meilleur modèle de regression[1]
		Equation	R^2	Somme des carrés	Equation	R^2	Somme des carrés	
Mononucleotide	Global	Y=0.004361X-0.01673	0.897	0.000446	$Y=0.0199e^{(0.06202X)}$	0.863	0.000591	Linéaire
	Insertion	Y=0.002969X-0.02785	0.934	0.000128	$Y=0.004245e^{(0.09733X)}$	0.902	0.000187	Linéaire
	Deletion	Y=0.001392X+0.01112	0.357	0.000709	$Y=0.01922e^{(0.03493X)}$	0.342	0.000726	Linéaire
Dinucleotide	Global	Y=0.002858X-0.02986	0.927	0.000131	$Y=0.002531e^{(0.1148X)}$	0.936	0.000115	Exponentielle
	Insertion	Y=0.00136X-0.01553	0.824	0.000080	$Y=0.0008494e^{(0.1262X)}$	0.816	0.000084	Linéaire
	Deletion	Y=0.001498X-0.01434	0.884	0.000060	$Y=0.001746e^{(0.106X)}$	0.906	0.000049	Exponentielle

[1] Le modèle de régression le mieux ajusté a été déterminé comme étant celui ayant le plus faible R^2 et la somme des carrés la plus faible entre les régressions linéaire et exponentielle.

1.2. Discussion

Depuis leur découverte dans les années 80, les séquences d'ADN microsatellites ont fait l'objet de nombreuses études fondamentales visant à décrire, comprendre et expliquer les mécanismes mutationnels et évolutifs responsables de leur fort polymorphisme dans les génomes de presque tous les organismes vivants (Ellegren, 2004). Il a été prouvé que le glissement de polymérase est le principal mécanisme introduisant des mutations microsatellites lors de la réplication de l'ADN (Ellegren, 2004; Kelkar et al., 2010; Levinson and Gutman, 1987a), qui peuvent être caractérisées par leurs taux, leur directionnalité (insertion ou délétion) et le nombre d'unités répétées glissées par mutation (appelé « pas » dans notre étude). Les taux de mutation des microsatellites peuvent être déterminés par diverses approches moléculaires, mathématiques et/ou computationnelles en utilisant des données obtenues à partir d'expériences *in vitro*, *ex/in vivo* et *in populo*. Il est à noter que, selon le type d'études ou d'analyses, le taux de mutation des microsatellites peut avoir des unités de temps différentes. Il peut être défini comme (i) mutations par génération dans des études utilisant les génotypes microsatellites issus de pedigrees, familles et tests de paternité (Ballantyne et al., 2010; Chandra et al., 2022; Steely et al., 2022; Whittaker et al., 2003) ou de sujets non apparentés de la même espèce (génomique des populations) ou d'espèces étroitement apparentées (génomique comparative) (Kelkar et al., 2008; Sainudiin et al., 2004; Webster et al., 2002; Wu and Drummond, 2011). Il peut également correspondre à (ii) des mutations par division cellulaire dans des études portant sur les mutations microsatellites dans les génomes des cellules d'un tissu (Christopher et al., 2019) ou dans des plasmides rapporteurs utilisés pour transformer des bactéries (Eckert and Yan, 2000; Levinson and Gutman, 1987b), des levures (Sia et al., 1997b; Wierdl et al., 1997) et des lignées cellulaires (Baptiste et al., 2013). Enfin, il peut être défini comme (iii) mutations par réplication dans des études *in vitro* basées sur « *HSV-tk forward mutation assay* » (Abdulovic et al., 2011; Baptiste et al., 2013; Eckert et al., 2002) ou des expériences de PCR (Shinde et al., 2003). L'unité de temps utilisée pour les taux de mutation des microsatellites, c'est-à-dire génération, division cellulaire ou réplication, est importante car les taux calculés à partir d'unités de temps différentes ne sont pas toujours comparables. Des taux de mutation plus élevés obtenus à partir des mutations par génération sont observés dans les microsatellites de la lignée paternelle par rapport à ceux de la lignée maternelle, probablement en raison d'un plus grand nombre de divisions des cellules germinales chez les mâles que chez les femelles (Kelkar et al., 2008; Sun et al., 2012; Wilson Sayres and Makova, 2011). En revanche, les taux de mutation calculés en utilisant les divisions cellulaires et les événements de réplication sont généralement comparables, car les divisions cellulaires approximent le nombre de réplications (Henderson and Petes, 1992). Ainsi, les études basées sur des expériences *in vitro* ont permis de déterminer les taux de mutation des microsatellites à partir du plus petit événement génétique, à savoir une réplication de l'ADN.

Dans cette première partie, nous avons développé une approche simple pour inférer les taux de mutation des microsatellites *in vitro* à partir des fréquences d'allèles microsatellites obtenues à partir d'expériences de PCR et RPA. Cette approche permet également d'identifier le modèle génétique le mieux adapté parmi les trente-deux proposés pour le mécanisme de mutation par glissement de la réplication des microsatellites à partir de nos données expérimentales. Ces modèles étaient en accord avec les résultats précédents des études décrivant les mutations microsatellites et/ou modélisant les mécanismes mutationnels des microsatellites. La présence de mutations microsatellites de plusieurs unités répétées a été décrite *in populo* et *in vivo* dans des études sur une génération (Huang et al., 2002; Sun et al., 2012) et après plusieurs divisions cellulaires (Sia et al., 1997b; Wierdl et al., 1997), , respectivement, ce qui a été confirmé *in vitro* à partir d'expériences « *HSV-tk forward mutation assay* » consistant en une unique réplication de l'ADN (Baptiste et al., 2013; Eckert et al., 2002). Ces mutations multi-pas consistaient principalement en 2 à 5 répétitions, augmentant avec l'accroisement du nombre de répétitions des microsatellites et diminuant de manière exponentielle à mesure que le nombre de

pas augmente (Baptiste et al., 2013; Huang et al., 2002; Sun et al., 2012; Whittaker et al., 2003). Notre approche a donc considéré seize possibilités pour les mutations microsatellites de 1 à 4 pas (insertion ou délétion) par événement mutationnel, dont les taux diminuaient exponentiellement avec l'augmentation du pas. De plus, la plupart des études ont également proposé une relation exponentielle croissante entre les taux de mutation des microsatellites et la longueur de l'allèle microsatellite (Baptiste et al., 2013; Eckert et al., 2002; Kelkar et al., 2010; Whittaker et al., 2003), tandis qu'un plus petit nombre ont suggéré une relation linéaire croissante (Shinde et al., 2003; Sun et al., 2012). Nous avons ainsi considéré deux relations (linéaire ou exponentielle) croissantes entre les taux de mutation des microsatellites et le nombre d'unités répétées, conduisant à un total de trente-deux modèles génétiques proposés pour le mécanisme de mutation des microsatellites dû au glissement de polymérase.

L'algorithme développé a utilisé une chaîne de Markov pour le calcul des fréquences d'allèles microsatellites calibrées pour les trente-deux modèles et les paramètres ont été ajustés avec l'algorithme SA (Kirkpatrick et al., 1983) couplé à une recherche en grille en utilisant le critère de minimisation du RMSE. Il est à noter que certaines études avaient également proposé des approches pour inférer les taux de mutation des microsatellites en utilisant différents modèles génétiques, cependant, elles avaient toutes été développées pour les données de génétique des populations (Whittaker et al., 2003; Wu and Drummond, 2011), tandis que d'autres approches développées pour les données *in vitro* visaient spécifiquement à identifier les génotypes microsatellites à partir des profils post-PCR incluant les artefacts *stutter* sans fournir de taux de mutation des microsatellites (Highnam et al., 2013; Raz et al., 2019).

Les performances de l'algorithme pour l'inférence des taux de mutation des microsatellites ont été évaluées en se basant sur les valeurs de RMSE et RR en utilisant des données artificielles générées dans différentes conditions. Le nombre de pas inclus dans les modèles a eu un effet sur les performances de l'algorithme : les modèles à un seul pas ont généralement montré de meilleurs résultats par rapport aux modèles à quatre pas (RMSE et RR inférieurs) dans toutes les conditions testées, ce qui était probablement dû au nombre réduit de paramètres à inférer dans les modèles à un seul pas (4 au lieu de 6). Des performances inférieures ont principalement été observées pour les taux de mutation avec de plus grands pas dans les modèles à quatre pas, mais cela ne devrait pas impacter les taux de mutation globaux en raison de la diminution exponentielle des taux de mutation avec l'augmentation des pas. En revanche, la relation entre les taux de mutation et le nombre de répétitions (qu'elle soit linéaire ou exponentielle) n'a pas eu un impact important sur les performances lorsque le même nombre de pas a été pris en compte dans les modèles. Parmi les paramètres testés, nous avons identifié que la précision de l'algorithme diminuait progressivement avec des valeurs de taux de mutation plus faibles à inférer, des différences accrues entre les taux d'insertion et de délétion ainsi qu'avec des valeurs plus élevées des coefficients α ou β et un nombre plus élevé de cycles de PCR, tandis que l'allèle de référence et le nombre de données considérées avaient peu d'influence sur les performances d'inférence. Le risque de rencontrer des conditions menant à de faibles performances de l'algorithme devrait être très peu probable avec des données expérimentales en conditions réelles de PCR, telles que des réactions de PCR entièrement efficaces au-delà de 50 cycles ou des polymérases avec des taux d'erreur microsatellites inférieurs à 10^{-5}. L'algorithme a aussi été utilisé pour l'identification des modèles génétiques à la base des profils microsatellites obtenus parmi les 32 modélisés. Nous avons proposé que le modèle obtenant la valeur de RMSE la plus basse pourrait être le plus proche modèle génétique sous-jacent. Cela a été illustré en utilisant des données artificielles générées avec les 32 modèles, et le RMSE calculé le plus bas a été capable d'identifier correctement chaque modèle génétique sous-jacent, évaluant notre approche. Enfin, trois modes ont été implémentés pour inférer les taux de mutation des microsatellites à partir d'expériences de PCR basées sur (i) des efficacités variables cycle par cycle, (ii) une efficacité fixe de 100 %, et (iii) des doublages de molécule. Ce dernier présente l'avantage d'être également compatible avec

les méthodes isothermes d'amplification d'ADN (Figure 38). Comparé au mode standard (i), le mode (ii) ne devrait être utilisé qu'avant la phase de plateau, tandis que le mode (iii) devrait fournir une bonne approximation des taux de mutation à chaque étape de l'amplification PCR et est compatible avec les méthodes d'amplification d'ADN isothermes, y compris RPA et potentiellement HDA (amplification dépendante de l'hélicase) (Vincent et al., 2004) et LAMP (amplification isotherme par boucle) (Notomi et al., 2000). Ce mode est particulièrement intéressant dans le cadre de ma thèse, puisqu'il me permettra d'étudier les taux de mutations microsatellites à partir d'expériences générées avec des méthodes autres que la PCR.

L'utilisation de l'algorithme sur les fréquences d'allèles microsatellites générés par PCR et RPA en utilisant des microsatellites mononucléotidiques (A/T) et dinucléotidiques (AC/TG) de 15, 19, 20 et 24 répétitions a permis de mieux comprendre le mécanisme et la dynamique du glissement de polymérase se produisant au niveau des séquences microsatellites au cours de ces deux méthodes de réplication *in vitro* de l'ADN (Figure 38). Les résultats des taux de mutation microsatellite avec les données de PCR ont montré une augmentation globale des taux avec une augmentation du nombre de répétitions, comme attendu, avec un biais systématique en faveur des délétions qui était beaucoup plus prononcé dans les microsatellites dinucléotidiques que dans les mononucléotidiques. E3d1u et L4d1u ont été identifiés comme les modèles génétiques de glissement de polymérase pour les microsatellites mono- et di-nucléotidiques dans notre configuration expérimentale de PCR, respectivement. Bien qu'il n'y ait pas de moyen simple de valider ces résultats, la régression des taux de mutation inférés en fonction du nombre de répétitions a montré une courbe croissante exponentiellement et linéairement pour les microsatellites mono- et dinucléotidiques, respectivement, ce qui était en accord avec le modèle génétique sous-jacent identifié par notre algorithme en ce qui concerne le type de relation liant les taux de mutation et le nombre de répétitions. Selon le type de microsatellites, les taux de mutation pourraient ainsi augmenter soit linéairement soit exponentiellement au cours de la PCR, ce qui n'a pas été mis en évidence dans une même étude. En effet, dans une étude précédente, la relation entre les taux de mutation et le petit nombre de répétitions des microsatellites mono- et dinucléotidiques ($A_{8\text{-}13}$ et $(CA)_{4\text{-}15}$) à partir des expériences de PCR a été identifiée comme linéaire (Shinde et al., 2003), tandis que d'autres études *in vitro* basées sur des expériences de *HSV-tk mutation assay* ont suggéré une relation exponentielle entre les taux d'erreur des microsatellites de la polymérase β et le nombre de répétitions (Baptiste et al., 2013; Eckert et al., 2002; Kelkar et al., 2010).

Nous avons également caractérisé pour la première fois les taux de mutation des microsatellites lors des expériences de RPA, qui ont montré des différences notables par rapport à la PCR (Figure 38). Ainsi, les taux globaux de mutation des microsatellites étaient plus faibles dans la RPA que dans la PCR, tandis que les taux de délétion étaient généralement légèrement supérieurs aux taux d'insertion. E2d1u et E3d2u ont été identifiés à partir des résultats de RMSE comme les meilleurs modèles génétiques sous-jacents pour les microsatellites mono- et di-nucléotidiques, respectivement, mais la relation exponentielle entre les taux de mutation et le nombre de répétitions n'a pas été entièrement validée par l'ajustement de courbes de régression. Ainsi, les résultats de l'ajustement de courbe ont montré que les taux de délétion et d'insertion des microsatellites dinucléotidiques augmentaient de manière exponentielle et linéaire, respectivement, indiquant que ces deux taux pourraient présenter une relation avec la longueur des répétitions qui est différente l'un de l'autre. De plus, avec les microsatellites mononucléotidiques, les taux d'insertion ont égalé les taux de délétion à 24 répétitions, et la tendance dans les deux droites de régression suggère que les taux d'insertion pourraient être supérieurs aux taux de délétion à un nombre de répétitions plus élevé.

La relation entre les taux de mutation des microsatellites et les unités de répétition pourrait être plus complexe que celle proposée par les modèles utilisés dans notre étude. Cela a également été suggéré par Wierdl *et al*., qui n'ont pas été en mesure d'ajuster leurs résultats de taux de mutation

obtenus chez la levure et la longueur des répétitions avec une simple fonction linéaire ou exponentielle (Wierdl et al., 1997). De plus, Xu et al. ont proposé des relations distinctes entre les taux de délétion ou d'insertion des microsatellites et la longueur des répétitions qui pourraient expliquer la distribution en forme de cloche des allèles microsatellites observée au niveau de la population, avec des taux de délétion augmentant de manière exponentielle avec la longueur des répétitions et des taux d'insertion restant constants (Xu et al., 2000). Nos modèles génétiques n'ont pas considéré de relations distinctes entre les taux d'insertion et de délétion et le nombre de répétitions, bien qu'elles aient pu être observées dans certains de nos résultats décrits ci-dessus à partir des analyses d'ajustement de courbe, et cela pourrait être une limitation de notre algorithme. Plus de modèles génétiques pourraient donc être inclus dans l'algorithme comme perspective de développement. Une autre perspective pourrait également être la capacité de notre code à gérer une combinaison de différents allèles microsatellites comme point de référence initial. Enfin, les résultats des données expérimentales ont mis en évidence plusieurs différences dans les taux de mutation des microsatellites et les mécanismes de glissement de polymérase entre la PCR et la RPA. Ces divergences pourraient être attribuées aux différences dans les polymérases utilisées, spécifiquement la Taq polymérase dans la PCR et la Bsu polymérase dans la RPA, ainsi qu'aux conditions expérimentales utilisées dans les deux méthodes, notamment la température de réaction, qui a montré diminuer drastiquement la formation d'artéfacts de glissement (Daunay et al., 2019).

En résumé, nous avons développé un algorithme qui permet l'inférence des taux de mutation des microsatellites à partir des fréquences d'allèles obtenues après amplification de l'ADN des séquences microsatellites en utilisant trois modes de calcul des taux de mutation. Il permet également l'identification du meilleur modèle génétique sous-jacent au glissement de polymérase parmi les 32 qui ont été modélisés, qui se distinguent par la relation entre les taux de mutation et la longueur des répétitions (linéaire ou exponentielle) et le nombre de pas autorisés par événement mutationnel (de 1 à 4). L'algorithme développé pourrait ainsi permettre une meilleure compréhension et caractérisation des mutations microsatellites se produisant *in vitro* à partir des méthodes d'amplification de l'ADN, ce qui pourrait potentiellement également fournir de nouvelles perspectives sur les mécanismes d'évolution de ces séquences *in vivo*. Il pourrait permettre plusieurs investigations dans l'étude fondamentale des microsatellites *in vitro*, et devrait être complémentaire à d'autres méthodes analysant les mutations microsatellites *in vitro* telles que l'essai de mutation directe *HSV-tk* (Baptiste et al., 2015; Eckert et al., 2002).

Ce projet a fait l'objet de l'écriture d'une publication en cours de finalisation à laquelle je suis associée en tant que 2ème auteur : Inferring microsatellite mutation rates from allele frequencies obtained from (non-)isothermal DNA amplification approaches using thirty-two genetic models. (Rojo Nofidiantsoa, Sophie I. Jeanjean, Nasreddine Rajoua, Antoine Daunay, Mourad Sahbatou, Jean-François Deleuze & Alexandre How-Kit.)

Ce projet a également fait l'objet d'une communication affichée à l'*European Society of Human Genetics Conference* (ESHG) en 2023 à Glasgow, ainsi qu'à l'*Internation Symposium on Human Genomics* (ISHG) en 2023 à Paris.

2. Projet 2 : Analyse détaillée des approches de séquençage de deuxième et troisième génération pour la détermination précise de la longueur des répétitions courtes en tandem et des homopolymères.

2.1. Résultats

2.1.1. Description des approches NGS utilisées dans notre étude comparative

La présente étude vise à évaluer et comparer la précision des approches de séquençage de deuxième et troisième génération pour identifier la longueur précise des microsatellites, incluant les microsatellites mononucléotidiques (A/T) et dinucléotidiques (AC/TG et AT/TA) avec 15 à 25 répétitions, insérés dans des plasmides, chacun contenant un code-barres spécifique à 4 chiffres (Tableau 16). Chaque préparation de plasmide était supposée pure, ne contenant que la séquence avec le nombre souhaité de répétitions, comme confirmé par le séquençage Sanger (Figure 39).

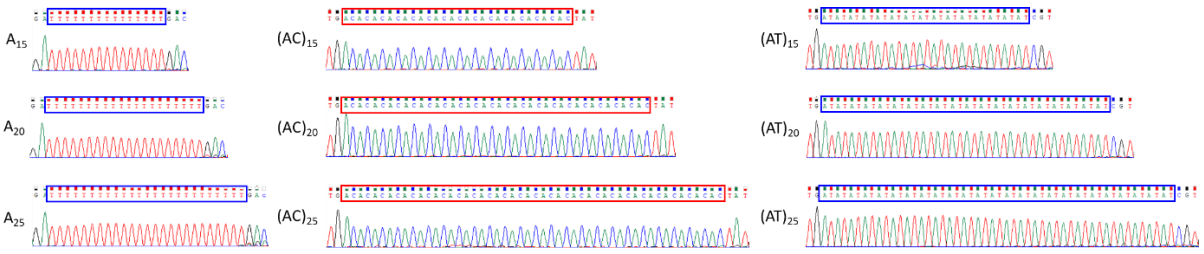

Figure 39 : Résultat du séquençage SANGER des 9 microsatellites clonés en plasmide pUC57-Mini pour le séquençage NGS.

La construction des inserts a été réalisée en suivant le modèle suivant (Tableau 16):

- Des sites de restrictions permettant la récupération de l'insert
- Un site de fixation d'amorce permettant une amplification par PCR
- Un index permettant la différentiation et le polymorphisme d'origine des microsatellites après amplification
- Le microsatellite
- Une zone flanquante afin d'obtenir une taille utilisable de l'insert

Je me suis concentrée sur les STRs mono- et di-nucléotidiques, les types de séquences microsatellites les plus courants dans le génome humain, et ai choisi un nombre de répétitions d'au moins 15, car il a été démontré que la précision de la détermination de la longueur des microsatellites à partir des données NGS devient de plus en plus difficile avec un nombre croissant de répétitions (Fungtammasan et al., 2015; Raz et al., 2019; Ross et al., 2013). Les STRs comportant plus de 25 répétitions n'ont pas été évalués dans l'étude afin de limiter l'influence des contraintes du séquençage *short-read* sur les analyses bioinformatiques, et pour permettre une comparaison avec le séquençage *long-read*. Pour les approches de séquençage *short-read* sur iSeq 100 et NextSeq 500 d'Illumina, des microsatellites extraits des plasmides par digestion par enzyme de restriction de type II (SSDP) ou des plasmides fragmentés aléatoirement (RFP) ont été utilisés comme échantillons d'ADN (Figure 40.A).

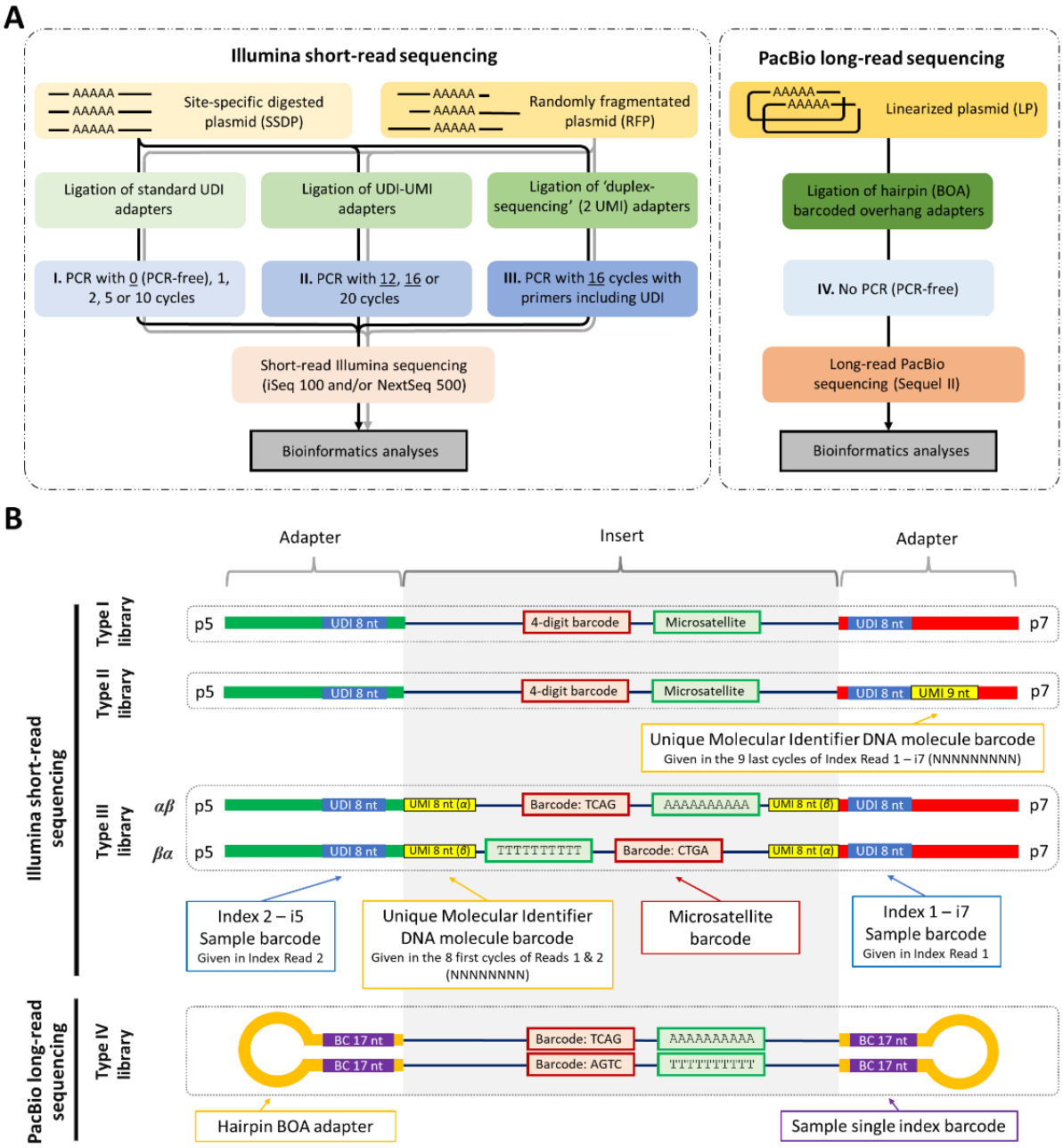

Figure 40 : Workflow des expériences NGS réalisées dans notre étude. A. Détail des différentes étapes effectuées dans notre étude pour la préparation des bibliothèques NGS pour short- et long-read, en utilisant des plasmides contenant des microsatellites comme échantillons d'ADN. Les conditions de PCR soulignées correspondent à celles appliquées à l'ADN fragmenté de manière aléatoire. B. Représentation schématique des différents composants moléculaires des quatre types de bibliothèques NGS utilisées dans notre étude. Les bibliothèques de séquençage à courtes lectures d'Illumina sont représentées en simple brin. En considérant la configuration double brin, les bibliothèques de type I et II incluaient des structures en 'Y' à chaque extrémité de l'ADN pour les bibliothèques sans PCR, en raison d'adaptateurs en 'Y' tronqués, tout en étant entièrement complémentaires dans les bibliothèques contenant de la PCR. Pour les bibliothèques de type III de «séquençage en duplex», les deux brins de la même molécule d'ADN double brin étaient marqués par des codes-barres à chaque extrémité de l'ADN, permettant leur identification — chaque brin présenterait une combinaison de codes-barres αβ ou βα dans les lectures 1 et 2 — et la correction des erreurs après la PCR et le séquençage.

Ces ADN ont été utilisés pour les préparations de librairies en utilisant des adaptateurs contenant 0, 1 ou 2 UMI (Unique Molecular Identifier) et incluant soit aucune PCR (sans PCR), soit de 1 à 20 cycles de PCR (Figure 40.A), aboutissant à trois types (I, II et III) de librairies (Figure 40.B). Pour le séquençage *long-read* sur PacBio Sequel II de Pacific Bioscience, un seul type de librairie sans PCR (Type IV) a été préparé à partir de plasmides linéarisés (≈2 kb) qui ont été directement ligaturés à des adaptateurs BOA standard pour la circularisation (Figure 40). Par la suite, plusieurs approches bioinformatiques, réalisées en collaboration avec Yimin Shen (Chercheur dans le laboratoire de bioinformatique du CEPH) ont été appliquées pour analyser les données de séquençage *short-* et *long-read*, permettant d'identifier la longueur des microsatellites (Figure 40).

2.1.2. Impact des cycles de PCR dans la préparation de librairies short-read Illumina sur l'estimation de la longueur des allèles microsatellites

La PCR est une technique d'amplification de l'ADN pratique et très répandue. Elle est utilisée à différentes étapes de la préparation des librairies *short-read* d'Illumina, que ce soit pour le séquençage de gènes ciblés (TGS, *Targeted Gene Sequencing*), le séquençage d'exome entier (WES, *Whole Exome Sequencing*) ou le séquençage de génome entier (WGS, *Whole Genome Sequencing*). *A contrario*, les protocoles sans PCR (*PCR-free*) sont moins polyvalents en raison de la plus grande quantité d'ADN requise et sont généralement limités au WES et WGS (Kozarewa et al., 2009; Yamaguchi et al., 2019). Afin de pouvoir évaluer l'impact de l'amplification par PCR sur la précision de la détermination de la longueur des microsatellites, nous avons préparé des librairies en réalisant la digestion SSDP et en appliquant de 0 à 20 cycles d'amplification par PCR avant le séquençage (Figure 40.A). La construction de l'insert d'ADN a été pensée de manière à faire débuter le séquençage des microsatellites dans le même cycle de séquençage, c'est-à-dire 34 et 36 cycles pour chaque répétition de microsatellites di- et mono-nucléotidiques, respectivement. Ainsi cela nous a permis de s'affranchir des biais dus à la position à laquelle le séquençage des microsatellites commence. Les taux d'erreur de séquençage, calculés à partir des données sans PCR, ont alors révélé une augmentation drastique des taux d'indels dans la séquence des microsatellites et des taux de substitution après la séquence des microsatellites (Figure 41). Ces tendances semblent également plus prononcées pour les plus longues répétitions (Figure 41). Les taux d'erreur de substitution élevés dans les séquences post-STR ont conduit à considérer la longueur des allèles de microsatellites en utilisant l'approche que nous avons développée (voir Matériels et Méthodes 5.1) plutôt que la séquence exacte obtenue des reads pour toutes les analyses ultérieures.

Pour chaque type de microsatellite, les résultats ont montré une plus grande proportion d'allèles authentiques dans les librairies sans PCR (90 % en moyenne, avec les microsatellites $(AT)_{20}$ et $(AT)_{25}$ présentant le pourcentage le plus bas d'allèles originaux, soit 89 % et 52 % respectivement). Cette proportion diminue ensuite avec l'augmentation du nombre de cycles de PCR (Figure 42). Une proportion plus faible de l'allèle original a également été observée pour les microsatellites avec un plus grand nombre de répétitions dans tous les types de librairies, tandis que l'effet de la PCR sur la réduction de l'allèle original affecte davantage les microsatellites A/T et AT/TA que AC/TG (Figure 42). De plus, le bruit de fond des *stutter* semble de plus en plus biaisé vers les délétions à mesure que le nombre de cycles de PCR augmente (Figure 42), comme attendu. Enfin, il est à noter que le bruit de fond des *stutter* dû au glissement de la polymérase est visible dès le premier cycle de PCR avec une diminution moyenne de 5 % (min=1,2 % avec $(AC)_{15}$ et max=8,9 % avec $(AT)_{25}$) de l'allèle original par rapport au protocole sans PCR (Figure 42).

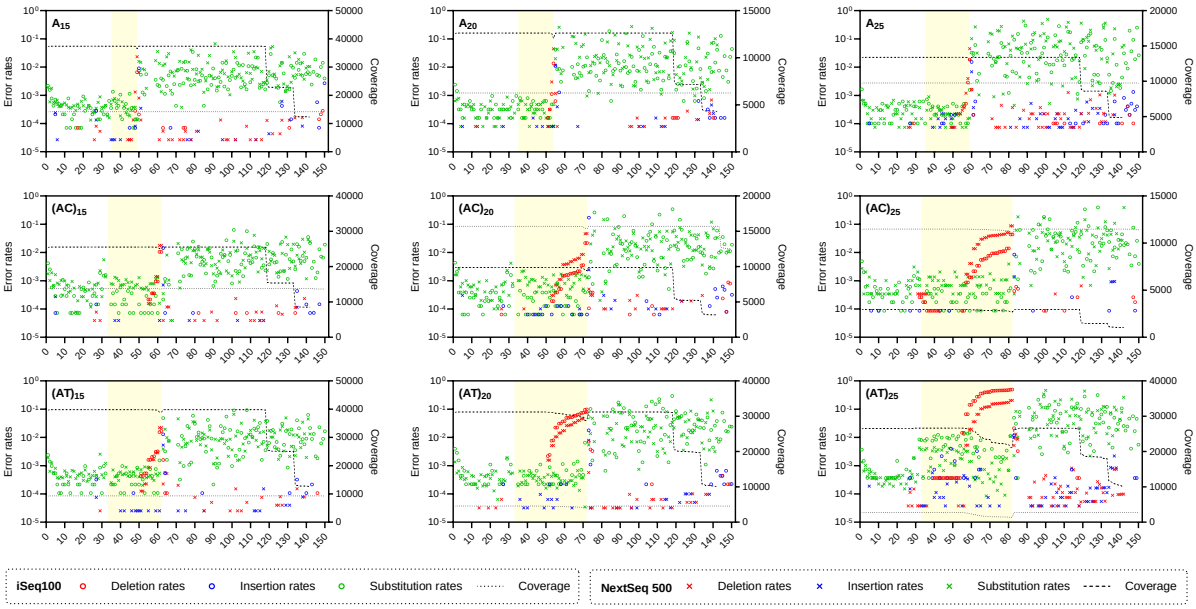

Figure 41 : Taux d'erreur de séquençage obtenus à partir des bibliothèques sans PCR de type I, en utilisant le séquençage short reads avec les instruments iSeq100 et NextSeq 500. L'axe des x correspond à la séquence provenant des lectures contenant des microsatellites, débutant au premier cycle de séquençage (34 et 32 nucléotides avant les microsatellites mono- et di-nucléotidiques, respectivement). Les séquences de microsatellites sont mises en évidence en jaune. Les données des répétitions d'expérience ont été regroupées selon le séquenceur utilisé. Pour le calcul des taux d'erreur, les reads brutes ont d'abord été alignées sur chaque séquence de référence. Ensuite, les couvertures (Cov) ont été calculées par le nombre total de A, T, G, C et N par position. Les taux de substitutions (Sub) ont été estimés par la formule Sub/Cov. Les taux de délétion (Del) et d'insertions (Ins) ont été définis respectivement par les formules Del/(Cov+Del+Ins) et Ins/(Cov+Del+Ins). L'identification des reads contenant des microsatellites a été effectuée à l'aide d'un script développé en interne, et le mappage a été réalisé avec Bowtie2 (v2.3.0). Les délétions, insertions et substitutions ont été comptées à l'aide d'IGVtools (v2.17.3). Les reads issues des données iSeq 100 ont une longueur de 151 pb, tandis que celles des données NextSeq 500 présentent trois longueurs différentes : 118 pb, 133 pb et 142 pb.

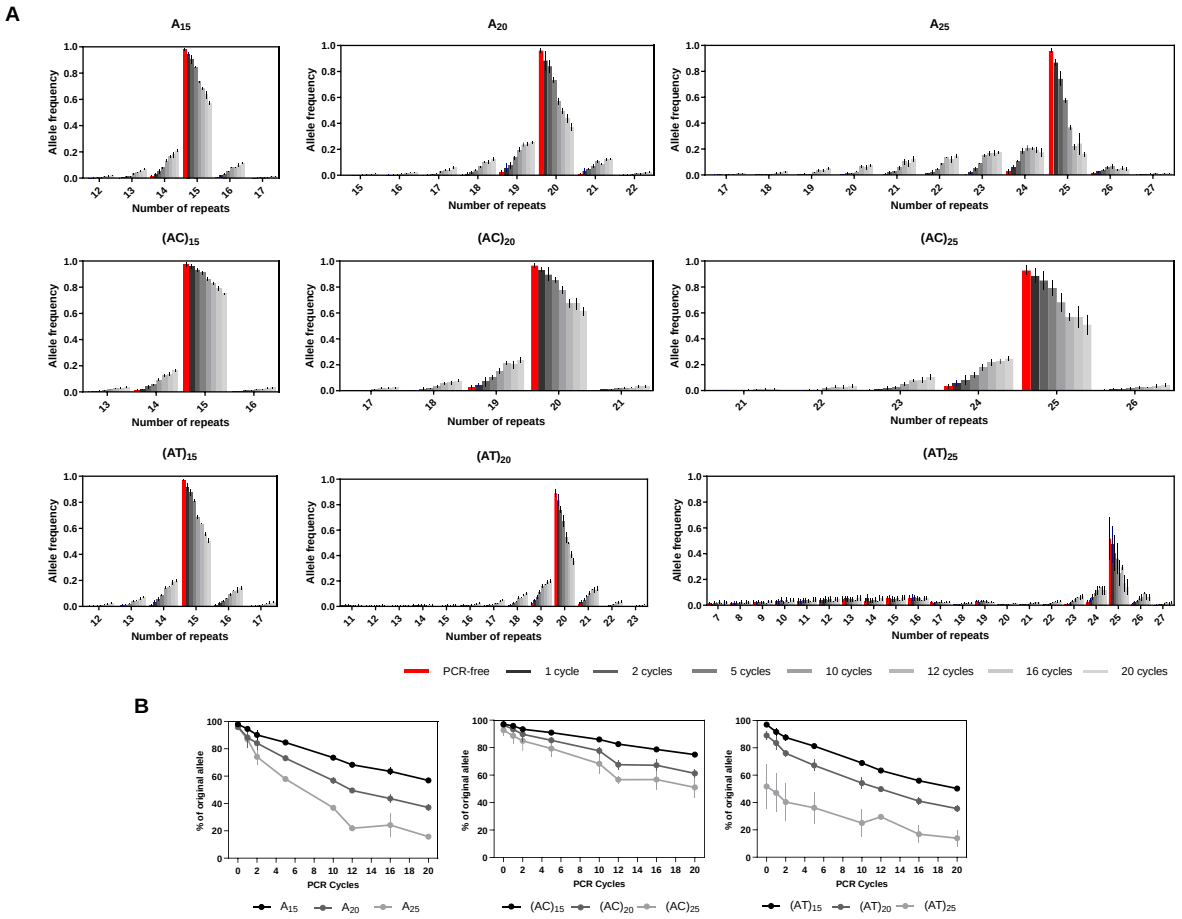

Figure 42 : Effet des cycles de PCR inclus dans la préparation des bibliothèques sur la longueur des allèles de microsatellites obtenue après séquençage en short-read Illumina utilisant la digestion SSDP. A. Fréquences des allèles de microsatellites obtenues avec des bibliothèques sans PCR et contenant de la PCR (de 1 à 20 cycles) et le séquençage en short-read Illumina utilisant la digestion SSDP comme modèles. B. Fréquences des allèles de microsatellites obtenues avec des bibliothèques sans PCR et contenant de la PCR (de 1 à 20 cycles) et le séquençage en short-read Illumina utilisant la digestion SSDP comme modèles. La figure présente uniquement les données des bibliothèques de type I et II, et chaque point inclut au moins cinq réplicats. Les données ont été générées sur les instruments Illumina iSeq100 et NextSeq 500.

En considérant les deux instruments de séquençage utilisés, l'iSeq100 et NextSeq 500, les résultats sont comparables, sauf pour $(AC)_{25}$ et $(AT)_{25}$ qui montrent un pourcentage plus élevé des allèles originaux avec l'iSeq100 et le NextSeq 500, respectivement (Figure 43). Enfin, la comparaison des données des librairies sans PCR réalisées par SSDP et RFP a montré de moins bons résultats pour celles réalisées par RFP, notamment avec les microsatellites les plus longs (Figure 44). Cela est probablement dû à l'augmentation du cycle de séquençage à partir duquel les microsatellites sont séquencés pour une grande proportion des reads, réduisant ainsi leur qualité. Dans l'ensemble, ces résultats indiquent que les protocoles sans PCR surpassent les protocoles contenant de la PCR, lorsqu'il s'agit de préserver la distribution de fréquence des allèles de microsatellites d'un échantillon, permettant ainsi de mesurer la longueur des microsatellites avec la plus grande précision.

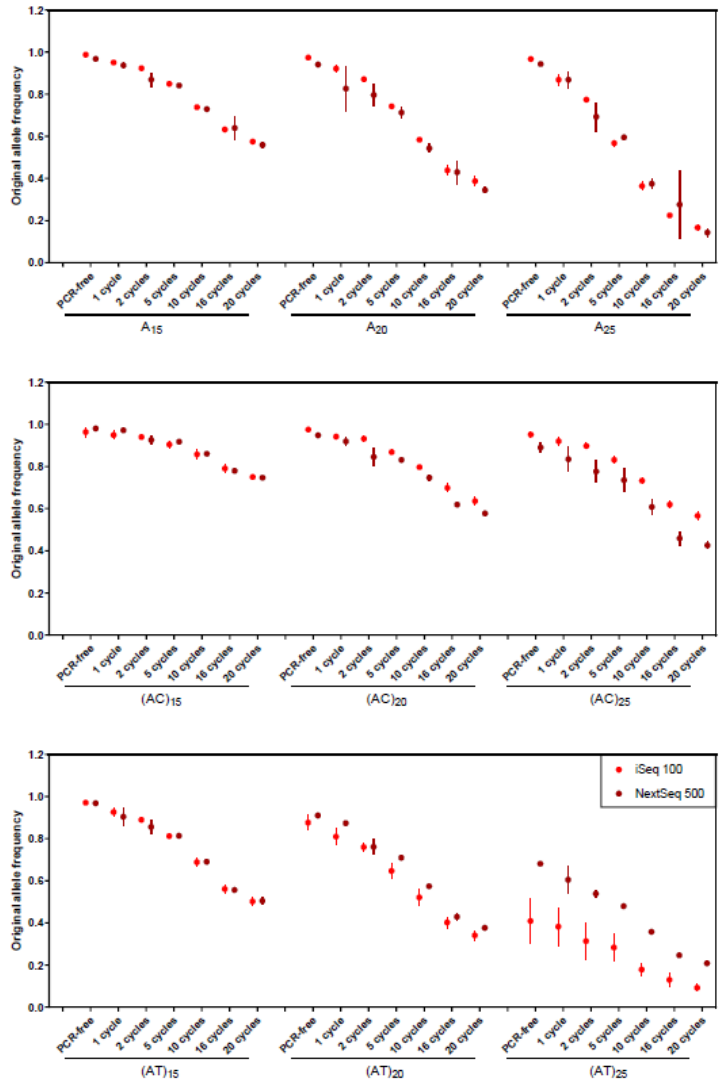

Figure 43 : Comparaison des performances des instruments Illumina iSeq 100 et NextSeq 500 pour la détermination précise de la longueur des neuf microsatellites étudiés en utilisant la digestion SSDP comme modèle. La figure présente uniquement les données des bibliothèques de type I et II, et chaque point inclut au moins cinq réplicats.

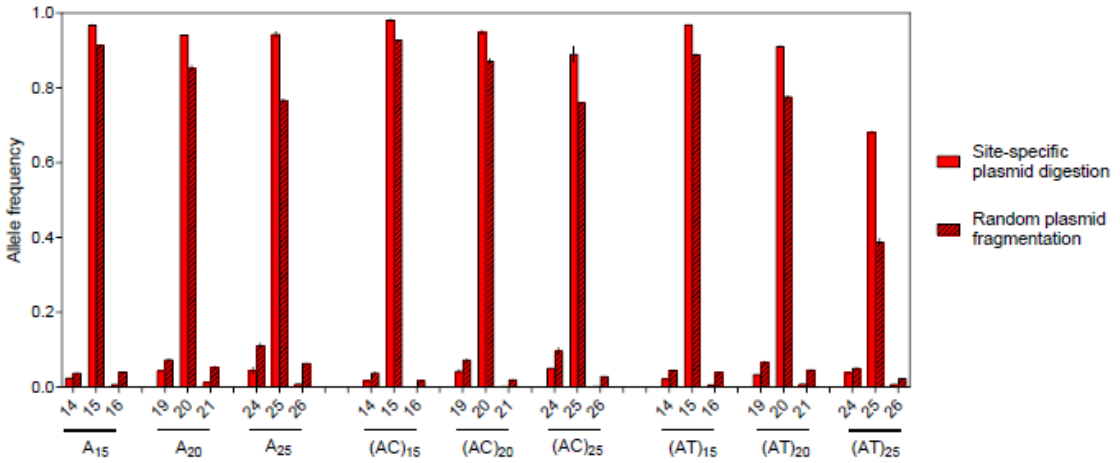

Figure 44 : Comparaison des performances de la détermination précise de la longueur des neuf microsatellites étudiés à partir de bibliothèques sans PCR de type I et du séquençage short-read Illumina NextSeq 500, en utilisant les digestions SSDP et RFP comme modèles. La figure comprend deux réplicats pour les modèles SSDP et trois réplicats pour les modèles RFP.

2.1.3. Détermination de la longueur des allèles des microsatellites par la correction d'erreur à partir d'UMI simple

Les identifiants moléculaires uniques (UMI, *Unique Molecular Identifiers*) peuvent être utilisés dans des protocoles à un ou deux UMI pour marquer les matrices d'ADN avant l'amplification de la librairie et le séquençage NGS (Figure 40). L'identification de plusieurs reads portant le même UMI (c'est-à-dire des duplicats de séquençage) permet ainsi d'identifier la molécule d'ADN unique à l'origine de ces reads et d'obtenir une séquence consensuelle qui réduit les erreurs de PCR et de séquençage (Kinde et al., 2011; Schmitt et al., 2012; Sloan et al., 2018).

Une stratégie utilisant un UMI unique de neuf nucléotides situé dans la séquence de l'adaptateur (librairie de type II) qui marquait les molécules d'ADN simple brin a d'abord été évaluée pour l'estimation correcte de la longueur des microsatellites (Figure 40). Pour l'identification des groupes de reads partageant le même UMI, aucune mutation dans la séquence de l'UMI n'a été autorisée pour réduire le risque d'agréger plusieurs reads provenant de plus d'une molécule d'ADN originale. De plus, la spécificité de l'identification de la molécule d'ADN originale ne pouvait pas être améliorée par les coordonnées du read de séquençage pour les données basées sur le protocole SSDP. Au moins 3 reads par groupe d'UMI ont été considérés pour la correction d'erreur et utilisés pour le développement de quatre modes de calcul de la longueur des microsatellites, nommés '*Max. Freq.*' (fréquence maximale), '*Max Length*' (longueur maximale), '*Mean Length*' (longueur moyenne) et '*Median Length*' (longueur médiane) (Figure 45.A). Parmi eux, le mode '*Max. Freq.*' s'est révélé le plus performant pour identifier la longueur de l'allèle original dans la plupart des conditions, et a donc été sélectionné pour des analyses plus détaillées (Figure 45.B).

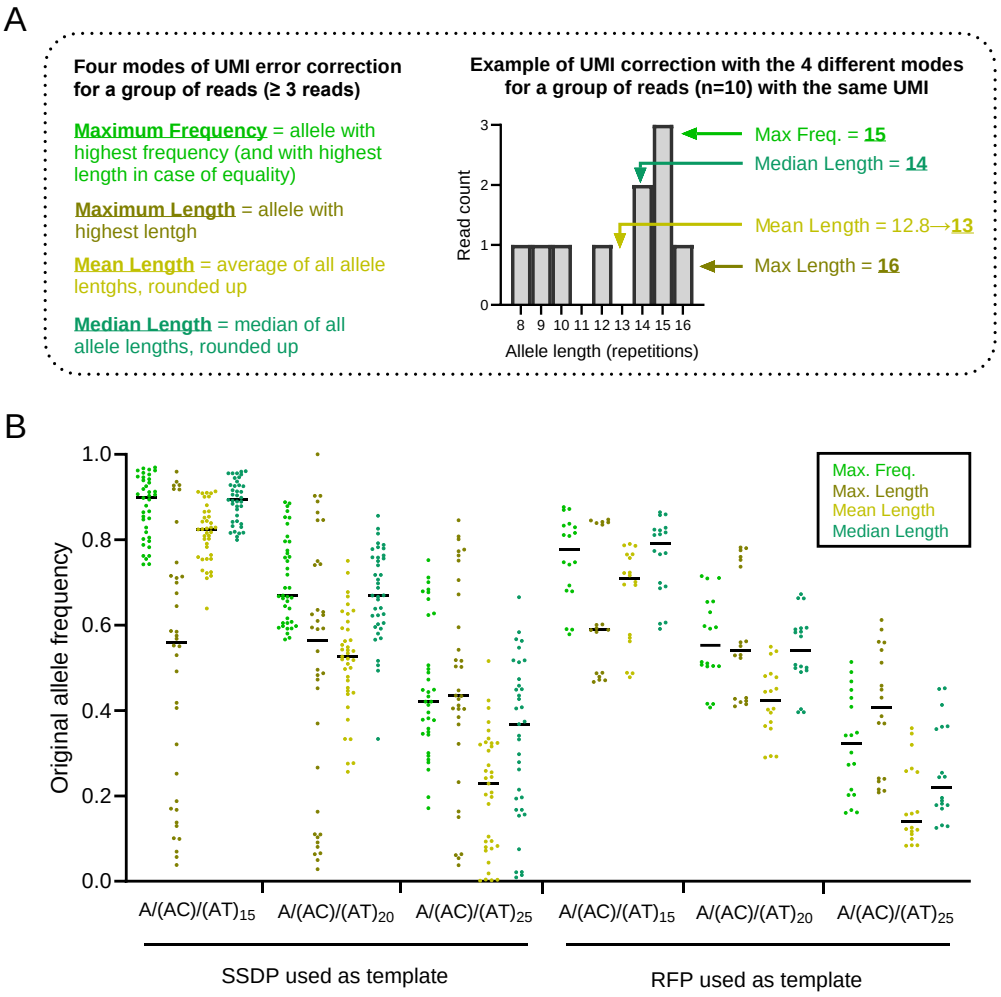

Figure 45 : Comparaison des différentes stratégies utilisées pour la correction d'erreur par UMI afin de déterminer avec précision la longueur des microsatellites. A. Description des quatre modes testés de correction d'erreur par UMI et exemple d'application de ces quatre modes. B. Fréquences des allèles originaux obtenues sur les données de la bibliothèque de type II après correction d'erreur par UMI. SSDP : plasmide digéré de manière site-spécifique ; RFP : plasmide fragmenté aléatoirement. Les données ont été générées sur un Illumina NextSeq 500.

Les résultats ont montré que la correction d'erreur par UMI augmentait la fraction de l'allèle original pour chaque type de séquence répétée analysée dans les conditions SSDP et RFP, bien que n'égalant pas les niveaux observés avec les librairies sans PCR (Figure 46). La réduction des taux d'erreur de longueur tendait à augmenter avec le nombre de cycles, et elle était généralement modeste (pas plus de deux fois) pour les microsatellites avec au moins 20 répétitions (Figure 46.A). Cependant, une réduction de presque 8 fois de l'erreur de longueur a été obtenue pour ceux avec 15 répétitions (Figure 46.A), suggérant que la correction d'erreur par UMI est plus efficace pour les STRs plus courts. L'augmentation du nombre minimal de reads par groupe d'UMI (de 3 à 7) a aussi été évaluée afin d'améliorer la précision de l'identification des allèles. De légères améliorations dans l'estimation de la longueur ont été observées en augmentant ce seuil pour les microsatellites avec 15 et 20 répétitions, mais pas pour ceux avec 25 répétitions (Figure 47). Ainsi, bien que les résultats obtenus pour les microsatellites avec 15 répétitions soient proches, la correction d'erreur UMI n'a pas présenté de niveaux plus élevés d'allèles originaux par rapport aux données sans PCR dans les conditions expérimentales

testées (Figure 46), indiquant que cette dernière approche est meilleure pour préserver et identifier la longueur originale des microsatellites avec 15 à 25 répétitions.

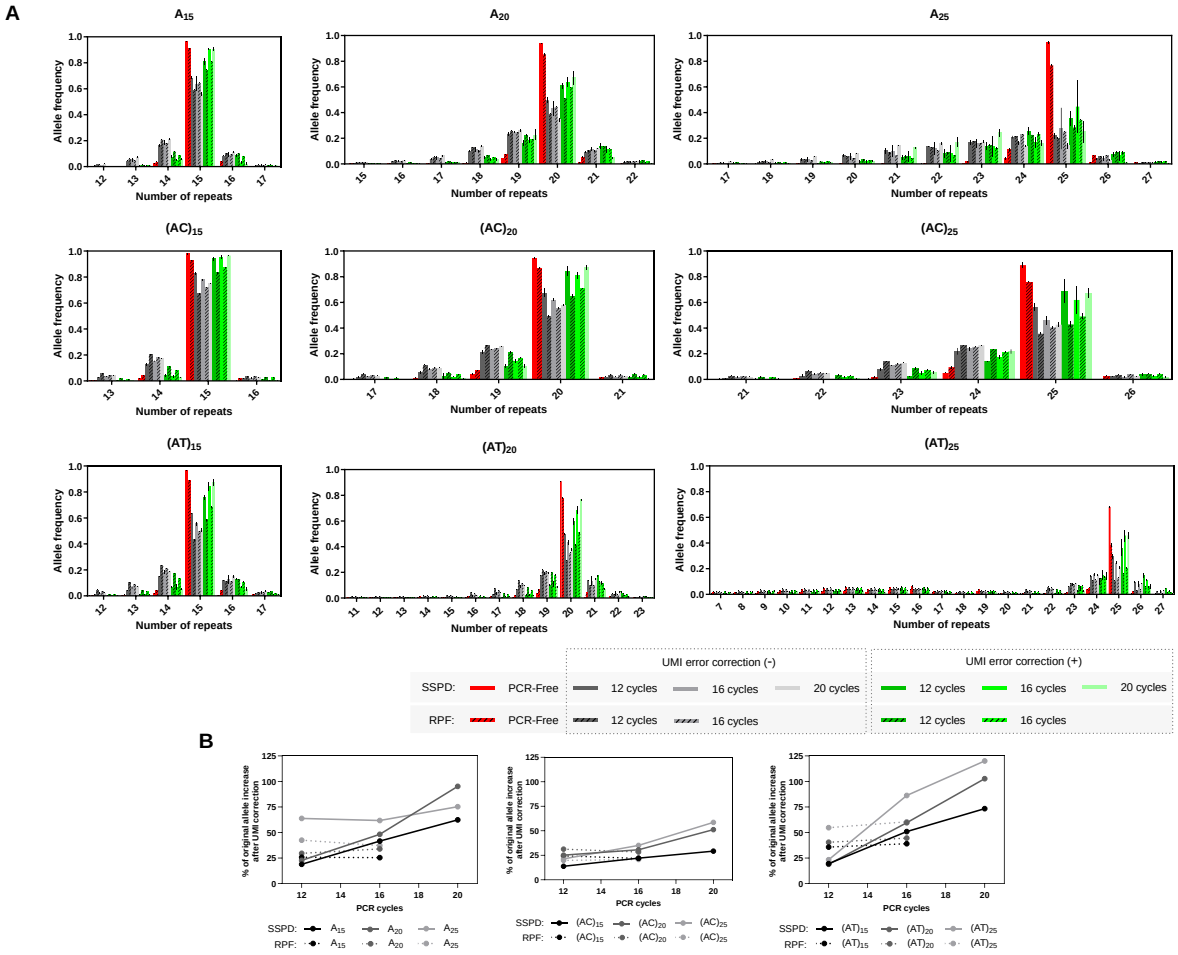

Figure 46 : Impact de la correction d'erreur par UMI sur les données de séquençage en short-read Illumina pour une détermination précise de la longueur des microsatellites. A. Fréquences des allèles de microsatellites obtenues avant et après correction d'erreur par UMI (mode Max. Freq.) à partir des bibliothèques de type II incluant 12, 16 et 20 cycles de PCR et utilisant SSDP et RFP comme modèles. B. Fréquences des allèles microsatellites obtenues avant et après la correction d'erreur par UMI (mode Max. Freq.) à partir des bibliothèques de type II (12, 16 et 20 cycles de PCR) utilisant SSDP et RFP comme modèles. Les données sans PCR provenant du séquençage en short-read Illumina sont également présentées dans les graphiques pour comparaison. Chaque point comprenait au moins des données expérimentales en triplicat, à l'exception des conditions sans PCR - SSDP et $(AC)_{25}$ - 12 cycles - SSDP - correction d'erreur par UMI (dupliqués). Toutes les données ont été générées sur un Illumina NextSeq 500.

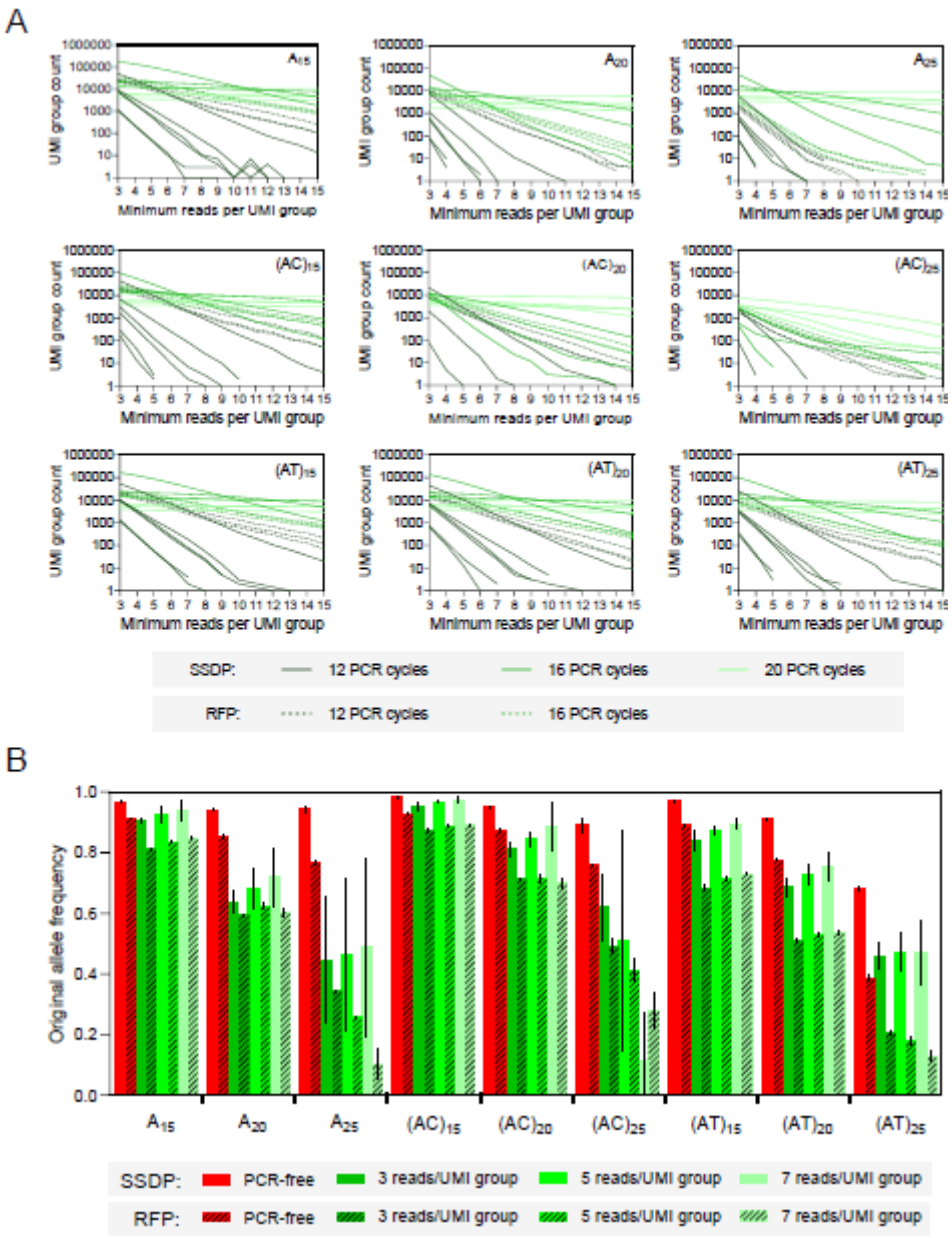

Figure 47 : Influence du nombre de reads par groupe d'UMI sur la précision de l'identification de la longueur des allèles microsatellites après correction d'erreur par UMI. A. Nombre de groupes d'UMI avec un nombre minimum de reads allant de 3 à 15 reads. B. Fréquence des allèles originaux obtenue après correction d'erreur par UMI en incluant différents seuils de reads (3, 5 et 7) par groupe d'UMI, en utilisant des données NGS provenant de bibliothèques de type II contenant 16 cycles de PCR. Les données sans PCR de type I provenant du séquençage en short-read Illumina sont également présentées dans le panneau B à titre de comparaison. SSDP : plasmide digéré de manière spécifique au site ; RFP : plasmide digéré de manière aléatoire. Toutes les données proviennent d'expériences en triplicat (à l'exception de la condition sans PCR - SSDP, qui provient d'expériences en duplicat). Les données ont été générées sur un Illumina NextSeq 500.

2.1.4. Amélioration de l'estimation de la longueur des allèles des microsatellites par correction d'erreur combinée à partir des données de séquençage en duplex

J'ai ensuite évalué la capacité du séquençage en duplex dit «*duplex sequencing*» à améliorer la détermination de la longueur des microsatellites. Le séquençage en duplex repose sur la présence de deux UMI de part et d'autre de chaque brin d'une molécule d'ADN bicaténaire (librairie de type III, Figure 40), ce qui permet leur identification après le séquençage NGS. Cela permet la construction d'une séquence consensus combinée à partir de deux familles de lectures (*αβ* et *βα*) qui représente chaque brin de l'ADN double brin, ce qui peut réduire les erreurs de PCR et de séquençage d'un facteur 10 supplémentaire par rapport aux méthodes à UMI unique, en particulier pour les mutations ponctuelles (Schmitt et al., 2012).

Dans un premier temps, la correction d'erreur par UMI standard a été effectuée en utilisant les mêmes critères que pour les librairies de type II, c'est-à-dire aucune mutation autorisée dans les codes-barres UMI et au moins 3 reads par groupe d'UMI, suivie d'une seconde étape de correction d'erreur combinée utilisant les tailles consensus *αβ* et *βα*. Les longueurs consensus des microsatellites générés à partir de la correction d'erreur par UMI standard étaient basées sur les quatre modes décrits précédemment (Figure 45.A), en considérant soit les lectures *αβ*, *βα* ou *αβ+βα* (pour *αβ+βα*, une taille consensus a été déterminée à partir d'au moins 3 lectures partageant les mêmes codes-barres *αβ* ou *βα*). Le mode '*Max. Freq.*' a de nouveau été sélectionné car il présentait un pourcentage plus élevé des allèles originaux dans la plupart des conditions (Figure 48).

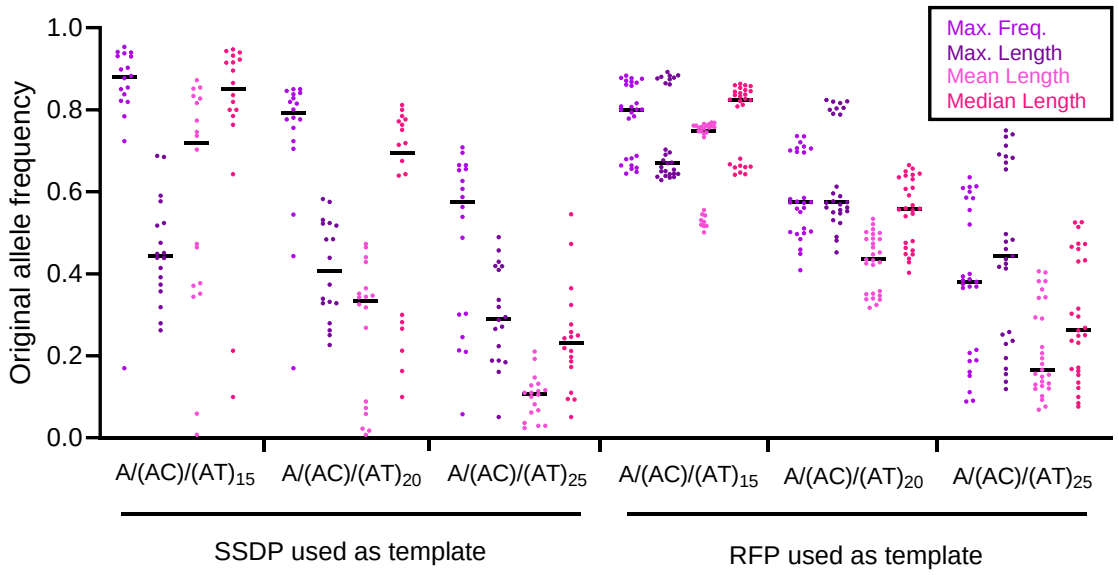

Figure 48 : Fréquences des allèles originaux obtenues sur les données de la bibliothèque de type III après correction d'erreur par UMI standard incluant les readss consensuelles αβ+βα, αβ et βα. SSDP : plasmide digéré de manière site-spécifique ; RFP : plasmide fragmenté aléatoirement. Les données ont été générées sur un Illumina NextSeq 500.

Comme pour les librairies de type II, les données non corrigées des librairies de type III ont présenté un pourcentage plus bas d'allèles originaux par rapport aux expériences sans PCR, en raison des artefacts de stutter introduits pendant la PCR qui étaient biaisés vers les suppressions (Figure 49.A). Après correction d'erreur par UMI standard, les proportions d'allèles originaux ont augmenté pour tous les microsatellites avec des valeurs comparables entre les tailles consensus *αβ*, *βα* et *αβ+βα*, n'indiquant

aucune différence majeure dans la correction d'erreur en fonction du type de famille de reads (Figure 49). En termes de réduction du taux d'erreur, la correction d'erreur par UMI standard était plus efficace pour les petites répétitions que pour les grandes répétitions, et dans les conditions de digestion SSDP par rapport aux conditions RFP (Figure 49.A). Cependant, le pourcentage d'allèles originaux après correction d'erreur n'atteignait pas les niveaux observés dans les expériences sans PCR (Figure 49.A).

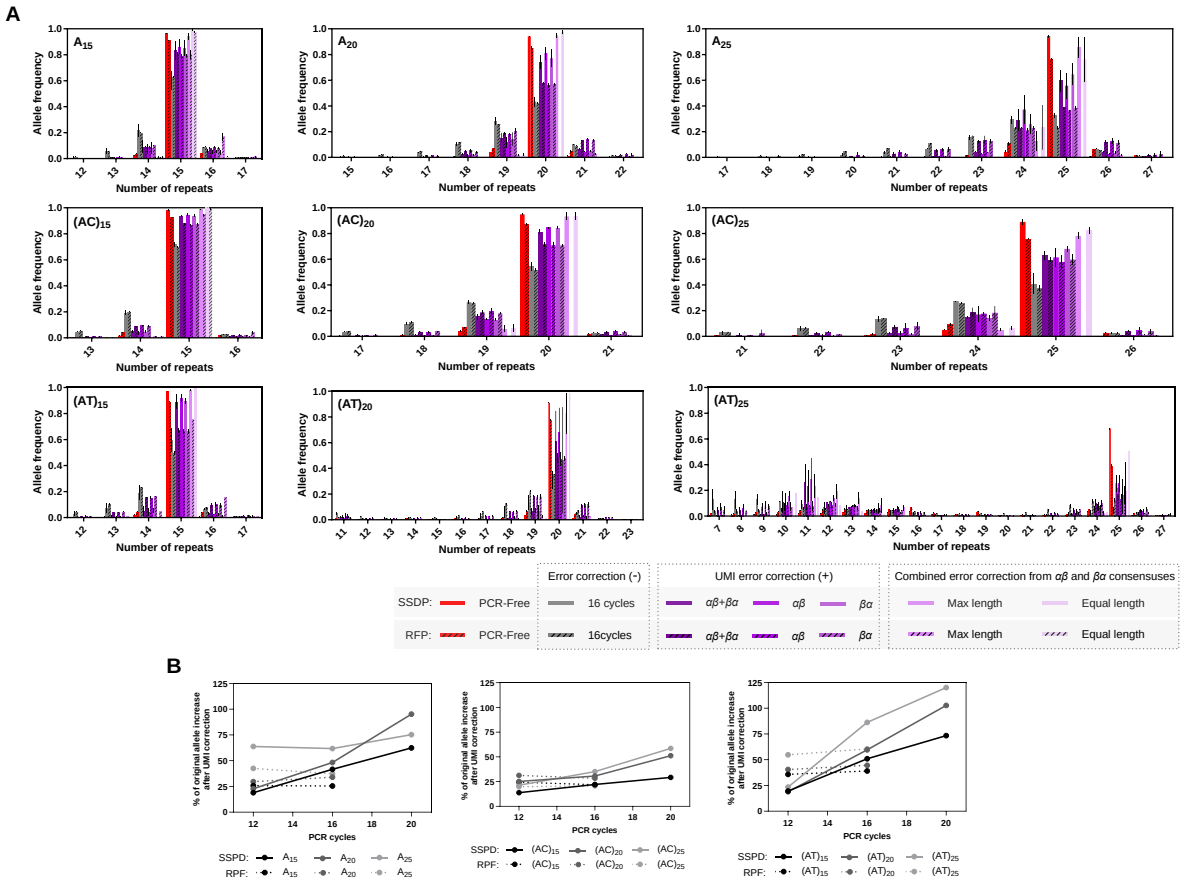

Figure 49 : Impact de la correction d'erreur dual-UMI du séquençage en duplex sur les données de séquençage short-read Illumina pour une détermination précise de la longueur des microsatellites. A. Fréquences des allèles microsatellites obtenues avant et après différents types de corrections d'erreur basées sur les UMI à partir des bibliothèques de type III incluant 16 cycles de PCR et utilisant SSDP et RFP comme modèles. B. Fréquences des allèles microsatellites obtenues avec des bibliothèques de type III (16 cycles de PCR) utilisant SSDP et RFP comme modèles, avant et après différentes corrections d'erreur basées sur les UMI. Les corrections d'erreur incluaient la correction d'erreur par UMI standard (mode Max. Freq.) obtenue soit à partir des reads αβ, βα, ou (αβ+βα), et la correction d'erreur combinée obtenue à partir des séquences consensuelles αβ et βα, soit basée sur la longueur maximale (le plus grand allèle est conservé comme consensus) ou sur la longueur égale (le consensus est atteint lorsque les allèles consensuels αβ et βα sont de même longueur). Chaque point provient de données expérimentales en duplicat (SSDP) ou en triplicat (RFP) ; cependant, pour la correction d'erreur combinée, certains points de données ont été partiellement ou totalement perdus (pas de barres dans les allèles n originaux). Toutes les données ont été générées sur un Illumina NextSeq 500.

La performance du séquençage en duplex a été évaluée également par la correction d'erreur combinée. Deux modes ont été testés, en considérant soit la longueur maximale, soit la longueur égale entre les allèles consensus *αβ* et *βα* de la même famille de lecture. Ainsi, le plus grand allèle est retenu comme consensus combiné avec le mode '*Max. Length*', tandis qu'un consensus combiné est obtenu

uniquement lorsque les deux allèles sont de même longueur avec le mode '*Equal Length*'. Pour certains microsatellites, les tailles consensus combinés n'ont pas pu être générées en raison du nombre réduit d'allèles consensus *αβ* et *βα* et de l'incapacité de les attribuer à la même famille. C'était le cas dans les expériences RFP de manière plus importante que dans les expériences SSDP et surtout pour les microsatellites avec un nombre croissant de répétitions (Figure 50).

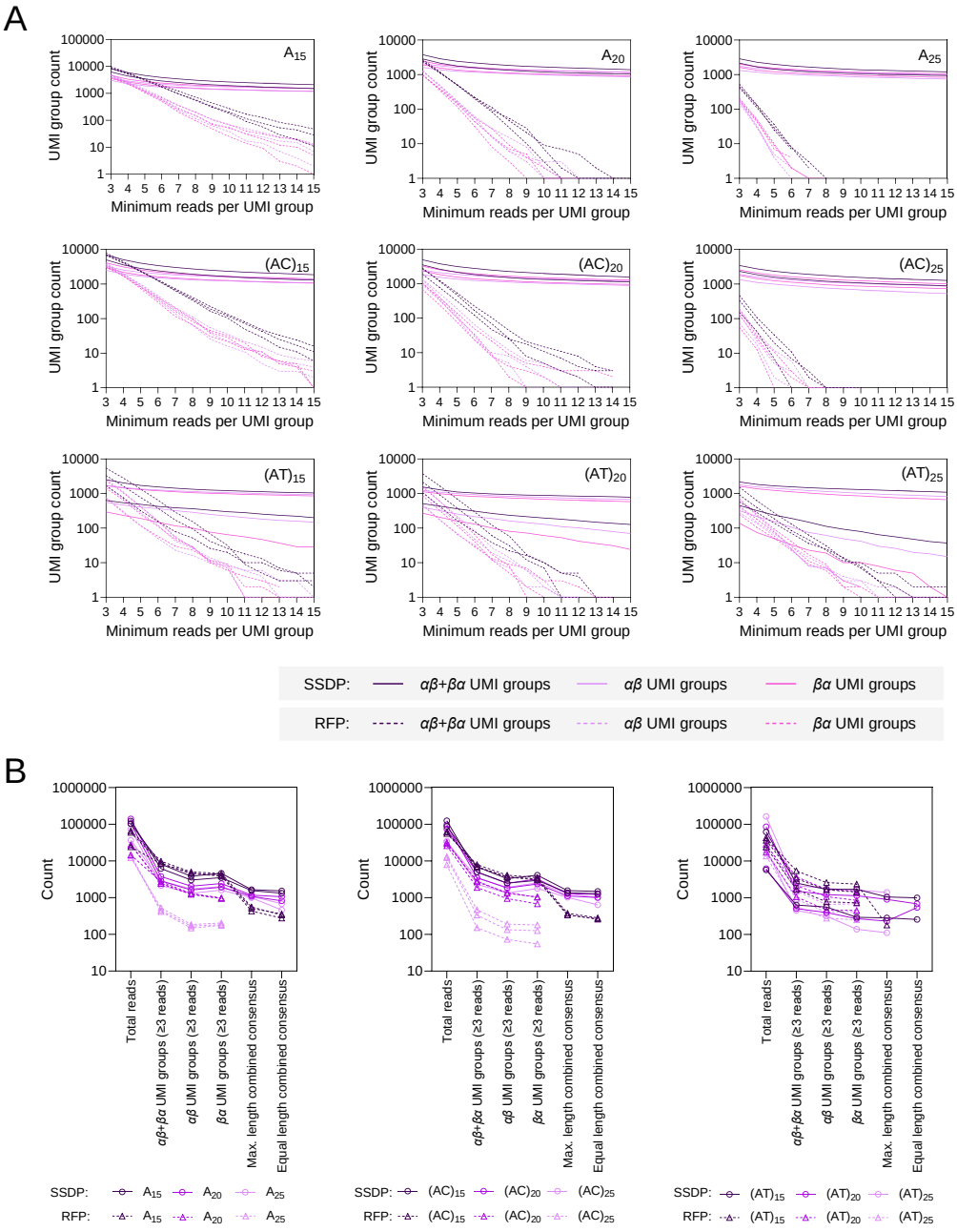

Figure 50 : Comptages des reads, groupes UMI, et consensus combinés obtenus par séquençage en short-read Illumina des bibliothèques de type III. A. Nombre de groupes UMI contenant un nombre minimum de reads allant de 3 à 15 reads. B. Comptages totaux des reads et des groupes de consensus pour chaque microsatellite, selon le type d'erreur standard UMI (αβ, βα et αβ+βα) et les modes de correction d'erreurs combinées (longueur maximale et longueur égale entre les allèles consensus αβ et βα). Les données sans PCR de type I issues du séquençage short reads Illumina sont également présentées dans le panneau B pour comparaison. SSDP : Plasmide digéré de manière spécifique au site. RFP : Plasmide digéré aléatoirement. Les données ont été générées à partir de duplicats pour SSDP et de triplicats pour RFP dans des expériences menées sur NextSeq 500 d'Illumina.

De plus, comme plusieurs familles de reads ont présenté des consensus *αβ* et *βα* discordants, l'approche '*Equal Length*' a permis moins fréquemment que l'approche '*Max. Length*' d'obtenir une taille consensus combinée. Après la correction des erreurs par l'approche combinée, le pourcentage d'allèles originaux a fortement augmenté pour chaque condition où cette correction était possible, de manière plus importante avec le mode '*Equal Length*' qu'avec le mode '*Max. Length*'. Cette correction a même permis d'atteindre 100% de l'allèle originel dans un réplicat SSDP pour $(AC)_{15}$ parmi 1496 séquences consensus combinées (Figure 49). Cette augmentation a donc surpassé la correction d'erreur standard par UMI, et parfois aussi les résultats « *PCR-free* » pour les microsatellites les plus petits (A_{15}, A_{20} et $(AC)_{15}$ dans les conditions SSDP et RFP et $(AT)_{15}$ dans la condition SSDP), réduisant le taux d'erreur de taille des allèles de moins d'une dizaine à plusieurs centaines de fois selon le type et le nombre de répétitions. Ainsi, le séquençage en duplex pourrait augmenter la précision de la mesure des tailles de microsatellites, mais il nécessitait un grand nombre de reads pour construire des séquences consensus combinées.

2.1.5. Précision et amélioration de l'identification de la longueur des allèles des microsatellites par séquençage en long-read PacBio

Le séquençage de troisième génération se caractérise par sa capacité à générer des reads très longs par rapport au séquençage de deuxième génération. Le séquençage en temps réel sur une seule molécule (SMRT, *Single-Molecule Real-Time*) développé par Pacific Biosciences est l'une des technologies de séquençage de troisième génération les plus novatrices et largement utilisées, basée sur des ZMW (*Zero-Mode Waveguide*) et le séquençage par synthèse sur une molécule unique (Iizuka et al., 2022; Logsdon et al., 2020). Pour évaluer la capacité de cette technologie à estimer correctement la longueur des microsatellites, les plasmides ont été linéarisés pour construire des librairies sans étape de PCR (librairie de type IV, Figure 40), puis séquencés directement sur une cellule SMRT 8M en utilisant un Sequel II.

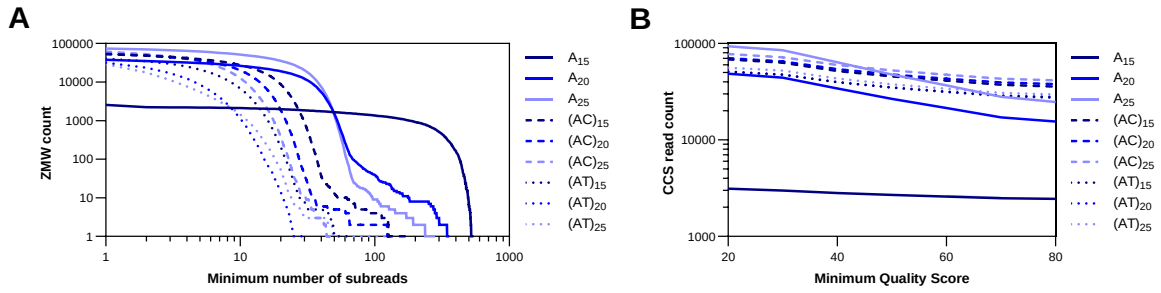

Figure 51 : Analyse 'brute' des données à partir du séquençage long-Read PacBio (Bibliothèques de Type IV) en utilisant différentes stratégies. A. Nombre de Zéro-Mode Waveguide (ZMW) par nombre minimal de sub-reads. B. Nombre de reads CCS par score de qualité minimal. Chaque point correspond à une seule expérimentation. Les données Type I PCR-free – SSDP (deux réplicats) issues du séquençage short-read Illumina (NextSeq 500) sont également présentées dans le panneau B pour comparaison.

L'analyse des reads bruts a indiqué que chaque plasmide avait plus de 1000 ZMW avec au moins 10 sub-reads exploitables avec un « Score de Code-barres » d'au moins 70, pouvant même atteindre plus de 100 sub-reads selon le type de répétition (Figure 51.A). Cela se reflétait dans la qualité des reads CCS (*Consensus Circular Sequencing*) qui montraient une proportion importante du score de qualité maximal, c'est-à-dire Q80 avec l'analyse SMRT Link (Figure 51.B). Il n'y avait pas de variations dans les taux de substitution avant et après la séquence microsatellite contrairement au séquençage en short-

read, ce qui est attendu car le sens du séquençage s'inverse après chaque sous-lecture avec le séquençage *long-read* PacBio (Figure 52). Cependant, des taux d'indels élevés ont été observés tout au long de la séquence avec cette chimie de séquençage, comme attendu, ainsi qu'une légère augmentation du taux de substitution uniquement dans les répétitions AT.

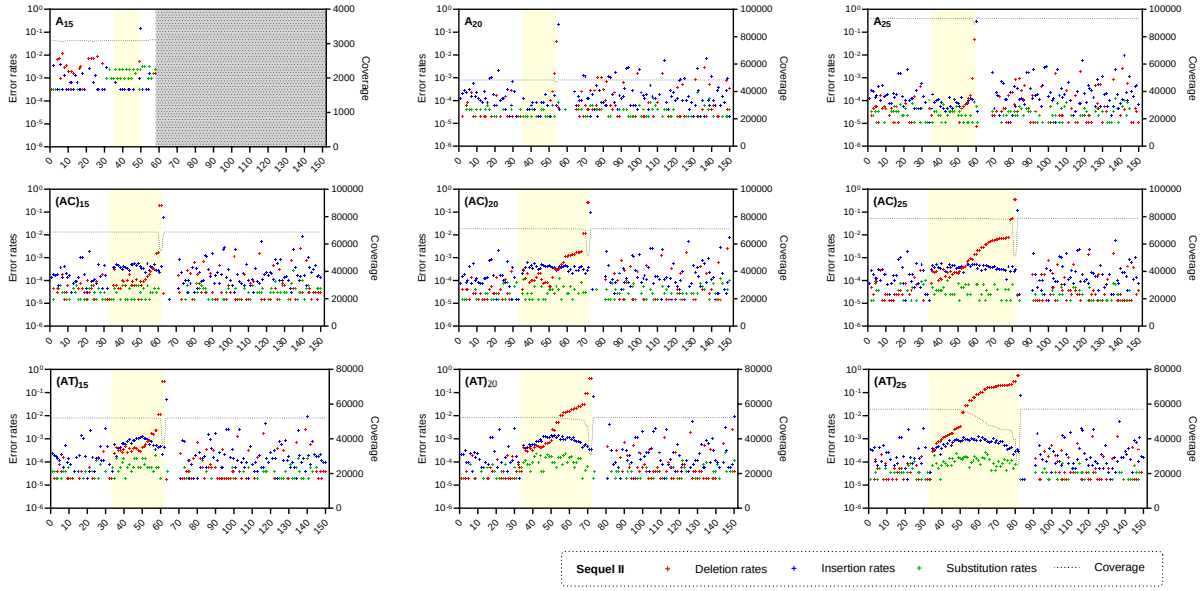

Figure 52 : Taux d'erreur de séquençage obtenus à partir de bibliothèques sans PCR de type IV, en utilisant le séquençage long-read avec le Sequel II. L'axe des x correspond aux mêmes séquences que celles présentées dans la figure 39, obtenues à partir de lectures CCS (Circular Consensus Sequencing) contenant des microsatellites. Les séquences de microsatellites sont mises en évidence en jaune. Une partie de la séquence a été manquante pour le plasmide A_{15} en raison de son code-barres spécifique, qui génère un second site PciI. Pour le calcul des taux d'erreur, les reads brutes ont été alignées sur chaque séquence de référence. La couverture (Cov) a été calculée comme le nombre total de A, T, G, C et N par position. Les taux de substitutions (Sub) ont été estimés par la formule Sub/Cov. Les taux de délétions (Del) et d'insertions (Ins) ont été définis respectivement par Del/(Cov+Del+Ins) et Ins/(Cov+Del+Ins). L'identification des fragments contenant des microsatellites a été réalisée à l'aide d'un script développé en interne. Les délétions, insertions et substitutions ont été comptées à l'aide d'IGVtools (v2.17.3).

Pour comparer la précision des tailles de microsatellites déterminées par le séquençage PacBio, nous avons d'abord utilisé différents seuils de scores de qualité des *reads* HiFi CCS allant de Q20 à Q80, c'est-à-dire une probabilité d'erreur de séquençage allant de 10^{-2} à 10^{-8} dans chaque *read* CCS. Comparés aux résultats *PCR-free* obtenus avec Illumina, les résultats des *reads* CCS présentaient toujours des pourcentages d'allèles originaux plus faibles, même pour les *reads* de la plus haute qualité, et ce pourcentage diminuait avec un nombre croissant de répétitions (Figure 53). Les tailles d'allèles incorrects étaient biaisées vers les insertions pour les homopolymères A/T et les délétions pour les STRs dinucléotidiques (Figure 53). Augmenter le seuil de score de qualité des *reads* CCS de Q20 à Q80 n'a pas augmenté le pourcentage d'allèles originaux pour les microsatellites dinucléotidiques, mais l'a légèrement augmenté pour les homopolymères. Cependant, cette différence légère pourrait être principalement due à la grande quantité de *reads* CCS Q80 dans notre analyse (Figure 51.B).

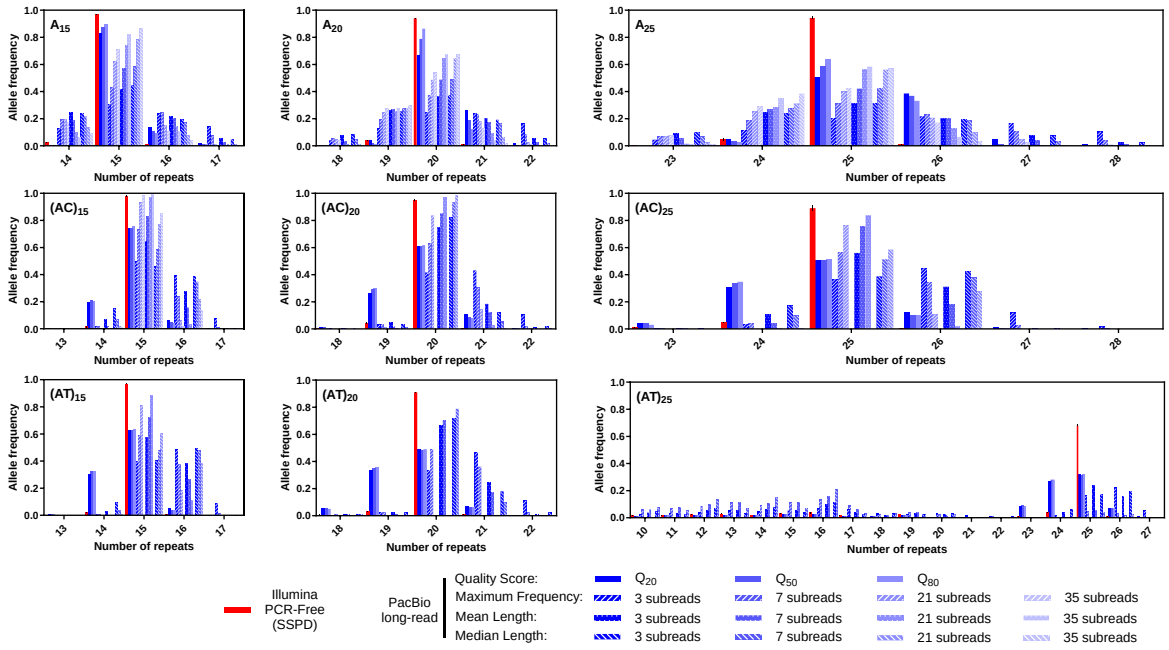

Figure 53 : Détermination précise de la longueur des microsatellites à partir du séquençage long-Read PacBio (Bibliothèques de Type IV) en utilisant différentes stratégies. Fréquences des allèles obtenues par séquençage long-read PacBio (bibliothèques de type IV) en utilisant différentes approches pour la détermination précise de la longueur des microsatellites. Chaque point correspond à une seule expérimentation. Les données Type I PCR-free – SSDP (deux réplicats) issues du séquençage short-read Illumina (NextSeq 500) sont également présentées dans le panneau B pour comparaison.

Nous avons ensuite tenté d'améliorer l'identification de la taille des allèles à partir du séquençage *long-read* en construisant des tailles consensus basées sur les quatre approches décrites précédemment (Figure 45.A), et en utilisant différents nombres de *sub-reads* allant de 3 à 35. Les approches '*Max Freq.*', '*Mean Length*' et '*Median Length*' ont présenté les meilleurs résultats, qui sont montrés dans la Figure 53. En augmentant le nombre de *sub-reads*, la proportion d'allèles originaux a augmenté avec les trois modes. Cependant, ces approches n'ont pas montré de meilleurs résultats que les *reads* CCS pour les homopolymères. En revanche, pour les microsatellites dinucléotidiques, les trois approches ont donné de meilleurs résultats que les *reads* CCS au-delà d'un seuil de *sub-reads* par taille consensus. Dans le cas des dinucléotides AC/TG, nous avons même pu augmenter les fractions d'allèles originaux au-dessus de celles dérivées des données « *PCR-free* » d'Illumina, avec la plus haute fraction étant de 0,994 pour $(AC)_{15}$ avec '*Mean Length*' en utilisant 35 *sub-reads* par taille consensus. Enfin on a pu aussi remarquer que les tailles d'allèles erronées étaient biaisées vers les insertions pour les STRs dinucléotidiques et équilibrées pour les homopolymères avec les trois approches utilisées, ce qui différait des *reads* CCS.

2.2. Discussion

Dans ce deuxième projet, une comparaison détaillée de plusieurs méthodes de séquençage de deuxième et troisième génération largement utilisées à ce jour ainsi que de diverses approches bio-informatiques, a été effectuée afin de déterminer de manière précise la taille des homopolymères et des répétitions STR mono- et di-nucléotidiques les plus fréquentes dans le génome humain. L'identification

précise des allèles et des génotypes microsatellites est connue pour être particulièrement difficile avec le séquençage NGS, en raison des artefacts de type *stutters* causés par le glissement de la polymérase lors des étapes de PCR et du taux d'erreur élevé lors du séquençage, notamment au niveau des séquences homopolymériques (Daunay et al., 2019; Fungtammasan et al., 2015; Raz et al., 2019; Ross et al., 2013; Stoler and Nekrutenko, 2021). Cette étude visait à identifier la meilleure approche de séquençage NGS pour caractériser la distribution réelle des allèles microsatellites présents dans un échantillon, en minimisant ou idéalement en supprimant les erreurs introduites lors de la préparation des librairies, du séquençage et de l'analyse bioinformatique. Cela pourrait être d'un intérêt particulier pour les échantillons contenant un mélange complexe d'allèles de longueur variable, comme dans les cancers MSI ou les cellules MMRD, ou même dans des applications médico-légales.

Plusieurs plasmides contenant des microsatellites en répétition mono- (A/T) ou di-nucléotidiques (AC/TG ou AT/TA) allant de 15 à 25 répétitions, considérés comme purs, ont été utilisés comme modèles pour la préparation des librairies et des expériences de séquençage NGS comparatives. Trois approches de séquençage ont été évaluées en *short-read* avec la technologie Illumina en utilisant les protocoles SSDP et RFP pour le type de matrice d'ADN et qui étaient basées sur des librairies de type I avec ou sans étape PCR (1 à 20 cycles), une correction d'erreur standard par l'utilisation d'un UMI unique marquant d'un code-barre un seul brin de l'ADN (librairies de type II), et le séquençage en duplex avec la possibilité d'une correction combinée d'erreurs (librairies de type III). Un quatrième type de librairies sans PCR avec le séquençage PacBio (type IV) a été évalué en utilisant des plasmides linéarisés (Figure 40). En ce qui concerne la condition SSDP, on peut considérer qu'elle était similaire aux expériences de séquençage ciblées pour des gènes où des loci pouvaient être capturés, amplifiés et séquencés à des coordonnées spécifiques (exemple de Padlock probes, smMIP, amplicon PCR) (Hiatt et al., 2013), tandis que la condition RFP était semblable aux expériences de séquençage du génome entier et de l'exome basées sur la fragmentation aléatoire d'un échantillon d'ADN (Zhao et al., 2021). Deux instruments ont été évalués pour le séquençage en *short-read* Illumina, l'iSeq 100 et le NextSeq 500. Bien que ce dernier soit moins recommandé pour les expériences de séquençage d'amplicons en raison de problèmes de diversité nucléotidique et de qualité de read, aucun problème de ce genre n'a été rencontré avec nos données NextSeq 500, qui ont montré des résultats similaires à ceux des données iSeq 100 (Figure 43). Cela est probablement dû à la grande quantité de PhiX utilisée et aux deux codes-barres dégénérés de 8 nucléotides séquencés lors des premiers cycles de chaque read dans les librairies de type III, dont le séquençage a été effectué sur la même *flow-cell* que les librairies de type I et II, offrant ainsi une diversité nucléotidique suffisante. Notre approche bioinformatique globale reposait sur l'identification des régions flanquantes des microsatellites pour estimer correctement leur longueur dans chaque read, une approche simple déjà utilisée dans certaines études (Gan et al., 2015; Herbreteau et al., 2021; Warshauer et al., 2013). Cette stratégie s'est avérée très efficace pour l'obtention d'allèle microsatellite consensus à partir de multiples *(sub-)reads* à partir de protocoles basés sur l'utilisation d'UMI (simples ou doubles) avec Illumina et PacBio, tandis que l'approche standard pour générer une séquence consensus échouait souvent en raison de trop nombreux indels et de la mauvaise qualité de *reads*, notamment après le microsatellite avec Illumina. Quatre modes ont été développés et évalués ('*Max. Freq.*', '*Max. Length*', '*Mean Length*' et '*Median Length*') pour la génération d'allèles consensus à partir des reads UMI d'Illumina (librairies de types II et III) et des *sub-reads* de PacBio, ainsi que deux modes supplémentaires ('*Max. Length*' et '*Equal Length*') pour le consensus combiné issu du séquençage en duplex.

L'analyse standard de la distribution de taille des microsatellites à partir de données en *short-reads* a montré l'effet croissant des *stutters* qui était principalement des délétions en fonction du nombre croissant de cycles de PCR et de répétition du motif (Figure 42). Les artefacts *stutters* étaient ainsi visibles dès le premier cycle PCR, indiquant qu'une seule étape d'élongation PCR peut altérer la fréquence des

allèles originaux présents dans un échantillon. En tant qu'approche standard, les *short-reads* générés à partir de librairies sans PCR présentaient le niveau le plus élevé d'allèles originaux, qui étaient plus élevés en condition SSDP qu'en condition RFP, bien qu'ils n'atteignissent pas 100%, comme idéalement souhaité. Ces résultats sont cohérents avec d'autres études, qui ont montré une proportion plus faible de reads erronés au niveau des séquences microsatellites dans les données WGS sans PCR par rapport aux données avec des étapes PCR (Fungtammasan et al., 2015).

La correction d'erreur à l'aide d'UMI unique ou double dans le séquençage en duplex sont deux approches qui améliorent la précision du séquençage et permettent la détection de variations génétiques rares, qui ont également été évaluées pour leur capacité à identifier de manière précise et correcte la taille des microsatellites (Figure 46 & Figure 49). La première limitation de ces approches était que la préparation des librairies nécessitait obligatoirement plusieurs cycles de PCR (12, 16 et/ou 20 cycles dans cette étude), ce qui a introduit de nombreux artefacts *stutters*. La correction standard d'erreur utilisée dans les librairies de types II et III basée sur le mode '*Max. Freq.*' a augmenté le pourcentage d'allèles originaux avec une réduction maximale du taux d'erreur de 10 fois, mais à un niveau inférieur à celui des librairies sans PCR. La correction combinée d'erreurs par séquençage en duplex a encore réduit les taux d'erreur et augmenté les proportions d'allèles originaux au même niveau ou légèrement au-dessus de ceux obtenus avec des données en *short-reads* en protocole sans PCR, mais cette correction n'était pas possible pour chaque microsatellite, notamment les plus longs en raison de la mauvaise qualité, de l'absence et/ou de données discordantes (Figure 49). La correction combinée d'erreur '*Equal Length*' a présenté des niveaux d'allèles originaux plus élevés qu'avec le mode '*Max. Length*', mais a atteint moins fréquemment une longueur consensus car elle nécessite deux allèles consensus concordants *αβ* et *βα*. Bien qu'elle soit couramment utilisée dans de nombreuses études basées sur la détection des mutations ponctuelles (Hiatt et al., 2013; Schmitt et al., 2012; Wang et al., 2019), l'impact de la correction standard et combinée d'erreur à l'aide d'UMI a été moins fréquemment décrit spécifiquement dans le contexte de la détermination de la longueur des microsatellites, en particulier dans les STR dinucléotidiques et les homopolymères (≥15 répétitions) longs. Les approches classiques basées sur les UMI ont déjà été évaluées soit dans des études sur le cancer investiguant des événements MSI rares, principalement dans les homopolymères à partir de l'ADN tumoral, sanguin ou plasmatique (Gallon et al., 2020, 2019; Georgiadis et al., 2019; Waalkes et al., 2018; Willis et al., 2019), ou dans des études médico-légales portant sur l'identification d'haplotypes à partir de mélanges d'ADN complexes utilisant principalement des STR tétranucléotidiques (Crysup et al., 2023; Woerner et al., 2021). Cependant, l'impact de la correction d'erreurs par UMI sur différents types et nombres de répétitions a rarement été décrit en détail. Récemment, un protocole dérivé du séquençage en duplex a été évalué sur des microsatellites mononucléotidiques de 8 à 18 répétitions. Il a présenté 0,48 % d'allèles erronés contre 5,5 % avant correction combinée d'erreur (Bae et al., 2023), un résultat proche de celui obtenu dans ce projet.

Enfin, la capacité du séquençage en *long-read* PacBio à déterminer correctement la taille des microsatellites a été évaluée dans ce projet. Les problèmes liés aux taux d'erreur élevés (indel) et à la précision de l'identification des allèles microsatellites au niveau des homopolymères et d'autres STR sont bien connus et ont déjà été rapportés dans certaines études de séquençage en *long-read* PacBio et Nanopore (Chiu et al., 2021; Dolzhenko et al., 2024; Fang et al., 2022; Mitsuhashi et al., 2019; Ross et al., 2013; Sacristán-Horcajada et al., 2021; Zhang et al., 2020). L'analyse des STR avec le séquençage de troisième génération est considérée comme très difficile à ce jour. D'après les expériences réalisées dans ce projet, les *reads* CCS avec la plus haute qualité (Q80) présentaient encore une grande proportion d'allèles erronés, bien qu'une légère amélioration ait été observée pour les homopolymères (Figure 51 & Figure 53). Ces résultats sont facilement surpassés par le séquençage *short-read* en protocole « *PCR-free* ». En établissant les allèles consensus à partir de différents nombres de *sub-reads*, la proportion

d'allèles originaux des STR dinucléotidiques a augmenté au-delà des valeurs obtenues à partir des *reads* CCS Q80 et parfois même celles obtenues par du séquençage *short-reads* en protocole « *PCR-free* ». Cela indique que la disponibilité et l'analyse des *sub-reads* pourraient être essentielles pour l'identification précise de la taille des STR dinucléotidiques. Cependant, cette amélioration n'a pas été observée pour les homopolymères. Pour les homopolymères, il est intéressant de noter que les *reads* CCS montraient un biais pour les allèles erronés vers les insertions, comme cela a été précédemment rapporté (Mitsuhashi et al., 2019; Ross et al., 2013; Sacristán-Horcajada et al., 2021), mais les allèles consensus générés à partir des *sub-reads* par l'approche développée au laboratoire présentaient des insertions et des délétions erronées relativement équilibrées. Par conséquent, il est possible que l'algorithme de traitement des *reads* CCS pourrait être responsable de ce biais de délétion et pas la chimie de séquençage SMT.

Ce projet fait l'objet d'un article en cours de révision dans Nucleic Acid Research auquel je suis associée en tant que co-1er auteur: A detailed analysis of second and third generation sequencing approaches for accurate length determination of short tandem repeats and homopolymers (Sophie I. Jeanjean, Yimin Shen, Lise M. Hardy, Antoine Daunay, Antonio Alberdi, Emmanuel Tubacher, Jean-François Deleuze & Alexandre How-Kit).

3. Projet 3 : Influence de la température sur les mutations des microsatellites

3.1. Résultats

Dans cette partie sera présentée l'intégralité des résultats décrivant l'impact de la température, de même que d'autres facteurs, qui sont intrinsèques aux microsatellites (nombre et type de répétitions), sur les mutations des microsatellites dans des contextes *in vitro* et *in vivo*. Dans un premier temps les résultats pour les méthodes *in vitro* seront explicités puis les méthodes *in vivo* viendront compléter cette étude.

3.1.1. Expérimentations en contexte in vitro

L'étude des mutations microsatellites dans un contexte *in vitro* implique d'effectuer des expériences de réplication d'ADN à l'aide d'un système pouvant être modulable au niveau de différents paramètres. Dans le cadre de ma thèse, j'ai voulu explorer l'effet de la température sur la formation des mutations microsatellites lors de la réplication de ces séquences. Afin de pouvoir évaluer son influence sur une large amplitude de températures, allant de 15°C à 65°C, trois méthodes d'amplifications de l'ADN ont été sélectionnées utilisables à des plages différentes de températures, principalement du fait des polymérases utilisées. La PCR, avec un cyclage en deux étapes, permet d'amplifier l'ADN de manière cyclique, dans une étape d'hybridation et d'élongation commune à des températures allant de 45°C à 65°C, à la suite d'une étape de dénaturation à 95°C. Les deux autres méthodes utilisées sont des méthodes d'amplifications isothermes de l'ADN où la réaction se produit à une température constante. Ainsi le WGA (*Whole Genome Amplification*) et la RPA (*Recombinase Polymerase Amplification*) permettent d'amplifier l'ADN à des températures allant de 15°C et 37,5°C pour la première et de 25°C et 45°C pour la seconde. L'ensemble de ces trois méthodes offre ainsi une couverture optimale pour les différentes températures que j'ai souhaité étudier (Figure 54). Les microsatellites utilisés sont soit des microsatellites artificiels (A_{15}, A_{20}, A_{24}, $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$) contenus dans des plasmides, soit des marqueurs microsatellites mono- et di-nucléotidiques du génome humain connus du laboratoire (NR24,

CAT25, BAT26, D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123). Dans le second cas, l'ADN utilisé provient d'un donneur sain de l'EFS âgé de moins de 40 ans, qui est homozygote pour le microsatellite concerné (à l'exception d'un donneur différent pour le marqueur D2S123), et dont le nombre de répétitions précis a été déterminé par séquençage Sanger et NGS.

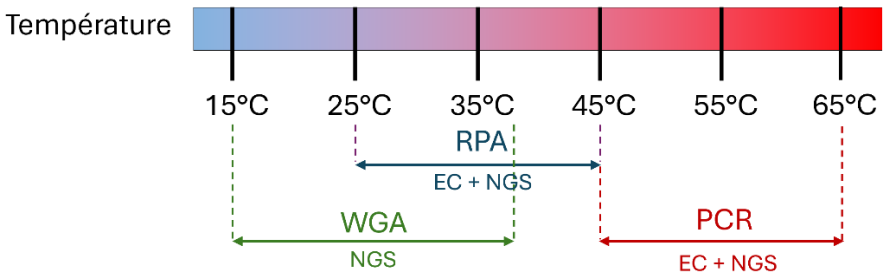

Figure 54 : Echelle des températures couvertes par les différentes méthodes d'amplification utilisées ainsi que les méthodes d'analyse utilisées.

Dans ces deux configurations, le nombre de molécules d'ADN avant et après amplification est connu par mesure de la fluorescence avant et après amplification, ce qui nous permet de calculer le nombre de doublement de molécules grâce à la formule détaillée dans le projet 1 ($nb\ doublement\ mol. = \log_2(\frac{N_f}{N_i})$ avec N_f, le nombre de molécules finales et N_i, le nombre de molécules initiales).

Les amplicons ainsi produits ont ensuite été analysés par électrophorèse capillaire sur un SeqStudio, à l'exception des ADN générés par la méthode du WGA, et par séquençage sur un iSeq100. La proportion de chaque allèle ainsi analysé a été mesurée et rapportée en pourcentage à partir des données d'électrophorèse capillaire et NGS. Pour les données NGS les fréquences alléliques contiennent un bruit de fond induit par le NGS, qui a été soustrait à l'aide de librairies construites à partir de plasmides purs permettant de mesurer uniquement ce bruit de fond, comme décrit dans la Figure 39 (Projet 2) (Figure 55).

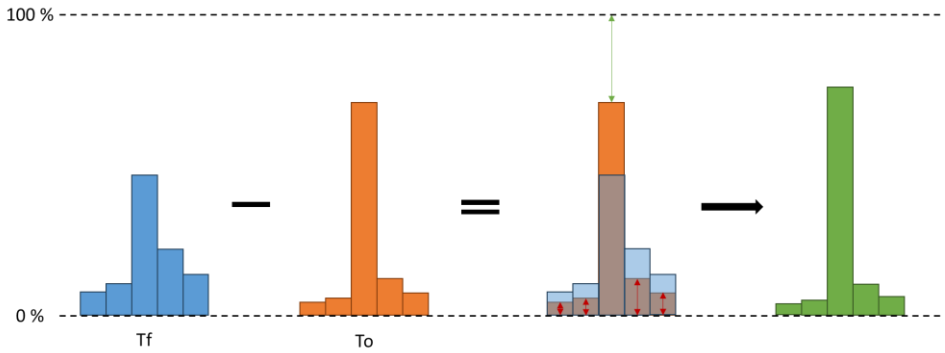

Figure 55 : Représentation schématique de l'effet de la soustraction du bruit de fond sur la répartition des fréquences alléliques. La sommes des fréquences restant à chaque étape égale à 100%

L'ensemble des résultats de fréquences alléliques ont été ensuite rapportés en fonction du nombre de doublement de molécule, ce qui permet de normaliser les variations d'événements de réplication. Il est à noter que ne possédant pas de séquences pour soustraire le bruit de fond pour les

marqueurs microsatellites du génome humain, ils ont été utilisés comme tels. Les données recueillies m'ont alors permis de calculer différentes mesures de fréquences de mutations par rapport au nombre de doublement de molécule, pour chaque température et microsatellites étudiés et de les visualiser en comparant soit la totalité des indels, insertions ou délétions d'un motif ou plus. Ces mesures exprimées en fréquence de délétion ou d'insertion par nombre de doublement de molécule permettent de normaliser les événements mutationnels, permettant de les comparer. A l'origine de ce projet, nous souhaitions utiliser l'algorithme d'inférence développé pour inférer les taux de mutations à partir des données de fréquences alléliques et de doublement de molécule. Cependant, dû au délai très long de calcul du programme développé dans le projet 1, qui nécessite beaucoup de ressources informatiques, je n'ai pas pu inclure directement les résultats des taux de mutations inférés au moment de la rédaction du manuscrit.

3.1.1.1. WGA

Le WGA (*Whole Genome Amplification*) est une technique permettant d'amplifier l'ADN génomique à partir d'une quantité très limitée d'échantillons. Le principe repose sur l'amplification aléatoire de l'ensemble du génome à l'aide d'une approche de type MDA (*Multiple Displacement Amplification*).

Le processus se déroule à température constante. Il commence par la dénaturation de l'ADN, séparant ainsi les deux brins d'ADN et rendant accessibles les séquences. Ensuite, des amorces dégénérées de type hexamère) se lient sur l'ADN fournissant des points de départ pour l'élongation. La polymérase Phi29 commence alors la synthèse de nouveaux brins d'ADN à température constante, tout en déplaçant les brins d'ADN déjà synthétisés, permettant ainsi de nouvelles liaisons d'amorces et une amplification continue. Le WGA permet alors d'obtenir suffisamment de matériel génomique pour des analyses ultérieures telles que le séquençage, la PCR ou les études de variation génétique, tout en préservant la représentation globale du génome original (Figure 56).

La polymérase Phi29, une enzyme dérivée du bactériophage Φ29, est une polymérase d'ADN avec une activité exonucléase 3'→5' (activité de relecture) qui offre une fidélité jusqu'à 1000 fois supérieure à celle de la polymérase Taq, avec un taux d'erreur estimé de $5x10^{-6}$ (Blanco and Salas, 1984; Dean et al., 2002; Pinard et al., 2006).

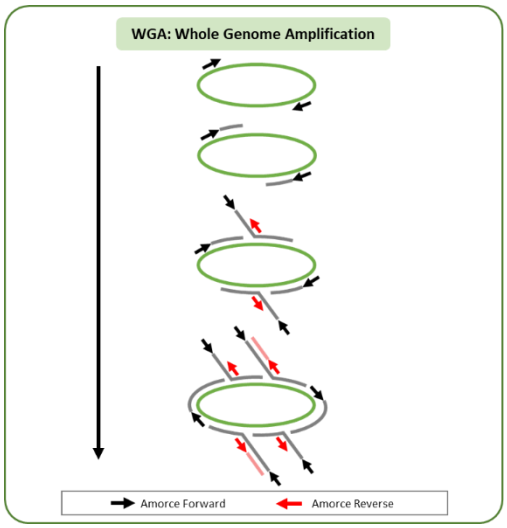

Figure 56 : Représentation schématique du processus d'amplification d'un plasmide par WGA. Adapté de (Vincent et al., 2004)

L'étude a été réalisée sur 6 microsatellites, 3 mono-nucléotides (A_{15}, A_{20}, A_{24}) et 3 di-nucléotides ($(AC)_{15}$, $(AC)_{20}$, $(AC)_{24}$) contenus dans un plasmide pUC57-Mini (Genscript) (Matériels et Méthodes 1.3.). Les 6 plasmides ont été mélangés de manière équimolaire avant de subir le WGA à l'aide d'hexamères dégénérés à différentes températures (15°C, 20°C, 25°C, 30°C, 35°C et 37,5°C). Les réactions ont été réalisées en quadruplicat en partant d'un nombre initial de plasmides de $3{,}68x10^6$ copies en suivant le protocole décrit dans la partie Matériels et Méthode 3.2.3. . Après amplification, les ADNs obtenus correspondent à des molécules d'ADN double brins linéaires répétant en tandem la séquence des plasmides et dont la taille est généralement supérieure à 10 kb.

Etant très visqueux et dans un petit volume réactionnel, l'ADN a été précipité et repris dans un volume de 500 µl permettant le dosage par fluorescence. On suppose que près de 100% de l'ADN est récupéré après précipitation du fait de sa concentration et de son haut poids moléculaire. Au final, entre 1,29 et 34 ng/µl d'ADN amplifié a été obtenu par échantillon. Cette quantification a permis le calcul du nombre de doublement de molécule allant de 18,5 à 22,3 (Figure 57).

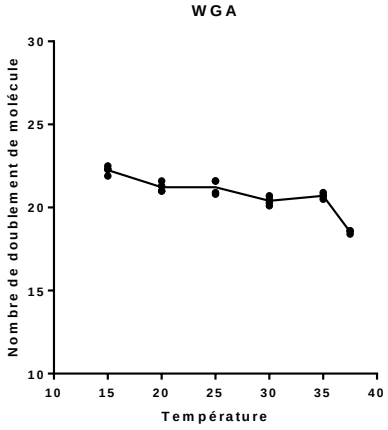

Figure 57 : Nombre de doublement de molécule après amplification d'un pool de 6 microsatellites par WGA en quadruplicats à différentes températures allant de 15 à 37,5 °C.

Dans le cas du WGA, la mesure des mutations microsatellites s'est faite exclusivement par NGS. Une excision des inserts des fragments d'ADN linéaires a été réalisée par triple digestion enzymatique. L'insert a été purifié et les librairies préparées avec ligature des adaptateurs et sans étape de PCR avant le séquençage, ce qui permet de maintenir l'intégrité du profil mutationnel introduit au cours du WGA (voir Travaux de recherche 2.1.1).

Les résultats obtenus (Figure 58) indiquent que pour tous les microsatellites étudiés, la fréquence de mutation rapportée par doublement de molécule (FMDM) augmente presque exponentiellement avec la température. Toutefois, il est important de remarquer la différence de comportement concernant les insertions et délétions selon le type de motifs et le nombre de répétitions. En effet, les résultats nous montrent que pour les microsatellites mono-nucléotides, les FMDM semblent subir une inversion de tendance de la délétion vers l'insertion (Figure 58.A) alors que pour les di-nucléotidiques, quel que soit la température, les fréquences tendent de préférence à l'insertion (Figure 58.B). De plus, la température à laquelle l'inversion a lieu varie selon le nombre de répétitions avec 32°C pour 15 répétitions, 28°C pour 20 et 35°C pour 24 (Figure 58.A.1). Enfin, il semblerait qu'il y a une prépondérance de mutations d'un motif (insertion et délétion) comparé à des insertions ou délétions plus grandes indépendamment du type de motif ou du nombre de répétions de ce dernier (Figure 58.A.2 et B.2) pour les motifs étudiés.

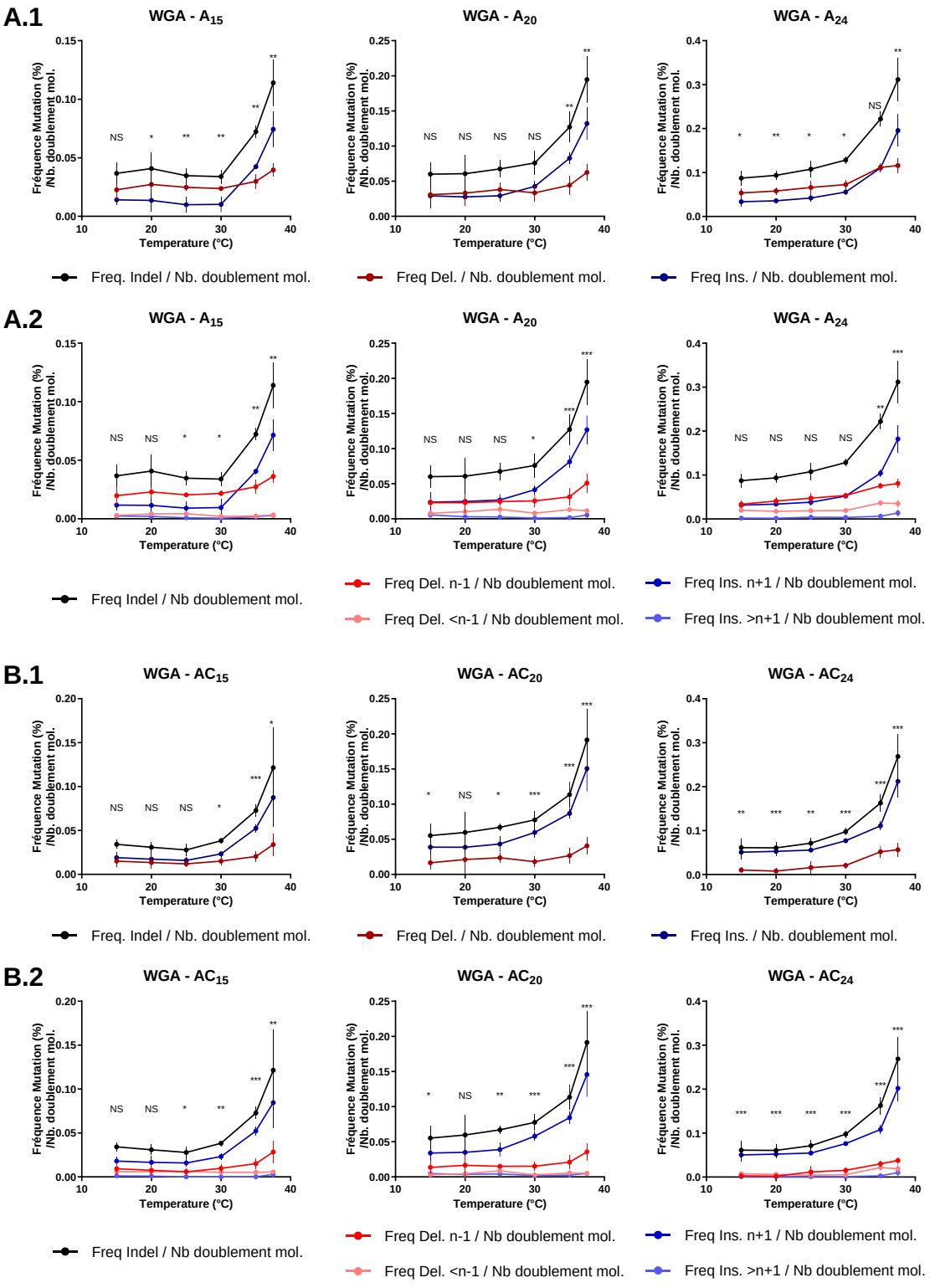

*Figure 58 : FMDM par rapport à la température obtenue par NGS sur WGA après soustraction du bruit de fond. A.1 Fréquence d'Indel (noir) (FIDDM), de délétion (rouge)(FDDM) et d'insertion (bleu)(FIDM) total en fonction du nombre de doublement de molécule pour les microsatellites A_{15}, A_{20} et A_{24}. A.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites A_{15}, A_{20} et A_{24}. B.1 FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$. B.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$. Marquage P.value : NS = >0.05, * = <0.05, ** = <0.01, *** = < 0.001*

Il est également intéressant de noter que la FIDM et la FDDM augmentent avec le nombre de répétitions (Figure 59). De plus, on observe également la préférence des mono-nucléotides pour l'insertion (Figure 59) contre une préférence pour les délétions pour les di-nucléotides (Figure 59).

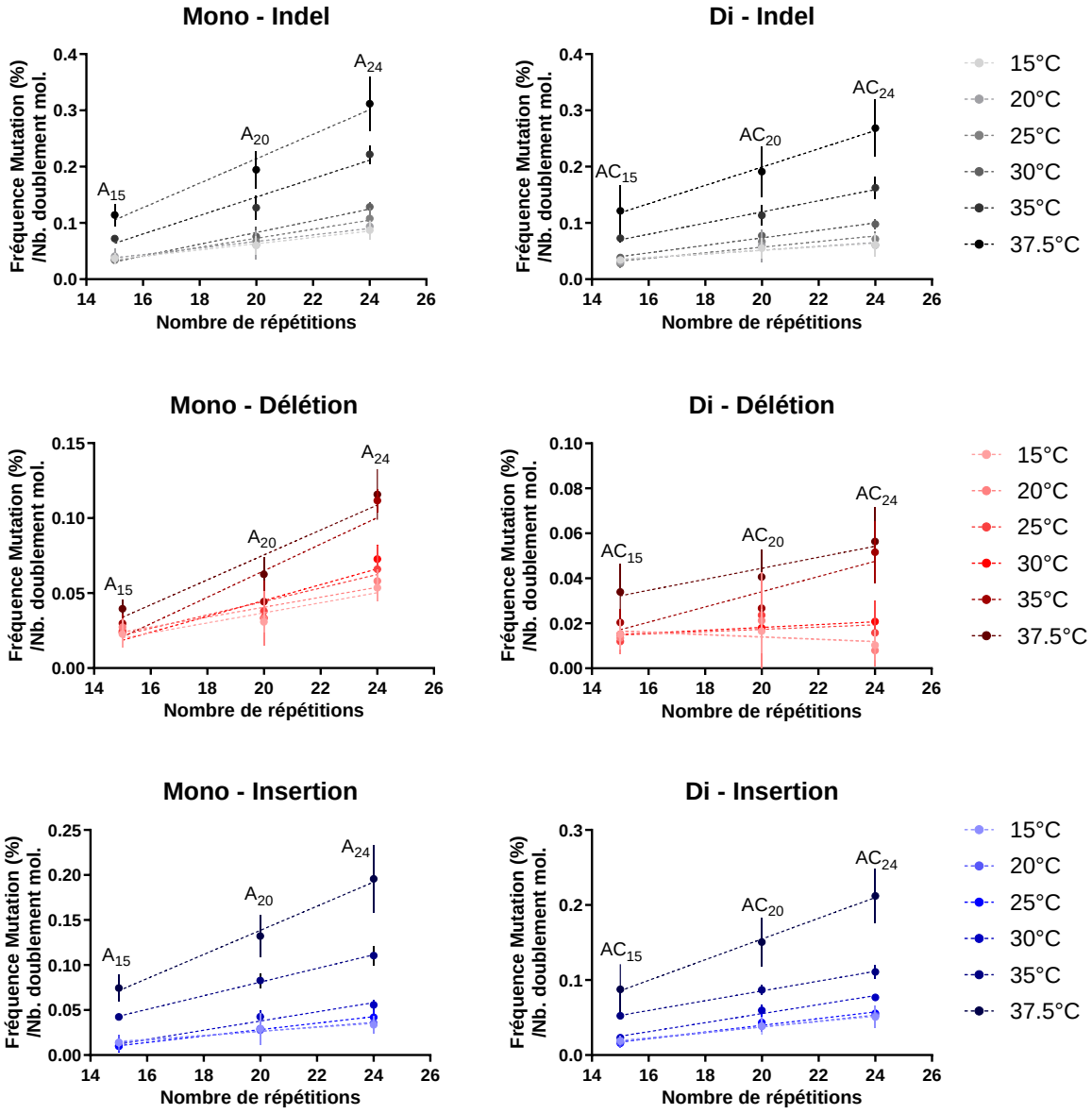

Figure 59 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et di-nucléotidiques $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$ de 15°C à 37,5°C obtenu par NGS sur WGA après soustraction du bruit de fond. Pour chaque graphique, une courbe de régression entre les points a pu être tracée.

3.1.1.2. RPA

La RPA (*Recombinase Polymerase Amplification*) est une technique d'amplification isotherme, permettant ainsi d'amplifier de l'ADN à une unique température. Le principe combine l'utilisation de recombinases, de protéines liant l'ADN simple brin (SSB, *Single-Strand DNA binding protein*) et d'une polymérase Bsu pour amplifier des séquences d'ADN spécifiques à une température généralement autour de 37-42°C (Figure 60).

La polymérase Bsu, aussi appelée polymérase Bsu DNA Pol I large fragment, est une enzyme dérivée de *Bacillus subtilis*. La particularité de cette polymérase réside dans ses capacités de fonctionnement à température constante, son activité 5'→3' exonucléase, sa bonne conjonction avec les recombinases et sa plus faible sensibilité aux impuretés et inhibiteurs. Elle a un taux d'erreur estimé à $3x10^{-4}$, similaire à celle de la Taq polymérase (Okazaki and Kornberg, 1964; Piepenburg et al., 2006).

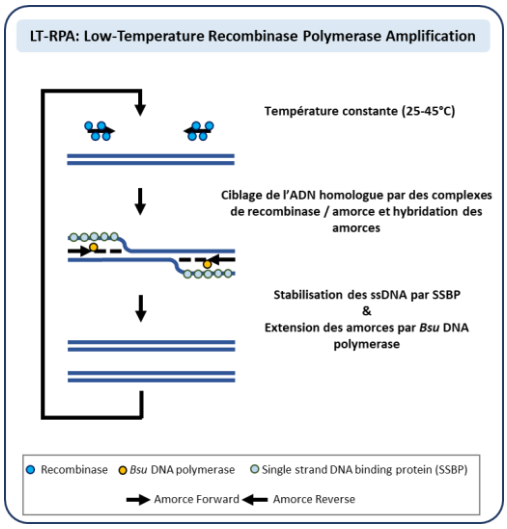

Figure 60 : Représentation schématique du processus d'amplification par LT-RPA. Adapté de (Piepenburg et al., 2006)

L'étude a été réalisée sur 6 plasmides pUC57-Mini (Genscript) différents, trois portant des microsatellites mono-nucléotides (A_{15}, A_{20} et A_{24}) et trois portant des microsatellites di-nucléotides ($(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$) (Matériels et Méthodes 1.3.) et sur 13 marqueurs microsatellitaires dont trois mono-nucléotidiques (NR24, CAT25 et BAT26) et 10 di-nucléotidiques (D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123, avec trois allèles analysés pour ce dernier $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ dont deux à partir d'un individu hétérozygote) stockés au laboratoire. Ces amplifications ont été réalisées à 25°C, 30°C, 35°C, 40°C, 42.5°C et 45°C avec pour chaque température, en quadripulicats, 92 000 ou 10 000 copies d'ADN par plasmide pour une analyse en électrophorèse capillaire ou NGS, respectivement, et avec 1562.5 copies d'ADN par marqueurs, le tout en utilisant le kit TwistAmp® Liquid Basic (Matériels et Méthodes 3.2.2.2.).

Les amplicons ainsi obtenus ont ensuite été analysés par électrophorèse capillaire ainsi que par NGS. Pour l'analyse par NGS, seulement 4 microsatellites et 8 marqueurs ont été analysés : A_{20}, A_{24}, $(AC)_{20}$, $(AC)_{24}$, D3S3630, D9S1838, D9S1825 D13S285, D14S1007, D18S61, D20S851 et D2S123 $(CA)_{21}$. La soustraction du bruit de fond du NGS ne pouvant être réalisée pour les marqueurs génomiques, les résultats provenant du NGS sont toutefois comparables à ceux de l'électrophorèse capillaire et ont ainsi pu être analysés de la même manière. Pour les amplicons des microsatellites analysés par électrophorèse capillaire, une digestion par l'enzyme BfaI a été réalisée sur leur intégralité afin de pouvoir s'affranchir

des problèmes possiblement liés à une polyadénylation incomplète chez les microsatellites mononucléotidiques A_n (Matériels & Méthodes 3.2.2.3.).

Il est visible que, pour tous les microsatellites, le nombre de doublement de molécule est relativement constant quelle que soit la température d'étude, à l'exception des résultats obtenus à partir de plasmides (Figure 61). Cette différence est due à la différence de quantité d'ADN utilisée au départ entre les deux expériences, ce qui était attendu.

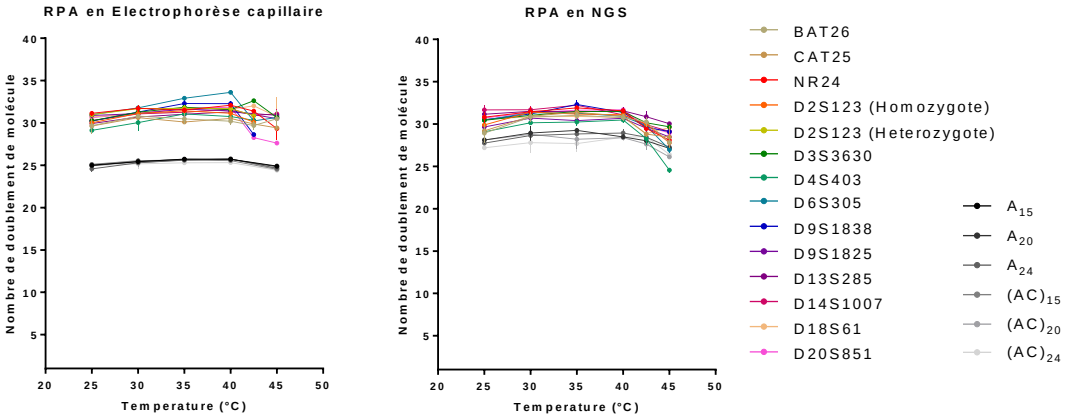

Figure 61 : Représentation graphique du nombre de doublement de molécule en fonction de la température selon la méthode d'analyse utilisée (Electrophorèse capillaire ou NGS) pour tous les microsatellites et marqueurs étudiés.

Concernant les résultats obtenus avec les plasmides, il est remarquable que la FMDM augmente avec la température pour tous les microsatellites, selon une tendance alternante entre le linéaire et l'exponentiel en fonction du microsatellite et de la méthode d'analyse.

Dans un premier temps, concernant les résultats obtenus par NGS, il est important de noter la différence de comportement concernant les insertions et délétions selon le type de motifs et le nombre de répétitions. En effet, les résultats nous montrent que pour les microsatellites mono-nucléotides, les FMDM semblent suivre une double inversion de tendance de la délétion vers l'insertion puis retour à la délétion (Figure 62.A) alors que pour les di-nucléotidiques, quelle que soit la température, la tendance s'oriente de préférence à la délétion (Figure 62.B). De plus, la température à laquelle cette double inversion a lieu varie selon le nombre de répétions avec 25°C puis 42.5°C pour 20 répétitions, contre 30°C puis 40-42.5°C pour 24 répétitions (Figure 62.A.1). Enfin, il semblerait qu'il y ait une légère prépondérance de mutations d'un motif (insertion et délétion) comparé à des insertions ou délétions plus grandes indépendamment du type de motif ou du nombre de répétions de ce dernier (Figure 62.A.2 & B.2) pour les motifs étudiés bien que ce soit moins marqué que pour les résultats concernant le WGA.

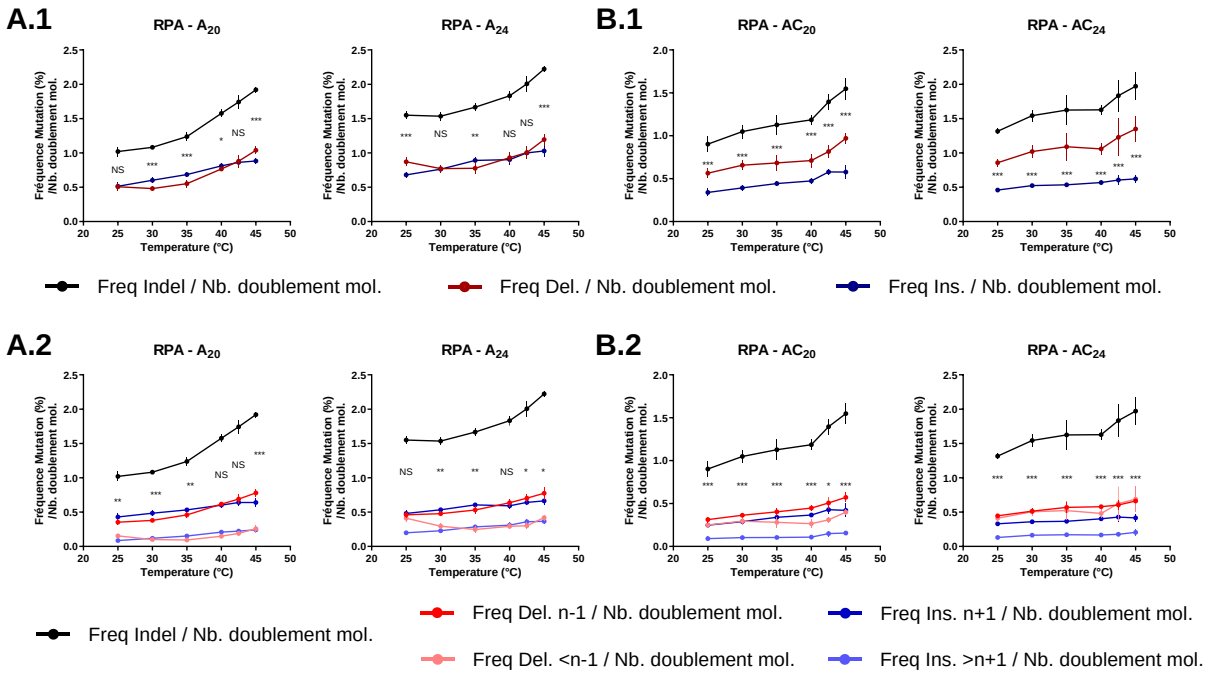

*Figure 62 : FMDM par rapport à la température obtenues par NGS sur RPA après soustraction du bruit de fond. A.1 FIDDM (noir), FDDM (rouge) et FIDM (bleu) total en fonction du nombre de doublement de molécule pour les microsatellites A_{20} et A_{24}. A.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites A_{20} et A_{24}. B.1 FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites $(AC)_{20}$ et $(AC)_{24}$. B.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites $(AC)_{20}$ et $(AC)_{24}$. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

Dans un second temps, concernant les résultats obtenus par électrophorèse capillaire, il est remarquable d'observer une tendance similaire à celle notée par l'analyse en NGS. Toutefois, pour les microsatellites mono-nucelotidiques, nous observons une simple inversion de tendance d'insertion vers la délétion d'autant plus marquée par le microsatellite A_{15} ne suivant pas cette mécanique (Figure 63.A). Enfin, concernant les microsatellites di-nucléotidiques, la tendance observée précédemment (Figure 62.B) est accentuée ici par l'écart grandissant en même temps que le nombre de répétitions (Figure 63.B).

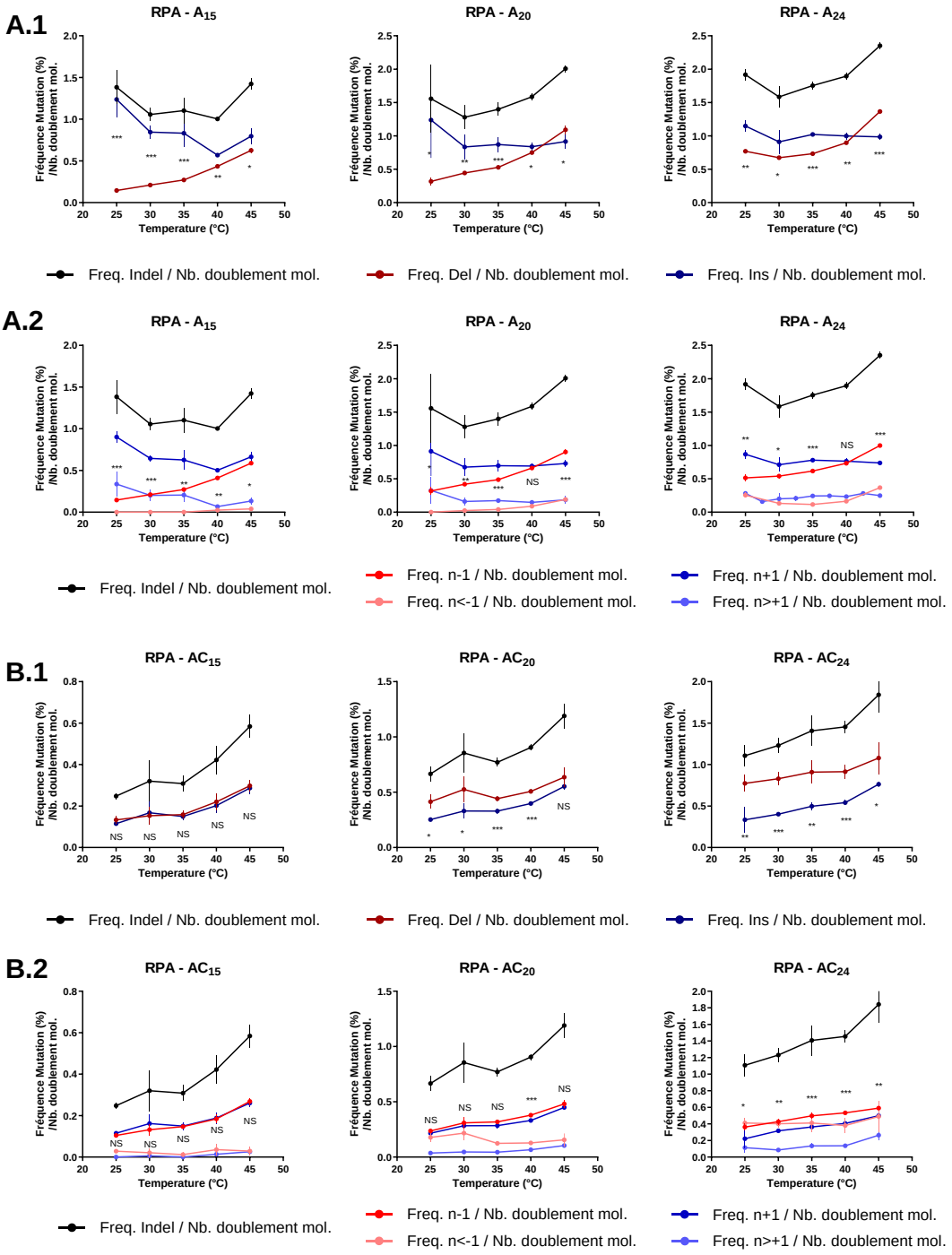

*Figure 63 : FMDM par rapport à la température obtenues par électrophorèse capillaire sur RPA. A.1 FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites A_{15}, A_{20} et A_{24}. A.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites A_{15}, A_{20} et A_{24}. B.1 FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$. B.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

Concernant les marqueurs génétiques analysés par NGS, la tendance semble être identique pour les mono-nucléotides concernant la double inversion de tendance, bien qu'il semble que plus le motif est long, plus l'inversion se déroule vers des températures élevées avec 28°C et 44°C pour NR24 (T_{23}), 34°C et 43°C pour CAT25 (T_{25}) et 37°C sans température de ré-inversion pour BAT26 (A_{26}) malgré un mouvement laissant supposer un retour à une FDDM supérieure à la FIDM (Figure 64.A.). De plus, il est important de noter que les mutations sont également favorisées pour des décalages d'un motif plutôt que plusieurs (Figure 64.B.), comme pour les plasmides.

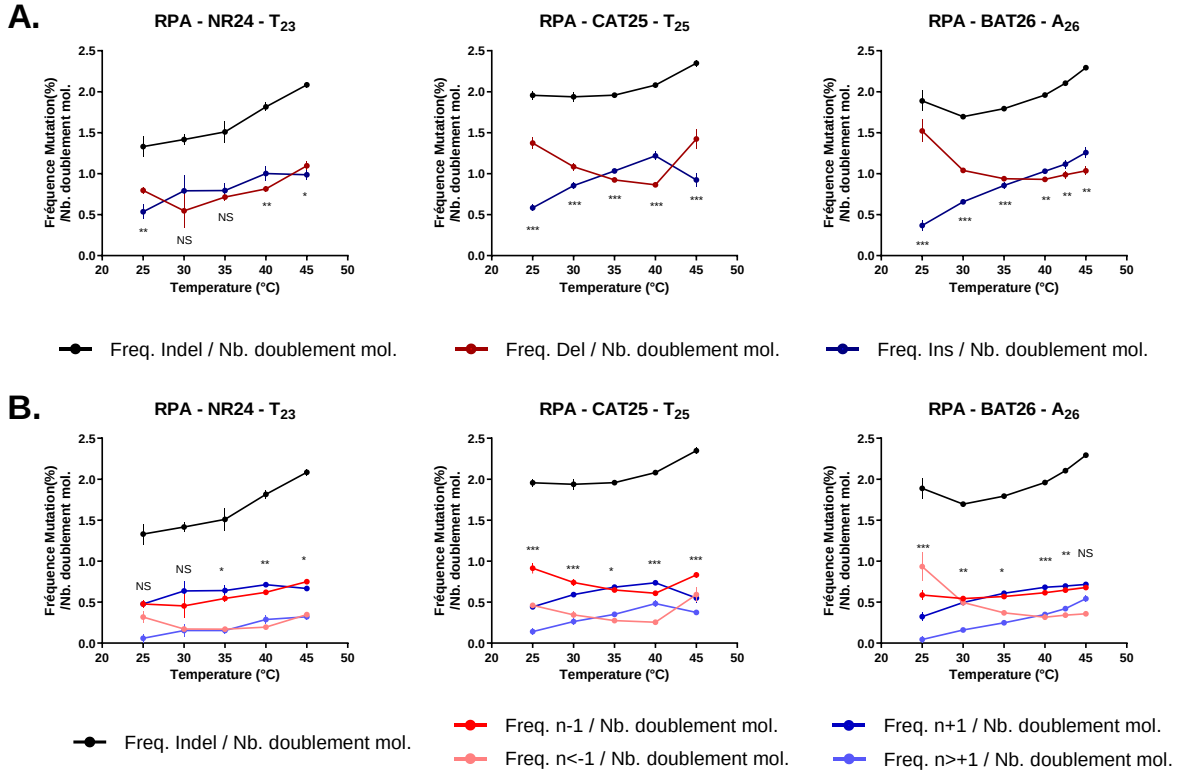

*Figure 64 : FMDM par rapport à la température obtenues par NGS sur RPA des marqueurs mono-nucléotidiques NR24, CAT25 et BAT26. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites NR24, CAT25 et BAT26. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites NR24, CAT25 et BAT26. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

Les marqueurs génétiques analysés par électrophorèse capillaire semblent également suivre une tendance similaire aux plasmides, avec pour les mono-nucléotides une simple inversion de tendance, bien que, contrairement au NGS, l'inversion se déroule toujours aux alentours de 30°C, sauf pour BAT26 (A_{26}) qui, malgré un mouvement laissant supposer une inversion, garde une FDDM supérieure à la FIDM (Figure 65.A.). De plus, il est important de noter que les mutations sont également favorisées pour des décalages d'un motif plutôt que plusieurs, sauf à 25°C (Figure 65.B.).

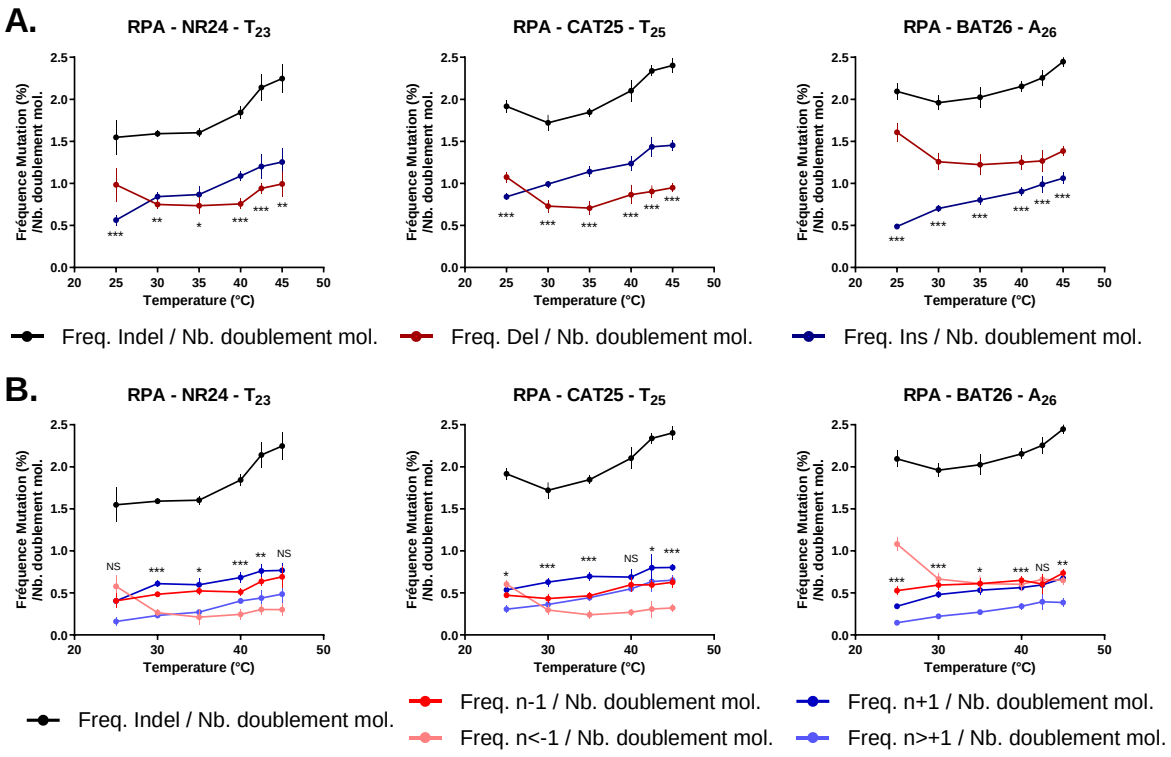

*Figure 65 : FMDM par rapport à la température obtenues par électrophorèse capillaire sur RPA des marqueurs mono-nucléotidiques NR24, CAT25 et BAT26. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites NR24, CAT25 et BAT26. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites NR24, CAT25 et BAT26. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

A présent, si l'on regarde les fluctuations des FMDM des marqueurs di-nucléotidiques en NGS, on peut observer de manière générale les mêmes phénomènes que sur les plasmides di-nucléotidiques. Toutefois, même s'il y a une différence entre FIDM et FDDM, l'écart est moins marqué. Il est notamment possible d'observer des marqueurs suivant une convergence, comme pour D3S3630, D18S61 et D13S285 allant même jusqu'au contact comme pour D14S1007 et D20S851 (Figure 66.A.). Pour ce qui est FMDM à un ou plusieurs motifs, la tendance est identique aux plasmides également, c'est-à-dire des FMDM supérieures pour les mutations à un motif, bien qu'il y ait des exceptions, la plus importante étant le marqueur D13S285 où la FDDM à un motif est inférieure en tout point à la FDDM à plusieurs motifs (Figure 66.B.).

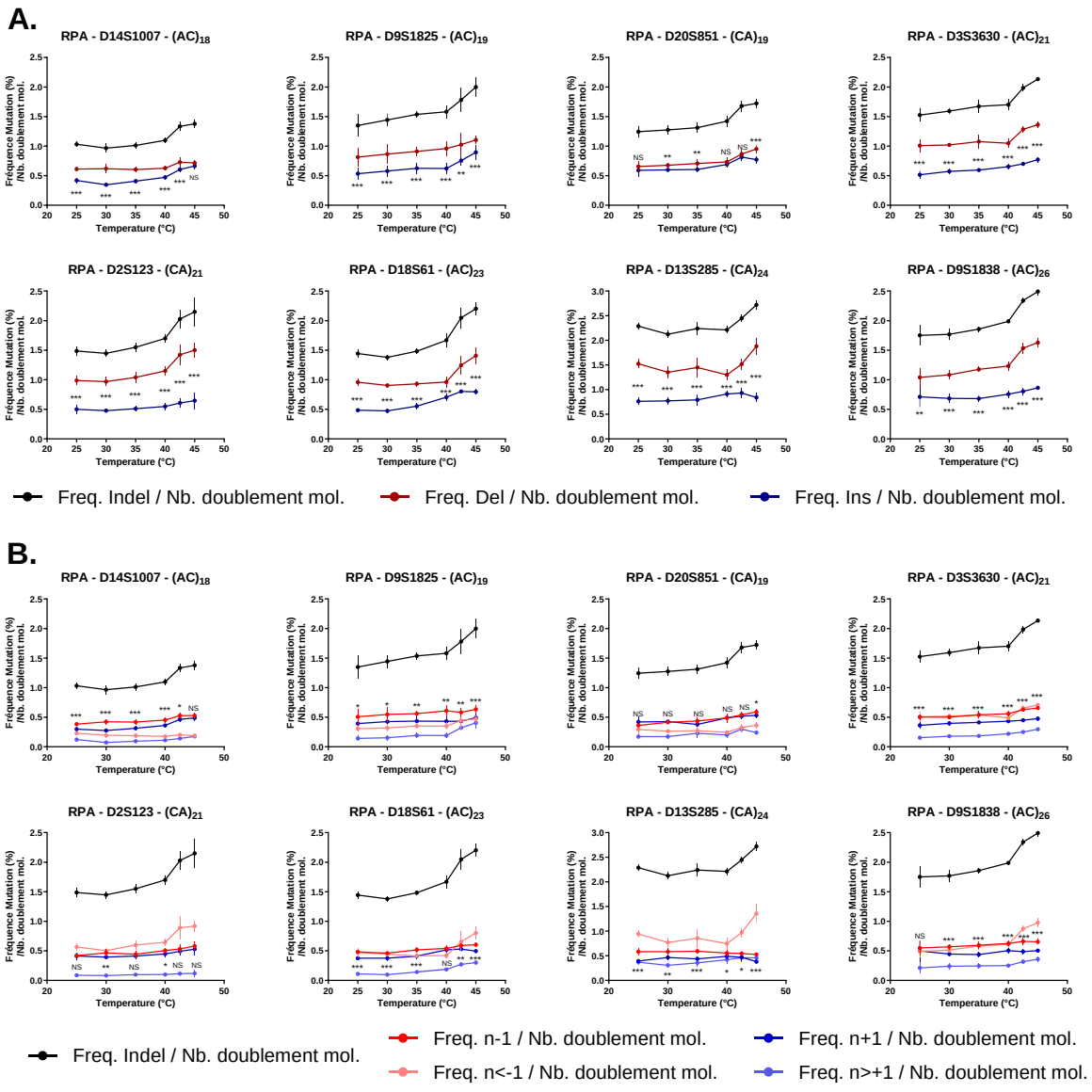

*Figure 66 : FMDM par rapport à la température obtenues par NGS sur RPA des marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair). Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

Concernant les FMDM des marqueurs di-nucléotidiques en électrophorèse capillaire, on peut observer les mêmes phénomènes qu'avec le NGS. On peut même observer des marqueurs suivant une convergence, comme pour D14S1007, D20S851, D3S3630, D2S123 sous sa forme $(CA)_{21}$ et D18S61 allant même jusqu'à l'inversion ou le contact comme pour D4S403 et D6S305 respectivement (Figure 67.A.). Pour ce qui est des FMDM à un ou plusieurs motifs, la tendance est identique au NGS (Figure 67.B.).

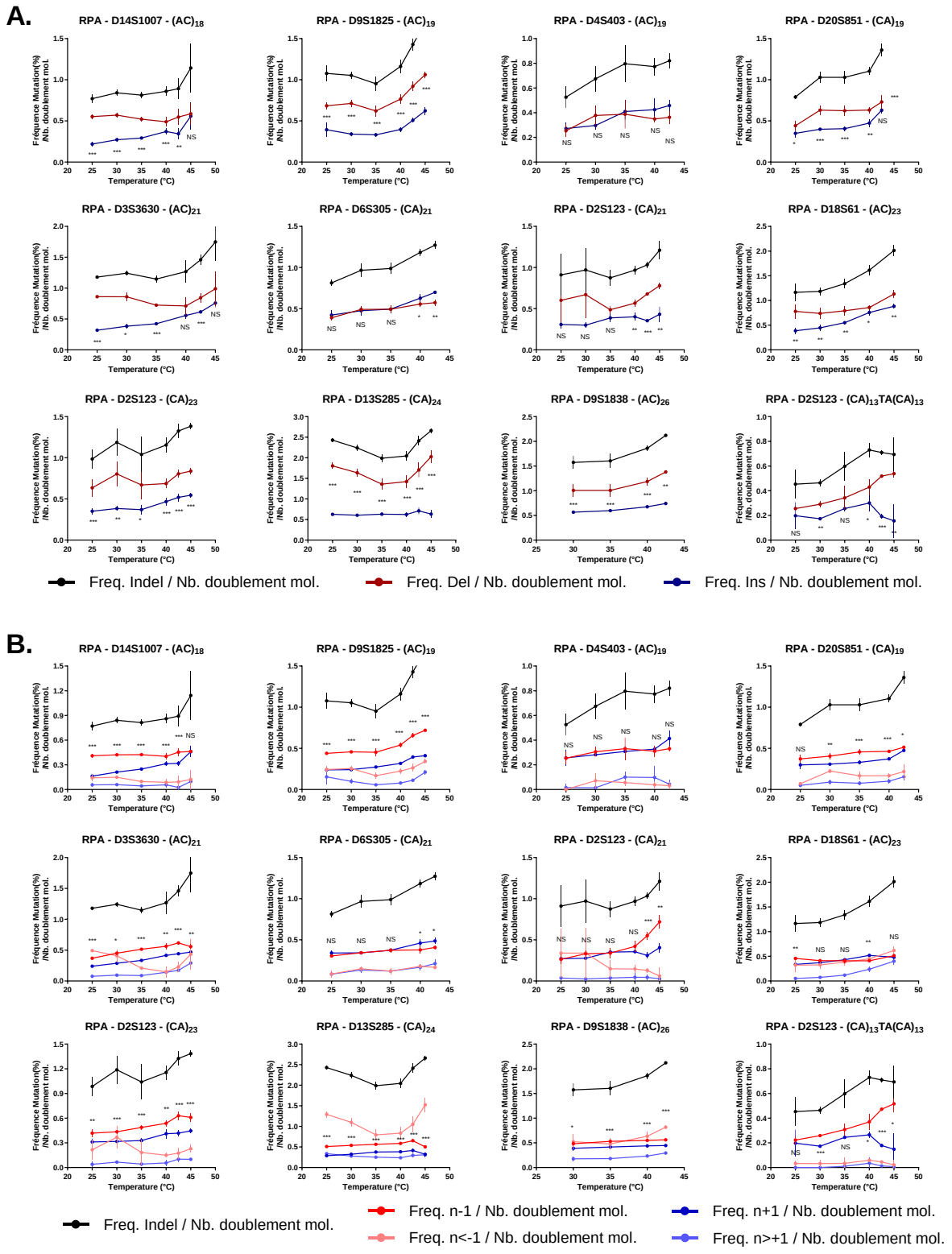

*Figure 67 : FMDM par rapport à la température obtenues par électrophorèse capillaire sur RPA des marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair). Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

Si l'on s'attarde sur la FMDM, il est intéressant de noter que la FDDM augmente avec le nombre de répétitions en analyse en NGS et par électrophorèse capillaire (Figure 68, Figure 69, Figure 70 & Figure 71), bien qu'il semble y avoir une tendance à la stagnation pour les di-nuclétoidiques pour le NGS (Figure 70). Concernant les FIDM, une inversion de tendance est visible entre 30°C et 35°C chez les mono-nucléotides en électrophorèse capillaire (Figure 69) là où elle maintient une augmentation en NGS et pour toutes les températures pour les di-nucléotides (Figure 70 & Figure 71). De plus, l'on observe une légère préférence des di-nucléotides pour la délétion (Figure 70 & Figure 71).

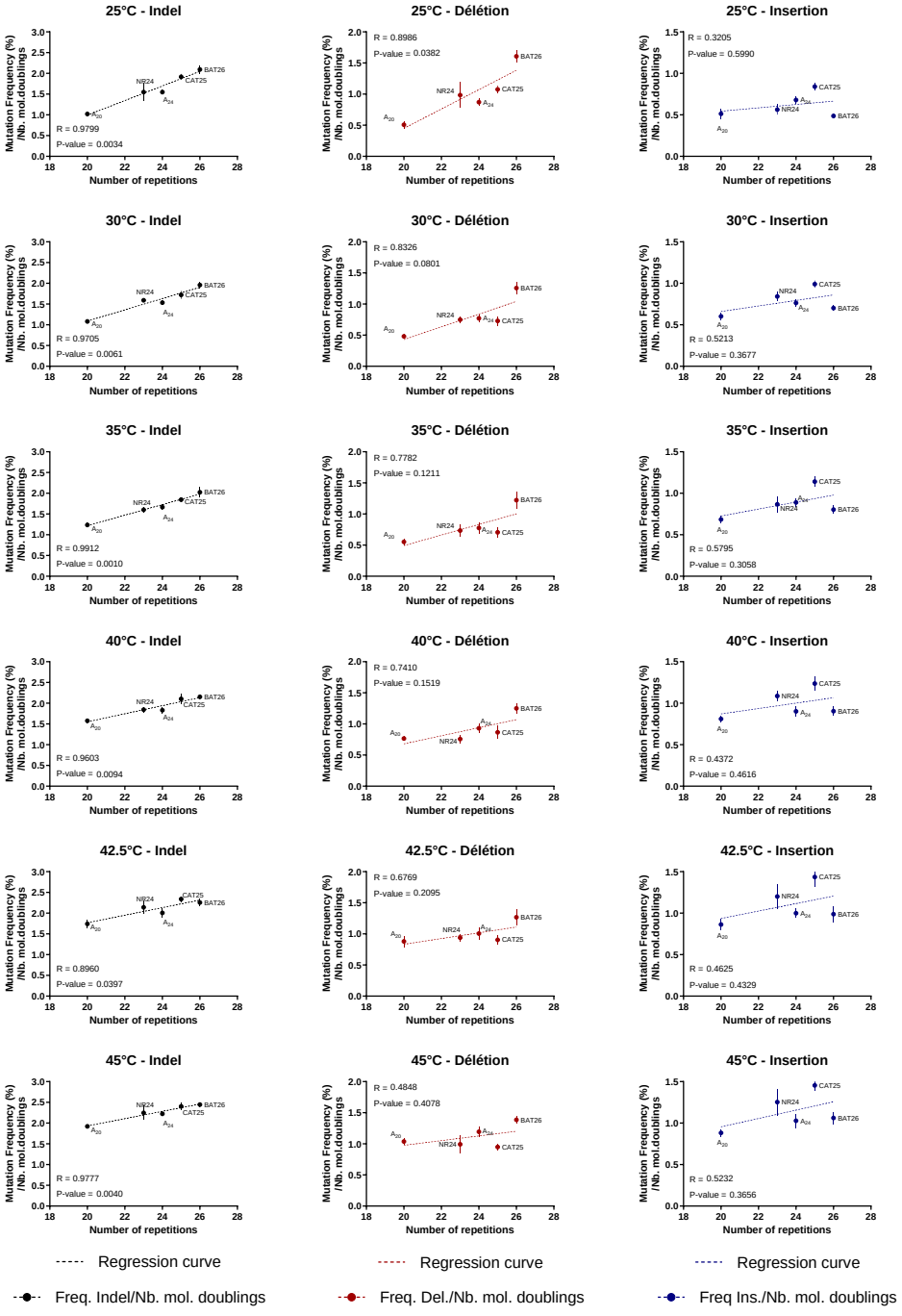

Figure 68 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et les marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 de 25°C à 45°C obtenu par NGS sur RPA avec soustraction du bruit de fond pour les plasmides. Pour chaque graphique, une courbe de régression entre les points a pu être tracée et le R calculé ainsi que la P.value.

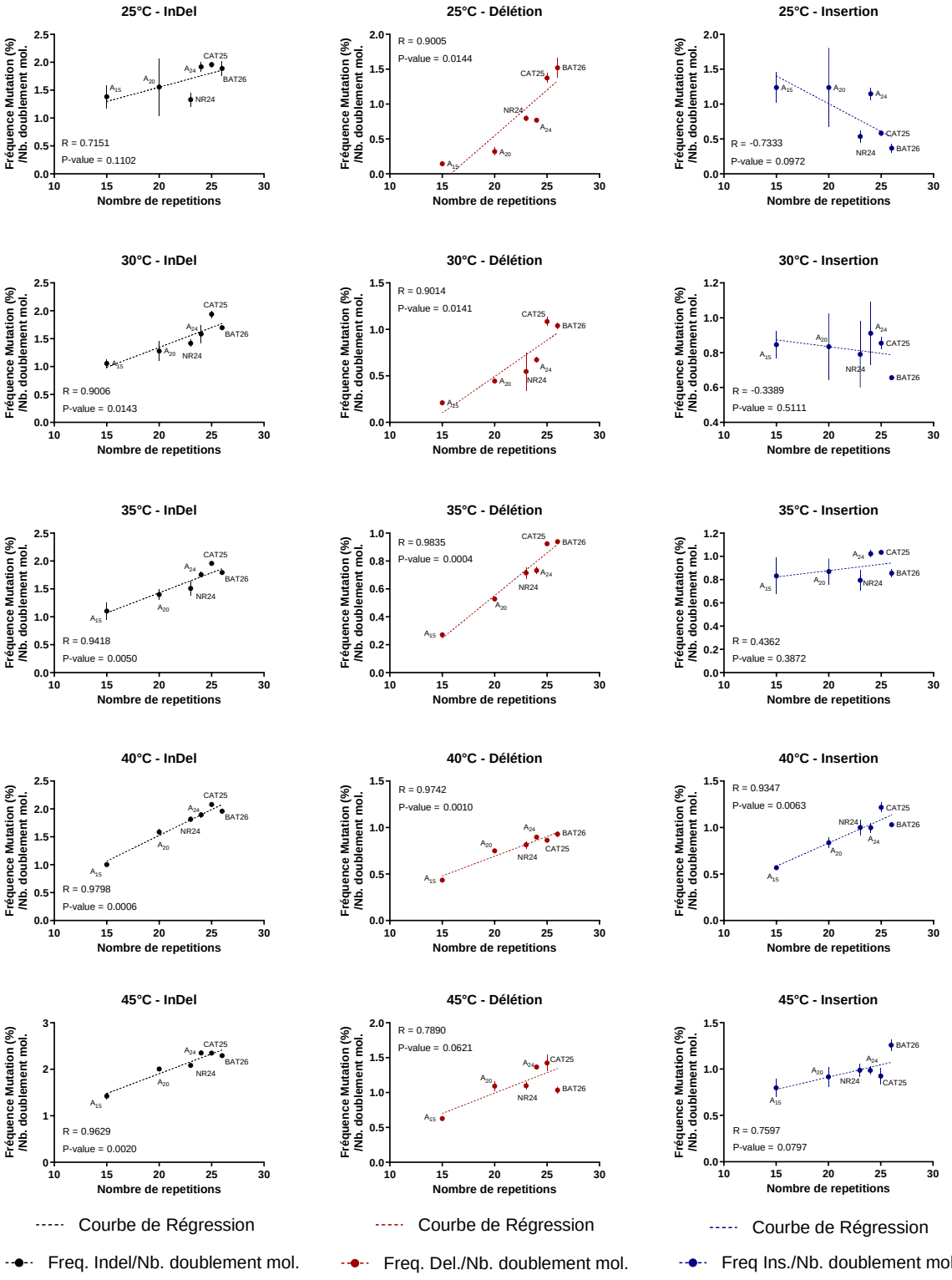

Figure 69 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et les marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 de 25°C à 45°C obtenu par électrophorèse capillaire sur RPA sans soustraction du bruit de fond pour les plasmides afin d'uniformisé les données comparées. Pour chaque graphique, une courbe de régression entre les points a pu être tracée et le R calculé ainsi que la P.value.

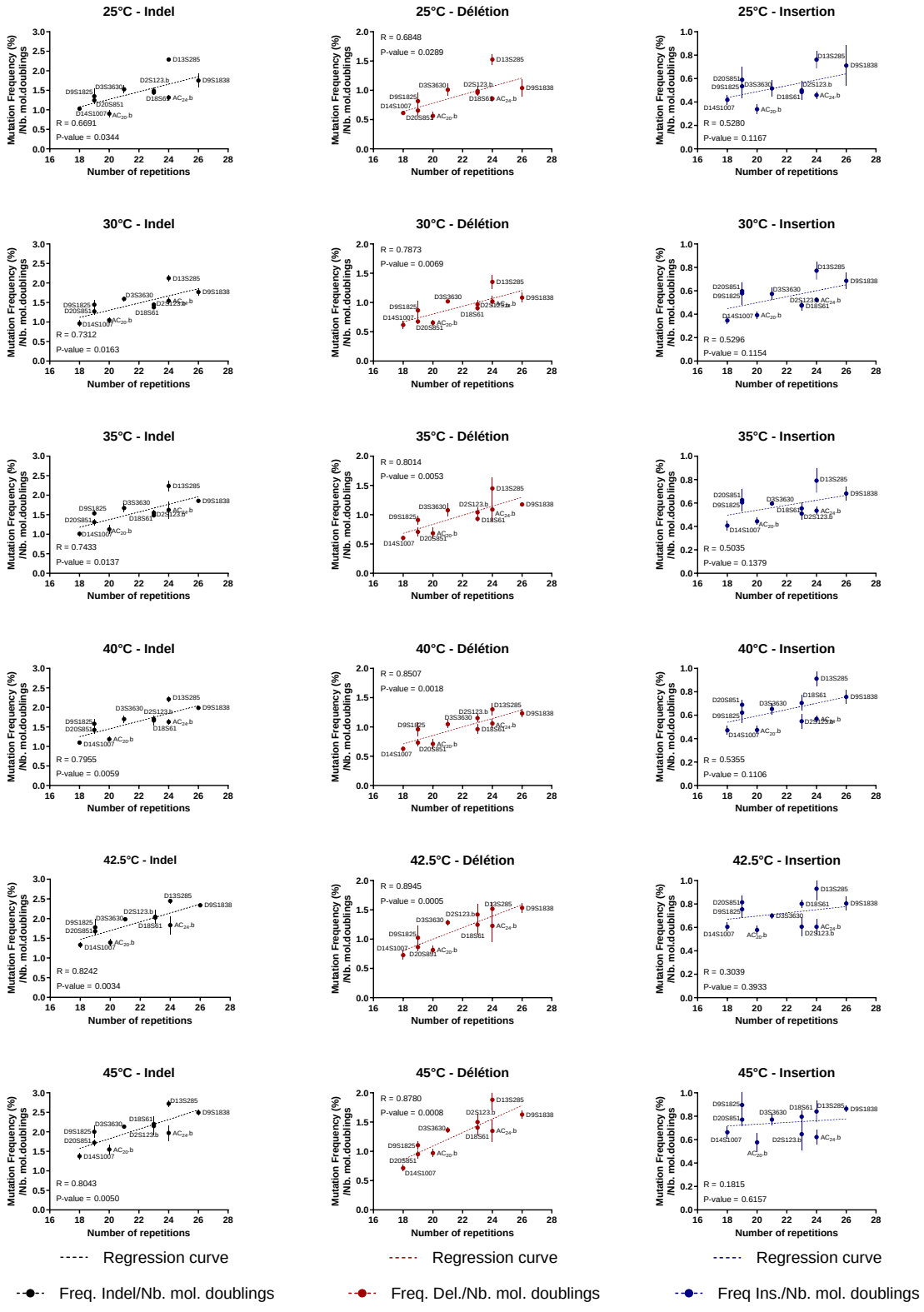

Figure 70 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites di-nucléotidiques $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$ et les marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ de 25°C à 45°C obtenues par NGS sur RPA avec soustraction du bruit de fond pour les plasmides. Pour chaque graphique, une courbe de régression entre les points a pu être tracée et le R calculé ainsi que la P.value.

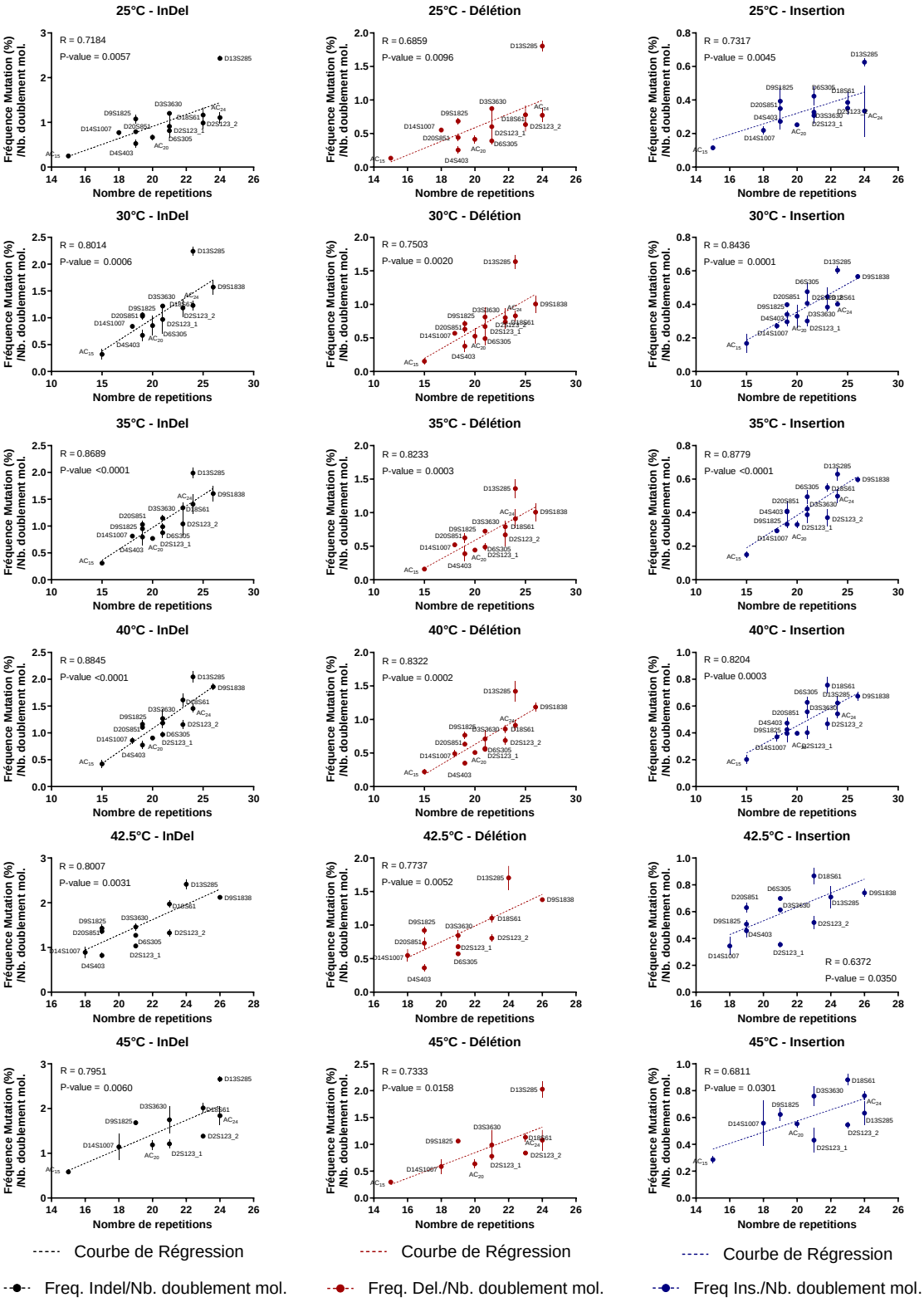

Figure 71 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites di-nucléotidiques $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$ et les marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ de 25°C à 45°C obtenues par électrophorèse capillaire sur RPA sans soustraction du bruit de fond pour les plasmides afin d'uniformisé les données comparées. Pour chaque graphique, une courbe de régression entre les points a pu être tracée et le R calculé ainsi que la P.value.

Enfin, il est intéressant de noter que de manière générale, les résultats par NGS ont des ecarts-type plus faibles que pour ceux par électrophorèse capillaire, les rendant ainsi plus répétables.

3.1.1.3. PCR

Pour cette étude, nous avons voulu étudier l'effet des températures d'amplification lors d'une PCR. Toutefois, là où une PCR classique se déroule en trois étapes (Matériels et Méthodes 3.2.1.), une PCR dite en 2 étapes a été réalisée ici où l'étape d'hybridation et d'élongation des amorces a été réalisée en même temps et à la même température. Cependant il faudra considérer que la réplication a pu débuter dès que les amorces se sont hybridées au moment de la descente en température, ce qui limite l'observation de l'effet unique de la température d'hybridation et d'élongation choisie sur la formation des mutations microsatellites. Ces PCR ont été réalisées avec la HotStart Taq polymérase, présentant un taux d'erreur de 10^{-5} d'après le fournisseur (Qiagen) (Figure 72).

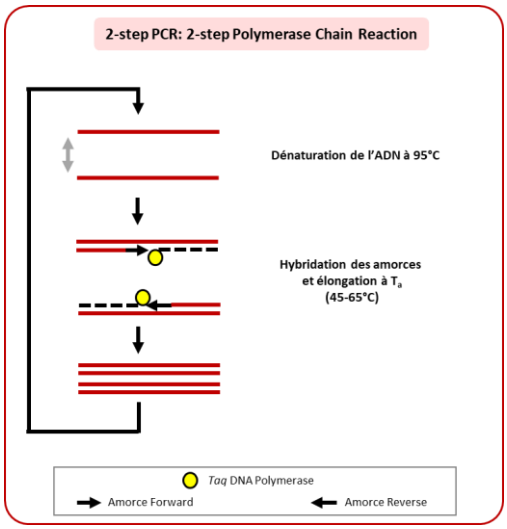

Figure 72 : Représentation schématique du processus d'amplification par PCR en 2 étape.

Pour la PCR, l'étude a été réalisée sur les 6 mêmes plasmides et 13 mêmes marqueurs que pour la RPA (A_{20}, A_{24}, $(AC)_{20}$, $(AC)_{24}$, NR24, CAT25, BAT26, D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présents comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$)) avec les mêmes retraits pour le NGS. Ces amplifications ont été réalisées à 45°C, 50°C, 55°C et 60°C en quadruplicats, avec 92 000 ou 10 000 copies d'ADN par plasmide pour une analyse en électrophorèse capillaire ou NGS, respectivement, et avec 1562,5 copies d'ADN par marqueurs, le tout en utilisant le protocole de PCR en temps réel présenté dans la partie Matériels et Méthodes 3.2.1.2.

Comme pour la RPA, les ADN ainsi obtenus ont ensuite été analysés par électrophorèse capillaire et NGS. La soustraction du bruit de fond du NGS ne pouvant être réalisée pour les marqueurs, les résultats provenant du NGS sont toutefois comparables à ceux de l'électrophorèse capillaire et ainsi être analysés de la même manière.

Il est à noter que la PCR subit une perte d'éfficacité à mesure que la température diminue, visible par le décalage du Ct. Toutefois, pour tous les microsatellites, le nombre de doublement de molécule est relativement constant quelle que soit la température d'étude (Figure 73).

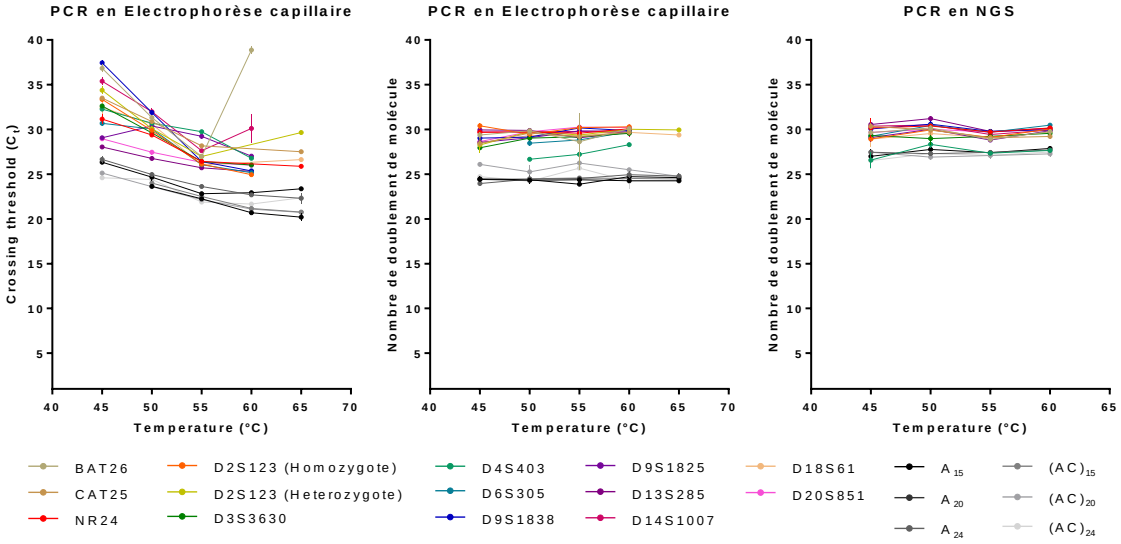

Figure 73 : Représentation graphique du nombre de doublement de molécule ou de la Ct (Crossing threshold) en fonction de la température selon la méthode d'analyse utilisée (Electrophorèse capillaire ou NGS) pour tous les microsatellites et marqueurs étudiés.

Ici, il est alors visible que pour tous les microsatellites des plasmides étudiés, la FMDM augmente légèrement, selon une tendance quasi linéaire, avec la température. Toutefois il est important de remarquer le comportement similaire concernant les FIDM et FDDM selon le type de motifs et le nombre de répétitions. En effet, les résultats nous montrent que pour tous les microsatellites mono- comme di-nucléotidiques, quelle que soit la température et la méthode d'analyse, les FMDM tendent de préférence à la délétion (Figure 74 & Figure 75 A.1 & B.1). Enfin, il semblerait qu'il y ait une légère prépondérance de mutations d'un motif (insertion et délétion) comparé à des insertions ou délétions plus grandes, indépendamment du type de motif ou du nombre de répétions de ce dernier (Figure 74 & Figure 75 A.2 et B.2) pour les motifs étudiés bien que ce soit moins marqué que pour les résultats concernant la RPA.

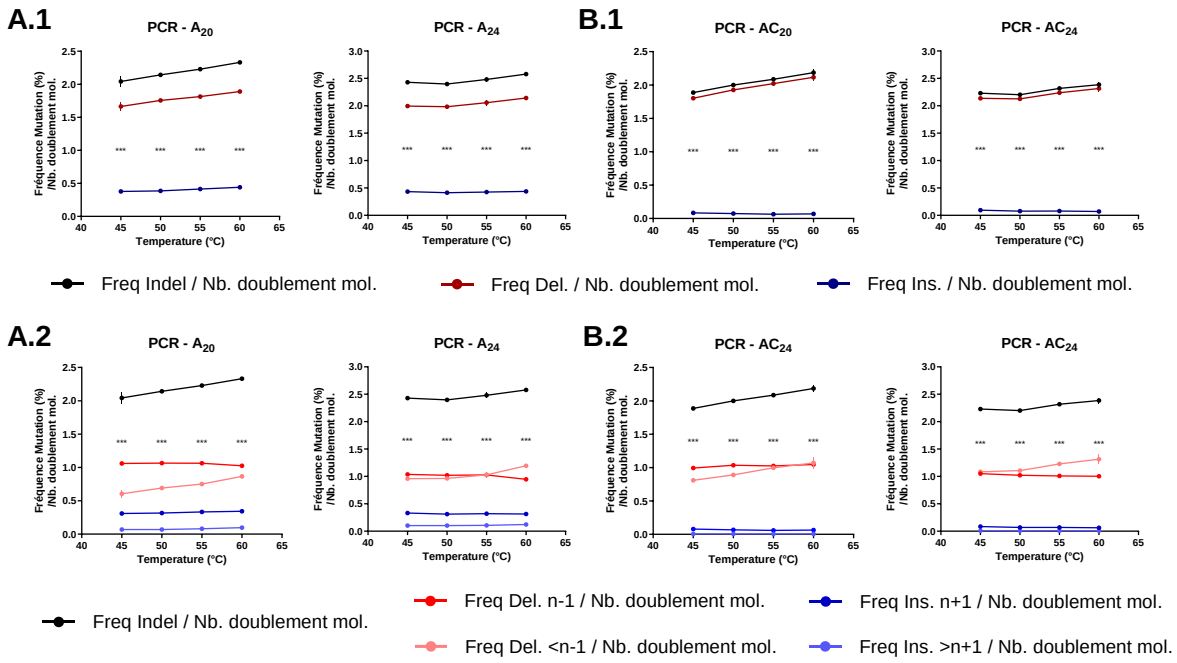

*Figure 74 : FMDM par rapport à la température obtenue par NGS sur PCR avec soustraction du bruit de fond. A.1 FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites A_{20} et A_{24}. A.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites A_{20} et A_{24}. B.1 FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites $(AC)_{20}$ et $(AC)_{24}$. B.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites $(AC)_{20}$ et $(AC)_{24}$. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

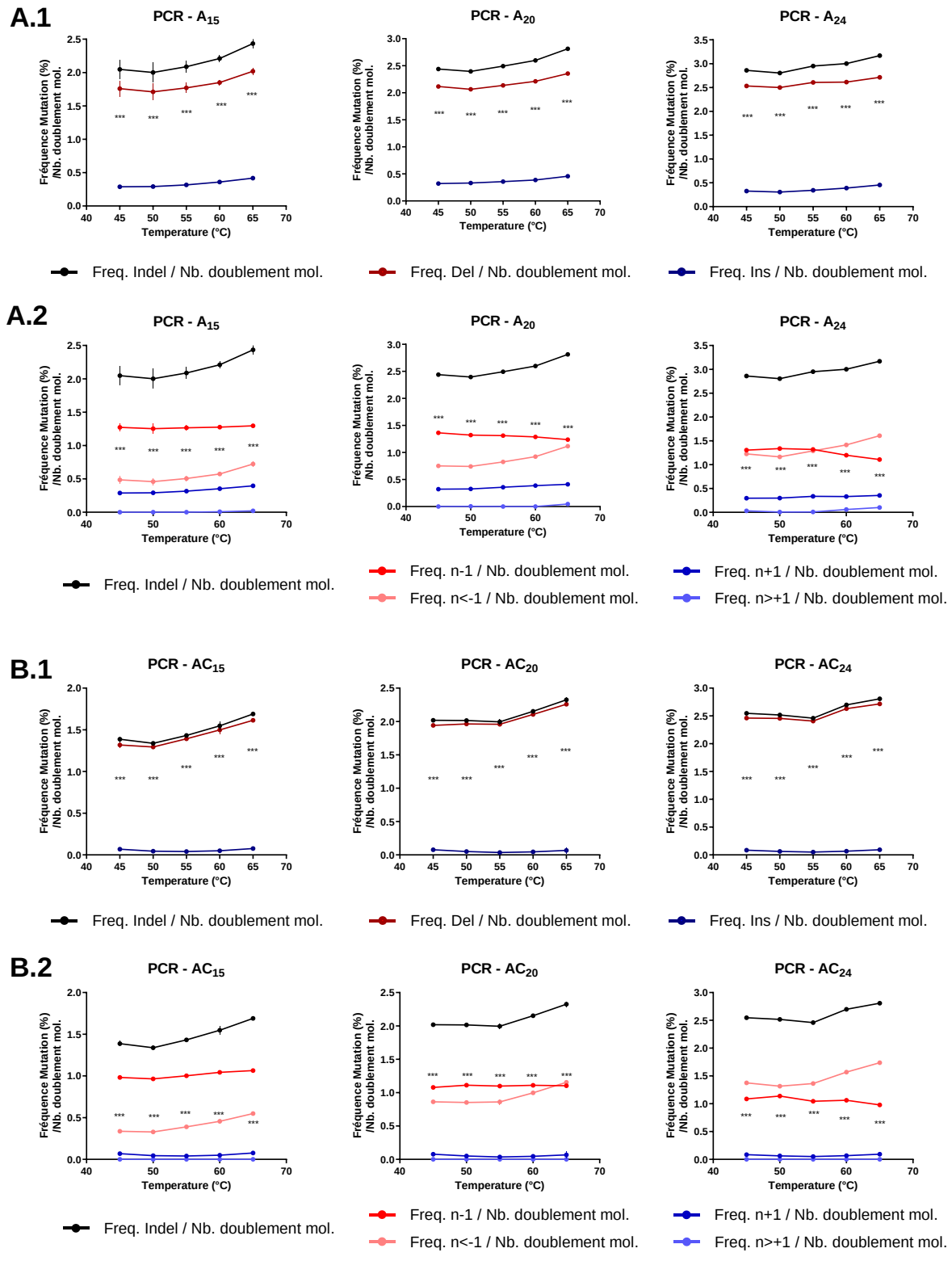

*Figure 75 : FMDM par rapport à la température obtenue par électrophorèse capillaire sur PCR. A.1 FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites A_{15}, A_{20} et A_{24}. A.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites A_{15}, A_{20} et A_{24}. B.1 FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$. B.2 FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

Concernant les marqueurs étudiés, comme pour les plasmides, la FMDM augmente faiblement avec la température. De plus, quel que soit le motif, le nombre de répétitions, la température et la méthode d'analyse, les FMDM tendent grandement vers la délétion (Figure 76, Figure 77, Figure 78 & Figure 79 A). Enfin, il semblerait qu'il y a une légère prépondérance de mutations d'un motif en insertion comparé à des insertions plus grandes indépendamment du type de motif ou du nombre de répétitions de ce dernier alors que pour les délétions, la prépondérance de mutations de plusieurs motifs prend le dessus à mesure que le nombre de répétitions des motifs augmente (Figure 76, Figure 77, Figure 78 & Figure 79 B)

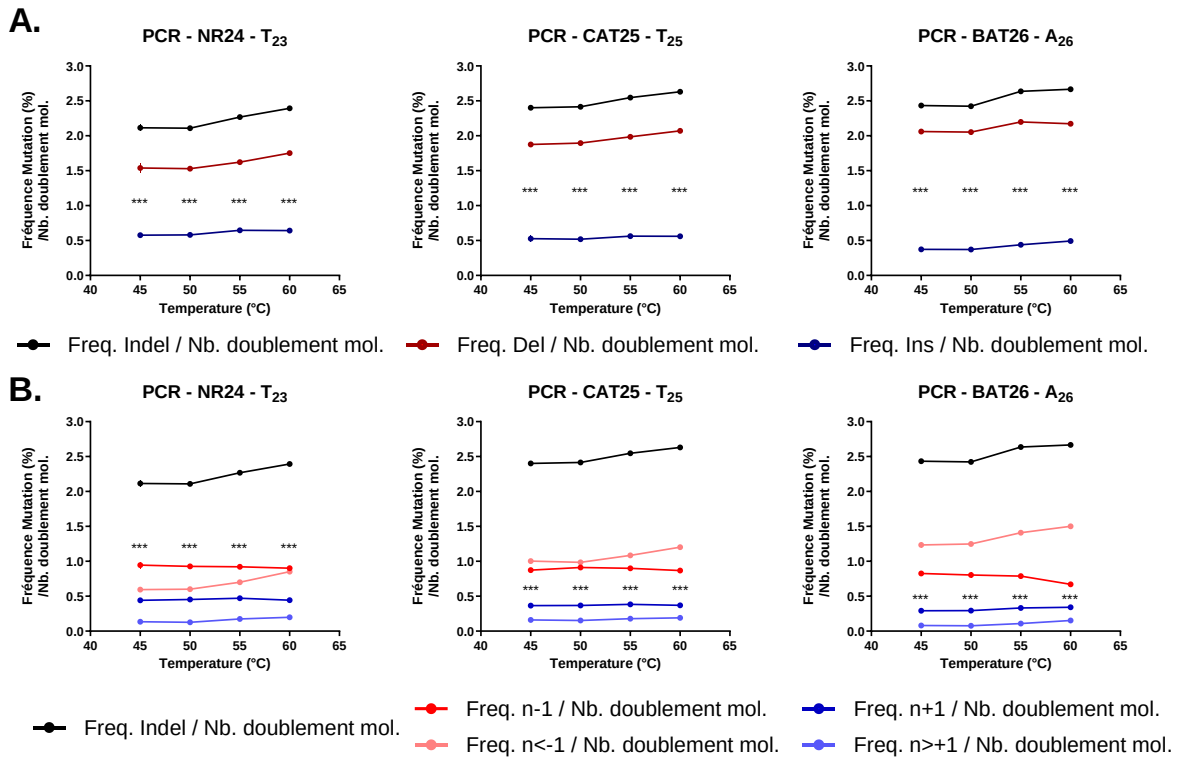

*Figure 76 : FMDM par rapport à la température obtenue par NGS sur PCR des marqueurs mono-nucléotidiques NR24, CAT25 et BAT26. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites NR24, CAT25 et BAT26. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites NR24, CAT25 et BAT26. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

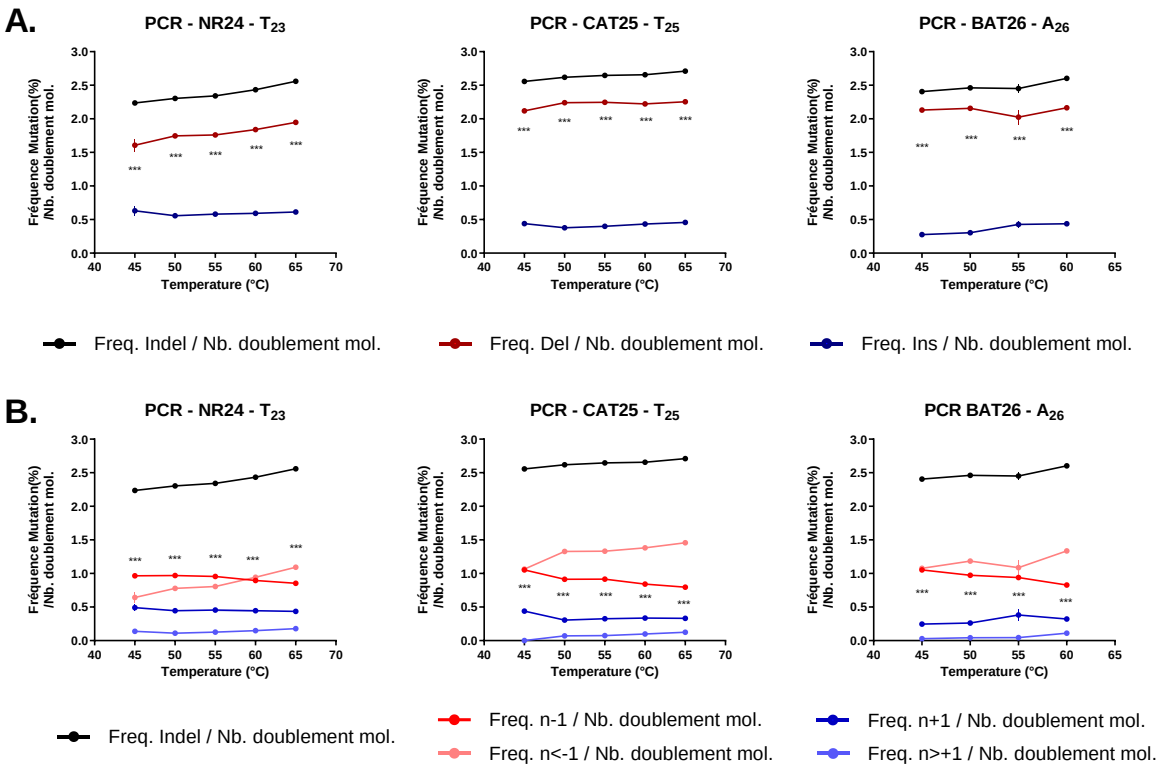

*Figure 77 : FMDM par rapport à la température obtenue par électrophorèse capillaire sur PCR des marqueurs mono-nucléotidiques NR24, CAT25 et BAT26. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les microsatellites NR24, CAT25 et BAT26. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair) pour les microsatellites NR24, CAT25 et BAT26. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

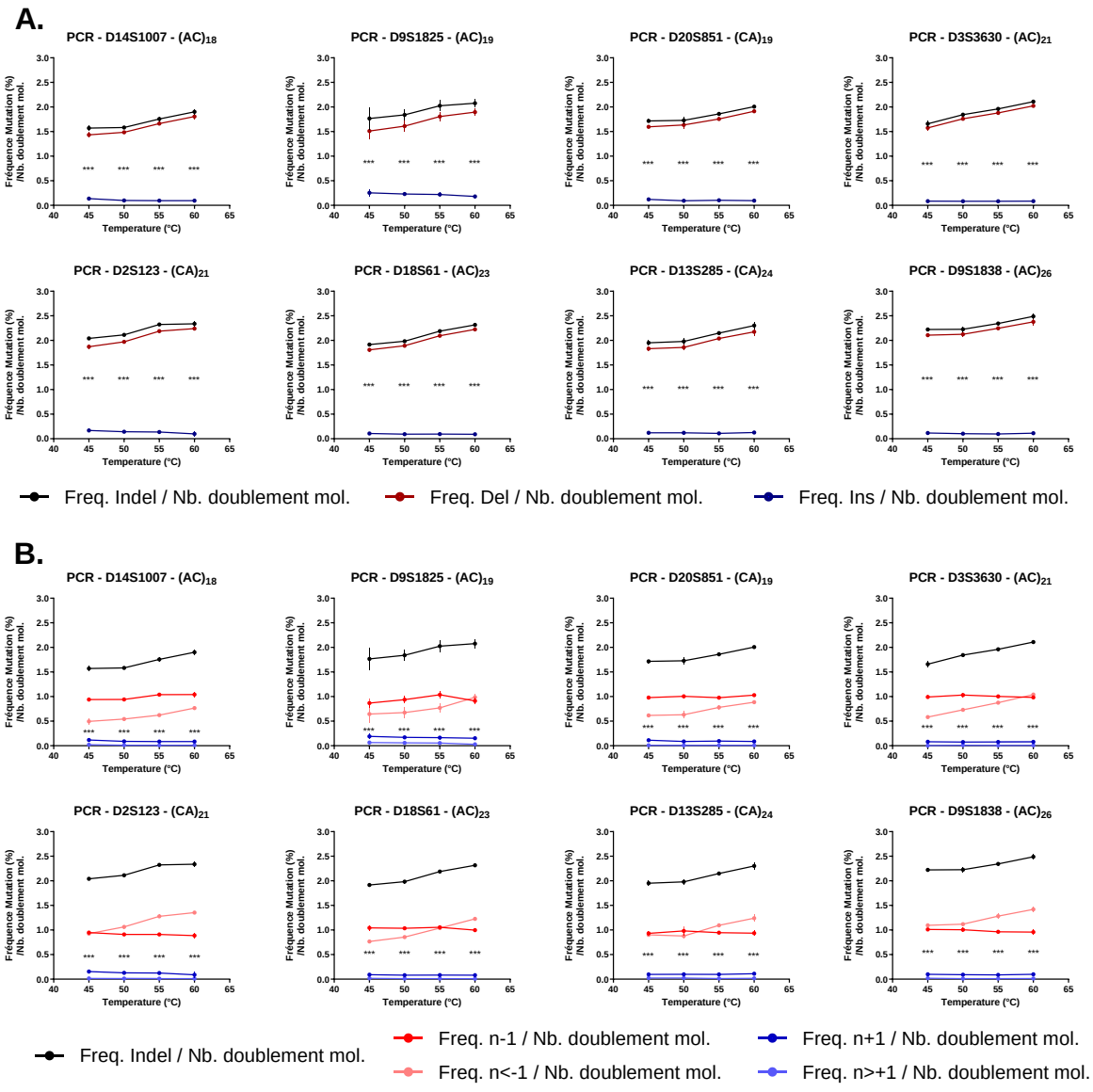

*Figure 78 : FMDM par rapport à la température obtenue par NGS sur PCR des marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair). Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

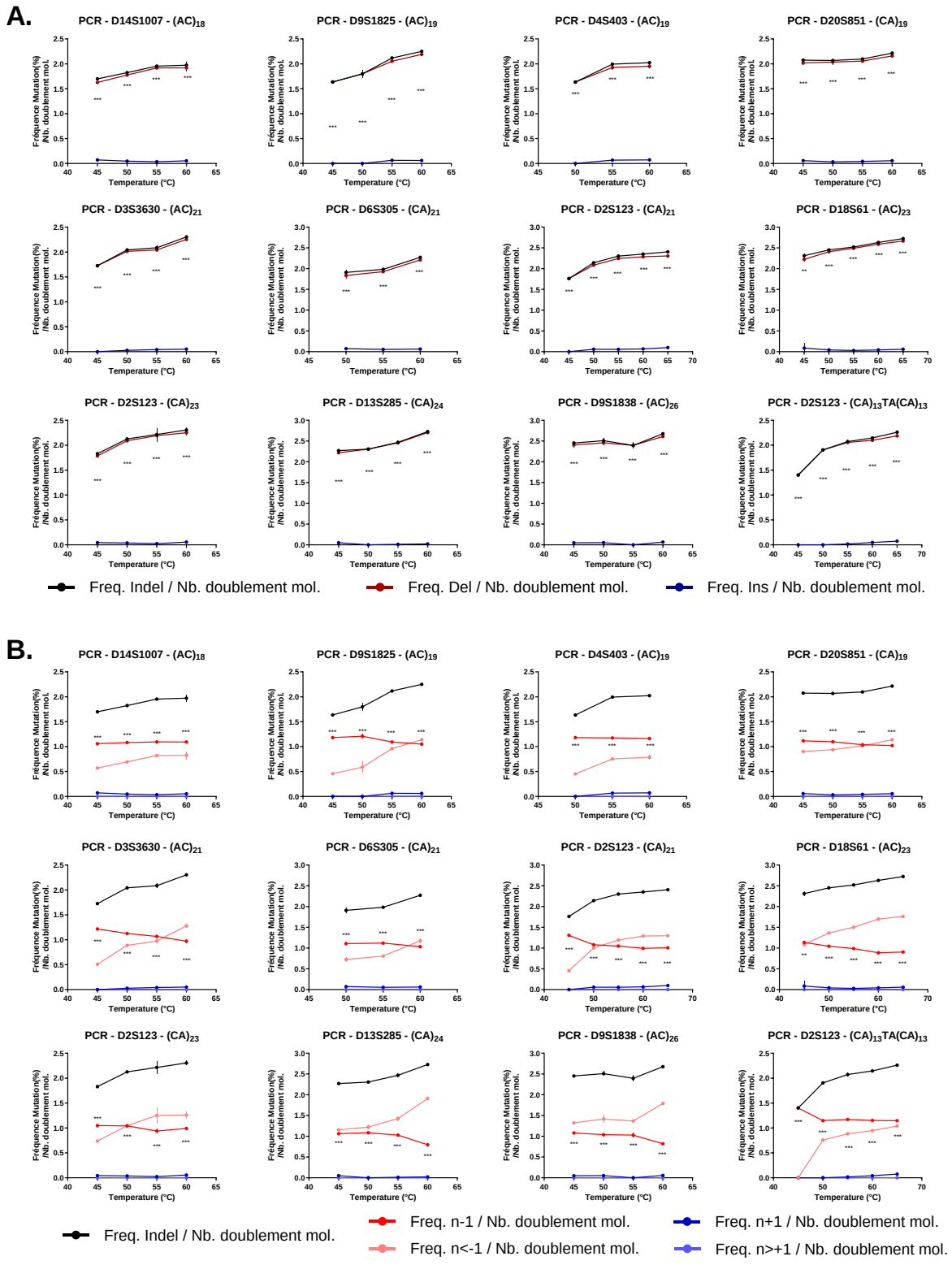

*Figure 79 : FMDM par rapport à la température obtenue par électrophorèse capillaire sur PCR des marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus d'un motif (rose), FIDM d'un motif (bleu) et FIDM de plus d'un motif (bleu clair). Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

Enfin, il est intéressant de constater que la FMDM augmente en délétion avec le nombre de répétitions (Figure 80, Figure 81, Figure 82 & Figure 83) alors que, concernant les insertions, elle suit une augmentation très légère, voire stagnante en NGS, pour les mono-nucléotides (Figure 80 & Figure 81) là où elle semble constante, voire à la baisse en NGS chez les di-nucléotides (Figure 82 & Figure 83). De plus, on observe une nette préférence pour la délétion quelle que soit la composition du motif (Figure 80, Figure 81, Figure 82 & Figure 83).

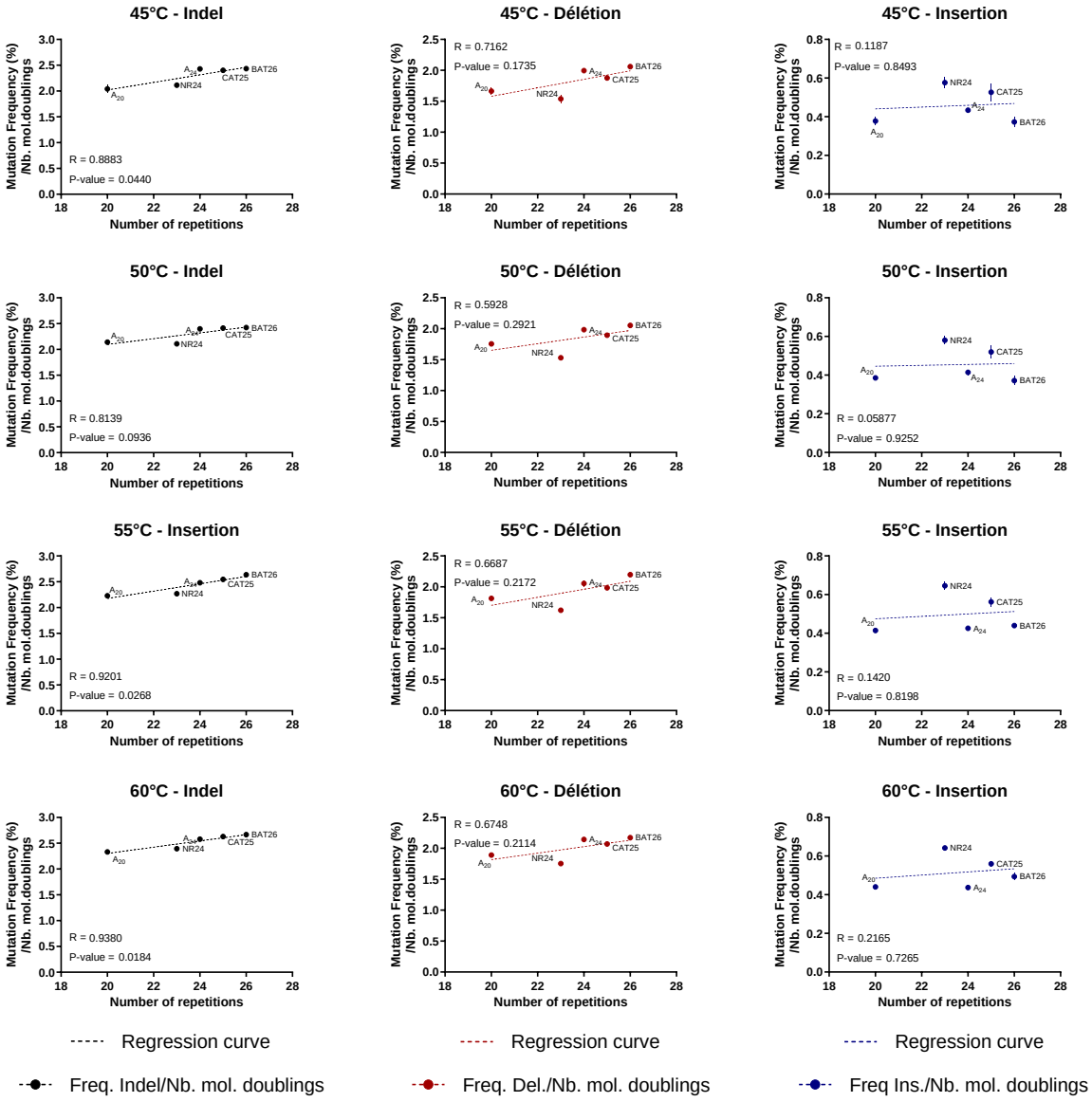

Figure 80 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et les marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 de 45°C à 65°C obtenu par NGS sur PCR avec soustraction du bruit de fond pour les plasmides. Pour chaque graphique, une courbe de régression entre les points a pu être tracé et le R calculé ainsi que la P.value.

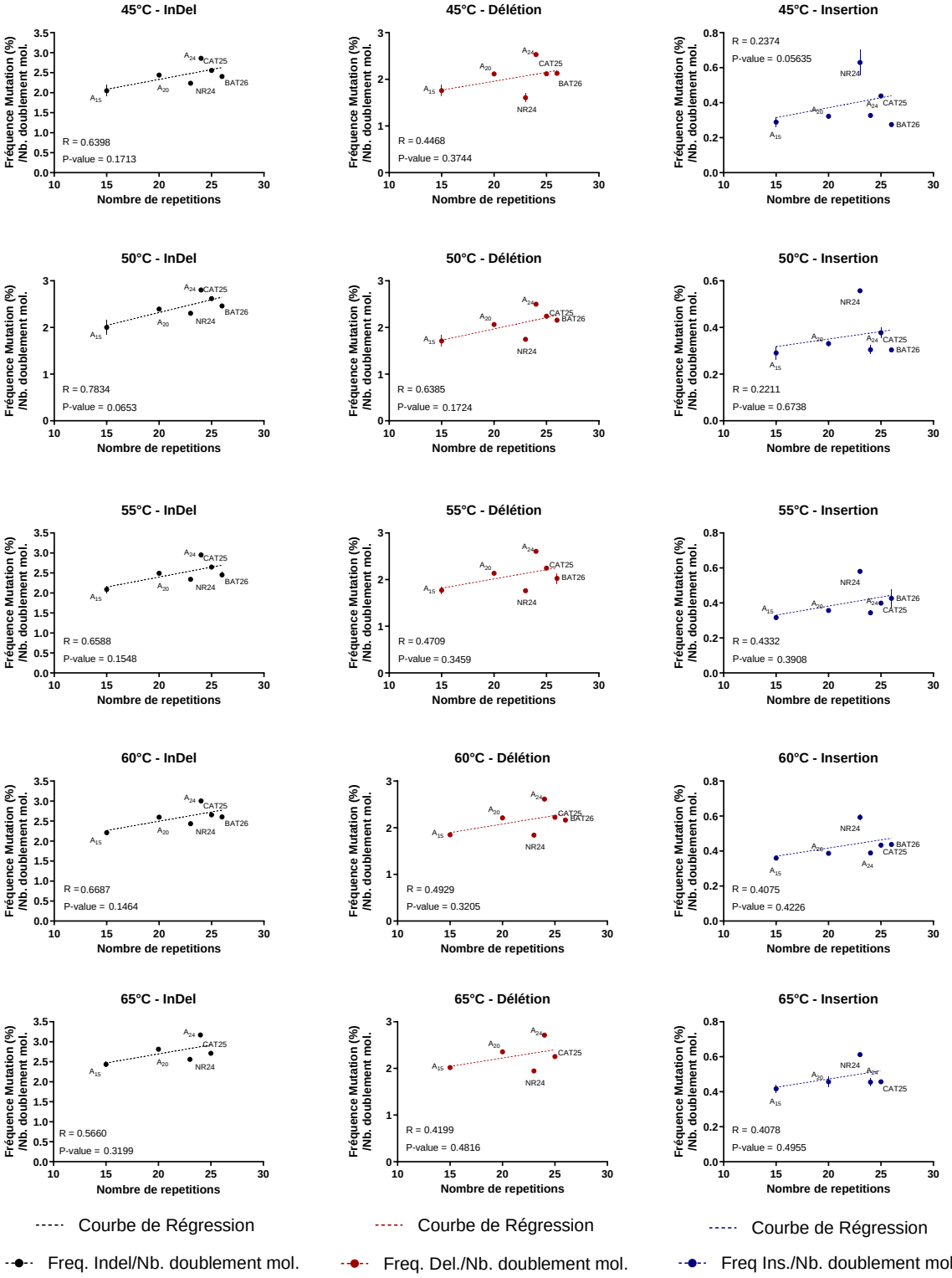

Figure 81 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites mono-nucléotidiques A_{15}, A_{20} et A_{24} et les marqueurs mono-nucléotidiques NR24, CAT25 et BAT26 de 45°C à 65°C obtenues par électrophorèse capillaire sur PCR sans soustraction du bruit de fond pour les plasmides afin d'uniformisé les données comparées. Pour chaque graphique, une courbe de régression entre les points a pu être tracée et le R calculé ainsi que la P.value.

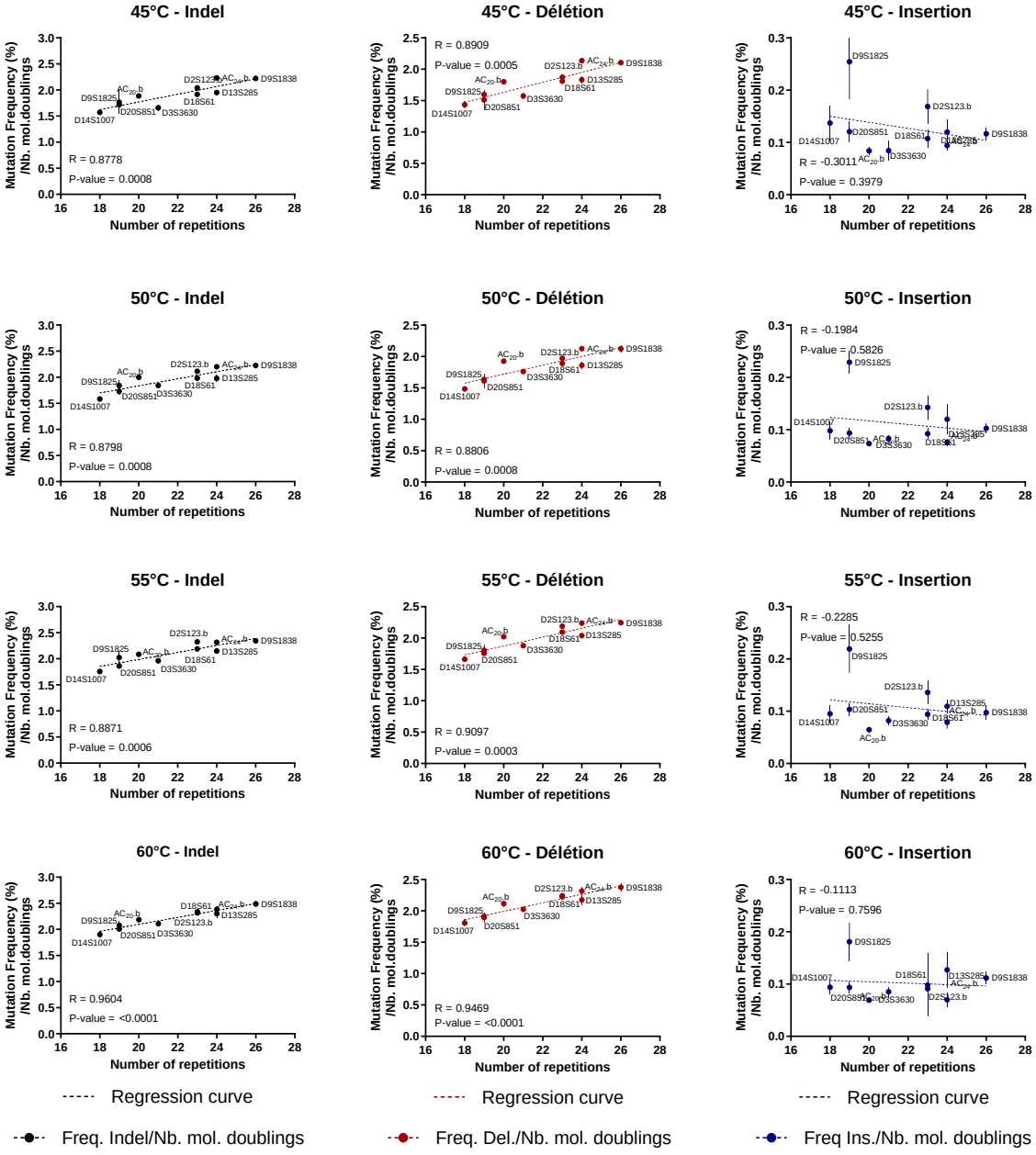

Figure 82 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites di-nucléotidiques $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$ et les marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ de 45°C à 65°C obtenues par NGS sur PCR avec soustraction du bruit de fond pour les plasmides. Pour chaque graphique, une courbe de régression entre les points a pu être tracée et le R calculé ainsi que la P.value.

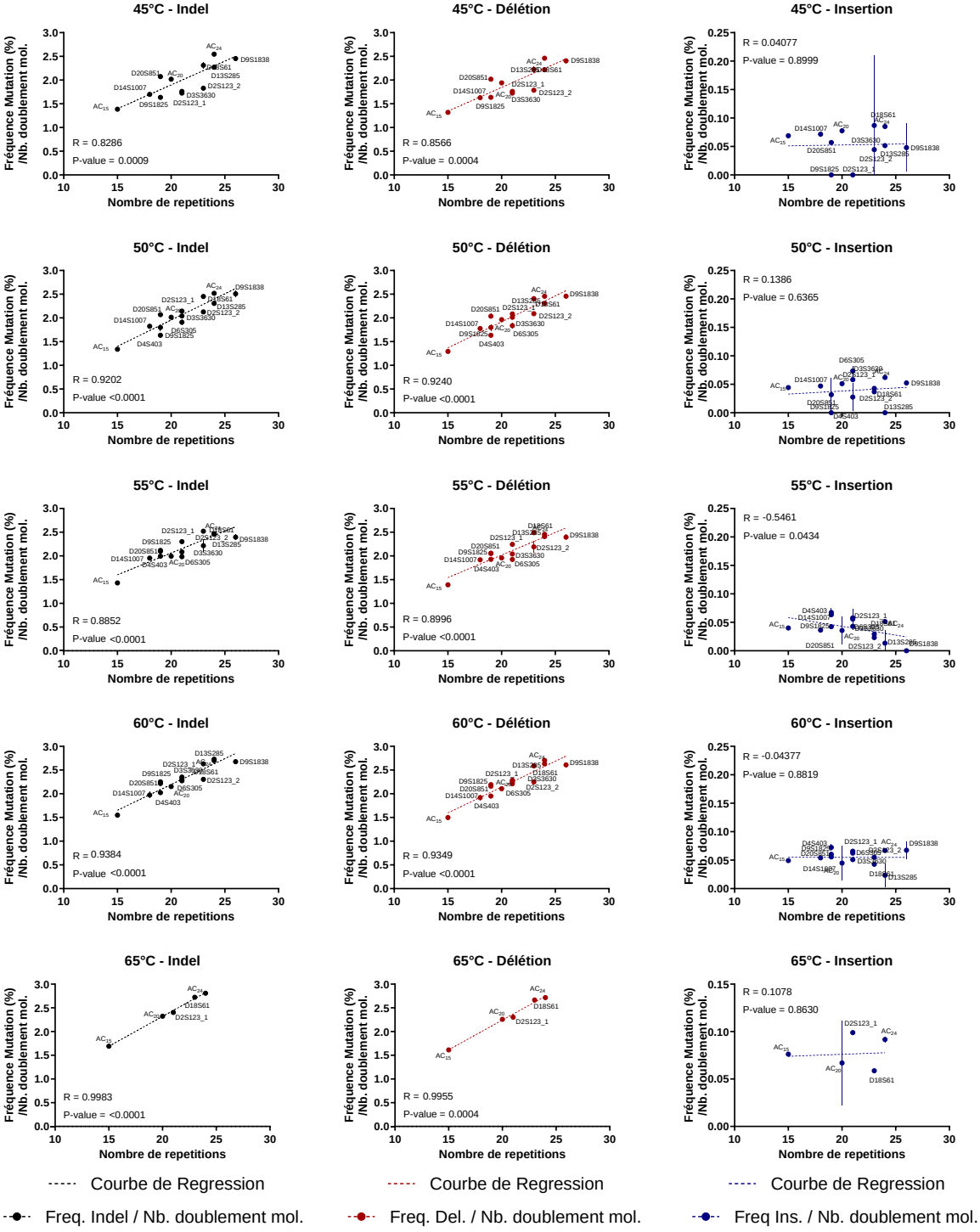

Figure 83 : FMDM par rapport au nombre de répétitions pour les Indels, les délétions et les insertions chez les microsatellites di-nucléotidiques $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$ et les marqueurs di-nucléotidiques D3S3630, D4S403, D6S305, D9S1838, D9S1825, D13S285, D14S1007, D18S61, D20S851 et D2S123 présent comme $(CA)_{21}$, $(CA)_{23}$ et $(CA)_{13}TA(CA)_{13}$ de 45°C à 65°C obtenues par électrophorèse capillaire sur PCR sans soustraction du bruit de fond pour les plasmides afin d'uniformisé les données comparées. Pour chaque graphique, une courbe de régression entre les points a pu être tracée et le R calculé ainsi que la P.value.

3.1.2. Expérimentations en contexte in vivo

3.1.2.1. Dans un organisme procaryote : Escherichia coli

Dans cette partie, j'ai voulu mesurer l'influence de la température sur la formation des mutations des microsatellites au sein d'un organisme procaryote, *Escherichia coli*. Cet organisme possédant un système de correction MMR, nous allons, en plus de la température, étudier l'impact de la défaillance de ce système sur la formation des mutations de deux types de microsatellites, A_{24} et $(AC)_{24}$. En plus d'une souche *E. coli* K-12 BW25113 Wild-type (WT), une deuxième souche, mais déficiente sur la sous-unité MutS du système MMR, et ainsi qu'une troisième, déficiente sur la sous-unité MutL du système MMR, ont été utilisées (Matériels et Méthodes 1.1.). Le protocole expérimental suivi consiste en l'insertion de plasmides dans des bactéries qui seront ensuite cultivées en milieu LB à différentes températures pendant plusieurs heures afin de permettre un nombre suffisant de réplication des plasmides, permettant ainsi aux mutations microsatellites de s'accumuler, et dont la mesure sera faite par NGS (Figure 84). Contrairement à l'étude des mutations microsatellites dans des systèmes *in vitro* (WGA, PCR et RPA), le nombre initial de plasmides insérés dans la bactérie et qui se répliquent n'est pas aussi simple à contrôler. Pour autant il est nécessaire que le nombre de plasmides transformant les bactéries et le nombre de réplication de l'ADN soient en nombre suffisant afin de générer un profil de mutation microsatellite typique, mesurable et reproductible, comme observé dans les systèmes *in vitro*, limitant ainsi les effets stochastiques et les profils atypiques, qui augmentent à mesure que le nombre de bactéries transformées et de plasmides insérés diminue. Ainsi, j'ai tout d'abord optimisé le protocole de transformation et de culture des bactéries, avant d'effectuer la production en conditions optimales.

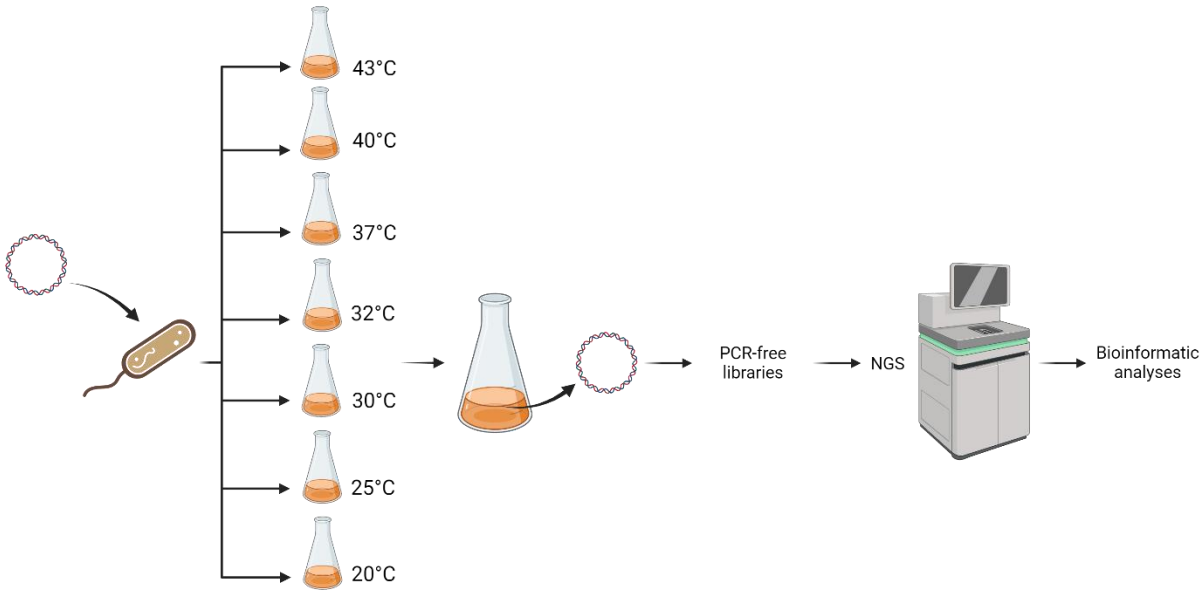

Figure 84 : Workflow Production Bactéries

3.1.2.1.1. Optimisations du protocole de transformation et de culture

Afin d'effectuer la phase de production des différentes cultures de bactéries transformées aux différentes températures d'intérêt, les conditions expérimentales ont été optimisées pour permettre une culture sur la nuit (Figure 85).

L'optimisation a tout d'abord consisté en la détermination de la bonne quantité de plasmide à utiliser pour la transformation, permettant d'avoir un nombre de plasmides initial en nombre suffisant

inséré dans les bactéries – la quantité à atteindre était d'au moins 10 000 copies - mais aussi à la bonne fenêtre de temps pour la culture. Une quantité initiale de plasmides de 30 à 70 ng ajoutée aux 200 µl de bactéries compétentes a été déterminée comme suffisante pour obtenir au moins 10 000 bactéries transformées par 1 plasmide. Ainsi, les bactéries rendues compétentes au laboratoire, en suivant le protocole présenté dans le Matériels et Méthodes 2.6.1, ont d'abord été décongelées sur glace. Entre 30 et 70 ng de plasmides ($(AC)_{24}$ ou A_{24}) ont été ajoutés à 200 µL de bactéries compétentes, suivi d'une incubation de 30 minutes sur glace. Après cette période d'incubation, la suspension a été soumise à un choc thermique à 42°C pendant 45 secondes, puis replacée sur glace pendant 5 minutes. 800 µL de milieu LB ont été ajoutés à la suspension, qui a été incubée à température d'intérêt (allant de 20 à 43°C) avec agitation à 250 rpm pendant 15 minutes permettant l'expression des gènes de résistance aux antibiotiques. J'ai utilisé pour cette étape une durée de 15 min après des tests après 0, 15, 30, 45 et 60 min (données non montrées), car elle ne permet pas aux bactéries de se diviser, ce qui fausserait la quantification des bactéries transformées et l'évaluation du nombre de divisions cellulaires réel.

Après cette incubation, 100 µL de la suspension ont été prélevés pour étalement sur les différents milieux solides utilisés (LB + Amp et LB + Amp + Kan) (Tableau 18) à une dilution allant de 1/10 à 1/10 000ème. Les boîtes ont été incubées à 37°C pendant la nuit pour permettre la croissance des colonies transformées et ainsi quantifier le nombre total de bactéries transformées, correspondant au nombre approximatif de plasmides insérés et qui vont subir des cycles de réplication, et déterminer également le taux de transformation.

Les 900 µL restants de la culture ont été inoculés dans 150 mL de milieu LB contenant les mêmes antibiotiques (LB + Amp ou LB + Amp + Kan) et incubés à des températures de 20°C, 25°C, 30°C, 32°C, 37°C, 40°C et 43°C sous agitation à 250 rpm pendant la nuit, en fonction des expériences d'optimisation et de production.

Afin d'optimiser la durée de culture et d'identifier la période optimale où effectuer le prélèvement final, nous avons effectué une cinétique de la culture en mesurant la DO à différents moments sur la souche WT et *mutS*, cultivée à différentes températures (25°C, 27°C, 40°C et 43°C). Les résultats de la cinétique montrent que le plateau est atteint après une vingtaine d'heures. Ces cinétiques, réalisées à chaque température d'étude, nous ont permis de déterminer une fenêtre d'arrêt d'environ 3 heures au bout de 15 h de cultures afin d'être en fin de phase exponentielle avec une D.O. en moyenne de 4 (Figure 85).

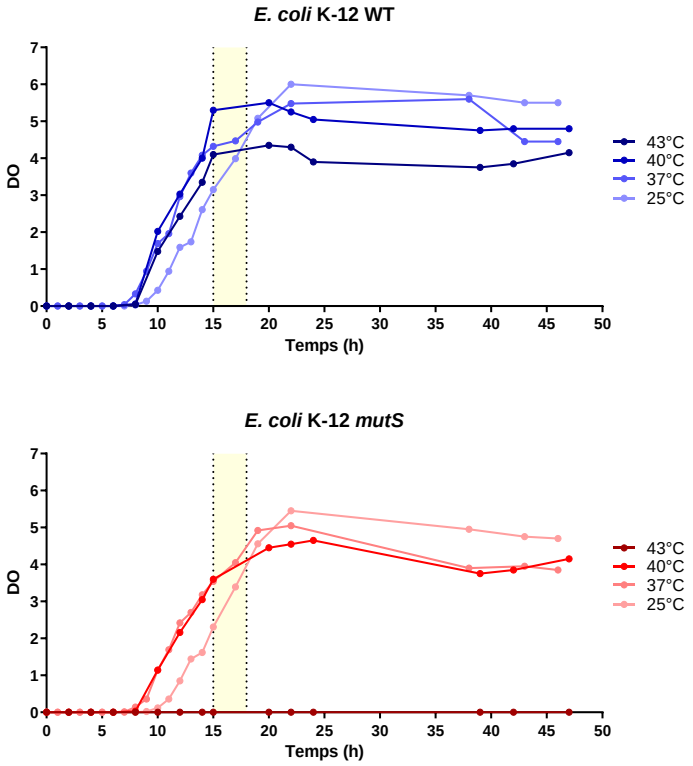

Figure 85 : Cinétique de détermination de la D.O. d'arrêt des cultures de la souche WT et du mutant mutS de E.coli pour une production optimale en restant dans la phase exponentielle (la zone beige représente le créneau d'arrêt de la culture).

Au vu des résultats, nous avons décidé que les cultures devaient être arrêtées à une D.O.600 autour de 4 au cours des phases de production, en étant mises directement à 4°C, avec le prélèvement de deux échantillons de 1,5 mL permettant d'effectuer des PCR sur culture, et trois échantillons de 50 mL pour les extractions plasmidiques. Le nombre de bactéries obtenu après la culture est calculé par mesure de DO, et confirmé par un étalement sur milieu solide d'environ 20% des cultures (LB+amp pour la souche WT et LB+amp+Kan pour les souches mutées) (Matériels et Méthodes 2.3.). Tous ces échantillons ont été centrifugés pendant 25 minutes à 4600 rpm pour recueillir le culot bactérien. De ce culot, les plasmides ont été extraits en suivant le protocole décrit dans le Matériels et Méthodes 3.1.1., purifiés par billes et analysés par NGS avec un NovaSeq 6000 (Figure 84).

Les résultats obtenus par NGS seront uniquement ceux des souches *mutS* et *mutL* car en soustrayant le bruit de fond du séquençage, des valeurs alternant autour de zéro pour les souches WT étaient obtenues, ce qui n'a pas permis d'étudier l'effet de la température sur le taux de mutation dans ces levures. Cela est dû au fait que le système MMR de ces souches, étant complètement opérationnel, corrige efficacement les erreurs de telle sorte qu'elles ne sont pas visibles par NGS avec le nombre de doublement cellulaire utilisés dans nos expériences.

3.1.2.1.2. Doublement cellulaire et doublement de molécule plasmidique

La phase de production a permis de réaliser toutes les cultures des souches transformées par les deux plasmides à toutes les températures à l'aide des protocoles optimisés de transformation et de

culture. A la fin de la production, les cultures ont bel et bien été arrêtées dans les conditions optimales que nous avions définies (Figure 86.A).

Nous avons pu déterminer que la quantité médiane de bactéries transformées avant la culture (T_0) était de 3 x 10^4, quelle que soit la souche ou le plasmide qui a permis la transformation (Figure 86.B). Toutefois, nous observons que le nombre de bactéries transformées après cultures (T_f) n'est pas le même selon la souche utilisée avec une valeur médiane de 1 x 10^{12} pour les WT, 9 x 10^{11} pour les mutS et 7 x 10^{11} pour les *mutL*, quel que soit le plasmide utilisé (Figure 86.C). La connaissance du nombre de bactéries initialement transformées et obtenues après culture nous a permis de calculer le nombre de doublement de bactéries au cours de la culture, déterminé par la formule : $log_2(\frac{T_f}{T_0})$. Malgré la légère différence du T_f, le nombre de doublement de bactéries calculé médian est de 25 pour toutes les conditions (Figure 86.D).

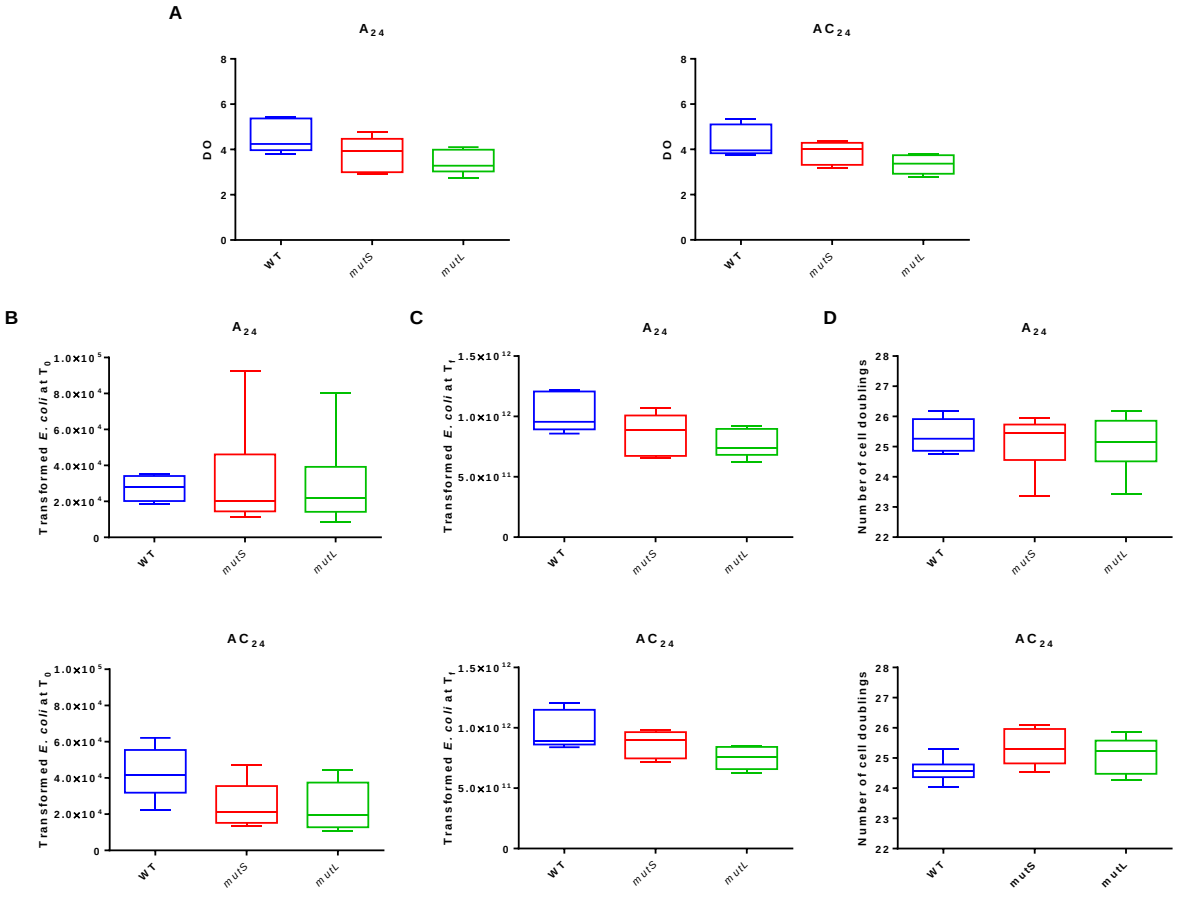

Figure 86 : Evolution de la quantité des bactéries transformées durant le processus de production. A. Valeurs de D.O. au moment de l'arrêt de la culture en fonction des souches transformés et du plasmide de transformation. B. Nombre de bactéries E. coli transformées à T0 selon la souche et le plasmide utilisées. C. Nombre de bactéries E. coli transformées à Tf selon la souche et le plasmide utilisées. D. Nombre de doublement de bactéries selon la souche et le plasmide utilisées.

Le nombre moyen de copies de plasmides par bactérie a pu également être déterminé par PCR en temps réel, qui a été effectuée directement sur un aliquote de la culture bactérienne en utilisant deux essais ciblant le plasmide (essai HT17 145) et un gène en copie unique dans le génome bactérien (gène *dnaC*). Les deux essais PCR ayant une efficacité très proche de 100%, le nombre moyen de plasmides

par bactérie a été calculé par la formule : $C_{t(HT17)} - C_{t(dnaC)}$. Ainsi, il est à noter que le nombre de copies de plasmide au sein des bactéries est relativement faible entre 20 et 32°C puis augmente avec la température avec une accélération vers 35°C pour toutes les souches et les plasmides (Figure 87.A).

Pour finir, le nombre de doublement de molécules a pu également être déterminé. Il a été défini comme le nombre de doublement cellulaire auquel est ajouté le nombre moyen de copies de plasmide par bactérie. Il est intéressant de noter que le nombre de doublement de molécules au sein des bactéries augmente avec la température avec une décélération et même une diminution à partir de 40°C pour toutes les souches et les plasmides (Figure 87.B). Le doublement de molécule médian obtenu est de 32 avec des variations allant de 30 à 34 quelle que soit la souche ou le plasmide étudié (Figure 87.C).

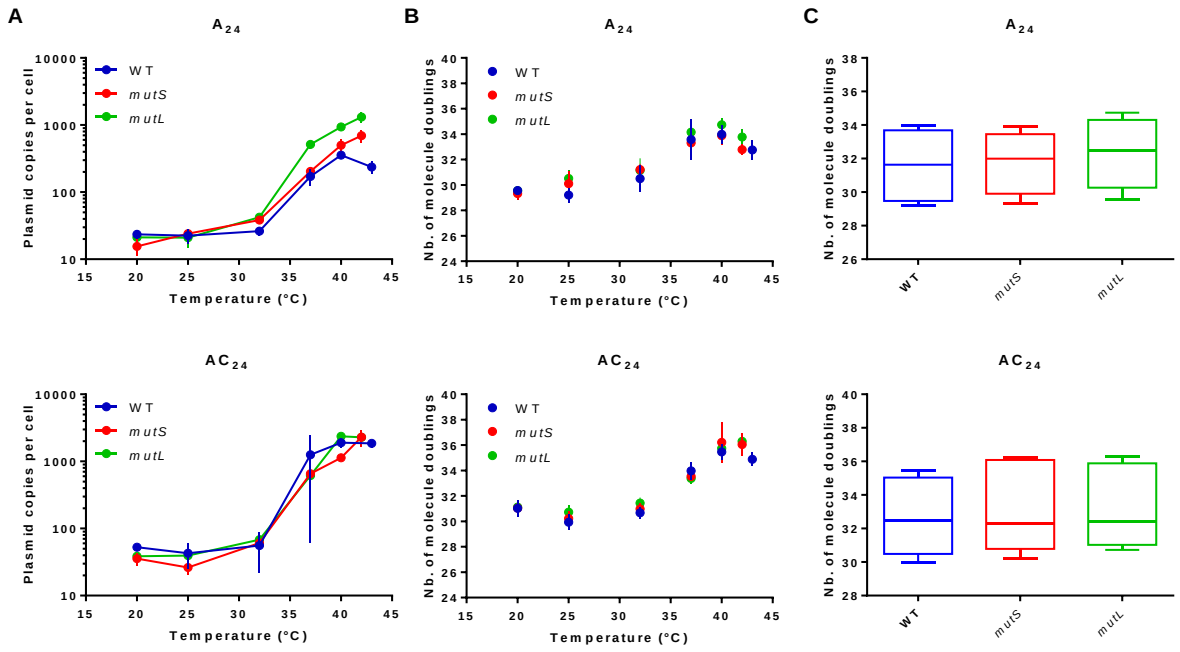

Figure 87 : Evolution de la représentation des plasmides au sein des bactéries transformées. A. Nombre de copies de plasmides par bactérie selon la température, la souche et le plasmide utilisés. B. Nombre de doublement de molécule selon la température, la souche et le plasmide utilisés. C. Nombre de doublement de molécule selon la souche et le plasmide utilisés.

3.1.2.1.3. Impact de la température sur les mutations microsatellites

Les résultats précédents ont permis de vérifier que les cultures se sont déroulées de manière homogène, et de déterminer le nombre de doublement de molécules. Suffisamment d'ADN a été obtenu pour effectuer l'étape suivante consistant à la préparation des librairies et le séquençage sur NovaSeq 6000. Les données de séquençage ainsi recueillies nous ont permis alors de pouvoir calculer les FMDM pour chaque température, souche et microsatellite et les visualiser en comparant la totalité des indels, les insertions, les délétions, d'un motif ou plus (Figure 88). Comme pour les expériences *in vitro*, des plasmides à T_0 ont également pu être séquencés afin de pouvoir normaliser les fréquences alléliques en leur retirant le bruit de fond et ainsi mesurer uniquement les mutations induites par paramètres étudiés.

Si l'on observe les différentes FMDM obtenues dans les conditions étudiées, on observe que quelle que soit la souche ou le motif, la FMDM augmente avec la température avec une courbe semblable à une sigmoïde (Figure 88.A). De plus, avec la sous-unité *mutS* ou *mutL* défaillante, les

mutations favorisent systématiquement les délétions principalement à un motif (Figure 88.B). Toutefois, il est intéressant de noter que la FMDM diminue de moitié entre les microsatellites A_{24} et $(AC)_{24}$.

On peut donc ici conclure que, pour un organisme procaryote tel que *E. coli*, la déficience d'une des sous-unités de son système MMR par rapport à l'autre ne change pas la directionnalité de la mutation. De plus, il est ici visible que pour un même nombre de répétitions la simplicité du motif A favorise la mutation contrairement à un motif di-nucléotidique tel que AC.

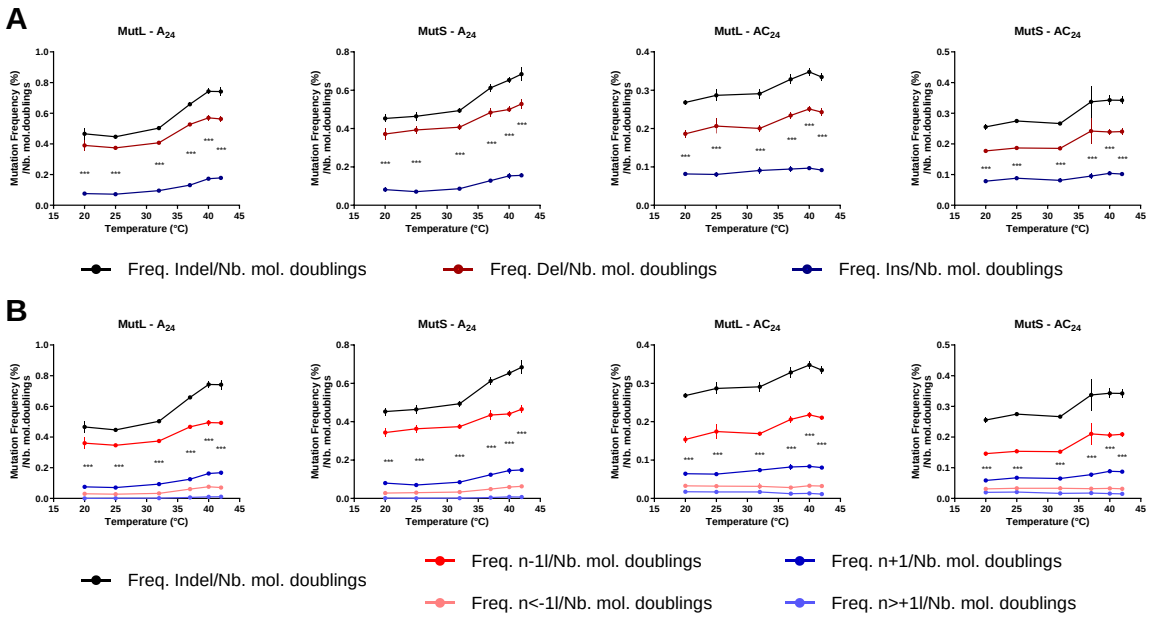

*Figure 88 : FMDM par rapport à la température de culture des différentes souches étudiées. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les souches mutL et mutS transformé par le plasmide A_{24} et $(AC)_{24}$. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus de un motif (rose), FIDM d'un motif (bleu) et FIDM de plus de un motif (bleu clair) pour les souches mutL et mutS transformé par le plasmide A_{24} et $(AC)_{24}$. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001*

3.1.2.2. Dans un organisme eucaryote : Saccharomyces cerevisiae

Dans cette partie de l'étude, j'ai voulu également mesurer l'influence de la température sur la formation des mutations des microsatellites dans un organisme eucaryote, *Saccharomyces cerevisiae*. Comme pour *Escherichia coli*, cet organisme possède un système de correction MMR. L'impact potentiel de la défaillance de ce système sur la formation de mutations a pu être étudié avec deux types de microsatellites, A_{24} et $(AC)_{24}$. La souche WT BY4741 de la Yeast Knockout (YKO) Collection provenant d'une lignée haploïde hétérothallique S288C, a été utilisée ainsi que deux souches mutées et déficientes soit sur la sous-unité *mlh1* de *mutL*, soit sur la sous-unité *msh2* de *mutS* du système MMR (Matériels et Méthodes 1.2.). De plus, *Saccharomyces cerevisiae* possédant un génome contenant des microsatellites, une étude de son propre génome a été par la suite réalisée.

3.1.2.2.1. Etudes plasmidiques

Le protocole expérimental suivi consiste en l'insertion de plasmides dans une levure qui seront ensuite cultivés en milieu YNB sans uracile à différentes températures pendant plusieurs heures afin de permettre un nombre suffisant de réplication des plasmides, permettant ainsi aux mutations

microsatellites de s'accumuler, et dont la mesure sera faite par NGS (Figure 89). Comme pour les bactéries, le nombre initial de plasmides insérés dans la levure et qui se répliquent n'est pas aussi simple à contrôler. Pour autant il est nécessaire que le nombre de plasmides transformant les levures et le nombre de réplication de l'ADN soient en nombre suffisant afin de générer un profil de mutation microsatellite typique, mesurable et reproductible, comme observé dans les systèmes *in vitro*, limitant ainsi les effets stochastiques et les profils atypiques, qui augmentent à mesure que le nombre de plasmides insérés diminue. Contrairement aux bactéries, il a été nécessaire d'inclure une étape d'amplification des plasmides extraits de levures dans *E. coli* (DH5α), car ils n'étaient pas directement utilisables pour la réalisation des librairies NGS. Comme précédemment, j'ai d'abord optimisé le protocole de transformation et de culture des levures, avant d'effectuer la production en conditions optimales.

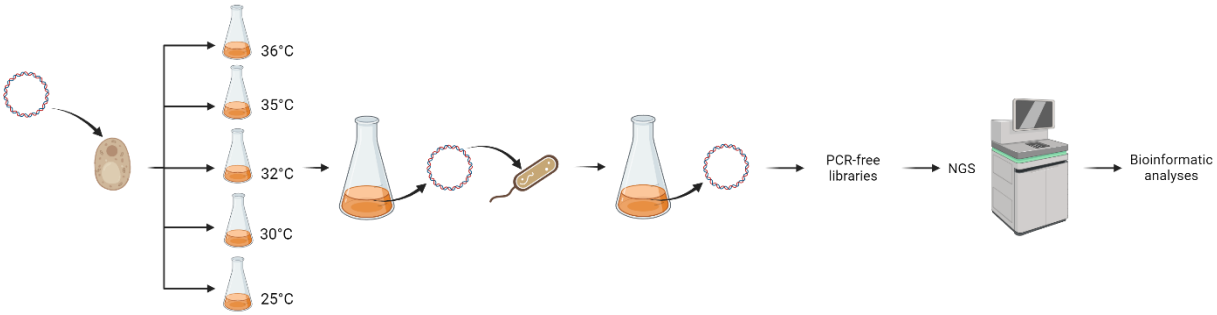

Figure 89 : Workflow Production Levures Plasmides

3.1.2.2.1.1. Optimisation du protocole de transformation et de culture

Les conditions expérimentales ont été optimisées pour permettre une culture sur plusieurs jours avec une récupération des cultures durant la journée (Figure 90). Ces conditions optimisées ont été ensuite appliquées à la phase de production.

L'optimisation a tout d'abord consisté en la détermination de la bonne quantité de plasmide à utiliser pour la transformation permettant d'avoir un nombre de plasmides initial en nombre suffisant inséré dans les levures – la quantité à atteindre était d'au moins 10 000 copies – mais aussi à la bonne fenêtre de temps pour la culture. Une quantité initiale de plasmides de 1 µg ajoutés aux 100 µL de levures compétentes a été déterminée comme suffisante pour obtenir au moins 10 000 levures transformées par 1 plasmide. Ainsi les levures rendues compétentes au laboratoire et transformées en suivant le protocole présenté dans le Matériels et Méthodes 2.6.2. et 2.7.2. , ont d'abord été décongelées sur glace. 10 µL de plasmides à 100 ng/µL de plasmides ($(AC)_{24}$ ou A_{24}) ont été ajoutés à 100 µL de levures compétentes, et 1 mL de Frozen-EZ Yeast Solution 3 a été ajouté au mélange, qui a été mélangé énergiquement. Le mélange a ensuite été incubé entre 25°C et 37°C pendant 45 minutes (après avoir testé 15, 30 et 45 minutes), durée qui ne permet pas à la levure de se diviser. Durant cette incubation, le mélange a été vortexé 2 à 3 fois pour assurer une bonne distribution des plasmides dans les cellules.

Après l'incubation, 100 µL d'une dilution au 1/500ème de la suspension transformée ont été étalés sur des boîtes de Petri contenant du milieu YNB avec et sans uracile (YNB Ura-), supplémenté en généticine (Gen+) pour les souches mutantes, puis incubés pendant 2 à 4 jours pour permettre la croissance des colonies transformées et ainsi déterminer le taux de transformation (compris entre 0,13 et 1,15 %).

Le volume restant de la transformation a été inoculé dans 150 mL du même milieu (YNB Ura-Gen+/-) et incubé à des températures de 25°C, 30°C, 32°C, 34°C et 36°C en fonction des expériences d'optimisation ou de productions.

Afin d'optimiser la durée de culture et d'identifier la période optimale où effectuer le prélèvement final, une cinétique de la culture a été effectuée en mesurant la DO à différents moments sur les trois souches (WT, *mlh1* et *msh2*), cultivées à différentes températures (25°C, 30°C et 37°C). Les résultats de la cinétique montrent que le plateau est atteint au bout d'une cinquantaine d'heures, à l'exception des souches à 37°C où les cultures étaient très lentes. Cette température a été écartée pour la suite des expériences. Ces cinétiques ont permis de déterminer une fenêtre d'arrêt entre 3 et 5 heures au bout de 1 à 2 jours de cultures afin d'être en fin de phase exponentielle avec une D.O. cible en moyenne de 4 pour les phases de production (Figure 90). Le nombre de levures obtenu à l'arrêt des cultures a été mesuré par la mesure DO et confirmé par étalement en milieu solide pour 20% des cultures (YNB Ura- pour la souche WT et YNB Ura- +Gen pour les souches mutées).

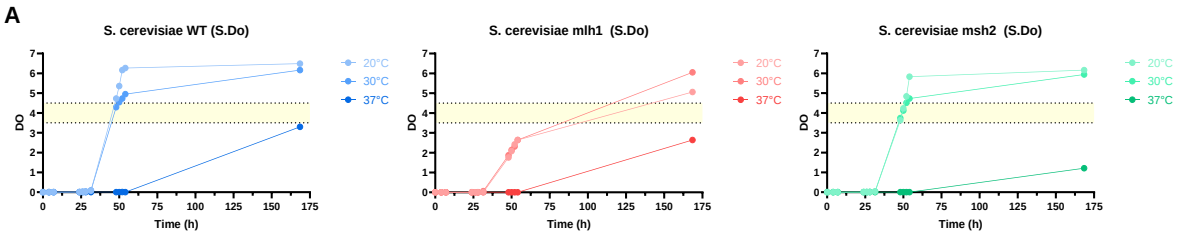

Figure 90 : Cinétique de détermination de la D.O. d'arrêt des cultures de la souches WT, et des mutant mlh1 et msh2 de S. cerevisiae pour une production optimale en restant dans la phase exponentielle (la zone beige représente la fourchette de D.O. d'arrêt de la culture)

Après arrêt des cultures, plusieurs aliquots ont été préparés, dont 3 de 50 ml centrifugés pendant 25 minutes à 46000 rpm pour recueillir les cellules transformées et 2 aliquotes de 1,5 ml pour la détermination du nombre de copies de plasmide par cellule par PCR en temps réel. Contrairement aux bactéries, les PCR directes sur les levures ne sont pas efficaces, et une co-extraction préalable d'ADN génomique et de plasmide a été nécessaire, effectuée à l'aide du Kit YeaStar Genomic DNA Kit sur 1,5 ml de culture. Par la suite, les plasmides ont été extraits en suivant le protocole d'extraction présenté dans le Matériels et Méthodes 3.1.2. Les plasmides extraits de levure donnant en quantité principalement de l'ADN génomique et peu d'ADN plasmidique pouvant être utilisés pour la préparation des librairies, il a été nécessaire de l'amplifier en bactérie. Pour un meilleur rendement et plus de reproductibilité, des bactéries compétentes commerciales (NEB, DH5-α) ont été utilisées.

Ce protocole a été aussi adapté afin i) de pouvoir mesurer le nombre de transformants ayant intégré un plasmide et ii) d'avoir un nombre de transformants initial le plus important possible (>1000), permettant d'avoir un profil représentatif du profil de mutation microsatellite généré dans la levure et limitant le nombre de réplication de l'ADN dans la bactérie. La modification du protocole inclut la réduction à 15 min du temps d'incubation dans du SOC (*Super Optimal broth with Catabolite répression*), ne permettant pas de division des bactéries, ce qui permet de connaître le nombre de bactéries ayant intégré un plasmide.

A 15 minutes d'incubation, 100 µL de la suspension ont été prélevés pour étalement sur milieux LB + Amp (Tableau 18) à une dilution de 1/10ème. Les boîtes ont été incubées à 37°C pendant la nuit pour permettre la croissance des colonies transformées et ainsi déterminer le taux de transformation.

Après l'incubation, les 900 µL restants de la culture ont été inoculés dans 150 mL de milieu LB contenant le même antibiotique (LB + Amp) et incubés à 37°C sous agitation à 250 rpm pendant la nuit. La culture a été arrêtée à une D.O.600 autour de 4. Enfin, la culture a été séparée en deux. Deux échantillons de 1,5 mL ont été prélevés pour effectuer des PCR sur culture, et deux échantillons de 50 mL ont été prélevés pour des extractions plasmidiques. Tous ces échantillons ont été centrifugés pendant 25 minutes à 4600 rpm pour recueillir le culot bactérien. De ce culot, les plasmides ont été extraits en suivant le protocole décrit dans le Matériels et Méthodes 3.1.4. purifiés par billes et analysés par NGS avec un NovaSeq 6000 (Figure 89).

Les résultats obtenus par NGS seront uniquement ceux des souches *mlh1* et *msh2* car en soustrayant le bruit de fond du séquençage, des valeurs alternant aux alentours de zéro pour les souches WT étaient obtenues, ce qui n'a pas permis d'étudier l'effet de la température sur le taux de mutation dans ces levures. Cela est dû au fait que le système MMR de ces souches, étant complètement opérationnel, corrige efficacement les erreurs de telle sorte qu'elles ne sont pas visibles par NGS avec le nombre de doublement cellulaire correspondant à nos expériences.

3.1.2.2.1.2. Doublement cellulaire et doublement de molécule plasmidique

La phase de production a été effectuée sur toutes les souches transformées par les deux plasmides et à toutes les températures. A la fin de la production, les cultures ont pu être arrêtées dans les conditions optimales que nous avions définies (Figure 91.A). Cependant, à certaines températures, aucune croissance de levure n'a été observée et ces points ont été enlevés de l'étude.

Après vérification par étalage à différentes étapes sur boîte de pétri (comme indiqué précédemment), nous avons déterminé qu'en moyenne nous avions 1,3 x 10^4 levures transformées avant la culture (T_0) quelle que soit la souche pour A_{24} et 8 x 10^3 pour $(AC)_{24}$ qui ont permis la transformation (Figure 91.B). Nous observons également que le nombre de levures transformées après cultures (T_f) est en moyenne de 2,2 x 10^{10} pour toutes les souches et plasmides sauf pour ma souche *msh2* avec A_{24} qui est à 3,5 x 10^{10} en moyenne (Figure 91.C). Enfin, malgré cette légère différence du T_f, le nombre de doublement cellulaire des levures est en moyenne à 21 pour toutes les conditions (Figure 91.D).

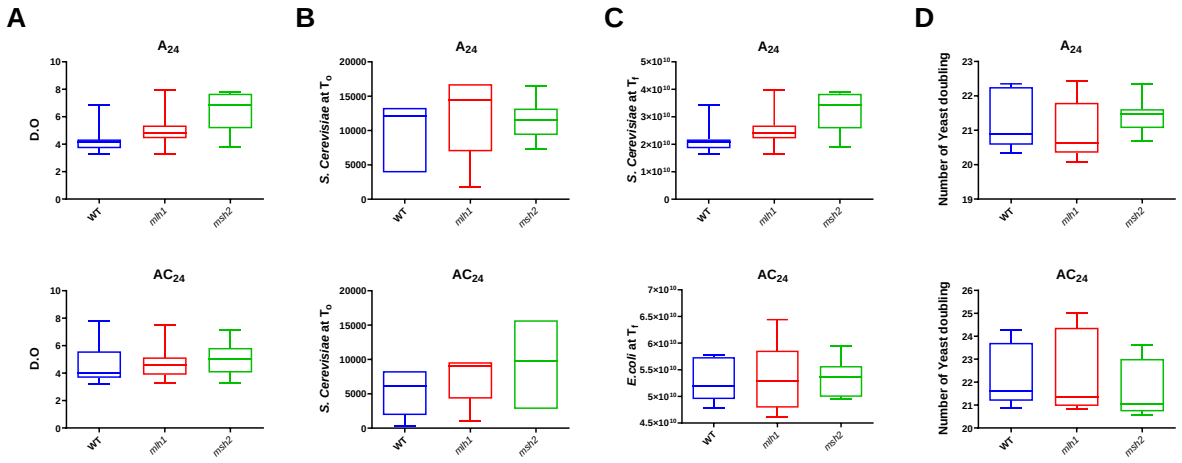

Figure 91 : Evolution de la quantité des levures transformées durant le processus de production. A. Valeurs de D.O. au moment de l'arrêt de la culture en fonction des souches transformées et du plasmide de transformation. B. Nombre de levures S. cerevisiae transformées à T0 selon la souche et le plasmide utilisées. C. Nombre de levures S. cerevisiae transformées à Tf selon la souche et le plasmide utilisées. D. Nombre de doublement de levures selon la souche et le plasmide utilisées.

Le nombre moyen de copies de plasmides par levure a pu également être déterminé par PCR en temps réel, qui a été effectuée en utilisant deux essais ciblant le plasmide (essai HT17-145) et un gène en copie unique dans le génome de levure (gène *RPL21A*). Concernant le nombre moyen de plasmides par levure, il est à noter qu'il est relativement faible mais reste constant aux alentours de 30 pour toutes les souches et les plasmides (Figure 92.A). Le nombre de doublement de molécule a pu également être déterminé et il est intéressant de noter que le nombre de doublement de molécule au sein des levures augmente avec la température pour le plasmide $(AC)_{24}$ alors qu'il est irrégulier pour A_{24} selon les températures (Figure 92.B). Le doublement de molécule médian obtenu est de 28,5 avec des variations allant de 27 à 34, quelle que soit la souche ou le plasmide étudié (Figure 92.C).

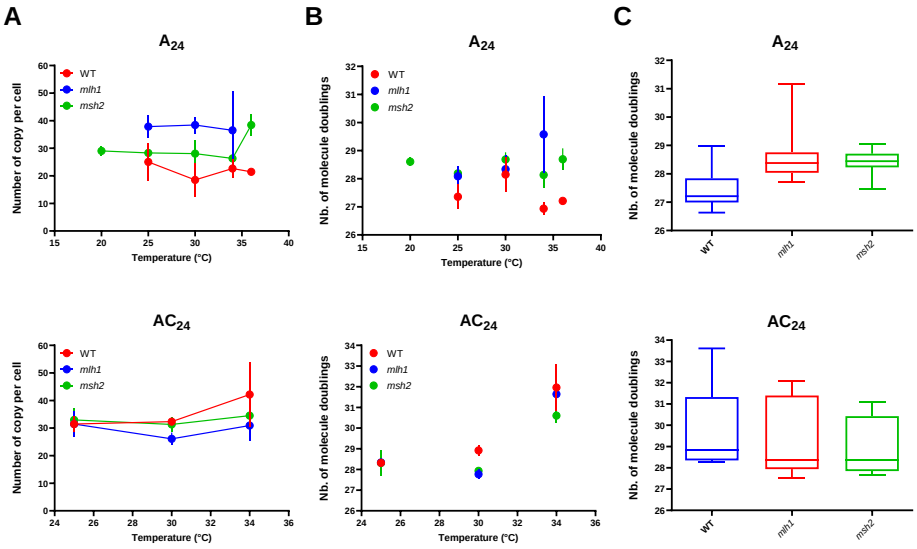

Figure 92 : Evolution de la représentation des plasmides au sein des levures transformées. A. Nombre de copies de plasmides par levure selon la température, la souche et le plasmide utilisés. B. Nombre de doublement de molécule selon la température, la souche et le plasmide utilisés. C. Nombre de doublement de molécule selon la souche et le plasmide utilisés.

En ce qui concerne la phase d'amplification du plasmide isolé de la culture de levure dans la bactérie DH5α, les données sont également homogènes, et montrent qu'en moyenne 2 500, 5 000 et 7 500 bactéries ont été transformées, et 5.10^{10}, $5{,}4.10^{10}$ et 5.10^{10} bactéries ont été obtenues après cultures dans les souches WT, *mlh1* et *msh2* respectivement (Figure 93.B & C). Le nombre de copies de plasmide par bactérie a aussi été mesuré par PCR sur bactérie. Ajouté au nombre de doublement cellulaire, un nombre moyen de doublement de molécule de 30 a été effectué dans la bactérie (Figure 93.F & G). Cependant, étant donné que cette bactérie présente un système MMR compétent, les mutations microsatellites introduites dans cette phase de culture devraient être négligeables, et ce nombre de doublement de molécule ne sera pas pris en compte pour le calcul des fréquences alléliques par nombre de doublement de molécule.

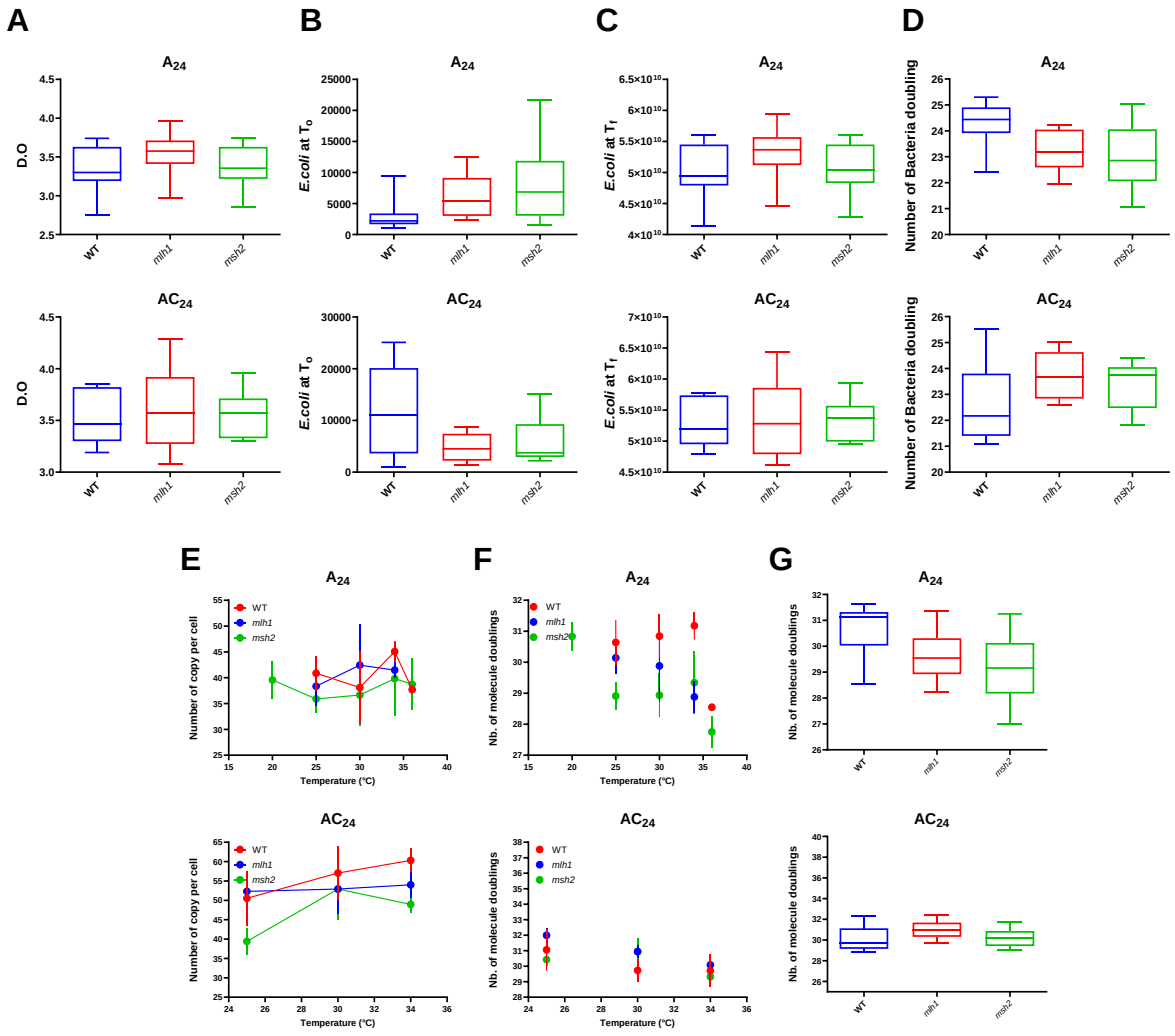

Figure 93 : Evolution de la quantité des bactéries transformées durant le processus de production. A. Valeurs de D.O. au moment de l'arrêt de la culture en fonction des souches d'origine du plasmide et du plasmide de transformation. B. Nombre de bactéries E.coli transformées à T_0 selon la souche d'origine et le plasmide utilisés. C. Nombre de bactéries E.coli transformées à T_f selon la souche d'origine et le plasmide utilisés. D. Nombre de doublement de bactéries selon la souche d'origine et le plasmide utilisés. E. Nombre de copies de plasmides par bactéries selon la température, la souche d'origine et le plasmide utilisés. F. Nombre de doublement de molécule selon la température, la souche d'origine et le plasmide utilisés. G. Nombre de doublement de molécule selon la souche d'origine et le plasmide utilisés.

L'ADN plasmidique extrait a pu être utilisé pour la préparation des librairies, suivi d'un séquençage sur NovaSeq 6000.

3.1.2.2.1.3. Impact de la température sur les mutations microsatellites

Comme pour les bactéries, les données de séquençage ainsi recueillies nous ont permis alors de calculer les FMDM pour chaque température, souche et microsatellite et de les visualiser en comparant la totalité des indels, des insertions, des délétions, d'un motif ou plus (Figure 94). Les plasmides à T_0 ont également été séquencés, afin de pouvoir normaliser les FMDM en retirant le bruit de fond des mutations induit par le séquençage et ainsi ne garder que les mutations propres aux paramètres étudiés. Comme pour les bactéries, les résultats obtenus par NGS montreront uniquement les valeurs pour les

souches *mlh1* et *msh2* car en soustrayant le bruit de fond du séquençage, nous retombions sur des valeurs alternant aux alentours de zéro pour les souches WT du fait d'un système MMR proficient et ne générant pas assez d'erreurs et de mutations pour qu'elles soient visibles au-dessus du bruit de fond, au nombre de doublement de molécule appliquée, induisant alors une perte de mesurabilité en NGS. Ces résultats permettent aussi de confirmer que l'étape d'amplification des plasmides extraits de levure dans les bactéries DH5-α n'a pas modifié de manière détectable le profil des microsatellites des plasmides obtenus dans les levures.

Enfin il est important de noter que nous avons subi les « caprices » des souches à certaines températures au cours de nos expériences, induisant des absences de résultats à 20°C notamment.

Ainsi, on observe au travers des conditions étudiées, que quelle que soit la souche ou le motif, la FMDM augmente avec la température, avec une tendance de type exponentielle qui est bien visible pour les configurations avec le plus de points, A_{24} *msh2* (Figure 94.A). De plus, avec la sous-unité *mlh1* ou *msh2* défaillante, les mutations favorisent systématiquement les délétions principalement à un motif (Figure 94.B). Toutefois, il est intéressant de noter une plus faible fréquence de mutations avec msh2 par rapport à mlh1 que ce soit chez A_{24} ou $(AC)_{24}$.

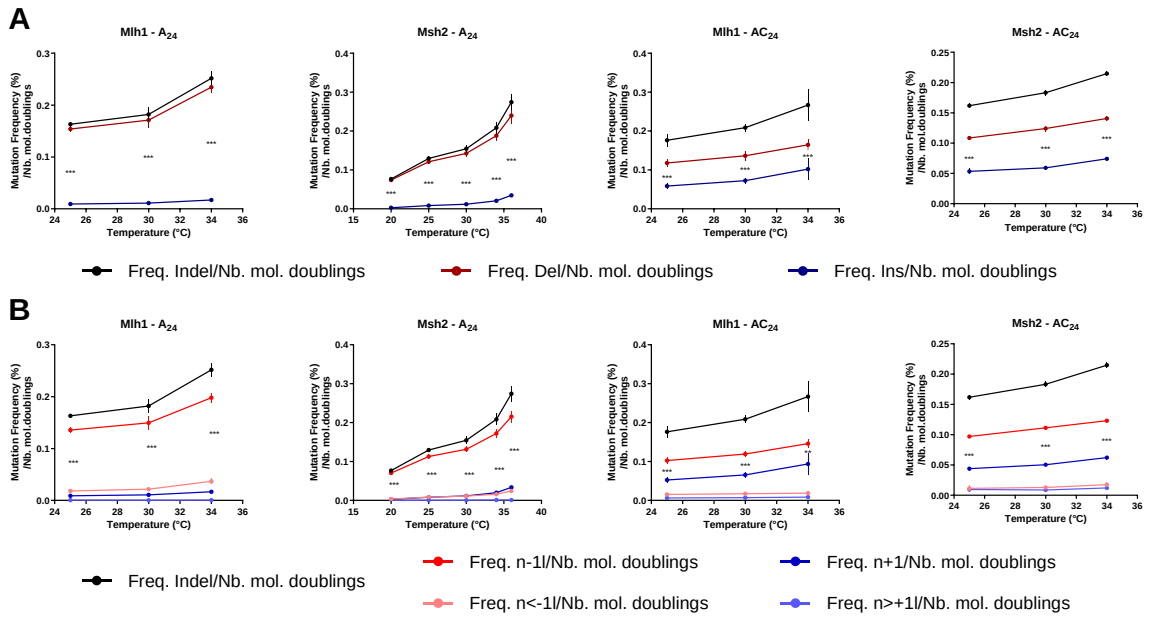

*Figure 94 : FMDM par rapport à la température de culture des différentes souches étudiées. A. FIDDM (noir), FDDM (rouge) et FIDM (bleu) total pour les souches mlh1 et msh2 transformé par le plasmide A_{24} et $(AC)_{24}$. B. FIDDM total (noir), FDDM d'un motif (rouge), FDDM de plus de un motif (rose), FIDM d'un motif (bleu) et FIDM de plus de un motif (bleu clair) pour les souches mlh1 et msh2 transformé par le plasmide A_{24} et $(AC)_{24}$. Marquage P.value : NS = >0.05 , * = <0.05, ** = <0.01 , *** = < 0.001.*

On peut donc ici conclure que, pour un organisme eucaryote tel que *S. cerevisiae*, la déficience d'une des sous-unités de son système MMR par rapport à l'autre n'influence pas la directionnalité de la mutation. De plus, il est ici visible que la simplicité du motif A ne favorise pas particulièrement la mutation contrairement à ce qui était visible chez *E. coli*, pour un même nombre de répétitions. Toutefois, la déficience de la sous-unité *mlh1* semble induire une plus forte mutabilité comparée à la perte de fonction de la sous-unité *msh2*.

3.1.2.2.2. Etude des mutations microsatellites dans le génome de S. cerevisiae

Contrairement au précédent protocole, celui suivi ici est basé sur des cultures de levures en milieu YPD à différentes températures pendant plusieurs heures afin de permettre un nombre suffisant de réplication des levures, permettant ainsi aux mutations microsatellites de s'accumuler dans le génome, puis de les mesurer par NGS (Figure 95). Afin de garantir l'obtention d'une quantité suffisante de matériels génomiques, j'ai tout d'abord optimisé le protocole de culture des levures, avant d'effectuer la production en conditions optimales.

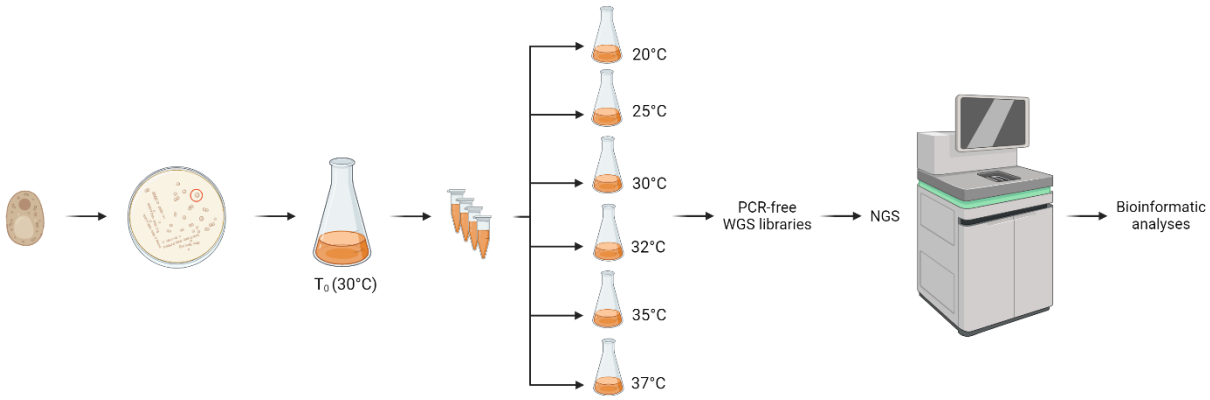

Figure 95 : Workflow Production Levures Génomiques

3.1.2.2.2.1. Optimisation du protocole de culture

Afin d'effectuer la phase de production des différentes cultures de levures aux différentes températures d'intérêt, les conditions expérimentales ont été optimisées pour permettre une culture sur la nuit (Figure 96)

Un clone a été initialement sélectionné par souche après culture sur boîte de petri avec du milieu YPD, supplémenté en généticine (Gen+) pour les mutants (Figure 107). Une préculture commune à chaque souche a été réalisée dans 50 mL de milieu YPD à 30°C sur la nuit, quantifiée par DO et aliquotée dans plusieurs tubes Eppendorf conservés à 4°C, à partir de laquelle toutes les expériences suivantes ont été réalisées.

Afin d'optimiser la durée de culture et d'identifier la période optimale où effectuer le prélèvement final, nous avons effectué une cinétique de la culture dans 150 ml de milieu YPD sur les trois souches (WT, *mlh1* et *msh2*), cultivées à différentes températures (20°C, 30°C, 35°C et 37°), à partir d'un inoculum de 20 000 cellules. Les résultats de la cinétique montrent que le plateau est atteint au bout d'une quarantaine d'heures. Ces cinétiques ont permis de déterminer une fenêtre d'arrêt d'environ 5 heures pour la phase de production au bout de 1,5 jours de cultures afin d'être en fin de phase exponentielle avec une D.O. cible en moyenne de 5 (Figure 96).

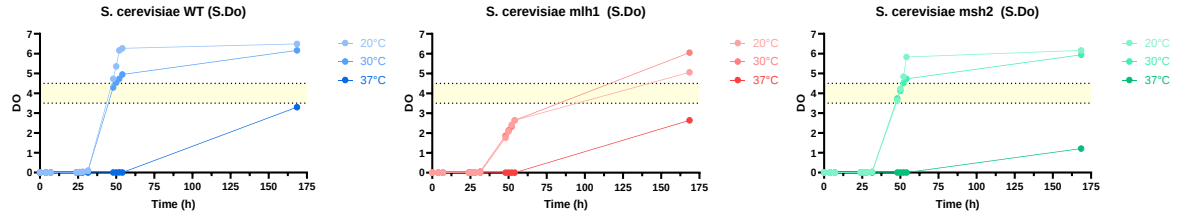

Figure 96 : Cinétique de détermination de la D.O. d'arrêt des cultures de la souches WT, et des mutant mlh1 et msh2 de S. cerevisiae pour une production optimale en restant dans la phase exponentielle (la zone beige représente la fourchette de D.O. d'arrêt de la culture)

Des aliquots de 50 ml des cultures de production subissent une centrifugation pendant 25 min à 46000 rpm pour recueillir les cellules et le culot obtenu a été conservé à -80°C avant isolation d'ADN génomique.

3.1.2.2.2. Doublement de cellule et de molécule

Après la phase d'optimisation, s'en est suivie la phase de production, consistant en la réalisation de toutes les cultures des souches à toutes les températures. A la fin de la production, les cultures ont bel et bien été arrêtées dans les conditions optimales que nous avions définies (Figure 97.A).

Après vérification par DO et comptage après étalement sur boîte de pétri (comme indiqué précédemment), nous avons déterminé qu'en moyenne nous avions 2,48 x 10^4 levures pour la souche WT, 2,3 x 10^4 levures pour la souche *mlh1* et 2,4 x 10^4 levures pour la souche *msh2* au début de la culture (Figure 97.B). Nous observons également que le nombre de levures après cultures (T_f) semble être en moyenne de 3 x 10^9 pour toutes les souches (Figure 97.C). Enfin, avec une si faible variation du T_f, le nombre de doublement de levures est en moyenne à 17 pour toutes les conditions (Figure 97.D).

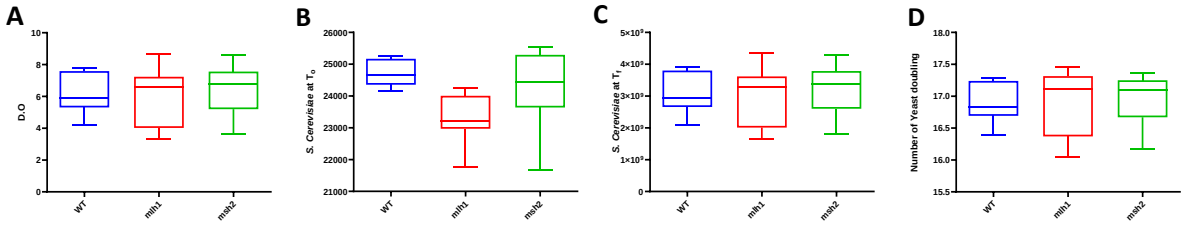

Figure 97 : Evolution de la quantité des levures durant le processus de production. A. Valeurs de D.O. au moment de l'arrêt de la culture en fonction des souches utilisées. B. Nombre de levures S. cerevisiae à T0 selon la souche utilisée. C. Nombre de levures S. cerevisiae à Tf selon la souche utilisée. D. Nombre de doublement de levures selon la souche utilisée.

3.1.2.2.3. Impact de la température sur les microsatellites du génome de la levure

L'ADNg extrait a permis la préparation de librairies WGS « *PCR-free* », avec un séquençage sur NovaSeq6000. La taille des inserts de toutes les librairies est homogène entre différentes souches (350 pb en moyenne) permettant une lecture totale des microsatellites (Figure 98.A). Il est également notable que la couverture est relativement homogène entre 400 et 800X (Figure 98.B). Fait intéressant, le nombre

de copies de l'ADN mitochondrial présent chez *S. cerevisiae* est plus important dans la préculture T_0, puis baisse dans les cultures et varie en fonction de la température (Figure 98.C).

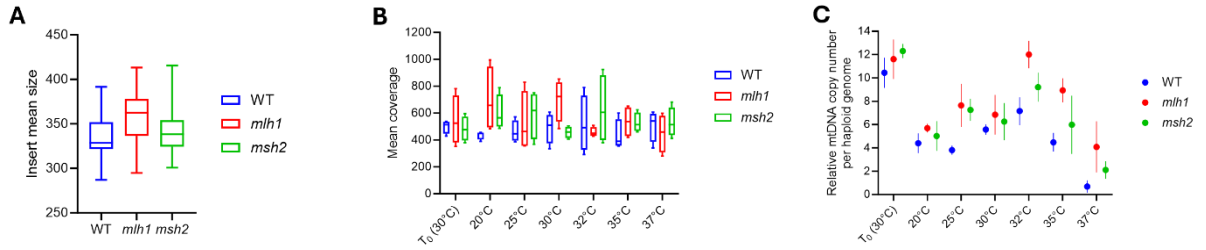

Figure 98 : Contrôle Qualité du séquence WGS de la levure. A. Taille moyenne des inserts analysés par WGS pour chaque souche utilisée. B. Couverture moyenne du séquençage en WGS des différentes souches utilisées en fonction de la température comparée à la culture de départ (T_0). C. Nombre de copies relatif de l'ADN mitochondrial de S. cerevisiae par génome en fonction de la température pour chaque souche utilisée comparé à la culture de départ (T_0).

Par rapport aux parties précédentes, l'analyse des microsatellites dans cette expérience est rendue plus complexe pour différentes raisons. Tout d'abord l'effet de la température sur la réplication de l'ADN est mesuré sur des cellules issues d'une préculture ayant déjà subi de nombreuses divisions cellulaires, et ayant déjà accumulé des mutations. Il n'y a donc pas un réel état T_0 qui représente un microsatellite avec 100% d'un allèle, notamment chez les souches mutées. La soustraction des profils microsatellites du T_0 ne retirerait pas uniquement le bruit de fond du séquençage NGS, mais également un profil réel de mutation microsatellite présent dans l'échantillon T_0.

Par rapport aux microsatellites de références, uniquement ceux communs aux trois souches et ayant un nombre de reads supérieurs à 200 ont été conservés, soit 1 022 microsatellites (Tableau 14). Seuls les microsatellites dont l'allèle majoritaire est commun aux trois souches ont été conservés pour l'analyse. De plus, tous microsatellites qualifiés de « très instables », c'est-à-dire ayant une fréquence d'allèle majeur (MAF) inférieure à 80% dans l'échantillon T_0 ont été retirés de l'étude (Tableau 15 & Figure 99.A). Ainsi, sur les 835 microsatellites restants, un score de MAF a été défini par $\sum_{T°=20°C}^{37°C} MAF\ (T_0) - MAF\ (T°)$. Nous avons ainsi séparé les microsatellites restants en 3 catégories. Dans un premier temps, nous avons les microsatellites « Stables », ayant une MAF comprise entre -2 et 2% (Tableau 15 & Figure 99.B). Dans un deuxième temps, il y a les microsatellites « Instables », ayant une MAF supérieure à 2% (Tableau 15 & Figure 99.C). Et enfin, nous retrouvons les microsatellites « Aberrants », ayant une MAF inférieure à -2% (Tableau 15 & Figure 99.D).

Tableau 14 : Les microsatellites au sein du génome de Saccharomyces cerevisiae. A. Répartition de l'ensemble des microsatellites de référence. B. Répartition des microsatellites conservés pour l'étude car communs aux trois souches et ayant un nombre de reads supérieur à 200.

A

Microsatellites	Number	Mean Size	SD	Repeat size
A/T	1047	14,97	4,10	12-42
C/G	3	14,67	1,15	14-16
AC/TG/CA/GT	29	10	4,75	7-31
AG/TC/GA/CT	13	9,92	6,75	7-32
TA/AT	245	10,08	2,97	7-20
TOTAL	**1337**	13,91	3,59	7-42

B

Microsatellites	Number	Mean Size	SD	Repeat size
A/T	757	15,18	4,61	12-42
C/G	3	14,67	1,15	14-16
AC/TG/CA/GT	29	10	4,75	7-31
AG/TC/GA/CT	13	9,92	6,75	7-32
TA/AT	220	10,02	3,08	7-20
TOTAL	**1022**	13,85	4,90	7-42

Tableau 15 : Catégorisation et répartition des microsatellites conservés pour l'étude. MAF (fréquence d'allèle majeur).

		Très Instable (MAF<80%) WT				Instable				Stable				Aberrant			
	Microsatellites	Number	Mean Size	SD	Size range	Number	Mean Size	SD	Repeat size	Number	Mean Size	SD	Repeat size	Number	Mean Size	SD	Repeat size
WT	A/T	36	17,86	5,20	11-28	40	15,38	5,56	12-33	199	14,76	5,32	12-42	361	13,66	2,51	12-30
	C/G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	AC/TG/CA/GT	3	8	1	7-9	3	9,67	1,53	8-10	13	10,00	6,44	7-31	8	9,88	2,64	7-13
	AG/TC/GA/CT	-	-	-	-	1	32,00	-	32	11	7,91	1,22	7-10	1	10	-	10
	TA/AT	3	12,33	2,52	10-15	27	10,00	2,88	7-16	86	8,81	1,91	7-13	85	10,22	3,15	7-20
	TOTAL	42				71				309				455			
mlh1	A/T	112	18,95	4,82	11-31	301	14,56	3,76	12-36	198	13,72	4,30	12-42	101	13,71	3,60	12-34
	C/G	3	14,67	1,15	14-16	-	-	-	-	-	-	-	-	-	-	-	-
	AC/TG/CA/GT	4	10,25	4,57	7-17	5	11,00	2,45	7-13	14	9,79	6,20	7-31	5	9,20	2,17	7-12
	AG/TC/GA/CT	-	-	-	-	-	-	-	-	9	10,56	8,11	7-32	4	8,50	1,73	7-10
	TA/AT	13	9,92	6,75	7-32	55	11,42	3,00	7-20	97	8,84	2,18	7-16	46	8,96	2,37	7-20
	TOTAL	132				361				318				156			
msh2	A/T	101	19,77	4,98	11-31	320	14,62	3,83	12-35	187	13,66	4,26	12-42	93	13,45	3,43	12-34
	C/G	3	14,67	1,15	14-16	-	-	-	-	-	-	-	-	-	-	-	-
	AC/TG/CA/GT	3	8	1	7-9	7	14,29	7,50	9-31	15	8,20	1,32	7-11	2	7,5	0,71	7-8
	AG/TC/GA/CT	-	-	-	-	1	10,00	-	10	12	9,92	7,05	7-32	-	-	-	-
	TA/AT	14	15,43	2,44	10-18	61	11,26	3,43	7-20	94	8,68	1,86	7-15	43	9,16	2,00	7-15
	TOTAL	121				389				308				138			

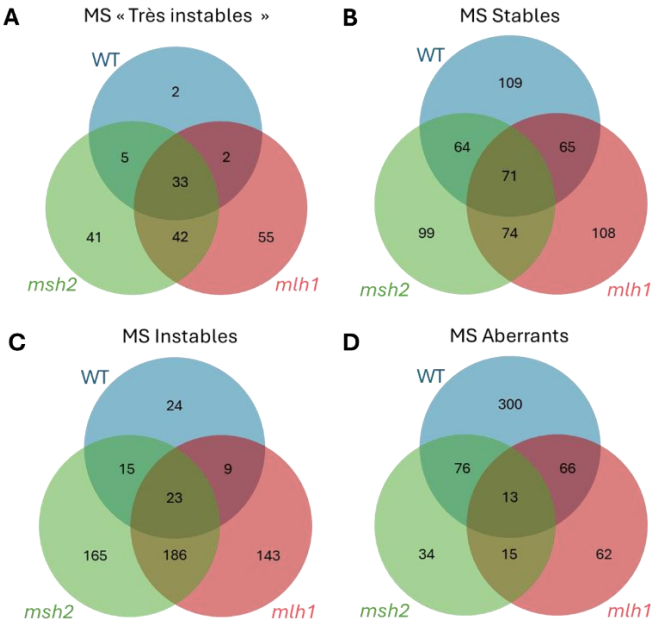

Figure 99 : Représentation en diagramme de la représentation des différentes catégories de microsatellites présent dans le génome de la levure a travers les trois souches étudié (WT, mlh1 et msh2). A. Représentation des microsatellites catégorisés de « Très instables ». B. Représentation des microsatellites catégorisés de «Stables», possédant une MAF comprise entre -2 et 2%. C. Représentation des microsatellites catégorisés de «Instables», possédant une MAF supérieur à 2%. D. Représentation des microsatellites catégorisés de «Abérrants», possédant une MAF inférieur à -2%.

Il est intéressant de remarquer que les microsatellites aberrants sont proportionnellement plus présents chez les souches WT alors que pour les mutants *mlh1* et *msh2*, la plus grosse proportion est occupée par les microsatellites dits instables (Figure 100). C'est d'ailleurs cette catégorie, « Instables », qui nous intéresse le plus car ce sont ces données qui nous montrent l'importance du rôle du système

MMR dans la mutation des microsatellites. Au moment de l'écriture de ce manuscrit, une analyse plus détaillée de ces résultats n'a pas pu être réalisée mais est en cours.

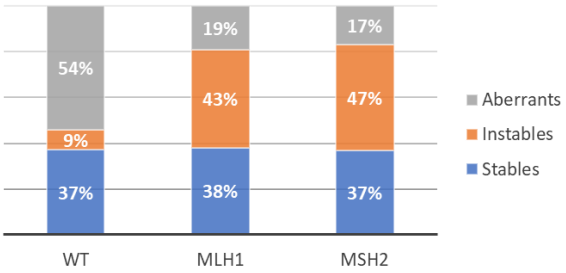

Figure 100 : Evaluation de la proportionnalité de représentation des différentes catégories de microsatellites pour chaque souche utilisée.

3.2. Discussion

Dans ce troisième projet, une étude détaillée de l'influence de différents paramètres sur la formation des mutations de microsatellites a été effectuée dans un contexte *in vitro* et *in vivo* en utilisant soit des microsatellites synthétiques contenus dans des plasmides, soit des marqueurs microsatellites présents dans le génome. Les mutations dans les microsatellites matures (>10 répétitions chez les microsatellites mono- et di-nucléotiques) (Ananda et al., 2013; Baptiste et al., 2013), qui sont connues pour être principalement induites par le mécanisme du glissement de la polymérase ont été ciblées dans cette étude, car elles permettent d'obtenir un taux de mutation important, permettant leur détection et analyse par les méthodes mises en place dans ma thèse. De manière originale, j'ai voulu principalement étudier l'impact d'un facteur exogène qui n'a été que très rarement étudié sur la formation des mutations microsatellites : la température.

Dans ce projet, j'ai voulu utiliser les méthodes développées dans les parties précédentes, à partir des fréquences alléliques et du nombre de doublement de molécules dans des contextes *in vitro* et *in vivo*, ce qui permettrait d'inférer directement le taux de mutation et d'identifier les modèles génétiques de glissement de polymérase les plus probables dans nos conditions expérimentales, s'éloignant des méthodes traditionnelles et de référence pour l'étude des mutations microsatellites *in vitro* et/ou *in vivo*. En effet, les méthodes traditionnelles d'étude des mutations microsatellites, comme la *HSV-tk forward mutation assay*, permettent de décrire et de mesurer les mutations et les taux de mutation des microsatellites *in vitro, in vivo* ou *ex vivo*. En permettant une seule élongation des séquences microsatellites *in vitro*, cette technique offre une mesure directe et équivalente des taux et fréquences de mutation, tandis que dans les systèmes vivants (bactéries, levures, lignées cellulaires humaines) des divisions cellulaires sont autorisées et les taux de mutations exprimés par nombre de divisions (Baptiste et al., 2013; Eckert et al., 1997; Sia et al., 1997b; Wierdl et al., 1997). En fonction des contextes *in vitro*, *in vivo* et/ou *ex vivo*, ces approches permettent également d'évaluer l'impact de différents paramètres, comme le type de microsatellite et le nombre de répétitions, le type de polymérase utilisé, le brin matrice utilisé lors de la réplication du microsatellite et la présence ou non d'un système MMR fonctionnel (Synthèse Bibliographique 5.2.6.). Ces méthodes ont l'avantage de permettre d'observer des mutations à faibles taux, comme lors d'expérience avec un système MMR compétent chez les souches WT, mais la mise en place de l'expérience est relativement lourde et fastidieuse, et ne permet pas d'analyser un nombre très important d'événements mutationnels, puisqu'il nécessite un séquençage Sanger individuel de plusieurs colonies mutantes, rarement plus de quelques dizaines par conditions dans les différentes études (Baptiste et al., 2013; Eckert et al., 1997; Sia et al., 1997b; Wierdl et al., 1997), ce qui augmente la

variabilité des résultats. Dans l'approche utilisée, j'ai pu mesurer les profils alléliques à partir de plusieurs millions d'événements mutationnels. Les expériences ont également considéré un nombre relativement important de molécules d'ADN au départ, ce qui a permis d'obtenir des profils de fréquences alléliques reproductibles et permettant une visualisation globale des événements mutationnels au cours de plusieurs cycles de réplication de l'ADN.

En travaillant sur une large gamme de températures allant de 15°C à 65°C, l'étude *in vitro* a permis d'analyser directement l'influence des variations de température, dont certaines extrêmes, sur les mutations microsatellites, ce qui aurait été impossible à reproduire dans un même type de cellules vivantes pour des raisons de viabilité. L'approche utilisée normalise les fréquences de mutations en fonction du nombre de doublements de molécule, permettant de comparer l'impact des variations de température sur la fréquence des mutations par doublement de molécule en fonction des différentes conditions expérimentales étudiées. Ce mode d'analyse comparant FMDM est effectué dans l'attente de l'inférence directe des taux de mutation à l'aide de l'algorithme développé dans le projet 1, qui n'a pas pu être effectuée au moment de la rédaction de ce manuscrit. Afin d'obtenir une quantification précise de la fréquence des mutations en fonction des variations de température, nous avons utilisé l'électrophorèse capillaire lorsque c'était possible. Cette approche est généralement considérée comme très fiable entre les profils générés et mesurés. Cependant, pour les microsatellites mono-nucléotidiques, il est parfois possible que l'ajout de l'adénine en 3' terminale par la polymérase soit incomplète et fausse le profil électrophorétique obtenu. Pour cette raison, les amplicons obtenus avec les microsatellites mono-nucléotidiques issus des plasmides ont été digérés par une enzyme de restriction avant électrophorèse capillaire. Cependant, l'absence de données de séquence ne peut pas prouver de manière que la taille observée par électrophorèse capillaire ne corresponde qu'à une variation du nombre de répétitions du microsatellite. Ainsi, les résultats des méthodes *in vitro* ont pu être confirmés par NGS, donnant directement accès aux séquences microsatellites, en utilisant le séquençage *short read* avec un protocole de préparation de librairies « *PCR-free* », qui permettait de maintenir au mieux les profils microsatellites originaux d'un échantillon, d'après les résultats de notre étude comparative de technologies présenté dans le projet 2 (Travaux de recherche 2.).

Les résultats des expériences *in vitro* ont montré qu'indépendamment du système d'amplification d'ADN utilisé (WGA, RPA et PCR), une augmentation de la fréquence de mutation par doublement de molécule est observée avec l'augmentation de la température (de 15 à 37°C, de 25 à 45°C, et de 45 à 65°C, respectivement) (Figure 58, Figure 62, Figure 64, Figure 66, Figure 74, Figure 76 & Figure 78), suggérant que l'augmentation de la température à laquelle se réplique l'ADN pourrait déstabiliser l'ADN bicaténaires en cours de synthèse, et induire ainsi plus de glissement de la polymérase (Huguet et al., 2010; Yakovchuk et al., 2006). Bien que ces résultats aient déjà été mentionnés avec des expériences PCR (Seo et al., 2014); pour autant, ils n'ont pas été mis en évidence à des températures plus basses (<45°C) avec les méthodes classiques d'études de mutation microsatellite tels que la HSV-tk, où aucune différence n'a pu être mise en évidence pour des expériences réalisées à 16°C et à 37°C (Baptiste et al., 2015). Les résultats de WGA obtenus avec les plasmides ont montré des fréquences relativement faibles de mutations par doublement de molécule par rapport à la RPA et à la PCR, mais qui augmentent de manière exponentielle avec la température. Cette augmentation est moins marquée avec la RPA, où une tendance exponentielle est toujours observée sur les plasmides, tandis qu'elle est plutôt linéaire par PCR (Figure 58, Figure 62, Figure 64, Figure 66, Figure 74, Figure 76 & Figure 78). Lorsque l'on s'intéresse au sens des mutations, une inversion entre délétions, qui étaient répondantes, et inversions, est observée dans les expériences de WGA vers 30°C avec les homopolymères. Pour les températures les plus fortes testées, les insertions sont très prédominantes en WGA avec tous les microsatellites. En ce qui concerne la RPA, les délétions prédominaient les insertions en général, mais sont relativement équilibrées sur toute la gamme de températures testées, avec une inversion entre délétions et insertions parfois observée

pour les homopolymères (Figure 62, Figure 64 & Figure 66). Enfin, la PCR montre un profil de mutation très reproductible sur l'ensemble des marqueurs génomiques et plasmidiques, avec une prédominance très forte des délétions, quoique moins marquées pour les homopolymères (Figure 74, Figure 76 & Figure 78). Ainsi, bien que la température ait entraîné des effets similaires et systématiques sur la formation de mutations microsatellites dans ces trois systèmes d'études *in vitro*, son effet diffère en fonction du système utilisé, que ce soit par l'intensité des mutations qu'elle a induites, ou par la prédominance du type de mutation : insertion ou délétion. Ces différences sont observables pour les microsatellites communs aux différents systèmes étudiés aux mêmes températures (25°C à 37°C pour le WGA et la RPA et 45°C pour la RPA et la PCR). Pour la PCR, il est possible que le cyclage en deux étapes ne permette pas de contrôler la température à laquelle l'élongation s'effectue, puisque la polymérase peut être active dès l'hybridation des amorces. Le profil obtenu serait donc constitué d'un mélange de molécules répliquées à la température choisie, auquel s'ajoute des molécules répliquées à plus forte température lors de la descente en température.

D'autre part, les trois systèmes d'études *in vitro* présentent également d'autres différences notables, notamment de milieux et conditions réactionnels. Les résultats obtenus impliquent donc la prise en compte de facteurs exogènes supplémentaires liés à l'utilisation de ces différentes méthodes d'amplification, et notamment les ADN polymérases. En effet, le choix de la polymérase dans les réactions d'amplification est un facteur majeur influençant la fidélité de la réplication des microsatellites. Certaines polymérases induisent des taux et des types de mutations variables dans les séquences de microsatellite, en raison de différences dans la structure et les mécanismes de relecture enzymatique (Baptiste et al., 2015; Çağlayan and Bilgin, 2012; Eckert et al., 1997). Dans notre étude, appliquant trois techniques d'amplifications distinctes, nous avons alors utilisé les trois polymérases différentes pour ces techniques (Phi29 pour le WGA, Bsu pour la RPA et Taq pour la PCR), chacune ayant des caractéristiques spécifiques. Au niveau des températures chevauchantes, il est possible, voire probable, que les différences observées dans les trois systèmes soient également causées par ces différentes polymérases, Phi29 (WGA) , la Bsu (RPA) et la Taq (PCR). Tandis que les FMDM sont similaires pour la Bsu et la Taq avec une valeur moyenne variant entre 0,5 et 3, la Phi29 présente des fréquences de mutation par doublement de molécule beaucoup plus faibles variant entre 0,05 et 0,4. La Phi29 est considérée comme une enzyme très processive, permettant de polymériser l'ADN sans interruption sur plus de 100 kb *in vitro* (Dean et al., 2002), tandis que le glissement de la polymérase nécessite une pause dans la synthèse de l'ADN et un décrochage de l'enzyme des brins en cours de synthèse (Viguera et al., 2001). Cela suggère que la polymérase Phi29 pourrait être moins sujette au glissement de la polymérase. L'effet de la température sur la fidélité de différentes polymérases provenant d'organismes psychrophiles, mésophiles et thermophiles a été également analysé à des températures allant de 5 à 72°C, et a montré une augmentation du taux d'erreurs (substitution et indel) dans des séquences non-microsatellites, avec l'augmentation de la température réactionnelle pour les polymérases issues d'organismes psychrophiles et mésophiles (Baptiste et al., 2015; Xue et al., 2021).

D'autres facteurs qui différaient dans nos trois systèmes d'amplification pourraient potentiellement avoir un impact sur la formation des mutations microsatellites, telles que la composition du tampon, la concentration en ions et dNTPs ou la présence d'agent visqueux, mais leurs effets potentiels n'ont été que très peu étudiés de manière comparative dans la littérature (Xue et al., 2021). Les résultats obtenus avec les trois systèmes d'amplification *in vitro* sont néanmoins à prendre avec précaution dans des analyses comparatives.

Concernant la partie *in vivo* du projet, tout comme pour la partie *in vitro*, elle a été effectuée selon une méthode différente des études de mutations microsatellites *in vivo* ou *ex vivo* classiquement utilisées, basées sur des systèmes rapporteurs à l'aide de « gènes suicides », où seules les cellules ayant un plasmide ayant muté pouvaient survivre en présence 5-fluoro-2'-deoxyuridine (Baptiste et al., 2013;

Sia et al., 1997b; Wierdl et al., 1997) ou de 5-fluoro-orotate. Cette méthode devait permettre une analyse simple et rapide des mutations microsatellites générées dans un système *in vivo*, si plusieurs conditions étaient satisfaites, notamment un nombre important de transformants initiaux par les différents plasmides, associé à un taux de mutation assez élevé pour générer suffisamment de mutations mesurables par NGS. Dans ces conditions, les mutations pouvaient être mesurables avec des profils reproductibles, ce qui pouvait permettre l'inférence des taux de mutations et du modèle de glissement de polymérase sous-jacent par l'algorithme développé en projet 1. Au cours de la réalisation de ce projet, l'application de cette approche avec les souches WT présentant un système MMR fonctionnel n'a pas pu donner de résultats analysables, en raison d'un nombre trop peu important d'événements mutationnels avec le nombre de doublement de molécule effectué. Cette approche est donc limitée à ce jour à l'étude de souches déficientes pour le système MMR.

Il est communément admis dans la littérature que les microsatellites des organismes procaryotes montrent une préférence pour la contraction alors que les microsatellites des organismes eucaryotes semblent montrer une préférence pour l'expansion (Ellegren, 2004). Toutefois, la directionnalité de ces mutations est visible dans des études de population ou de génome dans un contexte présentant des systèmes MMR fonctionnels, la mutation microsatellite résultant d'une erreur non corrigée par le système MMR. Dans le cadre des expériences effectuées avec les plasmides sur les souches à système MMR déficients, il a été d'abord remarquable de voir que la fréquence de mutation par doublement de molécule augmente chez *E. coli* et *S. cerevisiae* à mesure que la température de culture augmentait, de manière similaire à ce qui est observé dans les systèmes *in vitro* (Figure 88 & Figure 94), en particulier la RPA, que ce soit au niveau de la tendance des courbes, ou de la valeur de fréquence de mutations par doublement de molécule. En ce qui concerne le gène MMR muté, peu de différences ont été observées entre *mutL* et *mutS* chez *E. coli* et *msh2* et *mlh1* chez la levure, ce qui était cohérent avec la perte de fonction du système MMR, si un seul de ses membres est muté (Alotaibi et al., 2008; Huang and Li, 2020). Il est également notable que la fréquence des mutations était plus importante chez *E. coli* que chez la levure, et que concernant le type de mutation, elles étaient prépondérantes en délétion chez les deux organismes. Il peut être alors suggéré que chez les eucaryotes, le système MMR est si efficace pour corriger les délétions que les eucaryotes montrent une propension aux insertions alors que chez les procaryotes, le système MMR semble moins efficace à corriger les délétions, comme montré dans d'autres études où les différences de taux et de types de mutations ont été étudiées chez des souches WT et mutées pour le système MMR (Eckert and Hile, 2009; Reynes, 2021; Sia et al., 1997b; Wierdl et al., 1997). Cette différence d'efficacité de correction pourrait ainsi expliquer la raison pour laquelle les microsatellites sont en expansion dans les génomes d'eucaryotes et en contraction dans les génomes de procaryotes. Bien que la température joue sur le nombre de mutations dans ces organismes, il n'y a pas d'évidence d'inversion de types de mutations pour les températures testées, tant les délétions sont prépondérantes, contrairement à ce qui a été observé dans les expériences *in vitro*. Ces résultats sont également en accord avec les résultats d'études *in vivo* effectuées dans des systèmes MMR déficients chez l'homme ou la souris, tels que les cellules tumorales ou non, où une prépondérance des délétions est observée (Christopher et al., 2019; Cortes-Ciriano et al., 2017; Hause et al., 2016).

Les analyses des résultats concernant le *Whole Genome Sequencing* de levures n'ont été que peu approfondies dans cette étude pour diverses raisons. En effet le code informatique développé pour analyser de manière optimale les résultats ne permettrait pas l'analyse aisée de ces données. En effet, une difficulté concerne nos données des échantillons considérés T_0. Contrairement aux plasmides des précédentes études, ils contenaient déjà beaucoup de mutations, notamment dans les souches mutantes, puisqu'il s'agit de pré-cultures à partir de colonies isolées. Ainsi, notre méthode d'inférence ne permet pas à ce jour de traiter des données à partir d'un microsatellite qui n'est pas 100% pur, c'est-à-dire qui ne présente pas 100% de l'allèle originel, ce qui rend l'analyse particulièrement compliquée.

L'analyse de ces données est également difficile avec notre stratégie de soustraction de bruit de fond, puisque l'on soustrait, en plus du bruit, le profil mutationnel de l'échantillon T_0. De plus, certains échantillons présentaient après culture (de T_0 à T_f) des fréquences qui diminuaient pour certaines tailles d'allèles, ce qui aboutissait à une fréquence d'allèle négative, notamment à température élevée. Il est possible que ces allèles subissent des délétions prédominantes, à partir d'une certaine taille. Une stratégie d'analyse devra être développée pour l'exploitation optimale de ces données.

Pris collectivement, nos données d'expériences *in vivo* ont permis de mettre en évidence l'impact majeur de la température sur la formation des mutations microsatellites *in vivo* au niveau des microsatellites mono et di-nucléotidiques. À notre connaissance, l'impact de la température n'a jamais été étudié à ce jour sur la formation des mutations microsatellites, les expériences réalisées *in vivo* ou *ex vivo* utilisant des températures optimales de croissance. Pour autant, l'impact de la température a déjà été étudié dans des séquences non-microsatellitaires des gènes *rpoB* et *lacZ* chez *E. coli* avec un système MMR WT ou mutants (Hong et al., 2005). Leurs expériences ont montré une augmentation du taux de substitution de 30°C à 43°C, avec parfois une légère inflexion de 37°C à 43°C en fonction, notamment chez le WT et certaines souches mutantes. Plus récemment, des expériences similaires ont été menées à des températures de 25°C à 37°C avec une souche *E. coli* B REL606 *mutS* (Chu et al., 2018), ou de 37°C et 40°C avec des souches d'*E. coli* présentant un système MMR fonctionnel (Van Eldijk et al., 2024) et ont montré une augmentation du taux de mutation de 25°C à 37°C dans la première étude, et une augmentation ou diminution du taux de mutation dépendant du gène de résistance aux antibiotiques (ciprofloxacine, rifamicine et ampicilline) dans la seconde. Chez la levure, des expériences ont été menées sur différentes souches présentant un système MMR fonctionnel à des températures variables, pouvant aller de 30°C à des températures extrêmement hautes, à plus de 40°C et 50°C (Evans and Parry, 1975; Huang et al., 2018). Leurs résultats ont montré une augmentation d'événements mutationnels dans le génome de la levure aux températures plus élevées, mais qui résultent parfois d'une sélection positive leur permettant la survie à des fortes températures, limitant parfois la compréhension de l'effet unique de la température sur les mutations. A ce jour, peu d'études ont été réalisées sur des organismes multicellulaires. L'influence de températures allant de 15°C à 30°C a été notamment étudiée chez la drosophile, et n'a pas montré de variation des taux de mutation (Sheldon and Barker, 1964).

4. Discussion générale, Conclusion et Perspectives

Cette thèse s'inscrivait dans une démarche qui visait à approfondir les connaissances sur la compréhension des mécanismes mutationnels affectant les microsatellites en s'intéressant même à celles qui impactaient des méthodologies analytiques importantes pour leur étude. Ma thèse s'articulait autour de trois projets mêlant recherche fondamentale, technologique et méthodologique, qui s'inscrivaient dans deux objectifs principaux définis à la fin de l'introduction : (A) développer une approche combinant l'utilisation d'un système rapporteur et le séquençage haut débit permettant l'étude des mutations microsatellites et (B) étudier l'effet de facteurs, notamment exogènes, sur la formation des mutations microsatellites (Figure 24).

Dans le premier projet, nous avons souhaité développer une méthode d'inférence des taux de mutation des microsatellites à partir d'un profil de fréquence allélique obtenue à partir d'une amplification d'un microsatellite pur, c'est-à-dire comportant uniquement 100% d'allèle d'une taille définie, dans un système d'amplification de type PCR ou d'amplification isotherme de type RPA, et que nous avons appelé *Microsatellite Mutation Rate Calculator 32* (MSmrc32). Le taux de mutation est une mesure clé en génétique qui permet de décrire et caractériser les mécanismes mutationnels, par exemple à un locus donné. Nous avons également souhaité que cette méthode permette d'identifier le modèle

génétique à l'origine du glissement de la polymérase parmi 32 modèles qui ont été proposés à partir de la littérature. Ces modèles décrivent le mécanisme de glissement de polymérase qui peut se produire au cours de la réplication, prenant en compte plusieurs paramètres qui étaient souvent observés dans des études in vivo et in vitro et qui ont servi de contrainte à nos modèles. Les taux de mutations en insertion et en délétion ont été liés au nombre de répétitions par une relation croissante linéaire ou exponentielle, tandis que la mutation pouvait se produire de 1 à 4 pas, avec une diminution exponentielle des taux de mutation avec l'augmentation du nombre de pas. L'approche mathématique d'inférence reposait sur le critère de minimisation de l'erreur quadratique moyenne (RMSE) calculée entre la distribution des allèles microsatellites observées obtenue lors d'une expérience avec une distribution générée à partir d'un des 32 modèles, associé à un algorithme de recuit simulé (SA) couplé à une recherche en grille pour optimiser l'identification des paramètres. Cet algorithme a alors permis d'inférer les taux de mutation des microsatellites pour chacun des 32 modèles correspondant au RMSE le plus petit après optimisation, et aussi d'identifier le modèle génétique — le modèle avec le plus petit RMSE parmi les 32 — qui serait à l'origine de la distribution observée des allèles microsatellites. La méthode développée propose trois modes d'inférence adaptés à différents contextes expérimentaux que nous pouvons rencontrer dans nos expériences, et permettant plus de flexibilité à l'utilisateur : le mode I correspond à des efficacités variables cycle par cycle qui correspondent à un système d'amplification de l'ADN de type PCR en temps réel où les efficacités sont accessibles pour chaque cycle de la réaction à partir de données brutes, le mode II correspond à une efficacité unique et simplifiée de 100% adaptée à des données générées obligatoirement avant la phase de plateau de la PCR pour pouvoir inférer correctement les taux de mutation, et le mode III qui infère les taux de mutation à partir d'un nombre de doublements de molécule compatibles aux méthodes d'amplification isotherme tels que la RPA, la HDA ou la LAMP. À ce jour et à notre connaissance, la seule méthode similaire a été publiée en 2004 par Shinde et al. (Shinde et al., 2003) par une approche basée sur une méthode de quasi-vraisemblance ne prenant en compte qu'un seul modèle génétique et n'était applicable qu'à des données de PCR. L'algorithme développé a d'abord été testé sur des données simulées générées à partir des 32 modèles génétiques, permettant de comparer les taux de mutation simulés et inférés pour évaluer ses performances. Les résultats ont alors mis en évidence que les modèles avec un faible nombre de paramètres, comme les modèles à un seul pas, offrent des performances supérieures, en particulier pour les faibles taux de mutation. Les relations linéaires ou exponentielles entre les taux de mutation et la longueur des microsatellites n'ont montré que peu d'impact sur la précision globale. Par la suite, l'algorithme a été appliqué sur des données issues de PCR et de RPA sur des microsatellites mono- (A/T) et di-nucléotidiques (AC/TG). Les résultats ont révélé une augmentation des taux de mutation avec le nombre de répétitions, souvent biaisée en faveur des délétions, notamment dans les microsatellites di-nucléotidiques. Une dynamique distincte entre les méthodes a également été observée, avec des taux globaux de mutation plus faibles en RPA qu'en PCR, probablement en raison des différentes polymérases utilisées et des conditions expérimentales. Les modèles identifiés comme les plus probables pour les microsatellites mono- et di-nucléotidiques différaient selon la méthode : E3d1u et L4d1u pour la PCR, E2d1u et E3d2u pour la RPA. Ces résultats nous ont permis d'approfondir la compréhension des mécanismes de glissement de polymérase et des dynamiques mutationnelles *in vitro*, nous permettant alors de proposer un outil innovant pour l'analyse des mutations microsatellites *in vitro* à partir de méthodes d'amplification de l'ADN tels que la PCR et la RPA. Cette méthode offre ainsi un outil de caractérisation des dynamiques mutationnelles et pourrait être une alternative à d'autres approches telles que les essais HSV-tk. L'algorithme pourrait également ouvrir la voie à de nouvelles investigations sur les mécanismes d'évolution des microsatellites *in vivo*, mais aussi *ex vivo*. Parmi les perspectives de ce projet, bien que 32 modèles aient été définis, il est possible que d'autres soient pertinents et ils seraient donc intéressant d'en inclure davantage. Des modèles qui distinguent la relation entre le taux d'insertion et de délétion et le nombre de répétitions seraient peut-être intéressants à considérer, de même qu'un nombre de pas plus important, puisque

nous nous sommes limités à 4 dans ce projet. Par exemple, il aurait été intéressant de considérer une relation croissante exponentielle pour les délétions et linéaire pour les insertions, ou réciproquement. D'autre part, l'identification des modèles est basée sur le RMSE le plus petit obtenu par les 32 modèles. Bien que cette approche apparaisse comme reproductible et robuste dans un certain nombre de cas (sur les données PCR par exemple), elle l'est parfois moins (les données de RPA), ce qui génère des incertitudes sur le modèle génétique sous-jacent réel impliqué dans le glissement de la polymérase. Une approche plus robuste pour l'identification des modèles pourrait donc être explorée comme perspective. Au début de ma thèse, cet algorithme avait aussi pour but d'être utilisé sur les données expérimentales générées *in vivo*, dans la levure et la bactérie, ce qui n'a pas été possible faute de temps. Pour autant, l'algorithme est dans sa version actuelle initialisé sur un allèle unique pur (100% d'allèle de la taille indiquée), alors que les résultats du projet 3 obtenus à partir du génome de levure (expérience de WGS) nous ont montré des résultats difficiles à interpréter à cause d'un échantillon précultivé T_0 qui contenait lui-même un mélange d'allèles microsatellites de tailles différentes avant la mise en culture aux différentes températures. Améliorer notre logiciel pour la prise en compte d'un état initial qui n'est pas un T_0 pur ou qui représente une phase avec quelques cycles de réplication de l'ADN serait un atout majeur. Une perspective d'évolution importante du programme, permettrait des applications variées, et l'analyse de nos résultats de WGS.

Dans le second projet, nous avons voulu évaluer et comparer différentes technologies et protocoles de séquençage haut débit pour la détermination précise de la taille des microsatellites de type homopolymère A/T et microsatellites di-nucléotidiques AC/TG et AT/TA, qui sont les plus fréquents dans le génome humain. En effet, le génotypage et le séquençage des microsatellites ont toujours été fastidieux, et complexes du fait des artéfacts *stutter* produits par l'amplification PCR et des erreurs de séquençage dans ces régions répétées, et ce quelle que soit la technologie de séquençage utilisée. L'identification d'un protocole optimal de séquençage était importante dans le cadre de ma thèse car elle devait être appliquée à tous nos échantillons comportant des séquences de microsatellites amplifiés *in vitro* et *in vivo* pour l'étude des mutations, dans le but de refléter au mieux le profil allélique présent dans l'échantillon d'origine, minimisant le bruit de fond lié à l'étape de séquençage. Ce benchmark technologique était donc important pour la mise en place de notre système rapporteur pour l'étude des mutations microsatellites. Nous avons pour cela utilisé plusieurs plasmides contenant des microsatellites mono-(A/T) et di-(AC/TG et AT/TA) nucléotidiques de 15 à 25 répétitions, considérés comme purs, pour générer plusieurs bibliothèques NGS basées sur les technologies de séquençage de 2ème (Illumina) et de 3ème (PacBio) génération. Ces approches étaient pour la plupart dédiées à la réduction des erreurs de séquençage et à la détection des mutations rares et non microsatellitaires, et avaient donc été rarement appliquées à des microsatellites mono- et di-nucléotidiques matures. Pour le séquençage Illumina, nous avons évalué plusieurs protocoles, y compris des protocoles avec PCR (1 à 20 cycles) et sans PCR, ainsi que des protocoles avec un code barre-moléculaire "UMI" unique, localisé dans l'index, ou double (protocole "*duplex sequencing*") localisé à chaque extrémité de l'insert permettant la correction des erreurs de PCR et de séquençage, en utilisant deux séquenceurs : l'iSeq100 (approches sans correction UMI) et le NextSeq 500 (toutes les approches). Concernant le séquençage de 3ème génération, un protocole sans PCR a aussi été testé avec le séquenceur PacBio Sequel II. Les analyses bioinformatiques ont été également un élément clé de ce projet. Nous avons développé plusieurs stratégies pour identifier et estimer correctement la longueur des microsatellites à partir des données de séquençage. La stratégie consistait à considérer la taille des microsatellites plutôt que la séquence pour l'obtention de la taille des allèles microsatellites. Cela comprenait une approche axée sur l'identification des régions flanquantes des microsatellites afin d'estimer leur longueur dans chaque read. Cette méthode a permis de surmonter les problèmes d'artefacts *stutters* et de mauvaise qualité de reads, souvent rencontrés dans le séquençage NGS, en particulier avec les homopolymères et les répétitions complexes. En se focalisant sur la taille du microsatellite plutôt que sur sa séquence, nous avons pu

générer des consensus de taille par les approches basées sur les codes-barres moléculaires avec Illumina et les subreads PacBio, ce qui était impossible avec les séquences du fait d'un nombre trop important d'erreurs dans les séquences brutes. Nous avons aussi mis au point différents modes d'analyses pour générer des allèles consensus à partir des reads UMI ou des subreads issus du séquençage PacBio (*Max. Freq., Max. Length, Mean Length, Median Length*), ainsi que des méthodes combinées pour améliorer la précision de la détection des allèles issus d'expérience de "*duplex sequencing*". Les résultats de ce projet ont montré que le séquençage en *short-read* Illumina, notamment en protocole sans PCR, permet de générer les résultats les plus fiables pour la détermination de la taille des microsatellites, particulièrement dans les STR di-nucléotidiques et les homopolymères, malgré les avancées significatives du séquençage en *long-read* (PacBio), dont les bénéfices concernent les très longs microsatellites rencontrés dans les maladies génétiques à expansion de triplets. Toutefois, les *reads* PacBio ont permis une amélioration pour certains STR di-nucléotidiques ($(AC)_{15}$) avec une stratégie d'utilisation très poussée des *subread*s (35 *subread*s par consensus) peu compatible en routine. Les perspectives de ce projet sont multiples et concernent notamment l'analyse des microsatellites avec des motifs différents. On pourrait par exemple étendre notre étude à des microsatellites comportant des motifs plus grands tels que les microsatellites tri-, tétra- et penta-nucléotidiques, avec un nombre de répétitions plus important (30, 40, 50,…) ou réduit (10, 8, 5, …), avec des motifs différents pour des microsatellites mono- (C/G) et di-nucléotidiques (AG/TC et CG/GC), ou avec une proportion différente en GC dans les séquences flanquantes (20% à 80%). D'autre part, il serait intéressant d'analyser d'autres technologies et instruments de séquençage qui n'ont pas été évalués dans notre étude, dont certains ne sont disponibles que depuis peu. On peut citer le séquençage nanopore de troisième génération, la nouvelle chimie XLEAP SBS d'Illumina pour ses séquenceurs NextSeq 1000/2000 et NovaSeq X, des instruments PacBio Revio et Vega utilisant le séquençage par synthèse SMRT et du système PacBio Onso utilisant le séquençage par liaison, dont le fabricant affirme qu'il peut séquencer et résoudre avec précision les homopolymères et autres régions répétitives. Parmi les autres technologies notables, on peut aussi citer le système AVITI de séquençage par synthèse d'Element Biosciences, qui utilise l'amplification circulaire isotherme pour la génération de colonies, les instruments basés sur le séquençage par synthèse utilisant les nanobilles d'ADN (DNB) de MGI et la technologie de séquençage par synthèse d'Ultima Genomics. Nous prévoyons d'évaluer ces technologies dans de futures études afin d'améliorer le séquençage des microsatellites et la détermination de leur longueur. Il est certain que le séquençage des microsatellites devrait être grandement amélioré avec l'essor de nouvelles technologies de séquençage arrivant de plus en plus sur le marché.

Enfin, dans le troisième projet, nous avons mené une étude fondamentale approfondie pour analyser l'influence de divers paramètres, et notamment celle de la température, sur la formation des mutations microsatellites, à la fois *in vitro* et *in vivo*. L'approche *in vitro* s'appuyait sur des microsatellites synthétiques insérés dans des plasmides et des marqueurs microsatellites naturels du génome humain, tandis que l'approche *in vivo* était basée sur l'insertion de ces mêmes plasmides dans la bactérie *E. coli* et dans la levure *S. cerevisiae*, avec également une approche pangénomique pour la dernière espèce. Nous avons ciblé spécifiquement les mutations des microsatellites matures dans les plasmides et les marqueurs microsatellites humains, dont les mutations sont principalement dues au glissement de la polymérase, avec un taux d'erreur important qui permettait une mesure fiable et reproductible de ces erreurs, au-dessus du bruit de fond. Ces mutations, ayant un taux suffisamment élevé pour être analysé avec les méthodes développées dans les projets précédents, ont permis une étude détaillée et originale de l'impact de la température, rarement étudiée dans ce contexte. L'étude de ces mutations microsatellites a non seulement permis d'explorer les mécanismes à l'origine de leur apparition, mais aussi l'influence de conditions exogènes ou environnementales spécifiques, comme la température, sur leur évolution génétique, tant quantitative (quantité de mutations) que qualitative (directionnalité des mutations). L'approche a reposé sur l'utilisation de méthodes originales basées sur les fréquences

alléliques et le nombre de doublements de molécule d'ADN. Contrairement aux méthodes classiques comme l'essai HSV-tk, souvent limitées par leur complexité et le faible nombre d'événements analysés, les techniques utilisées dans ce projet ont permis de mesurer des profils alléliques à partir de plusieurs millions d'événements mutationnels. Ces approches de grande envergure ont permis une meilleure résolution dans l'analyse des effets de la température, en offrant une couverture plus large et plus précise des mutations, et d'obtenir des profils mutationnels reproductibles, car ils sont générés à partir d'un nombre important de molécules d'ADN initiales, étant donné les taux de mutations rencontrés. Cependant, l'une des limitations majeures de cette approche est que dans des contextes *in vivo* et WT, avec un système MMR fonctionnel, les mutations introduites étaient trop peu nombreuses, se confondant avec le bruit de fond de la méthode de détection choisie, le séquençage Illumina de librairies préparées sans étapes de PCR — et n'ont donc pas pu être analysées. Ainsi, dans des contextes WT, le système d'étude de mutation par HSV-tk demeure plus performant puisqu'il permet l'analyse des taux de mutations très faibles, comme rencontré dans un contexte WT (Eckert et al., 1997; Eckert and Yan, 2000). Les expériences ont couvert une large gamme de températures allant de 15°C à 65°C et ont révélé que l'augmentation de la température entraîne une hausse de la fréquence des mutations par doublement de molécule. Ce phénomène pourrait être lié à une déstabilisation accrue de l'ADN au cours de la synthèse, suggérant que des conditions environnementales qui modifient la stabilité de l'ADN pourraient avoir un impact direct sur la formation des mutations. Ces résultats ont été validés par électrophorèse capillaire et NGS, avec des protocoles préservant au mieux les profils microsatellites originaux. L'intégration de ces différentes techniques a permis de valider la robustesse des observations, tout en mettant en évidence des variations entre les conditions expérimentales. Les résultats ont également montré que l'effet de la température variait selon les méthodes d'amplification utilisées, comme le WGA, la RPA et la PCR, ainsi que les polymérases, c'est-à-dire la Phi29 polymérase, le Bsu polymérase et la Taq polymérase, respectivement. Ces variations influencent non seulement la fréquence des mutations mais aussi leur type, avec une prédominance de délétions ou d'insertions en fonction des conditions expérimentales. Cela souligne l'importance de tous les paramètres techniques dans l'étude des mutations microsatellites, et l'impact du type de polymérase ou du système d'amplification sur les résultats obtenus. Ce projet démontre l'importance de bien maîtriser ces paramètres dans le cadre de l'analyse génétique des microsatellites, et ouvre la voie à une meilleure compréhension des conditions dans lesquelles ces mutations surviennent. Concernant les expériences *in vivo*, réalisées sur des souches procaryotes (*E. coli*) et eucaryotes (*S. cerevisiae*) déficientes pour le système MMR, il a été observé que la température affecte significativement la fréquence des mutations microsatellites par doublement de molécule. Les mutations étaient majoritairement des délétions, indépendamment de la température, bien que leur fréquence augmentât avec celle-ci. Ces résultats *in vivo*, similaires à ceux obtenus *in vitro*, ne suggèrent aucune différence fondamentale entre les systèmes procaryotes et eucaryotes dans le biais mutationnel des microsatellites dans un contexte MMR déficient. Ces différences de biais observées à l'échelle des génomes eucaryotes et procaryotes pourraient être liées à d'autres paramètres, tels que l'efficacité variable du système de réparation des mésappariements, qui joue un rôle clé dans la gestion des erreurs de réplication dans l'ADN et corrigerait de manière plus efficace un type de mutation par rapport à un autre en fonction de l'espèce. Les résultats observés dans les systèmes procaryotes, qui montrent un biais plus prononcé vers les délétions, contrastent avec ceux observés dans les systèmes eucaryotes, où les expansions semblent en général prévaloir dans la plupart des organismes (Ellegren, 2004). Cependant, l'analyse des données de séquençage complet du génome de levure réalisée dans ma thèse a révélé des défis méthodologiques, notamment la complexité de l'inférence des mutations dans les microsatellites qui ont déjà subi des modifications au départ. Ces limitations, bien qu'importantes, mettent en évidence la nécessité d'affiner les méthodologies pour mieux comprendre les dynamiques des microsatellites dans des contextes biologiques complexes, et constituent l'une des perspectives du projet 1, comme évoqué plus haut.

En conclusion, ce projet a permis de mettre en lumière l'impact majeur de la température sur la formation des mutations microsatellites. L'approfondissement de la compréhension des facteurs environnementaux et des paramètres techniques influençant ces mutations pourrait représenter un progrès dans le domaine de la génétique. Ces travaux, combinant des approches innovantes et des méthodes classiques, offrent de nouvelles perspectives pour mieux comprendre les mécanismes sous-jacents aux mutations microsatellites, leur directionnalité et les facteurs environnementaux qui influencent leur dynamique dans différents contextes biologiques. Ces résultats pourraient motiver de futures recherches qui permettraient d'aider à mieux évaluer et anticiper l'impact des variations environnementales, notamment dans un contexte de réchauffement climatique, sur la stabilité et l'évolution des génomes, tant des espèces procaryotes qu'eucaryotes.

Lors d'études comparant les microsatellites du génome du chimpanzé et de l'humain, il a été observé que les séquences plus longues et certains types de répétitions mutent plus fréquemment, confortant un rôle de la composition et de la structure génomique dans la stabilité de ces séquences. En analysant spécifiquement les différences humaines, certaines répétitions en tandem se sont révélées plus étendues et fréquentes dans le génome humain, particulièrement dans des régions régulatrices associées à des gènes liés au développement cérébral et à l'immunité, ce qui pourrait être à l'origine de traits spécifiques à l'espèce humaine. De plus, certaines de ces expansions semblent être sous l'influence de la sélection naturelle, qui stabiliserait les microsatellites dans des régions génétiquement critiques. Ainsi, ces recherches proposent que les microsatellites et les répétitions en tandem contribuent potentiellement aux différences phénotypiques et fonctionnelles entre les humains et les chimpanzés, soulignant leur rôle dans la diversification génétique au cours de l'évolution des primates (Kelkar et al., 2008; Sainudiin et al., 2004; Sulovari et al., 2019; Webster et al., 2002). Toutefois, ayant une différence de température corporelle entre le chimpanzé et l'humain (36,5°C contre 37°C en moyenne respectivement), nous pouvons spéculer que certaines des différences de fréquences de mutations entre ces espèces pourraient être également expliquées par cette différence de température. Cette hypothèse est appuyée par les conclusions d'une étude comparative des génomes datée de 2014, où Amos et Filipe ont recherché des corrélations entre la fréquence de distribution de taille de certains motifs microsatellites dans le génome avec la température corporelle et le poids corporel adulte moyenné par sexe chez 69 espèces de mammifères (Amos and Filipe, 2014). Leurs résultats suggéraient qu'il existait bien une corrélation entre la température corporelle et la fréquence de distribution de taille des microsatellites et, par extension, au taux de mutation, où la température serait associée à une augmentation du taux de mutation des microsatellites. Même au sein de l'espèce humaine, ou d'autres homéothermes, il existe des différences de température corporelle plus ou moins importantes. En effet, entre un homme et une femme, entre deux femmes à différents moments de leur cycle hormonal, entre un caucasien et un asiatique, il existe des diversités de température variant jusqu'à 1°C (entre 36,1°C et 37,2°C) (Diamond et al., 2021; Fernández-Peña et al., 2023; Geneva et al., 2019), et la température varie également lors de processus physiologiques comme la défense contre des infections par l'augmentation de température avec la fièvre (Walter et al., 2016). Ces variations pourraient influer sur les mutations microsatellites. Enfin, si l'on se concentre sur les espèces dont la température corporelle n'est pas autorégulée telle que les espèces poïkilothermes comme les reptiles, amphibiens, poissons, insectes et invertébrés, ou même les plantes et les micro-organismes, comme étudié dans ma thèse, le simple effet du réchauffement climatique pourrait influer sur leurs températures corporelles ou cellulaires, ce qui pourrait favoriser les mutations microsatellites. Il serait également possible de penser que, si la température impacte le taux de mutations des microsatellites, elle pourrait impacter des mécanismes de mutations plus étendus et ainsi induire une pression de mutations évolutive dans certains cas.

En conclusion, les trois projets de ma thèse se sont inscrits dans deux grands objectifs consistant à la mise en place d'un système rapporteur permettant l'étude des mutations microsatellites *in vitro* et

in vivo compatibles au NGS et à la compréhension de l'effet de différents facteurs, notamment exogènes, sur le processus mutationnel des séquences microsatellites, qui ont été atteints au cours de ma thèse. Elles aboutissent sur des conclusions et hypothèses intéressantes à différents niveaux, qu'ils soient méthodologiques, technologiques ou fondamentaux, et qui ont des implications en biologie de l'évolution des génomes, notamment face aux perspectives de réchauffement climatique.

1. Matériel Biologique

1.1. Souches Bactériennes

Différentes souches de bactéries *E. coli* ont été utilisées, qui sont issues de la collection Dharmacon™ *E. coli* Keio Knockout Collection (Horizon Discovery, https://horizondiscovery.com/) générée à partir de la souche K-12.

La particularité de cette collection est que chaque mutant d'*Escherichia coli* K-12 porte une mutation précise d'un seul gène visant à faciliter l'étude fonctionnelle de la protéine ciblée due à la délétion du gène codant pour elle (Baba et al., 2006; Datsenko and Wanner, 2000).

La première étape dans la construction des mutants consistait à concevoir des cassettes de délétion spécifiques pour chaque gène cible. Chaque cassette comprenait un gène de résistance à un antibiotique (ici la kanamycine) flanqué de sites de reconnaissance FLP (« *FLP recognition target* », FRT). Ces sites FRT permettent l'excision ultérieure du gène de résistance par une recombinase FLP, laissant derrière eux une séquence FRT résiduelle. La cassette de délétion était amplifiée par PCR en utilisant des amorces portant des extensions homologues aux régions flanquantes du gène cible (Figure 101).

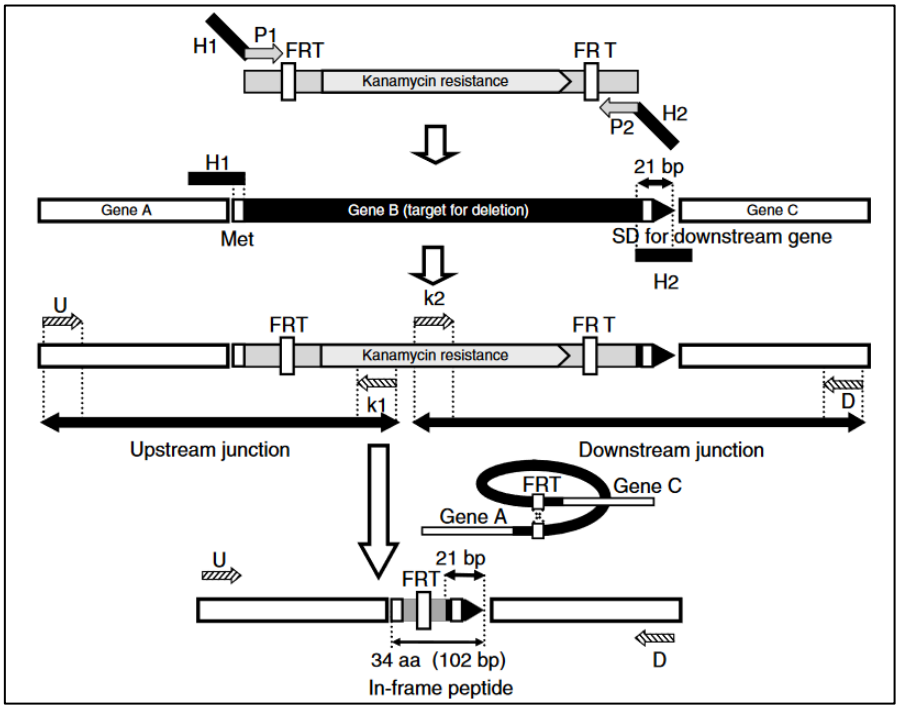

Figure 101 : Conception des amorces et construction des mutants à délétion d'un seul gène. Les amorces de knockout de gène ont des extrémités 3' de 20 nucléotides (nt) pour l'amorçage en amont (P1) et en aval (P2) des sites FRT flanquant le gène de résistance à la kanamycine dans le plasmide pKD13. Elles ont également des extrémités 5' de 50 nucléotides homologues aux séquences chromosomiques en amont (H1) et en aval (H2) pour cibler la délétion du gène (Tableau supplémentaire 2). H1 inclut le codon d'initiation du gène B (cible). H2 inclut les codons pour les six résidus C-terminaux, le codon stop, et 29 nucléotides en aval. Le même design d'amorce par rapport au gène B a été utilisé pour cibler les délétions, que le gène B se trouve dans un opéron avec les gènes A et C, comme montré, ou dans différentes dispositions chromosomiques. Les nouvelles jonctions créées entre la cassette de résistance et les séquences voisines en amont (gène A) et en aval (gène C) ont été vérifiées par PCR avec des amorces de kanamycine (k1 ou k2) et des amorces spécifiques au locus (U ou D). SD, séquence de liaison du ribosome Shine-Dalgarno. Schéma de (Baba et al., 2006)

Les cassettes de délétion amplifiées par PCR étaient introduites dans les cellules d'*E. coli* K-12 par électroporation, une technique permettant de rendre les cellules perméables à l'ADN exogène sous l'influence d'un champ électrique. Les cellules d'*E. coli* avaient préalablement été transformées avec un plasmide exprimant une protéine recombinase (λ Red recombinase), facilitant la recombinaison homologue entre les séquences flanquantes des cassettes et les régions correspondantes du génome bactérien. Après l'électroporation, les cellules étaient cultivées sur des milieux sélectifs contenant l'antibiotique correspondant, permettant uniquement la croissance des cellules ayant intégré la cassette de résistance. Les colonies résistantes obtenues étaient ensuite testées par PCR pour vérifier la présence correcte de la délétion du gène cible et l'intégration précise de la cassette de résistance.

Ces lignées ont pour fond la souche *E. coli* K-12 BW25113 (rrnB3 ΔlacZ4787 hsdR514 Δ(araBAD)567 Δ(rhaBAD)568 rph-1) :

- Non mutée (WT) : souche K-12 BW25113
- Mutée au niveau pour le gène *MutS* (ΔMutS) : souche K-12 BW25113 - JW2703
- Mutée au niveau pour le gène *MutL* (ΔMutL) : souche K-12 BW25113 - JW4128

Durant ma thèse je me suis également servie d'une autre souche de bactérie, la NEB® 5-alpha Competent *E. coli* (High Efficiency) (New England Biolabs) (fhuA2Δ(argF-lacZ)U169 phoA glnV44 Φ80Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17) qui est dérivée de la souche *DH5α*. Cette souche est résistante au phage T1 et déficiente en endonucléase afin de garantir une amélioration des capacités de transformation et de manipulation des plasmides.

1.2. Souches de Levures

Nous nous sommes servis de différentes souches de levure *S. Cerevisiae* de la Yeast Knockout (YKO) Collection issues de la lignée haploïde hétérothallique S288C :

- MAT a (his3Δ1 leu2Δ0 met15Δ0 ura3Δ0)
    - Non muté (WT) : souche BY4741
    - Muté au niveau du gène *MLH1* (Δmlh1) : souche BY4741-00750
    - Muté au niveau du gène *MSH2* (Δmsh2) : souche BY4741-06240

Les souches de levures utilisées font partie des souches dites « knockout » qui ont été créées en utilisant une stratégie basée sur la PCR. Deux réactions de PCR ont été réalisées.

La première PCR permet l'ajout d'étiquettes (séquences spécifiques) au gène de résistance aux antibiotiques. Les amorces UPTAG et DNTAG amplifient ainsi le gène *KanMX* dont l'expression confère résistance à la généticine (G418) chez la levure (Wach et al., 1994). Les amorces sont constituées de séquences génomiques flanquantes de l'ORF ciblée, de séquence commune à toutes les disruptions géniques pour l'amplification des « codes-barres moléculaires », d'une séquence unique pour le code-barres moléculaire TAG, et d'une séquence homologue à la cassette KanMX4 (Figure 102).

La seconde PCR a eu pour objectif d'incorporer les sites de recombinaison mitotique. Deux oligonucléotides spécifiques de l'ORF ciblée sont utilisés pour étendre la zone d'homologie, augmentant la spécificité de ciblage lors de la recombinaison mitotique de la cassette de disruption génique.

Chaque « cassette » contient une séquence unique connue sous le nom de "code-barres moléculaire" permettant une analyse en parallèle. Une étiquette d'ADN (séquence commune) flanquant les cassettes est également incorporée, permettant l'amplification des étiquettes uniques.

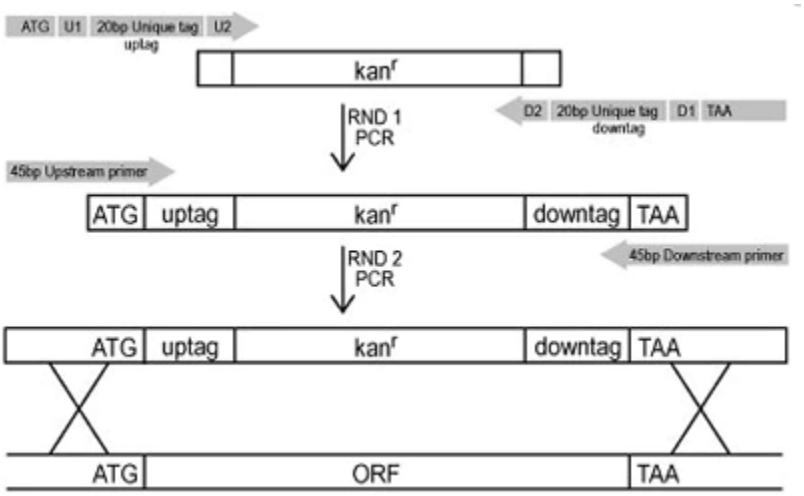

Figure 102 : Stratégie de délétion des souches. Schéma du fournisseur : SnapGene

1.3. Vecteurs et inserts

Différents vecteurs plasmidiques ont été utilisés au cours de cette thèse qui ont tous été synthétisés et produits en qualité industrielle par Genscript (200 µg pour les plasmides bactériens et 400 µg pour les plasmides de levures). Tous les plasmides ont été vérifiés par séquençage Sanger, et également digestion suivie d'une migration sur gel d'agarose par Genscript avant expédition.

Les types de microsatellites contenus dans ces plasmides sont les mono-nucléotidiques A, di-nucléotidiques AC et AT avec des répétitions allant de 15 à 25. Chaque plasmide utilisé à son état d'arrivée sera présenté dans ce manuscrit comme « pur » et représentant le T0.

1.3.1. Stratégie

L'objectif a été de pouvoir utiliser un même plasmide bactérien contenant des microsatellites pour les différentes parties du projet de thèse. Ces mêmes séquences artificielles ont également été insérées dans un plasmide de levure pour l'étude des mutations microsatellites dans cet organisme, avec une légère modification : deux index sont présents de part et d'autre de la séquence microsatellite. Ainsi, je vais vous présenter le design des séquences permettant les différentes applications de mon projet de thèse (Figure 103).

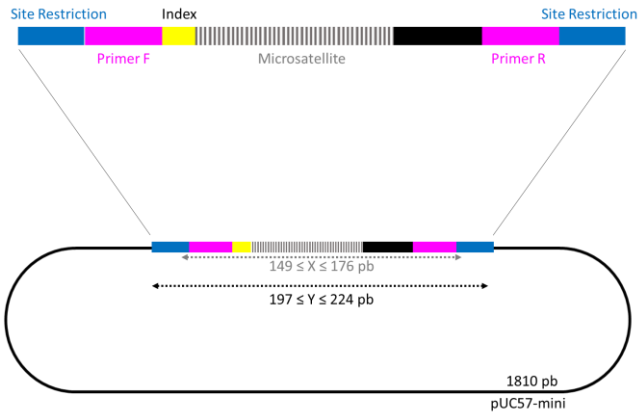

Figure 103 : Représentation schématique de la construction complète (plasmide et insert) pour les expérimentations in vitro et in vivo bactérienne.

1.3.2. Vecteurs

1.3.2.1. Bactérien

Pour toutes les expérimentations *in vitro* et les expériences *in vivo* dans les bactéries, le plasmide pUC57-Mini (Genscript) a été utilisé avec une insertion des séquences d'intérêt au niveau du site unique EcoRV (Figure 104).

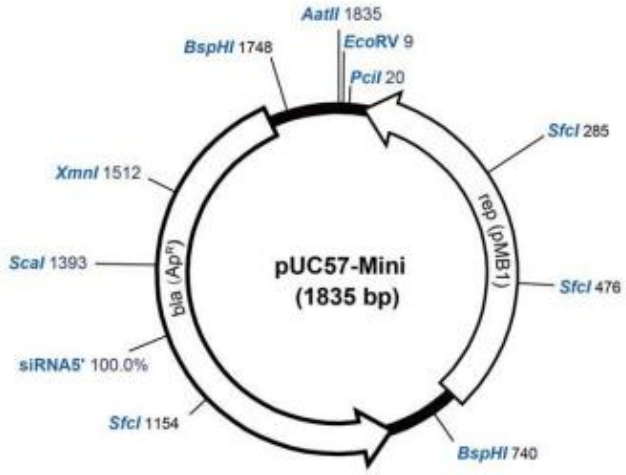

Figure 104 : Plan du vecteur pUC57-Mini.

1.3.2.2. Levures

Pour toutes les expérimentations *in vivo* dans les levures, le plasmide pESC-URA (Genscript) (Figure 105) a été utilisé avec quelques modifications. En effet, des séquences non essentielles du plasmide ont été enlevées par excision d'une cassette de 1 932b par les enzymes PsiI et MluI (f1 Ori - ADH1 terminator - FLAG tag - GAL1,10 promoter – Myc - CYC1 terminator) pour réduire la taille du plasmide et améliorer la transfection. L'insert a été cloné entre ces deux sites de digestion, incluant un signal de polyadénylation qui a été perdu précédemment par la double digestion du plasmide.

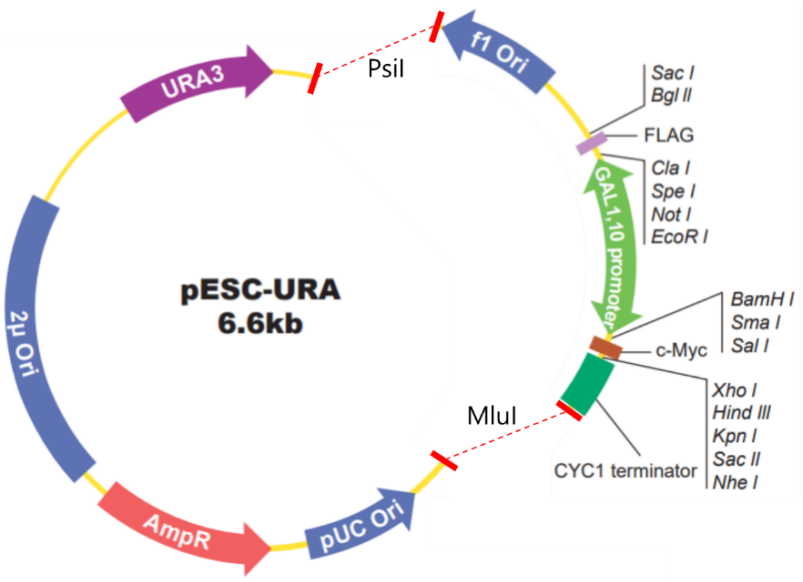

Figure 105 : Plan du vecteur pESC-URA avant modification.

1.3.3. Insert microsatellites

Les inserts de microsatellites ont été synthétisés par le fournisseur Genescript selon nos demandes avant d'être clonés dans leurs plasmides respectifs. Leurs séquences sont présentées dans le Tableau 16.

1.3.4. Oligonucléotides

Différents oligonucléotides ont été utilisés au cours de cette thèse permettant une vérification d'insert ou de souche, ou une amplification simple (Tableau 17).

Tableau 16 : Séquences des insert microsatellites (bleu : site de restriction, rose : primer HT17-145, jaune : index de reconnaissance du motif, gris : index de reconnaissance unique, vert : zone flanquante, ocre : signal de polyadénylation)

Bactéries	A15	TCACTGGGTCGATcctgtccatccattggaattgagttagaCATGTAGTCAAAAAAAAAAAAAAAtctctggtaagtgttcatgtgctcatgaa gtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTT
	A19	TCACTGGGTCGATcctgtccatccattggaattgagttagaTGATTAGTCAAAAAAAAAAAAAAAAAAAtctctggtaagtgttcatgtgctc atgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTT
	A20	TCACTGGGTCGATcctgtccatccattggaattgagttagaGTGATAGTCAAAAAAAAAAAAAAAAAAAAtctctggtaagtgttcatgtgct catgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTT
	A24	TCACTGGGTCGATcctgtccatccattggaattgagttagaTCCATAGTCAAAAAAAAAAAAAAAAAAAAAAAAtctctggtaagtgttca tgtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTT
	A25	TCACTGGGTCGATcctgtccatccattggaattgagttagaCAGCTAGTCAAAAAAAAAAAAAAAAAAAAAAAAAtctctggtaagtgtt catgtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTT
	AC15	TCACTGGGTCGATcctgtccatccattggaattgagtTGTATGTGACACACACACACACACACACACACACACACtatggtaagtgttcat gtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTT
	AC19	TCACTGGGTCGATcctgtccatccattggaattgagtTGCTTGTGACACACACACACACACACACACACACACACACACACACACtatg gtaagtgttcatgtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTT T
	AC20	TCACTGGGTCGATcctgtccatccattggaattgagtTTATCGTGACACACACACACACACACACACACACACACACACACACACACtaa tggtaagtgttcatgtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAG TTT
	nAC20	TCACTGGGTCGATcctgtccatccattggaattgagtTCTCGGTGACACACACACACACACACACACACACACACACACACACACACtaa tggtaagtgttcatgtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAG TTT
	AC24	TCACTGGGTCGATcctgtccatccattggaattgagtTCTTAGTGACACACACACACACACACACACACACACACACACACACACACA CACACACtatggtaagtgttcatgtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCA GTAGCCCAGTTT
	AC25	TCACTGGGTCGATcctgtccatccattggaattgagtTAGCAGTGACACACACACACACACACACACACACACACACACACACACACA CACACACACtatggtaagtgttcatgtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCC AGTAGCCCAGTTT
	AT15	TCACTGGGTCGATcctgtccatccattggaattgagtTAGTCGTGATATATATATATATATATATATATATATcgtggtaagtgttcatgtgctc atgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTT
	AT20	TCACTGGGTCGATcctgtccatccattggaattgagtTTCAGGTGATATATATATATATATATATATATATATATATATATcgtggtaagt gttcatgtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTT
	AT25	TCACTGGGTCGATcctgtccatccattggaattgagtTGTCAGTGATATATATATATATATATATATATATATATATATATATATATA TcgtggtaagtgttcatgtgctcatgaagtgtgcaatactttcctgcattagtagaatactagagatTTTCCGTCACAGATGCAGTTCCAGTAGCCC AGTTT
Levures	yA24	GAATTCAAGCCTGGATCCTTATAATGCATGTATACTAAACTCACAAATTAGAGCTTCAAGCTATTACGCCAGCTGAATTGGA GCGACCTCCGCCGCTACAGGGCGCGTCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCT CTTCGCTATTACGCCAGCTGAATTGGAGCGACCTCATGCTATACCTGAGAAAGCAACCTGACCTACAGGAAAGAGTTACTC AAGAATAAGAATTTTCGTTTTAAAACCTAAGAGTCACTTTAAAATTTGTATACACTTATTTTTTTTATAACTTATTTAATAATAA AAATCATAAATCATAAGAAATTCGCTTATTTAGAAGTGTCAACAACGTATCTACCATCACTGGGTCGATcctgtccatccattggaat tgagttagaCCTATAGTCAAAAAAAAAAAAAAAAAAAAAAATCtctggtaagtgttcatgtgctcatCCTATGTGCAATACTTTCCTG CATTAGTagaATACTAGAGATTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTTGGATCCCGTGACCTCGAG
	yAC24	GAATTCAAGCCTGGATCCTTATAATGCATGTATACTAAACTCACAAATTAGAGCTTCAAGCTATTACGCCAGCTGAATTGGA GCGACCTCCGCCGCTACAGGGCGCGTCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCT CTTCGCTATTACGCCAGCTGAATTGGAGCGACCTCATGCTATACCTGAGAAAGCAACCTGACCTACAGGAAAGAGTTACTC AAGAATAAGAATTTTCGTTTTAAAACCTAAGAGTCACTTTAAAATTTGTATACACTTATTTTTTTTATAACTTATTTAATAATAA AAATCATAAATCATAAGAAATTCGCTTATTTAGAAGTGTCAACAACGTATCTACCATCACTGGGTCGATcctgtccatccattggaat tgagtTCTGAGTGACACACACACACACACACACACACACACACACACACACACACACACACACACACACTATggtaagtgttcatgtgctcatC TGATGTGCAATACTTTCCTGCATTAGTagaATACTAGAGATTTTCCGTCACAGATGCAGTTCCAGTAGCCCAGTTTGGATCCCG TGACCTCGAG

Tableau 17 : Informations concernant les oligonucléotides.

Noms		Séquences	Applications	Taille amplicon	Fournisseurs	Purification
HT17_145	HT17_145_F	CCTGTCCATCCATTGGAATTGAGT	Contrôle & Projet 1 et 3	$(A)_n$: 142-152 $(AC)_n$: 150-170 $(AT)_n$: 150-170 yA_{24} : 151 $y(AC)_{24}$: 168	Eurogentec Eurofins	RP-Cartridge-Gold HPSF
	HT17_145_R	GGAACTGCATCTGTGACGGAAA				
FAM_HT17_145	FAM_HT17_145_F	FAM-CCTGTCCATCCATTGGAATTGAGT			Eurogentec Eurofins	HPLC RP-Cartridge-Gold
	HT17_145_R	GGAACTGCATCTGTGACGGAAA				
MutS	MutS_Ups_F1	TCGCCGTTCGCTTCTTCC	Bactéries Ctrl MutSd Projet 3		Eurogentec	RP-Cartridge-Gold
	MutS_Dwn_R1	TTGATAGCGTGTAGATGG				
MutL	MutL_Ups_F1	GAGCGACGATTACCAACA	Bactéries Ctrl MutLd Projet 3		Eurogentec	RP-Cartridge-Gold
	MutL_Dwn_R1	ATCAACGCTTATCAACTCTAC				
MutS_Kan_1	MutS_Ups_F1	TCGCCGTTCGCTTCTTCC	Bactéries Ctrl MutSd + Résistance Kanamycine Projet 3		Eurogentec	RP-Cartridge-Gold
	K1_R	CAGTCATAGCCGAATAGCCT				
MutS_Kan_2	K2_F	CGGTGCCCTGAATGAACTGC	Bactéries Ctrl MutSd + Résistance Kanamycine Projet 3		Eurogentec	RP-Cartridge-Gold
	MutS_Dwn_R1	TTGATAGCGTGTAGATGG				
MutL_Kan_1	MutL_Ups_F1	GAGCGACGATTACCAACA	Bactéries Ctrl MutLd + Résistance Kanamycine Projet 3		Eurogentec	RP-Cartridge-Gold
	K1_R	CAGTCATAGCCGAATAGCCT				
MutL_Kan_2	K2_F	CGGTGCCCTGAATGAACTGC	Bactéries Ctrl MutLd + Résistance Kanamycine Projet 3		Eurogentec	RP-Cartridge-Gold
	MutL_Dwn_R1	ATCAACGCTTATCAACTCTAC				
Y_MLH1	Y_MLH1_F	GGCGGCTGTTGATGTTAA	Levures Ctrl MLH1d Projet 3		Eurogentec	RP-Cartridge-Gold
	Y_MLH1_R	ACTATTGGCTGTCGTATATGAAT				
Y_MSH2	Y_MSH2_F	GTTGGATGCCTCCGCTATT	Levures Ctrl MSH2d Projet 3		Eurogentec	RP-Cartridge-Gold
	Y_MSH2_R	ATTGGTCAGTGGTTGCTTCA				
Y_RPL21A	Y_RPL21A_F	AGGCTTTTAACGGGCTGTAAG	Gène de levure présent en copie unique dans le génome (Projet 3)		Eurogentec	RP-Cartridge-Gold
	Y_RPL21A_R	AGATGCGTCCCAGCACATAA				
dnaC	dnaC_F1	AGAACAACTGCTTAACGAT	Gène bactérien présent en copie unique dans le génome (Projet 3)		Eurogentec	RP-Cartridge-Gold
	dnaC_R1	GGTCATCTCTTCCATATTACT				
BAT26	F_BAT26_F2	FAM-AACTCTTATCAGATGATTCCAA	Marqueurs génétiques Projet 3	195	Eurofins	HPLC HPSF
	BAT26_R2	GCTTCTTCAGTATATGTCAATG				
NR24	NR24_F1	GCTGAATTTTACCTCCTGAC	Marqueurs génétiques Projet 3	128	Eurofins	HPSF HPLC
	F_NR_24_R1	FAM-ATTGTGCCATTGCATTCCAA				
CAT25	F_CAT25_F2	FAM-CTGCTTATCTGAAACTTCCCAACT	Marqueurs génétiques Projet 3	122	Eurofins	HPLC HPSF
	CAT25_R2	GAGCTGAGATCGTGCCACTG				
D4S403	D4S403_F	AGGTGGCCCTGAGTAGGAGT	Marqueurs génétiques Projet 3	217-231	Eurofins	HPSF HPLC
	FAM_D4S403_R	FAM-TTTGAGGGAATGATTTGGGT				
D3S3630	FAM_D3S3630_F	FAM-AAGGGATAAGCTGCAAATCA	Marqueurs génétiques Projet 3	172-188	Eurofins	HPLC HPSF
	D3S3630_R	ACCAAATACAATTCATGAGACCTGA				
D14S1007	D14S1007_F	AGCTCCTATATGTCTTCACACAG	Marqueurs génétiques Projet 3	114-132	Eurofins	HPSF HPLC
	FAM_D14S1007_R	FAM-CTCCATTCCCATACGTCC				
D9S1838	FAM_D9S1838_F	FAM-ACCCAGCTACTGAGGAGGCTT	Marqueurs génétiques Projet 3	159-175	Eurofins	HPLC HPSF
	D9S1838_R	GCTTCTGCACTTTGTAGAACCAAAT				
D6S305	FAM_D6S305_F	FAM-CACCAGCGTTAGAGACTGC	Marqueurs génétiques Projet 3	204-230	Eurofins	HPLC HPSF
	D6S305_R	GCAAATGGAGCATGTCACT				
D2S123	F_D2S123_F	FAM-AAACAGGATGCCTGCCTTTA	Marqueurs génétiques Projet 3	197-227	Eurofins	HPLC HPSF
	D2S123_R	GGACTTTCCACCTATGGGAC				
D20S851	FAM_ D20S851_F	FAM-ACTTCAAGTTATGTGTGGCACAA	Marqueurs génétiques Projet 3	126-150	Eurofins	HPLC HPSF
	D20S851_R	GCCCAGACTCTGACACCTTT				
D13S285	FAM_D13S285_F	FAM-ATATATGCACATCCATCCATG	Marqueurs génétiques Projet 3	92-106	Eurofins	HPLC HPSF
	D13S285_R	GGCCAAAGATAGATAGCAAGGTA				
D9S1825	D9S1825_F	AGCTGCTTCCAAAACACA	Marqueurs génétiques Projet 3	127-145	Eurofins	HPSF HPLC
	FAM_ D9S1825_R	FAM-CTGTATAGGTTTCTGATGGACAAAG				
D18S61	F_D18S61_F1	FAM-ATTTCTAAGAGCACTCCCAAACT	Marqueurs génétiques Projet 3	157-183	Eurofins	HPLC HPSF
	D18S61_R1	ATATTTTGAAACTCAGGAGCAT				

Ctrl : Contrôle - RP-Cartridge-Gold : Chromatographie basée sur la différence d'hydrophobicité entre le produit à pleine longueur (qui contient encore le groupe 4,4'-diméthyltrityl (DMT) au niveau du 5' OH de la dernière base de l'oligonucléotide synthétique) et les séquences tronquées (sans le groupe DMT) (Eurogentec) - HPSF : High Purity Salt Free, Technologie de purification sur cartouche basée sur la filtration sur gel (Eurofins) - HPLC : High Performance Liquid Chromatography.

2. Milieux de culture

2.1. Milieux de culture bactériens

Le milieu de base utilisé avec les bactéries est le milieu *Lysogeny Broth* (LB) Lennox, composé pour 1 L de 10 g de tryptone, 5 g d'extrait de levure et 5 g de chlorure de sodium. Ce milieu fournit une source riche en nutriments essentiels tels que les acides aminés, les vitamines et les sels minéraux, nécessaires à la croissance et à la prolifération bactérienne. Comparés aux milieux Luria et Miller, les milieux LB Lennox sont favorisés pour la culture des souches dérivées de *E.coli* K-12 car ils apportent les conditions osmotiques optimales (le milieu Luria a deux fois moins de NaCl et le milieu Miller en a deux fois plus). Les milieux utilisés sont soit liquides pour la culture liquide soit gélifiés pour de la culture solide. Les milieux liquides ont été soit achetés directement en solution (LB Broth, Gibco™), ou préparés à partir de poudre ajoutée à de l'eau distillée/ultrapure tout comme les milieux solides (LB Broth Base (Lennox), LB Agar (Lennox L Agar), Invitrogen™), qui sont ensuite stérilisés 20 min à 121°C. Afin de pouvoir effectuer des sélections de clones, des antibiotiques ont été achetés directement en solution (Kanamycin Sulfate, Gibco™) ou préparés à partir de poudre (Ampicillin, sodium salt, irradiated, Gibco™), resuspendus avec de l'eau ultrapure et filtrés à 20 µm. Ils ont ensuite été ajoutés au milieu liquide au moment de l'utilisation ou au milieu gélifié lorsqu'il était encore liquide et refroidi entre 50 et 55°C (Tableau 18).

Tableau 18 : Détails de la composition des différents milieux de culture pour bactérie.

	Milieu liquide	Milieu gélosé
LB	20 g/L LB Broth	32 g/L LB Broth Agar
LB Kan^+	20 g/L LB Broth 30 µg/mL Kanamycine	32 g/L LB Broth Agar 30 µg/mL Kanamycine
LB Amp^+	20 g/L LB Broth 50 µg/mL Ampiciline	32 g/L LB Broth Agar 50 µg/mL Ampiciline
LB Kan^+ Amp^+	20 g/L LB Broth 30 µg/mL Kanamycine 50 µg/mL Ampiciline	32 g/L LB Broth Agar 30 µg/mL Kanamycine 50 µg/mL Ampiciline

2.2. Milieux de culture pour la levure

Le milieu de base utilisé avec les levures est le milieu *Yeast Extract Peptone Dextrose* (YEPD ou YPD) (BD Bouillon YPD, Difco™), composé pour 1 L de 10 g d'extrait de levure, 20 g de peptone et 20 g de glucose (dextrose). Ce milieu fournit une source riche en nutriments essentiels tels que les acides aminés, les vitamines et les sels minéraux, nécessaires à la croissance et à la prolifération des levures.

Certaines de nos souches et nos mutants étant auxotrophes à certains acides aminés, nous avons utilisé un milieu plus adapté à leurs sélections. Le milieu *Yeast Nitrogen Base* (YNB) (*Yeast Nitrogen Base*, Invitrogen™) est un milieu synthétique contenant le nécessaire au bon développement des levures à l'exception de la source de carbone et des acides aminés. Il a donc été supplémenté en glucose et en acide aminé selon les besoins.

Les milieux utilisés sont soit liquides pour la culture liquide, soit gélifiés pour de la culture solide. Les milieux liquides ont été préparés à partir de plusieurs poudres ajoutées à de l'eau distillée/ultrapure. Les milieux gélosés ont été préparés comme les milieux liquides supplémentés en agar (Agar gélifiant en poudre 500g, HIMEDIA™). Dans le cas des milieux YNB, des compléments de milieu *Yeast Synthetic Drop-Out* (S. Do) ont été ajoutés afin de déclencher l'autotrophie des levures recherchées, en omettant

certains acides aminés (*Yeast Synthetic Drop-out Medium Supplements without lysine*, *Yeast Synthetic Drop-out Medium Supplements without uracil*, Sigma-Aldrich™). Les milieux ont ensuite été stérilisés 20 min à 121°C. Afin de pouvoir effectuer une sélection de clones supplémentaires, des antibiotiques ont été achetés directement en solution (Antibiotique sélectif Geneticin™ (G418 sulfate), Gibco™) ou préparés à partir de poudre (Antibiotique sélectif Geneticin™ (G418 sulfate), Gibco™), resuspendue avec de l'eau ultrapure et filtrés à 20 µm. Ils ont ensuite été ajoutés au milieu liquide au moment de l'utilisation ou au milieu gélifié lorsqu'il était encore liquide et refroidi, à une température comprise entre 50 et 55°C (Tableau 19).

Tableau 19 : Détails de la composition des différents milieux de culture pour levure.

	Milieu liquide	Milieu gélosé
YPD	50 g/L YPG *Broth*	50 g/L YPG *Broth* 20 g/L Agar
YPD Gen^+	50 g/L YPG *Broth* 200 µg/mL *Geneticin*	50 g/L YPG *Broth* 20 g/L Agar 200 µg/mL *Geneticin*
YNB Lys^-	6.75 g/L YNB 1,92 g/L S. Do-Lys 2% Glucose	6,75 g/L YNB 20 g/L Agar 1,92 g/L S. Do-Lys 2% Glucose
YNB Lys^- Gen^+	6,75 g/L YNB 1,92 g/L S. Do-Lys 2% Glucose 200 µg/mL *Geneticin*	6,75 g/L YNB 20 g/L Agar 1,92 g/L S. Do-Lys 2% Glucose 200 µg/mL *Geneticin*
YNB Ura^-	6,75 g/L YNB 1,92 g/L S. Do-Ura 2% Glucose	6,75 g/L YNB 20 g/L Agar 1,92 g/L S. Do-Ura 2% Glucose
YNB Ura^- Gen^+	6.75 g/L YNB 1.92 g/L S. Do-Ura 2% Glucose 200 µg/mL *Geneticin*	6.75 g/L YNB 20 g/L Agar 1.92 g/L S. Do-Ura 2% Glucose 200 µg/mL *Geneticin*

2.3. Quantification des bactéries et levures en culture liquide par mesure de la densité optique

La mesure de la densité optique (DO) est une technique couramment utilisée pour évaluer la croissance des bactéries et des levures en culture. Cette méthode repose sur la capacité des cellules en suspension à disperser la lumière, permettant ainsi de quantifier indirectement la concentration cellulaire. La DO est mesurée à l'aide d'un spectrophotomètre, généralement à une longueur d'onde de 600 nm (DO600), où les cellules vivantes et mortes absorbent la lumière de manière similaire, donnant une indication de la biomasse totale.

L'absorption de la lumière est décrite par la loi de Beer-Lambert : $\frac{I_t}{I_o} = 10^{\varepsilon C l}$, plus généralement exprimée sous la forme du rapport $\log\left(\frac{I_0}{I_t}\right) = A = DO = \varepsilon . C . l$, communément appelée absorbance (A) ou densité optique DO, où la concentration de l'espèce absorbante (C) est exprimée en mol/L de l'espèce

absorbante, le trajet optique (l), exprimé en centimètres et le coefficient d'absorption molaire (ε) exprimé en $L.mol^{-1}.cm^{-1}$ (Figure 106).

Un blanc aura été réalisé avec le milieu de culture stérile utilisé pour ladite culture afin de normaliser l'absorption basale du milieu. Afin de permettre une bonne estimation de la concentration cellulaire dans la culture, l'absorbance au spectrophotomètre ne doit pas dépasser le seuil de linéarité de l'appareil se situant aux alentours de 0,6 (Figure 106). Pour ce faire, un prélèvement compris entre 100 µL et 1 mL, dépendant de la phase de la culture, a été réalisé puis déposé dans une semi-microcuve. Ce prélèvement a par la suite été dilué par l'ajout de milieu stérile jusqu'à obtenir entre 0,5 et 1 mL pour lecture dans le spectrophotomètre (Densimètres de 10 cellules Ultrospec, Harvard Biochrom).

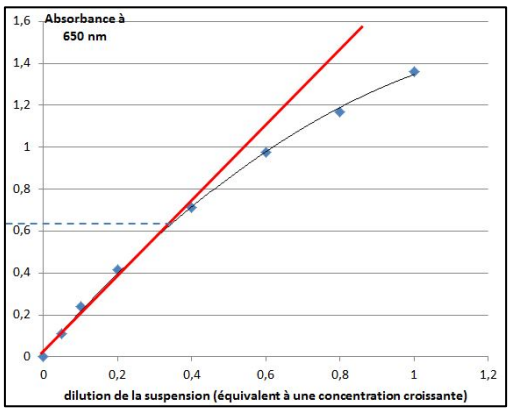

Figure 106 : Courbe de perte de linéarité entre densité optique et concentration bactérienne.

Dans le cas des bactéries, il a été estimé empiriquement qu'une DO égale à 1 correspondait à environ $1,5.10^9$ UFC/mL alors que pour les levures, elle correspondait à environ 1.10^8 UFC/mL. Ces valeurs ont été utilisées pour la quantification des micro-organismes dans la loi de Beer-Lambert.

2.4. Sélection de clones

Lors de manipulations sur des micro-organismes tels que les levures et les bactéries, il est important de garantir la reproductibilité génétique des expériences. Il est important de s'assurer que les micro-organismes que nous allions inoculer avaient le même pool génétique entre eux et donc provenaient d'un même micro-organisme de départ.

2.4.1. Isolement

Les micro-organismes (WT ou muté) étant fournis en tubes stockés à -80°C, il fallait s'assurer qu'ils étaient viables et sélectionner un clone unique.

Pour cela, un cône stérile de 100 µL a été utilisé pour gratter la surface des tubes de micro-organismes afin de ne pas décongeler l'intégralité du tube et ainsi garantir la bonne conservation du reste des micro-organismes. Ce qui fut gratté a été resuspendu dans 50 µL de milieu LB pour les bactéries et YPD pour les levures. Enfin, la suspension a été déposée dans un coin d'une boîte de petri contenant du milieu LB gélosé pour les bactéries et du milieu YPD gélosé pour les levures afin d'effectuer un isolement par la méthode des cadrans (Figure 107). Les isolements ont ensuite été cultivés à 37°C pour les bactéries et 30°C pour les levures dans une étuve entre 24h et 48h.

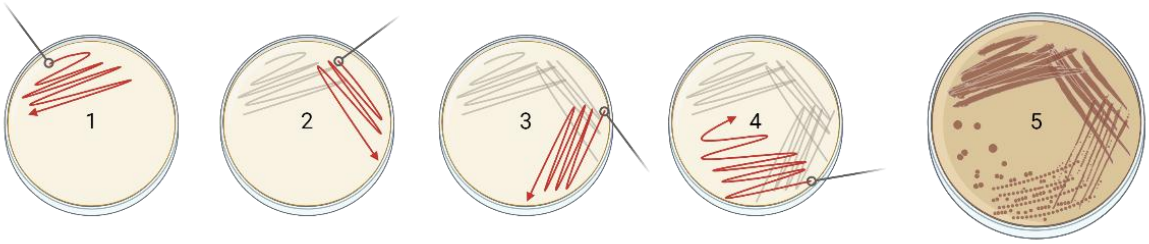

Figure 107 : Représentation schématique d'un isolement par la méthode des cadrans.

Cette technique d'isolement par cadrans permet d'obtenir des colonies isolées à partir d'une suspension concentrée des micro-organismes.

2.4.2. Strike

Une fois l'isolement réalisé, des colonies provenant d'un micro-organisme unique ont été identifiées. Toutefois, une colonie n'est pas suffisante pour réaliser l'intégralité des expériences. Pour obtenir une quantité plus importante d'un même clone, la colonie sélectionnée a été récupérée et étalée en ligne ou zigzag, appelée « *strike* » sur une gélose de LB pour les bactéries ou de YPD pour les levures (Figure 108). Les strikes ont ensuite été cultivé à 37°C pour les bactéries et 30°C pour les levures dans une étuve entre 24h et 48h. Cette technique permet d'obtenir un grand nombre de représentants d'un même clone de départ. Les bactéries et levures « strikées » sont conservées pendant au plus 3 mois à 4°C, au-delà, les organismes sont morts. Toutefois au bout de 2 mois, ils sont « strikés » à nouveau puis conservés à 4°C pour maintenir la viabilité des microorganismes.

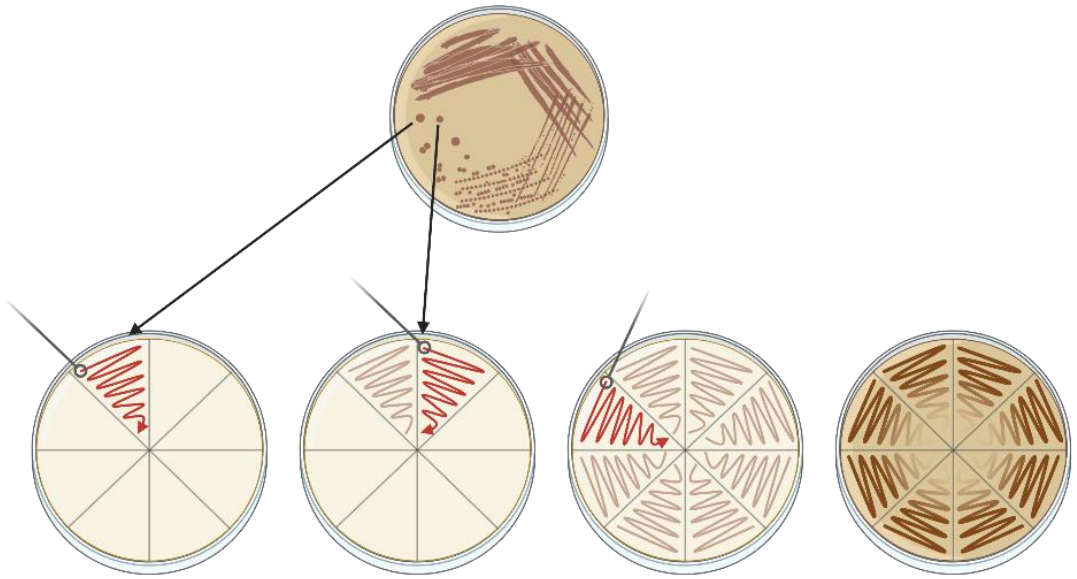

Figure 108 : Représentation schématique de strike de 8 colonies sur une boîte de petri.

2.5. Culture liquide de bactéries et levures non transformées en vue de la compétence

La mise en culture liquide de bactéries ou de levures nécessite une étape de préculture. En effet, les *strikes* étant conservés à 4°C, il est nécessaire de réacclimater les micro-organismes à leurs températures optimales de croissance dans un milieu nutritif optimal.

C'est pourquoi, dans des tubes à double clic, 3 mL de milieu LB pour les bactéries, ou 3 mL de milieu YPD pour les levures ont été ajoutés. Une petite surface d'un *strike* a été grattée avec une pointe de cône P200 µL et resuspendue dans le milieu du tube à double clic. Ce tube a été ensuite placé dans une étuve à 37°C pour les bactéries et 30°C pour les levures, avec une légère agitation à 250 rpm pour la nuit.

Une fois la préculture récupérée, une mesure de DO est réalisée et un volume équivalent à 1.10^8 UFC/mL de bactéries ou 1.10^7 UFC/mL de levures a été inoculé dans 150 mL de milieu LB pour les bactéries et YPD pour les levures. Ces suspensions ont été incubées à 37°C pour les bactéries et 30°C pour les levures sous une agitation de 250 rpm jusqu'à obtenir un DO de 6 pour les bactéries et 4 pour les levures.

En parallèle, un étalement de 100 µL de l'inoculât a été réalisé après dilution au 1/1 000 ou 1/10 000 sur boîte de pétri, contenant le milieu associé. Les boîtes ont été incubées à 37°C ou 30°C pendant la nuit pour permettre le dénombrement réel de l'inoculât.

Une fois la culture liquide ayant atteint sa DO d'arrêt, une aliquote de 100 µL a été diluée au 1/100 000 et étalée sur boîte de pétri, contenant le milieu associé. Les boîtes ont été incubées à 37°C ou 30°C pendant la nuit pour permettre le dénombrement réel de la culture finale. Enfin, le reste de la culture a été séparé en fraction de 50 mL et tous ces échantillons ont été centrifugés pendant 25 minutes à 4600 rpm pour recueillir le culot de micro-organisme, qui a ensuite été conservé à -80°C pour des utilisations futures.

2.6. Compétence

La compétence d'un micro-organisme tel que la bactérie ou la levure repose sur la capacité de cet organisme à devenir temporairement perméable à l'ADN exogène. Ainsi, il est possible de l'incorporer dans leur propre génome ou, dans le cas de plasmide, de s'insérer dans cet organisme. Ce processus commence par la reconnaissance et la liaison de l'ADN libre dans l'environnement à la surface cellulaire, suivie de son transport à travers la membrane cellulaire. Une fois à l'intérieur de la cellule, cet ADN exogène peut s'intégrer dans le génome via des mécanismes de recombinaison homologue, où il remplace ou complète des segments d'ADN existants. Dans le cas des plasmides, l'ADN ne s'intègre généralement pas ou peu dans le génome mais existe et se réplique de manière autonome dans le cytoplasme. L'état de compétence est généralement induit par des signaux environnementaux tels que la densité cellulaire élevée ou des conditions de stress (choc thermique ou chimique), qui activent les mécanismes nécessaires à l'absorption de l'ADN. Ces mécanismes permettent aux organismes de rapidement développer de nouvelles capacités et de s'adapter efficacement à des conditions changeantes, renforçant ainsi leur survie et leur compétitivité. Dans ma thèse, j'ai utilisé des bactéries et levures rendues compétentes chimiquement en vue d'une transformation thermique.

2.6.1. Bactéries E. coli

Pour rendre les bactéries Keio K12 compétentes, les bactéries ont été inoculées à partir d'un clone isolé en boîte de Petri, et cultivées dans 3 mL de milieu LB dans des tubes cristaux. La culture a été incubée toute la nuit à 37°C avec une agitation de 250 rpm. Le lendemain, 500 µL de cette suspension ont été dilués dans 55 mL de milieu LB dans un Erlenmeyer, puis incubés à 37°C sous agitation à 250 rpm jusqu'à atteindre une densité optique mesurée à 600 nm (D.O.600) comprise entre 0,5 et 0,7. Lorsque la DO nécessaire a été atteinte, la culture bactérienne a été placée sur glace pendant 20 minutes puis les cellules ont été sédimentées pendant 10 minutes à 4°C à 4000 rpm pour obtenir un culot bactérien. Ce culot a été remis en suspension dans 25 mL de $MgCl_2$ froid avec une concentration de à 0,1M. La suspension a été placée de nouveau sur glace pendant 30 minutes pour permettre une meilleure stabilisation des membranes cellulaires, puis sédimentée à nouveau pendant 10 minutes à 4°C à 4000 rpm. Le culot obtenu a été repris dans 5 mL de $CaCl_2$ froid à 0,1 M et la suspension a été incubée sur glace pendant au moins 2 heures. Ensuite, 1,75 mL de glycérol 80% ont été ajoutés à la suspension pour atteindre une concentration finale de 20% de glycérol, assurant ainsi la protection des cellules lors de la congélation.

La suspension finale a été aliquotée en fractions de 220 µL, qui ont été soit utilisées directement pour des transformations, soit plongées directement dans de l'azote liquide puis conservées à -80°C pour une utilisation ultérieure.

Concernant les bactéries 5-alpha compétentes, elles ont été achetées directement compétentes (NEB® 5-alpha Competent E. coli (High Efficiency) (New England Biolabs®))

2.6.2. Levures S. cerevisiae

Pour préparer des levures compétentes, le protocole du kit Frozen-EZ Yeast Transformation II Kit™ de Zymo Reaserch® a été suivi.

Les levures ont été cultivées à 30°C dans 10 mL de milieu YPD jusqu'à ce que la D.O.600 atteigne une valeur comprise entre 0,8 et 1,0. Une fois cette densité atteinte, la culture a été sédimentée pendant 4 minutes à 500 g pour obtenir un culot.

Après sédimentation, le surnageant a été éliminé, et 10 mL de Frozen-EZ Yeast Solution 1 ont été ajoutés au culot pour laver les cellules. La suspension a été sédimentée une seconde fois pendant 4 minutes à 500 g. Le surnageant a de nouveau été éliminé. Pour resuspendre le culot, 1 mL de Frozen-EZ Yeast Solution 2 a été ajouté.

La suspension a été soigneusement homogénéisée, puis aliquotée en fractions de 100 µL. Les aliquotes ainsi préparées ont été soit utilisées immédiatement pour des transformations, soit conservées par congélation lente à -80°C (les aliquotes sont mis dans une boîte en polystyrène puis placées dans un congélateur à -80°C) pour une utilisation ultérieure.

2.7. Transformation

2.7.1. Bactéries (Choc Thermique)

Durant ma thèse, deux types de bactéries compétentes ont été utilisés pour les transformations. Soit les bactéries qui ont été rendues compétentes au laboratoire pour les transformer avec des

plasmides purs T0, soit des bactéries compétentes commerciales NEB® 5-alpha Competent *E. coli* (High Efficiency) (New England Biolabs®) pour les transformer avec des plasmides extraits d'une première culture de levure.

Pour la transformation des bactéries K12 compétentes, les cellules bactériennes ont d'abord été décongelées sur glace. Entre 30 et 70 ng de plasmides ($(AC)_{24}$ ou A_{24}) ont été ajoutés à 200 µL de bactéries compétentes, suivi d'une incubation de 30 minutes sur glace. Après cette période d'incubation, la suspension a été soumise à un choc thermique à 42°C pendant 45 secondes, puis replacée sur glace pendant 5 minutes. 800 µL de milieu LB ont été ajoutés à la suspension, qui a été incubée à température ambiante pour les expériences standard, et de 20-45°C pour les expériences de variation de températures de culture, avec une agitation de 250 rpm pendant 15 minutes (lors des optimisations, nous avons aussi testé des temps d'incubation allant de 0 à 45 min) permettant l'expression des gènes de résistance aux antibiotiques.

Pour la transformation des bactéries compétentes commerciales DH5α, les cellules ont d'abord été décongelées sur glace. Entre 1 et 10 µL de plasmides ($(AC)_{24}$ ou A_{24}) ont été ajoutés à 100 µL de bactéries compétentes, suivis d'une incubation de 2, 10 ou 30 minutes sur glace. Après cette période d'incubation, la suspension a été soumise à un choc thermique à 42°C pendant 45 secondes, puis replacée sur glace pendant 2, 3 ou 5 minutes pour les phases d'optimisation, et fixée à 2 minutes pour la phase de production. Entre 200 et 900 µL de milieu SOC (*Super Optimal broth with Catabolite répression*) ont été ajoutés à la suspension, qui a été incubée à 37°C avec agitation à 250 rpm pendant 15 à 60 minutes pour les phases d'optimisation et fixée à 15 min pour les phases de production, permettant l'expression des gènes de résistance aux antibiotiques.

Après cette incubation, 100 µL de la suspension ont été prélevés et étalés sur les différents milieux solides utilisés (Tableau 18) à une dilution allant de 1/10 à 1/10 000ème. Les boîtes ont été incubées à 37°C pendant la nuit pour permettre la croissance des colonies transformées et ainsi déterminer le taux de transformation.

Les 900 µL restants de la culture ont été inoculés dans 150 mL de milieu LB contenant les mêmes antibiotiques (LB + Amp ou LB + Amp + Kan) et incubés à des températures entre 20°C et 45°C sous agitation à 250 rpm pendant la nuit. La culture a été arrêtée à une D.O. 600 autour de 4. Enfin, la culture a été séparée en deux. Deux échantillons de 1,5 mL ont été prélevés pour effectuer des PCR sur culture, et trois échantillons de 50 mL ont été prélevés pour des extractions plasmidiques. Tous ces échantillons ont été centrifugés pendant 25 minutes à 4600 rpm pour recueillir le culot bactérien, qui a ensuite été conservé à -80°C pour des utilisations futures.

2.7.2. Levures

Pour la transformation des levures compétentes, les cellules ont d'abord été décongelées sur glace. Entre 1 et 10 µL de plasmide à une concentration de 100 ng/µL ont été ensuite ajoutés à 50 ou 100 µL de levures compétentes. Pour faciliter la transformation, 500 ou 1000 µL de Frozen-EZ Yeast Solution 3 ont été ajoutés au mélange, qui a été mélangé énergiquement. Le mélange a ensuite été incubé à 30°C pour les expériences standard, et de 20-45°C pour les expériences de variation de températures de cultures pendant 45 minutes. Durant cette incubation, le mélange a été vortexé 2 à 3 fois pour assurer une bonne distribution des plasmides dans les cellules.

Après l'incubation, 100 µL d'une dilution au 1/500ème de la suspension transformée ont été étalés sur des boîtes de Petri contenant du milieu YNB sans uracile (YNB Ura^-) supplémenté en généticine

(Gen$^+$) (Tableau 19), puis incubés pendant 2 à 4 jours pour permettre la croissance des colonies transformées et ainsi déterminer le taux de transformations.

Le volume restant de la transformation a été inoculé dans 150 mL du même milieu (YNB Ura$^-$ Gen$^+$) et incubé à des températures allant de 20°C et 36°C. La culture a été poursuivie jusqu'à une D.O. 600 environ 6. Enfin, deux échantillons de 1,5 mL ont été prélevés pour effectuer des PCR sur culture, et deux échantillons de 50 mL ont été prélevés pour des extractions plasmidiques. Les échantillons de culture ont été centrifugés pendant 25 minutes à 46 000 rpm pour recueillir les cellules transformées. Le culot obtenu a été conservé à -80°C pour des utilisations ultérieures.

2.8. Détermination du taux de transformation et du nombre de doublement cellulaire

2.8.1. Taux de transformation

Calculer le taux de transformation est primordial afin de savoir, à partir de la quantité de micro-organismes utilisée, quelle est la réelle quantité de micro-organismes ayant intégré le plasmide et de ce fait déduire la « masse » génétique de départ de l'expérience. Si trop peu de micro-organismes sont transformés, l'échantillonnage de départ sera trop faible et ne sera pas représentatif. A l'inverse, si le taux est trop fort, il y aura un nombre de doublement trop faible avant saturation dans le milieu voire même la possibilité de l'intégration de plusieurs plasmides par organisme (contre un seul voulu). Pour cela, une comparaison, après étalement sur milieu stringeant, entre la quantité d'organisme ayant acquis la résistance, donnée par le plasmide, par rapport à la quantité d'organsime total présent nous permet une évaluation correcte du taux de transformation.

Afin de calculer le taux de transformation, la formule suivante a été appliquée :

$$Taux\ de\ Transformation = \frac{Nb.\ Organisme\ résistant\ (UFC/ml)}{Nb.\ Organisme\ total\ (UFC/ml)}$$

2.8.2. Nombre de doublement cellulaire

Calculer le nombre de doublement cellulaire est très important car il nous permettra de savoir combien de cycles de réplication les microsatellites ont subi au minimum avant l'arrêt de l'expérimentation. Il faudra y ajouter le nombre de réplication de plasmide au sein de la même cellule qui sera calculé à l'aide du nombre de copies de plasmide dans une cellule.

Afin de calculer le nombre de doublement cellulaire, la formule suivante a été appliquée :

$$Nombre\ doublement\ cellulaire = \log_2 \frac{Nb.Total\ Transfo.\ (UFC)}{Nb.Total\ Inoculé\ (UFC)}$$

3. Manipulation des acides nucléiques

3.1. Extraction d'ADN

3.1.1. ADN plasmidique de bactérie

Pour la purification de l'ADN plasmidique des bactéries, on a suivi le protocole du kit QIAGEN® Plasmid Plus Midi Kit™ (Figure 109).

Le culot préalablement obtenu a été remis en suspension dans 2 mL de Buffer P1, suivi d'un ajout de 2 mL de Buffer P2 à la solution. Le mélange a été homogénéisé par inversion et incubé à température ambiante pendant 3 min. Pendant cette incubation, le lysat est devenu visqueux. Il est important de noter que la lyse ne doit pas dépasser 5 minutes pour éviter une dégradation excessive des composants cellulaires. 2 mL de Buffer S3 ont été ensuite ajoutés à la suspension lysée. Le mélange a été inversé 4 à 6 fois pour assurer une neutralisation complète où il est devenu blanc et laiteux. Le lysat a été immédiatement transféré dans une cartouche QIAfilter. La cartouche a été laissée à température ambiante pendant 10 minutes. Après cette incubation, un piston a été inséré délicatement dans la cartouche pour filtrer le lysat dans un nouveau tube de collecte.

Une fois le lysat filtré, 2 mL de Buffer BB ont été ajoutés. Le mélange a été inversé 4 à 6 fois pour bien mélanger le tampon avec le lysat. Les colonnes de purification, munies de leurs tubes d'extension, ont été placées sur un système sous-vide de type QIAvac 24 Plus™ (QIAGEN®) réglé à environ -300 mbar. Le lysat a été transféré dans les colonnes et la sous-pression a été appliquée pour aspirer la solution. Pour le lavage des colonnes, 0,7 mL de Buffer ETR a été ajouté dans chaque colonne. La sous-pression de -300 mbar a été appliquée pour aspirer la solution de lavage. Ce processus a été répété avec 0,7 mL de Buffer PE pour un second lavage. Les colonnes ont ensuite été centrifugées pendant 1 minute à 10 000 g pour éliminer les résidus de lavage.

Après la centrifugation, les tubes de réception ont été remplacés par de nouveaux tubes propres. Pour l'élution de l'ADN, 200 µL de Buffer EB (10 mM Tris·Cl, pH 8) ont été ajoutés au centre de chaque colonne. La solution a été laissée en contact avec les membranes des colonnes pendant 1 minute pour permettre une élution optimale de l'ADN. Enfin, les colonnes ont été centrifugées pendant 1 minute à 10 000 g pour récupérer l'ADN purifié. L'éluât est conservé à -20°C.

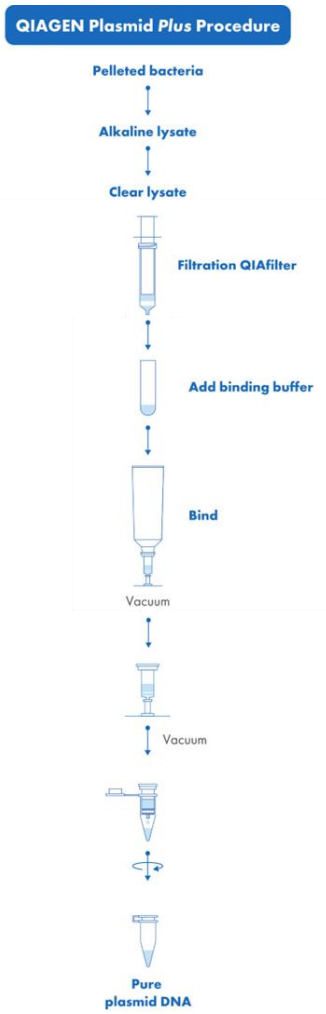

Figure 109 : Schéma représentatif de la procédure du Kit QIAGEN® Plasmid Plus Midi Kit™ (Adapté du schéma de leur site).

3.1.2. ADN génomique des levures

Pour la purification de l'ADN génomique de levures, on a suivi le protocole du kit Quick-DNA™ Fungal/Bacterial Miniprep Kit de Zymo Research® (Figure 110).

Le culot de levure préalablement obtenu, correspondant à environ 50-100 mg, a été resuspendu dans 200µL d'eau, 750 µL de BashingBead™ Buffer (tampon de lyse) contenant du détergent pour affaiblir les parois cellulaires. Pour renforcer la lyse, le tout a été mélangé à des billes de verre dans un tube ZR BashingBead™ Lysis, et les échantillons ont été soumis à un broyeur à billes à vitesse maximale pendant 5 min, ce qui permet une désintégration mécanique des parois cellulaires robustes. Le lysat a ensuite été centrifugé pendant 1 min à 10 000 g pour séparer les débris cellulaires insolubles de l'ADN en solution. 400 µL du surnageant ont été soigneusement transférés dans une colonne de purification Zymo-Spin™ III-F Filter et centrifugés à 8 000 g pendant 1 min pour séparer des quelques débris restants. Le filtrat a ensuite été mélangé à 1,2 mL de Genomic Lysis Buffer puis, en deux fois successives de 800 µL, le mélange a été transféré sur une colonne Zymo-Spin™ IICR Column et centrifugé pendant 1 min à 10 000g afin que l'ADN se lie à la membrane, tandis que les impuretés sont éliminées. Le filtrat a ensuite été jeté et 200 µL de DNA Pre-Wash Buffer a été ajouté à la colonne pour une première étape de lavage par centrifugation à 10 000 g pendant 1 min. 500 µL de g-DNA Wash Buffer a également été ajouté à la

colonne et centrifugé pendant 1 min à 10 000 g pour finaliser le lavage, puis placé dans un nouveau tube de 1,5 mL.

Enfin, l'ADN a été élué de la colonne en ajoutant 50 µL de DNA Elution Buffer et centrifugeant à 10 000 g pendant 30 sec.



Figure 110 : Schéma représentatif de la procédure du Kit Quick-DNA™ Fungal/Bacterial Miniprep Kit.

3.1.3. ADN génomique et plasmidique de levures

Pour la coextraction de l'ADN génomique et plasmidique de levure, on a suivi le protocole du kit YeaStar Genomic DNA Kit™ de Zymo Research®, initialement utilisé pour l'extraction d'ADN génomique à partir de petite quantité de biomasse de levures.

Une aliquote de 1,5 mL de culture de levure a été sédimentée à 600 g pendant 2 minutes, puis 200 µL de Solution 1 ont été ajoutés au culot, suivis de 3 µL de Zymolyase™. Les cellules ont été resuspendues soit par pichenettes, soit par un vortex léger pour assurer une bonne homogénéisation. Le mélange a été incubé à 37°C pendant une durée de 45 minutes pour permettre l'action de la Zymolyase™. 200 µL de Solution 2 ont ensuite été ajoutés au mélange, et la solution a été bien mélangée pour assurer une lyse complète des cellules. À ce stade, 400 µL de Solution 3 ont été ajoutés, et la solution a de nouveau été bien mélangée pour stabiliser l'ADN génomique et les plasmides libérés. La solution a été centrifugée à vitesse maximale pendant 3 minutes pour séparer les débris cellulaires du surnageant contenant l'ADNg. Le surnageant a été soigneusement transféré dans une colonne Zymo-Spin-I. Cette colonne a été centrifugée pendant 30 secondes pour permettre la liaison de l'ADNg à la membrane de la colonne, et le filtrat a été jeté. Pour le lavage de la colonne, 550 µL de Wash Buffer (préalablement mélangé avec de l'éthanol) ont été ajoutés à la colonne. La colonne a été centrifugée pendant 1 à 2 minutes puis placée dans un nouveau tube de 1,5 mL.

Pour éluer l'ADNg, 10 µL d'eau ont été ajoutés au centre de la membrane de la colonne. La colonne a été centrifugée pendant 1 minute pour récupérer l'ADNg purifié dans le tube de collecte.

3.1.4. ADN plasmidique des levures

Pour la purification de l'ADN plasmidique des levures, on a suivi le protocole du kit QIAGEN® Plasmid Plus Midi Kit™ présenté précédemment (cf. 3.1.1. ADN plasmidique des bactéries) avec quelques modifications pour une meilleure adaptation aux levures.

Le culot de levure préalablement obtenu a été resuspendu dans 4 mL de Buffer SCE (Tableau 20) fraîchement préparé. À cette suspension, 5 mg de Zymolase™ et 30 µL de β-mercaptoéthanol ont été ajoutés. Le mélange a été incubé à 37°C pendant 1 heure pour favoriser la digestion de la paroi cellulaire. Après incubation, la suspension a été centrifugée à 5000 rpm pendant 5 minutes pour recueillir les cellules digérées. Le culot obtenu a été resuspendu dans 4 mL de Buffer P1 et incubé à 37°C pendant 15 minutes. 2 mL de Lysis Buffer ont été par la suite ajoutés à la suspension. Le mélange a été doucement mais efficacement mélangé pour assurer une lyse complète, puis incubé à 65°C pendant 20 minutes. Pour neutraliser la solution, 3 mL de Buffer S3 ont été immédiatement ajoutés, et la suspension a été mélangée. La solution a ensuite été centrifugée à 5000 g pendant 45 minutes à 4°C. Le lysat obtenu, qui apparaît blanc et laiteux, a été transféré immédiatement dans une cartouche QIAfilter. La cartouche a été incubée à température ambiante pendant 10 minutes. Ensuite, un piston a été inséré délicatement dans la cartouche pour filtrer le lysat dans un nouveau tube de collecte.

Une fois le lysat filtré, 2 mL de Buffer BB ont été ajoutés. Le mélange a été inversé 4 à 6 fois pour bien mélanger le buffer avec le lysat. Les colonnes de purification, munies de leurs étendeurs, ont été placées sur un système sous-vide de type QIAvac 24 Plus™ (QIAGEN®) réglé à environ -300 mbar. Le lysat a été transféré dans les colonnes, et la sous-pression a été appliquée pour aspirer la solution. Pour le lavage des colonnes, 0,7 mL de Buffer ETR ont été ajoutés dans chaque colonne, et la sous-pression de -300 mbar a été appliquée pour aspirer la solution de lavage. Ce processus a été répété avec 0,7 mL de Buffer PE pour un second lavage. Les colonnes ont ensuite été centrifugées pendant 1 minute à 10 000 g pour éliminer les résidus de lavage.

Après la centrifugation, les tubes de réception ont été remplacés par de nouveaux tubes propres. Pour l'élution de l'ADN, 200 µL de Buffer EB (10 mM Tris·Cl, pH 8) ont été ajoutés au centre de chaque colonne. La solution a été laissée en contact avec les membranes des colonnes pendant 1 minute pour permettre une élution optimale de l'ADN. Enfin, les colonnes ont été centrifugées pendant 1 minute à 10 000 g pour récupérer l'ADN purifié.

Tableau 20 : Composition des Tampons Buffer SCE et Lysis Buffer.

Buffer SCE		
[Finale]	[Initiale]	Volume (mL) 1rx
1 M sorbitol	2 M	2.5
0.1 M NaAc	3 M	0.2
60 mM EDTA	0.5 M	0.6
H_2O		1.7
	Total	5

Lysis Buffer		
[Final]	[Initial]	Volume (mL) 1rx
3% SDS	10%	0.9
60 mM EDTA	0.5 M	0.4
H_2O		1.7
	Total	3

3.2. Méthode d'amplification

3.2.1. PCR

3.2.1.1. PCR classique

L'amplification en chaîne par polymérase ou « *Polymerase Chain Reaction* » (PCR), permet l'amplification *in vitro* de séquences d'ADN cibles par un couple d'amorces. Cette réaction se déroule par la répétition de trois étapes en cycle répété n fois : dénaturation de l'ADN, hybridation des amorces, élongation par la polymérase (Figure 111).

Les réactions PCR sont réalisées sur de l'ADN (génomique ou plasmidique) purifiée ou directement sur des bactéries en suspension dans leurs milieux de culture ou prélevées sur une gélose.

Le mélange réactionnel d'une PCR classique contient entre 10 ng et 100 ng de matrice, 0,2 mM de dNTPs, 0,3 mM du couple de primer, 1X de tampon (Tris-HCl pH 8.7, KCl, $(NH_4)^2SO_4$, 15 mM $MgCl_2$), 1.6 mM de $MgCl_2$ et 0.8 U de HotStarTaq DNA Polymerase (Qiagen®) dans un volume final de 20µL.

La PCR est réalisée dans un thermocycleur MasterCycler Pro (Eppendorf®) suivant le programme suivant :

- Une dénaturation initiale de 10 à 15 min à 95°C selon l'ADN matricielle.
- Un cycle de trois étapes à n répétitions (n allant de 10 à 50 selon l'application) :
    - La dénaturation de 30 sec à 95°C
    - L'hybridation des amorces de 30 sec dont la température optimale est identifiée préalablement par une PCR à gradient de température d'hybridation
    - L'élongation à 72°C (température optimale de l'enzyme) dont la durée varie en fonction de la taille du fragment à amplifier.

Une extension finale, à 72°C pendant 5 min, permettant de compléter les amplifications incomplètes et d'ajouter un A protrusif à chaque extrémité 3' du double brin d'ADN.

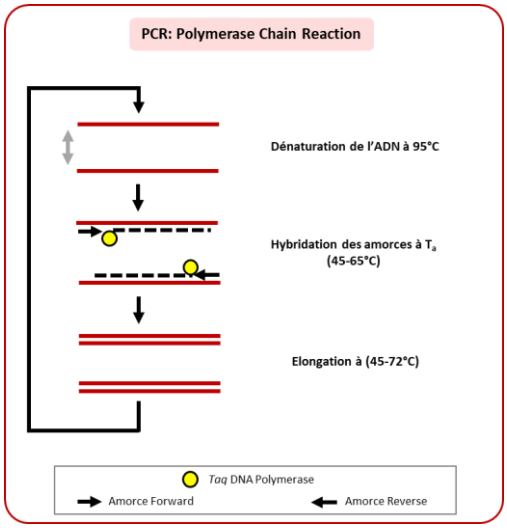

Figure 111 : Représentation schématique du processus d'amplification par PCR. Adapté de (Mullis et al., 1986)

Des PCR sur colonies bactériennes ont également été réalisées, afin de pouvoir vérifier sans nécessité d'extraction la présence des plasmides d'intérêt ou le contrôle des souches utilisées. Le protocole suivi est exactement le même à l'exception de la source d'ADN.

3.2.1.2. PCR en temps réel (rtPCR)

La PCR en temps réel est utilisée pour suivre en temps réel l'amplification de l'ADN dans un échantillon. Cette méthode combine la PCR classique avec la détection en temps réel de l'amplification des acides nucléiques. Elle permet différentes applications, telles que la quantification absolue ou relative d'un échantillon d'ADN, on parle alors de PCR quantitative (qPCR). Dans ma thèse, j'ai utilisé la PCR en temps réel pour différentes applications, telles que l'amplification de séquences microsatellites avant l'analyse de fragment ou le séquençage NGS d'amplicon, la détermination du nombre de copies de plasmides par génome haploïde dans un échantillon, dans la bactérie et la levure, et la quantification de librairies avant séquençage.

3.2.1.2.1. *Le mix réactionnel*

Contrairement à la PCR traditionnelle, qui détecte les produits d'amplification à la fin de la réaction, la rtPCR mesure l'accumulation de produits en temps réel, cycle par cycle. Dans ma thèse, j'ai utilisé pour la plupart de mes applications une réaction de PCR en temps réel à façon, basée sur l'utilisation d'un agent intercalant de l'ADN, le SYTO™ 9, qui est ajouté au mélange réactionnel. Le mélange réactionnel d'une PCR en temps réel contient entre 10 ng et 100 ng de matrice, 0,2 mM de dNTPs, 0,3 mM du couple de primer, 1X de tampon (Tris-HCl pH 8.7, KCl, (NH4)2SO4, 15 mM MgCl2), 1.6 µM de SYTO™ 9 (Invitrogen®), 1.6 mM de MgCl2 et 0.8 U de HotStarTaq DNA Polymerase (Qiagen®) dans un volume final de 20 µL. Les couples de primers utilisés dans ma thèse ne présentent généralement pas de modifications mais peuvent parfois porter un fluorochrome (FAM) lorsque qu'une expérience d'analyse de fragment est prévue en aval.

Le second mix réactionnel utilisé provient du kit de quantification de librairie en PCR quantitative KAPA (Roche). Dans ce mix réactionnel prêt à l'emploi, l'intercalant est le SYBR.

3.2.1.2.2. *Le cyclage PCR*

La PCR est réalisée dans un LightCycler™ 480 (Roche®) suivant le même programme que pour la PCR classique, avec une acquisition de l'intensité de fluorescence lors de l'élongation à 72°C. Des protocoles utilisant deux étapes par cycles ont aussi été utilisés au cours de ma thèse, avec l'étape d'hybridation et d'élongation en une seule étape à des températures allant de 45-65°C. Après les étapes de cyclage répétées, une courbe de fusion est ajoutée (Figure 112).

Pour la quantification des librairies par la KAPA, un programme unique de 5 min à 95°C, suivi de 35 cycles de 30 sec à 95°C et 45 sec à 60°C, finalisé par une courbe de fusion, est utilisé.

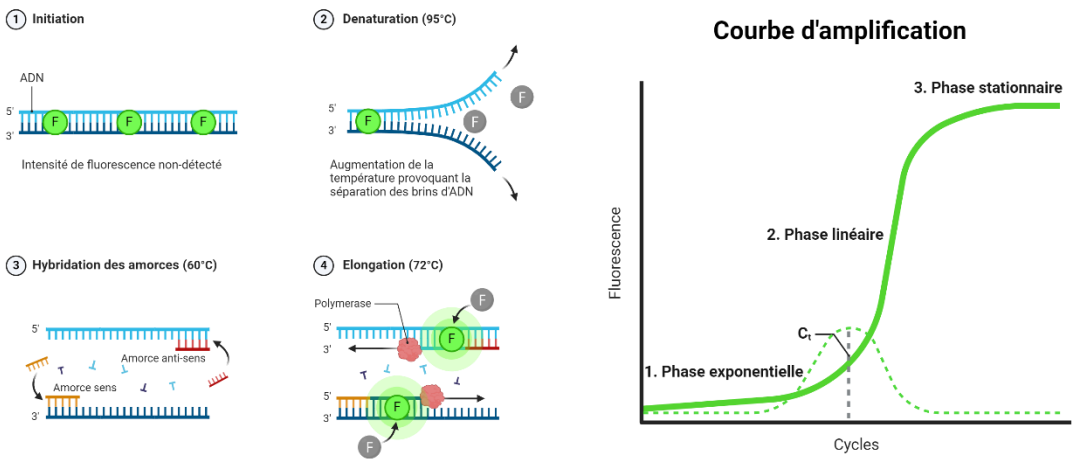

Figure 112 : Représentation schématique du processus de fluorescence par dye de la rtPCR donnant la qPCR.

3.2.1.2.3. *La courbe de fusion*

Une courbe de fusion est un graphique utilisé pour analyser la spécificité des produits amplifiés après la phase d'amplification. Un cycle de chauffe progressif de 65°C à 97°C par palier de 0.03 °C est appliqué en mesurant la fluorescence en continu. La courbe de fusion trace la diminution de la fluorescence en fonction de l'augmentation de la température, révélant ainsi les points de fusion (Tm) des produits. Les doubles brins d'ADN se dissociant à des températures spécifiques, dépendant de leur composition et longueur, un seul et net pic dans la courbe indique un produit d'amplification spécifique. *A contrario*, plusieurs pics ou une forme irrégulière peuvent indiquer la présence de produits non spécifiques ou de dimères d'amorces (Figure 113). Cette analyse est essentielle pour vérifier la pureté et l'identité des produits PCR, assurant la fiabilité des résultats.

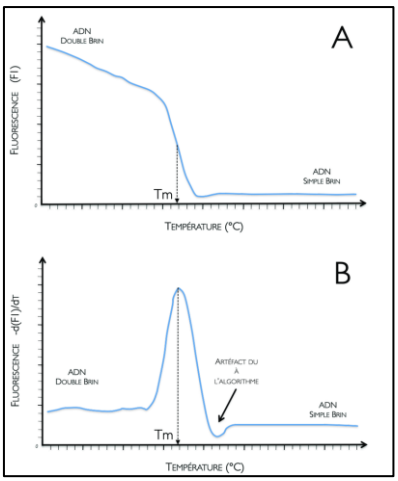

Figure 113 : Représentation graphique d'une courbe de fusion d'une amplification spécifique.

3.2.1.2.4. *Efficacité*

L'efficacité de la duplication de l'ADN est un paramètre crucial pour le calcul des quantités d'ADN amplifiées ou présentes à l'origine. Elle dépend des conditions de la PCR. Dans des conditions "standardisées" (même polymérase et même tampon de réaction), l'efficacité dépend principalement du

couple d'amorces utilisé et de la température à laquelle l'hybridation des amorces avec la matrice est réalisée. Il est souhaitable d'avoir des efficacités de duplication les plus proches de 100% car cela garantit que l'amplification de l'ADN se déroule dans des conditions optimales.

Afin de pouvoir calculer l'efficacité d'une PCR, deux calculs ont pu être effectués. Dans un premier temps, des gammes de dilutions de chaque plasmide avec le couple d'amorce d'HT17-145 ont été réalisées et l'efficacité correspondant à l'efficacité standard a pu être calculée avec la formule suivante :

$$Efficacité = 10^{-\frac{1}{pente\ de\ la\ gamme\ standard}} - 1$$

Dans un second temps, chaque courbe d'amplification a été récupérée et les données de fluorescence à chaque cycle ont été extraites. La formule suivante a ensuite été appliquée :

$$Efficacité\ cycle_i = log_2(\frac{intensité\ de\ fluorescence\ cycle_{i+1}}{intensité\ de\ fluorescence\ cycle_i})$$

(avec Int. fluo. étant la valeur de l'intensité de fluorescence et i le cycle auquel l'efficacité est calculée).

Ainsi, les efficacités de chaque couple de primer utilisé ont été vérifiées afin qu'elles soient au minimum de 80%.

3.2.1.3. PCR quantitative

Il est souvent nécessaire de pouvoir doser un échantillon d'ADN ou une librairie de manière précise. Le principe de cette PCR, « amélioration » de la PCR en temps réel, repose sur l'amplification d'une gamme de standardd d'ADN en parallèle des échantillons à doser. La valeur moyenne de Ct de chaque standard d'ADN permet de tracer une courbe standard en fonction de sa concentration connue. Cette courbe standard permet alors de convertir les Ct moyens des échantillons à doser en concentration (Figure 114).

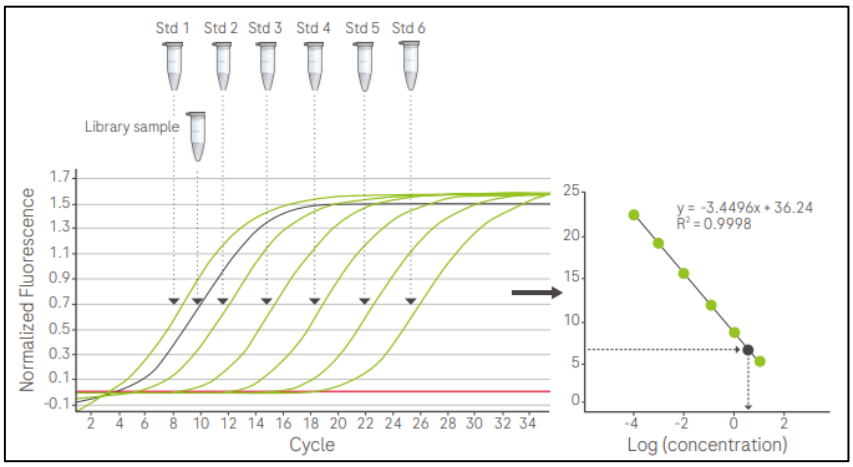

Figure 114 : Principe du kit de quantification de librairies KAPA (tiré de leur brochure).

Ceci a été réalisé expérimentalement en considérant comme échantillons différentes dilutions de 10 en 10 d'une même matrice en triplicat démarrant à $4{,}6.10^7$ copies jusqu'à 46 copies de la matrice. Et permettant ainsi de quantifier chaque échantillon précisément.

Dans le cas du dosage des librairies de séquençage en short-read Illumina, le kit de quantification de librairie KAPA (KAPA Library Quantification Kits™ – Roche®) a été utilisé.

Ce kit est composé de 6 standards d'ADN prédilué et des amorces de PCR spécifiques aux systèmes de séquençage, ici Illumina, ciblant les séquences communes des adaptateurs. C'est pourquoi, ici, après récupération des librairies, une dilution au 100 000ème a été réalisée afin d'être dans la gamme de dosage.

3.2.1.4. PCR digitale par gouttelettes (ddPCR)

La PCR digitale par gouttelettes (*Droplet Digital PCR* ou ddPCR) est une technique de quantification absolue de l'ADN. Le principe de cette technologie repose sur le partitionnement de la réaction de PCR en milliers de gouttelettes à l'échelle du nanolitre (0,85 nL avec la technologie Bio-Rad®). La création des gouttelettes est réalisée par une émulsion eau-huile, où chaque gouttelette agit comme un mini-réacteur individuel. Lors de l'amplification par PCR, chaque gouttelette est analysée et la distinction entre gouttelettes positives et négatives permet de déterminer si la réaction de PCR a eu lieu ou non. En comptant le nombre de gouttelettes positives et en utilisant la loi statistique de Poisson, il est possible de quantifier directement le nombre de copies de molécules d'ADN présentes dans l'échantillon initial avec une grande précision et sensibilité (Figure 115).

Au cours de ma thèse, les ddPCRs ont été réalisés avec le QX200 AutoDG Droplet Digital PCR System™ (Biorad®) comprenant un Automated Droplet Generator (AutoDG) et un QX200 Droplet Reader. Dans un premier temps, un mix réactionnel de PCR a été réalisé avec 1X de QX200 ddPCR EvaGreen Supermix, un mix des amorces associées aux séquences d'intérêt à 150 nM, environ 11 000 copies d'ADN, le tout dans 22 µL. Une fois la plaque centrifugée et exempte de toutes bulles, elle a été placée dans l'AutoDG afin de procéder à la formation des gouttelettes. Ensuite, la plaque a délicatement été placée dans un thermocycleur afin de suivre le programme de PCR suivant : 5 minutes à 95°C, puis 50 cycles de 45 secondes à 95°C et 1 minute à 60°C, enfin, 5 minutes à 4°C, puis 5 minutes à 90°C, le tout avec un ramp rate de 1°C/seconde. Enfin, la plaque a été scannée par le QX200 Droplet Reader afin de comptabiliser le nombre de gouttelettes positives, et ainsi de quantifier les molécules d'ADN.

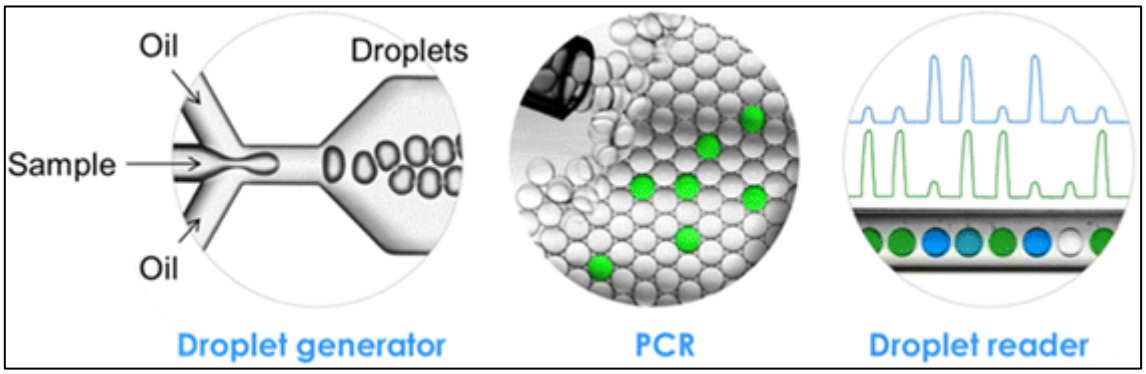

Figure 115 : Représentation schématique du Workflow de la ddPCR.

3.2.2. RPA

La RPA (Recombinase Polymerase Amplification) est une technique d'amplification isotherme de l'ADN qui combine l'utilisation de recombinases, de protéines liant l'ADN simple brin (SSB, *Single-Strand DNA binding protein*) et d'une polymérase Bsu pour amplifier des séquences d'ADN spécifiques à une température constante, généralement autour de 37-42°C.

La polymérase Bsu, aussi appelée polymérase Bsu DNA Pol I large fragment ou simplement Bsu polymerase, est une enzyme dérivée de *Bacillus subtilis*. La particularité de cette polymérase réside dans son activité « SDA » (*strand displacement activity*), son activité 5'→3' exonucléase, sa bonne conjonction avec les recombinases et sa plus faible sensibilité aux impuretés et inhibiteurs.

Un « cycle » d'amplification se déroule à température constante. L'ADN est ciblé par des complexes composés de recombinases enrobant les amorces et facilitant leur hybridation à la séquence cible double brin, et induisant ainsi une séparation des deux brins de l'ADN. Ensuite les SSB stabilisent l'ADN simple brin formé, puis la polymérase Bsu étend les amorces pour synthétiser de nouveaux brins d'ADN (Figure 116).

Dans le contexte de ma thèse, j'ai été amené à utiliser le RPA à des températures différentes des températures optimales allant ainsi de 25 à 45°C. Contrairement à la PCR traditionnelle qui nécessite des cycles de chauffage et de refroidissement, la LT-RPA se déroule à une température constante, facilitant l'amplification rapide et efficace de l'ADN.

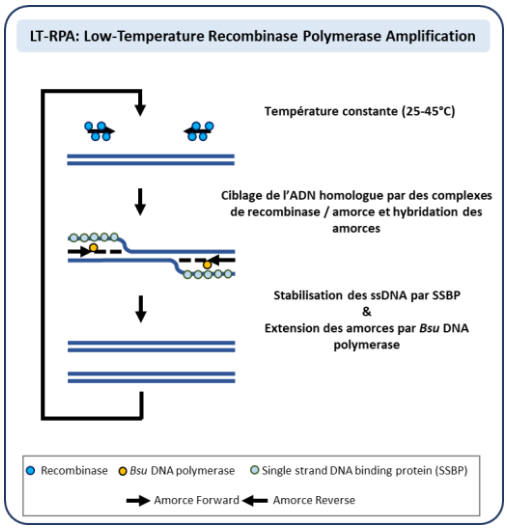

Figure 116 : Représentation schématique du processus d'amplification par LT-RPA. Adapté de (Piepenburg et al., 2006)

Afin de réaliser une LT-RPA, il existe deux kits commerciaux connus sous le nom de TwistAmp™ Basic et TwistAmp™ Liquid Kits, qui ne présentent aucune différence notable dans les résultats. Toutefois, alors que le kit TwistAmp™ Liquid est moins cher, le kit TwistAmp™ Basic est plus facile à manipuler.

3.2.2.1. TwistAmp® Basic Kit

Le kit TwistAmp® Basic (poudre) permet de réaliser cinq réactions de LT-RPA de 10 µL avec un seul tube. Il est crucial de respecter l'ordre d'ajout des réactifs.

Pour commencer, 29,5 µL de tampon de réhydratation sans amorce ont été ajoutés aux culots déshydratés contenant le mélange réactionnel (Recombinase, SSBP, dNTPs et Polymérase Bsu), puis 10,6 µL d'eau sans DNase ont été incorporés. Si davantage de tubes sont requis pour les expériences RPA, ils peuvent être combinés dans un même mélange après resuspension. Les amorces ont été ensuite ajoutées, soigneusement mélangées par pipetage, et 8,5 µL de ce mélange ont été déposés dans des tubes PCR de 0,2 mL. Une réaction de RPA typique de 10 µL inclut 8,02 µL du mélange suspendu et 0,48 µL de paires d'amorces à une concentration finale de 480 nM. Il est impératif de mélanger la réaction de

RPA délicatement par pipetage, sans créer de bulles. Ensuite, 1 µL d'ADN à 5 ng/µL par réaction RPA de 10 µL a été ajouté. Pour une concentration finale de 10,5 mM, 0,5 µL de 210 mM de $Mg(OAc)_2$, cofacteur essentiel à l'activité de la polymérase et des recombinases, ont été ajoutés. La réaction démarrant dès l'ajout du $Mg(OAc)_2$ au mélange, il est possible de le déposer sur la paroi ou le couvercle du tube.

Les tubes ont été vortexés et centrifugés, puis placés dans le thermocycleur préchauffé à une température comprise entre 25 et 45°C. Les conditions d'incubation de la RPA sont 40 minutes à température d'amplification, 2 minutes à 95°C et un refroidissement à 4°C.

Pour l'analyse, le produit LT-RPA, marqué par un fluorochrome (FAM) présent sur une des deux amorces PCR, a été examiné directement par électrophorèse capillaire.

3.2.2.2. TwistAmp® Liquid Basic Kit

Le kit TwistAmp® Liquid Basic permet de préparer au moins une réaction de RPA. Le protocole présenté est celui pour une réaction de RPA dans un volume final de 10 µL. Pour un plus grand nombre de réactions, un mélange réactionnel peut être préparé avec le volume excédentaire adapté.

Tout d'abord, le mélange réactionnel a été réalisé en ajoutant 5 µL de tampon de réaction 2 X, 1 µL de dNTPs à 20 mM pour une concentration finale de 200 µM, et 1 µL de mélange de base 10 X contenant les recombinases, SSBP et Polymérase Bsu. Le mélange a été vortexé et centrifugé brièvement. 0,5 µL de mélange de réaction de base 20 X ont été ensuite ajoutés, mélangés et centrifugés brièvement. Les amorces ont été ajoutées pour une concentration finale de 480 nM et le mélange a été complété avec de l'eau jusqu'à atteindre un volume final de 8,5 µL dans des tubes PCR de 0,2 mL. Le mélange a été soigneusement mélangé en pipetant sans créer de bulles. Par la suite, 1 µL d'ADN à 5 ng/µL a été ajouté. Pour une concentration finale de 10,5 mM, 0,5 µL de 210 mM $Mg(OAc)_2$, ont été ajoutés comme précédemment. Les autres étapes sont effectuées à l'identique du TwistAmp® Basic Kit.

3.2.2.3. Digestion contre polyadénilation

Afin de pouvoir s'affranchir des problèmes possiblement liés à une polyadénylation incomplète chez les microsatellites mononucléotidiques A_n, une digestion enzymatique a été réalisée. Dans un volume final de 20 µL, 25 unités de l'enzyme BfaI (NEB®) ont été mélangées à 1X de tampon Cutsmart et à la totalité des 10 µL d'ADN de la réaction de RPA. Le tout a ensuite été incubé à 37°C pendant 1 heure, puis l'enzyme a était inactivée par une incubation à 80°C pendant 20 minutes. L'ensemble a pu être ensuite directement analysé par électrophorèse capillaire sans purification.

3.2.3. WGA

Dans le cadre de ma thèse, j'ai appliqué cette méthode à un pool équimolaire des plasmides A_{15}, A_{20}, A_{24}, $(AC)_{15}$, $(AC)_{20}$ et $(AC)_{24}$ en utilisant le kit REPLI-G (Qiagen®). Dans un premier temps, 0,05 ng ($22.08x10^6$ copies) dans 2,5 µL d'ADN, soit 0,008ng par plasmide ($3,68x10^6$ copies) a été mélangé à 1X de Tampon D1 afin de dénaturer, par traitement alcalin l'ADN en incubant pendant 3 min à température ambiante. Ensuite 1X de Tampon N1 a été ajouté afin de neutraliser la solution et arrêter la dénaturation. Par la suite, 1X de REPLI-g Midi Reaction Buffer a été ajouté, ainsi que 1 µL de REPLI-g Midi DNA Polymerase et de l'eau afin d'atteindre 50 µL. Enfin, le tout a été incubé pendant 16 heures à 15°C, 20°C, 25°C, 30°C 35°C et 37.5°C (la température optimale recommandée étant à 32°C) puis inactivé 3 minutes

à 65°C et stocké à 4°C. Avant de futures analyses, le mix a été dilué avec 50 µL d'eau et une précipitation à l'éthanol a été réalisée, avec resuspension dans 500 µL à 37°C, afin de pouvoir doser au QuBit les pools amplifiés. Cette quantification nous a alors permis de calculer le nombre de doublement de molécules.

Soutenue par le système de tampon REPLI-g unique et optimisé, la polymérase Phi 29 surmonte facilement les structures secondaires telles que les boucles en épingle à cheveux, évitant ainsi les glissements, les arrêts et la dissociation de la polymérase pendant l'amplification. Cela permet de générer des fragments d'ADN allant jusqu'à 100 kb sans biais de séquence.

3.3. Méthode de dosage

A différentes étapes des manipulations, il est important de vérifier la quantité des ADN récupérés avant de futures analyses. Pour cela différentes méthodes ont été utilisées.

3.3.1. NanoDrop

Le NanoDrop (ThermoFisher®) est un spectrophotomètre pour des micro-volumes utilisé pour quantifier les acides nucléiques (ADN, ARN) et les protéines en mesurant leur absorption de la lumière. Son fonctionnement repose sur le principe que les molécules d'acides nucléiques et de protéines absorbent la lumière à des longueurs d'onde spécifiques.

Pour effectuer un dosage, une petite goutte de l'échantillon (1-2 µL) est placée directement sur le détecteur optique du NanoDrop 2000. L'instrument passe ensuite un faisceau de lumière à travers l'échantillon à 260 nm pour l'ADN. Le NanoDrop mesure l'absorbance de la lumière par l'échantillon à ces longueurs d'ondes spécifiques. En utilisant les valeurs d'absorbance et les coefficients d'extinction spécifiques aux molécules cibles, le logiciel intégré calcule la concentration de l'échantillon.

Un faisceau de lumière à 280 nm et à 230 nm traverse également l'échantillon. Le ratio d'absorbance 260/280 est utilisé pour évaluer la pureté. Un rapport d'environ 1,8 est généralement considéré comme "pur" pour l'ADN. Si le rapport est sensiblement plus bas, cela peut indiquer la présence de protéines, de phénol ou d'autres contaminants qui absorbent fortement près de 280 nm. Le ratio 260/230 est une deuxième mesure de la pureté de l'échantillon, car des contaminants absorbent à 230 nm (comme l'EDTA). Le rapport 260/230 doit être supérieur au rapport 260/280, car il se situe généralement entre 2 et 2,2. Un rapport plus bas pourrait indiquer une contamination.

3.3.2. Qubit

Le Qubit 3 Fluorometer™ (Invitrogen®) est un fluorimètre utilisé pour quantifier les acides nucléiques (ADN, ARN) et les protéines avec une grande précision et sensibilité analytique, qui dépendent du kit utilisé. Le fonctionnement repose sur l'utilisation de réactifs spécifiques qui se lient sélectivement aux molécules cibles et émettent une fluorescence proportionnelle à leur concentration.

Dans le cadre de ma thèse, le kit de dosage Qubit HS dsDNA assay Kit (Life Technologies®) a été utilisé, qui ne dose que l'ADN double brin à des concentrations d'ADN allant de 10 pg/µL à 100 ng/µl, grâce à un fluorophore propriétaire qui ne se fixe que sur les petits sillons de l'ADN double brin. Pour effectuer un dosage, les échantillons d'ADN ont d'abord été dilués puis mélangés avec le réactif Qubit correspondant dans des tubes. Le mélange a été incubé à température ambiante pendant 2 à 5 minutes

pour permettre la liaison et l'excitation du fluorophore. Il a été fait de même pour les deux standards. Ensuite, les standards puis les échantillons ont été placés dans le Qubit, où ils sont exposés à une source de lumière d'excitation bleue (430–495 nm), et rouge (600–645 nm). Le fluorophore lié aux molécules cibles émet une fluorescence, qui est détectée par le photomultiplicateur intégré dans l'appareil. Le Qubit convertit cette fluorescence en une mesure quantitative de la concentration de l'échantillon.

3.4. Méthodes de purification de l'ADN

Lors de la préparation de librairies, de nombreuses étapes de purification sont nécessaires afin de garantir la qualité du séquençage. Ici, ces étapes ont été réalisées par l'utilisation de billes magnétiques, par colonnes par l'utilisation du kit QIAquick PCR purification kit™ (Qiagen®) ou par précipitation.

3.4.1. Billes magnétiques

L'utilisation de billes magnétiques est une méthode permettant d'obtenir une haute pureté et un rendement élevé de nos échantillons. Les billes utilisées ont été les SPRIselect™ (Beckman Coulter®) pour les protocoles Illumina, et AMPure PB beads pour les protocoles PacBio, qui permettent aussi de purifier l'ADN selon une certaine taille (« *sizing* »).

Les billes magnétiques sont enrobées d'un matériau qui peut lier sélectivement les molécules d'ADN en présence d'une solution de liaison, généralement contenant du polyéthylène glycol (PEG) et du sel (comme le NaCl). Différents ratios permettent de sélectionner des tailles d'ADN spécifiques. Un ratio élevé (ex. 1.8X) favorisera la liaison des fragments d'ADN de plus petite taille, car il y aura plus de billes disponibles pour se lier aux fragments d'ADN, là où un ratio plus bas (par exemple 0.6X) permettra de lier préférentiellement les fragments d'ADN de plus grande taille, car les fragments plus petits resteront en solution (Figure 117).

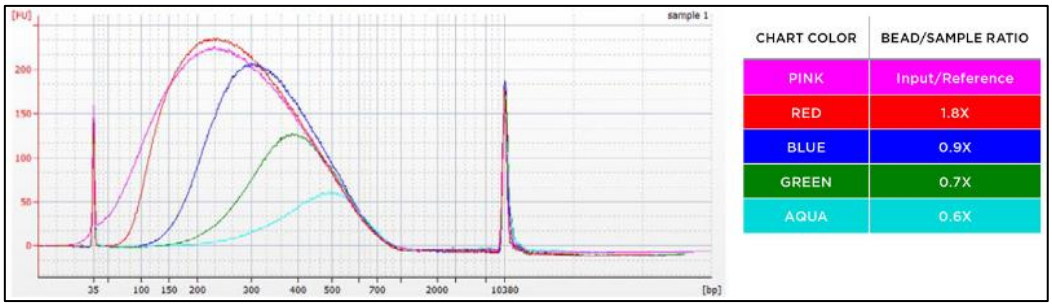

Figure 117 : Représentation des ratios de sélection de tailles par les billes AMPure.

Il est même possible de réaliser un « *double-sizing* » pour purifier et sélectionner des fragments d'ADN de taille très précise. Ce principe consiste à utiliser deux étapes de sélection. Dans la première étape, un ratio de billes relativement faible est utilisé pour lier et éliminer les fragments d'ADN de grande taille, laissant les fragments plus petits en solution. Ensuite, dans la seconde étape, un ratio de billes plus élevé est appliqué à la solution restante pour lier et récupérer les fragments d'ADN de petite taille souhaités (Figure 118).

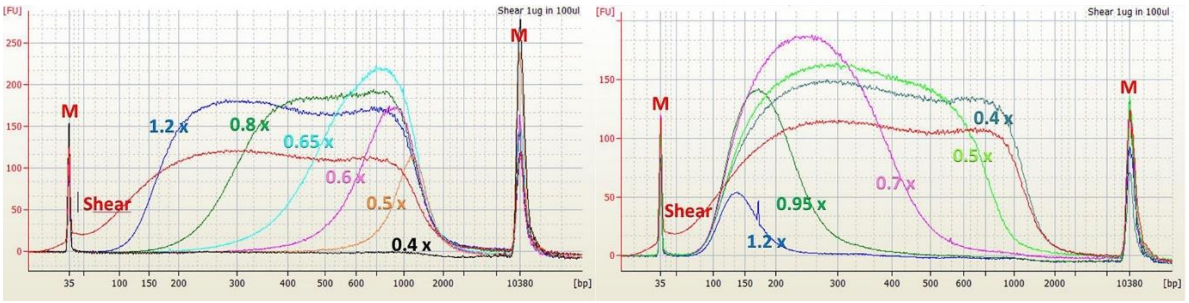

Figure 118 : Représentation des ratios de sélection de tailles par les billes SPRIselect (Left Side Selection & Right Side Selection).

Pour commencer, un volume de billes magnétiques correspondant de 0,85 à 1,8 fois le volume initial d'échantillons a été ajouté à la solution en fonction de la taille que l'on souhaite capturer par les billes. Cette étape est cruciale pour assurer un ratio adéquat de billes pour une liaison ciblée et efficace des séquences. Le mélange a été réalisé par aspiration et refoulement dix fois afin d'homogénéiser la suspension, garantissant ainsi une interaction optimale entre l'ADN et les billes magnétiques tout en limitant leurs dégradations. La suspension a été incubée pendant 5 minutes à température ambiante permettant une liaison spécifique et efficace de l'ADN aux billes magnétiques.

Après l'incubation, le tube a été placé sur un support aimanté. Cette étape permet de séparer les billes magnétiques du surnageant. Une fois les billes attirées par l'aimant, le surnageant a été soigneusement éliminé sans perturber les billes retenues par le champ magnétique. Pour éliminer les impuretés, un premier lavage a été effectué en ajoutant 200 µL d'éthanol à 70 % fraîchement préparé aux billes magnétiques toujours placées sur le support. Après quelques aspirations-refoulement, le surnageant contenant l'éthanol de lavage a été jeté. Ce lavage a été répété une fois supplémentaire, totalisant deux lavages à l'éthanol, afin de garantir une purification optimale. Les billes magnétiques ont été laissées sécher à l'air libre ou dans un environnement contrôlé pour éliminer toute trace résiduelle d'éthanol afin d'éviter toute contamination par l'éthanol dans les étapes suivantes.

Après séchage, entre 10 et 30 µL d'eau ont été ajoutés aux billes magnétiques pour l'élution des acides nucléiques. Le tube est ensuite retiré du support aimanté, et la suspension homogénéisée par aspiration – refoulement. Elle a été incubée pendant 2 minutes à température ambiante pour permettre la dissociation de l'ADN des billes. Enfin, le tube a été replacé sur le support aimanté et le surnageant, contenant les séquences purifiées, a été soigneusement récupéré dans un nouveau tube.

3.4.2. QIAquick PCR purification kit (Qiagen)

Le kit de purification QIAquick PCR purification kit (Qiagen®) a été utilisé pour la purification d'amplicon et de plasmide linéarisé.

Pour commencer, pour 1 volume d'ADN, 5 volumes de Buffer PB ont été mélangés et le pH ajusté avec de l'acétate de sodium 3 M si besoin. La solution a ensuite été placée dans une colonne QIAquick puis centrifugée pendant 30 sec à 13 000 rpm afin de fixer l'ADN. L'éluât a été jeté et pour le lavage des colonnes, 0,75 mL de Buffer PE a été ajouté dans chaque colonne. Les colonnes ont alors été centrifugées pendant 1 min à 13 000 rpm et l'éluât de lavage a été jeté. Les colonnes ont ensuite été re-centrifugées pendant 1 minute à 13 000 rpm pour éliminer les résidus de lavage.

Après la centrifugation, les tubes de réception ont été remplacés par de nouveaux tubes propres. Pour l'élution de l'ADN, 50 µL de Buffer EB (10 mM Tris·Cl, pH 8) ont été ajoutés au centre de chaque colonne. La solution a été laissée en contact avec les membranes des colonnes pendant 1 minute pour permettre une élution optimale de l'ADN. Enfin, les colonnes ont été centrifugées pendant 1 minute à 17 900 g pour récupérer l'ADN purifié. L'éluât est conservé à -20°C.

3.4.3. Précipitation à l'éthanol avec de l'acétate de sodium

Pour commencer, pour 1 volume d'ADN, 1/10 de volume d'acétate de sodium 3M, pH 5,2 a été ajouté. Ayant pour rôle de neutraliser les charges négatives des groupes phosphate présents dans l'ADN, réduisant ainsi sa solubilité dans l'eau. Par la suite, 2 à 2,5 volumes d'éthanol froid (100 %) ont été ajoutés au mélange réduisant encore plus la solubilité de l'ADN dans l'eau. Le mélange a été doucement inversé plusieurs fois pour homogénéiser la solution, sans vortexer afin de ne pas fragmenter l'ADN. Cela permet d'assurer une distribution uniforme des réactifs et un contact optimal entre l'ADN et les ions sodium. Une fois le mélange bien homogénéisé, il a été incubé à -20°C pendant 30 minutes. Cette incubation à basse température est cruciale pour maximiser la précipitation de l'ADN, car elle réduit la solubilité de l'ADN dans la solution alcoolique.

Après l'incubation, l'échantillon a été centrifugé à 15 000 g pendant 2 minutes à 4°C afin de culoter l'ADN. Le surnageant a ensuite été retiré très délicatement afin de ne pas toucher au culot. 500 µL d'éthanol froid à 70% ont été ajoutés pour laver le culot, puis recentrifugés à 15 000 g pendant 2 minutes à 4°C. Le surnageant a été retiré soigneusement et le culot a été séché à l'air libre pendant 10 min. Enfin, le culot d'ADN a été resuspendu avec 500 µL d'eau, dosé et conservé à -20°C.

3.5. Méthode d'analyse de l'ADN basée sur l'électrophorèse

Entre deux étapes d'expérimentations, des vérifications de la qualité des ADNs sont parfois effectuées. Chaque amplification, purification, digestion et action réalisée sur des séquences d'ADN peut apporter des artefacts ou des résidus qui nuiraient au bon déroulement des analyses ultérieures. Pour cela différentes méthodes peuvent être utilisées, permettant parfois la sélection de certaines gammes de molécules d'ADN selon leur taille.

3.5.1. Electrophorèse Horizontale

3.5.1.1. Gel d'agarose

Les acides nucléiques sont analysés par électrophorèse horizontale en gel d'agarose à différentes concentrations [de 0,8 à 2 %, (p/v)] contenant du colorant pour gel SYBR™ Safe à 60 ng/mL et dans un tampon TAE 1X (tris 40 mM, acide acétique 70 mM, EDTA 0,5 M, pH 8) ou 0,5X. Les échantillons sont déposés avec du tampon de charge 1X (6x DNA Loading Dye Buffer Double Blue, Solis Biodyne). La tension appliquée varie de 25 volts à 100 volts et les gels sont visualisés sous UV à 272 nm par le GelDoc Go Gel Imaging System (Biorad®). L'acquisition des images se fait par le logiciel Quantity One (Biorad®).

3.5.1.2. BluePippin

Le BluePippin™ (Sage Science®) est utilisé pour la séparation et la purification des fragments d'ADN en fonction de leur taille, basé sur une électrophorèse en gel. Le système utilise des cassettes de gel pré-coulées avec des puits de chargement et de récupération d'ADN. Le BluePippin permet une récupération précise des fragments d'une taille spécifique en utilisant un logiciel intégré pour définir des plages de taille. Des capteurs optiques détectent la migration des fragments marqués par un colorant fluorescent, et le système recueille automatiquement les fragments désirés dans des fractions distinctes. Cette méthode améliore la précision et l'efficacité de la sélection de taille.

Dans le cadre de ma thèse, le gel à 2 % agarose, 100-600 bp (2 % agarose, dye free w/ internal standards bluepippin 100bp-600 pb, Ozyme®) a été utilisé. Dans un premier temps, les échantillons d'ADN ont été préparés en les diluant (5 µg dans 30 µL) et les mélangeant avec 10 µL d'un standard contenant un colorant fluorescent et un marqueur de taille, pour permettre une bonne détection optique. Les cassettes de gel pré-coulées, spécifiques au BluePippin, ont ensuite été préparées en injectant le tampon de fonctionnement approprié dans les réservoirs des cassettes si besoin. Les échantillons d'ADN ont été chargés dans les puits de la cassette, puis la cassette a ensuite été insérée dans l'instrument BluePippin.

Une fois le programme de séparation lancé depuis le logiciel intégré, une tension électrique définie par le logiciel en fonction des paramètres de la cible est appliquée, provoquant la migration des fragments d'ADN à travers le gel. Le logiciel permet de définir des plages de taille spécifiques pour la collecte des fragments d'intérêt. Les capteurs optiques détectent la progression des fragments marqués, et une fois les fragments d'ADN atteignant la taille désirée identifiés, le système récupère automatiquement ces fragments dans des fractions distinctes, en interrompant le flux électrique aux moments appropriés pour éviter la contamination croisée. L'ADN élué est directement récupéré après le run dans un tampon (TE), et peut être quantifié, puis concentré à l'aide de billes en fonction des besoins. Ce protocole garantit une purification précise et efficace des fragments d'ADN de la taille souhaitée.

3.5.1.3. TapeStation

La 4150 TapeStation™ (Agilent®) est utilisée pour analyser la qualité et la quantité des acides nucléiques (ADN, ARN) et des protéines. Elle fonctionne en utilisant des bandes microfluidiques, ou ScreenTapes, qui contiennent des canaux préremplis avec un gel électrophorétique. Les échantillons sont chargés sur la ScreenTape, et l'instrument applique une tension électrique pour séparer les molécules selon leur taille. Un détecteur optique mesure la fluorescence des fragments marqués par un colorant. Le logiciel intégré analyse les données et génère des électrophorégrammes et des images de gel virtuelles, permettant de déterminer la concentration, la taille et l'intégrité des échantillons. La TapeStation est plus simple et plus rapide que le Bioanalyzer 2100™ et permet l'analyse simultanée de multiples échantillons avec un minimum d'intervention manuelle.

Dans le cadre de ma thèse, appliquant cette méthode uniquement sur de l'ADN, le kit High Sensitivity D1000™ (Agilent®) a été utilisé. Dans un premier temps, la ScreenTape D1000™ a été mise en place dans l'instrument. Les échantillons d'ADN (5 à 500 pg/µL), ont été préparés en mélangeant 2 µL avec 2 µL du réactif de charge fourni, composé de bromure d'éthidium et de tampon de chargement. Ce mélange a ensuite été chargé dans des puits individuels d'une barrette de micro-tubes. Simultanément, un ladder de taille D1000 a été préparé et chargé dans un puits spécifique de la barrette.

La barrette a alors été vortexée horizontalement pendant 1 minute à 2000 rpm, centrifugée pendant 1 minute, puis placée dans la TapeStation.

Une fois le processus lancé, la Tapestation dépose les échantillons dans la ScreenTape et le système automatise l'application d'une tension électrique pour séparer les fragments d'ADN à travers les canaux préremplis de gel de la ScreenTape. La détection se fait par fluorescence, chaque fragment étant marqué par le bromure d'éthidium qui est excité par un laser et détecté par le système optique de la TapeStation. Le logiciel intégré analyse les signaux fluorescents, génère les images réelles des gels de migrations et en déduit des électrophorégrammes (Figure 119). Le logiciel compare les résultats des échantillons avec le ladder pour déterminer la taille, la distribution et la concentration des fragments d'ADN.

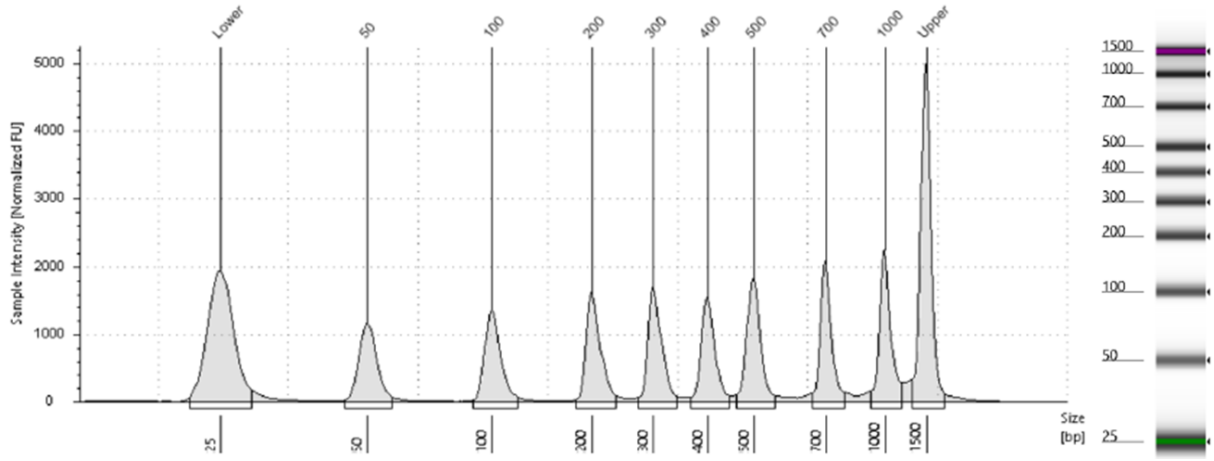

Figure 119 : Représentation des résultats sur TapeStation par gel et électrophorégramme virtuel.

3.5.2. Electrophorèse Capillaire

3.5.2.1. Bioanalyzer

Le Bioanalyzer Agilent 2100™ (Agilent®) est utilisé pour analyser la qualité et la quantité d'acides nucléiques (ADN, ARN) et de protéines. Il fonctionne en combinant l'électrophorèse capillaire microfluidique et la détection par fluorescence. Les échantillons sont chargés dans une puce microfluidique contenant des canaux et des puits miniaturisés. Une tension électrique est appliquée pour séparer les molécules en fonction de leur taille. La détection se fait par un laser qui excite les molécules marquées par un colorant fluorescent, permettant de générer un profil électrophorétique. Les résultats sont ensuite analysés par un logiciel dédié, fournissant des informations sur la concentration, la taille et l'intégrité des échantillons. Ce système est utilisé pour sa précision, sa rapidité et sa faible consommation d'échantillon.

Dans le cadre de ma thèse, le kit Agilent High Sensitivity DNA™ (Agilent®) a été utilisé pour sa précision et grande sensibilité, permettant d'analyser des ADN allant de 50 pb à 7000 pb à partir de concentrations d'ADN allant de 5 pg/µL à 5 ng/µl. Dans un premier temps, le mélange gel-colorant (gel-dye mix) a été préparé en combinant 15 µL du colorant fourni avec la matrice de gel, suivi d'un vortexage et d'une centrifugation rapide à travers un filtre pour garantir la pureté du gel. 9 µL du gel-colorant a ensuite été chargé dans un puits de la puce, puis introduit dans les puits du canal microfluidique de la puce à l'aide d'une seringue à pression, garantissant ainsi que les canaux ont été correctement remplis pour une électrophorèse efficace. La puce microfluidique a été chargée avec 1 µL des échantillons d'ADN (5 à 500 pg/µL), 1 µL du ladder de taille et 5 µL de marqueur interne, ce dernier permettant la

normalisation des résultats. Une fois la puce préparée, elle a été vortexée horizontalement pendant 1 min à 2400 rpm puis placée dans le Bioanalyzer 2100™ et le run a été lancé.

Comme pour une électrophorèse classique, une tension électrique a été appliquée aux canaux microfluidiques, provoquant la migration des fragments d'ADN à travers le gel. Un laser excite le colorant fluorescent lié aux fragments d'ADN, et la fluorescence émise a été détectée par un photodétecteur. Les signaux fluorescents sont convertis en données numériques, générant un électrophorégramme qui affiche des pics correspondant aux fragments d'ADN de différentes tailles. Le logiciel du Bioanalyzer analyse ces données en comparant les pics des échantillons à ceux du ladder de taille pour déterminer les tailles des fragments d'ADN et calcule également la concentration des fragments en fonction de l'intensité de la fluorescence. Les résultats sont présentés sous forme de gel virtuel et d'électrophérogramme (Figure 120), fournissant des informations précises sur la taille, la distribution et la concentration des fragments d'ADN dans chaque échantillon, permettant ainsi une analyse rapide, précise et sensible des échantillons d'ADN, même à des concentrations très faibles.

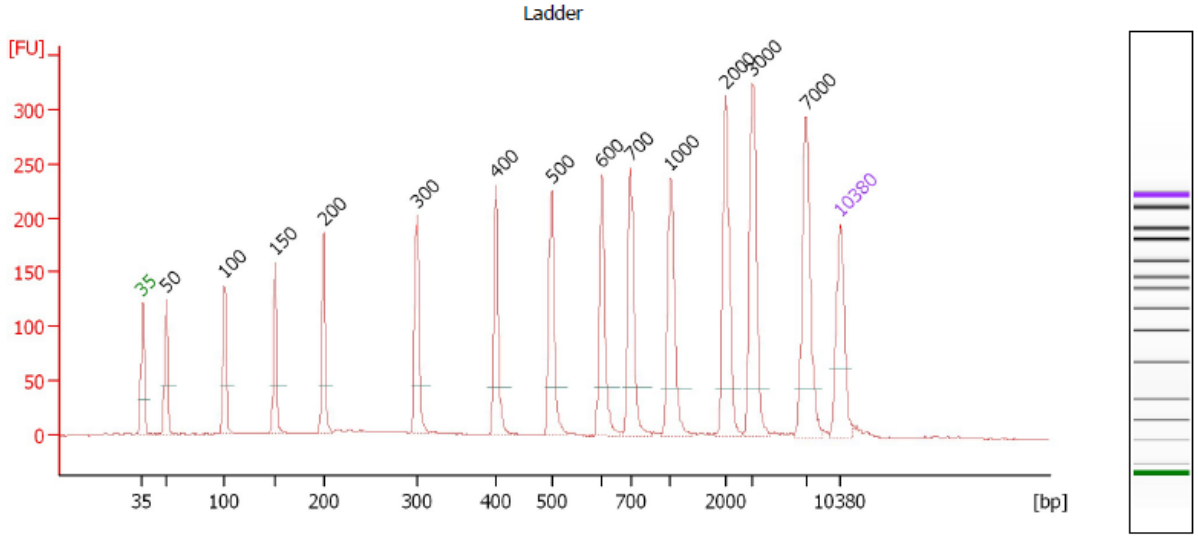

Figure 120 : Représentation des résultats sur Bioanalyzer par électrophorégramme et gel.

3.5.2.2. Fragment Analyzer 5200

Le 5200 Fragment Analyzer System™ (Agilent Technologies®) est un instrument de laboratoire utilisé pour analyser la taille, la concentration et l'intégrité des fragments d'ADN et d'ARN. Il fonctionne par électrophorèse capillaire automatisée. Les fragments sont marqués avec un colorant fluorescent, et lorsqu'ils passent devant un détecteur optique, la fluorescence est mesurée. Le logiciel du Fragment Analyzer convertit ces mesures en électrophorégrammes, affichant des pics correspondant à différents fragments. En comparant ces résultats à un ladder de taille connue, le système détermine la taille et la concentration des fragments dans chaque échantillon. Ce procédé permet une analyse rapide, précise et automatisée des acides nucléiques.

Dans le cadre de ma thèse, les kits HS Small Fragment Kit (1-1500 bp), HS NGS Fragment Kit (1-6000 bp), HS Large Fragment 50kb Kit et HS Genomic DNA 50kb Kit (Agilent Technologies®) ont été utilisés pour la vérification de la taille d'ADN génomique, de produit d'amplification et de digestion, avant et après purification et sélection de taille, et de librairies. Dans un premier temps, la quantité de gel nécessaire a été préparée en mélangeant le gel et un intercalant (Figure 121). La solution de conditionnement a également été préparée (Figure 121) et les deux tubes ont été chargés dans

l'appareil. Ensuite, 1 mL de Inlet Buffer, préalablement dilué à 1X, a été déposé dans les puits de la première ligne d'une plaque 96 puits, 1 mL de Storage Solution a été déposé dans les puits de la dernière ligne de cette même plaque, puis cette plaque a été placée dans le tiroir B de l'appareil. 200 µL de tampon TE Rinse à 0.25X ont été déposés dans la première ligne d'une nouvelle plaque 96 puits, puis placée dans le tiroir M de l'appareil. Enfin, 2 µL d'échantillon d'ADN (ou de ladder) ont été mélangés avec 22 µL de marqueur diluant puis déposés dans les puits d'une dernière plaque 96 puits, et placée dans le tiroir 1 de l'appareil.

5200 Fragment Analyzer system volume specifications

# of Samples to be Analyzed[1]	Volume of Intercalating Dye	Volume of Separation Gel[2]	Volume of 1x Conditioning Solution[2]
12	1.0 µL	10 mL	10 mL
24	1.5 µL	15 mL	15 mL
36	2.0 µL	20 mL	20 mL
48	2.5 µL	25 mL	25 mL
96	4.5 µL	45 mL	45 mL

[1] One sample well per separation is dedicated to the ladder.
[2] A 5 mL minimum volume in the tube is included.

Figure 121 : Protocole de préparation des Gels et solution de conditionnement.

Une fois les plaques et les capillaires prêts, le logiciel du 5200 Fragment Analyzer a été utilisé pour configurer le protocole d'analyse, y compris les paramètres de tension et le temps d'électrophorèse. L'analyse commence par l'application d'une tension électrique, provoquant la migration des fragments d'ADN à travers les capillaires. Les fragments marqués traversent un détecteur optique qui mesure la fluorescence, et ces données sont converties en électrophorégrammes par le logiciel. Les électrophorégrammes affichent des pics correspondant aux fragments présents dans chaque échantillon. En comparant ces pics aux marqueurs de taille, le logiciel détermine la taille et la concentration des fragments d'acides nucléiques. Ce protocole assure une analyse rapide, précise et automatisée, facilitant l'obtention de données détaillées sur les échantillons d'ADN.

3.5.3. SeqStudio

Le SeqStudio Genetic Analyzer ™(Applied Biosystems®) est un système de séquençage d'ADN de type Sanger, conçu pour des analyses génétiques rapides et précises, couplé à un système d'analyse de fragments par électrophorèse capillaire (Figure 122). Au cours de ma thèse, le SeqStudio a exclusivement été utilisé pour son système d'analyse de fragments, le séquençage de 1ère génération ayant été réalisé auprès du fournisseur des plasmides.

Le processus débute avec la préparation des échantillons d'ADN, qui sont amplifiés par PCR avec des amorces possédant un marqueur fluorescent. Le SeqStudio applique une tension électrique pour faire migrer les fragments d'ADN à travers quatre capillaires remplis d'un gel polymère, les fragments se séparant par taille. Un laser excite les marqueurs fluorescents, et un détecteur optique capte la fluorescence émise. Les signaux fluorescents sont convertis en données numériques par le logiciel intégré, qui analyse ensuite l'intensité et la position des pics fluorescents pour déterminer la taille et la quantité des fragments, permettant ainsi des applications variées, telles que la vérification de l'amplification d'ADN, le génotypage ou la détection de mutations. Le SeqStudio Genetic Analyzer™ est conçu pour être facile à utiliser, offrant des résultats rapides et fiables.

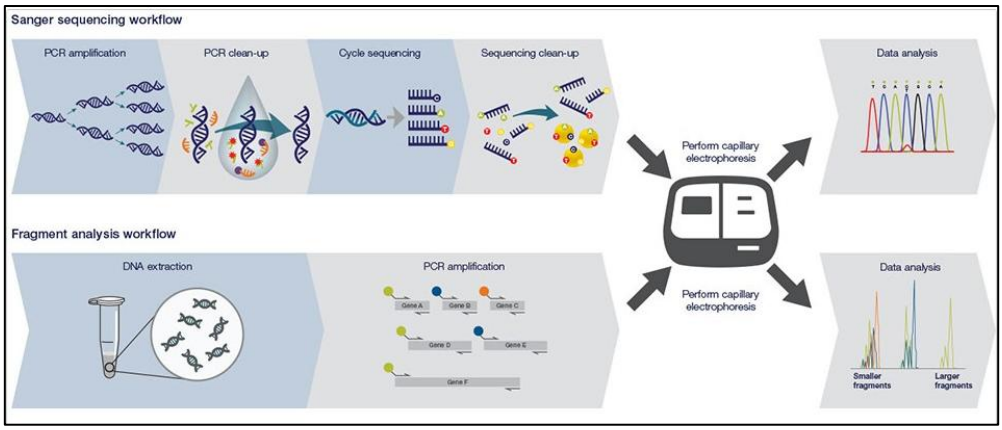

Figure 122 : Principe de double Workflow du SeqStudio Genetic Analyzer réalisant un séquençage SANGER et une analyse de fragment, le tout par électrophorèse capillaire.

Dans un premier temps, les échantillons d'ADN ont été préalablement amplifiés via PCR avec des amorces marquées par un fluorochrome. Les produits PCR ont été ensuite purifiés puis mélangés (0,5 µL) avec un marqueur de taille (0,125 µL) et de la formamide (qsp 10 µL), permettant la dénaturation des brins d'ADN. Les échantillons ont été transférés dans une plaque 96 puits puis chauffés à 95 °C pendant 10 minutes et placés en glace pendant 10 minutes pour réaliser la dénaturation et stabiliser l'ADN. La plaque est ensuite insérée dans le SeqStudio Genetic Analyzer.

L'instrument commence par l'injection des échantillons dans des capillaires remplis d'un gel polymère, où une tension électrique est appliquée pour provoquer la migration des fragments d'ADN. Un laser excite les marqueurs fluorescents attachés aux di-désoxynucléotides, et un détecteur optique capte la fluorescence émise par chaque fragment. Les données fluorescentes sont ensuite converties en signaux numériques par le logiciel intégré, qui génère des chromatogrammes représentant les séquences d'ADN et des électrophorégrammes représentant les tailles des séquences d'ADN.

4. Méthodes de séquençage

4.1. Séquençage de seconde génération

4.1.1. Préparation des librairies

4.1.1.1. Digestion / Fragmentation

4.1.1.1.1. *Site – spécifique (SSDP)*

Dans le cadre de ma thèse j'ai été amenée à préparer des librairies à partir des séquences d'intérêt contenant des microsatellites présents dans des plasmides (bactériens et/ou de levure). Ces séquences ont été isolées par digestion spécifique à l'aide d'enzymes de restriction de type II puis ont été par la suite utilisées pour la construction de librairies.

Le protocole de digestion enzymatique des plasmides bactériens a été soigneusement élaboré afin de garantir une récupération optimale des séquences ciblées. Le mélange réactionnel comprenait les éléments suivants : 1X de tampon de digestion, ici le rCutSmart™ Buffer 10X (New England Biolabs®), 10 unités (U) de l'enzyme HpyCH4V, 10 U de l'enzyme HpyCH4III, 50 U de l'enzyme MspI, et 5 µg de plasmides d'intérêt, le tout dans un volume final de 100 µL.

Ce mélange a été incubé pendant 60 minutes à 37°C pour permettre une digestion complète du plasmide. Après cette étape, 25 U de Klenow Fragment (3´→ 5´ exo-) (New England Biolabs®) et 0,05 mM dATP ont été ajoutés au mélange. Le mélange a ensuite été réincubé pendant 30 minutes supplémentaires à 37°C. Cette étape supplémentaire permet l'ajout d'un A aux extrémités de l'ADN (A-Tailing), ce qui permet les étapes de ligation ultérieures.

Le protocole de digestion enzymatique pour le plasmide de levures est le même que pour le plasmide bactérien si ce n'est l'ajout de 10 U de l'enzyme Fnu4HI au mélange réactionnel.

Les séquences d'ADN sont purifiées à l'aide de billes magnétiques à 0.85X (Matériels et Méthodes 3.4.1.), quantifiées et contrôlées par électrophorèse capillaire (BA, TP ou FA), avant l'étape de ligature des adaptateurs.

L'application de ce protocole assure une digestion précise et une préparation efficace du plasmide, ce qui est essentiel pour la construction réussie des librairies et pour le séquençage en aval.

4.1.1.1.2. Fragmentation aléatoire (RFP)

Dans le cadre du projet 2 de cette thèse, les plasmides contenant les séquences microsatellites ont été fragmentés aléatoirement pour ensuite être utilisés pour la construction de librairies.

4.1.1.1.2.1. Enzymatique

Ici, la fragmentation enzymatique fait partie intégrante du kit QIAseq FX DNA Library Kit ™(Qiagen®) utilisé pour la réalisation de la librairie présentée ci-après (voir paragraphe 4.1.1.2.2.).

4.1.1.1.2.2. Mécanique (xGen PRISM)

Ici, la fragmentation a été réalisée par l'utilisation d'un système de sonication, le Bioruptor Plus (Diagénode). La sonication est une technique qui utilise des ondes sonores de haute fréquence (ultrasons) pouvant briser les liaisons moléculaires permettant ainsi de fragmenter l'ADN. Ce processus repose sur la cavitation acoustique, où des microbulles se forment et implosent, générant des forces mécaniques qui rompent les liaisons moléculaires. Les échantillons, placés dans un bain sonore à température contrôlée, subissent des cycles d'ultrasons alternant phases actives (ON) et repos (OFF) pour éviter une dégradation thermique.

Un programme de 8 cycles en position basse de 30 sec ON et 90 sec OFF à 4°C a été réalisé sur deux pools équimolaires de plasmides, le premier de 900 ng et le second de 9 µg dans 100µL final dans des tubes 0.5 ml Bioruptor™ Plus Microtubes (Diagénode). Ce paramétrage permet de garder des fragments assez longs pour notre protocole. Par la suite les produits de fragmentation ont été concentrés avec des billes SPRI Select (Beckman Coulters) en 2X avec une élution dans 30µL (Matériels et Méthodes 3.4.1.). Puis la taille des fragments d'intérêt a été récupérée en utilisant le BluePippin (kit 1,5 % agarose 250 pb-1,5 kb) en calibrant une range de sélection à 300-600pb avec une élution dans 40µL (Matériels et Méthodes 3.5.1.2.).

4.1.1.2. Ligation des adaptateurs / indexes

4.1.1.2.1. QIAseq 1-step Amplicon Library (QIAGEN®)

Nous avons utilisé le kit QIAseq 1-Step Amplicon Library (Qiagen®) pour préparer certaines librairies de séquençage, incluant la préparation de librairies à partir d'amplicons générés par PCR et RPA, à partir d'inserts excisés de plasmides bactériens ou de levure, et également à partir d'inserts excisés de produit d'amplification de plasmides par WGA.

Après récupération des séquences, préalablement digérées et purifiées, un mélange réactionnel de ligation des adaptateurs a été réalisé. Ce mélange a été préparé en ajoutant 100 ng d'ADN, 5 µL de tampon 1-step Amplicon Library 4 X, 2 µL d'adaptateurs, pour une concentration finale de 1 µM, et 1 µL de mix d'enzyme du kit, le tout dans un volume final de 25 µL suivi d'un vortex et d'une centrifugation brève. Les tubes ont ensuite été placés dans le thermocycleur pendant 30 minutes à 25°C, suivies de 20 minutes à 65°C, et d'un refroidissement à 4°C. Enfin les produits ont été purifiés doublement par billes magnétiques avec un ratio de 0,8 X en suivant le même processus que présenté dans le paragraphe 3.4.1.

Les librairies ont été quantifiées en utilisant le kit Qubit HS dsDNA (Life Technologies®) et stockées à -20°C jusqu'au séquençage. La taille des librairies a été confirmée en utilisant le kit Agilent dsDNA 930 Reagent (75–20 000 pb) sur le système Fragment Analyzer 5300 (Agilent Technologies®).

Les adaptateurs utilisés lors de ces ligations sont en « Y » et ont été sélectionnés parmi la plaque d'index de séquençage QIAseq UDI (*Unique Dual Indexes*) Y-Adapter plate A (Qiagen), CDI (*Combined Dual Index*) adapters (Qiagen) ou xGen UDI-UMI (*Unique Molecular Indentifiers*) Adapters (Integrated DNA Technologies®). Les adaptateurs de type xGen UDI-UMI incorporent des codes-barres moléculaires qui seront lus dans la lecture de l'index 1, ce qui permettra la correction des erreurs de PCR et de séquençage à partir de la création d'une séquence consensus sur la base de différentes lectures de la même molécule initiale.

Pour les librairies préparées à partir de la digestion par des enzymes de restriction de plasmides purs (c'est-à-dire utilisées directement comme tel après la réception du fournisseur GenScript), 0 (sans PCR), 1, 2, 5 et 10 cycles de PCR ont été utilisés avec les librairies incluant des adaptateurs CDI, tandis que 12, 16 et 20 cycles de PCR ont été utilisés avec des librairies incluant des adaptateurs xGen™ UDI-UMI. Pour les librairies xGen™ UDI-UMI, approximativement 20 000 molécules d'ADN par type de microsatellite ont été utilisées comme matrice pour l'étape PCR.

Pour les librairies préparées à partir de produits PCR, RPA et WGA, et de plasmides non purs (c'est-à-dire introduits dans des levures et/ou bactéries après réception du fournisseur GenScript®), aucun cycle PCR n'a été utilisé dans le protocole de préparation des librairies.

4.1.1.2.2. QIAseq FX DNA Library Kit (Qiagen)

Le kit QIAseq FX DNA Library Kit (Qiagen®) a été utilisé pour préparer certaines librairies de séquençage, notamment pour les expériences de séquençage complet du génome de levure, mais aussi pour des expériences de séquençage de plasmides purs contenant différents types de microsatellites.

Dans un premier temps, 1 µg de pools équimolaires de plasmides ou d'ADN génomique de levure a été utilisé pour effectuer une fragmentation aléatoire (*shearing*) enzymatique en utilisant 5 µL de FX Buffer 10X, 2.5 µL de FX Enhancer et 10 µL du FX Enzyme Mix dans 50 µL final, pendant 4 minutes d'incubation à 32°C. Une fois la digestion terminée, 5 µL d'adaptateurs (QIAseq UDI (*Unique Dual Indexes*) Y-Adapter plate A (Qiagen®) ou xGen UDI-UMI (*Unique Molecular Indentifiers*) Adapters

(Integrated DNA Technologies®)) ont été ajoutés à la préparation. Ensuite, 45 µL de Ligation Master Mix, composé de 20 µL de Ligation buffer 5X et de 10 µL de DNA Ligase et 15 µL d'eau, ont été ajoutés au mix et incubés pendant 15 minutes à 20°C. Par la suite, les produits ont été purifiés deux fois par des billes magnétiques avec un ratio de 0.8 X en suivant le même processus que présenté dans le paragraphe 3.4.1.

Les librairies ont été quantifiées en utilisant le kit Qubit HS dsDNA (Life Technologies®) et stockées à -20°C jusqu'au séquençage. La taille des librairies a été confirmée en utilisant le kit Agilent dsDNA 930 Reagent (75–20000 pb) sur le système Fragment Analyzer 5300 (Agilent Technologies®).

Les adaptateurs utilisés lors de ces ligations ont été sélectionnés parmi la plaque d'index de séquençage xGen UDI-UMI (Integrated DNA Technologies®).

Enfin, les librairies pré-amplifiées ont été dimensionnées (550-950 pb) en utilisant une cassette d'électrophorèse sur gel à 2% sur un système de sélection de taille BluePippin (Sage Science®) selon les instructions du fabricant. Pour les expériences sur les plasmides purs, 0 (sans PCR), 12 et 16 cycles de PCR ont été utilisés pour l'amplification des librairies en utilisant approximativement 20 000 molécules d'ADN par type de microsatellite comme modèles pour la PCR. Pour le WGS de levure, 0 cycle PCR a été utilisé.

4.1.1.2.3. xGen™ cfDNA & FFPE DNA Library Prep v2 MC Kit (Integrated DNA Technologies)

Le kit xGen Prism DNA Library Prep Kit (actuellement xGen™ cfDNA & FFPE DNA Library Prep v2 MC Kit) (Integrated DNA Technologies®) a été utilisé pour préparer certaines librairies de séquençage à partir de plasmides purs qui ont été fragmentés mécaniquement. Ce protocole permet de réaliser des librairies permettant d'incorporer deux codes-barres moléculaires de part et d'autre de l'insert permettant d'identifier les deux brins d'une même molécule d'ADN double brin et de corriger les erreurs après séquençage. Il correspond à une approche similaire au « *Duplex Sequencing* » (Kennedy et al., 2014).

Après récupération des séquences purifiées, un mélange réactionnel de réparation des extrémités a été réalisé. Pour cela, 3 µL de tampon End Repair ont été ajoutés à 1,5 µL d'enzyme End Repair, mélangés et ajoutés aux 25 µL de séquences. Une incubation de 30 min à 20°C a ensuite été réalisée, suivie d'un refroidissement à 4°C. Un lavage de « *post-end repair* » a été réalisé. A partir des 30 µL d'ADN, 2,5 X de billes ont été ajoutées, mélangées, le tout incubé 10 minutes puis placé sur un support aimanté. Le surnageant a été enlevé et un mélange réactionnel nommé « Ligation 1 » a été ajouté. Ce mélange a été préparé en ajoutant 12,5 µL de tampon 1, 1 µL d'adaptateurs 1 (adaptateurs ne se pouvant se lier qu'à l'extrémité 3' de l'insert) (Figure 123), et 1.5 µL de mix d'enzyme 1 du kit, le tout dans un volume final de 15 µL. Ce volume a servi à la resuspension des billes contenant les ADNs par aspiration-refoulement. Les tubes ont ensuite été placés dans le thermocycleur avec des conditions d'incubation de 15 minutes à 20°C, suivies de 15 minutes à 65°C, et d'un refroidissement à 4°C afin de réaliser efficacement la ligation des adaptateurs. Par la suite, un mélange réactionnel nommé « Ligation 2 » a été ajouté. Ce mélange a été préparé en ajoutant 2,25 µL de tampon 2, 2 µL d'adaptateurs 2, 0,25 µL d'enzyme A et 0,5 µL d'enzyme B du kit, le tout dans un volume final de 5 µL. Après mélange par aspiration-refoulement, les tubes ont ensuite été placés dans le thermocycleur avec des conditions d'incubation de 30 minutes à 65°C, et d'un refroidissement à 4°C. Ainsi une activité « *gap-filling* » à travers l'UMI de l'adaptateur et suivi d'une ligation aux extrémités 5' de l'insert (Figure 123). Enfin les produits ont été purifiés par l'ajout de 50 µL de PEG/NaCl, mélange et mise sur support aimanté. En

suivant les mêmes directives que dans le paragraphe 4.3.2 un unique lavage a été effectué en ajoutant 160 µL d'éthanol à 80 % et les billes ont été élué, après séchage, par 10 µL d'eau.

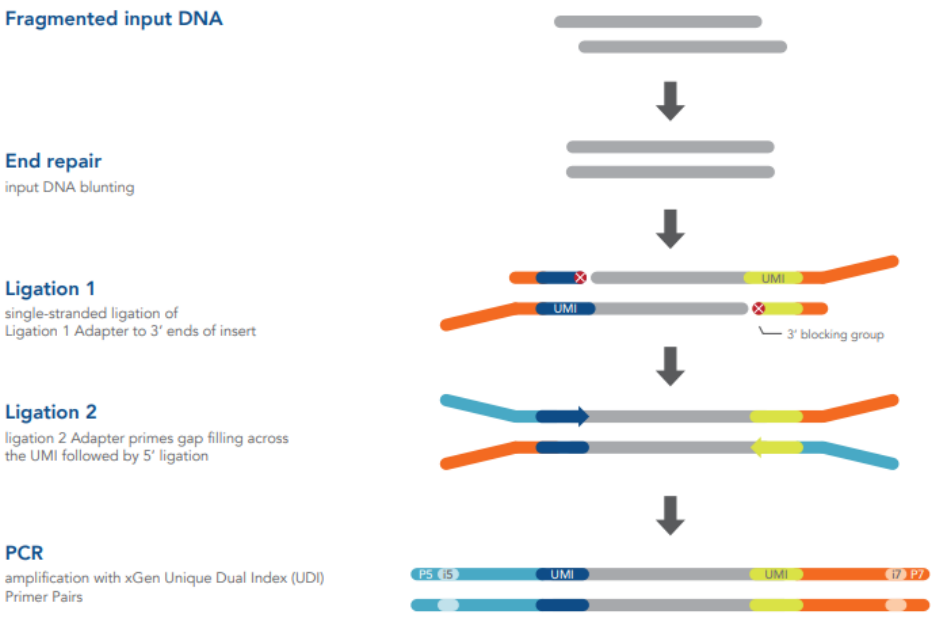

Figure 123 : Représentation schématique de la méthode de construction de librairie du xGen cfDNA & FFPE DNA Library Prep v2 MC Kit. (Integrated DNA Technologies).

Les librairies ont été quantifiées en utilisant le kit Qubit HS dsDNA (Life Technologies®) et stockées à -20°C jusqu'au séquençage. La taille des librairies a été confirmée en utilisant le kit Agilent dsDNA 930 Reagent (75–20 000 pb) sur le système Fragment Analyzer 5300 (Agilent Technologies®).

Dans cette méthode, des indexes de type UDI (marquant l'échantillon) ont ensuite été sélectionnés parmi la plaque xGen UDI primers (Integrated DNA Technologies®) nécessitant une phase de PCR pour introduire les index UDI et les séquences P5 et P7 nécessaires du séquençage short read.

16 cycles de PCR ont été utilisés pour l'amplification des librairies en faisant usage d'approximativement 20 000 molécules d'ADN par type de microsatellite comme modèles pour la PCR en utilisant les amorces xGen™ UDI (Integrated DNA Technologies®). Les librairies d'ADN xGen Prism finales incluaient deux UDI pour l'identification des échantillons et deux UMI de 8 nucléotides situés de chaque côté de l'insert pour la correction des erreurs de séquençage duplex.

4.1.2. Technologie Illumina en lecture courte – short read

Le séquençage Illumina est une technologie de séquençage de l'ADN dite *short-read* ou lecture courte basée sur la synthèse par incorporation de nucléotides fluorescents. On parle de séquençage par synthèse, tout comme le séquençage Sanger et le séquençage *long-read* de troisième génération PacBio. Les fragments d'ADN simple brins issus de librairies et portant les séquences p5 et p7 sont fixés sur une surface solide et amplifiés pour créer des clusters d'ADN identiques. Lors du séquençage, selon le séquenceur, des nucléotides marqués (ou non) avec des fluorochromes sont ajoutés. Une caméra haute résolution capture les signaux fluorescents émis par chaque nucléotide incorporé, permettant ainsi de déterminer la séquence des bases une à une. L'utilisation de cycles de lecture réversible et de

fluorochromes détachables permet de lire de longues séquences d'ADN de manière précise et efficace, avec une haute couverture et une faible marge d'erreur (Synthèse Bibliographique.5.3.2.).

4.1.2.1. Iseq 100

L'iSeq 100 est un séquenceur à petit débit commercialisé depuis 2018 allant jusqu'à 4 millions de reads en configuration single-end ou 8 millions en paired-end en 17h environ. Il fonctionne par un séquençage monocolore avec deux étapes chimiques et deux étapes d'imagerie par cycle de séquençage. Dans la chimie à un canal, l'adénine a une étiquette amovible et est marquée uniquement dans la première image. La cytosine possède un groupe de liaison pouvant se fixer à une étiquette et est marquée uniquement dans la seconde image. La thymine a une étiquette fluorescente permanente et est donc marquée dans les deux images, tandis que la guanine est toujours sombre (Figure 124).

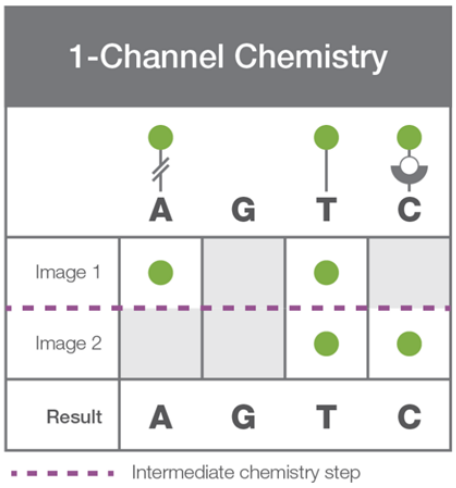

Figure 124 : Représentation schématique de la chimie du système monocolore.

Une fois que les librairies ont été réalisées avec différents couples d'index et dosées par KAPA, un pool équimolaire ou non a été réalisé. Le pool a ensuite été dilué pour atteindre 60 pM dans 100 µL auquel a été rajouté entre 30 et 50 % de PhiX à 60 pM. Enfin, 20 µL de ce pool global a été déposé dans la cartouche décongelée pendant au moins 8h à température ambiante et préchargée de réactifs. La flow cell est de type ordonné et est insérée dans la cartouche avant dépôt du pool de librairie. L'ensemble a été chargé dans l'instrument avant séquençage en 2x150 cycles et 2x8-10 cycles pour la lecture des index. Un fichier contenant les séquences des indexes des librairies a été préalablement uploadé dans le logiciel permettant à la suite du séquençage le démultiplexage informatique des séquences.

4.1.2.2. NextSeq 500

Le NextSeq 500 est un séquenceur à débit modéré commercialisé en 2014 allant jusqu'à 400 millions de reads en configuration paired-end en 12 h pour les runs cours ou 30 h pour les runs longs. Il fonctionne par un séquençage bicolore (Figure 125). Dans la chimie à deux canaux, la thymine a un colorant visible uniquement sur la première image (vert). La cytosine est marquée uniquement dans la seconde image (rouge). L'adénine a une étiquette fluorescente bicolore (vert + rouge) et est donc marquée dans les deux images, tandis que la guanine est toujours sombre (Figure 125).

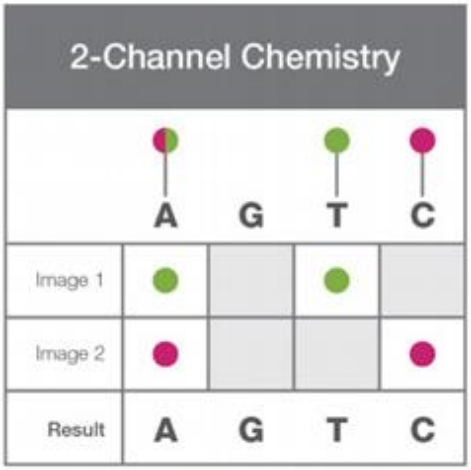

Figure 125 : Représentation schématique de la chimie du système bicolore.

Une fois que les librairies ont été réalisées avec différents couples d'index et dosées par KAPA, un pool équimolaire ou non a été réalisé. Le pool, dilué à 4 nM, a ensuite été dénaturé par le mélange de 5 µL du pool avec 5 µL de NaOH à 0,2 M., incubé à température ambiante pendant 1 minute. 5 µL de Tris-HCl, pH7 à 200 mM ont été ajoutés, mélangés et le tout a été vortexé pendant 1 minute à 280 g. Le pool de librairies dénaturé a ensuite été dilué avec 985 µL de Tampon HT1 pour atteindre 20 pM. Enfin, 97 µL du pool à 20 pM a été dilué avec 1203 µL de tampon HT1 afin d'obtenir 1.3 mL de pool à 1.5 pM. Concernant le PhiX, il a été dilué à 4 nM avec du RSB pour obtenir 25 µL puis a été dénaturé et dilué selon le même processus que le pool d'ADN. Enfin, 1 287 µL du pool a été mélangé avec 13 µL du PhiX, et l'ensemble a été déposé dans la cartouche décongelée à température ambiante pendant au moins 2h préchargée de réactifs. La flow cell est de type non-ordonnée, ce qui nécessite un dépôt précis de la librairie pour ne pas avoir une saturation en cluster. La cartouche avec la flowcell et le pool de librairies a été chargée dans l'instrument. Pour le séquençage, plusieurs cyclage ont été utilisés avec en read 1, 116, 118, 142, 150 et 149 cycles et en read 2, 35, 17, 133, 143 et 149 cycles respectivement. Enfin, pour la lecture des index, 18-8, 17-0, 17-8 2x8 cycles.

Un fichier contenant les séquences des indexes des librairies a été préalablement uploadé dans le logiciel permettant à la suite du séquençage, le démultiplexage informatique des séquences.

4.1.2.3. NovaSeq 6000

Le NovaSeq est un séquenceur à débit élevé commercialisé depuis 2017 allant jusqu'à 20 milliards de reads en configuration « *paired-end* » pour une durée de 13 à 48 h selon la configuration et la longueur des reads. Il peut fonctionner en bicolore et en quadricolore selon les réactifs et les configurations.

Dans le cadre de notre étude, un NovaSeq 6000 SP, bicolore comme le NextSeq 500, a été utilisé pour le séquençage des productions en bactérie et en levure (plasmide et WGS). Un cyclage de 2 x 150 bp a été réalisé, durant entre 24 et 30 heures. La réalisation du séquençage a été effectuée par Christian Daviaud sur la plateforme de production du CNRGH en mode « *ready-to-load* ».

4.2. Séquençage de troisième génération

4.2.1. Préparation des libraires

4.2.1.1. Linéarisation des plasmides

L'approche utilisée avec le séquençage PacBio en lecture longue a consisté à séquencer les plasmides purs de bactéries linéarisées. Le protocole de linéarisation des plasmides contient 1X de tampon de digestion, ici le NEBbuffer ™ r3.1 (New England Biolabs®), 10 U de l'enzyme de restriction PciI et 2 µg de chaque plasmide (sauf A15, car son barcode de 4nt unique ajoute un deuxième site de coupure). Ce mélange a été incubé pendant 60 minutes à 37°C pour permettre une digestion complète du plasmide puis l'enzyme a été inactivée pendant 20 minutes à 80°C. Les plasmides linéarisés ont été purifiés sur colonnes avec une élution dans 50 µL, quantifiés à l'aide du Qubit, regroupés de manière équimolaire par groupe de 4 et concentrés par précipitation à l'acétate de sodium, avant leur utilisation comme échantillon d'ADN pour la préparation des librairies.

4.2.1.2. SMRTbell prep kit 3.0 (Pacific Biosciences)

Nous avons utilisé le kit SMRTbell prep kit 3.0 (Pacific Biosciences®) pour préparer des librairies de séquençage qui ont été analysées sur le PacBio Sequel II.

Les librairies ont été préparées en utilisant des *Barcoded Overhang Adapters* (BOA) pour le multiplexage des amplicons (Pacific Biosciences®) et 200 ng de chaque pool d'ADN en utilisant le kit de préparation SMRTbell 3.0 (Pacific Biosciences®) selon les instructions du fabricant.

Dans un premier temps, une étape combinée de réparation des extrémités et d'ajout d'une queue avec une adénine protrusive a été réalisée. Les volumes présentés ci-après sont pour 8 réactions. La réparation a eu lieu en mélangeant 7,5 µL de DNA Prep Buffer, 1,3 µL de NAD et 2,6µL de DNA Damage Repair Mix dans un volume total de 25 µL auquel on ajoute 5 µL d'ADN puis incuber 30 min à 37°C. Concernant l'ajout de la queue, ont été ajoutés au mélange précédent, 2,5 µL de DNA Prep Buffer et 5 µL d'End Prep Mix dans un total de 25 µL, puis le tout a été incubé 30 min à 20°C puis 30 min à 65°C. Ensuite, pour effectuer la ligation des BOA à l'ADN, par tube, 1 µL d'adapteur, 5,3 µL de mix de ligation, 0,17 µL d'activateur de ligation et 0,17 µL d'additif de ligation ont été mélangés à l'ADN et incubés 1 h à 20°C. Les librairies ligaturées ont été purifiées en utilisant des billes AMPure PacBio® à 0,6X, quantifiées en utilisant le kit Qubit HS dsDNA (Life Technologies®) et stockées à -20°C jusqu'au séquençage. La taille des librairies a été confirmée en utilisant le kit Agilent dsDNA 930 Reagent (75–20000 pb) sur le système Fragment Analyzer 5300 (Agilent Technologies®).

4.2.2. Technologie PacBio Sequel II

Une fois que les librairies ont été préparées, elles sont transférées dans une plaque 96-puits et toute la procédure est réalisée par le système Sequel II.

Pour la réaction de séquençage long PacBio Sequel II, les amorces de séquençage v4 ont été hybridées et la polymérase d'ADN a été liée aux librairies multiplexées en utilisant le kit de liaison SMRTbell 3.0 (Pacific Biosciences)®, avant la purification des complexes ADN-protéine en utilisant des billes AMPure PacBio à 0,6X. Les librairies d'amplicons multiplexées ont été chargées sur une cellule SMRT et séquencées sur un instrument PacBio Sequel II utilisant la chimie C2 et un temps de lecture de 30 heures. Après la réaction de séquençage, les données de séquençage ont été démultiplexées par

SMRT Link (v9.0.0.92188) et les sub-reads brutes ainsi que les lectures de séquençage par consensus circulaire (CCS) ont été exportées pour les analyses bio-informatiques.

5. Bioinformatique

5.1. Projet 2

La totalité du développement et des analyses bioinformatiques concernant les résultats du projet 2 a été réalisée par le Dr. Yimin SHEN du CEPH. Chaque type de librairies a nécessité le développement d'un pipeline bio-informatique particulier en fonction des spécificités de chaque librairie, du type de technologie de séquençage et des analyses que l'on a souhaité réaliser. Les parties suivantes reprennent le détail de ces analyses.

5.1.1. Analyses des short-reads Illumina

5.1.1.1. Analyse standard de la longueur des allèles de microsatellites

Dans cette analyse, chaque échantillon contenait un mélange de microsatellites avec des séquences très similaires, à l'exception d'un code-barres spécifique de 4 nucléotides (Tableau 16). Même si une seule erreur ou mutation est acceptée dans ces séquences, leur identification peut être difficile. Les méthodes d'analyses classiques de similarité de séquences, de type blast, ou de cartographie, avec bowtie ou bwa, ne sont donc pas utilisables. Nous avons alors développé une méthode d'identification des microsatellites en ciblant la partie centrale de la séquence de part et d'autre des microsatellites avec des critères spécifiques (Figure 126.A). Cette méthode était basée sur la reconnaisance de 4 régions consécutives au niveau de la séquence comprenant (a) la région du code-barres de 4 nucléotides, (b) une séquence nucléotidique constante avant la séquence de microsatellite (5 pour les mononucléotides et 3 pour les dinucléotides), (c) le microsatellite, et (d) une séquence de 8 nucléotides après le microsatellite. La stratégie d'identification reposait sur ces 4 régions avec des seuils de mutation autorisés différents pour chaque région de 0, 1, 2 et 1 dans les régions a, b, c et d, respectivement (Tableau 16). Une fois les 4 régions identifiées, la longueur du microsatellite a pu être calculée et la distribution allélique représentée graphiquement.

Pour l'analyse de la taille des allèles, nous avons traité les fichiers FASTQ de l'iSeq100. Toutefois les fichiers FASTQ de NextSeq500 ont été extraits en démultiplexant les fichiers BCL après la fusion des 4 *lanes* via bcl2fastq (v2.20). Cette stratégie d'analyse des 4 régions a également été incluse dans le workflow d'analyse pour les corrections d'erreurs à partir d'un UMI unique (librairies de type II) ou d'UMI double (type III), ainsi que pour les séquences d'analyses des long-reads PacBio (type IV).

5.1.1.2. Correction des erreurs de PCR et de séquençage des longueurs d'allèles de microsatellites avec un UMI unique

Un UMI unique permet de regrouper les reads provenant de la même molécule d'ADN simple brin (librairie de type II). Pour analyser une séquence de microsatellite, les 4 régions décrites ci-dessus ont été identifiées et validées. En raison d'erreurs dans les reads et d'une faible diversité des séquences de référence, la séquence UMI n'était pas suffisante pour identifier tous les groupes d'UMI différents par l'approche de consensus d'UMI classique. Par conséquent, nous avons développé nos propres méthodes pour minimiser le risque d'erreur dans les traitements de données d'UMI unique (Figure 126.B).

Étant donné que les UMI simples n'étaient pas inclus dans les inserts (Figure 40.B), les données ont été extraites des fichiers BCL. Pour chaque *lane*, le code-barre a été déterminé et le démultiplexage effectué en incluant les données d'UMI via Picard (v2.8.2), puis les 4 lanes fusionnées par Samtools (v1.14) (Figure 126.B). Pour réduire l'erreur dans l'identification des groupes d'UMI, l'échantillon a été démultiplexé par type de microsatellite avant la cartographie en utilisant la méthode d'identification des microsatellites décrite ci-dessus. Les fichiers BAM de chaque microsatellite ont ensuite été convertis en FASTQ et alignés avec bwa (v0.7.17). Après avoir ajouté l'information UMI dans le fichier BAM aligné via Picard (v2.8.2), le groupe d'UMI a été identifié avec fgbio (v2.0.2) sans permettre de mutation pour réduire les erreurs. Pour chaque groupe d'UMI contenant au moins 3 lectures, nous avons identifié toutes les tailles de microsatellites et appliqué les quatre méthodes (('*Max. Freq.*' (fréquence maximale), '*Max Length*' (longueur maximale), '*Mean Length*' (longueur moyenne) et '*Median Length*' (longueur médiane) (Figure 45.A)) pour obtenir une taille consensus. Enfin, nous avons analysé le profil final des microsatellites basé sur la distribution des longueurs d'allèles de consensus.

5.1.1.3. Correction des erreurs de PCR et de séquençage des longueurs d'allèles de microsatellites avec un UMI double

Ici, comme pour l'analyse précédente, nous avons développé nos propres méthodes pour minimiser le risque d'erreur dans les traitements de données d'UMIs doubles (Figure 126.C)

Les reads des librairies de type III contiennent 2 séquences UMI (2 x 8 nt) à chaque extrémité de l'insert (Figure 40.B), ce qui augmente les combinaisons possibles d'UMI et réduit les erreurs d'identification des groupes d'UMI. Dans notre workflow (Figure 126.C), nous avons d'abord extrait les fichiers FASTQ en démultiplexant les fichiers BCL et en fusionnant les lanes via bcl2fastq (v2.20). Ensuite, les fichiers FASTQ ont été convertis en fichiers BAM non mappés par Picard (v2.8.2) et les informations concernant les UMI ont été extraites via fgbio (v2.0.2). Nous avons ensuite converti les fichiers BAM en FASTQ et effectué l'alignement avec bwa (v0.7.17). Après l'ajout des informations concernant les UMI dans les fichiers BAM mappés par Picard (v2.8.2), nous avons identifié le groupe d'UMI avec fgbio (v2.0.2) sans accepter une seule mutation. Nous avons analysé 3 types de reads (αβ standard, βα et αβ+βα) contenant au moins trois reads par groupe d'UMI et ainsi obtenu la longueur consensus d'un microsatellite pour chaque groupe d'UMI en utilisant les quatre méthodes (fréquence max., longueur max., longueur moyenne et longueur médiane). Nous avons également implémenté deux méthodes supplémentaires en comparant le consensus (fréquence max. et fréquence égale) obtenu des reads αβ et βα pour chaque groupe d'UMI. Ces méthodes de correction d'erreur combinées considéraient soit la longueur maximale entre les consensus αβ et βα (fréquence max.), soit sélectionnaient une longueur uniquement si les consensus αβ et βα étaient identiques (fréquence égale). Les profils finaux des microsatellites basés sur toutes ces longueurs d'allèles de consensus ont été utilisés pour une représentation graphique (Figure 126.C).

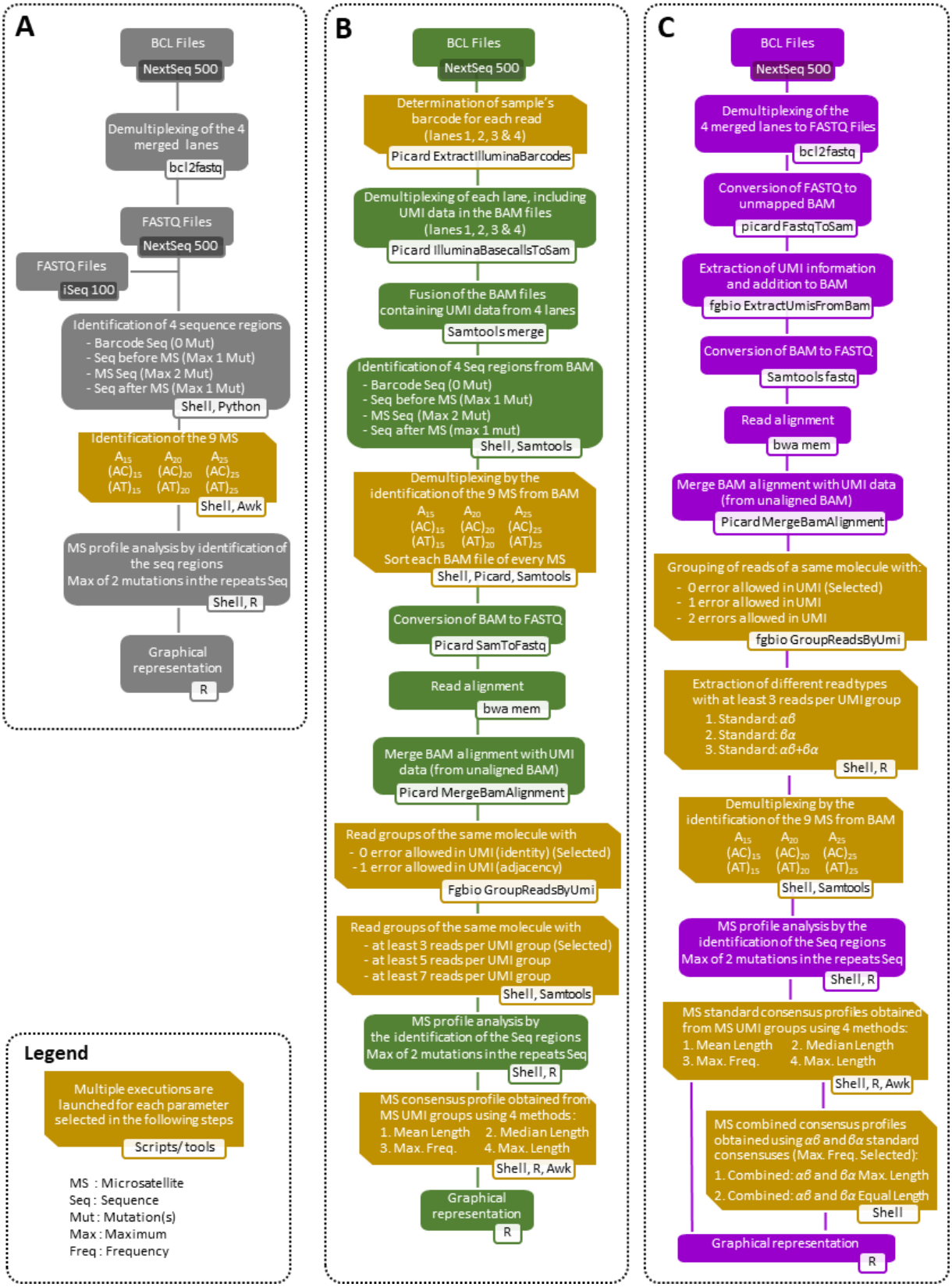

Figure 126 : Workflow des différentes analyses bioinformatiques réalisées à partir des données de séquençage en short reads Illumina. A. Workflow pour l'analyse de la longueur des allèles de microsatellites, utilisé avec les librairies de type I, II et III. B. Workflow de correction d'erreurs avec un UMI unique, utilisé avec les librairies de type II. C. Workflow de correction d'erreurs combinée et de séquençage en duplex, utilisé avec les librairies de type III. L'indication « (Selected) » précise les analyses, caractéristiques ou options dont les résultats sont présentés dans le manuscrit.

5.1.2. Analyses des long-reads PacBio

Deux types d'analyses ont été réalisées en utilisant les données CCS BAM et les données brutes de subreads BAM (Figure 127). La taille des fichiers de ces deux sources est très différente : par rapport aux fichiers CCS, d'environ 400 Mo, les fichiers de subreads brutes sont mille fois plus volumineux (environ 400 Go).

5.1.2.1. Analyse standard de la distribution des longueurs d'allèles de microsatellites

Pour évaluer la performance du profil en fonction de la qualité des lectures CCS, nous avons démultiplexé le fichier CCS BAM selon différents scores de qualité : QC20, QC30, QC40, QC50, QC60, QC70 et QC80, en utilisant SMRTLink (v11.0.0.146107). Comme pour le séquençage en short-read, notre échantillon contient un mélange de microsatellites aux séquences proches. Pour cette raison, nous avons utilisé la méthode d'identification des microsatellites basée sur la reconnaissance des 4 régions consécutives : a, b, c et d. Nous avons extrait les séquences CCS BAM non alignées après avoir identifié les neuf microsatellites, puis analysé leurs profils de taille d'allèles (Figure 127).

5.1.2.2. Analyse interne des données de subreads brutes pour la distribution des longueurs d'allèles de microsatellites

Nous avons également analysé les données de subreads brutes en utilisant les fichiers subreads BAM. Le démultiplexage a été réalisé selon le score de qualité du code-barre de 70 à l'aide de SMRTtools (v12.0.0.177059) lima. Nous avons ensuite sélectionné aléatoirement différents nombres de subreads par groupe : 3, 5, 7, 11, 15, 21, 25, 35 et 45. Tous les microsatellites ont été identifiés en utilisant l'approche d'identification basée sur les 4 régions consécutives : a, b, c et d. Enfin, pour chaque groupe de subreads, la longueur consensus a été obtenue en utilisant les quatre méthodes décrites précédemment (fréquence max., longueur max., longueur moyenne et longueur médiane), et nous avons analysé les profils de distribution des allèles.

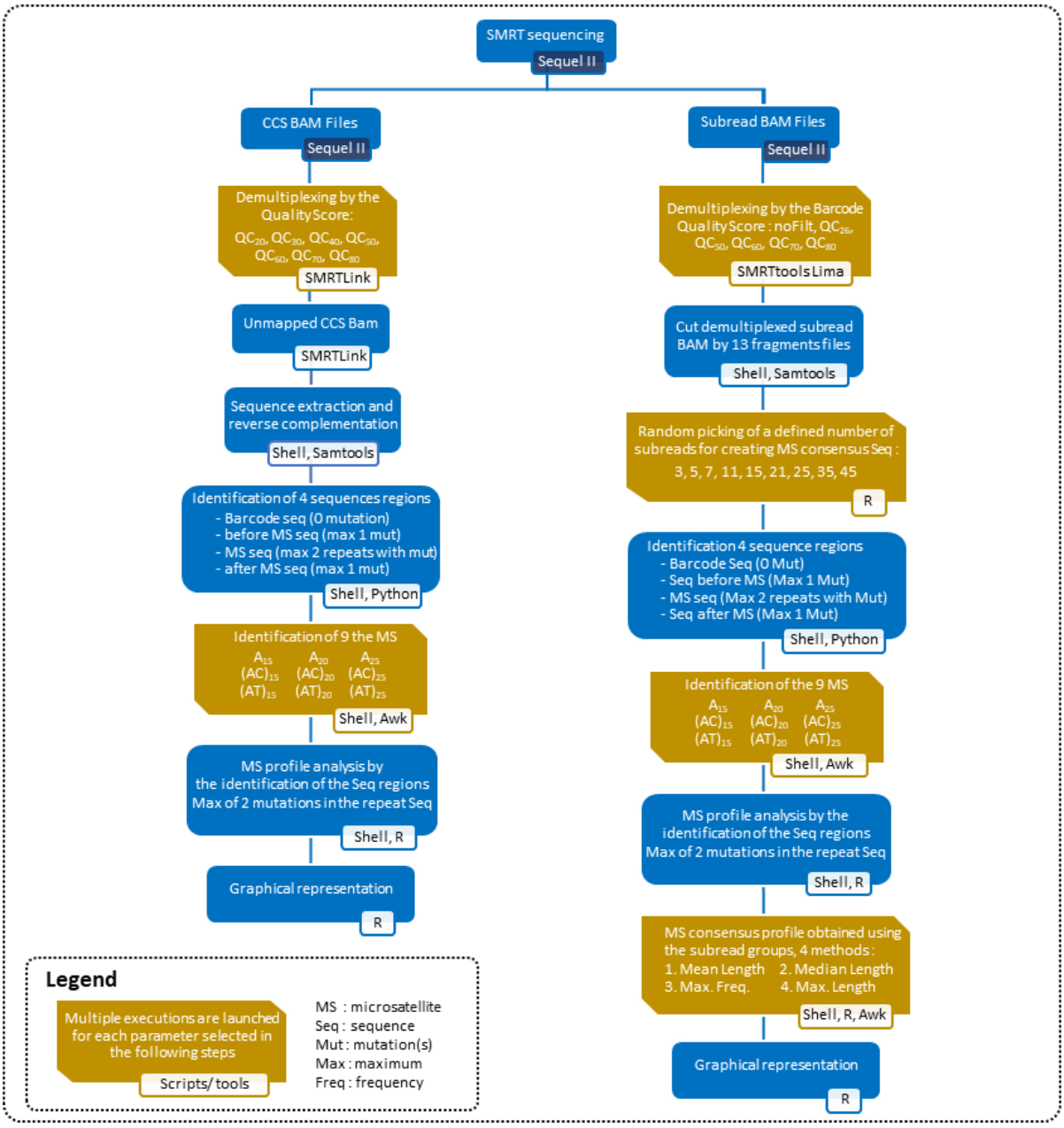

Figure 127 : Workflow des différentes analyses bioinformatiques réalisées à partir des données de séquençage en long-reads de PacBio (librairies de type IV).

5.2. Projet 3

La totalité des développements et des analyses bioinformatiques concernant les résultats du projet 3 a été réalisée par Emmanuel TUBACHER du CEPH en Perl. Les parties suivantes détaillent le protocole utilisé pour l'analyse des données.

5.2.1. R1R2

Dans cette analyse, l'objectif est de pouvoir définir l'évolution de la distribution de taille des microsatellites après des événements de réplication de l'ADN à partir d'un microsatellite originel pur

(100% d'allèles d'une taille connue), en se focalisant sur l'identification des séquences environnantes du microsatellite pour en déduire sa taille. Pour cela, aucun alignement par base n'est réalisé.

Dans un premier temps, une fois les séquences récupérées, on a analysé les 8 premières bases suivant le microsatellite en 3' afin de confirmer que le *read* contenait le microsatellite sur toute sa longueur en n'acceptant aucune erreur. Une deuxième séquence de reconnaissance du microsatellite a été analysée en 5', permettant de calculer le nombre de répétitions et le motif de départ. Lors de l'identification de la séquence en amont du microsatellite (en 5'), 3 erreurs sont autorisées en dehors de la séquence des index d'identification (Tableau 16).

Lors de la déduction de la taille, les proportions de bases doivent concorder à 80% avec celles du microsatellite attendu (pour un microsatellite $(A)_n$, il doit y avoir 80% de base A, mais pour un microsatellite $(AC)_n$, il doit y avoir environ 40% de base A et 40% de base C). Une fois ces bases validées sur le *read* 1 (R1) et le *read* 2 (R2), une comparaison est réalisée entre les 2 *reads*. Si la séquence du R1 correspond à la séquence du R2 alors l'analyse se poursuit, sinon, dans le cas d'une non-concordance (erreur ou absence d'un *reads*) les deux *reads* sont éliminés de l'analyse. Dans le cas particulier des microsatellites di-nucléotidiques, si la séquence de microsatellites est d'une longueur déduite impaire, la séquence est rejetée car il nous est alors impossible de savoir s'il y a eu dégradation d'un motif ou insertion partielle.

De par les critères que nous avons sélectionnés, la variabilité des runs, de la technologie et des microsatellites, le système d'analyse suit un principe ON/OFF. C'est-à-dire que soit le run est bon et sera facilement analysé, soit les données sont de mauvaise qualité et ne valident pas le premier critère. Les résultats et critères ne permettent pas d'entre deux.

5.2.2. BASIC

Dans cette analyse, à l'instar de l'analyse R1R2, l'objectif est de pouvoir définir le profil de distribution de taille des microsatellites après plusieurs cycles de réplication de l'ADN à partir d'un plasmide purs originel (100% d'une même taille) en se focalisant sur l'identification du microsatellite puis l'analyse de son environnement pour valider la qualité. Pour cela, aucun alignement par base n'est réalisé.

Un seul read est analysé, et dans un premier temps, avec la recherche de la séquence du microsatellite. Une fois identifié, nous avons analysé les séquences en 3' et en 5' du microsatellite afin de déterminer la taille d'origine et la qualité du run comme précédemment.

Cette méthode d'analyse est plus permissive sur l'environnement du microsatellite mais beaucoup plus astringente sur la qualité du microsatellite en lui-même comparé à l'analyse de R1R2.

5.2.3. Whole Genome Sequencing de Levure

Afin d'analyser correctement les données de séquençage du *Whole Genome Sequencing* (WGS) de levure réalisé sur un NovaSeq 6000 plusieurs étapes sont nécessaires. Premièrement, les fichiers FastQ obtenus ont dû être prétraités. Pour chaque session de séquençage, les données des deux lanes ont été fusionnées (*read* 1 et *read* 2 séparément) et réorganisées par nom de *read*, avec un renommage des noms de *read* dupliqués pour éviter les erreurs dans le processus d'alignement. Deuxièmement, les *reads* FastQ ont été alignées contre les séquences de référence de *mlh1* et *msh2*, ainsi que contre la séquence de *KanMX* (Walker et al., 2003), afin de vérifier le *knockout* effectif de ces gènes dans les échantillons

correspondants et leur remplacement par la cassette *KanMX* (Matériels et Méthodes 1.2.). Cette analyse a confirmé l'efficacité du processus de knockout : aucune lecture n'a été détectée pour *mlh1* et *msh2* dans les échantillons *knockouts*, tandis que des reads correspondant à *KanMX* étaient présentes. Troisièmement, nous avons récupéré 1 968 positions de microsatellites disponibles dans MSDB pour S288C (Avvaru et al., 2020) et extrait les séquences génomiques correspondantes en format FASTA (séquences des microsatellites et 30 bases en amont (5') et en aval (3')) à partir du génome de référence de la levure. Enfin, chaque session de séquençage en paires NOVASEQ a été alignée sur notre base de données des séquences de microsatellites de levure en utilisant BWA (Li and Durbin, 2009). Chaque correspondance a été analysée pour déterminer si le microsatellite était totalement inclus dans la séquence de lecture et pour estimer la taille du microsatellite en utilisant les informations de la chaîne CIGAR (*Compressed Idiosyncratic Gapped Alignment Report*). Les tailles des microsatellites ont ensuite été agrégées pour calculer les distributions de tailles pour chaque microsatellite. Les fréquences ont été calculées pour les microsatellites ayant plus de 200 lectures alignées.

6. Analyses statistiques et représentations graphiques

6.1. Schématisation

Les schémas et illustrations inclus dans cette thèse ont été générés à l'aide de Biorender.com (BioRender, 2023, https://www.biorender.com) ainsi que Microsoft PowerPoint, version 365 (Microsoft, 2023, https://www.microsoft.com/).

6.2. Représentation graphique

Les représentations graphiques des différents résultats inclus dans cette thèse ont été générés à l'aide de GraphPad Prism (version 10.4.0(621)) (GraphPad Software, 2023, https://www.graphpad.com).

6.3. Statistiques

Les analyses statistiques des différentes représentations graphiques, telles que les Mann-Whitenery, les regressions et les correlations pearson ont été réalisées à l'aide de GraphPad Prism, (version 10.4.0(621)) (GraphPad Software, 2023, https://www.graphpad.com).

7. Intelligence artificielle

L'intelligence artificielle par le biais de l'outil ChatGPT a été utilisée lors de la rédaction de cette thèse afin de faciliter la recherche bilbliographique, et la reformulation de certaines phrases ainsi que DEPPLWrite pour la correction de fautes de synthaxes.

Liste des communications scientifiques

1. Chapitre publié

Jeanjean, S.I., Renault, V., Daunay, A., Shen, Y., Hardy, L.M., Deleuze, J.-F., How-Kit, A., 2023. **LT-RPA: An Isothermal DNA Amplification Approach for Improved Microsatellite Genotyping and Microsatellite Instability Detection.** Methods Mol Biol 2621, 91–109. https://doi.org/10.1007/978-1-0716-2950-5_7

2. Article soumis

Titre: A detailed analysis of second and third generation sequencing approaches for accurate length determination of short tandem repeats and homopolymers. **Auteurs** : Sophie I. Jeanjean, Yimin Shen, Lise M. Hardy, Antoine Daunay, Antonio Alberdi, Emmanuel Tubacher, Jean-François Deleuze & Alexandre How-Kit. **Soumis à Nucleic Acids Research** le 12 juillet 2024, retour avec révision le 08 août 2024 et resoumission prévue pour fin décembre 2024.

3. Article en finalisation

Titre: Inferring microsatellite mutation rates from allele frequencies obtained from (non-) isothermal DNA amplification approaches using thirty-two genetic models. **Auteurs** : Rojo Nofidiantsoa, Sophie I. Jeanjean, Nasreddine Rajoua, Antoine Daunay, Mourad Sahbatou, Jean-François Deleuze & Alexandre How-Kit.

4. Posters

Titre: Inferring microsatellite mutation rates from allele frequencies obtained from (non-) isothermal DNA amplification approaches using thirty-two genetic models. **Auteurs** : Rojo Nofidiantsoa, Sophie I. Jeanjean, Nasreddine Rajoua, Antoine Daunay, Mourad Sahbatou, Jean-François Deleuze & Alexandre How-Kit. Poster présenté à **l'European Society of Human Genetics** (ESHG) Annual Meeting, Glasgow, United Kingdom en Juin 2023.

Titre: Inferring microsatellite mutation rates from allele frequencies obtained from (non-) isothermal DNA amplification approaches using thirty-two genetic models. **Auteurs** : Rojo Nofidiantsoa, Sophie I. Jeanjean, Nasreddine Rajoua, Antoine Daunay, Mourad Sahbatou, Jean-François Deleuze & Alexandre How-Kit. Poster présenté à **l'Internation Symposium on Human Genomics** (ISHG) Annual Meeting, Paris, France en Septembre 2023.

Bibliographie

Aaij, C., Borst, P., 1972. The gel electrophoresis of DNA. Biochim Biophys Acta 269, 192–200. https://doi.org/10.1016/0005-2787(72)90426-1

Abdul-Muneer, P.M., 2014. Application of Microsatellite Markers in Conservation Genetics and Fisheries Management: Recent Advances in Population Structure Analysis and Conservation Strategies. Genet Res Int 2014, 691759. https://doi.org/10.1155/2014/691759

Abdulovic, A.L., Hile, S.E., Kunkel, T.A., Eckert, K.A., 2011. The in vitro fidelity of yeast DNA polymerase δ and polymerase ε holoenzymes during dinucleotide microsatellite DNA synthesis. DNA Repair 10, 497–505. https://doi.org/10.1016/j.dnarep.2011.02.003

Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C., Wagner, J., McDaniel, J., Olson, N.D., Sauria, M.E.G., Vollger, M.R., Rhie, A., Meredith, M., Martin, S., Lee, J., Koren, S., Rosenfeld, J.A., Paten, B., Layer, R., Chin, C.-S., Sedlazeck, F.J., Hansen, N.F., Miller, D.E., Phillippy, A.M., Miga, K.H., McCoy, R.C., Dennis, M.Y., Zook, J.M., Schatz, M.C., 2022. A complete reference genome improves analysis of human genetic variation. Science 376, eabl3533. https://doi.org/10.1126/science.abl3533

Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., Durbin, R., 2011. Dindel: accurate indel calls from short-read data. Genome Res 21, 961–973. https://doi.org/10.1101/gr.112326.110

Alhopuro, P., Sammalkorpi, H., Niittymäki, I., Biström, M., Raitila, A., Saharinen, J., Nousiainen, K., Lehtonen, H.J., Heliövaara, E., Puhakka, J., Tuupanen, S., Sousa, S., Seruca, R., Ferreira, A.M., Hofstra, R.M.W., Mecklin, J.-P., Järvinen, H., Ristimäki, A., Orntoft, T.F., Hautaniemi, S., Arango, D., Karhu, A., Aaltonen, L.A., 2012. Candidate driver genes in microsatellite-unstable colorectal cancer. Int J Cancer 130, 1558–1566. https://doi.org/10.1002/ijc.26167

Alotaibi, H., Ricciardone, M.D., Ozturk, M., 2008. Homozygosity at variant MLH1 can lead to secondary mutation in NF1, neurofibromatosis type I and early onset leukemia. Mutat Res 637, 209–214. https://doi.org/10.1016/j.mrfmmm.2007.08.003

Amos, W., Filipe, L.N.S., 2014. Microsatellite frequencies vary with body mass and body temperature in mammals, suggesting correlated variation in mutation rate. PeerJ 2, e663. https://doi.org/10.7717/peerj.663

Ananda, G., Walsh, E., Jacob, K.D., Krasilnikova, M., Eckert, K.A., Chiaromonte, F., Makova, K.D., 2013. Distinct Mutational Behaviors Differentiate Short Tandem Repeats from Microsatellites in the Human Genome. Genome Biology and Evolution 5, 606–620. https://doi.org/10.1093/gbe/evs116

Angers, B., Bernatchez, L., 1997. Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. Molecular Biology and Evolution 14, 230–238. https://doi.org/10.1093/oxfordjournals.molbev.a025759

Anmarkrud, J.A., Kleven, O., Bachmann, L., Lifjeld, J.T., 2008. Microsatellite evolution: Mutations, sequence variation, and homoplasy in the hypervariable avian microsatellite locus HrU10. BMC Evolutionary Biology 8, 138. https://doi.org/10.1186/1471-2148-8-138

Antão-Sousa, S., Conde-Sousa, E., Gusmão, L., Amorim, A., Pinto, N., 2022. Estimations of Mutation Rates Depend on Population Allele Frequency Distribution: The Case of Autosomal Microsatellites. Genes 13, 1248. https://doi.org/10.3390/genes13071248

Armour, J., Jeffreys, A., 1992. Biology and applications of human minisatellite loci. Current Opinion in Genetics & Development 2, 850–856. https://doi.org/10.1016/S0959-437X(05)80106-6

Arning, L., Nguyen, H.P., 2021. Huntington disease update: new insights into the role of repeat instability in disease pathogenesis. Med Genet 33, 293–300. https://doi.org/10.1515/medgen-2021-2101

Athanasopoulou, K., Boti, M.A., Adamopoulos, P.G., Skourou, P.C., Scorilas, A., 2021. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. Life (Basel) 12, 30. https://doi.org/10.3390/life12010030

Avery, O.T., MacLeod, C.M., McCarty, M., 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med 79, 137–158.

Avvaru, A.K., Sharma, D., Verma, A., Mishra, R.K., Sowpati, D.T., 2020. MSDB: a comprehensive, annotated database of microsatellites. Nucleic Acids Research 48, D155–D159. https://doi.org/10.1093/nar/gkz886

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., Mori, H., 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular Systems Biology 2, 2006.0008. https://doi.org/10.1038/msb4100050

Bae, J.H., Liu, R., Roberts, E., Nguyen, E., Tabrizi, S., Rhoades, J., Blewett, T., Xiong, K., Gydush, G., Shea, D., An, Z., Patel, S., Cheng, J., Sridhar, S., Liu, M.H., Lassen, E., Skytte, A.-B., Grońska-Pęski, M., Shoag, J.E., Evrony, G.D., Parsons, H.A., Mayer, E.L., Makrigiorgos, G.M., Golub, T.R., Adalsteinsson, V.A., 2023. Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. Nat Genet 55, 871–879. https://doi.org/10.1038/s41588-023-01376-0

Bagshaw, A.T.M., 2017. Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. Genome Biology and Evolution 9, 2428–2443. https://doi.org/10.1093/gbe/evx164

Bakhtiari, M., Park, J., Ding, Y.-C., Shleizer-Burko, S., Neuhausen, S.L., Halldórsson, B.V., Stefánsson, K., Gymrek, M., Bafna, V., 2021. Variable number tandem repeats mediate the expression of proximal genes. Nat Commun 12, 2075. https://doi.org/10.1038/s41467-021-22206-z

Balaresque, P., 2007. Les microsatellites des génomes eucaryotes: De leur cycle de vie et de leur neutralité. Med Sci (Paris) 23, 729–734. https://doi.org/10.1051/medsci/20072389729

Balaresque, P., Toupance, B., Heyer, E., Crouau-Roy, B., 2003. Evolutionary Dynamics of Duplicated Microsatellites Shared by Sex Chromosomes. J Mol Evol 57, S128–S137. https://doi.org/10.1007/s00239-003-0018-z

Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y., van Duijn, K., Vermeulen, M., Brauer, S., Decorte, R., Poetsch, M., von Wurmb-Schwark, N., de Knijff, P., Labuda, D., Vézina, H., Knoblauch, H., Lessig, R., Roewer, L., Ploski, R., Dobosz, T., Henke, L., Henke, J., Furtado, M.R., Kayser, M., 2010. Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications. Am J Hum Genet 87, 341–353. https://doi.org/10.1016/j.ajhg.2010.08.006

Balloux, F., Ecoffey, E., Fumagalli, L., Goudet, J., Wyttenbach, A., Hausser, J., 1998. Microsatellite conservation, polymorphism, and GC content in shrews of the genus Sorex (Insectivora, Mammalia). Mol Biol Evol 15, 473–475. https://doi.org/10.1093/oxfordjournals.molbev.a025945

Bandipalliam, P., 2005. Syndrome of Early Onset Colon Cancers, Hematologic Malignancies & Features of Neurofibromatosis in HNPCC Families with Homozygous Mismatch Repair Gene Mutations. Familial Cancer 4, 323–333. https://doi.org/10.1007/s10689-005-8351-6

Baptiste, B.A., Ananda, G., Strubczewski, N., Lutzkanin, A., Khoo, S.J., Srikanth, A., Kim, N., Makova, K.D., Krasilnikova, M.M., Eckert, K.A., 2013. Mature Microsatellites: Mechanisms Underlying Dinucleotide Microsatellite Mutational Biases in Human Cells. G3 Genes|Genomes|Genetics 3, 451–463. https://doi.org/10.1534/g3.112.005173

Baptiste, B.A., Jacob, K.D., Eckert, K.A., 2015. Genetic evidence that both dNTP-stabilized and strand slippage mechanisms may dictate DNA polymerase errors within mononucleotide microsatellites. DNA Repair 29, 91–100. https://doi.org/10.1016/j.dnarep.2015.02.016

Baudrin, L.G., Deleuze, J.-F., How-Kit, A., 2018. Molecular and Computational Methods for the Detection of Microsatellite Instability in Cancer. Front. Oncol. 8, 621. https://doi.org/10.3389/fonc.2018.00621

Bauer, D.W., Butt, N., Hornyak, J.M., Perlin, M.W., 2020. Validating TrueAllele® Interpretation of DNA Mixtures Containing up to Ten Unknown Contributors. J Forensic Sci 65, 380–398. https://doi.org/10.1111/1556-4029.14204

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A.P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E.C., Galloway, J.G., Gladstein, A.L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W.W., Lohse, K., Matschiner, M., Nelson, D., Pope, N.S., Quinto-Cortés, C.D., Rodrigues, M.F., Saunack, K., Sellinger, T., Thornton, K., van Kemenade, H., Wohns, A.W., Wong, Y., Gravel, S., Kern, A.D., Koskela, J., Ralph, P.L., Kelleher, J., 2022. Efficient ancestry and mutation simulation with msprime 1.0. Genetics 220, iyab229. https://doi.org/10.1093/genetics/iyab229

Bayliss, C.D., Dixon, K.M., Moxon, E.R., 2004. Simple sequence repeats (microsatellites): mutational mechanisms and contributions to bacterial pathogenesis. A meeting review. FEMS Immunology & Medical Microbiology 40, 11–19. https://doi.org/10.1016/S0928-8244(03)00325-0

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573–580. https://doi.org/10.1093/nar/27.2.573

Bertola, D.R., Cao, H., Albano, L.M.J., Oliveira, D.P., Kok, F., Marques-Dias, M.J., Kim, C.A., Hegele, R.A., 2006. Cockayne syndrome type A: novel mutations in eight typical patients. J Hum Genet 51, 701–705. https://doi.org/10.1007/s10038-006-0011-7

Bhargava, A., Fuentes, F.F., 2009. Mutational Dynamics of Microsatellites. Molecular Biotechnology. https://doi.org/10.1007/s12033-009-9230-4

Biller, L.H., Syngal, S., Yurgelun, M.B., 2019. Recent advances in Lynch syndrome. Fam Cancer 18, 211–219. https://doi.org/10.1007/s10689-018-00117-1

Blanco, L., Salas, M., 1984. Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication. Proc Natl Acad Sci U S A 81, 5325–5329.

Boland, C.R., Goel, A., 2010. Microsatellite Instability in Colorectal Cancer. Gastroenterology 138, 2073-2087.e3. https://doi.org/10.1053/j.gastro.2009.12.064

Bonneville, R., Krook, M.A., Kautto, E.A., Miya, J., Wing, M.R., Chen, H.-Z., Reeser, J.W., Yu, L., Roychowdhury, S., 2017. Landscape of Microsatellite Instability Across 39 Cancer Types. JCO Precis Oncol 1, PO.17.00073. https://doi.org/10.1200/PO.17.00073

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N.D., Matschiner, M., Mendes, F.K., Müller, N.F., Ogilvie, H.A., Plessis, L. du, Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M.A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., Drummond, A.J., 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology 15, e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Boyer, J.C., Yamada, N.A., Roques, C.N., Hatch, S.B., Riess, K., Farber, R.A., 2002. Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Hum Mol Genet 11, 707–713. https://doi.org/10.1093/hmg/11.6.707

Brazda, V., Fojta, M., Bowater, R.P., 2020. Structures and stability of simple DNA repeats from bacteria. Biochemical Journal 477, 325–339. https://doi.org/10.1042/BCJ20190703

Bright, J.-A., Taylor, D., McGovern, C., Cooper, S., Russell, L., Abarno, D., Buckleton, J., 2016. Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. Forensic Science International: Genetics 23, 226–239. https://doi.org/10.1016/j.fsigen.2016.05.007

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., Rolf, B., 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62, 1408–1415.

Britten, R.J., Kohne, D.E., 1968. Repeated Sequences in DNA. Science 161, 529–540. https://doi.org/10.1126/science.161.3841.529

Buono, V., Burgio, S., Macrì, N., Catania, G., Hauffe, H.C., Mucci, N., Davoli, F., 2022. Microsatellite Characterization and Panel Selection for Brown Bear (Ursus arctos) Population Assessment. Genes (Basel) 13, 2164. https://doi.org/10.3390/genes13112164

Buschiazzo, E., Gemmell, N.J., 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28, 1040–1050. https://doi.org/10.1002/bies.20470

Butler, J.M., 2006. Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing. Journal of Forensic Sciences 51, 253–265. https://doi.org/10.1111/j.1556-4029.2006.00046.x

Çağlayan, M., Bilgin, N., 2012. Temperature dependence of accuracy of DNA polymerase I from *Geobacillus anatolicus*. Biochimie 94, 1968–1973. https://doi.org/10.1016/j.biochi.2012.05.019

Cann, H.M., 1992. CEPH maps. Current Opinion in Genetics & Development 2, 393–399. https://doi.org/10.1016/S0959-437X(05)80148-0

Cao, M.D., Balasubramanian, S., Bodén, M., 2015. Sequencing technologies and tools for short tandem repeat variation detection. Brief Bioinform 16, 193–204. https://doi.org/10.1093/bib/bbu001

Cao, M.D., Tasker, E., Willadsen, K., Imelfort, M., Vishwanathan, S., Sureshkumar, S., Balasubramanian, S., Bodén, M., 2014. Inferring short tandem repeat variation from paired-end *short reads*. Nucleic Acids Res 42, e16. https://doi.org/10.1093/nar/gkt1313

Carreras, C., Pascual, M., Cardona, L., Aguilar, A., Margaritoulis, D., Rees, A., Türkozan, O., Levy, Y., Aureggi, M., Khalil, M., 2007. The genetic structure of the loggerhead sea turtle (Caretta caretta) in the Mediterranean as revealed by nuclear and mitochondrial DNA and its conservation implications. Conservation Genetics 8, 761–775. https://doi.org/10.1007/s10592-006-9224-8

Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J., Deka, R., 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci U S A 94, 1041–1046.

Chandra, D., Mishra, V.C., Raina, A., Raina, V., 2022. Mutation rate evaluation at 21 autosomal STR loci: Paternity testing experience. Leg Med (Tokyo) 58, 102080. https://doi.org/10.1016/j.legalmed.2022.102080

Chao, E.C., Lipkin, S.M., 2006. Molecular models for the tissue specificity of DNA mismatch repair-deficient carcinogenesis. Nucleic Acids Res 34, 840–852. https://doi.org/10.1093/nar/gkj489

Chen, K.-C., 2012. Microsatellite DNA: Population Genetics and Forensic Applications. Tested Studies for Laboratory Teaching: Proceedings of the Conference of the Association for Biology Laboratory Education, K. McMahon, Editor 33.

Chiu, R., Rajan-Babu, I.-S., Friedman, J.M., Birol, I., 2021. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biol 22, 224. https://doi.org/10.1186/s13059-021-02447-3

Cho, D.H., Tapscott, S.J., 2007. Myotonic dystrophy: Emerging mechanisms for DM1 and DM2. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, The Muscular Dystrophies: Molecular Basis and Therapeutic Strategies 1772, 195–204. https://doi.org/10.1016/j.bbadis.2006.05.013

Christopher, J., Thorsen, A.-S., Abujudeh, S., Lourenço, F.C., Kemp, R., Potter, P.K., Morrissey, E., Hazelwood, L., Winton, D.J., 2019. Quantifying Microsatellite Mutation Rates from Intestinal Stem Cell Dynamics in Msh2-Deficient Murine Epithelium. Genetics 212, 655–665. https://doi.org/10.1534/genetics.119.302268

Chu, X.-L., Zhang, B.-W., Zhang, Q.-G., Zhu, B.-R., Lin, K., Zhang, D.-Y., 2018. Temperature responses of mutation rate and mutational spectrum in an Escherichia coli strain and the correlation with metabolic rate. BMC Evolutionary Biology 18, 126. https://doi.org/10.1186/s12862-018-1252-8

Chung, J., Maruvka, Y.E., Sudhaman, S., Kelly, J., Haradhvala, N.J., Bianchi, V., Edwards, M., Forster, V.J., Nunes, N.M., Galati, M.A., Komosa, M., Deshmukh, S., Cabric, V., Davidson, S., Zatzman, M., Light, N., Hayes, R., Brunga, L., Anderson, N.D., Ho, B., Hodel, K.P., Siddaway, R., Morrissy, A.S., Bowers, D.C., Larouche, V., Bronsema, A., Osborn, M., Cole, K.A., Opocher, E., Mason, G., Thomas, G.A., George, B., Ziegler, D.S., Lindhorst, S., Vanan, M., Yalon-Oren, M., Reddy, A.T., Massimino, M., Tomboc, P., Van Damme, A., Lossos, A., Durno, C., Aronson, M., Morgenstern, D.A., Bouffet, E., Huang, A., Taylor, M.D., Villani, A., Malkin, D., Hawkins, C.E., Pursell, Z.F., Shlien, A., Kunkel, T.A., Getz, G., Tabori, U., 2021. DNA Polymerase and Mismatch Repair Exert

Distinct Microsatellite Instability Signatures in Normal and Malignant Human Cells. Cancer Discovery 11, 1176–1191. https://doi.org/10.1158/2159-8290.CD-20-0790

Cilona, M., Locatello, L.G., Novelli, L., Gallo, O., 2020. The Mismatch Repair System (MMR) in Head and Neck Carcinogenesis and Its Role in Modulating the Response to Immunotherapy: A Critical Review. Cancers 12, 3006. https://doi.org/10.3390/cancers12103006

Coenye, T., Vandamme, P., 2005. Characterization of Mononucleotide Repeats in Sequenced Prokaryotic Genomes. DNA Research 12, 221–233. https://doi.org/10.1093/dnares/dsi009

Cohen, D., Chumakov, I., Weissenbach, J., 1993. A first-generation physical map of the human genome. Nature 366, 698–701. https://doi.org/10.1038/366698a0

Collura, A., Lagrange, A., Svrcek, M., Marisa, L., Buhard, O., Guilloux, A., Wanherdrick, K., Dorard, C., Taieb, A., Saget, A., Loh, M., Soong, R., Zeps, N., Platell, C., Mews, A., Iacopetta, B., De Thonel, A., Seigneuric, R., Marcion, G., Chapusot, C., Lepage, C., Bouvier, A.-M., Gaub, M.-P., Milano, G., Selves, J., Senet, P., Delarue, P., Arzouk, H., Lacoste, C., Coquelle, A., Bengrine-Lefèvre, L., Tournigand, C., Lefèvre, J.H., Parc, Y., Biard, D.S., Fléjou, J.-F., Garrido, C., Duval, A., 2014. Patients with colorectal tumors with microsatellite instability and large deletions in HSP110 T17 have improved response to 5-fluorouracil–based chemotherapy. Gastroenterology 146, 401-411.e1. https://doi.org/10.1053/j.gastro.2013.10.054

Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M., Park, P.J., 2017. A molecular portrait of microsatellite instability across multiple cancers. Nat Commun 8, 15180. https://doi.org/10.1038/ncomms15180

Crysup, B., Mandape, S., King, J.L., Muenzler, M., Kapema, K.B., Woerner, A.E., 2023. Using unique molecular identifiers to improve allele calling in low-template mixtures. Forensic Sci Int Genet 63, 102807. https://doi.org/10.1016/j.fsigen.2022.102807

Csink, A.K., Henikoff, S., 1998. Something from nothing: the evolution and utility of satellite repeats. Trends in Genetics 14, 200–204. https://doi.org/10.1016/S0168-9525(98)01444-9

Dantas, C.W.D., da Costa Neto, S.R., Alves, S.I.A., da Costa Pinheiro, K., De Los Santos, E.F.F., Ramos, R.T.J., 2024. SATIN: a micro and mini satellite mining tool of total genome and coding regions with analysis of perfect repeats polymorphism in coding regions. BMC Bioinformatics 25, 217. https://doi.org/10.1186/s12859-024-05842-2

Darwins, C., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, John Murray. ed.

Dashnow, H., Lek, M., Phipson, B., Halman, A., Sadedin, S., Lonsdale, A., Davis, M., Lamont, P., Clayton, J.S., Laing, N.G., MacArthur, D.G., Oshlack, A., 2018. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biology 19, 121. https://doi.org/10.1186/s13059-018-1505-2

Datsenko, K.A., Wanner, B.L., 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc. Natl. Acad. Sci. U.S.A. 97, 6640–6645. https://doi.org/10.1073/pnas.120163297

Daunay, A., Duval, A., Baudrin, L.G., Buhard, O., Renault, V., Deleuze, J.-F., How-Kit, A., 2019. Low temperature isothermal amplification of microsatellites drastically reduces stutter artifact formation and improves microsatellite instability detection in cancer. Nucleic Acids Research 47, e141–e141. https://doi.org/10.1093/nar/gkz811

Deamer, D., Akeson, M., Branton, D., 2016. Three decades of nanopore sequencing. Nat Biotechnol 34, 518–524. https://doi.org/10.1038/nbt.3423

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., Driscoll, M., Song, W., Kingsmore, S.F., Egholm, M., Lasken, R.S., 2002. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 99, 5261–5266. https://doi.org/10.1073/pnas.082089499

DePristo, M.A., Banks, E., Poplin, R.E., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery

and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491–498. https://doi.org/10.1038/ng.806

Derheimer, F.A., Kastan, M.B., 2010. Multiple roles of atm in monitoring and maintaining dna integrity. FEBS Lett 584, 3675–3681. https://doi.org/10.1016/j.febslet.2010.05.031

Descorps-Declère, S., Richard, G.-F., 2022. Megasatellite formation and evolution in vertebrate genes. Cell Reports 40. https://doi.org/10.1016/j.celrep.2022.111347

Devuyst, O., 2015. The 1000 Genomes Project: Welcome to a New World. Perit Dial Int 35, 676–677. https://doi.org/10.3747/pdi.2015.00261

Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petzl-Erler, M.L., Haines, G.K., Barch, D.H., 1998. Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 148, 1269–1284. https://doi.org/10.1093/genetics/148.3.1269

Di Rienzo, A., Peterson, A.C., Garza, J.C., Valdes, A.M., Slatkin, M., Freimer, N.B., 1994. Mutational processes of simple-sequence repeat loci in human populations. Proceedings of the National Academy of Sciences of the United States of America 91, 3166. https://doi.org/10.1073/pnas.91.8.3166

Diamond, A., Lye, C.T., Prasad, D., Abbott, D., 2021. One size does not fit all: Assuming the same normal body temperature for everyone is not justified. PLOS ONE 16, e0245257. https://doi.org/10.1371/journal.pone.0245257

Dolzhenko, E., English, A., Dashnow, H., De Sena Brandine, G., Mokveld, T., Rowell, W.J., Karniski, C., Kronenberg, Z., Danzi, M.C., Cheung, W.A., Bi, C., Farrow, E., Wenger, A., Chua, K.P., Martínez-Cerdeño, V., Bartley, T.D., Jin, P., Nelson, D.L., Zuchner, S., Pastinen, T., Quinlan, A.R., Sedlazeck, F.J., Eberle, M.A., 2024. Characterization and visualization of tandem repeats at genome scale. Nat Biotechnol 1–9. https://doi.org/10.1038/s41587-023-02057-3

Dorard, C., Thonel, A., Collura, A., Marisa, L., Svrcek, M., Lagrange, A., Jego, G., Wanherdrick, K., Joly, A.-L., Buhard, O., Gobbo, J., Penard-Lacronique, V., Zouali, H., Tubacher, E., Kirzin, S., Selves, J., Milano, G., Etienne-Grimaldi, M., Bengrine-Lefèvre, L., Duval, A., 2011. Expression of a mutant HSP110 sensitizes colorectal cancer cells to chemotherapy and improves disease prognosis. Nature Medicine 17, 9. https://doi.org/10.1038/nm.2457

Dou, H., Yang, H., Feng, L., Mou, P., Wang, T., Ge, J., 2016. Estimating the Population Size and Genetic Diversity of Amur Tigers in Northeast China. PLoS One 11, e0154254. https://doi.org/10.1371/journal.pone.0154254

Dragileva, E., Hendricks, A., Teed, A., Gillis, T., Lopez, E.T., Friedberg, E.C., Kucherlapati, R., Edelmann, W., Lunetta, K.L., MacDonald, M.E., Wheeler, V.C., 2009. Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. Neurobiol Dis 33, 37–47. https://doi.org/10.1016/j.nbd.2008.09.014

Du, L., Liu, Q., Zhao, K., Tang, J., Zhang, X., Yue, B., Fan, Z., 2020. PSMD: An extensive database for pan-species microsatellite investigation and marker development. Molecular Ecology Resources 20, 283–291. https://doi.org/10.1111/1755-0998.13098

Duval, A., Hamelin, R., 2002. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer Res 62, 2447–2454.

Eckert, K.A., Hile, S.E., 2009. Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. Mol. Carcinog. 48, 379–388. https://doi.org/10.1002/mc.20499

Eckert, K.A., Hile, S.E., Vargo, P.L., 1997. Development and use of an in vitro HSV-tk forward mutation assay to study eukaryotic DNA polymerase processing of DNA alkyl lesions. Nucleic Acids Research 25, 1450. https://doi.org/10.1093/nar/25.7.1450

Eckert, K.A., Mowery, A., Hile, S.E., 2002. Misalignment-mediated DNA polymerase beta mutations: comparison of microsatellite and frame-shift error rates using a forward mutation assay. Biochemistry 41, 10490–10498. https://doi.org/10.1021/bi025918c

Eckert, K.A., Yan, G., 2000. Mutational analyses of dinucleotide and tetranucleotide microsatellites in Escherichia coli: influence of sequence on expansion mutagenesis. Nucleic Acids Res 28, 2831–2838.

Ehrlich, M., Wang, R.Y., 1981. 5-Methylcytosine in eukaryotic DNA. Science 212, 1350–1357. https://doi.org/10.1126/science.6262918

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323, 133–138. https://doi.org/10.1126/science.1162986

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5, 435–445. https://doi.org/10.1038/nrg1348

Estoup, A., Wilson, I., Sullivan, C., Cornuet, J., Moritz, C., 2002. Inferring population history from microsatellite and enzyme data in serially introduced cane toads, Bufo marinus. Genetics 159, 1671–87.

Evans, W.E., Parry, J.M., 1975. The genetic effects of elevated temperature in the yeast, Saccharomyces cerevisiae. Heredity (Edinb) 35, 347–359. https://doi.org/10.1038/hdy.1975.105

Evrard, Tachon, Randrian, Karayan-Tapon, Tougeron, 2019. Microsatellite Instability: Diagnosis, Heterogeneity, Discordance, and Clinical Impact in Colorectal Cancer. Cancers 11, 1567. https://doi.org/10.3390/cancers11101567

Excoffier, L., Laval, G., Schneider, S., 2007. Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evol Bioinform Online 1, 47–50.

Fang, L., Liu, Q., Monteys, A.M., Gonzalez-Alegre, P., Davidson, B.L., Wang, K., 2022. DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. Genome Biol 23, 108. https://doi.org/10.1186/s13059-022-02670-6

Felton, K., Gilchrist, D., Andrew, S., 2007. Constitutive deficiency in DNA mismatch repair: is it time for Lynch III? Clinical Genetics 71, 499–500. https://doi.org/10.1111/j.1399-0004.2007.00801.x

Fernández-Peña, C., Reimúndez, A., Viana, F., Arce, V.M., Señarís, R., 2023. Sex differences in thermoregulation in mammals: Implications for energy homeostasis. Front. Endocrinol. 14. https://doi.org/10.3389/fendo.2023.1093376

Fernando, P., Polet, G., Foead, N., Ng, L., Pastorini, J., Melnick, D., 2006. Genetic Diversity, phylogeny and conservation of the Javan rhinoceros (Rhinoceros sondaicus). Conservation Genetics 7, 439–448. https://doi.org/10.1007/s10592-006-9139-4

Fonseca, N.A., Rung, J., Brazma, A., Marioni, J.C., 2012. Tools for mapping high-throughput sequencing data. Bioinformatics 28, 3169–3177. https://doi.org/10.1093/bioinformatics/bts605

Frith, M.C., Khan, S., 2018. A survey of localized sequence rearrangements in human DNA. Nucleic Acids Res 46, 1661–1673. https://doi.org/10.1093/nar/gkx1266

Fu, Y.X., Chakraborty, R., 1998. Simultaneous estimation of all the parameters of a stepwise mutation model. Genetics 150, 487–497. https://doi.org/10.1093/genetics/150.1.487

Fungtammasan, A., Ananda, G., Hile, S.E., Su, M.S.-W., Sun, C., Harris, R., Medvedev, P., Eckert, K., Makova, K.D., 2015. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 25, 736–749. https://doi.org/10.1101/gr.185892.114

Gallinger, S., Aronson, M., Shayan, K., Ratcliffe, E.M., Gerstle, J.T., Parkin, P.C., Rothenmund, H., Croitoru, M., Baumann, E., Durie, P.R., Weksberg, R., Pollett, A., Riddell, R.H., Ngan, B.Y., Cutz, E., Lagarde, A.E., Chan, H.S.L., 2004. Gastrointestinal cancers and neurofibromatosis type 1 features in children with a germline homozygous MLH1 mutation. Gastroenterology 126, 576–585. https://doi.org/10.1053/j.gastro.2003.11.008

Gallon, R., Mühlegger, B., Wenzel, S.-S., Sheth, H., Hayes, C., Aretz, S., Dahan, K., Foulkes, W., Kratz, C.P., Ripperger, T., Azizi, A.A., Baris Feldman, H., Chong, A.-L., Demirsoy, U., Florkin, B., Imschweiler,

T., Januszkiewicz-Lewandowska, D., Lobitz, S., Nathrath, M., Pander, H.-J., Perez-Alonso, V., Perne, C., Ragab, I., Rosenbaum, T., Rueda, D., Seidel, M.G., Suerink, M., Taeubner, J., Zimmermann, S.-Y., Zschocke, J., Borthwick, G.M., Burn, J., Jackson, M.S., Santibanez-Koref, M., Wimmer, K., 2019. A sensitive and scalable microsatellite instability assay to diagnose constitutional mismatch repair deficiency by sequencing of peripheral blood leukocytes. Hum Mutat 40, 649–655. https://doi.org/10.1002/humu.23721

Gallon, R., Sheth, H., Hayes, C., Redford, L., Alhilal, G., O'Brien, O., Spiewak, H., Waltham, A., McAnulty, C., Izuogu, O.G., Arends, M.J., Oniscu, A., Alonso, A.M., Laguna, S.M., Borthwick, G.M., Santibanez-Koref, M., Jackson, M.S., Burn, J., 2020. Sequencing-based microsatellite instability testing using as few as six markers for high-throughput clinical diagnostics. Hum Mutat 41, 332–341. https://doi.org/10.1002/humu.23906

Gan, C., Love, C., Beshay, V., Macrae, F., Fox, S., Waring, P., Taylor, G., 2015. Applicability of Next Generation Sequencing Technology in Microsatellite Instability Testing. Genes (Basel) 6, 46–59. https://doi.org/10.3390/genes6010046

Ganschow, S., Silvery, J., Kalinowski, J., Tiemann, C., 2018. toaSTR: A web application for forensic STR genotyping by massively parallel sequencing. Forensic Science International: Genetics 37, 21–28. https://doi.org/10.1016/j.fsigen.2018.07.006

Gelfand, Y., Rodriguez, A., Benson, G., 2007. TRDB—The Tandem Repeats Database. Nucleic Acids Res 35, D80–D87. https://doi.org/10.1093/nar/gkl1013

Gemayel, R., Vinces, M.D., Legendre, M., Verstrepen, K.J., 2010. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. Annu. Rev. Genet. 44, 445–477. https://doi.org/10.1146/annurev-genet-072610-155046

Geneva, I.I., Cuzzo, B., Fazili, T., Javaid, W., 2019. Normal Body Temperature: A Systematic Review. Open Forum Infectious Diseases 6, ofz032. https://doi.org/10.1093/ofid/ofz032

Georgiadis, A., Durham, J.N., Keefer, L.A., Bartlett, B.R., Zielonka, M., Murphy, D., White, J.R., Lu, S., Verner, E.L., Ruan, F., Riley, D., Anders, R.A., Gedvilaite, E., Angiuoli, S., Jones, S., Velculescu, V.E., Le, D.T., Diaz, L.A., Sausen, M., 2019. Noninvasive Detection of Microsatellite Instability and High Tumor Mutation Burden in Cancer Patients Treated with PD-1 Blockade. Clin Cancer Res 25, 7024–7034. https://doi.org/10.1158/1078-0432.CCR-19-1372

Gettings, K.B., Aponte, R.A., Vallone, P.M., Butler, J.M., 2015. STR allele sequence variation: Current knowledge and future issues. Forensic Sci Int Genet 18, 118–130. https://doi.org/10.1016/j.fsigen.2015.06.005

Girod, C., Vitalis, R., Leblois, R., Fréville, H., 2011. Inferring Population Decline and Expansion From Microsatellite Data: A Simulation-Based Evaluation of the Msvar Method. Genetics 188, 165–179. https://doi.org/10.1534/genetics.110.121764

Glenn, T.C., Stephan, W., Dessauer, H.C., Braun, M.J., 1996. Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. Mol Biol Evol 13, 1151–1154. https://doi.org/10.1093/oxfordjournals.molbev.a025678

Goldstein, D.B., Clark, A.G., 1995. Microsatellite variation in North American populations of Drosophila melanogaster. Nucleic Acids Res 23, 3882–3886.

Goldstein, D.B., Schlotterer, C. (Eds.), 1999. Microsatellites: Evolution and Applications. Oxford University Press, Oxford, New York.

Goldstein, D.B., Schlötterer, C., 1999. Microsatellites: Evolution and Applications. Genet. Res. 75, 123–125. https://doi.org/10.1017/S0016672300239141

Gugala, N.A., Ishida, Y., Georgiadis, N.J., Roca, A.L., 2016. Development and characterization of microsatellite markers in the African forest elephant (Loxodonta cyclotis). BMC Res Notes 9, 364. https://doi.org/10.1186/s13104-016-2167-3

Gulcher, J., 2012. Microsatellite markers for linkage and association studies. Cold Spring Harb Protoc 2012, 425–432. https://doi.org/10.1101/pdb.top068510

Gupta, P., Varshney, R.K., 2000. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat | Request PDF. Euphytica 113. https://doi.org/10.1023/A:1003910819967

Gusella, J.F., 1984. Genetic linkage of the Huntington's disease gene to a DNA marker. Can J Neurol Sci 11, 421–425. https://doi.org/10.1017/s0317167100045947

Gymrek, M., Golan, D., Rosset, S., Erlich, Y., 2012. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res 22, 1154–1162. https://doi.org/10.1101/gr.135780.111

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., Erlich, Y., 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet 48, 22–29. https://doi.org/10.1038/ng.3461

Haller, B.C., Messer, P.W., 2017. SLiM 2: Flexible, Interactive Forward Genetic Simulations. Molecular Biology and Evolution 34, 230–240. https://doi.org/10.1093/molbev/msw211

Hampel, H., Frankel, W.L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., Nakagawa, H., Sotamaa, K., Prior, T.W., Westman, J., Panescu, J., Fix, D., Lockman, J., Comeras, I., de la Chapelle, A., 2005. Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). N Engl J Med 352, 1851–1860. https://doi.org/10.1056/NEJMoa043146

Harfe, B.D., Jinks-Robertson, S., 2000. DNA mismatch repair and genetic instability. Annu Rev Genet 34, 359–399. https://doi.org/10.1146/annurev.genet.34.1.359

Hashem, V.I., Rosche, W.A., Sinden, R.R., 2002. Genetic assays for measuring rates of (CAG).(CTG) repeat instability in Escherichia coli. Mutat Res 502, 25–37. https://doi.org/10.1016/s0027-5107(02)00026-x

Hause, R.J., Pritchard, C.C., Shendure, J., Salipante, S.J., 2016. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med 22, 1342–1350. https://doi.org/10.1038/nm.4191

Henderson, S.T., Petes, T.D., 1992. Instability of simple sequence DNA in Saccharomyces cerevisiae. Mol Cell Biol 12, 2749–2757.

Herbreteau, G., Airaud, F., Pierre-Noël, E., Vallée, A., Bézieau, S., Théoleyre, S., Blons, H., Garinet, S., Denis, M.G., 2021. MEM: An Algorithm for the Reliable Detection of Microsatellite Instability (MSI) on a Small NGS Panel in Colorectal Cancer. Cancers (Basel) 13, 4203. https://doi.org/10.3390/cancers13164203

Hiatt, J.B., Pritchard, C.C., Salipante, S.J., O'Roak, B.J., Shendure, J., 2013. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res 23, 843–854. https://doi.org/10.1101/gr.147686.112

Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A., Mittelman, D., 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res 41, e32. https://doi.org/10.1093/nar/gks981

Holland, M.M., Parson, W., 2011. GeneMarker® HID: A Reliable Software Tool for the Analysis of Forensic STR Data. Journal of Forensic Sciences 56, 29–35. https://doi.org/10.1111/j.1556-4029.2010.01565.x

Hong, E.S., Yeung, A., Funchain, P., Slupska, M.M., Miller, J.H., 2005. Mutants with temperature-sensitive defects in the Escherichia coli mismatch repair system: sensitivity to mispairs generated in vivo. J Bacteriol 187, 840–846. https://doi.org/10.1128/JB.187.3.840-846.2005

How-Kit, A., Ye, K., 2023. Editorial: Microsatellite and microsatellite instability. Frontiers in Genetics 14, 1189212. https://doi.org/10.3389/fgene.2023.1189212

Hu, L.J., Laporte, J., Kress, W., Dahl, N., 1996. Prenatal diagnosis of X-linked myotubular myopathy: strategies using new and tightly linked DNA markers. Prenat Diagn 16, 231–237. https://doi.org/10.1002/(SICI)1097-0223(199603)16:3<231::AID-PD842>3.0.CO;2-M

Huang, C.-J., Lu, M.-Y., Chang, Y.-W., Li, W.-H., 2018. Experimental Evolution of Yeast for High-Temperature Tolerance. Molecular Biology and Evolution 35, 1823–1839. https://doi.org/10.1093/molbev/msy077

Huang, Q.-Y., Xu, F.-H., Shen, H., Deng, H.-Y., Liu, Y.-J., Liu, Y.-Z., Li, J.-L., Recker, R.R., Deng, H.-W., 2002. Mutation Patterns at Dinucleotide Microsatellite Loci in Humans. Am J Hum Genet 70, 625–634.

Huang, Y., Li, G.-M., 2020. DNA mismatch repair in the context of chromatin. Cell & Bioscience 10, 10. https://doi.org/10.1186/s13578-020-0379-7

Huguet, J.M., Bizarro, C.V., Forns, N., Smith, S.B., Bustamante, C., Ritort, F., 2010. Single-molecule derivation of salt dependent base-pair free energies in DNA. Proc Natl Acad Sci U S A 107, 15431–15436. https://doi.org/10.1073/pnas.1001454107

Iizuka, R., Yamazaki, H., Uemura, S., 2022. Zero-mode waveguides and nanopore-based sequencing technologies accelerate single-molecule studies. Biophys Physicobiol 19, e190032. https://doi.org/10.2142/biophysico.bppb-v19.0032

Imai, K., Yamamoto, H., 2008. Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. Carcinogenesis 29, 673–680. https://doi.org/10.1093/carcin/bgm228

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945. https://doi.org/10.1038/nature03001

Iossifov, I., Zheng, T., Baron, M., Gilliam, T.C., Rzhetsky, A., 2008. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. Genome Res 18, 1150–1162. https://doi.org/10.1101/gr.075622.107

Iyer, R.R., Pluciennik, A., Burdett, V., Modrich, P.L., 2006. DNA mismatch repair: functions and mechanisms. Chem Rev 106, 302–323. https://doi.org/10.1021/cr0404794

Iyer, R.R., Pluciennik, A., Napierala, M., Wells, R.D., 2015. DNA Triplet Repeat Expansion and Mismatch Repair. Annu Rev Biochem 84, 199–226. https://doi.org/10.1146/annurev-biochem-060614-034010

Jain, M., Olsen, H.E., Paten, B., Akeson, M., 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17, 239. https://doi.org/10.1186/s13059-016-1103-0

Jam, H.Z., Zook, J.M., Javadzadeh, S., Park, J., Sehgal, A., Gymrek, M., 2024. LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. Genome Biol 25, 176. https://doi.org/10.1186/s13059-024-03319-2

Jeanjean, S.I., Renault, V., Daunay, A., Shen, Y., Hardy, L.M., Deleuze, J.-F., How-Kit, A., 2023. LT-RPA: An Isothermal DNA Amplification Approach for Improved Microsatellite Genotyping and Microsatellite Instability Detection. Methods Mol Biol 2621, 91–109. https://doi.org/10.1007/978-1-0716-2950-5_7

Jeffreys, A.J., 1987. Highly variable minisatellites and DNA fingerprints. Biochemical Society Transactions 15, 309–317. https://doi.org/10.1042/bst0150309

Jeffreys, A.J., Tamaki, K., MacLeod, A., Monckton, D.G., Neil, D.L., Armour, J.A.L., 1994. Complex gene conversion events in germline mutation at human minisatellites. Nat Genet 6, 136–145. https://doi.org/10.1038/ng0294-136

Jeffreys, A. J., Wilson, V., Thein, S.L., 1985. Individual-specific 'fingerprints' of human DNA. Nature 316, 76–79. https://doi.org/10.1038/316076a0

Jeffreys, Alec J., Wilson, V., Thein, S.L., 1985. Hypervariable "minisatellite" regions in Human DNA. Nature 314, 67–73.

Jiménez, A., Holderegger, R., Csencsics, D., Quintanilla, L.G., 2010. Microsatellites reveal substantial among-population genetic differentiation and strong inbreeding in the relict fern Dryopteris aemula. Ann Bot 106, 149–155. https://doi.org/10.1093/aob/mcq094

Jin, Z., Sinicrope, F.A., 2021. Prognostic and Predictive Values of Mismatch Repair Deficiency in Non-Metastatic Colorectal Cancer. Cancers (Basel) 13, 300. https://doi.org/10.3390/cancers13020300

John, B., Gabor Miklos, G.L., 1979. Functional Aspects of Satellite DNA and Heterochromatin, in: International Review of Cytology. Elsevier, pp. 1–114. https://doi.org/10.1016/S0074-7696(08)61473-4

Jonchere, V., Marisa, L., Greene, M., Virouleau, A., Buhard, O., Bertrand, R., Svrcek, M., Cervera, P., Goloudina, A., Guillerm, E., Coulet, F., Landman, S., Ratovomanana, T., Job, S., Ayadi, M., Elarouci, N., Armenoult, L., Merabtene, F., Dumont, S., Parc, Y., Lefèvre, J.H., André, T., Fléjou, J.-F., Guilloux, A., Collura, A., de Reyniès, A., Duval, A., 2018. Identification of Positively and Negatively Selected Driver Gene Mutations Associated With Colorectal Cancer With Microsatellite Instability. Cell Mol Gastroenterol Hepatol 6, 277–300. https://doi.org/10.1016/j.jcmgh.2018.06.002

Jonchère, V., Montémont, H., Le Scanf, E., Siret, A., Letourneur, Q., Tubacher, E., Battail, C., Fall, A., Labreche, K., Renault, V., Ratovomanana, T., Buhard, O., Jolly, A., Le Rouzic, P., Feys, C., Despras, E., Zouali, H., Nicolle, R., Cervera, P., Svrcek, M., Bourgoin, P., Blanché, H., Boland, A., Lefèvre, J., Parc, Y., Touat, M., Bielle, F., Arzur, D., Cueff, G., Le Jossic-Corcos, C., Quéré, G., Dujardin, G., Blondel, M., Le Maréchal, C., Cohen, R., André, T., Coulet, F., de la Grange, P., de Reyniès, A., Fléjou, J.-F., Renaud, F., Alentorn, A., Corcos, L., Deleuze, J.-F., Collura, A., Duval, A., 2024. Microsatellite instability at U2AF-binding polypyrimidic tract sites perturbs alternative splicing during colorectal cancer initiation. Genome Biology 25, 210. https://doi.org/10.1186/s13059-024-03340-5

Kadyrov, F.A., Dzantiev, L., Constantin, N., Modrich, P., 2006. Endonucleolytic function of MutLalpha in human mismatch repair. Cell 126, 297–308. https://doi.org/10.1016/j.cell.2006.05.039

Karssene, Y., Nowak, C., Chammem, M., Cocchiararo, B., Nouira, S., 2019. Genetic diversity of the genus *Vulpes* (Red fox and Fennec fox) in Tunisia based on mitochondrial DNA and noninvasive DNA sampling. Mammalian Biology 96, 118–123. https://doi.org/10.1016/j.mambio.2018.09.008

Katti, M.V., Ranjekar, P.K., Gupta, V.S., 2001. Differential Distribution of Simple Sequence Repeats in Eukaryotic Genome Sequences. Molecular Biology and Evolution 18, 1161–1167. https://doi.org/10.1093/oxfordjournals.molbev.a003903

Kautto, E.A., Bonneville, R., Miya, J., Yu, L., Krook, M.A., Reeser, J.W., Roychowdhury, S., 2016. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget 8, 7452–7463. https://doi.org/10.18632/oncotarget.13918

Keerti, A., Ninave, S., 2022. DNA Fingerprinting: Use of Autosomal Short Tandem Repeats in Forensic DNA Typing. Cureus 14, e30210. https://doi.org/10.7759/cureus.30210

Kelkar, Y.D., Eckert, K.A., Chiaromonte, F., Makova, K.D., 2011. A matter of life or death: How microsatellites emerge in and vanish from the human genome. Genome Res. 21, 2038–2048. https://doi.org/10.1101/gr.122937.111

Kelkar, Y.D., Strubczewski, N., Hile, S.E., Chiaromonte, F., Eckert, K.A., Makova, K.D., 2010. What Is a Microsatellite: A Computational and Experimental Definition Based upon Repeat Mutational Behavior at A/T and GT/AC Repeats. Genome Biology and Evolution 2, 620–635. https://doi.org/10.1093/gbe/evq046

Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., Makova, K.D., 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. 18, 30–38. https://doi.org/10.1101/gr.7113408

Kennedy, S.R., Schmitt, M.W., Fox, E.J., Kohrn, B.F., Salk, J.J., Ahn, E.H., Prindle, M.J., Kuong, K.J., Shen, J.-C., Risques, R.-A., Loeb, L.A., 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc 9, 2586–2606. https://doi.org/10.1038/nprot.2014.170

Kim, T.-M., Laird, P.W., Park, P.J., 2013. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell 155, 858–868. https://doi.org/10.1016/j.cell.2013.10.015

Kimmel, M., Chakraborty, R., King, J.P., Bamshad, M., Watkins, W.S., Jorde, L.B., 1998. Signatures of population expansion in microsatellite repeat data. Genetics 148, 1921. https://doi.org/10.1093/genetics/148.4.1921

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. https://doi.org/10.1007/BF01731581

Kimura, M., Crow, J.F., 1964. The Number of Alleles That Can Be Maintained in a Finite Population. Genetics 49, 725. https://doi.org/10.1093/genetics/49.4.725

Kimura, M., Ohta, T., 1978. Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc Natl Acad Sci U S A 75, 2868–2872.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., 2011. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 108, 9530–9535. https://doi.org/10.1073/pnas.1105422108

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680. https://doi.org/10.1126/science.220.4598.671

Kleczkowska, H.E., Marra, G., Lettieri, T., Jiricny, J., 2001. hMSH3 and hMSH6 interact with PCNA and colocalize with it to replication foci. Genes Dev 15, 724–736. https://doi.org/10.1101/gad.191201

Kondelin, J., Gylfe, A.E., Lundgren, S., Tanskanen, T., Hamberg, J., Aavikko, M., Palin, K., Ristolainen, H., Katainen, R., Kaasinen, E., Taipale, M., Taipale, J., Renkonen-Sinisalo, L., Järvinen, H., Böhm, J., Mecklin, J.-P., Vahteristo, P., Tuupanen, S., Aaltonen, L.A., Pitkänen, E., 2017. Comprehensive Evaluation of Protein Coding Mononucleotide Microsatellites in Microsatellite-Unstable Colorectal Cancer. Cancer Res 77, 4078–4088. https://doi.org/10.1158/0008-5472.CAN-17-0682

Koren, S., Phillippy, A.M., 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current Opinion in Microbiology 23, 110–120. https://doi.org/10.1016/j.mib.2014.11.014

Kovalenko, M., Dragileva, E., St Claire, J., Gillis, T., Guide, J.R., New, J., Dong, H., Kucherlapati, R., Kucherlapati, M.H., Ehrlich, M.E., Lee, J.-M., Wheeler, V.C., 2012. Msh2 acts in medium-spiny striatal neurons as an enhancer of CAG instability and mutant huntingtin phenotypes in Huntington's disease knock-in mice. PLoS One 7, e44273. https://doi.org/10.1371/journal.pone.0044273

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., Turner, D.J., 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. Nat Methods 6, 291–295. https://doi.org/10.1038/nmeth.1311

Kozlowski, P., de Mezer, M., Krzyzosiak, W.J., 2010. Trinucleotide repeats in human genome and exome. Nucleic Acids Res 38, 4027–4039. https://doi.org/10.1093/nar/gkq127

Kroutil, L.C., Register, K., Bebenek, K., Kunkel, T.A., 1996. Exonucleolytic proofreading during replication of repetitive DNA. Biochemistry 35, 1046–1053. https://doi.org/10.1021/bi952178h

Krüger, S., Kinzel, M., Walldorf, C., Gottschling, S., Bier, A., Tinschert, S., von Stackelberg, A., Henn, W., Görgens, H., Boue, S., Kölble, K., Büttner, R., Schackert, H.K., 2008. Homozygous PMS2 germline mutations in two families with early-onset haematological malignancy, brain tumours, HNPCC-associated tumours, and signs of neurofibromatosis type 1. Eur J Hum Genet 16, 62–72. https://doi.org/10.1038/sj.ejhg.5201923

Kruglyak, S., Durrett, R.T., Schug, M.D., Aquadro, C.F., 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A 95, 10774–10778.

Kunkel, T.A., Erie, D.A., 2005. DNA mismatch repair. Annu Rev Biochem 74, 681–710. https://doi.org/10.1146/annurev.biochem.74.082803.133243

Lai, Y., Sun, F., 2003. The Relationship Between Microsatellite Slippage Mutation Rate and the Number of Repeat Units. Molecular Biology and Evolution 20, 2123–2131. https://doi.org/10.1093/molbev/msg228

Larrea, A.A., Lujan, S.A., Kunkel, T.A., 2010. SnapShot: DNA mismatch repair. Cell 141, 730.e1. https://doi.org/10.1016/j.cell.2010.05.002

Laugel, V., 2013. Cockayne syndrome: The expanding clinical and mutational spectrum. Mechanisms of Ageing and Development 134, 161–170. https://doi.org/10.1016/j.mad.2013.02.006

Le, D.T., Durham, J.N., Smith, K.N., Wang, H., Bartlett, B.R., Aulakh, L.K., Lu, S., Kemberling, H., Wilt, C., Luber, B.S., Wong, F., Azad, N.S., Rucki, A.A., Laheru, D., Donehower, R., Zaheer, A., Fisher, G.A., Crocenzi, T.S., Lee, J.J., Greten, T.F., Duffy, A.G., Ciombor, K.K., Eyring, A.D., Lam, B.H., Joe, A., Kang, S.P., Holdhoff, M., Danilova, L., Cope, L., Meyer, C., Zhou, S., Goldberg, R.M., Armstrong,

D.K., Bever, K.M., Fader, A.N., Taube, J., Housseau, F., Spetzler, D., Xiao, N., Pardoll, D.M., Papadopoulos, N., Kinzler, K.W., Eshleman, J.R., Vogelstein, B., Anders, R.A., Diaz, L.A., 2017. Mismatch-repair deficiency predicts response of solid tumors to PD-1 blockade. Science 357, 409–413. https://doi.org/10.1126/science.aan6733

Leclercq, S., Rivals, E., Jarne, P., 2010. DNA Slippage Occurs at Microsatellite Loci without Minimal Threshold Length in Humans: A Comparative Genomic Approach. Genome Biology and Evolution 2, 325–335. https://doi.org/10.1093/gbe/evq023

Lee, D., Ialicicco, M., Akkinepalli, H., Morreale, G., Liu, Y., Leung, H., Scippa, G.S., Greenland, A., Mackay, I., 2012. Genotyping SSR length variants by isothermal DNA amplification. Genome 55, 691–695. https://doi.org/10.1139/G2012-058

Leggett, R.M., Clark, M.D., 2017. A world of opportunities with nanopore sequencing. J Exp Bot 68, 5419–5429. https://doi.org/10.1093/jxb/erx289

Levinson, G., Gutman, G.A., 1987a. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Molecular Biology and Evolution. https://doi.org/10.1093/oxfordjournals.molbev.a040442

Levinson, G., Gutman, G.A., 1987b. High frequencies of short fameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escheychia coli K-12. Nucleic Acids Research 15, 5323–5338.

Li, G.-M., 2008. Mechanisms and functions of DNA mismatch repair. Cell Res 18, 85–98. https://doi.org/10.1038/cr.2007.115

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y.-C., Korol, A.B., Fahima, T., Beiles, A., Nevo, E., 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 11, 2453–2465. https://doi.org/10.1046/j.1365-294x.2002.01643.x

Li, Y.-C., Korol, A.B., Fahima, T., Nevo, E., 2004. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21, 991–1007. https://doi.org/10.1093/molbev/msh073

Lieberman, A.P., Fischbeck, K.H., 2000. Triplet repeat expansion in neuromuscular disease. Muscle Nerve 23, 843–850. https://doi.org/10.1002/(sici)1097-4598(200006)23:6<843::aid-mus2>3.0.co;2-8

Lim, K.M.X., Mahyuddin, A.P., Gosavi, A.T., Choolani, M., 2023. Genetics in prenatal diagnosis. Singapore Med J 64, 27–36. https://doi.org/10.4103/singaporemedj.SMJ-2021-433

Litt, M., Luty, J.A., 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44, 397–401.

Liu, Q., Zhang, P., Wang, D., Gu, W., Wang, K., 2017. Interrogating the “unsequenceable” genomic trinucleotide repeat disorders by long-read sequencing. Genome Medicine 9, 65. https://doi.org/10.1186/s13073-017-0456-7

Logsdon, G.A., Vollger, M.R., Eichler, E.E., 2020. Long-read human genome sequencing and its applications. Nat Rev Genet 21, 597–614. https://doi.org/10.1038/s41576-020-0236-x

Lynch, M., 2010. Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences 107, 961–968. https://doi.org/10.1073/pnas.0912629107

Lynch, M., 2007. The origins of genome architecture / [WWW Document].

Lyon, E., Laver, T., Yu, P., Jama, M., Young, K., Zoccoli, M., Marlowe, N., 2010. A simple, high-throughput assay for Fragile X expanded alleles using triple repeat primed PCR and capillary electrophoresis. J Mol Diagn 12, 505–511. https://doi.org/10.2353/jmoldx.2010.090229

MacDonald, M.E., Ambrose, C.M., Duyao, M.P., Myers, R.H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S.A., James, M., Groot, N., MacFarlane, H., Jenkins, B., Anderson, M.A., Wexler, N.S., Gusella, J.F., Bates, G.P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A.-M., Lehrach, H., Buckler, A.J., Church, D., Doucette-Stamm, L., O’Donovan, M.C., Riba-Ramirez, L., Shah, M., Stanton, V.P., Strobel, S.A.,

Draths, K.M., Wales, J.L., Dervan, P., Housman, D.E., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J.J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F.S., Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D., Harper, P.S., 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72, 971–983. https://doi.org/10.1016/0092-8674(93)90585-E

Maletzky, A., Kaiser, R., Mikulíček, P., 2009. Conservation Genetics of Crested Newt Species Triturus cristatus and T. carnifex within a Contact Zone in Central Europe: Impact of Interspecific Introgression and Gene Flow. Diversity 2. https://doi.org/10.3390/d2010028

MANEL, S., GUGERLI, F., THUILLER, W., ALVAREZ, N., LEGENDRE, P., HOLDEREGGER, R., GIELLY, L., TABERLET, P., 2012. Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. Mol Ecol 21, 3729–3738. https://doi.org/10.1111/j.1365-294X.2012.05656.x

Marshall, J.N., Lopez, A.I., Pfaff, A.L., Koks, S., Quinn, J.P., Bubb, V.J., 2021. Variable number tandem repeats – Their emerging role in sickness and health. Exp Biol Med (Maywood) 246, 1368–1376. https://doi.org/10.1177/15353702211003511

Mauger, F., Daunay, A., Deleuze, J.-F., Tost, J., How-Kit, A., 2016. Multiplexing of E-ice-COLD-PCR Assays for Mutation Detection and Identification. Clinical Chemistry 62, 1155–1158. https://doi.org/10.1373/clinchem.2016.258830

Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. Proc Natl Acad Sci U S A 74, 560–564.

McComish, B.J., Charleston, M.A., Parks, M., Baroni, C., Salvatore, M.C., Li, R., Zhang, G., Millar, C.D., Holland, B.R., Lambert, D.M., 2024. Ancient and Modern Genomes Reveal Microsatellites Maintain a Dynamic Equilibrium Through Deep Time. Genome Biology and Evolution 16, evae017. https://doi.org/10.1093/gbe/evae017

McGinnis, S., Madden, T.L., 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research 32, W20–W25. https://doi.org/10.1093/nar/gkh435

Medvedev, P., Stanciu, M., Brudno, M., 2009. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6, S13–S20. https://doi.org/10.1038/nmeth.1374

Mendel, G., 1866. Versuche über Pflanzen-Hybriden. Verhandlungen des Naturforschenden Vereins in Brünn 4, 3–47.

Messier, W., Li, S.-H., Stewart, C.-B., 1996. The birth of microsatellites. Nature 381, 483–483. https://doi.org/10.1038/381483a0

Miah, G., Rafii, M., Ismail, M., Puteh, A., Rahim, H., Islam, Kh., Latif, M., 2013. A Review of Microsatellite Markers and Their Applications in Rice Breeding Programs to Improve Blast Disease Resistance. IJMS 14, 22499–22528. https://doi.org/10.3390/ijms141122499

Mirkin, S.M., 2007. Expandable DNA repeats and human disease. Nature 447, 932–940. https://doi.org/10.1038/nature05977

Mitsuhashi, S., Frith, M.C., Mizuguchi, T., Miyatake, S., Toyota, T., Adachi, H., Oma, Y., Kino, Y., Mitsuhashi, H., Matsumoto, N., 2019. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biology 20, 58. https://doi.org/10.1186/s13059-019-1667-6

Miyatake, S., Koshimizu, E., Fujita, A., Doi, H., Okubo, M., Wada, T., Hamanaka, K., Ueda, N., Kishida, H., Minase, G., Matsuno, A., Kodaira, M., Ogata, K., Kato, R., Sugiyama, A., Sasaki, A., Miyama, T., Satoh, M., Uchiyama, Y., Tsuchida, N., Hamanoue, H., Misawa, K., Hayasaka, K., Sekijima, Y., Adachi, H., Yoshida, K., Tanaka, F., Mizuguchi, T., Matsumoto, N., 2022. Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. npj Genom. Med. 7, 1–15. https://doi.org/10.1038/s41525-022-00331-y

Montel, F., 2018. Séquençage de l'ADN par nanopores - Résultats et perspectives. Med Sci (Paris) 34, 161–165. https://doi.org/10.1051/medsci/20183402014

Morral, N., Nunes, V., Casals, T., Chillón, M., Giménez, J., Bertranpetit, J., Estivill, X., 1993. Microsatellite haplotypes for cystic fibrosis: mutation frameworks and evolutionary tracers. Hum Mol Genet 2, 1015–1022. https://doi.org/10.1093/hmg/2.7.1015

Mousavi, N., Shleizer-Burko, S., Yanicky, R., Gymrek, M., 2019. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res 47, e90. https://doi.org/10.1093/nar/gkz501

Müller, M., Leuschner, C., Weithmann, G., Weigel, R., Bat-Enerel, B., Steiner, W., Gailing, O., 2024. A genome-wide genetic association study reveals SNPs significantly associated with environmental variables and specific leaf area in European beech. Physiologia plantarum 176, e14334. https://doi.org/10.1111/ppl.14334

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., Erlich, H., 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51 Pt 1, 263–273. https://doi.org/10.1101/sqb.1986.051.01.032

Murray, V., Monchawin, C., England, P.R., 1993. The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. Nucleic Acids Res 21, 2395–2398.

Nachman, M.W., Crowell, S.L., 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297–304. https://doi.org/10.1093/genetics/156.1.297

Nadir, E., Margalit, H., Gallily, T., Ben-Sasson, S.A., 1996. Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. Proc Natl Acad Sci U S A 93, 6470–6475.

Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., McLellan, M.D., Wendl, M.C., Ding, L., 2014. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 30, 1015–1016. https://doi.org/10.1093/bioinformatics/btt755

Niu, W., Zhang, Y., Liu, H., Liang, N., Xu, L., Li, Y., Yao, W., Shi, W., Liu, Z., 2023. Single-Cell Profiling Uncovers the Roles of Endometrial Fibrosis and Microenvironmental Changes in Adenomyosis. J Inflamm Res 16, 1949–1965. https://doi.org/10.2147/JIR.S402734

Notomi, T., Okayama, H., Masubuchi, H., Yonekawa, T., Watanabe, K., Amino, N., Hase, T., 2000. Loop-mediated isothermal amplification of DNA. Nucleic Acids Res 28, e63.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S.J., Diekhans, M., Logsdon, G.A., Alonge, M., Antonarakis, S.E., Borchers, M., Bouffard, G.G., Brooks, S.Y., Caldas, G.V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L.G., Dishuck, P.C., Durbin, R., Dvorkina, T., Fiddes, I.T., Formenti, G., Fulton, R.S., Fungtammasan, A., Garrison, E., Grady, P.G.S., Graves-Lindsay, T.A., Hall, I.M., Hansen, N.F., Hartley, G.A., Haukness, M., Howe, K., Hunkapiller, M.W., Jain, C., Jain, M., Jarvis, E.D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V.V., Marschall, T., McCartney, A.M., McDaniel, J., Miller, D.E., Mullikin, J.C., Myers, E.W., Olson, N.D., Paten, B., Peluso, P., Pevzner, P.A., Porubsky, D., Potapova, T., Rogaev, E.I., Rosenfeld, J.A., Salzberg, S.L., Schneider, V.A., Sedlazeck, F.J., Shafin, K., Shew, C.J., Shumate, A., Sims, Y., Smit, A.F.A., Soto, D.C., Sović, I., Storer, J.M., Streets, A., Sullivan, B.A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B.P., Wenger, A., Wood, J.M.D., Xiao, C., Yan, S.M., Young, A.C., Zarate, S., Surti, U., McCoy, R.C., Dennis, M.Y., Alexandrov, I.A., Gerton, J.L., O'Neill, R.J., Timp, W., Zook, J.M., Schatz, M.C., Eichler, E.E., Miga, K.H., Phillippy, A.M., 2022. The complete sequence of a human genome. Science 376, 44–53. https://doi.org/10.1126/science.abj6987

Okazaki, T., Kornberg, A., 1964. ENZYMATIC SYNTHESIS OF DEOXYRIBONUCLEIC ACID. XV. PURIFICATION AND PROPERTIES OF A POLYMERASE FROM BACILLUS SUBTILIS. J Biol Chem 239, 259–268.

Orozco, D., 2023. TrueAllele® and STRmixTM: A Comparison of Two Probabilistic Genotyping Software Programs in Forensic DNA Profile Analysis. UC Davis.

Otten, A.D., Tapscott, S.J., 1995. Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure. Proceedings of the National Academy of Sciences of the United States of America 92, 5465. https://doi.org/10.1073/pnas.92.12.5465

Pearson, C.E., Nichol Edamura, K., Cleary, J.D., 2005. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 6, 729–742. https://doi.org/10.1038/nrg1689

Pei, J., Bao, P., Chu, M., Liang, C., Ding, X., Wang, H., Wu, X., Guo, X., Yan, P., 2018. Evaluation of 17 microsatellite markers for parentage testing and individual identification of domestic yak (Bos grunniens). PeerJ 6, e5946. https://doi.org/10.7717/peerj.5946

Peltomäki, P., Nyström, M., Mecklin, J.-P., Seppälä, T.T., 2023. Lynch Syndrome Genetics and Clinical Implications. Gastroenterology 164, 783–799. https://doi.org/10.1053/j.gastro.2022.08.058

Perrier, C., Guyomard, R., Bagliniere, J.-L., Nikolic, N., Evanno, G., 2013. Changes in the genetic structure of Atlantic salmon populations over four decades reveal substantial impacts of stocking and potential resiliency. Ecol Evol 3, 2334–2349. https://doi.org/10.1002/ece3.629

Piepenburg, O., Williams, C.H., Stemple, D.L., Armes, N.A., 2006. DNA Detection Using Recombination Proteins. PLoS Biol 4, e204. https://doi.org/10.1371/journal.pbio.0040204

Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., Egholm, M., Rothberg, J.M., Leamon, J.H., 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7, 216. https://doi.org/10.1186/1471-2164-7-216

Pinto, R.M., Dragileva, E., Kirby, A., Lloret, A., Lopez, E., Claire, J.S., Panigrahi, G.B., Hou, C., Holloway, K., Gillis, T., Guide, J.R., Cohen, P.E., Li, G.-M., Pearson, C.E., Daly, M.J., Wheeler, V.C., 2013. Mismatch Repair Genes Mlh1 and Mlh3 Modify CAG Instability in Huntington's Disease Mice: Genome-Wide and Candidate Approaches. PLOS Genetics 9, e1003930. https://doi.org/10.1371/journal.pgen.1003930

Piotti, A., Leonardi, S., Heuertz, M., Buiteveld, J., Geburek, T., Gerber, S., Kramer, K., Vettori, C., Vendramin, G.G., 2013. Within-Population Genetic Structure in Beech (Fagus sylvatica L.) Stands Characterized by Different Disturbance Histories: Does Forest Management Simplify Population Substructure? PLoS One 8, e73391. https://doi.org/10.1371/journal.pone.0073391

Podlutsky, A., Osterholm, A.M., Hou, S.M., Hofmaier, A., Lambert, B., 1998. Spectrum of point mutations in the coding region of the hypoxanthine-guanine phosphoribosyltransferase (hprt) gene in human T-lymphocytes in vivo. Carcinogenesis 19, 557–566. https://doi.org/10.1093/carcin/19.4.557

Polosina, Y.Y., Cupples, C.G., 2010. MutL: conducting the cell's response to mismatched and misaligned DNA. BioEssays 32, 51–59. https://doi.org/10.1002/bies.200900089

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., Lareu, M., 2013. An overview of STRUCTURE: applications, parameter settings, and supporting software. Front. Genet. 4. https://doi.org/10.3389/fgene.2013.00098

Poulos, R.C., Olivier, J., Wong, J.W.H., 2017. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. Nucleic Acids Res 45, 7786–7795. https://doi.org/10.1093/nar/gkx463

Pupko, T., Graur, D., 1999. Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units. J Mol Evol 48, 313–316. https://doi.org/10.1007/pl00006474

Putman, A.I., Carbone, I., 2014. Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecol Evol 4, 4399–4428. https://doi.org/10.1002/ece3.1305

Qiu, R., Sakato, M., Sacho, E.J., Wilkins, H., Zhang, X., Modrich, P., Hingorani, M.M., Erie, D.A., Weninger, K.R., 2015. MutL traps MutS at a DNA mismatch. Proc Natl Acad Sci U S A 112, 10914–10919. https://doi.org/10.1073/pnas.1505655112

Radko, A., Podbielska, A., 2021. Microsatellite DNA Analysis of Genetic Diversity and Parentage Testing in the Popular Dog Breeds in Poland. Genes (Basel) 12, 485. https://doi.org/10.3390/genes12040485

Raevaara, T.E., Gerdes, A.-M., Lönnqvist, K.E., Tybjærg-Hansen, A., Abdel-Rahman, W.M., Kariola, R., Peltomäki, P., Nyström-Lahti, M., 2004. HNPCC mutation MLH1 P648S makes the functional protein unstable, and homozygosity predisposes to mild neurofibromatosis type 1. Genes, Chromosomes and Cancer 40, 261–265. https://doi.org/10.1002/gcc.20040

Rajan-Babu, I.-S., Lian, M., Cheah, F.S.H., Chen, M., Tan, A.S.C., Prasath, E.B., Loh, S.F., Chong, S.S., 2017. FMR1 CGG repeat expansion mutation detection and linked haplotype analysis for reliable and

accurate preimplantation genetic diagnosis of fragile X syndrome. Expert Rev Mol Med 19, e10. https://doi.org/10.1017/erm.2017.10

Ramel, C., 1997. Mini- and Microsatellites. Environmental Health Perspectives 105.

Raz, O., Biezuner, T., Spiro, A., Amir, S., Milo, L., Titelman, A., Onn, A., Chapal-Ilani, N., Tao, L., Marx, T., Feige, U., Shapiro, E., 2019. Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. Nucleic Acids Res 47, 2436–2445. https://doi.org/10.1093/nar/gky1318

Renault, V., Tubacher, E., How-Kit, A., 2022. Assessment of Microsatellite Instability from Next-Generation Sequencing Data, in: Laganà, A. (Ed.), Computational Methods for Precision Oncology. Springer International Publishing, Cham, pp. 75–100. https://doi.org/10.1007/978-3-030-91836-1_5

Reynes, L., 2021. Connectivité et structure génétiques des populations d'Ericaria zosteroides (Fucales) et Laminaria rodriguezii (Laminariales) des côtes françaises.

Rich, J., Ogryzko, V.V., Pirozhkova, I.V., 2014. Satellite DNA and related diseases. Biopolym. Cell 30, 249–259. https://doi.org/10.7124/bc.00089E

Richard, G.-F., Kerrest, A., Dujon, B., 2008. Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. Microbiol Mol Biol Rev 72, 686–727. https://doi.org/10.1128/MMBR.00011-08

Richard, G.F., Paques, F., 2000. Mini- and microsatellite expansions: the recombination connection. EMBO Rep 1, 122–126. https://doi.org/10.1093/embo-reports/kvd031

Rose, O., Falush, D., 1998. A threshold size for microsatellite expansion. Molecular Biology and Evolution 15, 613–615. https://doi.org/10.1093/oxfordjournals.molbev.a025964

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., Jaffe, D.B., 2013. Characterizing and measuring bias in sequence data. Genome Biol 14, R51. https://doi.org/10.1186/gb-2013-14-5-r51

Rossi, N.A., Menchaca-Rodriguez, A., Antelo, R., Wilson, B., McLaren, K., Mazzotti, F., Crespo, R., Wasilewski, J., Alda, F., Doadrio, I., Barros, T.R., Hekkala, E., Alonso-Tabet, M., Alonso-Giménez, Y., Lopez, M., Espinosa-Lopez, G., Burgess, J., Thorbjarnarson, J.B., Ginsberg, J.R., Vliet, K.A., Amato, G., 2020. High levels of population genetic differentiation in the American crocodile (Crocodylus acutus). PLoS One 15, e0235288. https://doi.org/10.1371/journal.pone.0235288

Rothblum-Oviatt, C., Wright, J., Lefton-Greif, M.A., McGrath-Morrow, S.A., Crawford, T.O., Lederman, H.M., 2016. Ataxia telangiectasia: a review. Orphanet J Rare Dis 11, 159. https://doi.org/10.1186/s13023-016-0543-7

Rudd, S., Mewes, H.-W., Mayer, K.F.X., 2003. Sputnik: a database platform for comparative plant genomics. Nucleic Acids Research 31, 128. https://doi.org/10.1093/nar/gkg075

Rutledge, L.Y., Coxon, A., White, B.N., 2017. Genetic assessment of the San Clemente Island loggerhead shrike reveals evidence of historical gene flow with Santa Catalina Island. Global Ecology and Conservation 11, 42–52. https://doi.org/10.1016/j.gecco.2017.04.002

Sacristán-Horcajada, E., González-de la Fuente, S., Peiró-Pastor, R., Carrasco-Ramiro, F., Amils, R., Requena, J.M., Berenguer, J., Aguado, B., 2021. ARAMIS: From systematic errors of NGS long reads to accurate assemblies. Brief Bioinform 22, bbab170. https://doi.org/10.1093/bib/bbab170

Sainudiin, R., Durrett, R.T., Aquadro, C.F., Nielsen, R., 2004. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. Genetics 168, 383–395. https://doi.org/10.1534/genetics.103.022665

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74, 5463–5467.

Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M.A., Gemmell, N., 2013. Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements. PLoS One 8, e54710. https://doi.org/10.1371/journal.pone.0054710

Schächter, F., Faure-Delanef, L., Guénot, F., Rouger, H., Froguel, P., Lesueur-Ginot, L., Cohen, D., 1994. Genetic associations with human longevity at the APOE and ACE loci. Nat Genet 6, 29–32. https://doi.org/10.1038/ng0194-29

Schlötterer, C., 2004. The evolution of molecular markers — just a matter of fashion? Nat Rev Genet 5, 63–69. https://doi.org/10.1038/nrg1249

Schlötterer, C., 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma 109, 365–371. https://doi.org/10.1007/s004120000089

Schlötterer, C., 1998. Genome evolution: Are microsatellites really simple sequences? Current Biology 8, R132–R134. https://doi.org/10.1016/S0960-9822(98)70989-3

Schlötterer, C., Ritter, R., Harr, B., Brem, G., 1998. High mutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates. Mol Biol Evol 15, 1269–1274. https://doi.org/10.1093/oxfordjournals.molbev.a025855

Schmidt-Küntzel, A., Dalton, D.L., Menotti-Raymond, M., Fabiano, E., Charruau, P., Johnson, W.E., Sommer, S., Marker, L., Kotzé, A., O'Brien, S.J., 2018. Conservation Genetics of the Cheetah: Genetic History and Implications for Conservation. Cheetahs: Biology and Conservation 71–92. https://doi.org/10.1016/B978-0-12-804088-1.00006-X

Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., Loeb, L.A., 2012. Detection of ultra-rare mutations by next-generation sequencing. Proc. Natl. Acad. Sci. U.S.A. 109, 14508–14513. https://doi.org/10.1073/pnas.1208715109

Schug, M.D., Hutter, C.M., Wetterstrand, K.A., Gaudette, M.S., Mackay, T.F., Aquadro, C.F., 1998a. The mutation rates of di-, tri- and tetranucleotide repeats in Drosophila melanogaster. Mol Biol Evol 15, 1751–1760. https://doi.org/10.1093/oxfordjournals.molbev.a025901

Schug, M.D., Wetterstrand, K.A., Gaudette, M.S., Lim, R.H., Hutter, C.M., Aquadro, C.F., 1998b. The distribution and frequency of microsatellite loci in Drosophila melanogaster. Mol Ecol 7, 57–70. https://doi.org/10.1046/j.1365-294x.1998.00304.x

Scott, R.H., Mansour, S., Pritchard-Jones, K., Kumar, D., MacSweeney, F., Rahman, N., 2007. Medulloblastoma, acute myelocytic leukemia and colonic carcinomas in a child with biallelic MSH6 mutations. Nat Clin Pract Oncol 4, 130–134. https://doi.org/10.1038/ncponc0719

Sehgal, A., Ziaei Jam, H., Shen, A., Gymrek, M., 2024. Genome-wide detection of somatic mosaicism at short tandem repeats. Bioinformatics 40, btae485. https://doi.org/10.1093/bioinformatics/btae485

Senter, L., Clendenning, M., Sotamaa, K., Hampel, H., Green, J., Potter, J.D., Lindblom, A., Lagerstedt, K., Thibodeau, S.N., Lindor, N.M., Young, J., Winship, I., Dowty, J.G., White, D.M., Hopper, J.L., Baglietto, L., Jenkins, M.A., de la Chapelle, A., 2008. The Clinical Phenotype of Lynch Syndrome Due to Germ-Line PMS2 Mutations. Gastroenterology 135, 419-428.e1. https://doi.org/10.1053/j.gastro.2008.04.026

Seo, S.B., Ge, J., King, J.L., Budowle, B., 2014. Reduction of stutter ratios in short tandem repeat loci typing of low copy number DNA samples. Forensic Science International: Genetics 8, 213–218. https://doi.org/10.1016/j.fsigen.2013.10.004

Shah, S.N., Hile, S.E., Eckert, K.A., 2010. Defective Mismatch Repair, Microsatellite Mutation Bias, and Variability in Clinical Cancer Phenotypes. Cancer Research 70, 431–435. https://doi.org/10.1158/0008-5472.CAN-09-3049

Shatskikh, A.S., Kotov, A.A., Adashev, V.E., Bazylev, S.S., Olenina, L.V., 2020. Functional Significance of Satellite DNAs: Insights From Drosophila. Front Cell Dev Biol 8, 312. https://doi.org/10.3389/fcell.2020.00312

Sheldon, B.L., Barker, J.S., 1964. THE EFFECT OF TEMPERATURE DURING DEVELOPMENT OF MUTATION IN DROSOPHILA MELANOGASTER. Mutat Res 106, 310–317. https://doi.org/10.1016/0027-5107(64)90010-7

Shell, S.S., Putnam, C.D., Kolodner, R.D., 2007. The N Terminus of Saccharomyces cerevisiae Msh6 Is an Unstructured Tether to PCNA. Mol Cell 26, 565–578. https://doi.org/10.1016/j.molcel.2007.04.024

Shinde, D., Lai, Y., Sun, F., Arnheim, N., 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Research 31, 974–980. https://doi.org/10.1093/nar/gkg178

Shriver, M.D., Jin, L., Boerwinkle, E., Deka, R., Ferrell, R.E., Chakraborty, R., 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. Molecular Biology and Evolution 12, 914–920. https://doi.org/10.1093/oxfordjournals.molbev.a040268

Sia, E.A., Jinks-Robertson, S., Petes, T.D., 1997a. Genetic control of microsatellite stability. Mutat Res 383, 61–70. https://doi.org/10.1016/s0921-8777(96)00046-8

Sia, E.A., Kokoska, R.J., Dominska, M., Greenwell, P., Petes, T.D., 1997b. Microsatellite Instability in Yeast: Dependence on Repeat Unit Size and DNA Mismatch Repair Genes. Molecular and Cellular Biology 17, 2851–2858. https://doi.org/10.1128/MCB.17.5.2851

Sloan, D.B., Broz, A.K., Sharbrough, J., Wu, Z., 2018. Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods. Trends Biotechnol 36, 729–740. https://doi.org/10.1016/j.tibtech.2018.02.009

Smit, A., Hubley, R., Green, P., 2013. RepeatMasker Open-4.0. [WWW Document]. URL https://www.repeatmasker.org/ (accessed 9.3.24).

Smith, K.N., Nicolas, A., 1998. Recombination at work for meiosis. Curr Opin Genet Dev 8, 200–211. https://doi.org/10.1016/s0959-437x(98)80142-1

Solignac, M., Vautrin, D., Baudry, E., Mougel, F., Loiseau, A., Cornuet, J.-M., 2004. A microsatellite-based linkage map of the honeybee, Apis mellifera L. Genetics 167, 253–262.

Song, Q.J., Shi, J.R., Singh, S., Fickus, E.W., Costa, J.M., Lewis, J., Gill, B.S., Ward, R., Cregan, P.B., 2005. Development and mapping of microsatellite (SSR) markers in wheat. Theor Appl Genet 110, 550–560. https://doi.org/10.1007/s00122-004-1871-x

Spampinato, C.P., Gomez, R.L., Galles, C., Lario, L.D., 2009. From bacteria to plants: a compendium of mismatch repair assays. Mutat Res 682, 110–128. https://doi.org/10.1016/j.mrrev.2009.07.001

Städele, V., Vigilant, L., 2016. Strategies for determining kinship in wild populations using genetic data. Ecol Evol 6, 6107–6120. https://doi.org/10.1002/ece3.2346

Stankiewicz, P., Lupski, J.R., 2002. Molecular-evolutionary mechanisms for genomic disorders. Curr Opin Genet Dev 12, 312–319. https://doi.org/10.1016/s0959-437x(02)00304-0

Steely, C.J., Watkins, W.S., Baird, L., Jorde, L.B., 2022. The mutational dynamics of short tandem repeats in large, multigenerational families. Genome Biol 23, 253. https://doi.org/10.1186/s13059-022-02818-4

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A.D., Cooper, D.N., 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133, 1–9. https://doi.org/10.1007/s00439-013-1358-4

Stoffel, E.M., Mangu, P.B., Gruber, S.B., Hamilton, S.R., Kalady, M.F., Lau, M.W.Y., Lu, K.H., Roach, N., Limburg, P.J., 2015. Hereditary Colorectal Cancer Syndromes: American Society of Clinical Oncology Clinical Practice Guideline Endorsement of the Familial Risk–Colorectal Cancer: European Society for Medical Oncology Clinical Practice Guidelines. J Clin Oncol 33, 209–217. https://doi.org/10.1200/JCO.2014.58.1322

Stoler, N., Nekrutenko, A., 2021. Sequencing error profiles of Illumina sequencing instruments. NAR Genom Bioinform 3, lqab019. https://doi.org/10.1093/nargab/lqab019

Strand, M., Prolla, T.A., Liskay, R.M., Petes, T.D., 1993. Destabilization of tracts of simple repetitive DNA in Yeast by mutations affecting DNA mismatch repair. Nature 365, 274–276.

Strauss, B.S., Sagher, D., Acharya, S., 1997. Role of proofreading and mismatch repair in maintaining the stability of nucleotide repeats in DNA. Nucleic Acids Res 25, 806–813. https://doi.org/10.1093/nar/25.4.806

Subramanian, S., Mishra, R.K., Singh, L., 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biology 4, R13. https://doi.org/10.1186/gb-2003-4-2-r13

Sulovari, A., Li, R., Audano, P.A., Porubsky, D., Vollger, M.R., Logsdon, G.A., Human Genome Structural Variation Consortium, Warren, W.C., Pollen, A.A., Chaisson, M.J.P., Eichler, E.E., Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., Fan, X., Wen, J., Handsaker, R.E., Fairley, S., Kronenberg, Z.N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A.M., Hastie, A.R., Antaki, D., Anantharaman, T., Audano, P.A., Brand, H., Cantsilieris, S., Cao, H., Cerveira, E., Chen, C., Chen, X., Chin, C.-S., Chong, Z., Chuang, N.T., Lambert, C.C., Church, D.M., Clarke, L., Farrell, A., Flores, J., Galeey, T., Gorkin, D.U., Gujral, M., Guryev, V., Heaton, W.H., Korlach, J., Kumar, S., Kwon, J.Y., Lam, E.T., Lee, J.E., Lee, J., Lee, W.-P., Lee, S.P., Li, S., Marks, P., Viaud-Martinez, K., Meiers, S., Munson, K.M., Navarro, F.C.P., Nelson, B.J., Nodzak, C., Noor, A., Kyriazopoulou-Panagiotopoulou, S., Pang, A.W.C., Qiu, Y., Rosanio, G., Ryan, M., Stütz, A., Spierings, D.C.J., Ward, A., Welch, A.E., Xiao, M., Xu, W., Zhang, C., Zhu, Q., Zheng-Bradley, X., Lowy, E., Yakneen, S., McCarroll, S., Jun, G., Ding, L., Koh, C.L., Ren, B., Flicek, P., Chen, K., Gerstein, M.B., Kwok, P.-Y., Lansdorp, P.M., Marth, G.T., Sebat, J., Shi, X., Bashir, A., Ye, K., Devine, S.E., Talkowski, M.E., Mills, R.E., Marschall, T., Korbel, J.O., Eichler, E.E., Lee, C., 2019. Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc. Natl. Acad. Sci. U.S.A. 116, 23243–23253. https://doi.org/10.1073/pnas.1912175116

Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., Stefansson, K., 2012. A direct characterization of human mutation based on microsatellites. Nat Genet 44, 1161–1165. https://doi.org/10.1038/ng.2398

Szövényi, P., Arroyo, K., Guggisberg, A., Conti, E., 2009. Effects of Pleistocene glaciations on the genetic structure of Saxifraga florulenta (Saxifragaceae), a rare endemic of the Maritime Alps. TAXON 58, 532–543. https://doi.org/10.1002/tax.582017

Tae, H., Kim, D.-Y., McCormick, J., Settlage, R.E., Garner, H.R., 2014. Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. Bioinformatics 30, 652–659. https://doi.org/10.1093/bioinformatics/btt595

Takuno, S., Miyagi, R., Onami, J., Takahashi-Kariyazono, S., Sato, A., Tichy, H., Nikaido, M., Aibara, M., Mizoiri, S., Mrosso, H.D.J., Mzighani, S.I., Okada, N., Terai, Y., 2019. Patterns of genomic differentiation between two Lake Victoria cichlid species, Haplochromis pyrrhocephalus and H. sp. 'macula.' BMC Evolutionary Biology 19, 68. https://doi.org/10.1186/s12862-019-1387-2

Tang, H., Kirkness, E.F., Lippert, C., Biggs, W.H., Fabani, M., Guzman, E., Ramakrishnan, S., Lavrenko, V., Kakaradov, B., Hou, C., Hicks, B., Heckerman, D., Och, F.J., Caskey, C.T., Venter, J.C., Telenti, A., 2017. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am J Hum Genet 101, 700–715. https://doi.org/10.1016/j.ajhg.2017.09.013

Tanudisastro, H.A., Deveson, I.W., Dashnow, H., MacArthur, D.G., 2024. Sequencing and characterizing short tandem repeats in the human genome. Nat Rev Genet 25, 460–475. https://doi.org/10.1038/s41576-024-00692-3

Tautz, D., 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research 17, 6463–6471. https://doi.org/10.1093/nar/17.16.6463

Tautz, D., Schlötterer, C., 1994. Simple sequences. Current Opinion in Genetics & Development 4, 832–837. https://doi.org/10.1016/0959-437X(94)90067-1

Taylor, J., Durkin, J., Breden, F., 1999. The death of a microsatellite: A phylogenetic perspective on microsatellite interruptions. Molecular biology and evolution 16, 567–72. https://doi.org/10.1093/oxfordjournals.molbev.a026138

Teeter, K.C., Payseur, B.A., Harris, L.W., Bakewell, M.A., Thibodeau, L.M., O'Brien, J.E., Krenz, J.G., Sans-Fuentes, M.A., Nachman, M.W., Tucker, P.K., 2008. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res 18, 67–76. https://doi.org/10.1101/gr.6757907

Tekaia, F., Dujon, B., Richard, G.-F., 2013. Detection and Characterization of Megasatellites in Orthologous and Nonorthologous Genes of 21 Fungal Genomes. Eukaryot Cell 12, 794–803. https://doi.org/10.1128/EC.00001-13

The Cancer Genome Atlas Network, Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., Kovar, C.L., Lewis, L.R., Morgan, M.B., Newsham, I.F., Reid, J.G., Santibanez, J.,

Shinbrot, E., Trevino, L.R., Wu, Y.-Q., Wang, M., Gunaratne, P., Donehower, L.A., Creighton, C.J., Wheeler, D.A., Gibbs, R.A., Lawrence, M.S., Voet, Douglas, Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E.S., Gabriel, S., Getz, G., Ding, L., Fulton, R.S., Koboldt, D.C., Wylie, T., Walker, J., Dooling, D.J., Fulton, L., Delehaunty, K.D., Fronick, C.C., Demeter, R., Mardis, E.R., Wilson, R.K., Chu, A., Chun, H.-J.E., Mungall, A.J., Pleasance, E., Gordon Robertson, A., Stoll, D., Balasundaram, M., Birol, I., Butterfield, Y.S.N., Chuah, E., Coope, R.J.N., Dhalla, N., Guin, R., Hirst, C., Hirst, M., Holt, R.A., Lee, D., Li, H.I., Mayo, M., Moore, R.A., Schein, J.E., Slobodan, J.R., Tam, A., Thiessen, N., Varhol, R., Zeng, T., Zhao, Y., Jones, S.J.M., Marra, M.A., Bass, A.J., Ramos, A.H., Saksena, G., Cherniack, A.D., Schumacher, S.E., Tabak, B., Carter, S.L., Pho, N.H., Nguyen, H., Onofrio, R.C., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Getz, G., Meyerson, M., Protopopov, A., Zhang, J., Hadjipanayis, A., Lee, E., Xi, R., Yang, L., Ren, X., Zhang, H., Sathiamoorthy, N., Shukla, S., Chen, P.-C., Haseley, P., Xiao, Y., Lee, S., Seidman, J., Chin, L., Park, P.J., Kucherlapati, R., Todd Auman, J., Hoadley, K.A., Du, Y., Wilkerson, M.D., Shi, Y., Liquori, C., Meng, S., Li, L., Turman, Y.J., Topal, M.D., Tan, D., Waring, S., Buda, E., Walsh, J., Jones, C.D., Mieczkowski, P.A., Singh, D., Wu, J., Gulabani, A., Dolina, P., Bodenheimer, T., Hoyle, A.P., Simons, J.V., Soloway, M., Mose, L.E., Jefferys, S.R., Balu, S., O'Connor, B.D., Prins, J.F., Chiang, D.Y., Neil Hayes, D., Perou, C.M., Hinoue, T., Weisenberger, D.J., Maglinte, D.T., Pan, F., Berman, B.P., Van Den Berg, D.J., Shen, H., Triche Jr, T., Baylin, S.B., Laird, P.W., Getz, G., Noble, M., Voet, Doug, Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Shukla, S., Lawrence, M.S., Zhou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Park, R.W., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P.J., Chin, L., Thorsson, V., Reynolds, S.M., Bernard, B., Kreisberg, R., Lin, J., Iype, L., Bressler, R., Erkkilä, T., Gundapuneni, M., Liu, Y., Norberg, A., Robinson, T., Yang, D., Zhang, W., Shmulevich, I., de Ronde, J.J., Schultz, N., Cerami, E., Ciriello, G., Goldberg, A.P., Gross, B., Jacobsen, A., Gao, J., Kaczkowski, B., Sinha, R., Arman Aksoy, B., Antipin, Y., Reva, B., Shen, R., Taylor, B.S., Chan, T.A., Ladanyi, M., Sander, C., Akbani, R., Zhang, N., Broom, B.M., Casasent, T., Unruh, A., Wakefield, C., Hamilton, S.R., Craig Cason, R., Baggerly, K.A., Weinstein, J.N., Haussler, D., Benz, C.C., Stuart, J.M., Benz, S.C., Zachary Sanborn, J., Vaske, C.J., Zhu, J., Szeto, C., Scott, G.K., Yau, C., Ng, S., Goldstein, T., Ellrott, K., Collisson, E., Cozen, A.E., Zerbino, D., Wilks, C., Craft, B., Spellman, P., Penny, R., Shelton, T., Hatfield, M., Morris, S., Yena, P., Shelton, C., Sherman, M., Paulauskis, J., Gastier-Foster, J.M., Bowen, J., Ramirez, N.C., Black, A., Pyatt, R., Wise, L., White, P., Bertagnolli, M., Brown, J., Chan, T.A., Chu, G.C., Czerwinski, C., Denstman, F., Dhir, R., Dörner, A., Fuchs, C.S., Guillem, J.G., Iacocca, M., Juhl, H., Kaufman, A., Kohl III, B., Van Le, X., Mariano, M.C., Medina, E.N., Meyers, M., Nash, G.M., Paty, P.B., Petrelli, N., Rabeno, B., Richards, W.G., Solit, D., Swanson, P., Temple, L., Tepper, J.E., Thorp, R., Vakiani, E., Weiser, M.R., Willis, J.E., Witkin, G., Zeng, Z., Zinner, M.J., The Cancer Genome Atlas Network, Genome Sequencing Center Baylor College of Medicine, Genome Sequencing Center Broad Institute, Genome Sequencing Center Washington University in St Louis, Genome Characterization Center BC Cancer Agency, Genome-Characterization Center Broad Institute, Genome-Characterization Center Brigham and Women's Hospital and Harvard Medical School, Genome-Characterization Center University of North Carolina, C.H., Genome-Characterization Centers University of Southern California and Johns Hopkins University, Genome Data Analysis Center Broad Institute, Genome Data Analysis Center Institute for Systems Biology, Genome Data Analysis Center Memorial Sloan-Kettering Cancer Center, Genome Data Analysis Center University of Texas MD Anderson Cancer Center, Genome Data Analysis Centers, U. of C., Santa Cruz and the Buck Institute, Biospecimen Core Resource International Genomics Consortium, Nationwide Children's Hospital Biospecimen Core Resource, Tissue source sites and disease working group, 2012. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337. https://doi.org/10.1038/nature11252

Thierry, A., Bouchier, C., Dujon, B., Richard, G.-F., 2008. Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in Candida glabrata. Nucleic Acids Res 36, 5970–5982. https://doi.org/10.1093/nar/gkn594

Tishkoff, D.X., Boerger, A.L., Bertrand, P., Filosi, N., Gaida, G.M., Kane, M.F., Kolodner, R.D., 1997. Identification and characterization of Saccharomyces cerevisiae EXO1, a gene encoding an exonuclease that interacts with MSH2. Proc Natl Acad Sci U S A 94, 7487–7492.

Tóth, G., Gáspári, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10, 967–981. https://doi.org/10.1101/gr.10.7.967

Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., Turner, S.W., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res 38, e159. https://doi.org/10.1093/nar/gkq543

Trivedi, S., 2006. Comparison of simple sequence repeats in 19 Archaea. Genet Mol Res 5, 741–772.

Truninger, K., Menigatti, M., Luz, J., Russell, A., Haider, R., Gebbers, J.-O., Bannwart, F., Yurtsever, H., Neuweiler, J., Riehle, H.-M., Cattaruzza, M.S., Heinimann, K., Schär, P., Jiricny, J., Marra, G., 2005. Immunohistochemical Analysis Reveals High Frequency of PMS2 Defects in Colorectal Cancer. Gastroenterology 128, 1160–1171. https://doi.org/10.1053/j.gastro.2005.01.056

Tumendemberel, O., Zedrosser, A., Proctor, M.F., Reynolds, H.V., Adams, J.R., Sullivan, J.M., Jacobs, S.J., Khorloojav, T., Tserenbataa, T., Batmunkh, M., Swenson, J.E., Waits, L.P., 2019. Phylogeography, genetic diversity, and connectivity of brown bear populations in Central Asia. PLoS One 14, e0220746. https://doi.org/10.1371/journal.pone.0220746

Umar, A., Buermeyer, A.B., Simon, J.A., Thomas, D.C., Clark, A.B., Liskay, R.M., Kunkel, T.A., 1996. Requirement for PCNA in DNA Mismatch Repair at a Step Preceding DNA Resynthesis. Cell 87, 65–73. https://doi.org/10.1016/S0092-8674(00)81323-9

Ummat, A., Bashir, A., 2014. Resolving complex tandem repeats with long reads. Bioinformatics 30, 3491–3498. https://doi.org/10.1093/bioinformatics/btu437

Urzi, F., Šprem, N., Potočnik, H., Sindičić, M., Konjević, D., Ćirović, D., Rezić, A., Duniš, L., Melovski, D., Buzan, E., 2021. Population genetic structure of European wildcats inhabiting the area between the Dinaric Alps and the Scardo-Pindic mountains. Sci Rep 11, 17984. https://doi.org/10.1038/s41598-021-97401-5

Usdin, K., House, N.C.M., Freudenreich, C.H., 2015. Repeat instability during DNA repair: Insights from model systems. Crit Rev Biochem Mol Biol 50, 142–167. https://doi.org/10.3109/10409238.2014.999192

Van Eldijk, T.J.B., Sheridan, E.A., Martin, G., Weissing, F.J., Kuipers, O.P., Van Doorn, G.S., 2024. Temperature dependence of the mutation rate towards antibiotic resistance. JAC Antimicrob Resist 6, dlae085. https://doi.org/10.1093/jacamr/dlae085

Vergnaud, G., Denoeud, F., 2000. Minisatellites: mutability and genome architecture. Genome Res 10, 899–907. https://doi.org/10.1101/gr.10.7.899

Vieira, M.L.C., Santini, L., Diniz, A.L., Munhoz, C. de F., 2016. Microsatellite markers: what they mean and why they are so useful. Genet Mol Biol 39, 312–328. https://doi.org/10.1590/1678-4685-GMB-2016-0027

Viguera, E., Canceill, D., Ehrlich, Sd., 2001. Replication slippage involves DNA polymerase pausing and dissociation. The EMBO Journal 20, 2587. https://doi.org/10.1093/emboj/20.10.2587

Vincent, M., Xu, Y., Kong, H., 2004. Helicase-dependent isothermal DNA amplification. EMBO Rep 5, 795–800. https://doi.org/10.1038/sj.embor.7400200

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23, 4407–4414.

Waalkes, A., Smith, N., Penewit, K., Hempelmann, J., Konnick, E.Q., Hause, R.J., Pritchard, C.C., Salipante, S.J., 2018. Accurate Pan-Cancer Molecular Diagnosis of Microsatellite Instability by Single-Molecule Molecular Inversion Probe Capture and High-Throughput Sequencing. Clinical Chemistry 64, 950–958. https://doi.org/10.1373/clinchem.2017.285981

Wach, A., Brachat, A., Pöhlmann, R., Philippsen, P., 1994. New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. Yeast 10, 1793–1808. https://doi.org/10.1002/yea.320101310

Walker, M.E., Gardner, J.M., Vystavelova, A., McBryde, C., de Barros Lopes, M., Jiranek, V., 2003. Application of the reuseable, KanMX selectable marker to industrial yeast: construction and evaluation of heterothallic wine strains of Saccharomyces cerevisiae, possessing minimal foreign DNA sequences. FEMS Yeast Research 4, 339–347. https://doi.org/10.1016/S1567-1356(03)00161-2

Walser, J.-C., Furano, A.V., 2010. The mutational spectrum of non-CpG DNA varies with CpG content. Genome Research 20, 875. https://doi.org/10.1101/gr.103283.109

Walter, E.J., Hanna-Jumma, S., Carraretto, M., Forni, L., 2016. The pathophysiological basis and consequences of fever. Crit Care 20, 200. https://doi.org/10.1186/s13054-016-1375-5

Wang, M., 2023. A Review on Forensic DNA Analysis.

Wang, R.Y., Gehrke, C.W., Ehrlich, M., 1980. Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. Nucleic Acids Res 8, 4777–4790.

Wang, T.T., Abelson, S., Zou, J., Li, T., Zhao, Z., Dick, J.E., Shlush, L.I., Pugh, T.J., Bratman, S.V., 2019. High efficiency error suppression for accurate detection of low-frequency variants. Nucleic Acids Res 47, e87. https://doi.org/10.1093/nar/gkz474

Wang, Yunhao, Zhao, Y., Bollas, A., Wang, Yuru, Au, K.F., 2021. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 39, 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Warshauer, D.H., Lin, D., Hari, K., Jain, R., Davis, C., Larue, B., King, J.L., Budowle, B., 2013. STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci Int Genet 7, 409–417. https://doi.org/10.1016/j.fsigen.2013.04.005

Watson, J.D., Crick, F.H.C., 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 171, 737–738.

Weber, James L., 1990. Human DNA polymorphisms and methods of analysis. Current Opinion in Biotechnology 1, 166–171. https://doi.org/10.1016/0958-1669(90)90026-H

Weber, J. L., 1990. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics 7, 524–530. https://doi.org/10.1016/0888-7543(90)90195-z

Weber, J.L., May, P.E., 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44, 388–396.

Weber, J.L., Wong, C., 1993. Mutation of human short tandem repeats. Hum Mol Genet 2, 1123–1128. https://doi.org/10.1093/hmg/2.8.1123

Webster, M.T., Smith, N.G.C., Ellegren, H., 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc Natl Acad Sci U S A 99, 8748–8753. https://doi.org/10.1073/pnas.122067599

Wei, L., Nakajima, S., Böhm, S., Bernstein, K.A., Shen, Z., Tsang, M., Levine, A.S., Lan, L., 2015. DNA damage during the G0/G1 phase triggers RNA-templated, Cockayne syndrome B-dependent homologous recombination. Proc Natl Acad Sci U S A 112, E3495–E3504. https://doi.org/10.1073/pnas.1507105112

Weinberger, C.S., Vianna, J.A., Faugeron, S., Marquet, P.A., 2021. Inferring the impact of past climate changes and hunting on the South American sea lion. Diversity and Distributions 27, 2479–2497. https://doi.org/10.1111/ddi.13421

Weissenbach, J., 1993a. A second generation linkage map of the human genome based on highly informative microsatellite loci. Gene 135, 275–278. https://doi.org/10.1016/0378-1119(93)90077-G

Weissenbach, J., 1993b. Microsatellite polymorphisms and the genetic linkage map of the human genome. Current Opinion in Genetics & Development 3, 414–417. https://doi.org/10.1016/0959-437X(93)90114-5

Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G., Sibly, R.M., 2003. Likelihood-based estimation of microsatellite mutation rates. Genetics 164, 781–787.

Wierdl, M., Dominska, M., Petes, T.D., 1997. Microsatellite Instability in Yeast: Dependence on the Length of the Microsatellite. Genetics 146, 769–779. https://doi.org/10.1093/genetics/146.3.769

Wijfjes, R.Y., Boesten, R., Becker, F.F.M., Theeuwen, T.P.J.M., Snoek, B.L., Mastoraki, M., Verheijen, J.J., Güvencli, N., Denkers, L.-A.M., Koornneef, M., Eeuwijk, F.A. van, Smit, S., Ridder, D. de, Aarts, M.G.M., 2024. Allelic variants confer Arabidopsis adaptation to small regional environmental differences. https://doi.org/10.1101/2023.09.18.558200

Wilder, J., Hollocher, H., 2001. Mobile Elements and the Genesis of Microsatellites in Dipterans. Molecular Biology and Evolution 18, 384–392. https://doi.org/10.1093/oxfordjournals.molbev.a003814

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., Erlich, Y., 2017. Genome-wide profiling of heritable and de novo STR variations. Nat Methods 14, 590–592. https://doi.org/10.1038/nmeth.4267

Willis, J., Lefterova, M.I., Artyomenko, A., Kasi, P.M., Nakamura, Y., Mody, K., Catenacci, D.V.T., Fakih, M., Barbacioru, C., Zhao, J., Sikora, M., Fairclough, S.R., Lee, H., Kim, K.-M., Kim, S.T., Kim, J., Gavino, D., Benavides, M., Peled, N., Nguyen, T., Cusnir, M., Eskander, R.N., Azzi, G., Yoshino, T., Banks, K.C., Raymond, V.M., Lanman, R.B., Chudova, D.I., Talasaz, A., Kopetz, S., Lee, J., Odegaard, J.I., 2019. Validation of Microsatellite Instability Detection Using a Comprehensive Plasma-Based Genotyping Panel. Clinical Cancer Research 25, 7035–7045. https://doi.org/10.1158/1078-0432.CCR-19-1324

Wilson Sayres, M.A., Makova, K.D., 2011. Genome analyses substantiate male mutation bias in many species. Bioessays 33, 938–945. https://doi.org/10.1002/bies.201100091

Wimmer, K., Etzler, J., 2008. Constitutional mismatch repair-deficiency syndrome: have we so far seen only the tip of an iceberg? Hum Genet 124, 105–122. https://doi.org/10.1007/s00439-008-0542-4

Wimmer, K., Kratz, C.P., Vasen, H.F.A., Caron, O., Colas, C., Entz-Werle, N., Gerdes, A.-M., Goldberg, Y., Ilencikova, D., Muleris, M., Duval, A., Lavoine, N., Ruiz-Ponte, C., Slavc, I., Burkhardt, B., Brugieres, L., on behalf of the EU-Consortium Care for CMMRD (C4CMMRD), 2014. Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium 'Care for CMMRD' (C4CMMRD). J Med Genet 51, 355–365. https://doi.org/10.1136/jmedgenet-2014-102284

Woerner, A.E., Mandape, S., King, J.L., Muenzler, M., Crysup, B., Budowle, B., 2021. Reducing noise and stutter in short tandem repeat loci with unique molecular identifiers. Forensic Sci Int Genet 51, 102459. https://doi.org/10.1016/j.fsigen.2020.102459

Wu, C.-H., Drummond, A.J., 2011. Joint Inference of Microsatellite Mutation Models, Population History and Genealogies Using Transdimensional Markov Chain Monte Carlo. Genetics 188, 151–164. https://doi.org/10.1534/genetics.110.125260

Wu, Q., Zang, F., Ma, Y., Zheng, Y., Zang, D., 2020. Analysis of genetic diversity and population structure in endangered *Populus wulianensis* based on 18 newly developed EST-SSR markers. Global Ecology and Conservation 24, e01329. https://doi.org/10.1016/j.gecco.2020.e01329

Xu, Xin, Peng, M., Fang, Z., Xu, Xiping, 2000. The direction of microsatellite mutations is dependent upon allele length. Nat Genet 24, 396–399. https://doi.org/10.1038/74238

Xue, Y., Braslavsky, I., Quake, S.R., 2021. Temperature effect on polymerase fidelity. Journal of Biological Chemistry 297. https://doi.org/10.1016/j.jbc.2021.101270

Yakovchuk, P., Protozanova, E., Frank-Kamenetskii, M.D., 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res 34, 564–574. https://doi.org/10.1093/nar/gkj454

Yamaguchi, I., Watanabe, T., Ohara, O., Hasegawa, Y., 2019. PCR-free whole exome sequencing: Cost-effective and efficient in detecting rare mutations. PLoS One 14, e0222562. https://doi.org/10.1371/journal.pone.0222562

Ye, J., McGinnis, S., Madden, T.L., 2006. BLAST: improvements for better sequence analysis. Nucleic Acids Research 34, W6–W9. https://doi.org/10.1093/nar/gkl164

Zavodna, M., Bagshaw, A., Brauning, R., Gemmell, N.J., 2014. The Accuracy, Feasibility and Challenges of Sequencing Short Tandem Repeats Using Next-Generation Sequencing Platforms. PLOS ONE 9, e113862. https://doi.org/10.1371/journal.pone.0113862

Zhang, H., Jain, C., Aluru, S., 2020. A comprehensive evaluation of long read error correction methods. BMC Genomics 21, 889. https://doi.org/10.1186/s12864-020-07227-0

Zhang, X.R., Meng, H.T., Shi, J.F., Wang, N., Wang, Z.Y., Zhang, B., Zhu, B.F., Guo, Y.X., 2023. Efficiency evaluation of common forensic genetic markers for parentage identification involving close relatives. Forensic Sci Int 345, 111594. https://doi.org/10.1016/j.forsciint.2023.111594

Zhang, Y.-M., Mao, Y., Xie, C., Smith, H., Luo, L., Xu, S., 2005. Mapping Quantitative Trait Loci Using Naturally Occurring Genetic Variance Among Commercial Inbred Lines of Maize (Zea mays L.). Genetics 169, 2267–2275. https://doi.org/10.1534/genetics.104.033217

Zhao, Y., Fang, L.T., Shen, T.-W., Choudhari, S., Talsania, K., Chen, X., Shetty, J., Kriga, Y., Tran, B., Zhu, B., Chen, Z., Chen, W., Wang, C., Jaeger, E., Meerzaman, D., Lu, C., Idler, K., Ren, L., Zheng, Y., Shi, L., Petitjean, V., Sultan, M., Hung, T., Peters, E., Drabek, J., Vojta, P., Maestro, R., Gasparotto, D., Kõks, S., Reimann, E., Scherer, A., Nordlund, J., Liljedahl, U., Foox, J., Mason, C.E., Xiao, C., Hong, H., Xiao, W., 2021. Whole genome and exome sequencing reference datasets from a multi-center and cross-platform benchmark study. Sci Data 8, 296. https://doi.org/10.1038/s41597-021-01077-5

Zhu, Y., Strassmann, J.E., Queller, D.C., 2000. Insertions, substitutions, and the origin of microsatellites. Genet Res 76, 227–236. https://doi.org/10.1017/s001667230000478x