Quantitative study of structural variations of chromosomes in saccharomyces cerevisiae

Alexandre Gillet-Markowska

Université Pierre et Marie Curie

Complexité du Vivant, ED515

Laboratoire de biologie computationnelle et quantitative (UMR7238) / Equipe de Biologie des Génomes

Thèse de doctorat de Génétique

Par Alexandre Gillet-Markowska

Étude quantitative des variations structurelles des chromosomes chez *Saccharomyces cerevisiae*

Dirigée par Gilles Fischer

Présentée et soutenue publiquement le 21/09/2015

Devant un jury composé de :

Pr. Dominique Higuet	Président
Dr. Marie-Claude Marsolier-Kergoat	Rapporteur
Dr. Gaël Yvert	Rapporteur
Pr. Olivier Lespinet	Examinateur
Dr. Gilles Fischer	Directeur de thèse

Résumé de la thèse

Les variations structurelles (SV) des chromosomes qui incluent les délétions, les duplications, les inversions, les translocations et les insertions sont d'importantes contributrices aux maladies génétiques, à la transformation des cellules malignes et plus généralement au polymorphisme des chromosomes. Nous avons développé un outil bio-informatique (Ulysses) qui permet désormais d'obtenir une image fine de ces SV qui se produisent dans les génomes. Nous avons montré avec des données de séquençage du génome humain simulées ou réelles, ainsi qu'avec des données de séquençage du cancer du sein, qu'Ulysses est le premier logiciel permettant la détection de SV rares dans un échantillon avec une bonne spécificité. Nous avons en effet montré qu'avec Ulysses nous pouvons maintenant tirer parti de la forte couverture physique de librairies MP et ainsi détecter des SV en faibles fréquences dans un échantillon. Ces résultats nous ont permis de mettre en place une nouvelle approche pour quantifier le nombre, les types et la localisation de SV qui se produisent *de novo* dans des populations monoclonales de levures. Ce travail démontre pour la première fois qu'il est possible de détecter des SV subclonales (SSV) dans des cultures de levures. Notre approche de détection des SSV présente l'avantage d'être beaucoup plus rapide à effectuer que des expériences d'accumulation de mutations. Elle peut donc être facilement réalisée dans différents fonds génétiques, souches ou espèces. Nous avons pu quantifier le taux et les types de SSV qui se produisent dans le génome de la levure. Nos estimations des taux des SSV se produisant naturellement dans des cellules sauvages est de l'ordre de 10^{-3} SSV /cellule/division ce qui implique que des populations réellement clonales de plus de 1 000 cellules n'existeraient pas. De plus, compte tenu du taux de mutation ponctuel chez la levure ($3{,}3.10^{-10}$ mutation/cellule/génération) et de la taille du génome (12,5Mb), il se produit donc en moyenne $3{,}3.10^{-10}$ x $12{,}5x10^{6}$ = $3{,}9.10^{-3}$ mutation/cellule/division. Les mutations ponctuelles et les SV se produisent donc à des fréquences équivalentes dans le génome. Parallèlement, nous avons montré que le niveau d'instabilité des individus dépend de facteurs génétiques de prédisposition. Pour les identifier, nous avons développé des systèmes génétiques de mesure des taux de SV chez la levure qui vont nous permettre à l'avenir d'identifier les gènes contrôlant l'instabilité chromosomique par analyse de liaison à grande échelle.

Remerciements

Le travail présenté dans ce mémoire a été réalisé au sein du laboratoire de "Biologie Computationnelle et quantitative" UMR 7238 UPMC / CNRS. Je tiens à remercier Gilles Fischer de m'avoir accueilli dans son équipe, d'avoir suivi ce projet tout au long de son développement, de m'avoir initié aux techniques de génétique des levures, de m'avoir introduit à la communauté des levuristes pendant les congrès et de rester à l'écoute pour me soutenir et me conseiller.

Je tiens à remercier Madame Marie-Claude Marsolier-Kergoat (CEA Institut de Biologie et de Technologies de Saclay) et Messieurs Gael Yvert (ENS Lyon), Olivier Lespinet (Université Paris-Sud) et Dominique Higuet (Université Pierre et Marie Curie) d'avoir accepté d'examiner mon travail en cette période où le thermomètre et la couleur du ciel ne s'y prêtent pas.

Merci à tous les membres du groupe "BiG", passés et présents, et plus particulièrement à Hélène avec qui j'ai passé de très bon moments à la paillasse mais aussi lors de nos pauses café (je vous rassure, maximum 2 fois par jour) mais qui me donne également du fil à retordre au squash ! Merci à Nicolas qui m'a initié à un nombre incalculable d'astuces au labo et qui aura toujours été fidèle au poste lorsqu'il s'agit de mettre à genoux le serveur Amadea ou Big Daddy ! Ce travail a également bénéficié de la contribution technique d'Ingrid Lafontaine et environnementale de Nikos Vakirlis et Guénola Santus.

Un grand merci à la clique des MC (Anne, Claire, Elodie, Juliana, Mathilde et Hugues qu'est jamais là ;)) *pour s'être mobilisés* pour ces soirées, pique-niques et weekend ! J'en profite pour remercier collectivement tous les membres du laboratoire qui ont contribué de près ou de loin à ce travail.

Jean-Philippe Meyniel de chez ISoft m'a été d'un grand secours tout au long de ma thèse. Ses compétences en Amadea m'ont aidé plus d'une fois lors de ce travail.

Un grand merci aux copains de promo (ils se reconnaîtront) dont le soutien tout ce temps en dehors du labo a pu avoir tellement de formes.

Un grand merci à Vince également pour toutes ces années de soutien et d'amitié !

Une pensée particulière à Mahé, mes parents, mes sœurs, et toute ma famille pour leur soutien et leur confiance journalière.

Table des matières

Abréviations

ADN : Acide désoxyribonucléique

ADNr: ADN ribosomique

ALF: A-like Faker

ARNm: ARN messager

BAF: B allele frequency

BIR: Break-Induced Replication

BFB: Break-Fusion-Bridge (cycle de cassure-fusion-pont)

CDB: Cassure Double Brin

CLS: 'Chronological Life Span'

CNV : Copy Number Variant (Variation du nombre de copies)

CSB: cassures simples brin

CTF: Chromosome Transmission Fidelity

DEL: Délétion

DMD: dystrophie musculaire de Duchenne

DUP: Duplication en tandem

eQTL: QTL d'expression

FISH: hybridation in situ en fluorescence

Gb: Giga-bases

GCR: Gross Chromosomal Rearrangement

GWAS: Genome wide association studies

HU: hydroxyurée

INS: Insertion

INV: Inversion

LD: Luria-Delbrück

LTR: Long Terminal Repeat

MA: Mutation Accumulation

MAL: Mutation Accumulation Lines (Lignées accumulatrices de mutations)

Mb: Mega-bases

MMEJ: Microhomology-Mediated End Joining

MMR: MisMatch Repair

NGS: Next Generation Sequencing

NHEJ: Non-homologous end joining

NRT: Non Reciprocal Translocation

pb : Paires de bases

PPV: Positive Predictive Value

QTL: Quantitative Trait Loci

RD: Read Depth

RH: Recombinaison Homologue

RLS: Replicative Life Span

RP: Read-Pair

RT: Reciprocal Translocation

SFR: Site Fragile Rare

SNP: Single Nucleotide Polymorphysm

SR: Split-Read

SV: Structural Variation (Variation structurelle)

TAR : Transcription-Associated Recombination

Contexte scientifique - Génomique et plasticité des génomes

En 1920, le botaniste allemand Hans Winkler publiait un livre marquant de l'histoire de la génétique intitulé "Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche" (Winkler, 1920). La traduction littérale du titre est "Étendue et causes de la parthénogenèse dans les règnes des plantes et des animaux". Cependant, ce livre n'est pas resté dans l'histoire de la science pour sa description de la parthénogenèse mais pour une idée essentielle de la page 165:

> *Ich schlage vor, für den haploiden Chromosomensatz, der im Verein mit dem zugehörigen Protoplasma die materielle Grundlage der systematischen Einheit darstellt den Ausdruck: das Genom zu verwenden und Kerne, Zellen und Organismen, in denen ein gleichartiges Genom mehr als einmal in jedem Kern vorhanden ist, homogenomatisch zu nennen, solche dagegen, die verschiedenartige Genome im Kern führen, heterogenomatisch.*

Voici ma traduction du début de la citation (d'après la traduction anglaise de Joshua Lederberg et d'Alexa McCray (Lederberg and McCray, 2001)) :

> *Je propose d'utiliser le terme 'Genom' pour décrire le lot de chromosomes haploïdes, qui avec le reste des composants protoplasmiques pertinents, définissent les éléments fondateurs des espèces.*

C'est pour cette invention du mot génome (une contraction de « génos» (« race » en grec) et de « chromosome ») que le livre de Hans Wrikler est sous le feu des projecteurs depuis près de 100 ans. Aujourd'hui, les concepts de « Génomes » et « Génomiques » - l'étude des génomes - sont partout et se sont même répandus dans la culture populaire.

Les prémisses de l'histoire de la génomique ne commencent cependant que bien après l'invention du mot génome. C'est en 1977 que Fred Sanger publie le premier génome séquencé, celui du virus phiX174 (Sanger et al., 1977). Cette contribution du Dr Sanger lui a valu de recevoir le prix Nobel de Chimie en 1980. Le perfectionnement des techniques mises au point par Sanger ont permis de séquencer au cours des années suivantes des génomes de taille croissante : bactériophage lambda (48 502pb) (Sanger et al., 1982), Vaccinia (192kb) (Goebel et al., 1990)…

Les premiers succès de séquençage reposaient cependant en grande partie sur la taille modeste des génomes viraux. En 1989, André Goffeau a mis en place un consortium pour séquencer le génome de la levure de boulanger *Saccharomyces cerevisiae* dont les 16 chromosomes totalisent $12{,}5.10^6$ bases. Ce fût une collaboration entre pas moins de 94 laboratoires. La taille du génome (12Mb) de la levure de boulanger, 60 fois plus grand que les plus gros génomes jusqu'alors séquencés, explique la nécessité de cet effort collectif. Et à cette époque, le séquençage de cet organisme modèle semblait être un premier pas logique vers l'étude du génome humain, lui-même 100 fois plus grand (3 Gb). Les connaissances en génétique et en biologie moléculaire suggéraient déjà que de nombreuses protéines effectuant des processus cellulaires élémentaires étaient conservées chez tous les eucaryotes, ce qui renforçait l'intérêt de ce projet. C'est ainsi qu'en 1996 était publiée la première séquence d'un génome eucaryote (Goffeau et al., 1996), soit seulement 1 an après que le groupe de Craig Venter ait publié les génomes de 2 bactéries : Haemophilus influenzae (1,4 Mb) (Fleischmann et al., 1995) et Mycoplasma genitalium (580 kb) (Fraser et al., 1995). L'achèvement du séquençage du génome de levure marqua un tournant dans l'histoire du séquençage puisque jusqu'à présent seulement des « petits » génomes (génomes viraux, de bactéries ou d'organelles) avaient été séquencés.

Le séquençage du génome de la levure n'était cependant pas le seul projet d'envergure car, dès septembre 1997, 4 projets de large échelle furent entrepris. Ainsi le génome de la mouche du vinaigre (*Drosophola Melanogaster*) était séquencé à 6%, celui du nématode (*Caenorhabditis elegans*) complété à 71%, celui de la souris (*Mus musculus*) complété à moins d'un pourcent et le génome humain complété à seulement 1,5%. La publication du génome du nématode (97 Mb) en 1998 en a fait le premier représentant du règne animal à être séquencé (The C. Elegans Consortium, 1998).

Ces projets constituaient à l'époque de réels exploits dans la mesure où ils intervenaient 10 ans avant le changement de paradigme technologique apporté par les technologies de séquençage de seconde génération (NGS pour 'Next Generation Sequencing'). Ce n'est en effet qu'en 2004 que '454 Life Science^{MD} ' lançait sur le marché son pyroséquenceur. La première version de la machine réduisait le coût de séquençage d'un facteur 6 par rapport au séquençage Sanger en capillaire des machines du groupe ABI^{MD}. Deux ans plus tard, en 2006, Solexa^{MD} lançait sur le marché son séquenceur Genome Analyzer^{MD} capable de séquencer 1Gb (composées de petites lectures de 36-50 pb) et ABI^{MD} son système SOLiD^{MD} (Sequencing by Oligonucleotide Ligation and Detection) de rendement un peu plus faible que le Genome Analyzer^{MD} mais qui était capable de produire des séquences plus longues (~300pb). La mise sur le marché de ces technologies dites de seconde génération a

permis de diminuer 100 fois le coût du séquençage par base entre 2004 et 2006 (de 0,01$ en 2004 à 0,0001$ en 2006).

Aujourd'hui, près de 20 ans après le séquençage du génome de la levure, des milliers de génomes appartenant à l'ensemble des règnes du vivant ont été séquencés grâce aux technologies NGS. La base de données GOLD du Joint Genome Institute (JGI), qui recense tous les organismes séquencés (génomes entièrement complétés ou partiellement assemblés (draft), compte désormais (au 20 mars 2015) 11 313 génomes eucaryotes, 39 969 génomes bactériens, 986 génomes d'archae et 4 410 génomes viraux. Un des éléments majeurs ressortant de ce séquençage massif des génomes au cours des 10 dernières années a été la découverte de l'ampleur des réarrangements chromosomiques existant entre les génomes. Ainsi, le séquençage des génomes humain et murin a montré qu'il existe un bien plus grand nombre de réarrangements que prévu entre ces 2 génomes (Pevzner and Tesler, 2003). En effet, ces différences entre les 2 espèces s'expliqueraient par la présence d'un nombre important de micro-réarrangements chromosomiques donnant ainsi un aperçu de la dynamique de l'évolution des génomes de mammifères.

Avec le progrès des techniques de séquençage, la génomique est peu à peu rentrée dans l'ère du re-séquençage de génomes et du séquençage de différentes souches d'une même espèce. Ceci a permis de montrer que cet intense polymorphisme des chromosomes se retrouve également de manière intra-spécifique. Ainsi, chaque être humain possède des différences de séquence par rapport au génome de référence pouvant aller de quelques kilo-bases à plusieurs méga-bases chez certains individus, sans pour autant que cela soit associé à une pathologie particulière. Cependant, des réarrangements chromosomiques ont tout de même été identifiés comme étant responsable de maladies génétiques (Stankiewicz and Lupski, 2010) montrant ainsi leur fort potentiel phénotypique. Les cancers sont notamment des pathologies associées à des remaniements chromosomiques. En effet, la majorité des cancers chez l'homme se caractérise par une instabilité chromosomique élevée qui se traduit par le gain ou la perte de chromosomes entiers et par des remaniements de leur structure. En clinique, ce processus d'instabilité est considéré comme un important contributeur à la transformation des cellules malignes et à la constitution d'une hétérogénéité intra-tumorale, dont les conséquences cliniques néfastes sur la résistance thérapeutique et le pronostic vital sont bien connues (Heng et al., 2013).

Une nouvelle étape dans l'appréciation de la plasticité des génomes a été franchie ces dernières années puisque l'on sait désormais que de nombreux tissus issus d'un même individu

sont susceptibles de présenter un mosaïsme chromosomique spécifique (O'Huallachain et al., 2012; Piotrowski et al., 2008), c'est-à-dire que l'on peut identifier des réarrangements chromosomiques tissus-spécifiques. Le séquençage de cellules individuelles a notamment permis de révéler que jusqu'à 80 évènements de rétrotransposition ont lieu par neurone et que 15% des neurones contiennent de larges réarrangements chromosomiques (Cai et al., 2014; Coufal et al., 2009). Des travaux sur des cellules sanguines ont également montré que la quantité de réarrangements chromosomiques est positivement corrélée avec l'âge (Jacobs et al., 2012; Laurie et al., 2012).

En 20 ans, la génomique est passée d'une ère où séquencer un génome humain nécessitait des moyens humains, techniques et financiers extraordinaires (3 milliards de dollars et plusieurs années) à une ère où il est possible pour n'importe quel laboratoire de séquencer plusieurs génomes pour quelques milliers d'euros en quelques semaines. L'énorme progrès technique du séquençage nous a permis d'obtenir un aperçu de la forte plasticité des génomes entre les espèces et beaucoup plus récemment à l'échelle d'une population, d'un individu, d'un tissu ou d'une cellule. Cependant, même si l'on connait désormais assez bien la phénoménologie des réarrangements chromosomiques, les deux enjeux majeurs pour ces prochaines années sont d'obtenir une description beaucoup plus détaillée des taux et des différents types de réarrangements chromosomiques auxquels sont naturellement soumis les génomes et de caractériser les déterminants génétiques qui contrôlent la prédisposition à l'instabilité chromosomique.

INTRODUCTION

Partie 1 - Identification des variations structurelles dans les génomes

1.1 - Qu'est-ce qu'une variation structurelle ?

La génétique telle qu'inventée par Gregor Mendel est aujourd'hui à un tournant. D'après Jim Lupski – généticien moléculaire au Baylor College of Medicine au Texas: "En génétique, l'état du courant de pensé Mendélien est dans le même état que la physique Newtonienne lorsqu'Einstein est arrivé : une nouvelle ère semble s'ouvrir devant nous". D'après les règles de transmission du patrimoine génétique de Mendel, chaque individu diploïde hérite d'exactement 2 copies de chaque gène, un de chaque parent (avec pour exception les chromosomes sexuels). Cependant, la possibilité que le nombre de copies de certains fragments d'ADN dans les génomes soit variable défie ces règles simples de la transmission de l'information génétique au cours des générations. Les variations du nombre de copie de l'ADN font partie d'une classe de réarrangements chromosomiques plus large appelée Variations Structurelles ou SV pour Structural Variation en anglais. Elles forment un groupe intermédiaire entre les mutations ponctuelles qui n'affectent qu'une seule base et les aneuploïdies qui font varier le nombre d'un ou plusieurs chromosomes dans un génome. Les SV sont des réarrangements de l'ADN de 50 paires de bases minimum[1] qui regroupent plusieurs catégories de réarrangements chromosomiques : les SV balancées qui ne font pas varier le nombre de copies de l'ADN comme les inversions et les translocations réciproques et les SV non balancées (ou CNV pour Copy Number Variants) qui incluent les duplications, les délétions et les translocations non réciproques (Figure 1). Bien que ce type d'évènement soit connu depuis longtemps (Bridges, 1936), l'importance de ces variations sur le polymorphysme chromosomique chez l'homme n'a été démontrée qu'à partir de 2004 dans 2 publications majeures portant sur l'impact quantitatif des SV sur le génome humain (Iafrate et al., 2004; Sebat et al., 2004).

[1] La taille minimum des SV (non valable pour les translocations) est définie arbitrairement à 50 paires de bases minimum. Cette limite arbitraire avait initialement été fixée à 1000 pb mais a été réduit à 50pb grâce au progrès des techniques de détection (voir partie 1.2.3).

Il est désormais reconnu que la proportion de nucléotides affectée par des SV est équivalente ou plus importante que celle affectée par des mutations ponctuelles ou SNP ('Single Nucleotide Polymorphysme') (Kloosterman et al., 2015). Cependant, alors que les SNP sont directement identifiables dans les séquences courtes d'ADN produites par les séquenceurs, et donc facilement quantifiables, les SV peuvent s'étendre sur plusieurs mégabases. Depuis les travaux pionniers de Dobzhansky sur les inversions chez la drosophile (Dobzhansky and Sturtevant, 1938) démontrant que ces dernières sont polymorphes dans les populations, les chercheurs ont dépensé beaucoup d'énergie pour étudier l'impact quantitatif et qualitatif des SV dans les génomes. Cependant, la formation des SV implique souvent des séquences répétées qui compliquent leur caractérisation. L'identification des SV représente donc un défi de la biologie moderne et elle fait l'objet d'une attention particulière d'une large communauté de généticiens, bio-informaticiens, physiciens, mathématiciens et médecins.

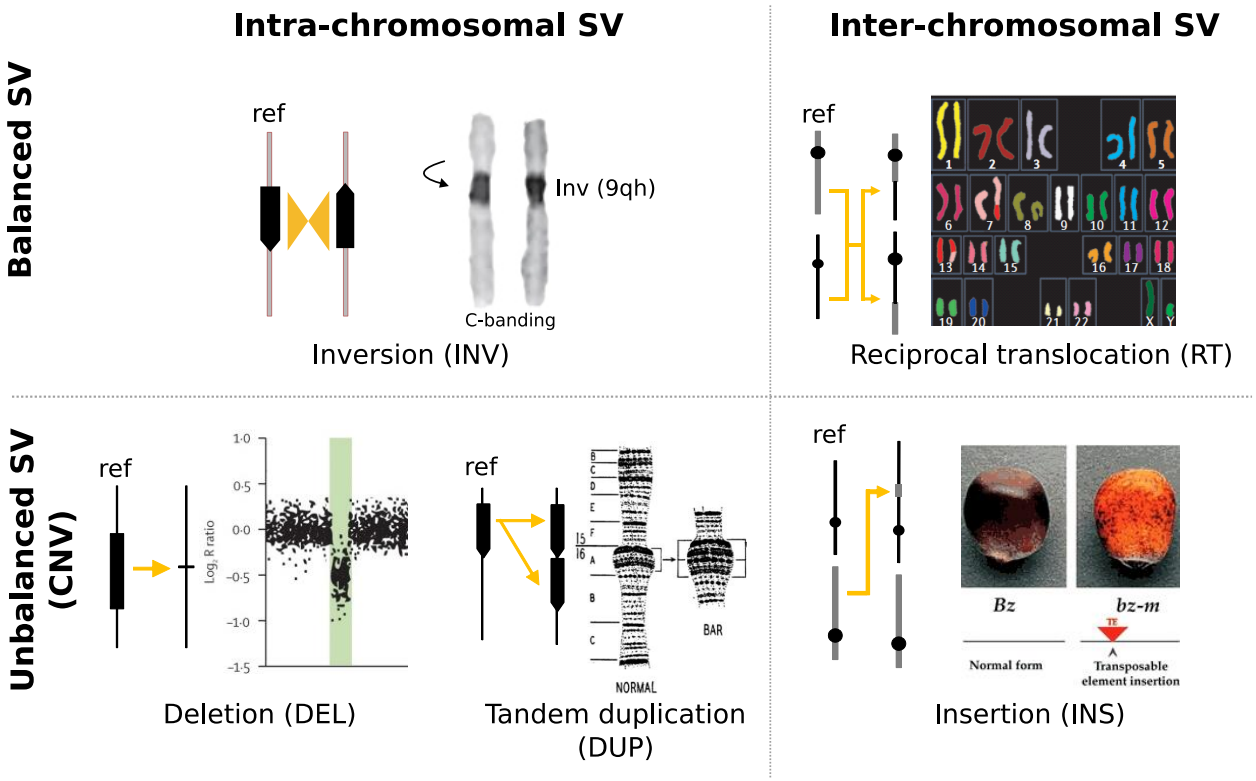


Figure 1: Classification des Variations structurelles dans les génomes. Les SV se subdivisent en deux groupes : les SV balancées qui ne font pas varier le nombre de copies et les SV non balancées qui font varier le nombre de copies de l'ADN. Dans ces deux catégories, on distingue les SV intra et inter-chromosomiques. L'inversion est illustrée par une expérience de coloration en bande C qui permet de visualiser l'inversion d'un fragment chromosomique. La délétion est illustrée par une baisse de l'intensité des sondes d'une puce à ADN. La translocation réciproque de bras chromosomique est illustrée par la technique de peinture des chromosomes qui permet de visualiser les bras transloqués. Crédits images: Feuk et al. 2006 ; Bridges 1936

1.2 - Méthodes d'identification des variations structurelles dans les génomes

1.2.1 - Cytogénétique

Caryotypes simples

Le terme cytogénétique fait référence à l'étude des aspects cellulaires de l'hérédité, plus particulièrement à la description de la structure des chromosomes et à l'identification d'aberrations chromosomiques. Le point de départ de cette discipline remonte au début du 20ème siècle avec la découverte par Robertson de l'une des premières SV chez le criquet (*Caelifera acrididae* et *Caelifera tetrigidae*) (Robertson, 1916). Dans ce travail, l'auteur décrit, grâce à l'observation de chromosomes métaphasiques, des translocations non réciproques entre 2 chromosomes acrocentriques. Ce type de translocation, qui résulte en la formation d'un grand chromosome contenant les deux bras *q* et en la perte du néo-chromosome formé des 2 bras *p*, porte aujourd'hui son nom.

C'est cependant à la mouche du vinaigre du genre *Drosophila* que nous devons l'essor des techniques de cytogénétiques. En 1934, Painter adapte la méthode de coloration des chromosomes par l'acetocarmine au marquage des chromosomes polytènes des glandes salivaires de la drosophile (Painter, 1934). Cette méthode permet d'obtenir une détection rapide des inversions chez la drosophile (Tan, 1935). Une carte des inversions existantes sur le chromosome 3 de *Drosophila pseudoobscura* a pu être produite dès la fin des années 30 (Dobzhansky and Sturtevant, 1938) (figure 2). Le marquage des chromosomes polythènes de glande salivaire a également permis la découverte d'une des premières SV associée à un phénotype (Bridges, 1936). La SV décrite dans cet article publié dans le journal Science est une duplication en tandem (DUP) associée au phénotype 'bar', bien connu des généticiens drosophilistes. Les mouches porteuses de cette duplication ont un œil plus fin que l'œil des mouches sauvages (figure 1). Le segment dupliqué peut à son tour se dupliquer et donner lieu au phénotype hyper-bar avec un œil encore plus fin. Cette DUP est probablement la première à avoir démontrée le fort potentiel phénotypique des SV. Historiquement, la cytogénétique servait au diagnostic des maladies génétiques humaines associées aux grands réarrangements chromosomiques. Cependant, plusieurs décennies supplémentaires ont été nécessaires pour mettre au point le caryotype complet à 46 chromosomes du génome humain (Ford and Hamerton, 1956). La description de caryotypes présentant une anomalie du nombre de chromosomes et associés à divers syndromes a ensuite rapidement suivi : syndrome de Down en 1958 (Lejeune et al., 1959), syndrome de Klinefelter et de Turner en 1959 (Ford et al., 1958; Jacobs and Strong, 1959). Puis, une découverte majeure a été faite dans le domaine de la cancérologie avec l'identification du chromosome de Philadelphie qui était régulièrement

trouvé dans le sang de patient atteint de Leucémie myéloïde chronique (Nowell and Hungerford, 1960).

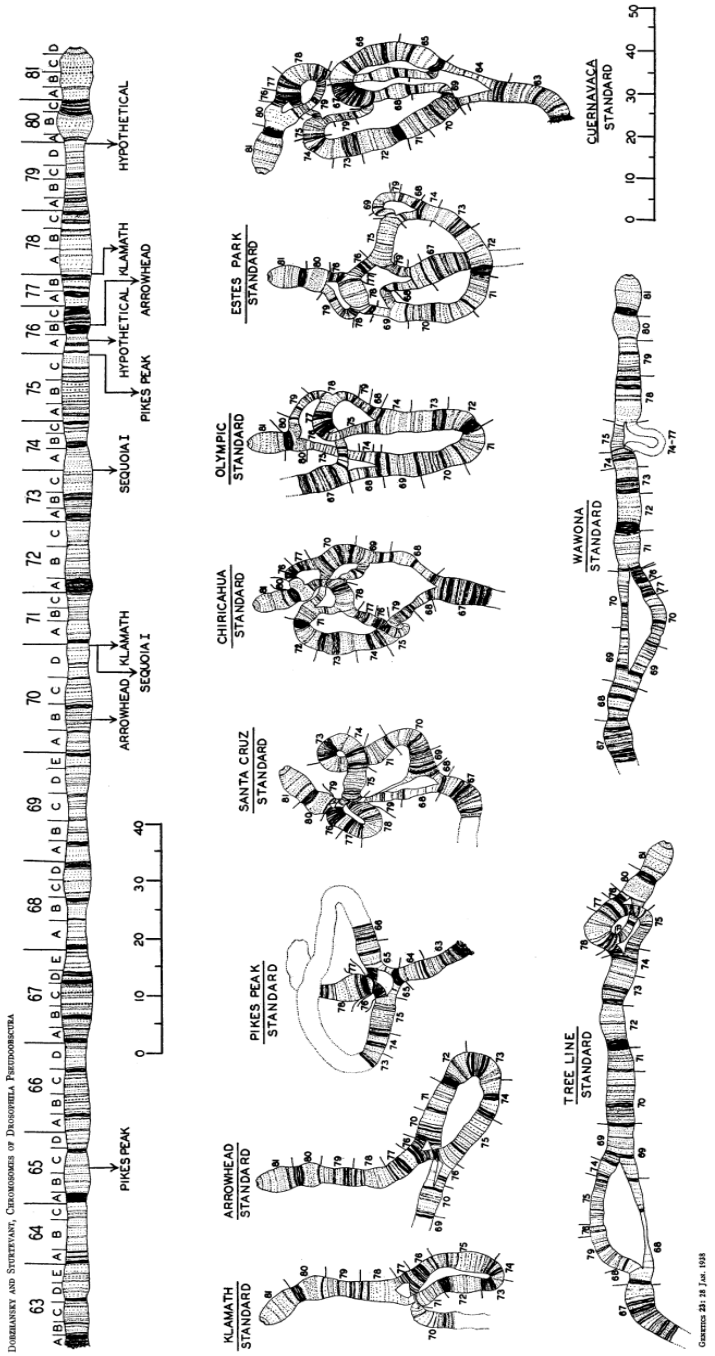


Figure 2: Haut - Cartographie des inversions sur le chromosome 3 de *Drosophila pseudoobscura*. Bas - configuration du chromosome 3 dans les hétérozygotes pour chaque inversion. Crédits: Dobzhansky & Sturtevant 1936

Colorations en bandes

Une analyse plus détaillée des caryotypes est devenue possible à partir des années 60 avec l'invention de la coloration en bandes. En utilisant des fluorochromes couplés à un agent alkylant (comme la quinacrine), un motif de marquage spécifique de chaque chromosome pouvait être obtenu permettant ainsi d'identifier chaque chromosome (Caspersson et al., 1970). Dans les années 70, la coloration Giemsa (qui révèle les zones riches en adénine et en thymine) a rapidement remplacé la coloration à la quinacrine. Les améliorations de la résolution des marquages en bandes ont ainsi permis de découvrir que le chromosome de Philadelphie ne résultait pas d'une délétion sur le chromosome 22 mais d'une translocation entre les chromosomes 9 et 22 (Rowley, 1973) (Figure 3).

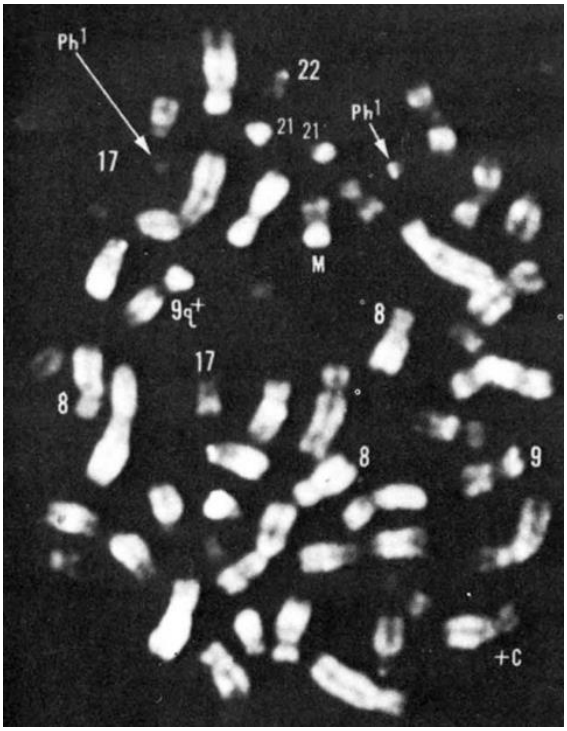


Figure 3: Chromosomes métaphasiques colorées à la quinacrine. Les deux chromosomes de Philadelphie aussi nommes t(9 ;22)(q34 ;q11) sont indiqués par le sigle "Ph". Cette translocation entre les chromosomes 9 et 22 aboutit à la fusion des gènes BRC et ABL ce qui forme le gène chimérique BRC-ABL créant ainsi une mutation de l'oncogène ABL. Crédits: Rowley 1973

Comme l'indique l'étude de Zhao et al. récemment publiée (Zhao et al., 2015) dans laquelle 200 000 caryotypes ont été réalisés afin de diagnostiquer des translocation Robertsoniennes, la coloration de Giemsa est toujours couramment utilisée. Cependant, cette technique de caryotypage classique se limite à la détection d'aberrations chromosomiques majeures comme des aneuploïdies et des polyploïdies ou des SV macroscopiques de plusieurs méga-bases.

Cytogénétique moléculaire

Les techniques de cytogénétique moléculaire reposent majoritairement sur l'utilisation de l'hybridation in situ en fluorescence (FISH) (Bauman et al., 1980). Le principe de cette technique est d'utiliser la microscopie à fluorescence pour révéler la présence et la localisation de sondes d'ADN ou d'ARN marquées et spécifiques d'un locus déterminé sur des chromosomes métaphasiques, des noyaux en interphase ou des fibres de chromatine.

Les techniques de cytogénétique moléculaire classiques utilisées sur des chromosomes métaphasiques permettent la détection de SV sub-microscopiques. Cependant, l'utilisation de brins de chromatine (Heng et al., 1992) plutôt que de chromosomes métaphasiques permet de passer d'une résolution d'environ 5Mb à environ 5kb. La réduction de la taille des sondes, d'abord à l'aide de fosmides puis plus récemment à l'aide de banques d'oligonucléotides (Yamada et al., 2011) a également permis d'améliorer la précision des marquages des chromosomes.

La coloration simultanée de l'ensemble des chromosomes par des sondes d'ADN, dite peinture chromosomique, a permis d'obtenir à la fin des années 80 le premier caryotype humain basé sur la technique du FISH (Cremer et al., 1988). De nombreuses améliorations ont ensuite été apportées à cette technique jusqu'à permettre la coloration combinée des chromosomes avec plusieurs fluorochromes ('Multiplex-FISH', 'spectral-karyotyping' et 'combined binary ratio labelling') (Schrock et al., 1996; Speicher et al., 1996; Tanke et al., 1999). Ces techniques sont particulièrement utiles pour la détection des translocations chromosomiques (figure 1B) et des aneuploïdies. Il a ainsi été possible de développer la technique du FISH multicolore pour des applications spécifiques, notamment pour passer au crible des régions particulières du génome comme les télomères (Brown et al., 2001). L'identification de SV intra-chromosomique a quant à elle été facilitée par des techniques de coloration en bandes multicolores comme le 'Cross-Species Color Segmentation' (Müller et al., 1998).

L'évolution de la cytogénétique moléculaire a permis de répondre à une autre difficulté rencontrée par les cytogénéticiens cliniques : la préparation de chromosomes métaphasiques à partir de cellules tumorales est particulièrement difficile. Une première technique consiste à hybrider sur des chromosomes métaphasiques de cellules saines de l'ADN issu de cellules tumorales ou bien de cellules saines marqué par des fluorochromes différents (Kallioniemi et al., 1992). Cette technique appelée hybridation génomique comparative (CGH pour *comparative genomic hybridization*) permet de visualiser les variations du nombre de copies dans l'ADN à l'échelle du génome complet. Cependant, elle comporte de nombreuses limitations : (i) elle ne permet pas de détecter des réarrangements balancés comme les translocations

réciproques et les inversions, (ii) les anomalies de la ploïdie ne sont pas détectées non plus et (iii) elle est globalement peu informative sur le type de SV responsable des variations du nombre de copies.

En plus des progrès des techniques de marquage des chromosomes, la cytogénétique à beaucoup profité des progrès du matériel et des logiciels de détection de la fluorescence. L'invention des caméras CCD ('*Cooled charge couppled-device*') et de filtres plus efficaces et spécifiques pour les microscopes ont ainsi énormément contribué à l'amélioration de la sensibilité et de la résolution de l'imagerie. Le marquage des fibres de chromatine est aujourd'hui plus facile grâce à une procédure automatisée de peignage moléculaire (Michalet et al., 1997). Le développement des logiciels de traitement d'image ces dernières années a également eu un impact non-négligeable sur notre capacité à analyser les images de cytogénétique moléculaire.

1.2.2 - Puces à ADN

L'avènement des puces à ADN dans les années 1990 a permis d'outrepasser les 2 limites majeures des outils de cytogénétique moléculaire : leur manque de résolution et le faible débit des expériences. Bien que les puces à ADN aient été initialement développées pour mesurer l'expression différentielle de gènes, elles ont rapidement été utilisées dans des domaines multiples tels que la découverte de CNV, le génotypage des SNP, l'étude de la méthylation de l'ADN, de l'épissage alternatif, des interactions protéines-ADN et petits ARN-ADN (techniques associées à l'immunoprécipitation de chromatine). La détection des SV est cependant restreinte à 2 types de puces qui sont les puces CGH et les puces à SNP.

Découverte de SV non balancés avec les puces CGH

Les premières expériences de puces CGH (Solinas-Toldo et al., 1997) avaient pour objectif d'améliorer la résolution de l'hybridation génomique comparative classique. Dans les puces CGH, les chromosomes en métaphase sont remplacés par des sondes d'ADN fixées sur une puce (Figure 4). Le principe de ces puces est le suivant : l'échantillon à tester, constitué de fragments d'ADN ou d'ARN marqués par un fluorochrome, est hybridé sur la surface de la puce de manière compétitive avec un échantillon de référence marqué par un fluorochrome différent. Apres l'hybridation des 2 échantillons, l'intensité des 2 fluorochromes (Cy5 et Cy3) est mesurée grâce à un scanner de puces. La première étape de l'analyse consiste à calculer pour chaque sonde le ratio entre l'intensité mesurée pour l'échantillon test et l'intensité mesurée pour l'échantillon référence. Ce ratio est le plus souvent exprimé en échelle

logarithmique de base 2 (on parle alors de log2 ratio). On peut ensuite tracer ces ratios le long des chromosomes. Théoriquement, un ratio de 0 (log2 (2/2)= 0) signifie qu'il existe le même nombre de copie de la sonde dans les 2 échantillons d'ADN, un log2 ratio de 0,58 (log2 (3/2) = 0,58) signifie un gain de copie dans l'échantillon de test par rapport à la référence et inversement, un log2 ratio de -1 (log2 (1/2) = -1) indique la perte d'une copie dans l'échantillon de test. Afin de limiter les effets des CNV présents dans l'échantillon de référence, ce dernier est généralement un mélange de l'ADN d'un grand nombre d'individus (N=~100).

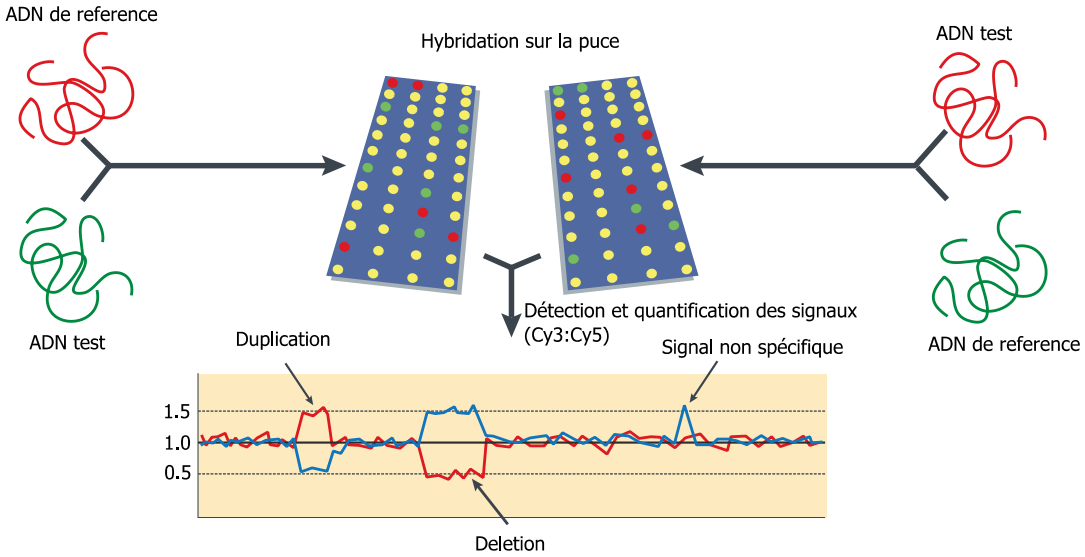


Figure 4: Principe des puces CGH. L'échantillon à tester est marqué par fluorochrome de la référence, l'un avec Cy3, l'autre avec Cy5. Un « dye-swap » qui consiste à réaliser l'expérience une seconde fois en inversant les fluorochromes, peut être effectué pour éviter les biais de marquages différentiels entre les 2 fluorochromes. Le mélange d'ADN test et de référence est ensuite hybridé sur la surface de la puce. La lecture de l'intensité des fluorochromes au niveau de chaque sonde permet ensuite d'analyser les variations du nombre de copies de l'ADN. Adapté de Feuk 2006

Les sondes des puces dessinées pour le génome humain étaient initialement constituées de milliers de larges séquences d'ADN (80-200kb) sélectionnées sur des intervalles de 1Mb et clonées dans des chromosomes bactériens artificiels (BAC). En 2004, la première puce dite 'tiling array' où l'ensemble du génome était représenté a été créée (Ishkanian et al., 2004). Cette puce comprenait plus de 430 000 séquences de BAC chevauchantes et dont la résolution rendait possible la découverte de CNV. La résolution de ces puces s'est ensuite considérablement améliorée grâce au remplacement des BAC par des ADN complémentaires, puis des amplicons de PCR et, à partir du milieu des années 2000, des oligonucléotides.

SNP array

Tout comme les puces CGH, les puces SNP ont fait l'objet d'énormes progrès depuis leur invention. Au départ, ces puces ne pouvaient génotyper que quelques milliers de SNP (Bignell et al., 2004) alors qu'aujourd'hui ce sont plusieurs millions de SNP qui peuvent être étudiés sur une même puce. Contrairement aux puces CGH qui reposent sur la co-hybridation de 2 échantillons sur une même puce, les puces SNP ne reçoivent qu'un seul échantillon. L'analyse des CNV à l'aide de puce SNP repose sur 2 paramètres : la comparaison des valeurs de la fluorescence de l'échantillon à une référence (log2 ratio) et la « fréquence de l'allèle B » (BAF pour '*b allele frequency*') qui est la proportion totale du signal d'un locus (allèles A et B) expliquée par le seul allèle A (Alkan et al., 2011). Les génotypes AAB et ABB correspondent à des valeurs de BAF spécifiques (mais aussi les génotypes plus complexes comme AAAB, AABB et BBBA) permettant de déduire le nombre de copies de l'ADN pour des variations allant de 0

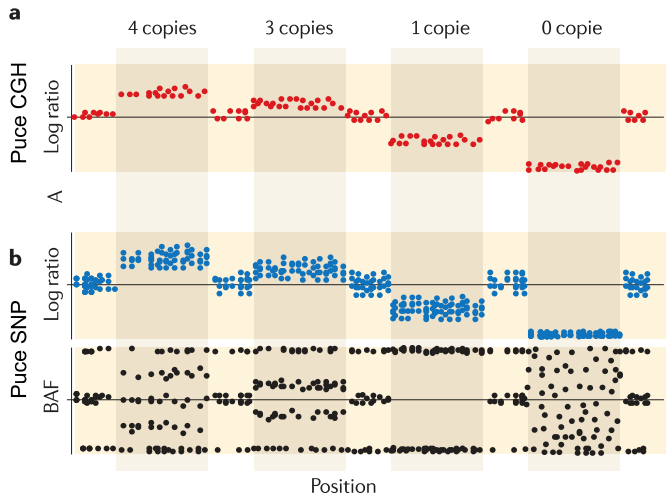


Figure 5: Puces CGH *versus* puce SNP. (a) Dans les puces CGH c'est le ratio du signal entre l'échantillon test et l'échantillon de référence qui est converti en log2 ratio pour indiquer la variation du nombre de copies. (b) Les puces SNP arrivent à une métrique similaire en comparant l'intensité de l'échantillon en cours d'analyse a une collection de références ou au reste de la population analysée. Cette approche ne permet pas d'obtenir un ratio signal/bruit aussi bon que les puces CGH. En revanche, les puces SNP offrent une métrique supplémentaire appelé la BAF (b allele frequency) qui est la proportion de la totalité du signal allélique (A+B) expliquée par le seul allèle A. Cette mesure a en revanche un très bon ratio signal/bruit ce qui permet de détecter des variations du nombre de copies allant de 0 à 4. Crédits : Alkan et al. 2011

à 4 copies (figure 5). Afin d'améliorer la détection des CNV, les fabricants de puces ont commencé à ajouter vers le milieu des années 2000 des sondes situées dans des zones du génome dans lesquelles aucune variation du nombre de copie de l'ADN n'a été répertoriée. Ils ont par ailleurs également augmenté la densité des sondes dans les régions propices aux CNV. Par exemple, la puce humaine Affymetrix^{MD} 6.0 contient environ 50% de sondes dédiées à la détection de CNV (McCarroll et al., 2008).

De nombreux algorithmes que nous ne passerons pas en revue ici sont disponibles pour détecter les CNV à partir de ces puces SNP (Dellinger et al., 2010; Glessner et al., 2013; Pinto et al., 2011). Une seconde approche bio-informatique existe qui consiste à utiliser les données de GWAS (Genome wide assocation studies ou étude d'association pangénomique) présentes dans les bases de données pour appeler des CNV. Dans ce cas, le signal d'un SNP dans un allèle par rapport au signal standard d'un allèle normalisé permet de déduire la délétion ou la duplication d'un segment d'ADN (Glessner et al., 2012).

Cependant, certains réarrangements chromosomiques comme les translocations réciproques et les inversions ne peuvent pas être détectés grâce aux puces à ADN. En effet, ils ne font pas varier le nombre de copies des fragments d'ADN. Le seul élément caractéristique de leur présence est leurs jonctions dont la détection est sous le seuil de détection des puces à ADN. La seconde limitation des puces à ADN concerne le faible taux de vrais positifs. Ceci s'illustre parfaitement dans le cas des puces SNP lors de la comparaison de 3 des algorithmes de détection des CNV les plus utilisés (PennCNV, QuantiSNP, et Birdsuite) : seulement 1,5% des CNV trouvés par ces 3 logiciels sont en commun (Kim et al., 2012). De plus, le taux de vrais positifs après validation expérimentale (qPCR) est seulement de ~38% lorsque le CNV est détecté par un seul logiciel, 57% lorsqu'il l'est par 2 et 71% lorsqu'il est détecté par 3 logiciels conjointement. Il reste donc 30% d'erreur dans les 1,5% de CNV identifiés malgré leur confirmation par 3 logiciels différents.

1.2.3 - Séquençage de seconde génération

L'avènement des technologies de séquençage de seconde génération a révolutionné la détection des SV dans les génomes et a rapidement remplacé les puces à ADN, si bien que ces dernières ne sont désormais utilisées principalement qu'en milieu clinique, à des fins diagnostiques. La détection de SV à partir des données de séquençage de seconde génération (NGS) présente néanmoins des difficultés techniques majeures qui ont nécessité l'invention de nouvelles méthodes d'analyses. Ces techniques peuvent être regroupées en 4 grandes classes : l'assemblage, les alignements partiels des lectures (ou '*Split-reads*', SR), la profondeur de la couverture ('*Read-Depth*', RD) et les lectures appariées ('*Read-Pair*', RP) (Figure 6).

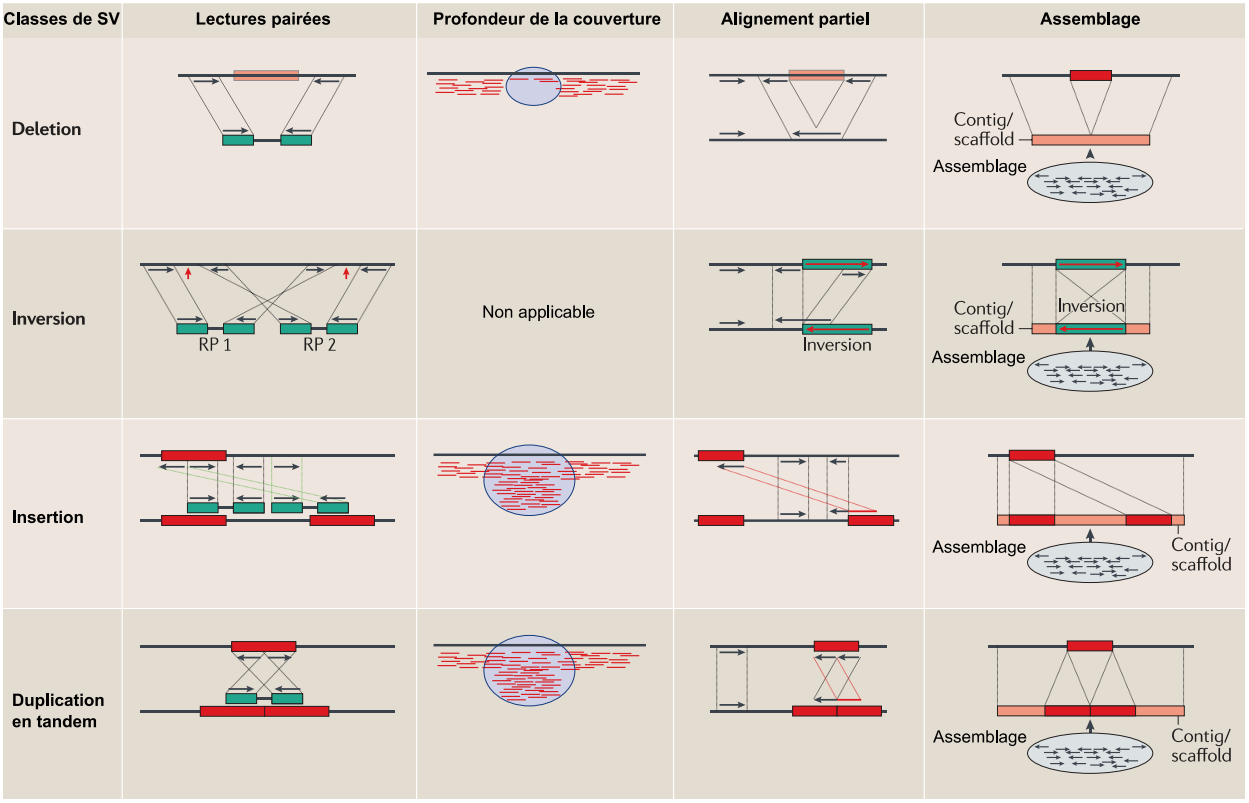


Figure 6: Méthodes de détection des SV. Il existe 4 approches analytiques de la détection des SV. Les lectures appariées peuvent être utilisées pour détecter n'importe quel type de SV car chacun est caractérisé par une cartographie particulière des lectures : les délétions provoquent une augmentation de la taille d'insert des RP, les inversions induisent des groupes de RP intrachomosomiques ayant une orientation [-,-] a une jonction et [+,+] a l'autre, les insertion induisent des RP dont une des deux lectures a un pied dans l'insertion et l'autre dans le chromosome receveur, enfin les duplications en tandem provoquent des RP qui ont une orientation opposée a l'orientation majoritaire de la librairie. La profondeur de séquençage ne permet de qualifier que des CNV, les délétions faisant diminuer la couverture de séquençage et les insertions/duplications l'augmentent. Le principe de l'alignement partiel est d'identifier la jonction des SV dans des lectures. Cette dernière méthode permet donc en théorie de détecter n'importe quel type de jonctions pour lesquelles on est capable de cartographier de façon unique les lectures. Enfin, la méthode de l'assemblage permet également en théorie de reconstruire la structure de n'importe quel SV même si en pratique les séquences répétées affectent beaucoup cette capacité. Crédits : Alkan et al. 2011.

1.2.3.a - Assemblage de séquences

Cette méthode est en théorie la plus précise pour la découverte de SV. Le principe est de reconstruire *de novo* le génome d'intérêt en utilisant les données de séquençage, et ce, avec ou sans connaissances préalable de la structure du génome. Ainsi, après ré-assemblage des séquences, une simple comparaison avec le génome de référence permet de lister les différences structurelles du génome d'intérêt (figure 6). En pratique, un tel assemblage, qui repose sur des séquences courtes d'une centaine de paires de bases au maximum, est très difficile. Différents algorithmes ont ainsi été développés. Les assembleurs classiques (SOAPdenovo (Luo et al., 2012), ALLPATHS-LG (Gnerre et al., 2011)…) utilisent des graphes de Bruijn pour assembler un génome. En observant les bifurcations qui séparent les couleurs (les

bulles), la séquence de variations structurelles peut être déduite. Bien que la détection de bulles soit adaptée à la découverte de SV balancées, la détection des CNV est plus complexe. En effet, une duplication de l'ADN introduit en générale une copie quasi parfaite dans le génome. La répétition ainsi engendrée pose des problèmes aux assembleurs basés sur les graphes de Bruijn qui ne peuvent pas déterminer quelles séquences d'ADN entrent et sortent de ces répétitions, ce qui empêche la détection de bulles. La répétition n'est donc pas intégrée à la reconstruction d'un autre fragment du génome ('contig') mais est rapportée comme un 'contig' isolé. Ceci est d'autant plus problématique que les génomes eucaryotes contiennent souvent une fraction importante de séquences répétées et de séquences dupliquées. L'un des premier assembleurs (Cortex) dédié à la découverte de CNV , qui par ailleurs ne nécessitait pas l'utilisation d'un génome de référence pour l'assemblage, fût publié en 2012 (Iqbal et al., 2012).

Bien que cette méthode de détection des SV par assemblage ait permis dès 2008 de détecter des SV chez l'homme (Kidd et al., 2008), d'autres limites interviennent et elle est finalement l'une des méthodes les moins employées aujourd'hui. Tout d'abord elle est complexe et nécessite d'importantes ressources en calculs. De plus, elle ne permet pas en général de déterminer l'haplotype et par conséquent les SV sont homozygotes.

1.2.3.b - Profondeur de Séquençage

Cette méthode RD est basée sur l'hypothèse qu'il existe une corrélation directe entre la valeur de la couverture locale de l'ADN et le nombre de copies du locus. Plus précisément, les méthodes basées sur la profondeur du séquençage supposent que la distribution des lectures le long du génome est uniforme (typiquement suivant une loi de Poisson) et toute déviation de cette distribution correspond à une délétion ou une duplication (figure 6).

En pratique, les lectures issues du séquenceur sont alignées sur un génome de référence (Langmead et al., 2009; Li and Durbin, 2009), puis la couverture de séquençage dans une fenêtre de taille déterminée est calculée. Cette couverture brute est ensuite corrigée pour compenser les biais dus au contenu en GC (Boeva et al., 2011; Janevski et al., 2012) avant d'appliquer un algorithme de segmentation qui va identifier toutes les zones continues et de tailles variables du génome présentant la même couverture. D'un point de vue expérimentale, il existe ensuite 3 situations dans lesquelles on peut utiliser la méthode de la profondeur de séquençage : un échantillon unique, des échantillons pairés (échantillon/contrôle) et le cas des échantillons multiples. Dans le cas d'un échantillon unique, un algorithme est appliqué directement sur ces données de couverture pour établir quelles fenêtres dévient

significativement de la couverture moyenne et peuvent représenter des CNV. Dans le cas d'une paire échantillon/contrôle, c'est la déviation du nombre de copies de l'échantillon test par rapport à l'échantillon contrôle qui est testée. Enfin, dans le cas d'échantillons multiples, c'est une déviation de chaque échantillon par rapport à la moyenne de la population des échantillons qui est utilisé.

Cette méthode basée sur la profondeur du séquençage se limite principalement à la détection de CNV. Elle ne permet pas de détecter les SV balancées même si certains algorithmes possèdent des stratégies permettant de détecter les micro-chutes de couverture caractéristiques des jonctions des SV balancées (Sindi et al., 2012). De plus, cette méthode ne permet pas de détecter facilement les petits CNV (inferieurs a 1kb) bien que cette question ait été adressée par plusieurs études (Abyzov et al., 2011; Yoon et al., 2009).

La méthode de la profondeur de la couverture permet cependant d'évaluer le plus précisément possible le nombre de copies de séquences d'ADN par opposition aux techniques des lectures appariées et des alignements partiels (voir ci-après). L'intérêt de la communauté pour cette approche de détection des CNV est notamment illustré par l'implémentation récente d'un algorithme de détection dédié à une plateforme de calcul intensif hautement parallélisée reposant sur des processeurs graphiques (GPU) (Manconi et al., 2015).

1.2.3.c - Alignements partiels des lectures

Parmi l'ensemble des techniques, celle des alignements partiels des lectures (ou SR pour '*Split-Read*') propose la meilleure résolution puisqu'elle permet de cartographier les limites des SV à l'échelle de la paire de bases. Cette technique a d'abord été utilisée pour identifier des indels (insertions/délétions) notamment dans le génome humain (Mills et al., 2006). Elle a ensuite été adaptée aux lectures des séquenceurs de seconde génération. Le principe est d'identifier des lectures qui ne peuvent pas être cartographiées dans toute leur longueur sur le génome de référence mais dans lesquelles on peut identifier 2 régions qui s'alignent dans 2 positions distinctes du génome. Initialement, un des défauts de cette méthode était le coût en calcul de la recherche des alignements partiels. Mais, l'utilisation de 2 lectures appariées, (voir ci-après) dont l'une est alignée sur le génome de référence, a permis de réduire l'espace de recherche de la lecture non cartographiée (Ye et al., 2009). Cette optimisation a contraint alors à ne pouvoir aligner des lectures qu'à proximité d'une lecture déjà complètement alignée, limitant ainsi les algorithmes classiques (Gustaf (Trappe et al., 2014), Pindel (Ye et al., 2009), SVSeq2 (Zhang et al., 2012)) à détecter des SV relativement petits. Il a fallu un nouvel

algorithme : Prism (Jiang et al., 2012), issu de l'algorithme modifié de Needleman–Wunsch, (alignement global de deux chaînes de caractères) pour passer outre cette limitation.

1.2.3.d - Lectures appariées

La technique des lectures appariées consiste à générer des paires de séquences situées à une distance physique connue puis à les cartographier sur un génome de référence. Le plus souvent, l'identification des SV à partir de ces paires de séquences (ou RP pour 'Read Pair') se fait en identifiant des groupes de RP chevauchantes et discordantes (figure 6 et 7) (voir le paragraphe suivant pour les méthodes d'identification des groupes de RP). Une RP discordante correspond à une paire dont un des paramètres décrivant sa cartographie par rapport au génome de référence est inattendu. Ces paramètres de discordance sont définis par une combinaison de 3 critères : la taille de l'insert (la distance entre les 2 lectures cartographiées sur le génome de référence), l'orientation des 2 séquences de la paire et le chromosome sur lequel elles sont cartographiées. Chaque type de SV simple possède sa propre signature qui se traduit par sa combinaison caractéristique de ces 3 critères (figure 7). Un des grands avantages de la méthode des RP est qu'elle permet, comme la méthode du SR, la détection de tous les types de SV (figure 9). En revanche, elle ne permet pas de détecter des duplications, des insertions et des délétions dont la taille est inférieure à la taille d'insert moyenne des lectures (généralement ~300pb). Il est toutefois possible de contourner ce problème en utilisant des RP sans rechercher et grouper les RP discordantes. Cette autre approche est basée sur un modèle statistique qui teste la probabilité qu'une paire de séquences ait une taille d'insert différente du reste des RP. Elle est en général adaptée pour détecter de petites délétions <100pb (Chen et al., 2009; Marschall et al., 2012).

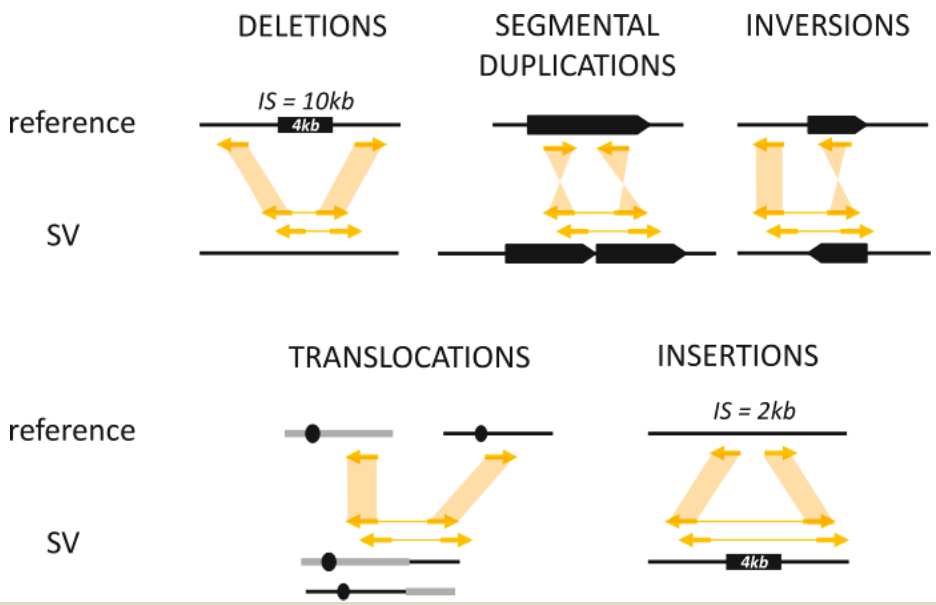


Figure 7: Détection des SV par la méthode des RP. Les molécules d'ADN dont les extrémités sont séquencées sont représentées par un trait orange avec deux flèches aux extrémités. Ces fragments d'ADN ont tous environ la même taille physique. La taille d'insert et l'orientation des lectures ne sont cependant définies qu'après la cartographie des RP sur le génome de référence. Chaque type de SV possède sa propre signature facilement détectable : les délétions se caractérisent par une augmentation de la taille d'insert par rapport au reste des RP de la librairie, c'est l'inverse pour les insertions. Les duplications segmentales en tandem produisent des RP avec une orientation opposée à l'orientation majoritaire de la librairie. Les inversions ont 2 jonctions et produisent des RP avec une orientation [-,-] à l'une et des RP [+,+] à l'autre. Les translocations se caractérisent elles par des RP inter-chromosomiques.

La problématique liée à l'identification des groupes de RP discordantes pour décrire des SV se rapporte à la théorie des graphes. Chaque RP forme le sommet d'un graphe et le chevauchement entre 2 RP décrit une arrête. L'identification des cliques maximales[2] qui forment les SV est un problème de type NP-difficile qui a été abordé de plusieurs manières dans la littérature. VariationHunter (Hormozdiari et al., 2010) a par exemple introduit une solution basée sur des graphes d'intervalles qui permet d'identifier des 'clusters' (groupes) de RP chevauchantes en temps polynomial. GASV (Sindi et al., 2009) propose une approche géométrique de la détection des SV qui permet l'application d'algorithmes utilisant une "ligne de balayage" virtuelle. De nombreux autres algorithmes proposent des implémentations moins formelles qui ne garantissent pas l'identification de l'ensemble des cliques maximales (Gillet-Markowska et al., 2015; Quinlan et al., 2010; Rausch et al., 2012a).

La méthode des lectures appariées est actuellement la méthode la plus utilisée pour détecter des SV. La démonstration de sa faisabilité remonte à 2003 lorsque Volik et collègues ont réalisé le séquençage de ~8000 BAC dans lesquels des fragments de 140kb issus de la lignée cellulaire tumorale mammaire MCF-7 étaient clonés (Volik et al., 2003). Bien que cela ne

[2] Groupe de RP toutes compatibles entre elles pour définir la même SV

représente que 0,37X la couverture du génome humain, les auteurs ont pu identifier environ 400 jonctions de réarrangements chromosomiques dont une partie a été confirmée par FISH. Ensuite, la mise sur le marché de la plateforme de séquençage 454 FLX^{MD} a révolutionné l'étude des SV en permettant de générer un grand nombre de paires de lectures, de l'ordre du million, espacées de 3, 8 ou 20kb et de longueur équivalente aux produits d'un séquençage Sanger (500 à 1000pb). L'intérêt de cette plateforme a rapidement été démontré par la cartographie de plus de 1000 SV chez 2 individus humains différents (Korbel et al., 2007). Aujourd'hui, la technologie dominante pour créer des RP est l'Illumina Paired-End (PE). Bien que cette technologie produise des lectures plus courtes que la plateforme 454 FLX^{MD} (maximum 250pb), elle permet d'obtenir un nombre de RP qui est de 2 ordres de grandeurs plus élevé (~250 millions de RP). La taille d'insert d'une banque PE classique est de l'ordre de 300 pb (annexe 1), mais un protocole alternatif appelé Mate Pair (MP) permet d'obtenir des tailles d'inserts allant jusqu'à 20kb (annexe 2). Bien que le protocole expérimental de production des banques MP soit plus complexe, ces dernières possèdent un gros avantage sur les banques PE : les SV ont tendance à produire des séquences répétées à leurs jonctions (voir chapitre II), ce qui rend difficile la cartographie des lectures dans ces régions. Ce problème peut être contourné en utilisant les banques MP qui augmentent les chances de produire des RP en dehors de ces zones répétées mais qui permettent quand même de détecter les SV. Une autre alternative est d'utiliser des algorithmes permettant d'inférer des SV à partir de lectures pouvant être cartographiées à des *loci* multiples (Hormozdiari et al., 2010). Ce type d'approches est cependant moins spécifique que les approches qui reposent sur une cartographie unique.

1.2.3.e - Limitations et approches multiples

Aucune des 4 méthodes de détection des SV évoquées précédemment n'est universelle et permet la détection de SV de tous types et de toutes tailles. Les méthodes basées sur la profondeur de séquençage sont celles qui permettent le mieux d'analyser les CNV mais elles souffrent cependant d'un manque de résolution (>1kb). Les outils qui utilisent les lectures appariées peuvent détecter tous types de SV mais sont en général très sensibles aux éléments répétés. Bien que les algorithmes de SR soient plus spécifiques que les algorithmes de RP, leur sensibilité est toutefois très dépendante de la longueur des lectures. Enfin, bien que les outils d'assemblages des génomes semblent prometteurs, une amélioration des technologies de séquençage et notamment de la longueur des lectures sera nécessaire afin

de pouvoir assembler plus facilement et plus correctement les génomes dans les régions répétées. Afin d'essayer de combiner les avantages de chaque méthode, un certain nombre d'algorithmes implémentant 2 ou 3 techniques de détection des SV ont été proposés ces dernières années (pour revue (Pirooznia et al., 2015)).

1.2.4 - Evolutions futures des techniques de détection des SV avec les nouvelles technologies de séquençage en molécule unique

La principale technologie utilisée aujourd'hui, Illumina[MD], produit avec un faible taux d'erreur un très grand nombre de séquences de petites longueurs pour un coût modéré. Les paramètres les plus intéressants à améliorer pour les biologistes sont (i) les coûts car le nombre d'expériences de séquençages réalisables en laboratoire dépend majoritairement du prix, et (ii) la longueur des séquences car plus les séquences sont longues plus elles contiennent d'information biologique. Les technologies les plus prometteuses de séquençage de troisième génération donnent accès à des séquences qui sont au moins 10 fois plus longues qu'avec Illumina® mais pour un coût encore environ 5 à 10 fois plus élevé Figure 8.

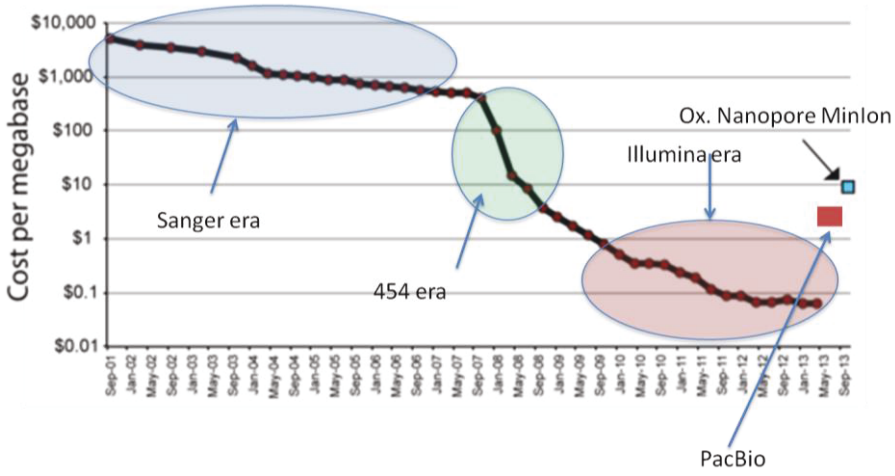


Figure 8: Coût du séquençage par mégabase depuis septembre 2001. Notons que l'échelle est logarithmique. Entre 2001 et 2007, la baisse du prix des séquences a diminué exponentiellement avec le temps. A partir de l'arrivée du séquençage NGS avec la technologie 454, la chute des prix s'est encore accélérée. La baisse du prix s'est toutefois tassée ces 4 dernières années. Il est cependant probable que nous soyons à une période charnière pendant laquelle les nouvelles technologies de séquençages de longues molécules uniques (PacBio et Minion) vont rapidement augmenter leur débit pour réduire massivement leur coût par mégabase. Crédits : http://evomics.org/2014/01/sequencing-technology-wheres-my-minion/

PacBio RSII[MD]

En avril 2013, Pacific Biosciences a commercialisé la version 2 de sa machine permettant de séquencer en temps réel des molécules d'ADN unique (pas d'amplification en cluster comme pour illumina) . Leur technologie de séquençage intitulée SMRT[MD] pour Single

Molecule, Real-Time fournit des lectures d'une longueur moyenne de 10 à 15kb avec un taux d'erreur par base de seulement 0,001 après correction (kit de biochimie P4-C4). Le débit limité de cette machine (~40 000 séquences) semble restreindre son utilisation à de petits génomes ou à des besoins spécifiques comme la transcriptomique. Cependant, au début de cette année, le laboratoire de Evan Eichler a publié un travail dans lequel ils ont séquencé en utilisant cette technologie une lignée cellulaire de môle hydatiforme (ou grossesse môlaire) avec une couverture d'environ 40X (Chaisson et al., 2015). Ces données leur ont permis de résoudre 55% des 160 trous persistants dans l'assemblage de l'euchromatine du génome humain (~1Mb), qui sont majoritairement des régions riches en GC contenant plusieurs kilobases de répétitions en tandem. L'aspect le plus remarquable de ce projet reste l'identification des SV dans le génome. En effet, l'équipe a identifié 26 079 indels (>50bp) résolus à l'échelle de la paire de base. Par comparaison, moins de 15% des SV découvertes par ce projet ont pu l'être avec la technologie Illumina classique. Ce travail démontre ainsi les bénéfices importants apportés par des lectures plus longues pour l'assemblage des régions répétées ainsi que pour la caractérisation des SV. Notons par ailleurs que des outils d'analyses de SV pour ce type de données commencent à voir le jour (Wang et al., 2015)

Oxford Nanopore

Proposé il y a 2 décennies (Kasianowicz et al., 1996), le séquençage par nanopores est en train de devenir une réalité. Le principe de la technique proposée par Oxford Nanopore est simple : un complexe de protéine est assemblé atour d'un nanopore situé dans une membrane. Une des protéines est une hélicase qui sépare la double hélice pour faire passer un ADN simple brin dans le port protéique. Un système électronique mesure ensuite la variation de courant induite par le passage de chaque base à travers le pore. Chacun des 4 types de bases possède sa propre signature électronique ce qui permet de séquencer le brin d'ADN. Des brins d'ADN de n'importe quelle longueur peuvent ainsi être séquencés. Des séquences allant jusqu'à 95kb ont été produites. Le projet est actuellement dans une phase de bêta-test publique. Au cours de cette phase, Oxford Nanopore fournit aux laboratoires un mini-séquenceur (MinION) dont le débit est relativement limité (~30 000 séquences par run). Le nombre de séquences étant directement proportionnel au nombre de nanopores, le débit des appareils pourra aisément être augmenté dans le future. Bien que le taux d'erreur par base soit encore de 10 à 15%, les premiers assemblages complets de génomes bactériens ont été rapportés (Szalay and Golovchenko, 2015) rendant cette technologie très prometteuse.

Partie 2 - Origine moléculaire des SV dans les génomes

Les cassures double brin (CDB) sont l'une des sources majeures d'instabilité dans les génomes. Les CDB de l'ADN sont en effet des lésions potentiellement toxiques pour le génome car elles interrompent la continuité physique des chromosomes. Sans réparation, toute l'information génétique allant de la cassure jusqu'au télomère est perdue. Dans certaines conditions, les CDB sont toutefois utilisées par les cellules pour promouvoir la diversité génétique, par exemple lors la méiose ou de la recombinaison V(D)J. Qu'elles soient programmées par les cellules ou issues d'évènements mutagènes, les CDB restent réparées par les mêmes mécanismes. Dans la première partie de ce chapitre, nous reviendrons sur les différentes origines des CDB dans les génomes et les voies de signalisation impliquées dans leur détection et leur prise en charge. Puis dans une seconde partie, nous évoquerons les mécanismes de réparation des CDB qui peuvent conduire à la production de SV. Notons dès à présent que même si ce chapitre ne traite pas exclusivement de la levure de boulanger (*S. cerevisiae*), il insiste plus particulièrement sur les mécanismes moléculaires existants chez cette espèce. Lorsque ce ne sera pas précisé, les protéines seront donc celles de cet organisme modèle. Dans le cas contraire, l'espèce sera précisée.

2.1 - Origines de l'instabilité chromosomique

Il existe plusieurs sources de CDB dans les génomes qui se classifient en 2 catégories : les agents exogènes et les agents endogènes.

2.1.1 - Agents exogènes

Les principaux agents exogènes pouvant être à l'origine de CDB dans les génomes sont les rayonnements (ionisant et UV) et les agents chimiques comme la bléomycine ou la camptothécine. Les rayonnements issus du soleil sont une source quasi constante de CDB dans les cellules. Ces dernières doivent donc être réparées en permanence pour maintenir l'intégrité du génome. A titre d'exemple, il est estimé qu'un individu effectuant un vol Paris-New York (~10h de vol) subit un rayonnement qui conduit à la formation de 0,05 CDB par cellules en moyenne.

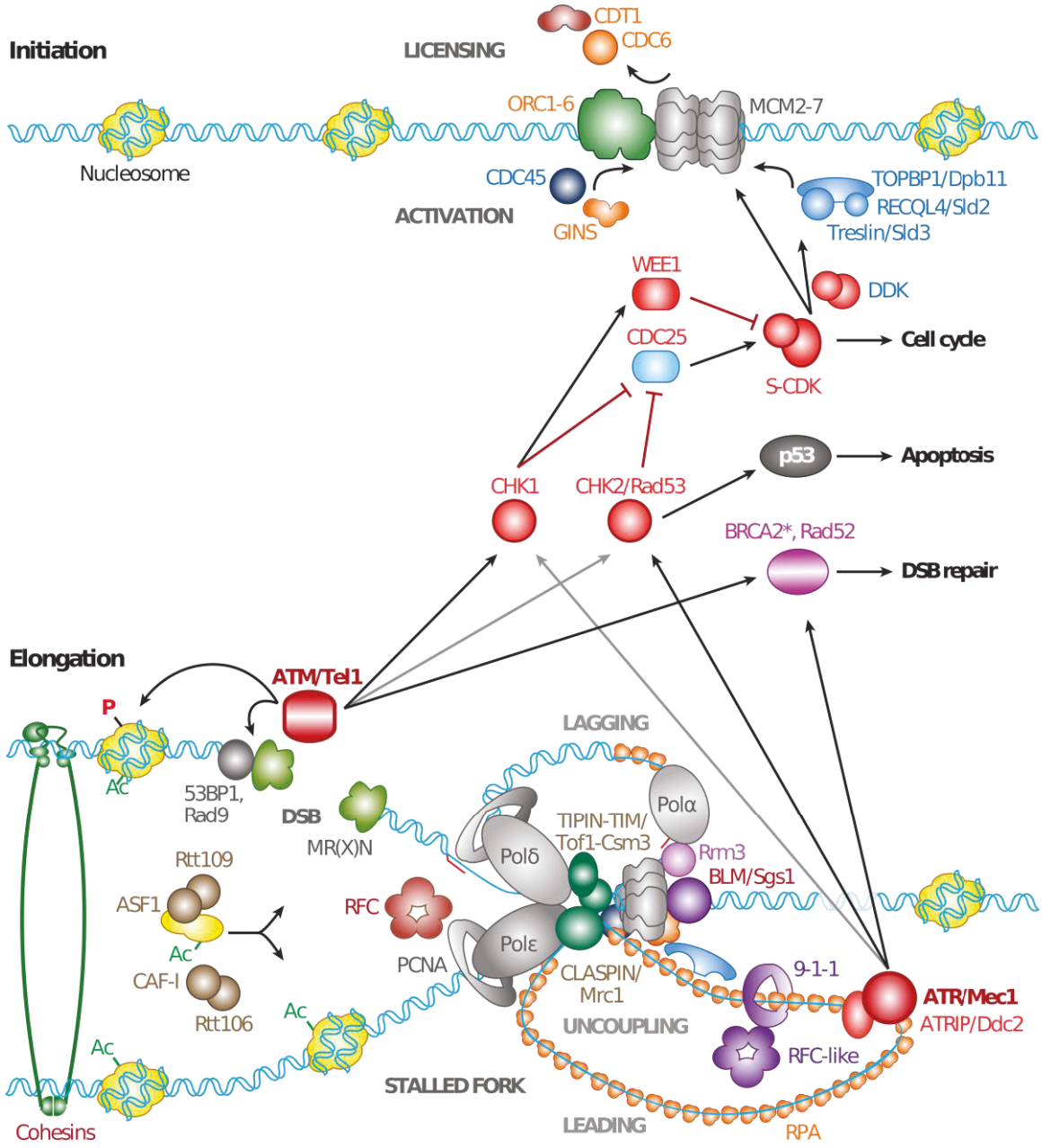


Figure 9 : Protéines impliquées dans la régulation de la réplication chez les eucaryotes et éléments contrôlant la stabilité du génome en *trans*. La formation du complexe pré-réplicatif nécessite le chargement de plusieurs protéines dont le complexe ORC, CDC6p ou encore le complexe d'hélicase MCM2-7. Le déclanchement d'une origine dépend de l'activation du complexe MCM2-7 par les CDK (cyclin dependent kinases) et les DDK (Dbf4 dependent kinases) qui déclenchent la réplication de l'ADN. Le brin direct est synthétisé par la polymérase epsilon et le brin indirect par la polymérase delta. Un arrêt anormal de la fourche de réplication peut provoquer un découplage entre la synthèse des deux brins ce qui conduits à de longues portions d'ADN simple brin sur le brin tardif qui vont être la cible de la protéine RPA dont la présence va activer la kinase Mec1. Une CDB (ici au niveau du brin précoce) provoque l'activation de la kinase Tel1. Tel1 et Mec1 phosphorylent plusieurs effecteurs qui vont activer le point de contrôle des dommages à l'ADN. Ce sont ensuite les kinases Chk1p et Rad53p qui sont responsables de l'inactivation des CDK. Crédits : Aguilera et al. 2013

Un CT-scan (imagerie médicale à rayon-X) induit quant à lui environ 0,3 cassures double

brin par cellule (Ciccia and Elledge, 2010). Lors d'une cassure double brin, 2 complexes de protéines se fixent indépendamment à la cassure : le complexe MRX et le complexe Ku (pour revue (Gobbini et al., 2013)). Brièvement, le complexe Ku induit la réparation de l'ADN via le mécanisme NHEJ et le complexe MRX permet lui la résection de l'ADN pour initier la réparation par recombinaison homologue (voir seconde partie de ce chapitre). Ces 2 complexes se fixent de manière compétitive et sont régulés positivement et négativement par plusieurs facteurs (Rad9, Sae2...), eux-mêmes régulés par des phosphorylations CDK dépendantes (Bothmer et al., 2010; Huertas et al., 2008; Wu et al., 2008). L'activité endonucléolytique du complexe MRX et de Sae2 sur le brin 5' est cruciale pour résoudre les cassures dues aux agents ionisants, à la bléomycine, la camptothécine ou les agents méthylants. Dans ces cas, des adduits protéines-ADN ou des structures d'ADN altérés doivent être enlevés pour continuer la réparation de la cassure. Une résection efficace de l'ADN nécessite ensuite le recrutement de l'exonucléase 5'-3' Exo1 et de l'hélicase Sgs1 (Mimitou and Symington, 2008). Le complexe MRX est également responsable de l'activation du point de contrôle de l'ADN Tel1-Mec1 (ATM chez l'homme). Cette voie de signalisation est à l'origine d'une cascade de phosphorylation de plus d'une vingtaine de protéines qui vont conduire à l'arrêt du cycle cellulaire et à l'activation des mécanismes de réparation (voir figure 9 et pour revue (Thompson and Schild, 2002)). Les CDB ne sont pas des évènements mutagènes *per se* mais sont la porte d'entrée des SV dans les génomes. Les voies de signalisation, de détection et de prise en charge des cassures du type de celles décrites dans ce paragraphe sont extrêmement efficaces mais il existe cependant des mécanismes de réparation non fidèles qui peuvent conduire à la formation de SV (voir partie b).

2.1.2 - Agents endogènes

Les sources endogènes d'instabilité des chromosomes sont liées principalement à la physiologie cellulaire et au déroulement de la phase S, et comprennent la réplication de l'ADN, les points de contrôles du cycle cellulaire et le modelage des nucléosomes.

Physiologie et métabolisme cellulaire

De nombreux aspects du métabolisme cellulaire incluant le pH, la pression osmotique, le groupement sulfhydryle (-SH), les espèces réactives de l'azote et de l'oxygène peuvent probablement être aussi à l'origine de l'instabilité des chromosomes. L'un des cas les mieux étudiés est celui du métabolisme de l'oxygène. L'importance des dérivés réactifs de l'oxygène sur la stabilité du génome est visible chez certains mutants de levures comme celui de la

peroxyrédoxine Tsa1. Les levures mutantes pour cette peroxydase thiol spécifique (dont le rôle principal est de détecter et d'éliminer les dérivés réactifs de l'oxygène) présentent une augmentation du taux de gros réarrangements chromosomiques (GCR[3], (Chen and Kolodner, 1999), voir chapitre III. De plus, un double mutant pour Tsa1 et pour un gène de la réparation des cassures double brin comme Mre11, Rad51 ou Rad6 est synthétique létal dans un contexte d'aérobie mais pas d'anaérobie. Ces données suggèrent donc que ces doubles mutants accumulent de trop nombreuses lésions et cassures de l'ADN pour être viables.

Vieillissement

Le vieillissement est l'un des aspects physiologiques et métaboliques majeurs qui participent à l'instabilité du génome. Le vieillissement se mesure par 2 aspects de la vie des cellules : le vieillissement réplicatif (ou RLS pour 'replicative life span') et le vieillissement chronologique (ou CLS pour 'chronological life span'). Chez la levure, le RLS correspond au nombre total de bourgeon qu'une levure mère peut produire. Chez la cellule de mammifères, il correspond au nombre maximal de divisions qu'elle peut effectuer. Il a ainsi été montré que lors du vieillissement réplicatif, le taux de perte d'hétérozygotie (LOH pour 'Loss of Heterozygoty') augmente jusqu'à 100 fois chez la levure (McMurray and Gottschling, 2003). Ces pertes d'hétérozygotie dues à l'âge ne sont pas le résultat d'une mauvaise ségrégation des chromosomes lors de la méiose mais sont causées par des recombinaisons mitotiques, suggérant ainsi que les vieilles cellules ont une capacité altérée à se répliquer leur ADN. La perte de la stabilité de l'ADN ribosomique avec l'âge des cellules est aussi cohérente avec cette hypothèse (Lindstrom et al., 2011).

Le second aspect du vieillissement des cellules est le vieillissement chronologique qui correspond au temps que des cellules peuvent passer en phase stationnaire toute en restant viables. L'accumulation d'événements de rétrotransposition de Ty pourrait notamment être un contributeur du vieillissement chronologique des cellules puisque les mutants qui réduisent la fréquence de rétrotransposition réduisent également la fréquence de LOH et d'aneuploidies liée au vieillissement (Maxwell et al., 2011). Le processus de vieillissement étant un processus complexe, il est cependant probable qu'il ait des causes multifactorielles.

[3] Classes de SV qui incluent des duplications, des extensions de répétitions, des translocations, des inversions et des délétions

Instabilité génomique liée à la phase S

L'instabilité génomique peut avoir pour origine différentes étapes du cycle de l'ADN depuis la réplication jusqu'à la ségrégation des chromosomes en mitose. Les erreurs de réplication ou de réparation sont cependant les causes les plus communes de l'instabilité chromosomique.

Dysfonctionnement du point de contrôle de phase S et de la réparation post-réplicative

L'activation du point de contrôle de phase S permet d'assurer la stabilité du génome en stoppant la phase S, via l'inhibition des origines de réplication tardives, et en activant la réparation de l'ADN avant la ségrégation des chromosomes en mitose (Bartek et al., 2004). Ce point de contrôle peut par exemple être déclenché lors du blocage de la synthèse du brin précoce. Ceci induit un découplage entre la synthèse du brin tardif et du brin précoce au niveau de la fourche de réplication (Pagès and Fuchs, 2003) qui conduit à l'accumulation de longs fragments d'ADN simple brin couverts par la protéine RPA. Cette accumulation est un signal suffisant pour déclencher le point de contrôle de phase S. Les premières protéines kinases à entrer en jeu dans ce point de contrôle sont Tel1 et Mec1. Elles régulent la sélection des mécanismes de réparation (figure 9). Ces 2 kinases régulent de nombreux effecteurs, incluant Chk1 et Rad53, qui ralentissent la réplication et inhibent les origines de réplications tardives. L'inactivation du point de contrôle de phase S provoque donc une forte instabilité génomique. Par exemple, l'inactivation des gènes codants pour les protéines Mec1, Rad53 ou Chk1 accroît les taux de GCR jusqu'à être 200 fois par rapport à ceux des cellules sauvages (Myung and Kolodner, 2002; Myung et al., 2001a).

Lorsqu'une fourche de réplication rencontre un adduit de l'ADN, la cellule peut déclencher des mécanismes de tolérance des dommages de l'ADN qui permettent à la cellule de terminer la réplication. Ces adduits peuvent en effet être répliqués par des polymérases translésionnelles (fidèles ou non) ou en effectuant un changement de matrice sur la chromatide sœur lorsque c'est possible. Les levures déficientes pour les voies de réparation post-réplicative présentent un fort taux de GCR mitotiques (Friedberg, 2005).

La recombinaison homologue est la voie de réparation majeure des cassures double brin pendant la réplication (voir partie b). Elle est nécessaire pour le redémarrage des fourches de réplication cassées et est une voie de réparation alternative en l'absence de réparation post-réplicative. La recombinaison homologue est active durant les phases S et G2 du cycle cellulaire (Heyer et al., 2010). Lorsque la recombinaison homologue n'est pas fonctionnelle, par exemple dans des mutants Rad51, Rad52 ou MRX, le taux de GCR est augmenté (Myung et al.,

2001b). Un des cas d'instabilité génomique les plus connus chez la levure et chez l'homme est celui de l'inactivation de l'hélicase Sgs1 (BLM chez l'homme) qui est impliquée dans la dissolution des jonctions de Holliday (Heyer et al., 2010).

Déclanchement des origines de réplication

Une réduction du nombre d'origines de réplication actives peut être à l'origine d'une instabilité chromosomique. Par exemple, les levures qui ne possèdent pas l'inhibiteur de CDK Sic1 déclenchent la réplication à un nombre restreint d'origines, ce qui augmente la distance inter-fourches. Ces mutants ont une phase S plus longue avec une accumulation d'ADN simple brin aux fourches et se caractérisent par un nombre élevé de gros réarrangements chromosomiques (GCR, voir chapitre 3) ainsi que des aneuploïdies fréquentes (Lengronne and Schwob, 2002). Un autre exemple classique existe chez les mammifères et leur site fragile commun FRA3B qui est l'un des plus enclins à former des CDB dans le génome. La fragilité extrême de ce site dans certains types cellulaires s'explique par l'absence d'initiation de la réplication dans une région de 700kb autour de ce site, ce qui force la fourche de réplication à parcourir une distance sept fois plus grande qu'habituellement (Letessier et al., 2011).

Le déclanchement des origines de réplication en dehors de la phase S via le chargement de l'hélicase MCM2-7 sur le complexe pré-réplicatif peut causer la re-réplication de portions de l'ADN et induire une instabilité génomique (figure 9). Chez la levure, des amplifications locales de l'ADN peuvent se produire suite à une dérégulation de Cdc6 et MCM7-2 au niveau des régions de rétrotransposition des éléments Ty. Ces évènements de rétro-transposition se produisent via des évènements de recombinaison dépendants de Rad52, ce qui suggère que la dérégulation de l'initiation des origines de réplication est propice à l'instabilité génomique et produit des cassures doubles brins(Green et al., 2010).

Progression défectueuse de la fourche de réplication

Une progression défectueuse des fourches de réplication peut conduire à la formation des cassures simple (CSB) ou double brin (CDB), des GCR et des aneuploïdies. Les premières indications qu'un contrôle génétique substantiel est exercé par la cellule pour prévenir l'instabilité génomique furent obtenues de l'étude du phénotype hyper recombinogène de mutants de la réplication chez *Escherichia coli* (Zieg et al., 1978) et *Saccharomyces cerevisiae* (Hartwell and Smith, 1985). Depuis, un très grand nombre de gènes qui affectent la progression de la fourche de réplication et le taux de GCR a été identifié. Parmi les gènes mutateurs conduisant aux plus forts phénotypes, on trouve *RAD27* qui code pour une

endonucléase impliquée dans la maturation des fragments d'Okazaki au cours de la réplication. Les mutants *rad27* ont un taux élevé de mutations et présentent de petites duplications entre de courts éléments répétés (Tishkoff et al., 1997). Ces mutants ont également un taux de recombinaison environ 100 fois plus élevé (Vallen and Cross, 1995). La mutation *rad27* associée à une mutation d'un gène du groupe d'épistasie Rad52p est synthétique létale, ce qui montre que les voies de réparation par recombinaison homologue sont nécessaires à la survie de la cellule. De plus, *rad27* est également synthétique létal avec les gènes du point de contrôle de phase S (*RAD9*, *RAD24*, *MEC3*...). Il a donc été proposé que les mutants rad27 accumulent de nombreuses cassures simple et double brin qui doivent être réparées pour éviter la létalité (Debrauwère et al., 2001; Vallen and Cross, 1995)

Chez la levure, il a été montré qu'un autre élément capable de moduler la progression des fourches de réplication est la disponibilité des dNTP. L'hydroxyurée (HU), qui inhibe la ribonucléotide réductase[4], affecte en effet brutalement la vitesse de progression des fourches de réplication. Inversement, une surexpression de la ribonucléotide réductase accélère la progression des fourches de réplication (Poli et al., 2012). Un traitement par l'hydroxyurée des levures sauvages conduit ainsi à une augmentation d'un facteur 2,5 des SV subtélomériques et cette augmentation est jusqu'à 700 fois plus élevée dans un double mutant mec1Δ sgs1Δ (Cobb et al., 2005).

La dernière source majeure d'instabilité liée à la progression des fourches de réplication est causée par les lésions ou adduits de l'ADN. Ces derniers sont provoqués par des agents exogènes (cf paragraphes précédents) ou des erreurs du NER[5] et BER[6]. Ceci est par exemple illustré chez la levure par les allèles dit *rem* pour 'recombination and mutation' qui sont des allèles semi-dominants hyper-mutateurs et hyper-recombinogènes du gène du NER *Rad3* (Montelone et al., 1988). Dans ce mutant, l'ADN simple brin généré après excision d'une base n'est pas complété efficacement, ce qui conduit à une CDB lorsque la fourche de réplication traverse ce site (Moriel-Carretero and Aguilera, 2010).

Séquences répétées

Les répétitions de l'ADN sont l'un des éléments qui favorisent le plus l'instabilité des chromosomes. Les répétitions en tandem peuvent subir de longues extensions à cause d'un

[4] Protéine qui transforme les ribonucléotides en désoxyribonucléotide

[5] Réparation par excision de nucléotides

[6] Réparation par excision de base

glissement des fourches de réplication (Pearson et al., 2005) ou bien à cause des cassures simple ou double brin (Ireland et al., 2000; Schweitzer and Livingston, 1998). Chez la levure, mais aussi chez *E. coli* et les cellules humaines, les répétitions en tandem sont difficiles à répliquer et peuvent bloquer les fourches de réplication (Miret et al., 1998). Les cellules présentant un défaut dans un des nombreux gènes de la réplication comme PCNA, Rad27 ou les polymérases sont aussi très sujettes aux contractions et expansions des répétitions.

Sites fragiles et zones de réplication lente

Les sites fragiles ont été définis cytogénétiquement comme étant des régions des chromosomes métaphasiques présentant des constrictions ou des cassures causées par l'inhibition de la réplication sous l'effet par exemple de l'aphidicoline (Durkin and Glover, 2007). Il en existe deux types : les sites fragiles rares (SFR) trouvés dans moins de 5% de la population et les sites fragiles commun (SFC). Les SFR sont soit des répétitions de trinucléotides CGG sensibles à l'acide folique soit des sites composés de minisatellites riches en AT. Les SFC ne présentent eux pas de séquences répétées mais une séquence riche en AT. Une de leurs caractéristiques est la possibilité d'induire la CDB par traitement par l'aphidicoline. Les conséquences de l'instabilité génomique des sites fragiles sont diverses et vont d'une augmentation du taux des délétions et des translocations à l'amplification de la région du site fragile (pour revue (Franchitto, 2013; Glover et al., 2005)). Les mécanismes de points de contrôles participent probablement à la stabilité de ces régions puisque la fragilité est augmentée dans les mutants Mec1, Rad9 et Rad53 (Lahiri et al., 2004). L'identification des sites fragiles chez la levure a permis de démontrer la relation entre les sites fragiles et la réplication de l'ADN. En effet, les mutants Mec1 accumulent des fourches bloquées et des cassures de l'ADN aux zones de réplications lentes où doivent se rejoindre les fourches de réplication en phase S (Cha and Kleckner, 2002). De même, une diminution du niveau de polymérase alpha induit des translocations chromosomiques et des délétions qui impliquent des évènements de recombinaison entre 2 Ty (Lemoine et al., 2005).

Structures d'ADN non conformes (différentes de l'ADN-B)

Les répétitions sont susceptibles de conduire à la formation de structures d'ADN non B comme l'ADN Z, des triples hélices d'ADN, des structures en épingles, et des G-quadruplexes. Chez la levure, le G-quadruplexe est un cas d'école en tant que structure non B connue comme étant un hotspot d'instabilité. C'est l'hélicase Pif1, en se fixant sur les G-quadruplexes

pour les dérouler, qui semble être chargée de réguler la stabilité du génome au niveau de ces structures (Ribeyre et al., 2009). Ces structures sont potentiellement des barrières à la progression des fourches, mais aussi des substrats pour des nucléases spécifiques. Par exemple, le site fragile formé des répétitions GAA/TCC chez la levure stimule fortement la formations de CDB et les réarrangements chromosomiques de type GCR (voir chapitre 3) (Kim et al., 2008). Cette augmentation de l'instabilité pourrait être causée par la formation de structures d'ADN triple brin qui sont des substrats efficaces du mécanisme de réparation des mésappariements MMR (mismatch repair).

Collision entre la réplication et la transcription

La transcription est à l'origine d'un phénomène appelé TAR pour 'transcription-associated recombination' (TAR) qui augmente la fréquence des recombinaisons mitotiques. Ce phénomène serait dû à la difficulté pour les fourches de répliquer les zones fortement transcrites provoquant alors l'arrêt des fourches et des CDB (Aguilera and Gómez-González, 2008; Bermejo et al., 2012). Chez la levure, par exemple, les promoteurs des gènes spécifiquement exprimés au cours de la phase S ont été associés à des événements de recombinaisons impliquant des répétitions de l'ADN (Prado and Aguilera, 2005). Dans cette étude, les auteurs ont en effet observé que l'instabilité est très favorisée par les collisions entre les ADN et ARN polymérases. Il a également été montre chez la levure qu'après exposition des cellules à l'hydroxyurée, les CDB se produisent préférentiellement à proximité des gènes dont la transcription est induite par hydroxyurée suggérant que ces CDB résultent de la collision des fourches de réplications avec des facteurs de transcriptions ou la machineries de transcription (Hoffman et al., 2015).

2.2 - Mécanismes moléculaires de formation des SV

Le principe de formation d'une SV dans un génome est de juxtaposer des régions chromosomiques initialement distantes. Ces changements s'opèrent globalement par 2 types de mécanismes : les mécanismes de recombinaison homologue (RH) et les mécanismes indépendants de la RH. La RH nécessite 2 éléments importants : une grande séquence d'identité d'ADN (~50 pb chez *E. coli* et *S. cerevisiae* et ~300 pb chez les mammifères) ainsi qu'une protéine catalysant les échanges de brin (*RecA* chez les procaryotes et les protéines de la famille *Rad51* chez les eucaryotes). En effet, les voies de signalisation utilisant la RH

impliquent l'invasion (catalysée par *RecA/Rad51*) d'un duplex de séquences homologues par l'extrémité 3' d'un ADN simple brin. Par opposition, les mécanismes RH indépendants utilisent uniquement de la microhomologie ou pas d'homologie du tout. Ces mecanismes RH indépendants sont soit replication-dependants comment le MMBIR ou replication-independants comme le MMEJ ou le NHEJ.

2.2.1 - Mécanismes de recombinaison homologue

La RH est à la base de plusieurs mécanismes de réparation fidèle des cassures de l'ADN qui utilisent une séquence similaire sur le chromosome homologue ou sur la chromatide sœur pour réparer la séquence endommagée. Les SV peuvent être produites par des mécanismes de RH non pas parce qu'ils sont imprécis mais parce le processus de réparation peut utiliser des séquences homologues ectopiques.

Chez la levure, le noyau peut être globalement divisé en 3 compartiments qui se comportent différemment vis-à-vis de la RH : (i) le nucléole dont de nombreuses protéines sont exclues pour supprimer la recombinaison entre les répétitions de rDNA (Torres-Rosell et al., 2007) (ii) la périphérie du noyau à laquelle la RH est généralement diminuée a l'exception des pores nucléaires ou la conversion génique est stimulée par au niveau des lésions d'ADN persistantes (Nagai et al., 2008), (iii) et le nucléoplasme qui est permissif pour la RH (pour revue (Gasser and Taddei, 2012)).

Choix du partenaire de recombinaison

La majorité des mécanismes décrits plus haut et pouvant conduire à la formation des SV résulte de l'implication d'un partenaire non allélique pour effectuer la recombinaison. Ceci est contrôlé par les cellules par différents mécanismes. Premièrement, le MMR inhibe la recombinaison entre séquences homéologues (> or = 99% identiques). Le MMR évite également l'utilisation d'un partenaire de réparation de taille réduite. De plus, la chromatide sœur (en phase S ou G2) est en général le partenaire de réparation privilégié. En effet, les cohésines, qui sont des protéines assemblées au niveau des CDB, permettent de rassembler les 2 chromatides ensemble et ainsi de faciliter la réparation des CDB (Sjögren and Nasmyth, 2001). En d'autres termes, elles permettent de limiter l'utilisation du NAHR qui est potentiellement mutagène.

Chez les mammifères, il a été montré que le choix du partenaire de recombinaison est influencé par le maintien de la proximité physique entre les 2 extrémités des cassures double brin. Ceci favorise une réparation fidèle des CDB plutôt que la formation d'une SV (Soutoglou et al., 2007).

Modèles de réparation des cassures doubles brin

Les deux modèles de réparation des CDB sont illustrés à la figure 10a : la voie de résolution des doubles jonctions de Holliday qui peut conduire soit à un enjambement (ou 'crossing-over') soit à de la conversion génique et le SDSA (pour 'synthesis-dependent strand annealing') qui ne génère pas de crossing-over.

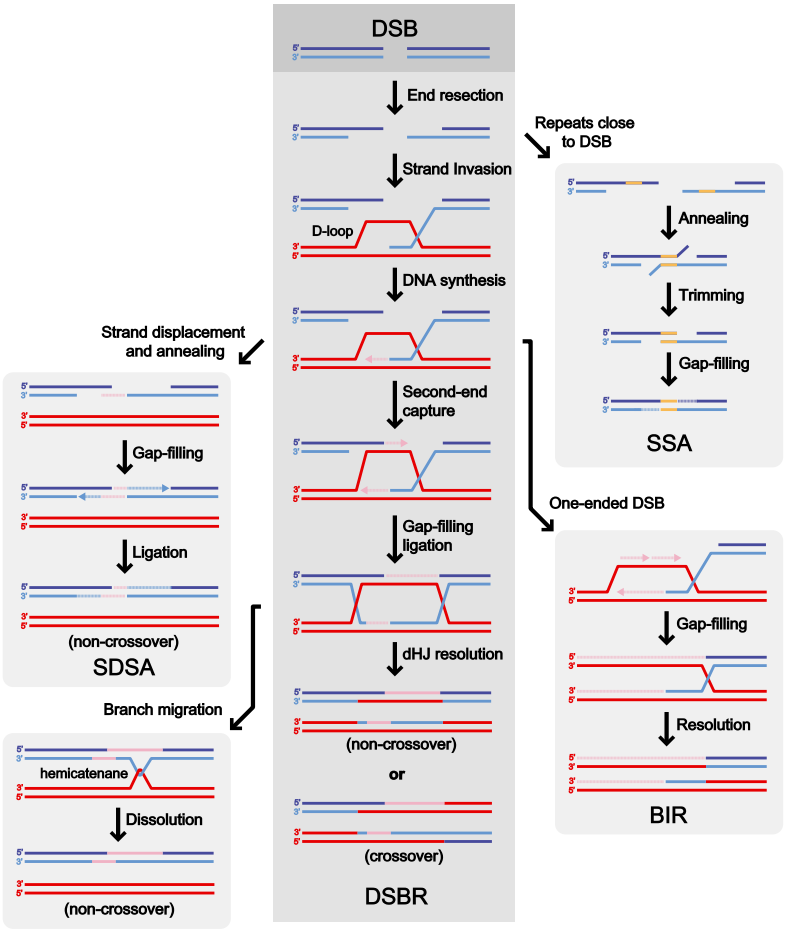


Figure 10: Mécanismes de réparation des DSB impliquant la recombinaison homologue. A la suite d'une CDB, il peut y avoir résection des extrémités 5' par le complexe MRX qui permet ensuite une étape d'invasion d'une séquence homologue et de synthèse de l'ADN. La voie classique consiste en la capture du second partenaire de recombinaison et formation d'une double jonction de Holliday qui peut être résolue avec ou sans enjambement. Dans le cas du SDSA, le brin capturé est religué avec l'autre extrémité de la CDB. En cas de cassure double brin avec une seule extrémité au niveau des fourches de réplications, la CDB est réparé par le mécanisme BIR. Enfin, la réparation par SSA se produit en cas de répétition de séquences répétées de part et d'autre de la cassure double brin. Dans ce cas les deux partenaires peuvent s'hybrider directement après résection conduisant à une délétion. Crédits : Mathiasen 2014

Le mécanisme de SDSA se produit lorsque les extrémités d'ADN simple brin envahissent transitoirement la séquence donneuse et se ré-apparient avec la séquence receveuse. Le SDSA semble être un mécanisme qui permet d'éviter les crossing-over et les pertes d'hétérozygotie. Cependant, il peut toutefois produire des duplications et des délétions quand l'ADN donneur contient des répétitions en tandem (Pâques and Haber, 1999).

Chez tous les eucaryotes, il existe un biais dans la résolution des jonctions de Holliday qui favorise la résolution sans crossing-over en mitose (Pâques and Haber, 1999). Ceci peut s'expliquer par une fréquence plus importante de SDSA en mitose. Les hélicases et topoisomérases Srs2p, Sgs1p et Top3p sont d'ailleurs connues pour favoriser la résolution sans crossing-over en déroulant la boucle de déplacement ('D-loop') (Ira et al., 2003). Un autre facteur permettant de restreindre les crossing-overs est la longueur de l'homologie impliquée dans la recombinaison homologue. Lorsque cette dernière n'est pas suffisante, la double jonction de Holliday ne peut probablement pas se former correctement (Prado and Aguilera, 2003).

Lorsque l'envahissement des deux extrémités d'ADN conduit à la formation de deux jonctions de Holliday, un crossing-over pourra avoir lieu au cours de leur résolution. Les crossing-over peuvent conduire à une perte d'hétérozygotie si les chromatides portant les mêmes allèles ségrégent ensemble lors de la mitose. Si un crossing-over se forme entre des séquences homologues ectopiques du même chromosome, on parle de recombinaison homologue non allélique ou NAHR pour 'non-allelic homologous recombination'. Le NAHR produit des délétions, des duplications et des translocations réciproques (voir figure 11). Un crossing-over intra-chromosomique entre des séquences dupliquées inversées produit lui une inversion.

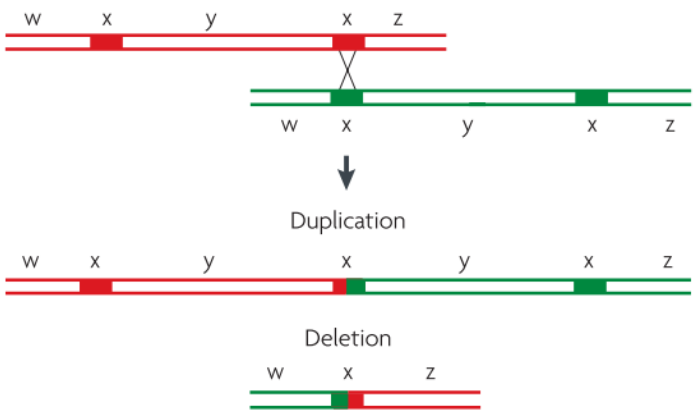


Figure 11: SV produites par recombinaison homologue non allélique (NAHR). Les NAHR peuvent conduire à des délétions et des duplications ainsi que des translocations lorsque la recombinaison a lieu entre des séquences présentent sur différents chromosomes. Crédits : Hastings et al. 2009.

Le BIR

La RH peut conduire à la reformation d'une nouvelle fourche de réplication capable de recopier le chromosome jusqu'à son extrémité. C'est le cas si une des 2 extrémités de la cassure double brin est perdue ou si au cours de la réplication une seule extrémité d'ADN double brin est formée au niveau d'une fourche de réplication cassée (figure 10). Ce processus est appelé BIR pour 'break-induced replication' (Ira and Haber, 2002; Kraus et al., 2001; Signon et al., 2001). Les étapes d'invasion de brin dépendent alors des protéines de la recombinaison homologue comme Rad51, Rad52, Rad54, Rad55 et Rad57 (Davis and Symington, 2004). Cependant, si la recombinaison se produit à partir d'une séquence homologue ectopique, différents types de SV peuvent se produire : des délétions, des duplications ou des translocations et des réarrangements complexes.

Le SSA

Après résection des 2 extrémités d'une CDB, il peut y avoir appariement des 2 ADN simple brin si une séquence répétée est présente de part et d'autre de la cassure (figure 10). Ce mécanisme appelé SSA pour 'single-strand annealing' produit des petites délétions jusqu'à 10kb maximum (Pâques and Haber, 1999). Plus la distance entre les 2 répétitions est importante, moins il est probable que la résection de l'ADN soit suffisante pour conduire à l'hybridation des 2 répétitions. Ce mécanisme doit donc jouer un rôle limité dans la formation des SV.

En plus des mécanismes de réparation utilisant la RH, il existe d'autres mécanismes de réparation des CDB qui n'utilisent pas ou très peu d'homologie. Dans leurs cas, la probabilité de former des SV est beaucoup plus élevée. Ces mécanismes se divisent en 2 catégories : ceux dépendant et ceux indépendant de la réplication.

2.2.2 - Mécanismes de réparation non-homologue et non-réplicative

Le NHEJ

Le NHEJ pour 'non-homologous end joining' ne nécessite pas d'homologie pour réparer des CDB. Ce mécanisme est le mécanisme de réparation préférentiel des CDB en G1 alors que la RH est dominante en S/G2 (Mathiasen and Lisby, 2014). Ce mécanisme permet de re-liguer les 2 extrémités d'une CDB (pour revue : (Daley et al., 2005)) (figure 12). Brièvement, après CDB, le complexe protéique Ku70 / Ku80 est recruté et recrute à son tour la ligase 4, Lif1p et Pol4p qui vont permettre la religature (Deriano and Roth, 2013). Ce mécanisme peut parfois

conduire à de petites délétions de quelques paires de bases (1-4 pb) et dans certains cas à l'insertion d'un fragment d'ADN (d'origine mitochondriale ou un transposon) à la jonction (Haviv-Chesner et al., 2007). Chez certaines espèces comme la mouche du vinaigre, l'analyse de ces jonctions indique que le NHEJ est responsable de plus de 90% des SV (Zichner et al., 2013). Chez les mammifères, le NHEJ est le mécanisme principale de réparation utilisé en phase G0 et G1 (Rothkamm et al., 2003).

Le mécanisme NHEJ intervient aussi dans le phénomène appelé *chromothripsis* qui a été découvert récemment dans des cellules cancéreuses et qui conduit à la formation de réarrangements chromosomiques massifs curieusement restreint à un ou quelques chromosomes (Stephens et al., 2011). L'analyse des jonctions suggère que les fragments chromosomiques sont raboutés par le NHEJ mais le mécanisme responsable de la restructuration complète des chromosomes est encore largement méconnu. Une étude récente rapporte cependant que ces évènements se produisent dans des structures particulières extranucléaires appelées micronoyaux (Zhang et al., 2015). Elle montre qu'en cas de mauvaise ségrégation d'un chromosome, ce dernier se retrouve enfermé dans un micronoyau. Au cycle cellulaire suivant, ce chromosome est répliqué dans le micronoyau. C'est la réplication défectueuse du chromosome à l'intérieur du micronoyau qui conduirait à de multiples cassures double brin, elles-mêmes seraient réparées via le NHEJ. Ces évènements ne sont donc pas le résultat de l'accumulation de SV au cours des divisions cellulaires mais le produit d'un seul évènement au cours d'une seule division cellulaire.

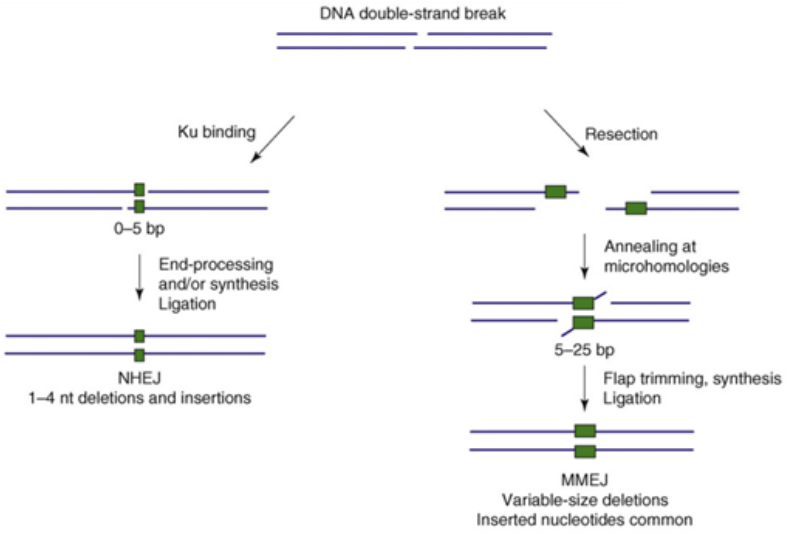


Figure 12: Réparation des CDB par le NHEJ et le MMEJ. Le NEHJ nécessite la fixation du complexe Ku qui est un inhibiteur du complexe MRX. Dans ce mécanisme, les 2 extrémités sont raboutes avec éventuellement la formation de petites délétions/insertions de quelques nucléotides a la jonction. En cas de résection des extrémités de la cassure, la présence de microhomologie de part et d'autre de la cassure peut conduire à la réparation par le MMEJ.

Un mécanisme particulier lié à la formation de réarrangements chromosomiques par le NHEJ est le cycle de cassure-fusion-pont (ou BFB pour Break-Fusion-Bridge). Brièvement, ce mécanisme résulte de la fusion de chromatides sœurs d'un chromosome ou de la fusion (probablement par le NHEJ) entre 2 chromosomes ayant perdu leurs télomères à cause d'une CDB (Figure 13). McClintock a proposé que cette fusion génère un chromosome dicentrique qui, tiré vers les 2 pôles, forme un pont anaphasique (McClintock, 1951). Ceci conduit à une cassure de ce dernier durant la ségrégation des chromosomes. Les chromosomes peuvent finalement être réparés par une nouvelle fusion avec une chromatide sœur (produit des duplications), par fusion avec un autre chromosome (translocation) ou par addition d'un télomère (le cycle s'arrête).

Ce mécanisme génère de grandes duplications inversées et des cycles répétés conduisant à une amplification de ces répétitions inversées. De grands cycles d'amplifications ont été observés dans des cellules de mammifères et ce mécanisme pourrait également jouer un rôle important dans la réorganisation des génomes tumoraux (Tanaka and Yao, 2009; Tanaka et al., 2006).

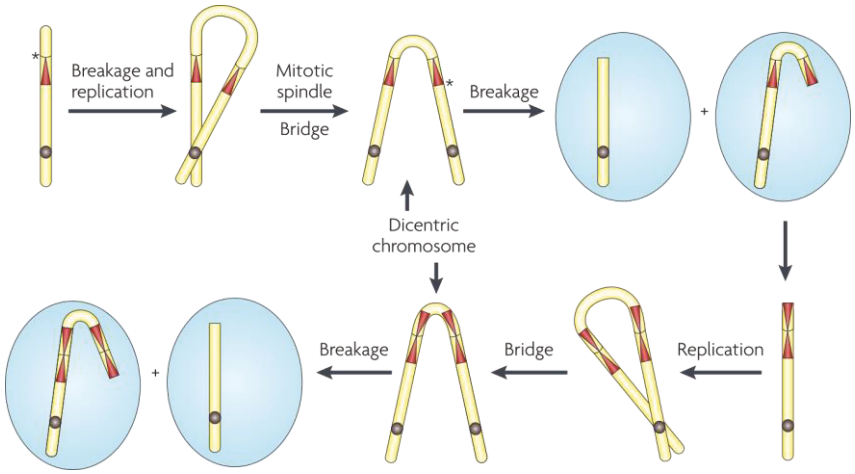


Figure 13: Cycle de cassure-fusion-pont. Le cycle peut être initié par une cassure (indiquée par la petite étoile). Les terminaisons des 2 chromosomes homologues peuvent alors fusionner et former un chromosome dicentromérique. La tension créée par la ségrégation des chromosomes peut créer une nouvelle cassure qui va provoquer une délétion dans une des cellules filles et une duplication dans l'autre cellule fille. Le cycle peut ensuite se reproduire. Crédits : Tanaka et al. 2009.

Le MMEJ

Le MMEJ pour microhomology-mediated end joining (aussi appelé aNHEJ pour alternative NHEJ) répare les CDB en utilisant de la microhomologie (5-25 pb) (pour revue (McVey and Lee, 2008))(figure 13). Le MMEJ ressemble au SSA mais contrairement à ce dernier, il ne

nécessite presque pas d'homologie, ni la protéine Rad52. De plus, contrairement au NHEJ, le MMEJ ne nécessite pas la fixation du complexe Ku pour réparer la CDB. En revanche, le MMEJ est dépendant du complexe MRX, des endonucléases Rad1 et Rad10, de Sae2, des polymérases Pol4, Rev3, Rad30 et Pol32 et est partiellement dépendant de la ligase 4 (Lee and Sang, 2007; Ma et al., 2003). L'insertion fréquente de nucléotides à la jonction lors de la réparation et la dépendance de ce mécanisme à de nombreuses polymérases suggèrent que le MMEJ a recours à différentes étapes de synthèses. Brièvement, le modèle de réparation par le MMEJ est le suivant : lors d'une CDB, le complexe MRX, Sae2 et Exo1 sont recrutés. Après résection de l'ADN, il peut y avoir directement une hybridation des 2 ADN simples brin grâce à de la microhomologie, puis une synthèse des bases d'ADN manquantes et une ligature de l'ADN. Des étapes supplémentaires de synthèse d'ADN par des polymérases translésionnelles peuvent également avoir lieu avant la religature et conduire dans ce cas à des insertions, en plus de la délétion due à la résection.

2.2.3 - Mécanismes de réparation non-homologue dépendant de la réplication

Erreurs de réplication par glissement

Les erreurs de réplication par glissement ('replication-slippage') se produisent au niveau de séquences répétées qui ont environ la même longueur que les fragments d'Okazaki (~150pb chez la levure). Lors de la réplication, ces zones forment des boucles d'ADN simple brin au niveau du brin tardif, ce qui produit des duplications de séquences (figure 13) (Albertini et al., 1982). Chez *E. coli*, les glissements de la réplication peuvent se produire en l'absence de RecA et dépendent à la fois de la longueur de l'homologie et de la distance entre les répétitions (Albertini et al., 1982; Lovett et al., 1994; Shimizu et al., 1997). Ce mécanisme est plus fréquent chez des mutants de la réplication. Ceci est probablement dû à la déstabilisation de la fourche de réplication (Bierne et al., 1997). De plus, chez des mutants du MMR, le taux de glissement de la réplication est augmenté lorsque les répétitions ne sont pas parfaites (Lovett and Feschenko, 1996). Les conditions de formation de ce type de duplication étant relativement spécifique (en termes de taille et de distance des séquences répétées), ce mécanisme est principalement à l'origine de l'instabilité des microsatellites (Schlötterer and Harr, 2001).

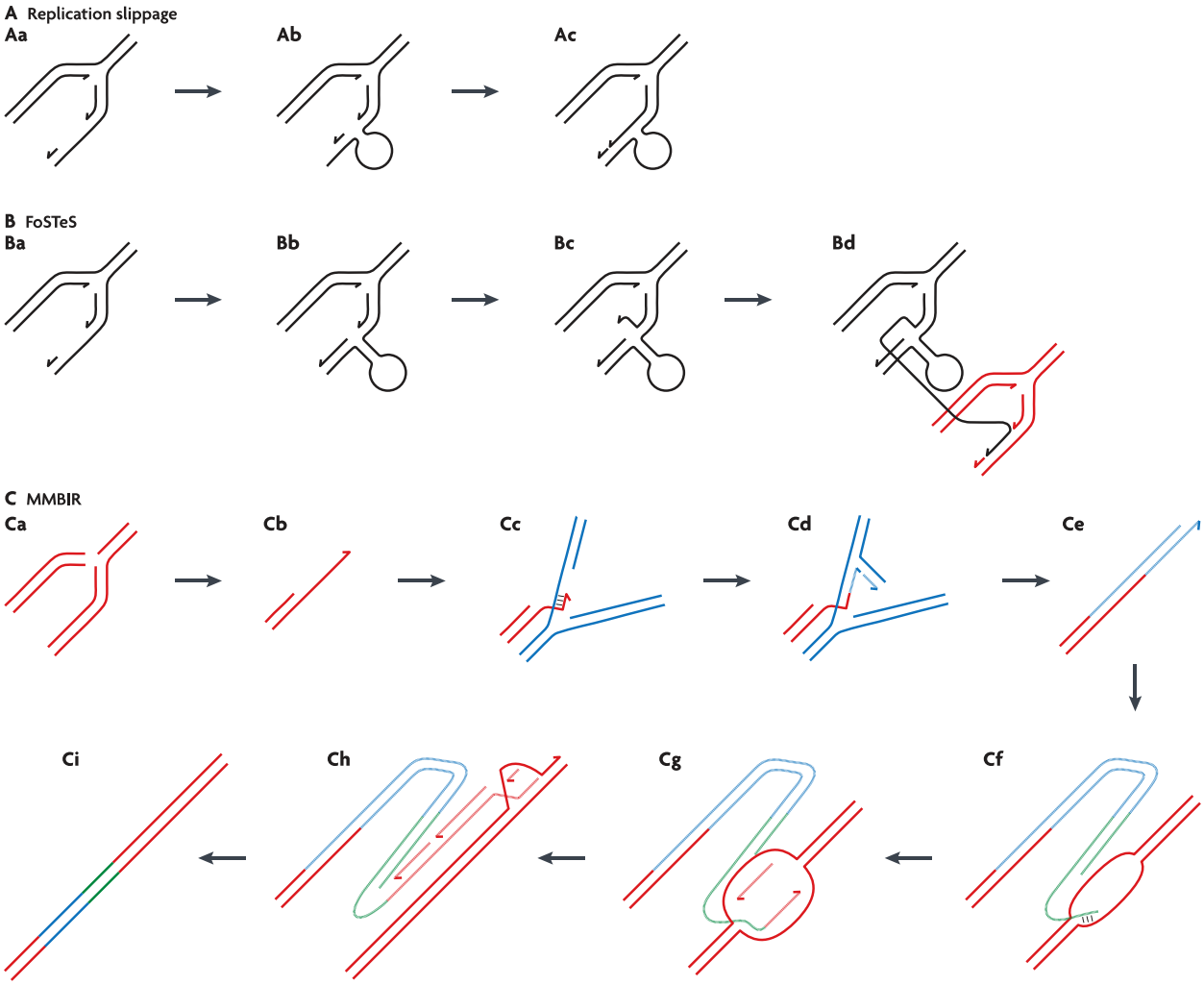


Figure 14: Mécanismes RH-indépendants de formation des SV. (A) Une erreur de réplication par glissement se produit lorsqu'un bout du brin tardif forme une boucle et que l'extrémité du primer ARN se fixe sur une autre séquence présentant un peu d'homologie. (B) Arrêt des fourches de réplication et changement de matrice : Une structure secondaire peut se former sur un segment du brin tardif ce qui peut bloquer la progression de la fourche de réplication. L'extrémité 3' ne possède alors plus de matrice et peut aller s'hybrider a une autre fourche de réplication présentant de l'homologie. (C) Lors de la cassure d'une fourche de réplication, une extrémité double brin est formée. L'extrémité de la fourche cassée va alors pouvoir s'hybrider à une séquence ectopique via la microhomologie et ce par un mécanisme Rad51 indépendant. Cette hybridation va initier la synthèse d'ADN et former une nouvelle fourche de réplication Pol32 dépendante. Crédits : Hastings et al. 2009.

Arrêt des fourches de réplication et changement de matrice (FoSTeS)

L'étude des amplifications des gènes *lac* chez *E. coli* a permis d'observer que le changement de matrice n'est pas forcement limité à une seule fourche de réplication (Slack et al., 2006). Ce mécanisme désormais appelé FoSTeS pour 'fork stalling and template switching' (Lee et al., 2007) implique l'arrêt de la fourche de réplication puis un changement de matrice d'un fragment d'Okazaki qui va envahir une autre fourche de réplication à proximité (figure 14). Slack et collègues avaient proposé cette hypothèse car la longueur moyenne des amplicons de leur étude était de ~20kb, ce qui est trop long pour une seule fourche de réplication (Slack et al., 2006). De plus, les jonctions entre les amplicons ne présentent que de la microhomologie

(4-15pb), ce qui indique que la RH n'est pas impliquée dans ce mécanisme et renforce l'idée d'un mécanisme réplicatif. Les propriétés physiques des amplicons, la microhomologie aux jonctions et la structure parfois complexe des amplicons chez *E. coli* sont également caractéristiques de nombreuses duplications et délétions chez l'homme (Zhang et al., 2009). Ceci suggère que le FoSTeS pourrait être un mécanisme important de formation des CNV dans les génomes.

Le BIR microhomologie dépendant (MMBIR, MMIR)

Un autre mécanisme similaire au BIR peut se produire par l'intermédiaire de la microhomologie (Payen et al., 2008). Ce dernier est appelé MMBIR pour microhomology-mediated break-induced replication ou MMIR pour Microhomology/ Microsatellite-induced Replication. Brièvement, lors de la cassure d'une fourche de réplication (figure 14), une extrémité double brin est formée. L'extrémité de la fourche cassée va alors pouvoir s'hybrider à une séquence ectopique via la microhomologie et ce par un mécanisme Rad51 indépendant. Cette hybridation va initier la synthèse d'ADN et former une nouvelle fourche de réplication Pol32 dépendante. L'hybridation par microhomologie peut se produire n'importe où en amont ou en aval de la fourche cassée. Ce mécanisme produit donc des délétions et des duplications, mais aussi des inversions si le brin s'hybride en sens opposé. De plus, si l'hybridation se fait sur un autre chromosome, cela conduit à des translocations non réciproques. Lorsque plusieurs hybridations suivies d'une réplication se produisent en série, cela conduit à la formation de SV complexes (Smith et al., 2007). Le MMBIR est probablement le mécanisme majeur expliquant les SV non récurrentes dans les génomes (Bauters et al., 2008), c'est-à-dire des SV qui ne sont pas produites par NAHR.

Partie 3 - Impact des variations structurelles dans les génomes

3.1 - Valeur sélective des SV dans les génomes

La valeur sélective ou fitness d'un individu décrit sa capacité à se reproduire dans la population. Dans le cas des SV, on peut résumer la valeur sélective en 2 facteurs : l'impact des SV sur le phénotype (et donc le taux de survie) et l'impact des SV sur la fécondité des individus. Nous allons analyser dans un premier temps pourquoi la majorité des SV sont délétères dans les génomes et appréhenderons ces causes a plusieurs niveau allant du niveau moléculaire jusqu'à l'étiologie des maladies génétiques chez l'homme. Ensuite, nous verrons que bien que la majorité des SV possède un phénotype délétère, ils présentent un potentiel adaptatif fort.

3.1.1 - Valeur sélective négative des SV les génomes

Altération de la quantité d'ARNm par les CNV

Une des caractéristique clé des SV est qu'elles peuvent affecter de grandes portions d'ADN et donc des unités fonctionnelles comme des gènes et des éléments régulateurs de l'expression des gènes. La large diversité des SV laisse imaginer de nombreux impacts moléculaires sur l'expression et la régulation des gènes comme ceux illustrés à la figure 15.

Les délétions et duplications de gènes ou d'éléments régulateurs peuvent directement affecter le niveau d'ARN messager (ARNm) (figure 15 c). La conséquence des CNV sur le niveau d'expression a par exemple été étudié chez la mouche (Stenberg et al., 2009), la souris (Chaignat et al., 2011) et l'homme (Vazquez-Mena et al., 2012). Ces études ont conclu qu'il existe globalement une assez bonne corrélation entre le nombre de copies d'un gène et le niveau d'ARNm. Chez la levure, le niveau de protéines est largement corrélé avec les aneuploïdies des chromosomes (Pavelka et al., 2010). Chez l'homme, l'analyse de cellules dérivées de patients présentant des aneuploïdies complètes ou partielles des chromosomes a donné les mêmes résultats (Aït Yahya-Graison et al., 2007; Ricard et al., 2010). Cependant ces résultats doivent être nuancés puisque à l'échelle des gènes chez l'homme, le niveau d'expression n'est pas systématiquement réduit de moitié lorsqu'un des 2 allèles d'un gène est

délété ni augmenté à 3/2 dans le cas des trisomies. De plus, tous les gènes ne sont pas affectés par les CNV et certains ont même un niveau d'expression inverse au nombre de copies (Aït Yahya-Graison et al., 2007), ce qui suggère l'existence de modèles complexes de régulation de l'expression des gènes (pour revue (Weischenfeldt et al., 2013)). De façon intéressante, l'étude de l'expression des gènes chez des patients porteurs de duplications ainsi que la cartographie des QTL d'expression chez la souris ont démontré que les CNV affectent non seulement l'expression des gènes mais aussi la variance de leur expression (Aït Yahya-Graison et al., 2007; Henrichsen et al., 2009). Ceci suggère que des boucles de rétroaction dont le rôle est de maintenir le niveau d'expression constant peuvent ne plus fonctionner de manière optimale dans le cas de CNV.

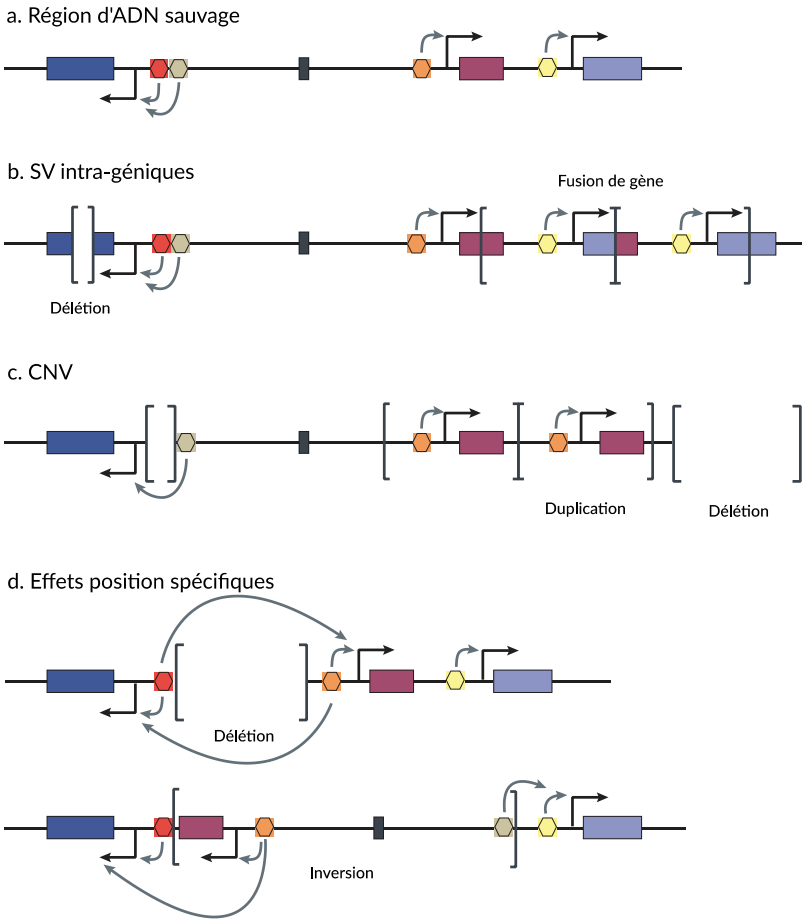


Figure 15: Conséquences fonctionnelles des SV. Les SV peuvent avoir différents effets. (b) Les délétions et les duplications peuvent altérer le dosage génique (b-d), fusionner des régions de l'ADN et créer des transcrits chimériques (b-d). Les SV peuvent également avoir un effet hors des gènes (d) en affectant des séquences régulatrices. Adapté de Weischenfeldt et al. 2013.

Altération des éléments de régulation par les SV

Les grandes délétions et les inversions peuvent également placer à proximité des gènes des séquences régulatrices qui vont perturber le programme d'expression génique (figure 15

d). L'annotation des éléments cis-régulateur dans les génomes n'est cependant pas aussi complète que l'annotation des gènes. De plus, les éléments cis-régulateurs sont fréquemment composés de multiples modules individuels qui contrôlent des aspects spatio-temporels distincts de l'expression des gènes et peuvent se trouver à des distances très grandes des promoteurs qu'ils régulent (pour revue (Wittkopp and Kalay, 2011)). Ceci peut expliquer l'impact tissus/cellules-spécifique des SV évoqué précédemment.

Bien que l'effet des SV est en général plus fort lorsque ces dernières incluent des gènes, beaucoup de QTL d'expression sont attribuables à des SV situées dans des régions inter-géniques (Schlattl et al., 2011; Stranger et al., 2007; Yalcin et al., 2011). Ceci suggère que les SV peuvent altérer la régulation des gènes en modifiant leur régulation. D'ailleurs, de nombreux eQTL sont couramment identifiés comme étant attribuables à des SV cellule-spécifiques, tissus-spécifiques ou spécifiques de certaines étapes du développement (Cahan et al., 2009; Chaignat et al., 2011; Henrichsen et al., 2009) ce qui conforte l'hypothèse d'une expression des gènes contexte dépendante.

La β-thalassémie qui est causée par la délétion du locus du contrôle de la β-globuline (Kioussis et al., 1983) et l'holoprosencéphalie[7] qui est due à une translocation qui sépare le gène SSH de son élément régulateur contrôlant son expression durant le développement du prosencéphale (Belloni et al., 1996) sont des cas d'écoles de l'altération d'éléments régulateurs. La surexpression d'un gène à cause d'une SV a également été associée à des maladies génétiques comme l'hypoplasie congénitale unilatérale des doigts. Dans ce cas, il a été montré que la morphogénèse des doigts requière un niveau précis des protéines de la voie de signalisation BMP et que la duplication du gène *BMP2* provoque ainsi des malformations (Capdevila and Belmonte, 2001; Dathe et al., 2009).

Les modèles moléculaires évoqués dans les paragraphes précédents ne constituent cependant que des hypothèses et il est probable que les SV conduisent à des phénomènes plus complexes par opposition à une simple hausse ou baisse d'expression. En effet, l'impact fonctionnel réel des SV dans les génomes est ainsi encore controversé dans la littérature. Certaines études ont par exemple rapporté que les SV interviennent davantage que les SNP sur les différences phénotypiques entre individus (Conrad et al., 2010; Keane et al., 2011). D'autres études ont identifié quant à elles des effets très variables des SV sur la présence d'eQTL (Cahan et al., 2009; Henrichsen et al., 2009; Stranger et al., 2007; Yalcin et al., 2011).

[7] Malformation cérébrale complexe due à un défaut de clivage médian du prosencéphale, à l'origine de manifestations neurologiques et d'anomalies faciales de degré variable.

De plus, ces modèles sont soutenus par des données expérimentales qui restent encore relativement restreintes. Par exemple, nos connaissances sur l'haplo-insuffisance ou le dosage génique sont encore limitées. Des cribles chez la levure de boulanger (Deutschbauer et al., 2005), la drosophile (Lindsley et al., 1972) et le nématode (Hodgkin, 2005) suggèrent ainsi que la proportion de gènes affectés par les CNV en temps normal est relativement faible (~3% chez la levure et la mouche). Cependant cette proportion peut atteindre 12 à 20% chez la levure en cas de carence en nutriments (Delneri et al., 2008) sans que l'on sache exactement comment.

Impact des CNV sur le phénotype des individus

Les paragraphes précédents suggèrent que bien que les duplications ne provoquent pas de perte de matériel génétique comme les délétions, il est possible que la plupart d'entre elles soient délétères. Ceci est notamment illustré par les exemples de maladies génétiques associées aux duplications (cf. paragraphes suivants et pour revue (Girirajan et al., 2011)). En parallèle, d'autres résultats tendent vers l'idée que les duplications ont une valeur adaptative négative. La forte différence entre les taux de duplications calculés empiriquement avec les lignées accumulatrices de mutation et ceux calculés par les méthodes bio-informatiques (voir partie « Calcul des taux de SV dans les génomes ») peut s'expliquer par une sélection négative des SV (Katju et al., 2009; Lipinski et al., 2011). De plus, chez la drosophile, le taux de duplication dans le génome est plus faible qu'attendu sous l'hypothèse que les duplications seraient neutres (Langley et al., 2012). Ces données suggèrent donc que les problèmes dus au dosage génique jouent un rôle majeur dans le coût en fitness des duplications.

Il est en général admis que les délétions sont plus nocives que les duplications. En effet, plusieurs études des CNV à large échelle chez l'homme ont mis en évidence que les délétions se produisent moins fréquemment que les duplications (Conrad et al., 2006; Locke et al., 2006). Ceci suggère que la délétion de séquences codantes est plus nocive pour le fitness cellulaire et qu'elle est sous forte sélection purificatrice. Conrad et collègues ont comparé les fréquences relatives des délétions chez l'homme dans les introns et les séquences inter-géniques (Conrad et al., 2010). Ils ont observé un fort déficit des délétions introniques suggérant qu'elles sont contre-sélectionnées en raison de leur rôle important dans l'épissage des gènes. Ces résultats expliquent peut-être pourquoi les délétions sont également détectées moins fréquemment que les duplications dans les expériences d'accumulation des mutations chez la levure et le nématode (Lipinski et al., 2011; Lynch et al., 2008). En effet, bien que l'effet

de la sélection soit limité au maximum dans les MAL, les mutations les plus délétères ont toujours moins de chances d'être fixées lors de chaque goulot d'étranglement.

Impact des SV balancées sur l'isolement reproductif des individus

De nombreuses données indiquent que les SV balancées comme les translocations réciproques et les inversions jouent un rôle majeur de barrière reproductive dans les populations en réduisant la recombinaison entre les hétéro-caryotypes (Delneri et al., 2003; Hou et al., 2014; Schaeffer et al., 2003).

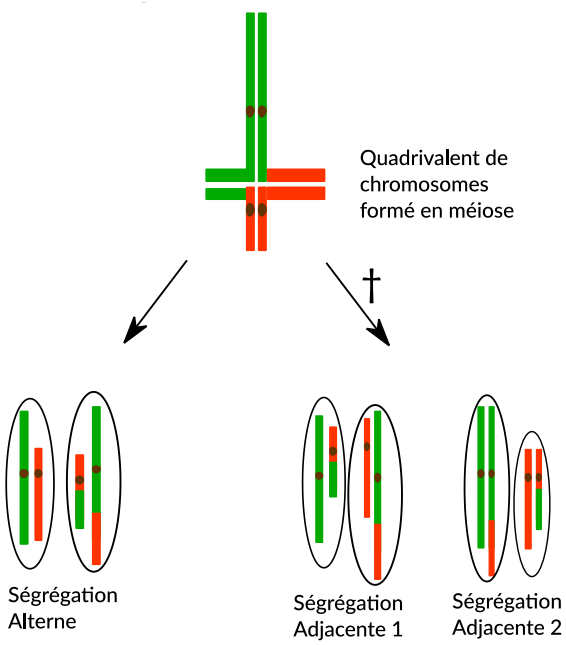


Figure 16: Gamètes résultant de la ségrégation méiotique de chromosomes ayant subi une translocation réciproque. Dans ce cas les chromosomes homologues ne peuvent pas s'apparier correctement en méiose et forment une structure quadrivalente. La ségrégation alterne produit des gamètes viables. En revanche les ségrégations adjacentes conduisent a une perte d'information génétique et ne sont donc pas viables. Des ségrégations 3 :1 peuvent également être observées et créer des aneuploïdies.

Par exemple, les translocations réduisent la viabilité des gamètes de 50% par translocation (Loidl et al., 1998). Cette baisse s'explique par la formation en méiose de structure de chromosomes quadrivalents dont la ségrégation, avec ou sans perte d'information génétique, définit la viabilité des gamètes (figure 16). Cependant, alors qu'on s'attend à ce qu'elles soient fortement contre-sélectionnées, les translocations réciproques sont relativement courantes chez les plantes et ont été identifiées dans de nombreuses espèces de l'agriculture comme le seigle (Benito et al., 1994), le soja (Mahama and Palmer, 2003), le colza (Osborn et al., 2003), le pêcher (Jáuregui et al., 2001)... Dans la quasi-totalité des cas, ces translocations ont été associées à une forte baisse du rendement et de la fertilité des plantes. Les gamètes issus de ségrégations adjacentes conduisent à la formation de graines ou de pollen stériles (Jáuregui et

al., 2001; Kakeda and Miyahara, 1995). De plus, les translocations détectées jusqu'ici l'ont été dans le cadre de l'analyse du rendement. En parallèle, un seul exemple de translocation neutre (chez l'orge) a jusqu'ici été identifié (Farré et al., 2012). Cinq translocations ont également été identifiées dans une souche naturelle Malaisienne de S. cerevisiae qui réduit la viabilité des spores issues des hybrides avec d'autres souches a environ 3% (Marie-Nelly et al., 2014). Une translocation réciproque qui provoque une augmentation de la transcription du gène SSU1 a également été identifiée dans une souche du vin et des souches naturelles (Hou et al., 2014).

Cancers et maladies génétiques chez l'homme

Comme nous l'avons vu dans le chapitre I, les maladies génétiques et le cancer illustrent particulièrement bien l'impact phénotypique négatif des SV dans les génomes. En effet, les SV sont la source de nombreuses maladies génétiques rares comme la maladie de Charcot-Marie (Lupski et al., 1991), les syndromes de Koolen de Vries et de la microdélétion 16p11.2 (Koolen et al., 2012; Walters et al., 2010), l'alpha-thalassémie (Flint et al., 1986) ou encore la dystrophie musculaire de Duchenne (Francke et al., 1985). Le séquençage des génomes de familles dont un membre est atteint par une maladie génétique a montré que la plupart des SV formées *de novo* ont curieusement une origine paternelle. Il existe une forte corrélation de cette origine avec l'âge du père (Batista et al., 1994; Hehir-Kwa et al., 2011).

Chez l'homme, les maladies où la prévalence des SV est la plus forte sont les cancers. Ces derniers se caractérisent par un processus dynamique de remaniement du génome qui se traduit par le gain ou la perte de chromosomes entiers ou par des remaniements de leur structure. La majorité des cancers sont par conséquent des maladies complexes qui impliquent de nombreux évènements mutationnels (figure 17). Historiquement, le passage de la période de la cytogénétique, où les chercheurs cherchaient à identifier des SV macroscopiques dans les cellules tumorales, à une période pendant laquelle on a cherché « les gènes du cancer » a été déclenché par l'identification du chromosome de Philadelphie (figure 3). La formation de ce chromosome est le résultat d'une translocation qui conduit à la formation d'un gène de fusion BCR-ABL qui est responsable de la leucémie myéloïde chronique (Rowley, 1998). L'ère des « gènes du cancer » est pourtant un échec important de la génétique des 10 dernières années. Aujourd'hui, après le séquençage de milliers de tumeurs, nous constatons que la vaste majorité des mutations ponctuelles détectées ne sont pas partagées entre les patients (Pleasance et al., 2010; Stratton et al., 2009; Yates and Campbell, 2012). D'ailleurs, ce résultat aurait pu être prédit dès le début de l'ère du séquençage des génomes tumoraux, lorsque l'on a réalisé que

seule une infime proportion des cancers peut s'expliquer facilement par un schéma d'acquisition de mutations en série comme dans les leucémies myéloïdes chroniques (Gabor Miklos, 2005). Par opposition, la plupart des tumeurs solides semble évoluer de façon aléatoire, avec des altérations du caryotype ponctuées dans le temps (Heng et al., 2006).

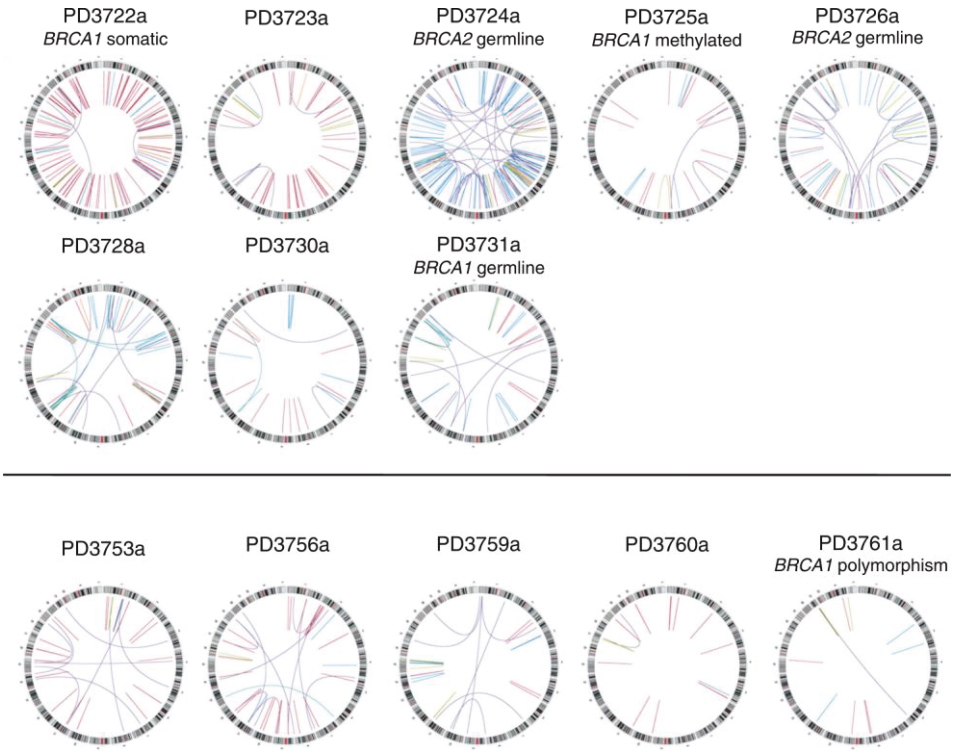


Figure 17: SV dans les génomes de 13 cas de cancers de l'ovaire. Les chromosomes du génome de référence sont dessines en cercles. Les SV sont représentées par des lignes qui connectent leurs jonctions. Lignes orange : inversions, lignes bleues : duplication, lignes violettes : translocations, lignes rouges : duplications en tandem, lignes bleu foncées : délétions. Crédits : Mc Bridge et al. 2012

La majorité des cancers chez l'homme se caractérise donc par une forte instabilité chromosomique (ou CIN pour 'chromosomal Instability'). La variété des SV décrites dans les cellules cancéreuses est extrêmement étendue et inclue les inversions, les translocations, les duplications, les délétions et les réarrangements complexes. Ce processus d'instabilité fait donc l'objet d'une attention particulière dans le domaine clinique. C'est un important contributeur à la transformation des cellules malignes et à la constitution d'une hétérogénéité intra-tumorale, dont les conséquences cliniques néfastes sur la résistance thérapeutique et le pronostic vital sont bien connues. Cependant, les SV qui sont observées dans chaque cancer sont souvent de nature différente. Les cancers du sein et des ovaires présentent par exemple un taux de duplication segmentale beaucoup plus élevé que n'importe quel autre type de tumeurs (McBride et al., 2012; Ng et al., 2012). Les cancers du pancréas sont caractérisés par des cycles de cassure-fusion-pont très fréquents (Campbell et al., 2010), les cancers de la prostate montrent eux un fort taux de réarrangements balancés complexes (Berger et al., 2011).

Certains autres cancers comme les sarcomes et les neuroblastomes montrent une forte fréquence de chromothripsis (Molenaar et al., 2012; Rausch et al., 2012b). De plus, les déterminants génétiques qui favorisent l'acquisition spécifique d'un type particulier de réarrangements sont encore largement méconnus. En 2011, la première étude de séquençage de cellules uniques dans un cancer (Navin et al., 2011) fut publiée. Dans cet article les auteurs se sont notamment intéressés à la progression tumorale dans le cancer du sein. Ces données montrent que dans ce type de cancer, les SV font irruption soudainement dans le génome et que ces évènements de réarrangements sont suivis par des périodes stables puis d'expansion de la tumeur. Une autre étude utilisant le séquençage de cellules uniques a permis de caractériser la réponse au traitement par l'Abiratérone[8] chez des patients atteints de métastases de cancer de la prostate (Dago et al., 2014). Elle a notamment démontré que ce n'est pas un seul mais plusieurs clones indépendants qui aboutissent à une amplification de certains gènes comme l'oncogène MYC, par des CNV différents.

3.1.2 - Potentiel adaptatif des SV

Bien que les SV produisent un réagencement des chromosomes qui est en général délétère, certaines SV peuvent, dans des conditions environnementales particulières, conduire à une diversification biologique. Parmi les cas les plus célèbres de SV avec une valeur adaptative positive chez l'homme, on retrouve celui de l'augmentation du nombre de copies de l'amylase[9] AMY1 qui a permis une adaptation à un régime riche en amidon (Perry et al., 2007). De façon intéressante, la domestication du chien par l'homme s'est également accompagnée d'une augmentation du nombre de copies du gène AMY1 canin permettant aux chiens de bénéficier d'un régime riche en amidon, contrairement à son ancêtre le loup (Axelsson et al., 2013). L'augmentation du nombre de copies du gène CCL3L1 qui provoque une forte baisse de la susceptibilité à l'infection par le virus HIV chez l'homme est un autre exemple célèbre de SV avec une valeur adaptative forte (Gonzalez et al., 2005). Pour finir, l'augmentation du nombre de copies de segments entiers d'ADN a aussi été associée à l'adaptation des microorganismes à de nouveaux environnements ou à des milieux où les ressources sont limitées (Payen et al., 2013; Sonti and Roth, 1989), à la résistance des insectes aux insecticides

[8] Inhibiteur sélectif de la biosynthèse des androgènes

[9] Constituant du suc pancréatique et de la salive, requis pour le catabolisme des glucides à longue chaîne comme l'amidon

(Newcomb et al., 2005) et leur tolérance au métal (Maroni et al., 1987), à la résistance des parasites aux médicaments (Nair et al., 2007), et à l'augmentation de la résistance des vertébrés aux pathogènes bactériens (Jackson et al., 2007).

Bien que les délétions soient généralement nocives pour le fitness des cellules, il semble toutefois qu'elles puissent également jouer un rôle important dans l'adaptation. Par exemple, des expériences chez la salmonelle ont montré que de nombreuses délétions provoquent une augmentation de la vitesse de croissance, ce qui suggère que de nombreux gènes présentent une charge négative pour le fitness des cellules en conditions de laboratoire (Koskiniemi et al., 2012). Toujours chez la salmonelle, il a été montré que la purge des pseudogènes dans le génome se produit à une fréquence plus rapide qu'elle ne devrait si l'on considère que les délétions sont neutres (Kuo and Ochman, 2010). Pour finir, l'étude des bactéries du genre Methylobacterium a conclu que la perte de gènes peut être adaptative (Lee and Marx, 2012). Les délétions fréquentes de gènes chez les bactéries a d'ailleurs conduit à l'élaboration de la théorie de la « Reine Noire » qui stipule que les gènes dont le produit peut être facilement acquis à partir d'autres organismes sont contre sélectionnés (Morris et al., 2012).

Chez les plantes, l'analyse de l'enrichissement des GO termes[10] chez l'orge et le soja a démontré que les CNV sont plus fréquentes dans les gènes appartenant aux catégories de type 'mort cellulaire' et 'modification des protéines' (McHale et al., 2012; Muñoz-Amatriaín et al., 2013). La majorité de ces gènes de type 'mort cellulaire' sont des gènes de type 'R' codant pour des protéines de type NBS-LRR (nucleotide-binding site leucine-rich repeat) et qui sont impliqués dans la résistance aux maladies, dans la reconnaissance et dans l'initialisation de la réponse aux pathogènes (Dangl and Jones, 2001; Eitas and Dangl, 2010). La catégorie 'modification des protéines' compte également des gènes de type 'R'. Les zones enrichies en CNV, associées aux gènes de types 'R', constituent ainsi des clusters de gènes de résistances aux pathogènes (Chełkowski et al., 2003). On peut citer notamment les mécanismes de résistance à la Rouille (Collins et al., 2001), à l'échaudure des arbres (Garvin et al., 2000) ou encore au mildiou (Wei et al., 2002) qui reposent sur ce type de loci. Dans le cas du mildiou, le locus contient 32 gènes. La colocalisation des CNV avec ces mécanismes de défense contre les pathogènes de la plante suggère que les CNV jouent un rôle dans la régulation des voies de signalisation de résistance aux pathogènes en modifiant le niveau d'expression des gènes.

[10] Description structurée des gènes et des produits géniques dans le cadre d'une ontologie commune à toutes les espèces

La drosophile fait partie des premiers organismes chez qui le potentiel adaptatif de SV a été démontré grâce à l'étude de la valeur adaptative des inversions. Depuis les travaux pionniers de Dobzhansky, nous savons que la fréquence des inversions dans le génome de la mouche est corrélée aux transitions de latitude (ou clines) et de saison, ce qui suggère qu'une pression de sélection intense est maintenue sur ces inversions (Dobzhansky, 1947; Schaeffer, 2008). De nombreux modèles expliquant le maintien des inversions dans les populations ont été proposés. Ces derniers reposent sur le fait que les inversions suppriment la recombinaison entre les régions réarrangées et permettent donc de maintenir un déséquilibre de liaison favorable à certaines combinaisons d'allèles (Hoffmann and Rieseberg, 2008; Kirkpatrick and Barton, 2006). Cet effet des inversions sur la protection des allèles est par exemple illustré chez le papillon *Heliconius numata*. Les diverses populations de cette espèce possèdent différentes inversions d'un superlocus de 400kb qui produit un fort déséquilibre de liaison figure 18. Ce mécanisme procure un avantage sélectif fort puisqu'il permet de maintenir un schéma de coloration des ailes proche de celui d'un autre papillon qui, lui, est toxique pour les prédateurs (Joron et al., 2011).

Des expériences d'évolution à court terme chez la levure ont permis d'isoler des souches transloquées dans différents contextes d'étude (Adams et al., 1992; Dhar et al., 2011; Dunham et al., 2002; Koszul et al., 2004). L'un des exemples les plus marquants est l'identification chez une souche du vin et d'autres souches naturelles d'une translocation causant la surexpression du gène SSU1 (qui code pour une pompe à sulfite) grâce à la création d'un nouveau site de fixation des facteurs de transcription (Hou et al., 2014; Pérez-Ortín et al., 2002). Nous avons expliqué dans la partie précédente que les SV balancées comme les inversions et les translocations réciproques peuvent conduire à la réduction de la fertilité des individus et à une diminution de la recombinaison méiotique, ce qui conduit à une isolation reproductive. Bien que la majorité des SV soit délétères, il a été montre qu'elles peuvent jouer un rôle crucial dans l'adaptation à des conditions environnementales spécifiques (Anderson et al., 1991; Joron et al., 2011; Lyon, 2003; Stefansson et al., 2005). Il a par exemple été montré chez la levure que le coût de ces SV en méiose peut être compensé par l'apport d'un fort avantage sélectif pour la croissance végétative dans certaines conditions (Avelar et al., 2013).

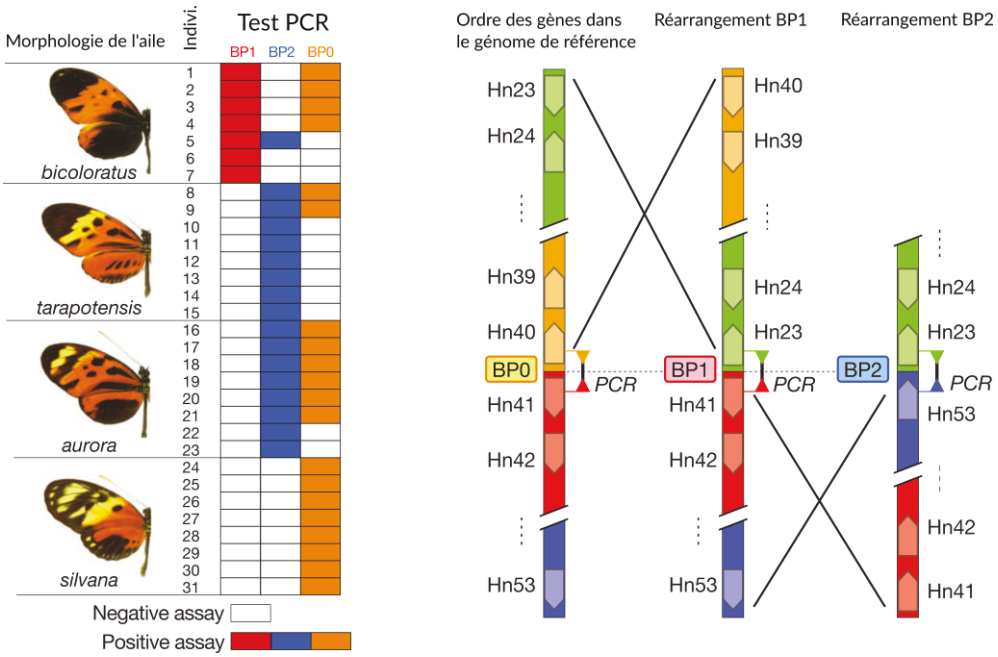


Figure 18: Inversions au superlocus du papillon *Heliconius numata* associées aux dessins des ailes. Le schéma de gauche représente les résultats d'amplification par PCR de la jonction de l'inversion BP0, BP1 et BP2 dans 31 papillons classes en fonction des dessins des ailes. Chaque schéma de coloration d'aile est associé à une combinaison particulière de ces inversions. A droite, structure des *loci* testés par PCR. Adapté de Joron et al 2011

La délétion qui provoque la bêta-thalassémie chez l'homme confère une résistance au paludisme malgré le phénotype délétère évident de cette délétion, ce qui explique le maintien de cet allèle dans les populations endémiques pour le paludisme. Ces exemples permettent de voir les SV non pas sous le jour de leur caractère délétère mais plutôt de leur potentiel adaptatif. (Marroni et al., 2014) décrivent le maintien de la plasticité chromosomique dans les populations comme un moyen d'innovation constant (figure 19). Dans ce modèle, les génomes sont considérés comme des entités nécessitant des éléments de bases (le « Core Genome ») pour pouvoir fonctionner. C'est l'équivalent des gènes essentiels chez la levure. Une seconde partie des génomes est superflu dans la majorité des conditions expérimentales mais fournit un potentiel adaptatif grâce au remodelage chromosomique des SV. C'est l'équivalent des régions subtélomériques chez les levures. Cependant, les déterminants génétiques de la plasticité des génomes sont encore largement inconnus et l'existence d'une pression de sélection qui favoriserait le maintien de cette plasticité n'a pas été démontré. L'étude du « fitness » des SV nécessite cependant de pouvoir détecter l'ensemble des SV qui se produisent dans les génomes. Nous analyserons donc dans la partie suivant l'impact quantitatif des SV dans les génomes.

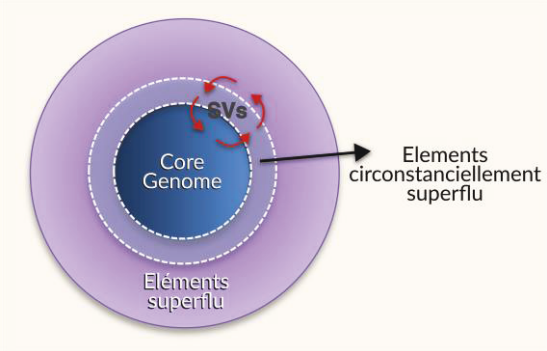


Figure 19: Les éléments essentiels au fonctionnement d'un organisme forment le « core genome ». La distinction entre le core génome et le reste du génome n'est cependant pas immuable. Les SV peuvent permettre aux organismes de s'adapter aux conditions environnementales grâce par exemple aux variations du nombre de copies.

3.2 - Impact quantitatif des SV dans les génomes

3.2.1 - Les SV, une source majeure de polymorphisme dans le génome humain

Bien que l'étude des maladies génétiques et du cancer représentent un attrait médical évident, ils ne représentent en réalité qu'une faible partie du polymorphisme des SV du génome humain. L'avènement des puces à ADN a permis de mesurer le polymorphisme très important généré par les SV dans le génome humain (Abecasis et al., 2010; Iafrate et al., 2004; Korbel et al., 2007; Redon et al., 2006; Sebat et al., 2004; Tuzun et al., 2005). Ainsi les premières analyses du polymorphisme chez l'homme montrèrent qu'environ 12% du génome est susceptible de comporter des CNV.

L'une des premières études à avoir démontré à très large échelle l'impact des SV sur le polymorphisme dans les populations humaines reposait sur l'analyse de 2500 individus sains ou malades avec des puces SNP Illumina^{MD} (Itsara et al., 2009). En faisant l'hypothèse que ces individus étaient représentatifs de la population générale, cette étude a fourni un aperçu de la taille et de la fréquence des SV. Cette étude estimait le nombre moyen de CNV par personne entre 3 et 7 CNV, ce qui représenterait environ 540 kb en moyenne. De plus, 65 à 80% de la population compterait au moins un CNV de plus de 100kb, 5 à 10% compterait un CNV d'au moins 500kb et 1% de la population serait porteuse d'un CNV d'au moins 1Mb. La très vaste majorité de ces grand CNV sont très rares dans la population (<< 1%), ce qui est certainement lié à leur désavantage sélectif important.

Le nombre de SV détectés jusqu'à présent dans la population humaine est très important. La base de données DGV (Database of Genomic Variants) recense 3 026 547 SV d'après 62 études dont une écrasante majorité est constituée de CNV. En effet, la détection des CNV est plus facile et remonte à plus longtemps grâce aux puces à ADN. Un des premiers projets à avoir détecté des SV à large échelle, avec une résolution suffisante pour analyser la séquence des jonctions et modéliser les mécanismes moléculaires de leur formation, est le '1000 Human genome projet' (Abecasis et al., 2010). Cette étude, pour laquelle plusieurs milliers de génomes ont été séquencés, a ainsi permis de déterminer que 89% des délétions se produisent par des mécanismes qui ne dépendent pas de la recombinaison homologue, la majorité d'entre elles ne présentant aucune homologie aux jonctions. En revanche, les insertions et les duplications (en dehors de la mobilité des transposons de type Alu) se produisent à 52% via le NAHR et à 28% par des phénomènes d'amplifications des répétitions de type 'replication slippage'.

Aujourd'hui, la formation *de novo* des SV dans les génomes est généralement étudiée grâce aux techniques de séquençage haut-débit. Chez l'homme, une approche classique consiste à séquencer des trios familiaux père-mère-enfant. Les études récentes (Kloosterman et al., 2015) s'efforcent de multiplier les signaux de détection des SV et de valider indépendamment leur existence (PCR, Sanger, MiSeq, IonTorrent). Kloosterman et collègues ont ainsi détecté 41 SV dans 250 familles (231 trios, 8 quartets dizygotes (parents + jumeaux dizygotes), 11 quartets monozygotes (parents + vrai jumeaux)).

Le polymorphisme chez l'homme est également détectable à l'échelle des cellules et des tissus. La plasticité des chromosomes a par exemple été étudiée à l'échelle de la cellule unique dans le sang (Jacobs et al., 2012; Laurie et al., 2012). Dans ces travaux les auteurs ont montré que la quantité de réarrangements chromosomiques est positivement corrélée à l'âge. Une autre étude de séquençage de cellule individuelles a quant à elle démontré qu'environ 7% des spermatozoïdes sont aneuploïdes chez l'homme (Wang et al., 2012). Enfin, D'autres études de ce type ont permis de révéler que jusqu'à 80 évènements de rétrotransposition ont lieu par neurone et que 15 à 41% des neurones du cortex frontale contiennent de larges réarrangements chromosomiques (Cai et al., 2014; Coufal et al., 2009; McConnell et al., 2013). Il est toutefois étonnant que des fréquences si élevées n'aient pas été détectées précédemment et ces études font l'objet d'une vive polémique. Cette polémique a notamment été alimentée par une publication du laboratoire du Dr Angelika Amon qui présentait cette même technique de séquençage de cellules individuelles utilisée sur des neurones mais sans

pouvoir reproduire ces résultats (Knouse et al., 2014). Dans cette étude, 96 neurones individuels de souris furent séquencés et une seule SV dans 1 seul neurone a pu être détectée. Les auteurs ont de plus séquencés 89 neurones issus de biopsies de lobe frontaux de 4 patients et n'ont trouvé que 2 aneuploïdies (soit 2.2% des cellules). Bien que cette étude n'infirme pas l'existence des SV dans les tissus, elle remet donc en question leur fréquence et rappelle l'importance de valider expérimentalement la découverte des SV en raison des nombreux biais potentiels pouvant produire des faux positifs.

3.2.2 - Etude du polymorphisme des chromosomes dans les organismes modèles

Ces dernières années, le séquençage de nombreuses espèces a été réalisé pour explorer la diversité génétique, notamment au sein d'organismes modèles animaux ou végétaux classiques (souris, drosophile, nématode, levure).

SV chez la souris (Mus musculus)

L'étude des SV chez la souris a profité des progrès des technologies de séquençage. En 2010, le séquençage de 2 lignées pures de souris permet de découvrir 7 196 SV par rapport au génome de référence (C57BL/6J), les 2/3 étant dues à des insertions de transposons (Quinlan et al., 2010). Dans le tiers restant, 60% des SV sont des délétions par rapport à la souche de référence et 26% des insertions. Les inversions et les duplications en tandem ne représentent respectivement que 6 et 9% des SV. Un grand nombre de jonctions de ces SV (3316) ont été résolues à l'échelle du nucléotide, ce qui a permis de montrer que 16% d'entre elles sont des cas complexes, qui se sont probablement produits suite à des évènements de changement de matrices lors de la réplication (FoSTeS). En 2011, le 'Mouse Genomes Project' a publié la séquence complète de 18 lignées de laboratoire ainsi que la carte détaillée des SV et des évènements de rétro-transpositions dans chaque lignée (Keane et al., 2011; Yalcin et al., 2011). Ce projet a permis de découvrir 710 000 SV répartis sur seulement 1% du génome de la souris. Il permet en outre de se rendre compte de la puissance du séquençage par rapport aux puces CGH : alors que seulement 121 délétions avaient précédemment été découvertes dans la souche DBA/2J (Agam et al., 2010), le 'Mouse Genomes Project' en a lui identifié 16 318 dans cette même souche.

SV chez la mouche du vinaigre (Drosophila melanogaster)

La forte incidence des inversions sur les génomes de drosophile est connue depuis longtemps (cf chapitre1). Le type de réarrangement le plus commun chez la mouche est l'inversion paracentrique[11]. C'est ainsi que bien avant l'avènement du séquençage nouvelle génération, plus de 500 inversions avaient déjà été répertoriées chez *D. melanogaster* (Aulard et al., 2004).

A la fin des années 2000, les premières études de génomique des populations donnèrent un aperçu de l'impact des SV avec une résolution de quelques centaines de paires de bases dans une quinzaine de lignées différentes (Dopman and Hartl, 2007; Emerson et al., 2008). Plus récemment, le projet DGRP pour 'Drosophila melanogaster Genetic Reference Panel' a dressé une carte des SV grâce au séquençage de 39 lignées (Zichner et al., 2013). Ce consortium a identifié 8 962 délétions et 916 duplications en tandem avec des tailles allant de 50 à 165 327 pb pour les délétions (médiane 178 pb) et 78 à 129 958 pb pour les duplications (médiane 2 111 pb) (Zichner et al., 2013). Aujourd'hui, ce projet a permis d'identifier un total de 1 296 080 SV chez la drosophile en séquençant plus de 200 souches. L'étude des jonctions de ces SV suggère qu'elles se produisent en majorité (88%) via le NHEJ.

SV chez le nématode (Caenorhabditis elegans)

Comme chez la drosophile, les inversions sont des mutations répandues chez le nématode. On ne peut en revanche pas parler d'inversions paracentriques puisque chez le nématode les chromosomes sont holocentriques[12]. De plus, alors que les translocations sont mystérieusement rares dans le génome de la mouche, la comparaison des génomes de *Caenorhabditis elegans* et de *C. briggsae* a montré qu'il existe 1 translocation pour 5 inversions (Stein et al., 2003).

Les SV intraspécifiques chez cet organisme ont d'abord été identifiées grâce aux puces CGH réalisées sur 2 souches relativement divergées CB4856 (Hawaii) et JU258 (Madeira) (Maydan et al., 2007). Cette étude avait permis d'identifier 141 délétions dans le génome d'Hawaii et 122 dans celui de Madeira par rapport à la souche de référence Bristol N2. En 2010, les mêmes auteurs utilisaient des puces CGH pour détecter les SV dans 12 souches de nématodes (Maydan et al., 2010). Cette étude permit de démontrer que le nématode possède

[11] Qui n'inclus pas le centromère

[12] Chez les espèces possédant des chromosomes holocentriques, le rôle de centromère est assuré par les chromosomes dans toute leur longueur

une majorité de délétions par rapport aux duplications. Au total, 510 délétions affectant 1 136 gènes furent identifiées, représentant plus de 5% des gènes de la souche de référence N2. Finalement, le séquençage de la souche Hawaii montra en fait que le nombre de délétions serait 10 fois plus élevé que ce qui avait été trouvé précédemment pour ce génome (1 430 délétions) (Vergara et al., 2014). Alors qu'il avait été initialement postulé par Maydan et collègues que la majorité des délétions chez le nématode se produisent via le NAHR (Maydan et al., 2010), ces derniers résultats indiquent que seul 13% des délétions portent en fait la signature de ce mécanisme.

SV dans les génomes de plantes

Chez les plantes, bien que l'appellation SV (et plus particulièrement les CNV) soit utilisée pour évoquer le polymorphisme entre les cultivars[13], l'existence de l'hétérogénéité intra-cultivar est également un phénomène reconnu. La contribution des SV chez les plantes a été étudiée ces 5 dernières années chez de nombreuses espèces comme Arabidopsis (Cao et al., 2011; Lu et al., 2012), le maïs (Springer et al., 2009; Swanson-wagner et al., 2010), le soja (McHale et al., 2012), l'orge (Muñoz-Amatriaín et al., 2013) et le riz (Hurwitz et al., 2010). Quelques études se sont même intéressées à l'exploration des variations génétiques d'un même individu (Debolt, 2010; McHale et al., 2012; Ossowski et al., 2010). Le maïs fût la première plante dont les CNV ont été largement étudiées. Nous notons 2 études majeures (Springer et al., 2009; Swanson-wagner et al., 2010) qui ont utilisé des puces CGH dessinées à partir de la souche de référence B73. Springer et al ont détecté les CNV dans la lignée Mo17 alors que Swanson-wagner et collègues se sont concentrés sur les régions codantes de 19 lignées de maïs et 14 lignées sauvages de la téosinte de Balsas (Zea mays subsp. parviglumis[14]). Dans les lignées B73 et Mo17, 400 régions présentant des CNV ont été identifiées. Bien que des CNV aient été détectées sur presque tous les chromosomes du maïs, elles ne sont pas uniformément distribuées dans le génome. Les régions présentant peu voire pas du tout de CNV ont été principalement trouvées à proximité des centromères. Chez le soja, plusieurs centaines de CNV de taille moyenne (environ 20kb) et dont la distribution n'est pas uniforme dans le génome (presque pas de CNV sur les chromosomes 5 et 11) ont été détectées dans le génome de 3 cultivars distincts (Archer, Minsor and Noir 1) (McHale et al., 2012). Enfin, chez Arabidopsis, 80 (Cao et al., 2011) et plus récemment 180 (Long et al., 2013) lignées ont été

[13] Variété de plante obtenue en culture, généralement par sélection

[14] Sous-espèce du genre Zea considérée comme l'ancêtre du maïs (Zea mays)

séquencées dans le cadre du 1001 Genomes Project. Dans la première étude, les souches ont été sélectionnées dans 6 zones géographiques distinctes du globe et l'analyse de la profondeur de séquençage a permis d'identifier 1 059 CNV dont la taille varie entre 1 et 13kb et qui couvre 2.2 Mb du génome de référence (~1,5% du génome).

SV chez la levure de boulanger (Saccharomyces cerevisiae)

L'une des premières SV découvertes chez la levure fût une grande délétion sur le chromosome 3 qui comprenait le locus de type sexuel MAT (Hawthorne, 1963). Cette délétion, dite « de Hawthorne », est une délétion de 100kb induite par la recombinaison ectopique des répétitions en orientation directe de MAT et de la cassette HMR (Herskowitz, 1988) figure 20. L'isolement de cette délétion a été possible dans des cellules diploïdes puisque la diploïdie permet de s'affranchir partiellement des gènes essentiels (délétion léthale dans des cellules haploïdes). La taille des délétions visibles dans des cellules haploïdes en est donc limitée. Des délétions plus petites ont également été isolées dès les années 70 chez la levure (voir ci-après : « Systèmes génétiques et tests de fluctuations »).

Les premiers pas de la génomique des populations n'ont pas été très fructueux pour la découverte des CNV chez la levure. En 2009, le séquençage par le 'Saccharomyces Genome Resequencing Project' (SGRP) de 36 souches de *S. cerevisiae* à basse couverture (1-4X) n'a pas permis de détecter des CNV (Liti et al., 2009). Plus tard, 14 de ces souches furent re-séquencées à une couverture suffisante pour permettre leur assemblage *de novo* (Bergström et al., 2014). L'analyse des génomes des 4 principales lignées phylogénétiques de *S. cerevisiae* a ainsi montré que ces génomes sont largement colinéaires. Les régions subtélomériques[15], bien que plus difficiles à assembler, présentent cependant une grande variabilité avec un grand nombre de SV couvrant 423kb du génome. Cela représente 32% des séquences subtélomériques chez *S. cerevisiae*. Par comparaison, il a été mesuré que seul 0,7% du génome non subtélomérique est affecté par des CNV, ce qui représente une dynamique 42 fois moins élevée. L'exemple le plus frappant de CNV découverte dans cette étude est celui de la souche du vin YJM981 qui compte environ 1000 copies du gène YRF1, alors que les autres souches n'en possèdent qu'entre 10 et 40. Ceci est un exemple extrême d'un phénomène d'amplification de l'élément subtélomérique Y' servant de mécanisme alternatif pour la

[15] Région allant du dernier gène essentiel sur le chromosome à la fin du chromosome

maintenance des télomères suite à l'échec du système de maintenance classique (Lundblad and Blackburn, 1993).

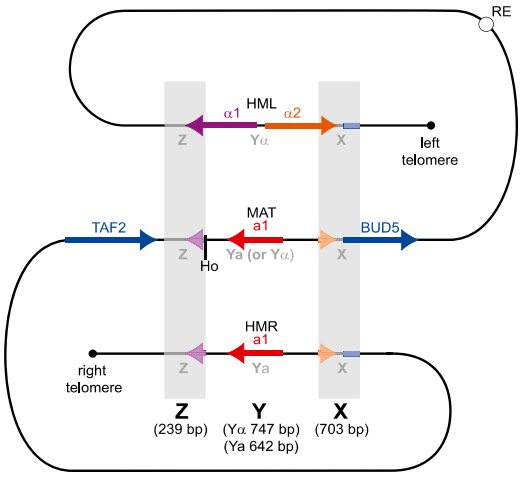


Figure 20: Organisation des *loci HML*, *MAT* et *HMR* sur le chromosome 3 chez *S. cerevisiae*. Les régions Z et X sont présentes en 3 copies en orientation parallèle et permettent l'échange des cassettes de type sexuel par conversion génique. Crédits : Gordon et al.

Plus récemment, (Marie-Nelly et al., 2014) ont démontré en utilisant des données de contact chromosomique Hi-C que la souche malaysienne UWOPS03-464.4 de *S. cerevisiae* possèdent 8 jonctions interchromosomiques qui résultent de plusieurs évènements de translocations réciproques. Bien que l'existence de translocations entre différentes espèces de levures étaient déjà connu (Fischer et al., 2000, 2006), ce travail confortée l'idée que même des souches très proches peuvent accumuler un grand nombre de réarrangements chromosomiques.

3.2.3 - Calcul des taux de SV dans les génomes

Il existe plusieurs approches pour estimer les taux d'apparition de SV dans les génomes : plusieurs techniques basées sur le séquençage entier des génomes (génomique comparative, séquençage de trio…), l'analyse de la fréquence des SV dans les populations, les lignées accumulatrices de mutations et les systèmes génétiques de sélection positive.

Taux de fixation des réarrangements chromosomiques basés sur le séquençage des génomes

La génomique comparative permet d'étudier la vitesse de fixation des réarrangements chromosomiques dans les génomes en comparant leur synténie, c'est-à-dire la conservation de l'ordre des gènes entre 2 espèces. Ces taux de fixation dans les génomes correspondent à la

fixation de SV initialement polymorphes. Ils donnent donc une idée de la dynamique évolutive des génomes mais pas des taux de formation de novo des SV. De plus, les taux inférés de cette manière ne peuvent pas être considérés comme universels puisqu'ils varient entre les espèces et au cours de l'évolution des espèces (Bourque et al., 2005; Fischer et al., 2006).

La comparaison de génomes d'espèces proches dont on connais la distance évolutive, à une échelle du million d'année (*D. melanogaster* et *D. sophophora* par exemple), a montré que chez les invertébrés, le taux de réarrangements chromosomiques est presque 2 fois plus élevé que chez les vertébrés : la mouche du vinaigre acquière entre 0,05 et 0,07 jonctions par mégabases (Mb) par million d'année (Ma) contre environ 0,03/Mb/Ma pour les rongeurs (Bourque et al., 2004; Ranz et al., 2001). Cette différence est certainement imputable à l'histoire des 2 espèces : la taille efficace[16] des populations drosophiles est plus grande que celle des rongeurs (10^6 contre 10^5) et les drosophiles se multiplient plus vite (10 générations par an contre 2 pour les rongeurs (Eyre-Walker and Keightley, 2007)). Le nématode accumule lui un nombre relativement élevé de réarrangements chromosomiques avec 0,5 à 0,7 jonctions/Mb/Ma (Stein et al., 2003) ce qui pourrait être imputable à une taille efficace de population encore plus grande et à un temps de génération encore plus court que la mouche. De plus, le nématode possède des chromosomes holocentriques capables de tolérer plus de réarrangements chromosomiques car il n'y aura jamais de création de chromosomes acentriques ou dicentriques, ce qui favorise l'accumulation de SV (Dernburg, 2001). Chez les vertébrés, parmi les génomes de l'homme, de la souris et du poulet, c'est celui du poulet qui accumule le moins de réarrangements (Burt et al., 1999). De même chez le poisson, un faible taux d'évolution des chromosomes a été mis en évidence en comparant l'ordre des gènes ches l'homme à celui du Tétraodon et du Fugu (International Chicken Genome Sequencing Consortium, 2004). Notons en particulier que les rongeurs subissent un taux d'évolution bien plus rapide que les primates (3,2-3,5 réarrangements /Mb/Ma chez les rongeurs contre 1,6 réarrangements /Mb/Ma chez l'homme). Il a également été montre que les taux de réarrangements sont en moyenne 3 fois plus élevés chez les vertébrés que chez les levures (2 réarrangements/Ma). Cependant, compte tenus du fait que les génomes des vertébrés sont en moyenne 200 fois plus gros que les génomes de levures, les taux de réarrangements chromosomiques normalises sont en moyenne 50 fois plus élevés dans les génomes de levures que dans les génomes de vertébrés (Drillon and Fischer, 2011).

[16] Nombre d'individus d'une population idéale pour lequel le degré de dérive génétique est équivalent à celui de la population réelle.

Les estimations précédentes, qui sont basées sur la génomique comparative, indiquent la vitesse d'évolution des génomes. Mais elles dépendent aussi de plusieurs paramètres comme le temps de divergence, le temps de génération et la taille des populations efficaces. De façon classique la comparaison des génomes de plusieurs individus de la même espèce, par exemple ceux d'individus appartenant à une même famille, permet d'estimer les taux de formation *de novo* des SV par génération. Cela consiste à séquencer le génome d'un enfant et de ses deux parents. Les SV présentes uniquement chez l'enfant permettent alors d'estimer le taux de mutation. L'une des premières études s'appuyant sur des pédigrées étendus[17] a permis de mesurer un taux de CNV de $1{,}2.10^{-2}$ CNV (>100kb) par génome et par génération. Plus récemment, l'étude de 250 familles a permis d'obtenir une image plus globale puisque toute la gamme de taille des SV (>20pb) a été analysée (Kloosterman et al., 2015) au lieu de se limiter aux SV de taille supérieures à 100pb ou 1kb. Le taux de SV rapporté dans cette dernière étude est mécaniquement plus élevé (0,16 SV par génération) que ceux rapportés précédemment.

Taux des SV basés sur la fréquence des SV dans les populations

Une autre méthode de calcul des taux de SV dans les génomes fait écho à la découverte de maladies génétiques liées à des SV spontanées chez l'homme. Dans les années 70-80, l'étude du caryotype de 5 000 enfants morts nés a permis d'établir le taux de translocations Robertsoniennes à $3{,}54 \times 10^{-4}$ par gamète par génération (Jacobs, 1981). Des taux de SV peuvent également être calculés en utilisant une propriété particulière des fréquences des SV dans les populations, découverte par Haldane (Haldane, 1935). Ce dernier a démontré que pour les gènes du chromosome X, le taux de mutation peut être estimé par (1 – f)x/3, où f représente la fertilité des mâles affectés par rapport aux non affectés et x est la fréquence des mâles affectés dans la population. Si la mutation étudiée provoque la létalité ou la stérilité, le taux de mutation devient x/3 (ce qui est logique puisque les mâles représentent 1/3 du réservoir d'allèles pour une maladie dont la prévalence est à l'équilibre). Cette méthode a été appliquée à la dystrophie musculaire de Duchenne (DMD), qui est une maladie liée au chromosome X dans laquelle les mâles sont stériles (van Ommen, 2005). L'incidence chez les mâles étant de 1 :3 500, il en résulte que la fréquence de mutation conduisant à cette maladie est de $\sim 10^{-4}$. Comme 65% des mutations responsables de cette maladie sont des délétions et

[17] trio parents-enfants + des individus de la descendance pour analyser la transmission des CNV. Les CNV doivent être transmis à environ 50% de la descendance pour ne pas être considérés comme des faux positifs

9% des duplications, cela donne donc des taux de $6{,}67.10^{-5}$ délétions et 10^{-5} duplications par locus DMD et par génération. Le segment d'ADN dupliqué ou délété représentant 1/2000ème du génome humain, ces taux extrapolés à l'échelle du génome complet donnent 0,13 délétions et 0,02 duplications par génome et par génération. Rien n'indique cependant que la totalité des duplications et des délétions de ce locus induisent la maladie et les taux calculés seraient donc sous-estimés. Des taux équivalents ont toutefois pu être obtenus pour d'autres maladies comme la maladie de Charcot-Marie (Lupski, 2007).

Lignées accumulatrices de mutations

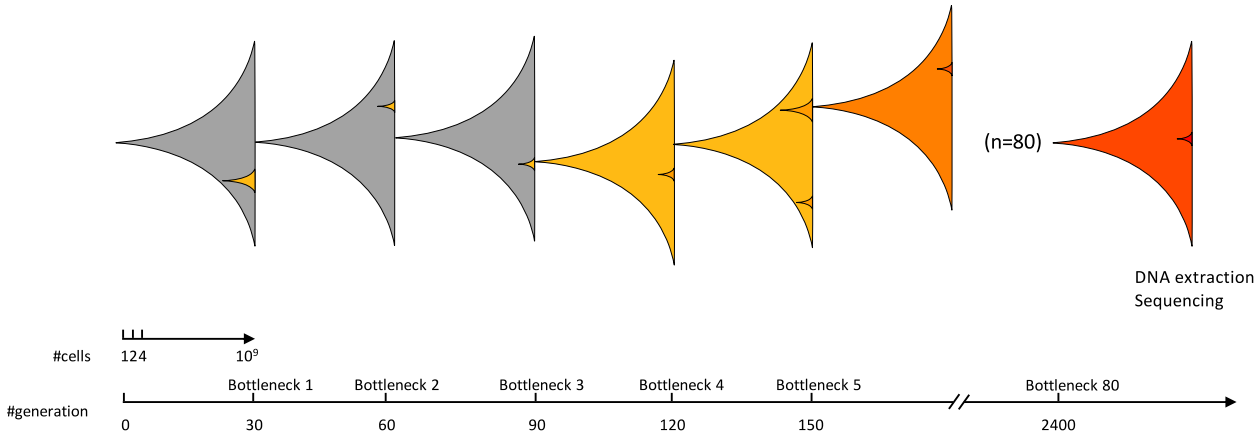


Figure 21: Lignées accumulatrices de mutations (MAL). Le principe des MAL est de réaliser des cultures successives et d'essayer de fixer régulièrement des mutations subclonales dans le génome par dérive génétique. Dans cet exemple, on cultive des cellules pendant 30 générations avant de réaliser un goulot d'étranglement d'une seule cellule. A chaque goulot d'étranglement, une mutation peut être fixée dans le génome. Le nombre final de mutations observées dans le génome dépend du nombre de génération total et du nombre de goulots d'étranglements réalisés (ici 80 goulots d'étranglements pour 2400 génération au total).

Le principe des lignées accumulatrices de mutation (ou MAL pour mutation accumulation lines) est de cultiver des cellules pendant un grand nombre de générations et de réaliser des goulots d'étranglement fréquents pour limiter les effets de la sélection naturelle et fixer des mutations par dérive génétique (Figure 21). La taille de la population efficace (N_e) est en général minimale pour réduire l'efficacité de la sélection et permettre l'accumulation de mutation (Halligan and Keightley, 2009; Mukai, 1964; Ohnishi, 1977). L'utilisation des MAL pour calculer des taux de mutation présente l'avantage de permettre une quantification des SV à l'échelle du génome complet par séquençage haut-débit.

Les MAL ont été utilisés pour calculer les taux de certains types de SV chez des organismes modèles comme la levure de boulanger, le nématode ou la drosophile. Chez la levure, l'expérience a été initialement réalisée dans des cellules haploïdes (4 lignées en parallèles, ~4800 générations (Lynch et al., 2008)), ce qui a permis d'estimer le taux de délétion et de duplication à respectivement $1{,}2.10^{-2}$ et 2.10^{-2} SV par cellule et par division. Ces taux ont cependant été inférés à partir d'un nombre restreint de SV (11 duplications et 4 délétions) et ils

ne représentent donc probablement qu'une estimation grossière des taux réels. Une étude plus récente, où 145 lignées diploïdes ont été cultivées pendant 2062 générations, n'a permis d'identifier que 3 CNV. Ceci n'est pas surprenant compte-tenu de la stabilité des génomes diploïdes chez la levure, ces dernières pouvant réparer les mutations par recombinaison avec le chromosome homologue (Nishant et al., 2010; Zhu et al., 2014).

Le vers *C. elegans* a été le premier organisme multicellulaire pour lequel une estimation du taux de duplication et de délétion a été réalisé à partir de MAL (Lipinski et al., 2011). Dans cette étude, 10 lignées de vers ont été cultivées, passées par des goulots d'étranglement d'un seul vers pendant ~430 générations, puis le génome des souches a été analysé par puce CGH. Au total 14 duplications ont été identifiées et validées par PCR quantitative. Ces duplications concernent en fait 30 gènes et surviennent à un taux de duplication compris entre 1,25 et 3,4 10^{-7} par gène et par génération. Les auteurs de ce travail pensent cependant que ces taux sont sous-estimés en raison d'une densité insuffisante des sondes sur la puce pour détecter les petites duplications et de l'utilisation de sondes uniques dans le génome, ce qui ne permet pas de tenir compte du caractère recombinogènes des séquences répétées dans les génomes. Chez la drosophile, les génomes de 8 lignées MAL ont été séquencés (Schrider et al., 2013). Bien que l'organisme et la technologie de détection des SV utilisée (Illumina PE) soient très différents, le taux de duplication par gène est étonnamment proche de celui estimé chez le nématode (compte tenu de leur éloignement évolutif): 1,25-3,75.10^{-7} par gène/génération.

Tests de fluctuations et systèmes génétiques de mesure des taux de SV

Il existe de nombreux systèmes génétiques qui permettent de mesurer des fréquences de SV dans les génomes. Cependant l'utilisation de ces systèmes est limitée à 4 niveaux. i) Ils ne sont utilisables que dans des organismes modèles dans lesquels ils peuvent être construits facilement. ii) Ils sont en général limités à un nombre restreint de *loci* qui ne sont pas nécessairement représentatifs du génome complet. iii) Ils ne mesurent souvent qu'une ou quelques catégories de SV à la fois. iv) l'utilisation d'un phénotype pour sélectionner les cellules mutantes produit une image potentiellement biaisée du processus normal de mutation.

Les systèmes génétiques possèdent toutefois des avantages notables. Ils permettent de mesurer les fréquences de presque n'importe quel type de SV pour peu qu'on puisse imaginer un système de mesure adéquat. Ils permettent aussi d'effectuer des mesures dans différents fonds génétiques pour faire des études de génétique inverse. Enfin ils permettent de s'affranchir du séquençage encore relativement coûteux. Des systèmes génétiques de mesure

des fréquences de SV ont été développés dans plusieurs organismes modèles dont la bactérie, la levure et la mouche.

Test de fluctuation

Les différents systèmes génétiques précédents permettent de mesurer la fréquence d'une SV mais ne donnent pas directement son taux de formation. Le taux de mutation correspond à la fréquence à laquelle une mutation se produit par unité de temps. Ainsi, dans une culture, le nombre de cellules mutantes à un temps *t* dépend directement du moment *ti* où la mutation s'est produite. C'est cette propriété fondamentale des mutations aléatoires qu'ont exploité Luria et Delbrück en établissant une méthode, le test de fluctuation, qui permet de mesurer les taux de mutation (Luria and Delbrück, 1943).

Le principe expérimental du test de fluctuation, tel qu'imaginé par Luria, est de réaliser un grand nombre de cultures en parallèle et de compter le nombre de mutants qui sont apparus dans chaque culture arrivée à saturation. Le nombre de mutations indépendantes dont découlent les mutants peut ensuite être obtenu par plusieurs méthodes, qui reposent toutes sur la modélisation réalisée par Delbrück de la distribution du nombre de mutants dans les cultures (complétée plus tard par Lea et Coulson (Lea and Coulson, 1949)). La version la plus simple du modèle de Luria-Delbrück (LD) est formulée par les hypothèses suivantes :

1. Les cellules poussent de façon exponentielle
2. La probabilité d'apparition de chaque mutation est indépendante des mutations précédentes
3. La probabilité de mutation est constante au cours de la vie d'une cellule
4. Les vitesses de croissance des cellules mutantes et sauvages sont identiques
5. La proportion de cellules mutantes est toujours faible par rapport aux cellules sauvages
6. Le taux de réversion des mutations est négligeable
7. La mort cellulaire est négligeable
8. Tous les mutants sont détectables par le dispositif expérimental
9. Aucun mutant n'est produit après la sélection des cellules mutantes
10. Le nombre initial de cellules est négligeable par rapport au nombre final de cellules dans la culture
11. La probabilité de mutation est constante au cours de la culture

Les méthodes qui permettent d'estimer le nombre moyen de mutations par culture à partir de ce modèle (ou d'une variante) sont appelées des estimateurs. Les deux estimateurs les plus

classiques sont le terme zéro de la loi de Poisson et la méthode MSS-ML (Ma-Sandri-Sarkar Maximum Likelihood).

❖ *Terme zéro de la loi de Poisson*

La distribution du nombre de mutants dans une culture ne suit pas une loi de Poisson mais c'est par contre le cas du nombre de mutations. Parmi les multiples cultures réalisées en parallèle, celles qui ne présentent pas de mutants sur les boites de pétri correspondent à des cultures qui n'ont pas subi d'évènements mutationnels sélectionnables. Le nombre moyen de mutation par culture peut alors être calculé à partir de la proportion de cultures sans mutant, en utilisant le terme zéro de la loi de Poisson appelé p_0 (Luria and Delbrück, 1943)

$p_0 = e^{-m}$ avec m le nombre moyen de mutations par culture

$m = -\ln p_0$

Cette méthode permet de s'affranchir de l'hypothèse numéro 4 du modèle de LD. En effet, puisque cette méthode utilise la proportion du nombre de culture sans mutants pour calculer m, elle n'est pas influencée par une vitesse de croissance différentielle des mutants (appelée b). Cette méthode est cependant expérimentalement plus contraignante que la méthode du MSS car elle nécessite de se placer dans des conditions où m est compris entre 0,3 et 2,3 et où au moins 30% des cultures ont au moins 1 mutant. De plus, pour un même nombre de cultures, cette méthode est moins précise que la méthode MSS.

❖ *Ma-Sandri-Sarkar Maximum Likelihood (MSS-ML)*

Comme pour la majorité des problèmes de statistique paramétrique[18], les méthodes de maximum de vraisemblance permettent d'obtenir les meilleurs résultats. L'algorithme MSS a été décrit en 1992 (Sarkar et al., 1992). C'est un algorithme récursif qui permet de calculer efficacement la distribution de LB pour un m donné :

$p_0 = e^{-m}$

$p_r = \frac{m}{r} \sum_{i=0}^{r-1} \frac{p_i}{(r-i+1)}$ avec r le nombre de mutants

Cet algorithme peut être utilisé pour identifier la meilleure valeur de m en utilisant la fonction de maximum de vraisemblance suivante (Ma et al., 1992) :

$f(r|m) = \prod_{i=1}^{C} f(r_i|m)$ avec C la proportion de cultures avec i mutants

[18] On parle de test paramétrique lorsque l'on stipule que les données sont issues d'une distribution paramétrée

Grâce à cette fonction, une procédure peut être mise en place pour trouver la meilleure valeur de *m*. Contrairement à la méthode du terme zéro de la loi de Poisson, cet estimateur possède l'avantage d'être utilisable presque quel que soit le nombre de mutants par culture. Il a toutefois été montré que cet estimateur devient instable pour les grandes valeurs de *m* et une procédure de « winsorisation » est nécessaire pour les cultures avec plus de 150 (Foster, 2006) ou 500 (Hamon and Ycart, 2012) mutants (c'est-à-dire que toutes les cultures avec un grand nombre de mutants sont classées dans la même catégorie).

- Déviations du modèle de LD

Certaines hypothèses du modèle de LD comme l'absence de mort cellulaire, la division exponentielle des cellules et le fitness identique des cellules mutantes et sauvages sont en général fausses. Par exemple, l'absence de prise en compte de la mort cellulaire dans l'estimation de *m* a été étudiée dans plusieurs travaux et il a été montré qu'elle entraîne la sous-estimation de *m* si elle n'est pas corrigée (Angerer, 2001; Tan, 1983; Ycart, 2014).

Un second type de déviation est appelé « postplating growth » ou croissance post-étalement. Il s'agit probablement de la génération de mutants secondaires qui permettent aux cellules de continuer à se diviser sur le milieu de sélection, conduisant à l'apparition de mutants après l'étalement des cellules. Ce phénomène a été modélisé par Lang et Murray (Lang and Murray, 2008).

De plus, un biais expérimental classique des tests de fluctuation consiste à étaler sur milieu sélectif des fractions de chaque culture et non la culture en totalité pour évaluer le nombre de mutants. Cet échantillonnage plus ou moins fort en fonction de la fraction étalée, provoque une perte d'information puisque l'ensemble des mutants de chaque culture n'est pas nécessairement détecté. Cet effet appelé efficacité d'étalement ('plating efficiency') est maximum lorsqu'aucun mutant n'est détecté (ce qui correspond à l'hypothèse qu'aucune mutation n'a eu lieu) alors que la culture contient des mutants. Ceci a été largement étudié et des corrections du modèle de LD ont été proposées pour le caractériser (Jones, 1993; Stewart, 1991; Zheng, 2008)

Dans la majorité des cas, les mutants générés à partir de l'utilisation de systèmes génétiques ont un fitness différent des cellules sauvages. Dans ce cas, un modèle de Luria-Delbrück à 2 paramètres : LD(m, b), où b représente le fitness des cellules mutantes par rapport à celui des cellules sauvages, peut être utilisé. Ces paramètres peuvent être estimés conjointement par maximum de vraisemblance ou par la méthode de la fonction génératrice (Hamon and Ycart, 2012; Zheng, 2002).

Les tests de fluctuations décrits précédemment permettent de calculer des taux de mutations à partir de la fréquence des mutants identifiés dans de nombreux systèmes génétiques. Par souci de concision, nous analyserons ici les systèmes génétiques existants pour la levure de boulanger uniquement.

Systèmes génétiques de mesure des taux des délétions

Les premiers tests de détection des délétions chez la levure n'ont pas utilisé des systèmes construits *de novo*, mais ont cherché à utiliser des gènes dont on peut sélectionner la perte comme *HIS4* (Fink and Styles, 1974). Ce système génétique relativement complexe est basé sur l'observation qu'une mutation polaire[19] d'un des deux premiers segments (A et B) ne poussent pas sur un milieu avec de l'histidinol. Une délétion en-phase de cette mutation (mais qui conserve le segment C) peut restaurer le phénotype sauvage. De plus, les cellules qui subissent une délétion ne peuvent plus donner des révertants incapables de pousser sur un milieu avec de l'histidinol. Ces premiers systèmes ont permis de mesurer des fréquences de délétion de l'ordre de 10^{-8} DEL/division. Le système *CYC1* a d'ailleurs permis d'identifier un 'hotspot' de mutation dans une souche particulière de levure. Bien que dans les souches classiques sauvages le taux de délétion soit relativement faible, Liebman et ses collaborateurs ont identifié une souche de levure dans laquelle la délétion simultanée de 3 gènes (dont *CYC1*) est observée à une fréquence beaucoup plus élevée (10^{-5} DEL/division) (Liebman et al., 1979). Il sera démontré plus tard que dans cette souche, le gène *CYC1* est flanqué de 2 rétrotransposons Ty en orientation directe provoquant cette délétion (Liebman et al., 1981). Un cas analogue concerne la délétion du locus du gène *SUP4* causé par la présence d'éléments LTR (Long Terminal Repeat) issus de Ty1 et Ty2 et situés de part et d'autre du gène (Rothstein et al., 1987). Dans les 2 cas, la délétion peut être expliquée par un évènement d'enjambement inégal entre les séquences répétées ou par le SSA.

En parallèle, plusieurs systèmes génétiques furent développés chez la levure pour mesurer le taux de délétions dans le génome. Un de ces systèmes est basé sur la recombinaison de 2 hétéro-allèles du gène *HIS3* (Schiestl, 1989). Dans ce système, des allèles *his3Δ3'* et *his3Δ5'*, qui partagent 400pb d'homologie, sont placés dans cet ordre et à quelques kilo-bases l'un de l'autre, en orientation directe sur le même chromosome. La recombinaison entre ces 2 allèles produit une délétion du fragment d'ADN les séparant ('pop-out'), dont la fréquence mesurée par les auteurs est de 10^{-4}/DEL/division.

[19] Mutation qui affecte la transcription ou la traduction de la partie du gène se trouvant en aval. Par exemple, les mutations non-sens, les décalages du cadre de lecture.

Un autre système de mesure des délétions utilise un allèle non fonctionnel du gène *URA2* (*ura2 15-30-72*). Ce dernier possède 2 substitutions et 1 insertion nucléotidique qui décale le cadre de lecture. Ces mutations tombent dans les domaines GATase[20] et CPSase[21] qui, contrairement au dernier domaine de la protéine (ATCase[22]), sont redondants avec la voie de biosynthèse de l'arginine. Le décalage du cadre de lecture est donc suffisant pour rendre la souche auxotrophe pour l'uracile. Une délétion des domaines GATase et CPSase peut cependant conduire à la réactivation de la transcription et de la traduction du domaine terminal ATCase[23] de la protéine et ainsi restaurer la prototrophie de l'uracile (Exinger and Lacroute, 1979; Tourrette et al., 2007). Ces délétions ne peuvent se produire que via des mécanismes utilisant la microhomologie ou bien ne nécessitant pas d'homologie. Dans ce cas, la fréquence de délétion mesurée par les auteurs est beaucoup plus faible : 10^{-9} DEL/division.

Système de mesure des taux des gros réarrangements chromosomiques (GCR)

Le système GCR pour 'Gross Chromosomal Rearrangement' a été développé par Chen et Kolodner en 1999 (Chen and Kolodner, 1999). Le principe de ce système est de placer 2 marqueurs (*URA3* et *CAN1*) localisés à environ 10kb l'un de l'autre et de co-sélectionner la perte du phénotype procuré par ces 2 gènes. La probabilité d'obtenir 2 mutations inactivant chacun des 2 gènes en une seule étape de sélection étant trop faible, les mutants de phénotype prototrophes pour l'uracile et sensibles à la canavanine ont donc pour origine une délétion du locus qui peut être associée à d'autres types de réarrangements. Ce système peut mesurer différents types de SV : des délétions des 2 cassettes, des pertes de régions terminales suivis de l'addition d'un télomère et des translocations non réciproques avec les autres chromosomes. Les taux de translocations et de délétions interstitiels mesurés avec ce système étaient de 10^{-9}/division avec une jonction impliquant de la microhomologie ou pas d'homologie du tout. La majorité de ces évènements sont donc produits par le NEHJ. Dans la publication originale, les 2 gènes *URA3* et CAN1 étaient placés 10kb après le dernier gène essentiel sur le chromosome 5 (*PCM1*). L'insertion d'un élément répété entre *CAN1* et *PCM1* (*HXT13* ou un Ty) permet d'augmenter le taux de translocation non réciproque d'un facteur 3

[20] Glutamine amidotransférases: catalyse l'enlèvement d'un groupe ammoniac NH3 d'une glutamine (Q)

[21] Carbamoyl phosphate synthétase: catalyse la synthèse de carbamoyl phosphate à partir de glutamine et de bicarbonate (HCO_3)

[22] Aspartate carbamoyltransférase: catalyse la première étape de la biosynthèse des pyrimidines

[23] Aspartate carbamoyltransférase: catalyse la première étape de la biosynthèse des pyrimidines

(Chan and Kolodner, 2011). Ceci démontre encore une fois que le taux de SV est très dépendant du contexte génomique et qu'un taux global à l'échelle du génome peut difficilement être extrapolé à partir d'une mesure effectuée à un locus spécifique.

Système de mesure des taux des duplications

Les systèmes de mesures des duplications se classent en 2 catégories : ceux qui mesurent la duplication d'un locus unique et ceux qui mesurent le taux de duplication d'un locus déjà présent en multiples copies. Chez la levure, les 2 cas d'école des systèmes de mesure d'amplification du nombre de copies sont l'ADNr et le gène *CUP1* qui peuvent s'amplifier par enjambement inégal (Petes, 1980). Certaines de ces SV sont extrêmement fréquentes. Les enjambements inégaux au niveau de l'ADNr se produisent en mitose à une fréquence de 10^{-2} par division cellulaire (Szostak and Wu, 1980) et en méiose à une fréquence de ~10^{-1} par division cellulaire (Petes, 1980).

Comme pour les délétions, de nombreux systèmes génétiques de mesure de la fréquence des duplications segmentales ont été mis au point chez la levure. Des cellules qui ont subi une duplication du gène *ADH2* ou *ADH4* peuvent par exemple être sélectionnées sur un milieu contenant de l'antimycine A[24] (Dorsey et al., 1992). Les grandes duplications (>90kb) sur le bras droit du chromosome 15 peuvent être identifiées dans des mutants *rpl20AΔ*, en sélectionnant les cellules qui présentent une reprise de croissance normale (grâce à la duplication de *RPL20B*) (Koszul et al., 2004; Payen et al., 2008). Les cellules portant la duplication d'un segment de l'allèle *ura2-15-30-72* peuvent également être isolées via la sélection de mutants URA+ (Schacherer et al., 2005). D'autres systèmes plus complexes avec plusieurs cassettes ont été construits. Notons en particulier la caractérisation de la cassette contenant les gènes *CUP1* et *SFA1* qui permet d'identifier des amplifications en sélectionnant les cellules sur un milieu contenant un taux élevé de cuivre et formaldéhyde (Narayanan et al., 2006; Zhang et al., 2013). Le laboratoire du Dr Argueso dans le Colorado a ensuite produit un système génétique complètement quantitatif, formé de la cassette *SFA1*^{V208I}*-CUP1-KanMX4*, qui permet de suivre précisément le nombre de copies d'un locus dans le génome (données non publiées). Un système reposant sur l'utilisation d'un allèle partiellement fonctionnel de *LEU2* et sur la sélection de cellules prototrophes LEU+ permet de détecter des amplifications de ce gène (~10 copies) (Erhart and Hollenberg, 1983; Watanabe and Horiuchi, 2005). Enfin,

[24] Inhibiteur de la chaîne respiratoire des mitochondries

un autre système basé sur le gène *ade3-2* permet de sélectionner des cellules dont la couleur passe du rose au rouge lorsque ce locus se duplique (Green et al., 2010; Koshland et al., 1985).

SITUATION DU SUJET DE THESE

Lorsqu'une SV apparaît dans une cellule ou un individu d'une population, sa fréquence est alors minimale (1/le nombre total de cellules). La première étape vers la fixation d'une SV apparue de *no novo* dans la population est d'échapper à la dérive génétique. Ainsi, pour avoir une chance de se fixer dans la population, ces SV doivent atteindre un seuil critique de fréquence (~1/*s*, où *s* [0 < s < 1] est la valeur sélective de la SV) au-delà duquel la croissance des cellules mutantes est quasiment déterministe à cause de la faible influence de la dérive génétique (Levy et al., 2015). Cependant, comme la valeur sélective moyenne des populations augmente au cours du temps, même les SV avantageuses peuvent être perdues par dérive génétique si elles n'ont pas atteint une fréquence suffisante (Desai and Fisher, 2007). Compte tenu du fait que la majorité des SV est vraisemblablement neutre ou délétère, elles devraient être purgées rapidement des populations. Pourtant, la génomique comparative montre que les génomes d'espèces apparentées diffèrent par des centaines de réarrangements chromosomiques. Il existe par exemple 80 réarrangements chromosomiques entre le rat et la souris (Bourque et al., 2005; Ma et al., 2006). De même, certaines espèces de levures - pourtant phylogénétiquement proche - présentent de nombreux réarrangements chromosomiques (Dujon et al., 2004; Fischer et al., 2000, 2006). De plus, à l'échelle intraspécifique, la forte incidence des SV dans les populations naturelles a par exemple été démontrée depuis longtemps chez la Drosophile grâce à la cartographie d'inversions dont les fréquences peuvent atteindre plusieurs dizaines de pour cent (Dobzhansky, 1947; Schaeffer et al., 2008). Malgré le fait que les SV provoquent une barrière reproductive dans les populations (Delneri et al., 2003; Schaeffer et al., 2003) et que les translocations réduisent la viabilité des gamètes de 50% (Loidl et al., 1998), nous avons vu l'exemple d'une souche malaisienne de *S. cerevisiae* qui compte près de 8 jonctions interchromosomiques ce qui provoque un isolement reproductif fort avec les autres souches de l'espèce (Marie-Nelly et al., 2014). En résumé, bien que les SV soient majoritairement purgées des populations à cause de leur désavantage sélectif lié à une mauvaise valeur adaptative et à leur impact sur la fertilité méiotique, un nombre significatif de réarrangements chromosomiques atteignent la fixation dans les génomes d'espèces proches.

Ces observations impliquent que la plasticité des génomes est vraisemblablement sous-estimée et qu'une proportion importante des SV ne dépasse jamais des fréquences très faibles dans les populations. L'objectif principal de mon projet de thèse entamé lors de mon M2 a été de quantifier précisément l'impact des SV dans les génomes de levures en essayant d'évaluer les taux de formation des SV à l'échelle du génome complet. Ce projet nommé Ulysses a été

imaginé par Gilles Fischer il y a environ 6 ans lorsque la technologie de séquençage Illumina commençait à s'imposer grâce à un débit de séquençage permettant d'envisager des applications qui jusqu'alors restaient hors de portée. Le principe du projet Ulysses était d'utiliser la puissance de séquençage de ces nouvelles machines afin d'essayer de détecter la formation *de novo* de tous les types de SV et à l'échelle du génome complet dans des cultures monoclonales de levures. Le nom Ulysses fait référence à la mythologie grecque dans laquelle Ulysse (ou Ulysses en anglais) est attiré par les sirènes – des créatures chimériques, mi-femmes et mi-oiseau – mais parvient finalement à résister à leur pouvoir de séduction. Ulysses fait donc référence aux sirènes qui sont utilisées ici comme une métaphore qui renvoie aux jonctions des SV qui correspondent à la jointure de 2 régions du génome qui sont habituellement distantes.

Comme tout autre problème où l'on cherche à détecter un signal faible, la détection de SV subclonales est confrontée au problème du bruit expérimental. Ce bruit est dû à l'utilisation de banques de séquençage d'ADN génomiques particulières dites Mate-Pair qui sont indispensables pour identifier des SV avec des fréquences inférieures à 1/1 000. La construction de ces banques génère en effet beaucoup de séquences d'ADN chimériques résultantes de ligations intermoléculaires qui sont topologiquement identiques aux paires de séquences discordantes décrivant les vraies SV. La détection de SV rares avec ces banques a donc nécessité la création d'un logiciel qui soit capable d'évaluer la vraisemblance de chaque SV en tenant compte de la proportion de chimères de séquençage. Avant mon arrivée au laboratoire, Gilles Fischer et Ingrid Lafontaine avaient commencé à travailler sur un outil bio-informatique de détection des SV et une première version du logiciel permettait de détecter les délétions. Durant ma thèse, j'ai travaillé aussi bien avec Gilles pour concevoir les stratégies de détection des différents types de SV, qu'avec Ingrid pour implémenter la détection de tous les types de SV, ainsi qu'avec Hugues Richard pour l'implémentation de modèles statistiques permettant de filtrer les faux positifs.

Parallèlement au développement bio-informatique d'Ulysses, j'ai poursuivi le versant biologique du projet en travaillant à la quantification du nombre, des taux et de la localisation des SV dans des populations subclonales de levures. Ainsi, afin de valider les taux de SV mesurés par l'approche Ulysses, j'ai construit des systèmes génétiques permettant de mesurer les taux d'inversions, de duplications et de translocations réciproques à un locus spécifique dans le génome de *S. cerevisiae*. Dans le cadre de ce projet, j'ai notamment développé un site web (www.lcqb.upmc.fr/bzrates) qui implémente l'algorithme 'Generating Function' (Hamon and Ycart, 2012) de mesure des taux de mutation à partir de données de tests de fluctuation et

qui permet de prendre en compte la croissance différentielle des mutants par rapport aux cellules sauvages. J'ai également entrepris une approche de génétique inverse avec pour objectif d'identifier des gènes contrôlant la stabilité du génome en mesurant dans différents systèmes génétiques les taux de formation de différents types de SV dans une vingtaine de fonds génétiques différents. Enfin, en collaboration avec plusieurs membres du laboratoire, j'ai entrepris d'évaluer les déterminants de la stabilité des chromosomes par une approche innovante de recherche de QTL dans 5 lignées pures de *S. cerevisiae*.

RESULTATS ET DISCUSSION

1 - Ulysses : détection de SV présentes en faibles proportions grâce à des librairies de séquençage avec une grande taille d'insert

Il existe plusieurs méthodes de détection des SV dans les génomes (cf. Partie 1 de l'introduction). Toutefois, aucune d'entre elles ne permet la détection des SV de toutes tailles (Figure 22).

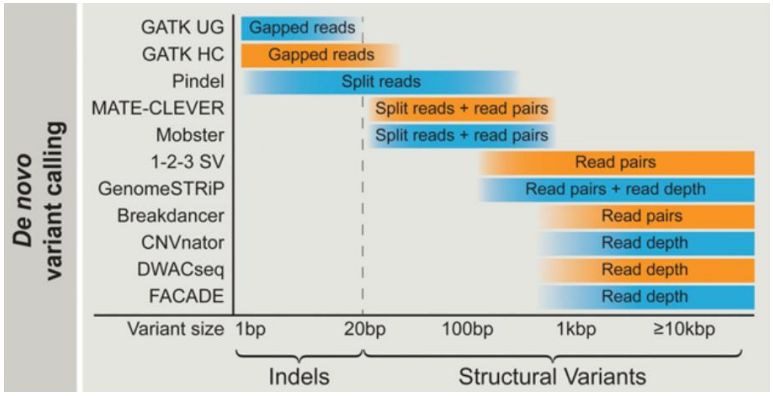


Figure 22: Gamme des tailles des SV détectables par différents logiciels utilisant différentes méthodes de détection. Seules les methodes utilisant des approches Read Pairs ou Read Depth (Profondeur de séquençage) permettent de détecter de larges SV. D'après Klossterman et al. 2015.

Les méthodes les plus adaptées pour la détection de grandes SV (>5kb) sont la profondeur de séquençage (RD) et les lectures appariées (RP). Cependant, bien que la méthode d'analyse de la RD permette de détecter la variation du nombre de copies, elle n'est pas adaptée à la détection de SV balancées. De plus, cette méthode ne permet pas de détecter des réarrangements chromosomiques présents en faibles proportions dans un échantillon. En effet, si trop peu de cellules sont porteuses de la SV, l'influence de ces cellules sur la variation de la couverture locale sera négligeable car elle ne sera pas significativement différente des variations locales de la couverture dues entre autres au contenu en GC et à l'amplification par PCR de la librairie.

La méthode des RP permet de détecter tous les types de SV car chacune possède sa propre signature qui se traduit par sa combinaison unique de critères de cartographie des paires de lectures (voir partie 1 de l'introduction et figure 7). Je ne décrirai pas ici la signature

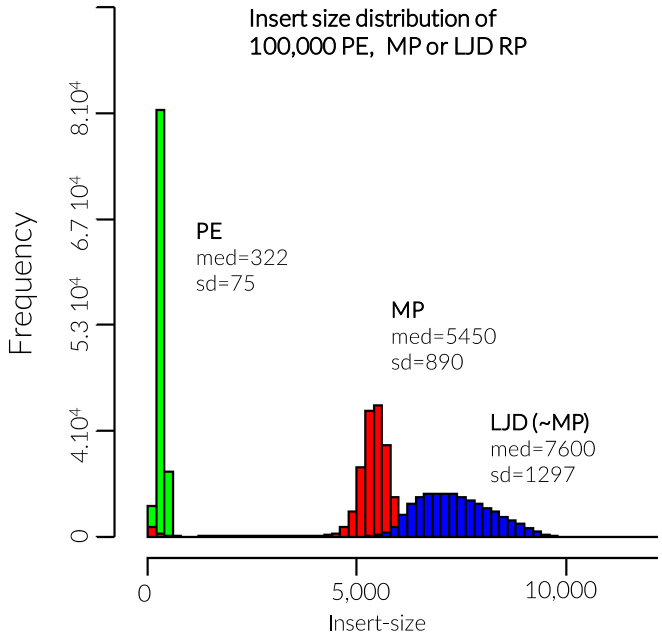


Figure 23: Distribution des tailles d'insert de 3 librairies de séquences appariées. PE=Paired-End, MP=Mate-Pair, LJD = Long Jumping Distance (MGW Eurofins). L'augmentation de la taille d'insert est concomitante avec l'augmentation de la dispersion de la taille des séquences. La contamination des librairies MP par des séquences PE est également visible à gauche de l'histogramme en rouge.

de chaque type de SV mais elle sera évoquée dans l'article n°1, présenté à la fin de cette partie. La capacité de la méthode RP à détecter des SV rares provient de l'utilisation de librairies de type Mate-Pair (MP) dans lesquelles la distance entre les 2 lectures est plus grande que dans les librairies Pair-end classiques (figure 23). En effet, un insert plus grand présente plusieurs intérêts : le premier est de faciliter la détection des SV qui impliquent des éléments répétés qui rendent la cartographie des lectures difficile. L'augmentation de la taille d'insert permet en effet enjamber de ces régions répétées et donc de cartographier les lectures dans des régions uniques. Le second intérêt repose sur l'augmentation de la couverture physique qui a pour effet d'augmenter la sensibilité. Cette dernière repose sur le principe de considère que l'ensemble des nucléotides entre les deux lectures d'une RP sont échantillonnés par le séquençage et qu'ils apportent de l'information même s'ils ne sont pas séquencées (figure 24). Par exemple, même si la jonction d'une délétion d'un fragment d'ADN n'est pas séquencée, toutes les RP qui chevaucheront cette DEL auront une taille d'insert plus grande ce qui permet donc quand même de détecter la délétion proportionnellement à l'augmentation de la taille d'insert.

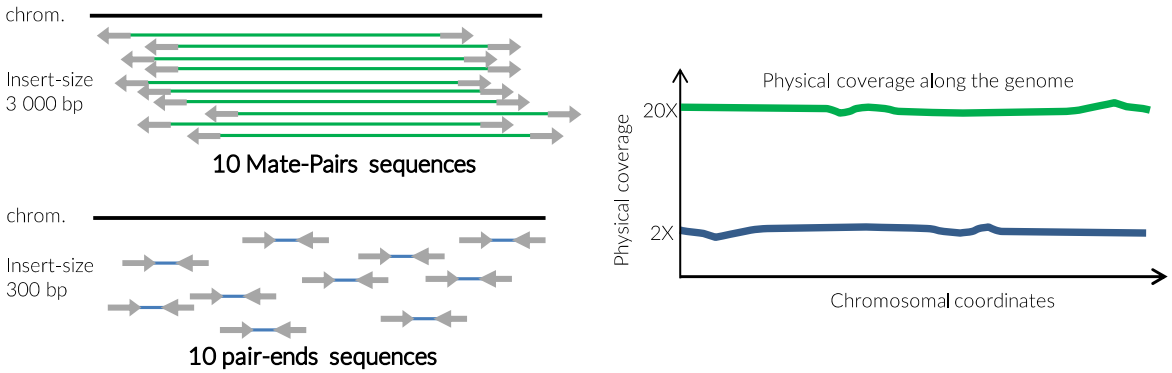


Figure 24: La couverture physique est proportionnelle à la taille d'insert des RP. Pour un nombre de RP identique, la couverture physique est plus élevée dans une librairie PE que MP.

Bien qu'elles soient théoriquement plus adaptées à la détection des SV rares, les librairies MP présentent cependant un inconvénient majeur qui est de produire beaucoup de chimères lors de la fabrication des banques. Au cours de la création de la banque de séquençage, l'ADN génomique est nébulisé en fragments de grandes tailles qui sont plus tard circularisés après l'ajout d'adaptateurs biotinilés. Cette étape est critique puisque des circularisations intermoléculaires peuvent se produire et ainsi créer des paires de séquences chimériques (figure 25). Ces chimères sont topologiquement identiques aux RP discordantes qui décrivent les SV et constituent donc un bruit expérimental important (entre 1 et 5% des RP en moyenne). Leur quantité peut être limitée en réduisant la concentration d'ADN dans le tube lors de l'étape de circularisation. Cependant, la réduction du nombre de molécules d'ADN diminue mécaniquement la diversité des séquences de la librairie. Ceci a pour conséquence d'augmenter le nombre de séquences dupliquées et donc de provoquer une perte de couverture physique. L'obtention d'une couverture physique élevée est par conséquent le résultat d'un équilibre entre la diversité des séquences de la librairie et le nombre de chimères. Afin de minimiser le nombre de FP détecté dans ces librairies, il est donc indispensable de pouvoir filtrer les chimères lors de la détection des SV.

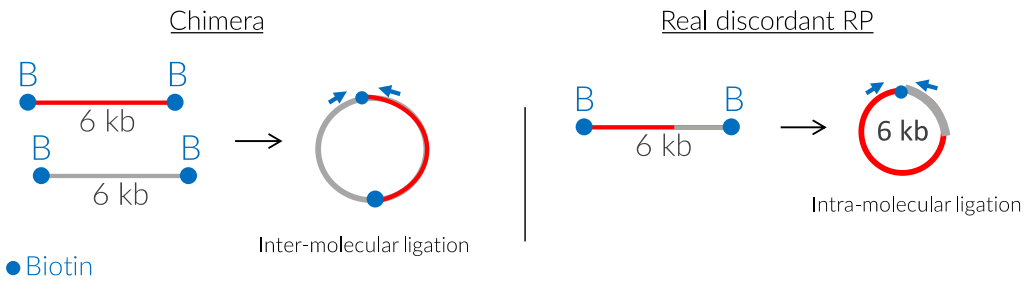


Figure 25: Illustration de l'origine des séquences chimériques lors de la création des banques MP. Les RP chimériques sont le résultat de ligations intermoléculaires. Elles peuvent être identiques à de vraies SV comme dans cet exemple.

Au cours de ce projet, nous avons développé un outil informatique, Ulysses (www.lcqb.upmc.fr/ulysses/ et https://github.com/gillet/ulysses/), qui est à l'heure actuelle le seul algorithme à pouvoir tirer parti de la très grande couverture physique des librairies Mate-Pair pour détecter des réarrangements présents en très faible proportion dans un échantillon. Brièvement, Ulysses utilise la construction de cliques maximales composées de séquences discordantes chevauchantes pour identifier des SV putatives. Ces séquences discordantes se caractérisent par des paramètres de cartographies inattendues par rapport au génome de référence (orientation, taille d'insert, chromosomes différents). Lors de la recherche de SV rares dans une banque MP, le nombre absolu de ces séquences discordantes résultantes de vraies SV est négligeable par rapport au nombre total de chimères de la librairie. Toute la difficulté de l'identification des vraies SV dans ces librairies réside donc dans notre capacité à différentier les chimères des vraies SV. Bien que les chimères et les RP discordantes des SV soient topologiquement identiques, les chimères sont *a priori* réparties de manière aléatoire dans le génome alors que les séquences discordantes qui correspondent à de vraies SV sont concentrées dans des régions qui correspondent aux jonctions des SV. Dans cette situation, nous avons établi des modèles statistiques qui établissent la vraisemblance de chaque SV compte tenu du nombre absolu de séquences discordantes (qui correspond au nombre de séquences discordantes provenant des SV ainsi que des séquences discordantes qui sont en fait des chimères et que l'on ne peut *a priori* pas distinguer) dans la librairie. Je ne rentrerai pas dans la description des modelés statistiques qui ont été établis en collaboration avec Hugues Richard (MC UPMC) et qui sont décrits dans l'article à la fin de cette section.

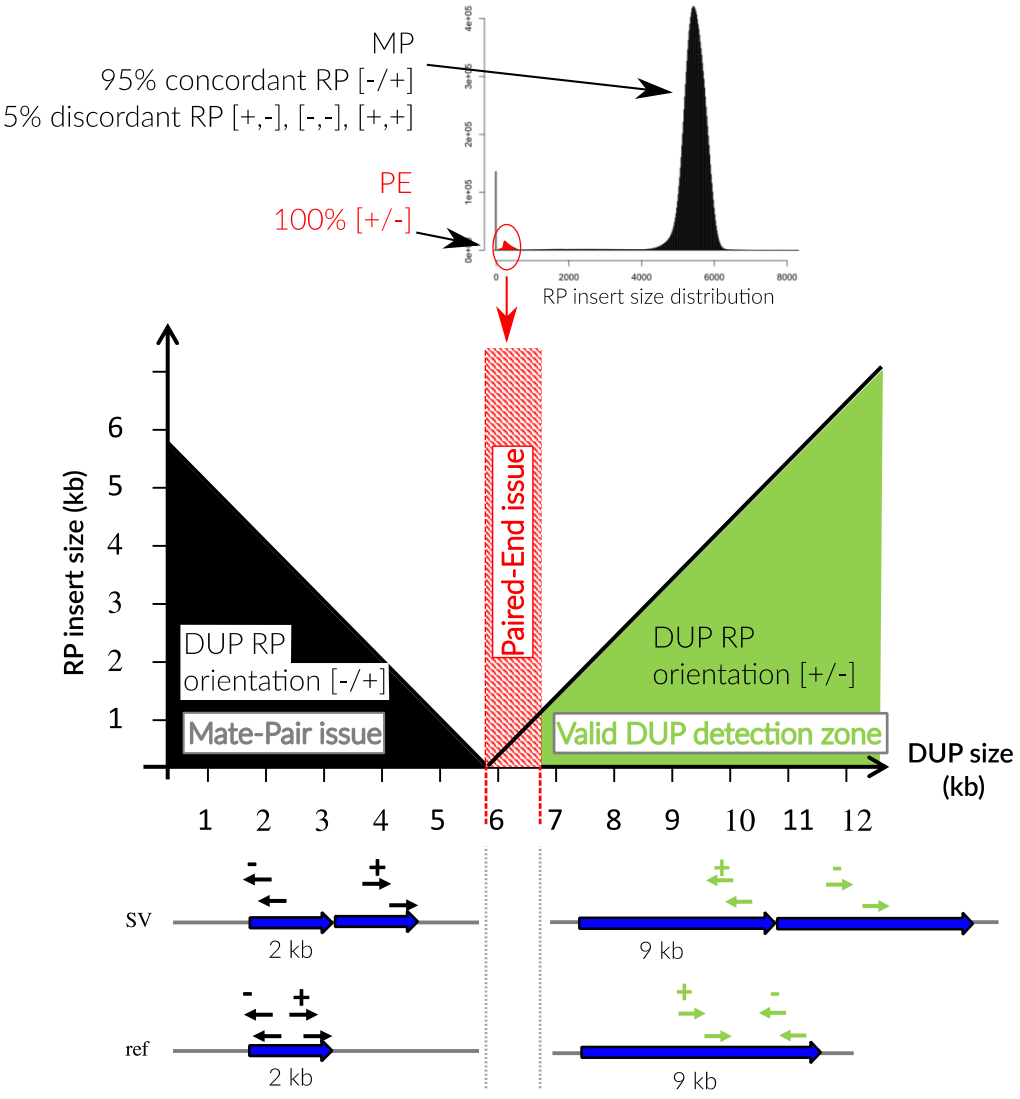


Figure 26: Impact de la contamination des séquences PE dans les librairies MP pour la détection des duplications en tandem. Les séquences PE ont une orientation opposée aux séquences MP et ont une taille d'insert de quelques centaines de paires de bases au lieu de plusieurs kb. Cependant, la signature moléculaire d'une duplication en tandem est de produire après cartographie sur le génome de référence des lectures en orientation opposée. Si ces duplications sont plus grandes que la taille d'insert de la librairie, ces DUP sont détectables (zone verte du graphique). Si la taille de cette duplication est proche de la taille d'insert de la librairie MP, les séquences discordantes produites par la duplication seront identiques aux séquences PE en termes de taille d'insert (zone hachurée rouge). Les petites duplications produisent quant à elles de petites séquences MP. Leur détection dépend alors de la qualité de la distribution des RP illustrée en haut de la figure.

Le développement du logiciel Ulysses s'est fait conjointement à la réalisation de nos expériences de détection des SV subclonales dans des populations monoclonales de levures (voir partie suivante). En effet, bien que nos premières expériences de détection des SV subclonales dans des populations n'aient pas permis de détecter de nombreuses SV, elle ont en réalité contribué au développement du logiciel Ulysses en permettant des aller-retours entre la validation expérimentale des SV et leur détection *in silico*. Elle a notamment contribué à la mise en place d'un module de filtrage des séquences PE. Ces séquences ne sont pas de type MP, car elles ne résultent pas de la circularisation d'un grand fragment d'ADN mais correspondent à

une contamination par des petits fragments d'ADN de type PE qui ont des orientations opposées. Ces derniers sont topologiquement identiques aux duplications en tandem et doivent donc être filtrés (figure 26).

Afin de tester les performances d'Ulysses sur un vrai jeu de données, j'ai dans un premier temps utilisé des données du 1000 Human Genome Project et comparer les résultats obtenus aux performances de 3 autres logiciels de détection des SV. Les critères d'évaluation des performances étaient les suivants : la spécificité de l'algorithme, sa sensibilité et sa capacité à analyser un génome humain séquencé à haute couverture avec des ressources informatiques modestes. Pour cela, je devais sélectionner un génome humain pour lequel on dispose de donnée MP ainsi que d'un *gold standard* de SV validées expérimentalement. Cependant, à partir de ces 2 critères, le champ des jeux de données utilisables et j'ai donc décidé de sélectionner l'individu NA12878 pour lequel nous disposions de 3 jeux de données de : une librairie MP à 30X de couverture en séquence, une librairie PE à basse couverture (5X) et une librairie PE à très haute couverture (200X). L'individu NA12878 a été intensivement étudié et un effort particulier a été consenti pour générer expérimentalement un *gold standard* des délétions présentes dans ce génome (Mills et al., 2011). Avec les données MP, j'ai montré qu'Ulysses possède la meilleure sensibilité brute par rapport à 3 autres logiciels couramment utilisés dans la littérature (BreakDancer, GASVpro, Delly) (Chen et al., 2009; Rausch et al., 2012a; Sindi et al., 2012) avec 167 335 DEL détectées (Article 1, figure 2A, page 95). Parmi elles, 1 278 sont considérées comme des vrais positifs (VP) d'après le *gold standard*. Le second meilleur algorithme, Breakdancer, ne détecte que 197 vraies DEL pour 21 333 DEL au totale. Après application des modules statistiques, Ulysses filtre 99,88% des DEL totales et 86,31% des VP. Bien que la perte de signal soit massive, Ulysse détecte toujours après filtrage 175 vraies DEL pour 367 DEL au total soit une valeur prédictive (PPV, Positive Predictive Value) qui passe de 0,76% à 45% (augmentation de 59 fois). En ce qui concerne la sensibilité après filtrage des faux positifs (FP), Ulysses est juste derrière Breakdancer qui détecte lui 197 DEL. Cependant, la PPV de BreakDancer est de seulement 0,009 car ce logiciel est incapable de filtrer efficacement les FP ce qui rend en pratique ses résultats inexploitables sans des étapes manuelles de filtrage. Afin d'observer les performances de chacun des algorithmes en fonction de la couverture en RP des SV j'ai séparé les DEL identifiées par chaque logiciel et par classe de couvertures. À basse couverture (entre 2 et 20 RP), c'est-à-dire pour les DEL qui ont potentiellement une fréquence allélique faible, j'ai montré qu'Ulysses possède une valeur prédictive nettement meilleure que les autres logiciels (Article 1, figure 2B, page 95). De plus, bien qu'Ulysses ait été conçu pour travailler avec des données MP, j'ai également montré avec

des librairies PE qu'Ulysses possède une meilleure spécificité. En résumé, nous avons montré que sur des données de séquençage réelles du génome humain, Ulysses présente une sensibilité équivalente à celle des logiciels existants, mais que sa spécificité pour la détection des DEL est nettement supérieure.

Il a ensuite été nécessaire d'évaluer les performances d'Ulysses selon les types de SV et, pour cela j'ai réalisé des simulations de SV dans le génome humain. J'ai écrit un script permettant de générer aléatoirement des délétions (DEL), des inversions (INV), des duplications en tandem (DUP), des insertions (INS), des translocations réciproques (RT) et des translocations non réciproques (NRT) dans un génome complet. J'ai ainsi pu générer 3 jeux de données avec des couvertures de 10X, 30X et 60X. Chacun de ces jeux de données est composé de séquences MP qui correspondent à 3 chromosomes humains comprenant différents types de SV à des fréquences alléliques variables allant de 6% à 52%. Ces jeux de données m'ont permis d'observer que pour les SV présentes à des fréquences alléliques élevées, Ulysses possède, tout comme les autres logiciels, une excellente valeur prédictive (Article 1, figure 3A, page 95). En revanche, pour les SV présentes à de faibles fréquences alléliques, Ulysses est le seul logiciel à atteindre une PPV de ~70% alors que le second meilleur logiciel (Delly) n'obtient qu'une PPV de 11%. Ces simulations ont également montré que la spécificité d'Ulysses pour la détection des SV interchromosomiques (RT, NRT, INS) rares (proportion relative par rapport aux séquences sauvages < 0,03) est jusqu'à 210 fois meilleure que les autres logiciels. Cet écart de performance s'explique notamment par le fait que lorsqu'une SV implique la formation de 2 jonctions dans le génome (RT et INS), son annotation en tant que SV par Ulysses repose à la fois sur la détection et sur la liaison de ces 2 jonctions alors que l'annotation faite par les autres logiciels ne requière que l'une ou l'autre. Pour finir, j'ai fait varier entre 0.1% et 5% la proportion de séquences chimériques dans les jeux de données simulés et étudié la capacité d'Ulysses et des autres logiciels à détecter des SV (Article 1, figure 3B, page 95). Quelle que soit la proportion de chimères, Ulysses présente une meilleure PPV que les autres logiciels lors de la détection de SV présentes en faibles proportions alléliques. Cet écart se réduit en augmentant la fréquence des SV et devient nul pour les SV ayant une proportion égale à 0,52 par rapport aux séquences sauvages.

Puisque les données sur le génome humain et les simulations ont montré qu'Ulysses permet d'obtenir une image fine des SV dont la fréquence allélique est très faible, nous avons voulu tester son efficacité pour détecter l'hétérogénéité clonale en termes de SV. Nous avons pour cela testé notre approche sur des données issues d'une patiente atteinte d'un cancer du sein. Dans la cas de ce type de tumeur, il a été montré que les duplications segmentales et les

amplifications de l'ADN sont des évènements précoces et récurrents de la tumorigénèse (Inaki et al., 2014). Nous avons en effet confirmé ces résultats en détectant un fort niveau de duplications somatiques dans l'échantillon tumoral (Article 1, supplementary figure 4, page 95).

En conclusion, Ulysses est le premier logiciel permettant la détection de SV rares dans un échantillon avec une bonne spécificité. Associé à l'utilisation de librairies de séquençages avec une grande taille d'insert, ce logiciel est capable de détecter tous les types de SV. Compte-tenu de nos connaissances croissantes sur la plasticité des génomes, sur l'instabilité des cellules cancéreuses ou encore sur l'hétérogénéité des tissus, Ulysses présente un intérêt scientifique majeur. L'article qui suit est notre présentation du logiciel Ulysses qui fût publiée dans le journal *Bioinformatics* en 2015.

Article 1

Ulysses: Accurate detection of low-frequency structural variations in large insert-size sequencing libraries

Gillet-Markowska, A., Richard, H., Fischer, G. & Lafontaine, I.

Bioinformatics 2015, doi: 10.1093/bioinformatics/btu730

Bioinformatics, 31(6), 2015, 801–808
doi: 10.1093/bioinformatics/btu730
Advance Access Publication Date: 7 November 2014
Original Paper

OXFORD

Genome analysis

Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries

Alexandre Gillet-Markowska[1,2], Hugues Richard[1,2], Gilles Fischer[1,2,*] and Ingrid Lafontaine[1,2,*]

[1]Sorbonne Universités, UPMC University Paris 06, UMR 7238, Biologie Computationnelle et Quantitative and [2]CNRS, UMR7238, Laboratory of Computational and Quantitative Biology, F-75005 Paris, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on June 17, 2014; revised on October 24, 2014; accepted on October 27, 2014

Abstract

Motivation: The detection of structural variations (SVs) in short-range Paired-End (PE) libraries remains challenging because SV breakpoints can involve large dispersed repeated sequences, or carry inherent complexity, hardly resolvable with classical PE sequencing data. In contrast, large insert-size sequencing libraries (Mate-Pair libraries) provide higher physical coverage of the genome and give access to repeat-containing regions. They can thus theoretically overcome previous limitations as they are becoming routinely accessible. Nevertheless, broad insert size distributions and high rates of chimerical sequences are usually associated to this type of libraries, which makes the accurate annotation of SV challenging.

Results: Here, we present Ulysses, a tool that achieves drastically higher detection accuracy than existing tools, both on simulated and real mate-pair sequencing datasets from the 1000 Human Genome project. Ulysses achieves high specificity over the complete spectrum of variants by assessing, in a principled manner, the statistical significance of each possible variant (duplications, deletions, translocations, insertions and inversions) against an explicit model for the generation of experimental noise. This statistical model proves particularly useful for the detection of low frequency variants. SV detection performed on a large insert Mate-Pair library from a breast cancer sample revealed a high level of somatic duplications in the tumor and, to a lesser extent, in the blood sample as well. Altogether, these results show that Ulysses is a valuable tool for the characterization of somatic mosaicism in human tissues and in cancer genomes.

Availability and implementation: Ulysses is available at http://www.lcqb.upmc.fr/ulysses.

Contact: ingrid.lafontaine@upmc.fr or gilles.fischer@upmc.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Our current understanding of the structural and functional impact of SV onto the biology of genomes has largely benefited from the development of the second generation of DNA sequencing technologies. The computational detection of SV has mainly relied on the development of four methodological strategies, the 'read-depth' method (Campbell *et al.*, 2008; Alkan *et al.*, 2009; Chiang *et al.*, 2009; Yoon *et al.*, 2009; Mills *et al.*, 2011), the 'split-read' method (Lam *et al.*, 2010; Zhang *et al.*, 2011; Jiang *et al.*, 2012), the *de novo* genome assembly (Wang *et al.*, 2011) and the 'Paired-End' method [PEM (Chen *et al.*, 2009; Korbel *et al.*, 2009; Lee *et al.*, 2009; Hormozdiari *et al.*, 2010; Quinlan *et al.*, 2010; Zeitouni

et al., 2010; Qi and Zhao, 2011; Marschall et al., 2012; Sindi et al., 2012; Hart et al., 2013)]. Some detection tools have increased SV detection specificity and breakpoint resolution by combining several of these detection strategies (Ye et al., 2009; Medvedev et al., 2010; Abyzov and Gerstein, 2011; Handsaker et al., 2011; Rausch et al., 2012; Yang et al., 2013).

Computational analyses using these approaches demonstrated that large SV are major contributors to the genomic polymorphism between individuals (Conrad and Hurles, 2007; Korbel et al., 2007; Kidd et al., 2008; Mills et al., 2011). Polymorphic SV were shown to contribute to both common diseases and rare genomic disorders and to alter normal gene function during cancer development (Fanciulli et al., 2007; Hollox et al., 2008; Stephens et al., 2009; Pinto et al., 2010; Girirajan et al., 2011). New approaches have also started to reveal the quantitative importance of somatic SV in healthy tissues such as neuron or blood cells (Singer et al., 2010; Laurie et al., 2012; McConnell et al., 2013; Voet et al., 2013).

However, the true level of somatic mosaicism probably remains underestimated, owing to the limitations inherent in classical short-range PE libraries. More SV are now theoretically accessible thanks to the recent development of long-range Mate Pair (MP) libraries in which the two reads can be separated by several kilobases. MP libraries present major advantages over classical PE libraries because large inserts can span over large repeated regions often involved in SV formation and because MP libraries provide, for the same number of reads, a much higher physical coverage of the genome. Higher physical coverage triggers the possibility of uncovering SV that are present at low frequency in mosaic genomes. However, MP libraries involve a ligation step during the library construction which generates a large amount of chimerical Read Pairs (RPs), making those library prone to higher rates of false-positive SV. In addition, MP libraries suffer from wide insert size (IS) distributions, which bring additional noise to the detection of deletion and insertion events. These limitations explain why currently available SV detection tools that were developed for short-range PE libraries perform badly on MP data.

Here, we report a new PEM-based software, called Ulysses, specifically designed to detect SV in MP datasets. Ulysses comprises a SV scoring module, which improves SV detection accuracy in MP libraries. Our algorithm can annotate the full spectrum of SV, including deletions (DEL), segmental duplications (DUP), inversions (INV), small insertions (sINS, with a size smaller than the library IS), large insertions (INS), reciprocal translocations (RTs) and non-reciprocal translocations (NRT). Benchmarks on real MP sequencing datasets from the 1000 Human Genome project, on MP simulated datasets as well as on a breast cancer tumor MP library showed that Ulysses outperforms three commonly used detection tools [Breakdancer (Chen et al., 2009), GASVpro (Sindi et al., 2012) and Delly (Rausch et al., 2012)] for all types of SV and notably for low frequency structural variants in MP libraries. In addition, Ulysses is on par with or outperforms the three other tools on PE datasets, making it a highly versatile detection tool.

2 Methods

2.1 Overview of Ulysses

Ulysses is a PEM-based algorithm, which comprises two independent parts: library parsing (Steps 1–2) and SV detection (Steps 3–5, Fig. 1A). The algorithm automatically tunes parameters for SV detection from the set of statistical properties derived from the library parsing. Then, Ulysses builds simple undirected graphs describing groups of discordant RP that consistently support the existence of the same structural variation (SV). The problem of using cliques to predict SV has already been exactly solved and largely implemented in the past few years (Lee et al., 2008; Hormozdiari et al., 2009; Sindi et al., 2009). Cliques are defined in Ulysses in a way closer to (Rausch et al., 2012). However, because a parameter ensures that all IS within a clique are in a comparable size range (ISc_n, see below), Ulysses adds new constraints on cliques. Note that our clustering rules only reflect our implementation rather than an exact solution to the problem of cliques identification. Finally, Ulysses assesses the statistical significance of each candidate variant in a principled manner using for each type of SV an explicit model for the generation of chimerical RP (Fig. 1A). The five main steps are detailed below and further details can be found in the Supplementary Material.

2.1.1 Step 1—Statistics of the library and detection parameters

Starting from a library alignment file (BAM format), Ulysses derives summary statistics [read-pair orientation, empirical IS distribution (f_l), IS median (μ) and median absolute deviation (σ)] from 1 million of randomly sampled RP. The median and median absolute deviation estimates were preferred over mean and standard deviation as they are more robust to outliers. These values are used to set SV detection parameters described below (d_n, ISc_n, $p_{d_n/\ell_k}(\tau)$ and p_{IS}, Fig. 1B).

Two descriptors, d and ISc, are defined to assess whether two RP are consistent (Fig. 1B). Given two RP of size IS_1 and IS_2, spanning the genomic intervals $[l_1, r_1]$ and $[l_2, r_2]$, respectively, we consider:

– their maximal interdistance: $d = \max(|l_1 - l_2|, |r_1 - r_2|)$,
– their IS difference: $ISc = |IS_1 - IS_2|$.

Two RP are consistent if they satisfy $d \leq \mu + n\sigma$ and $ISc \leq n\sigma$ (note that ISc only applies to intra-chromosomal RP). For further reference, we will name those thresholds d_n and ISc_n (see Section 2.5). In addition, two probabilities, $p_{d_n/\ell_k}(\tau)$ and p_{IS}, are defined to assess the statistical significance of groups of consistent RP (Fig. 1B):

– $p_{d_n/\ell_k}(\tau)$ the probability that τ RP have consistent positions in a chromosome of length ℓ_k,
– $p_{IS}(\tau)$ the probability that τ RP have consistent IS.

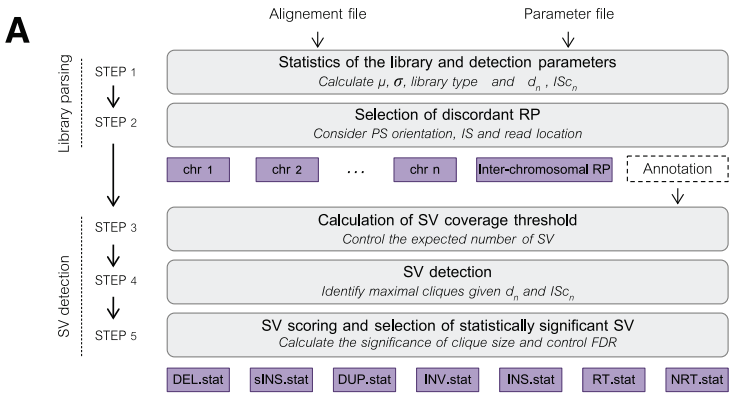
Details about the computation of both probabilities are given in Supplementary Material.

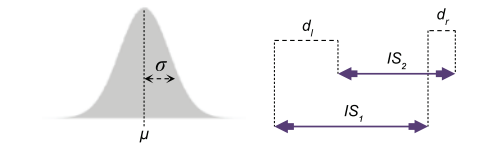
2.1.2 Step 2—Selection of discordant RP

To define SV, Ulysses relies on the identification of discordant RP (with mapping quality $\geq$20), i.e. RP that map incongruously onto the reference genome (Fig. 1C). RP are considered as discordant when they fulfill at least one of the three following criteria:

– Incongruous RP orientation: any RP not in a [−,+] orientation (throughout the text, read orientations are given for MP and must be reversed for PE libraries).
– IS deviating from the expected range: any RP with an IS outside the range $[\mu - n\sigma, \mu + n\sigma]$.
– Incongruous read location: any RP with the two reads on two different chromosomes.

At the end of the library parsing part (Steps 1 and 2), a set of alignment files containing all identified discordant RP is produced (Fig. 1A). The following SV detection part (Steps 3–5) is independent and can be run separately.





	Parameter	Threshold	Probability (τ RP)
Interdistance	$d_n = \mu + n.\sigma$	$\max(d_l, d_r) \leq d_n$	$p_{d_n/\ell_k}(\tau)$
IS difference	$ISc_n = n.\sigma$	$\lvert IS1 - IS2 \rvert \leq ISc_n$	$p_{IS}(\tau)$

C

	SV topology	RP orientation	IS	Read location	Probability π_{SV} of τ consistent RP
Single junction SV	DEL (Ref (chr k) / SV)	[-,+]	$\geq \mu + n.\sigma$	Same chr.	$\pi_{DEL/sINS/DUP} = p_{d_n/\ell_k}(\tau) \cdot p_{IS}(\tau)$
	sINS (Ref (chr k) / SV)	[-,+]	$\leq \mu - n.\sigma$	Same chr.	
	DUP (Ref (chr k) / SV)	[+,-]	Not constrained	Same chr.	
	NRT (Ref (chr k, l) / SV)	[-,-] [+,+] [-,+] [+,-]	-	Diff. chr	$\pi_{NRT} = p_{d_n/\ell_k}(\tau) \cdot p_{d_n/\ell_l}(\tau)$
Two junction SV	INV (Ref (chr k) / SV)	[-,-] [+,+]	Not constrained	Same chr.	$\pi_{INV} = p_{2d_n/\ell_k}(2\tau) \cdot p_{IS}(2\tau) \cdot p_{d_n/\ell_k}(\tau{-}1) \cdot p_{IS}(\tau{-}1)$
	RT (Ref (chr k, l) / SV)	[-,-] [+,+] [+,+] [-,-] [-,+] [+,-] [+,-] [-,+]	-	Diff. chr	$\pi_{RT} = p_{2d_n/\ell_k}(2\tau) \cdot p_{2d_n/\ell_l}(2\tau)$
	INS (Ref (chr k, l) / SV)	[-,-] [+,+] [+,+] [-,-] [-,+] [+,-] [+,-] [-,+]	-	Diff. chr	$\pi_{INS} = p_{2d_n/\ell_k}(2\tau) \cdot p_{2d_n/\ell_l}(\tau)$

ℓ_k length of chromosome k

Fig. 1. Ulysses design. (**A**) Flowchart of the Ulysses algorithm indicating all five processing steps (grey) and output files (light purple). The program is composed of two independent parts, the library parsing and the SV detection, comprising two and three steps, respectively. The five steps are detailed in Section 3 (Overview of Ulysses). (**B**) Consistency parameters. Distribution of IS (top left) and schematic representation of two overlapping RP (purple arrows, top right) allow defining the interdistance and the IS difference parameters, thresholds and probabilities of having τ discordant RP (see Supplementary Material). ISc_n is a threshold only applicable to intra-chromosomal RP. (**C**) SV detection characteristics. For each type of SV, the description of discordant RP, including their topology (left panel), properties (middle panel) and probabilities (right panel) is detailed. Mapping read orientations are intended for MP libraries [−,+] and should be reversed for PE libraries. The π_{SV} formula gives the probability for τ discordant RP to be consistent (see the paragraph 'Overview of Ulysses'). *k* and *l* are two different chromosomes of sizes ℓ_k and ℓ_l. Note that π_{SV} can be directly expressed as a product between $p_{d_n/\ell_k}(\tau)$ and $p_{IS}(\tau)$ because the same value of *n* is used for both d_n and ISc_n and therefore there are only two degrees of freedom to fully define the three parameters $|l_1 - l_2|$, $|r_1 - r_2|$ and $|IS_1 - IS_2|$

2.1.3 Step 3—Calculation of SV coverage minimal threshold

The minimal number τ of consistent RP that is required to support each type of SV is set in order to limit the expected number of candidate SV (N_{SV}). This is especially justified when libraries have a wide distribution of ISs and/or a high number of chimerical reads that will generate a high number of false positives (FPs) (as it is often the case with MP libraries).

For each type of SV, if π_{SV} is the probability that x discordant RP are consistent, the number of expected candidate SV supported by x RP is $N_{SV} = \binom{D_{RP}}{x} \cdot \pi_{SV}$ with D_{RP} being the total number of

discordant RP of this type. The probability π_{SV} depends on the topology of the SV and can be explicitly formalized using the probabilities $p_{d_n/\ell_k}(\tau)$ and $p_{IS}(\tau)$ defined above (Fig. 1C). In practice, N_{SV} is limited to be at most 10 000. The value of τ (default $\tau = 2$) is automatically increased while N_{SV} is above this limit.

2.1.4 Step 4—SV detection

Discordant RP are categorized into one of the seven SV types reported in Figure 1C. RP are then sorted by chromosomes and coordinates, and used to build a simple undirected graph. Each discordant RP represents one vertex of the graph and one edge is drawn between each pair of consistent RP. Ulysses defines a SV as a maximal clique in the graph where all consistent RP are connected to each others. Note that the same vertex can belong to two or more different maximal cliques (meaning that a given RP can be used to define different SV).

DEL, sINS and DUP are SV that produce a single new DNA junction in the rearranged chromosome when compared with the reference genome. For those single-junction SV, we directly use the corresponding maximal cliques to identify candidates SV. Each INV, INS or RT produces two new DNA junctions when compared with the reference genome. These two-junction SV are consequently detected by two maximal cliques, which, in Ulysses, need to be interconnected (by RP orientation and relative coordinates, Fig. 1C and Supplementary Material). This requirement increases the specificity of the detection in comparison with methods that consider only one junction, even for the detection of two-junction SV. All combinations of compatible pairs of cliques are conserved, thus allowing the same RP to be re-used in several different pairs. In addition, when the position of the centromeres is provided, the relative orientation of the two cliques is checked and an RT event will be reported only when both of the rearranged molecules contain a single centromere. Otherwise, the corresponding SV is classified as an INS. For both INS and RT, the number of RP must be equally distributed between the two cliques (ratio > 0.1). Otherwise, the corresponding SV is classified as a NRT.

2.1.5 Step 5—SV scoring and selection of statistically significant SV

Ulysses evaluates whether the number m of RP that supports each candidate variation is significant. This step is essential to filter out false-positive predictions from the artefacts in the library. The significance level of each candidate is then corrected by controlling the false-discovery rate (FDR).

Deletions and small insertions: The statistical significance of each candidate deletion is estimated by calculating the probability b_m to randomly sample at least m RP identifying this deletion, given the local physical coverage C. Given the smallest RP describing the deletion with an IS s, each RP drawn from the IS distribution has a probability prop_{IS} to be consistent with the smallest RP ($\text{prop}_{IS} = \sum_{l \geq s} f_l$, see Supplementary Material). The probability b_m will result from the binomial sampling of the RP:

$$b_m = \sum_{i \geq m} \binom{C}{i} \text{prop}_{IS}^{i} \cdot (1 - \text{prop}_{IS})^{C-i}.$$

In practice, C is estimated over the region corresponding to the DEL position (considering at most the 10 first kilobases of the deleted region). The significance of small insertions can be evaluated in the same manner, by flipping around the IS distribution.

Segmental duplications, non-reciprocal translocations and two-junction SV: For those SV, the statistical significance is computed with a binomial distribution as being the probability ν_m that at least m RP among all combinations of discordant RP of each type are consistent, given the probability π_{SV} (see Fig. 1A):

$$\nu_m = \sum_{i \geq m} \binom{C_m}{i} \pi_{SV}^{i} \cdot (1 - \pi_{SV})^{C_m - i},$$

where $C_m = \binom{D_{RP}}{m}$ is the number of ways of having m RP among a total of D_{RP} discordant ones on the SV considered (see Step 3 and Fig. 1C).

ν_m is calculated for each chromosome independently (or each pair of chromosomes for inter-chromosomal rearrangements).

After SV scoring, we set up a P-value cut-off by controlling the FDR (default value 0.01). Q-values are estimated using a bootstrap method (Storey *et al.*, 2004). During our tests with experimental and simulated data, this approach greatly improved the specificity of the detection (see Section 3).

2.2 Sequencing datasets

The details and the accession numbers for all real sequencing datasets used in this study (NA12878 and breast cancer BT71) are provided in Supplementary Methods.

For simulated sequencing datasets, a MP ($\mu = 2000$ bp, $\sigma = 1487$ bp, normal distribution, read-length = 50 bp) and a PE ($\mu = 233$ bp, $\sigma = 10$ bp, normal distribution, read-length = 50 bp) Illumina-like datasets were generated with wgsim 0.3.1–*r*13 (Li *et al.*, 2009) with default parameters using a 162 Mb human genomic region as a reference (GRCh37/hg19 chromosomes 20, 21 and 22) at three different levels of coverage (10×, 30× and 60×). The MP dataset was generated with variable proportions of random chimerical RP (0.1%, 1%, 2.5% and 5%) in order to sample different levels of experimental artefacts that can derive from the circularization step during MP library construction. Chimerical RP were generated by randomly sampling the RP and by shuffling the mates. As a consequence, chimerical RP can have any orientations and be either intra or inter-chromosomal. The simulated reads were remapped on the reference genome using BWA aln 0.7.3*a*–*r*367 with default parameters (Li *et al.*, 2009). Artificial SV were then added to both PE and MP simulated library datasets after the remapping step in order to avoid mapping artefacts that could bias SV detection. Artificial SV were designed with variable numbers of discordant RP (with 4 and 8 RP in the 10× dataset; with 4, 8, 16 and 32 RP in the 30× dataset and with 4, 8, 16, 32 and 64 RP in the 60× dataset). These numbers correspond to a relative coverage varying from 0.06 (4 RP in the 60× library dataset: $4/(60+4) = 0.06$) up to 0.52 (64 RP in 60× dataset: $64/(60+64) = 0.52$). Each set of artificial SV comprises 50 SV of each type (DUP, DEL, INV, INS, RT and NRT) for each relative coverage, generating a total of 600, 1200 and 1500 SV for 10×, 30× and 60× datasets, respectively. DUP, DEL and INV were generated with random sizes, varying from 1 to 50 kb. The detection of one SV was considered as true positive (TP) when at least one of the RP that defined the simulated SV was recovered. All simulated datasets are available at http://www.lcqb.upmc.fr/ulysses/simulations.

2.3 Benchmark with other detection tools

All benchmarks analyses were performed using the AMADEA Biopack platform developed by ISoft (http://www.isoft.fr/bio/biopack_en.htm). Breakdancer v1.2.6 (Chen *et al.*, 2009), GASVpro-HQ v1.2 (Sindi *et al.*, 2012) (referred to as GASVpro) and Delly v0.5.6 (Rausch *et al.*, 2012) were configured to detect SV defined by at least two consistent RP (which is the lowest limit to define a

clique). All other parameters remained set by default. The sensitivity (Sn) of a detection tool is the proportion of true SV that are detected ([TPs]/[TPs + false negatives]) and its precision (or positive predictive value, PPV) is the proportion of true SV among all detected SV ([TPs]/[TPs + FPs]). The F1-score is an estimator of the trade-off between Sn and PPV ($2 \times Sn \times PPV/(Sn + PPV)$). Supplementary Tables S1, S2 and S3 report Sn, PPV, F1-scores and running times of Ulysses, Delly, Breakdancer and GASVpro on all simulated datasets.

Note that GASVpro cannot perform SV detection for MP datasets (both simulated and experimental). The execution of GASVpro on these data was stopped after more than 166 h of computation on a single CPU after splitting the sequence files by chromosomes (core i7 870, 32GB RAM). As a consequence, GASVpro does not appear in any of the MP analyses.

2.4 Ulysses detection parameter *n*

The performance of Ulysses was tested as a function of the n value, which controls the main detection parameters for all types of SV (the interdistance d_n and IS difference ISc_n, see Section 3). For both MP and PE simulated datasets, F1-scores remain highly stable for n values ranging from 3 to 10 (Supplementary Fig. S1). For real sequencing dataset (NA12878 MP 30×), a n value set to 6 provides a good compromise between Sn and PPV (Supplementary Fig. S2). This value of $n = 6$, set by default in the parameter file, is therefore well suited for most applications and does not require any manual adjustment by the user.

3 Results

3.1 Performance of Ulysses on MP sequencing datasets

The performance of Ulysses was compared with three other widely used SV detection tools: Breakdancer, GASVpro and Delly. Given that BreakDancer and Delly do not discriminate between INS, RT and NRT, these three types of SV were merged into a single class of events called inter-chromosomal (INTER) for analysis. The performance of each method was estimated with PPV, representing the total number of TPs divided by the total number of detected SV.

3.1.1 Performance on real sequencing data from the 1000 Human Genome Project

The three tools were benchmarked on a real MP sequencing dataset (30× coverage) coming from a single individual (NA12878, see Section 2). We focused on DEL detection because a robust GS containing 2209 DEL was available (see Section 2). Results are presented only for Ulysses, Delly and BreakDancer (SV detection could not finish in reasonable time with GASVpro on MP datasets, even with sequences split by chromosome, see Section 2).

The results on Figure 2A are presented as ROC curve showing the number of TPs DEL as a function of the total number of predicted DEL (TP + FP). The main interest of our approach lies in the scoring method that allows to filter out a large number of FP cliques that directly result from chimerical RP. The scoring module filters out 99.88% of the 166 057 (167 335 − 1278) FP detected. It also filters out 86.31% of the 1278 TP, keeping as statistically significant 175 TP after filtering. Note that the final sensitivity of Ulysses after the scoring step (175 TP) is very close to that of BreakDancer (197 TP) and higher than that of Delly (48 TP). As a result, the scoring method strongly increases Ulysses PPV from 0.76% to 45% (a 59-fold increase), which explains the massive gain in specificity of Ulysses over the other tools (Fig. 2A, Supplementary Table S1).

We have checked that the higher detection accuracy of Ulysses did not result from a biased distribution of DEL sizes in the GS that could have favoured Ulysses over the other tools. The median deletion size of DEL from the GS is 4.9 kb, whereas the median deletion sizes of DEL detected by Ulysses, Delly and Breakdancer are of 13.2, 21.5 and 5.5 kb, respectively. Thus, Ulysses and Delly do not detect small DEL from the GS. BreakDancer is able to predict such small DEL at the cost of precision, with the concomitant detection of about 21 000 FPs.

In order to further characterize Ulysses higher detection accuracy, we plotted the relative precision (proportion of TP DEL or PPV) and the FDR (proportion of FP DEL) of the three tools as a function of DEL physical coverage (Fig. 2B). This plot shows that Ulysses precision remains relatively constant (between 0.24 and 0.57) over the whole range of coverages whereas BreakDancer precision increases with increasing coverage (from 0.0024 to 0.54). Delly achieves correct precision only for intermediate coverage values. Compared with the other tools, Ulysses precision is particularly higher for the lowest physical coverage values (between 1 and 20 RP). These results show that Ulysses is particularly efficient for the detection of low frequency SV.

3.1.2 Detection of somatic mosaicism in tumour sample

We used 8 kb MP data from a luminal A breast tumor to analyse the somatic mosaicism present in cancer cells (data taken from Inaki et al. (2014)). In this paper, the authors suggest that tandem duplications appear to be early events in tumour evolution, especially in the genesis of amplicons. We performed the differential detection of tandem DUP in the blood and the tumour samples with Ulysses and Delly (Supplementary Fig. S4). Nearly all DUP detected in the blood sample with allele frequencies higher than 0.2 were also found in the tumour, as expected for germline SV. Note that Uysses detected 49 germline DUP whereas Delly only found 5 such DUP. We found nearly no evidence of high allele frequency (AF >0.2) somatic DUP in the blood sample whereas in the tumor sample, we detected a significant number of somatic DUP that probably occur early during tumour development as suggested by their AF higher than 0.2 (28 with Ulysses versus 3 with Delly, Supplementary Fig. S4A and B, respectively). The highest level of somatic mosaicism was found by Ulysses as low frequency DUP (AF <0.2) in the tumour sample. Ulysses detected 4551 low frequency DUP whereas Delly found only 256 such DUP (Supplementary Fig. 4A and B, respectively). In this analysis, somatic mosaicism was also detected in the blood sample, as already reported (Jacobs *et al.*, 2012; Laurie *et al.*, 2012), but to a lesser extent than in the tumour sample (Supplementary Fig. S4).

3.1.3 Performance on simulated SV

A MP sequencing dataset with wide IS distributions ($\mu = 2$ kb, $\sigma = 1487$ bp) was simulated at different sequencing coverage values (10×, 30× and 60×, see Section 2), using three human chromosomes as reference. Fifty simulated SV of each type (DUP, DEL, INV, INS, RT and NRT) were added to the libraries at relative physical coverages varying from 0.06 up to 0.52. Intra-chromosomal SVs (DUP, DEL and INV) were generated with random sizes, varying from 1 kb to 50 kb. Note that simulated SV represent the TP SV in this dataset. In addition, variable proportions of random chimerical RP (from 0.1% to 5% of the reads) were added to the library in order to simulate the typical experimental noise that derives from the circularization step during MP library construction.

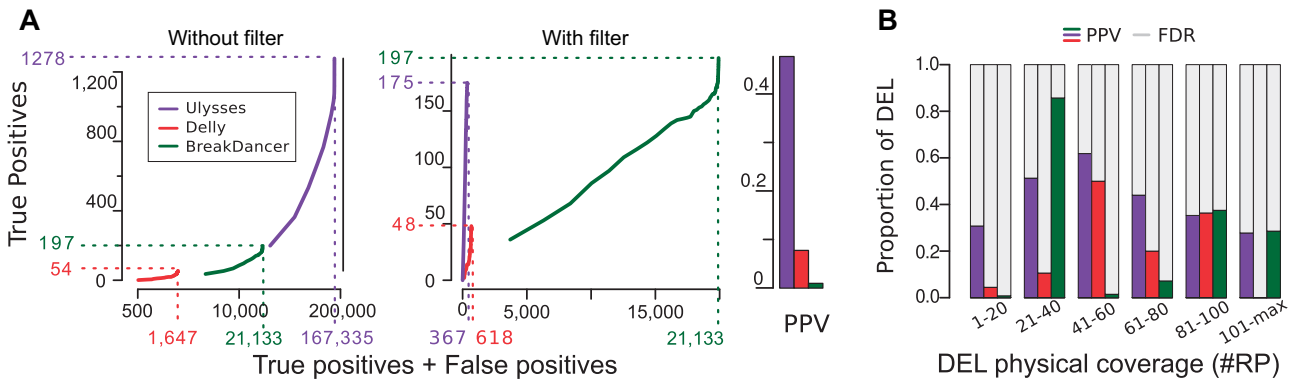


Fig. 2. Deletions in the NA12878 MP 30X dataset. (**A**) Left and middle panels: ROC-like curves representing the number of TP DEL, without and with statistical filter (see Gold Standard in Section 2) as a function of the total number of predictions (TPs + FPs) using the relative SV coverage as a cumulative varying threshold. GASVpro could not be run for the MP library (see text). Right panel: detection accuracy with filter represented as PPV. (**B**) Relative precision (proportion of true-positive DEL) and FDR (proportion of false-positive DEL) of the three tools as a function of DEL physical coverage in MP30X NA12878 dataset. Blueviolet, red and dark green bars represent the precision (PPV) of Ulysses, Delly and Breakdancer, respectively. Grey bars represent the proportion of false-positive DEL (FDR, 1-PPV)

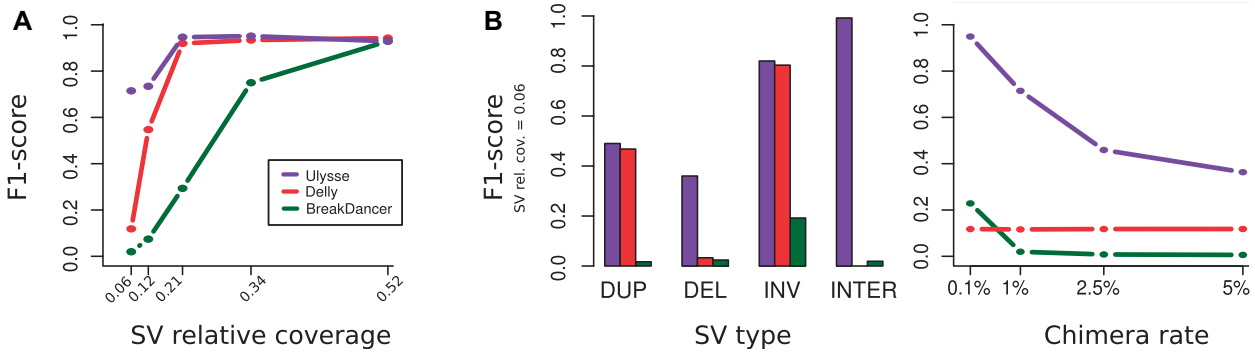


Fig. 3. Results on a 60X MP dataset with RP distribution $\mu = 2$ kb, $\sigma = 1487$ bp and 1% of chimerical RP. (**A**) F1-score as a function of SV relative coverage. Note that GASVpro is absent from the MP simulations because the execution of the program on this dataset exceeded 166 h of computation (see Supplementary Table S2). (**B**) Left panel: F1-score for the different types of SV with a relative coverage of 0.06. The INTER class of SV gathers results from INS, RT and NRT. Right panel: F1-score value are represented as a function of the proportion of chimerical RP present in the dataset (SV relative coverage = 0.06)

No major difference was found across the range of sequencing coverage values. Therefore, the results, presented on Figure 3, only focus on the 60× coverage dataset (see Supplementary Figs. S5–S9 and Supplementary Tables S2 for 10× and 30× results). We first assessed the performance of the three algorithms across variable SV physical coverage values (Fig. 3A). For SV present at high relative coverage (0.52, corresponding to a heterozygous SV in a diploid cell), all three algorithms perform globally well, with similar accuracy as estimated by the F1-scores. However, as soon as the SV relative coverage decreases, the performance of Delly and BreakDancer rapidly deteriorates (Fig. 3A). At the lowest SV relative coverage (0.06), Ulysses retains a detection accuracy of 71% whereas Delly and BreakDancer drop down to 11% and 2%, respectively. We found that detection accuracy of Delly and BreakDancer deteriorates because of high over-prediction of FP SV (see ROC curves in Supplementary Figs. S5–S9). These results show that Ulysses is the only algorithm able to detect low frequency SV (i.e. physical coverage between 0.06 and 0.34) with high accuracy.

Next, we compared the three tools for each type of SV, across all physical coverage values. Again, for the highest relative coverage (0.52), all tools achieve comparable accuracies on all types of SV (Supplementary Fig. S10). For all SV types, the performance of Ulysses compared with the other tools gradually improves with decreasing relative coverage. For the lowest relative coverage value, Ulysses outperforms both Delly and BreakDancer for all types of SV (Fig. 3B left panel).

Because MP libraries often comprise high proportions of chimerical RP, we also tested whether the conclusions drawn here for a dataset containing 1% of chimerical RP also apply to datasets with other levels of chimeras (between 0.1% and 5%). Again, for the lowest SV coverage value (0.06), Ulysses shows the highest F1-scores over the entire range of chimera (Fig. 3B right panel). For relative coverage values from 0.12 to 0.34, the difference between the tools gradually reduces and totally vanishes for the highest coverage (0.52, Supplementary Fig. S10).

3.2 Performance on PE libraries

We also tested Ulysses performance on classical PE libraries, with smaller ISs, and compared it with Delly, BreakDancer and GASVpro. For real sequencing dataset (NA12878), Ulysses and BreakDancer achieve the best detection accuracies at low coverage (5×). For a high coverage 200× dataset, Ulysses clearly outperforms all tools (Supplementary Fig. S11A). Noticeably, GASVpro, which also includes a SV scoring and filtering module, performs very

poorly on low coverage data but reached the second best PPV on the high coverage data. For simulated dataset, RP were generated with characteristics similar to those described above for MP data (in terms of sequencing coverage, SV type and physical coverage, see Section 2). Both Ulysses and BreakDancer behave well at all SV relative coverages and across all types of SV whereas Delly and GASVpro show lower F1-scores for low frequency DEL (Supplementary Fig. S11B).

4 Discussion

Ulysses uses a PEM approach that relies on the identification of groups of discordant RP to detect the full spectrum of SV. This strategy, also implemented in the three other detection tools tested here, is highly sensitive but usually lacks specificity when used alone on MP data with wide IS distribution and high proportion of chimerical RP. To overcome these limitations, we developed in Ulysses a scoring module that statistically assesses the genuineness of all candidates SV, given an explicit model for the generation of chimerical RP. To deal with wide IS distribution, Ulysses evaluates IS consistency between RP and filters out groups of RP with inconsistent IS. To deal with high proportions of chimerical RP (up to 5%), Ulysses uses statistics based on the relative coverage of candidate SV. These two parameters automatically adjust to the characteristics of the library such that no manual tuning is required to achieve detection with good accuracy across all types of SV. As a result, Ulysses is the only tool that performs equally well on MP and PE data. On a real MP sequencing dataset from the 1000 Human Genome Project (NA12878, 30×), Ulysses achieves a better sensitivity than the other tools and by several orders of magnitude the best precision. Ulysses also achieves the highest detection accuracy on PE datasets, showing that globally, Ulysses outperforms the other tools no matter the type of sequencing library.

Ulysses scoring module brings a major benefit by enabling accurate detection of low coverage variants. Low coverage SV can correspond to rearrangements occurring in genomic regions which are difficult to sequence and where the local coverage could dramatically drop (some regions were shown to be consistently prone to low coverage). Alternatively, low coverage SV could also result from rearrangements only present in a small subset of the cell population. This is of particular interest when analyzing samples that contain polymorphic somatic mutations such as cancer samples. The analysis of somatic mosaicism in a breast tumour sample revealed that Ulysses achieves an efficient detection of both germline and somatic DUP. We also showed that Ulysses is able to detect somatic mosaicism in a blood sample. Furthermore, recent insights onto somatic mosaicism showed that subclonal cell heterogeneity is not restricted to cancer cells and could be common between cells from a single tissue sample. Given such an unsuspected level of somatic genome plasticity, the availability of a SV detection tool like Ulysses, with high accuracy for rare SV is of primary importance.

Acknowledgements

The authors thank our colleagues from LCQB for fruitful discussions and Jean-Philippe Meyniel from ISoft for invaluable tips and advices for pipeline developments in AMADEA.

Funding

This work was supported by the Agence Nationale pour la Recherche grant 2010 BLAN1606 and by an ATIP grant from the Centre National de la Recherche Scientifique (CNRS). H.R. was partly supported by a Japan Society for the Promotion of Science (JSPS) fellowship [PE11014].

Conflict of interest: none declared.

References

Abyzov,A. and Gerstein,M. (2011) AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics (Oxford, England)*, **27**, 595–603.

Alkan,C. *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.*, **41**, 1061–1067.

Campbell,P.J. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Chen,K. *et al.* (2009) Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Chiang,D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103.

Conrad,D.F. and Hurles,M.E. (2007) The population genetics of structural variation. *Nat. Genet.*, **39**(Suppl. 7), S30–S36.

Fanciulli,M. *et al.* (2007) Fcgr3b copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.*, **39**, 721–723.

Girirajan,S. *et al.* (2011) Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet.*, **7**, e1002334.

Handsaker,R.E. *et al.* (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.*, **43**, 269–276.

Hart,S.N. *et al.* (2013) SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations. *PLoS One*, **8**, e83356.

Hollox,E.J. *et al.* (2008) Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.*, **40**, 23–25.

Hormozdiari,F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, **19**, 1270–1278.

Hormozdiari,F. *et al.* (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, **26**, i350–i357.

Inaki,K. *et al.* (2014) Systems consequences of amplicon formation in human breast cancer. *Genome Res.*, **11**, 1–13.

Jacobs,K.B. *et al.* (2012) Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.*, **44**, 651–658.

Jiang,Y. *et al.* (2012) PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics (Oxford, England)*, **28**, 2576–2583.

Kidd,J.M. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.

Korbel,J.O. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

Korbel,J.O. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.*, **10**, R23.

Lam,H.Y.K. *et al.* (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.*, **28**, 47–55.

Laurie,C.C. *et al.* (2012) Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.*, **44**, 642–650.

Lee,S. *et al.* (2008) A robust framework for detecting structural variations in a genome. *Bioinformatics (Oxford, England)*, **24**, i59–i67.

Lee,S. *et al.* (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Publ. Group*, **6**, 473–474.

Li,H. *et al.* (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Marschall,T. *et al.* (2012) Clever: clique-enumerating variant finder. *Bioinformatics*, **28**, 2875–2882.

McConnell,M.J. *et al.* (2013) Mosaic copy number variation in human neurons. *Science*, **342**, 632–637.

Medvedev,P. *et al.* (2010) Detecting copy number variation with mated short reads. *Genome Res.*, **20**, 1613–1622.

Mills,R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Pinto,D. *et al.* (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368–372.

Qi,J. and Zhao,F. (2011) inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res.*, **39**(Web Server issue), W567–W575.

Quinlan,A.R. *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.*, **20**, 623–635.

Rausch,T. *et al.* (2012) Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, I333–I339.

Sindi,S. *et al.* (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics (Oxford, England)*, **25**, i222–i230.

Sindi,S.S. *et al.* (2012) An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol., **13**, R22.

Singer,T. *et al.* (2010) Line-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci., **33**, 345–354.

Stephens,P.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.

Storey,J.D. *et al.* (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Statist. Methodol.*, **66**, 187–205.

Voet,T. *et al.* (2013) Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Res.*, **41**, 6119–6138.

Wang,J. *et al.* (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods*, **8**, 652–654.

Yang,L.X. *et al.* (2013) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, **153**, 919–929.

Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, **25**, 2865–2871.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

Zeitouni,B. *et al.* (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics (Oxford, England)*, **26**, 1895–1896.

Zhang,Z.D. *et al.* (2011) Identification of genomic indels and structural variations using split reads. *BMC Genomics*, **12**, 375.

SUPPORTING METHODS

SV detection: Identification of two-junction SV

For INV, Ulysses seeks for pairs of compatible maximal cliques of consistent RP with [-,-] and [+,+] orientations (Figure 1C). Two maximal clique, C1 and C2, are considered as compatible if (i) they have opposite orientations (by convention, maximal cliques with [-,-] and [+,+] orientations are designated C1 and C2, respectively), (ii) the largest coordinate among left reads in C1 is smaller than the smallest coordinate among right reads in C2 and (iii) the distance between the smallest coordinate in C1 and the largest coordinate in C2 must be smaller than $2\times d_n$. Note that the two compatible maximal cliques can overlap if the inverted DNA segment is smaller than $2\times d_n$.

For INS and RT, a similar procedure is applied to all inter-chromosomal maximal cliques. Compatible maximal cliques of RP must have opposite orientations (Figure 1C). By convention, C1 is the maximal clique that contains the RP with smallest coordinate on the chromosome with the smallest lexicographical name. RT and INS are topologically identical and generate similar types of C1 and C2 maximal cliques. However, not all INS are compatible with the criteria required to define a RT. Thus, for RT the left and right coordinates of the reads on each of the two chromosomes involved in the translocation must be bounded within $2\times d_n$ bp and must not overlap while for INS, coordinates are bounded for only one of the two chromosomes (the recipient of the inserted fragment) and can overlap on one chromosome (the donor chromosome) if the inserted DNA segment is smaller than $2\times d_n$. In addition, when the position of the centromeres is provided to the program, the relative orientation of the two cliques is checked and an RT event will be reported only when both rearranged molecules contain a centromere. Otherwise, the corresponding SV is classified as an INS. For both INS and RT, the number of RP must be equally distributed between the two cliques (ratio > 0.1). Otherwise, the corresponding SV is classified as a non-reciprocal translocation (NRT, Fig. 1C).

Computing the probabilities of having τ consistent RP

RP are defined as being consistent based on their maximal inter-distance *d*, and their maximal IS difference *ISc* (Figure 1B). In order to compute the chance to observe τ consistent RP by chance, we calculate the related probabilities $p_{d_n/\ell_k}(\tau)$ and $p_{IS}(\tau)$ separately before calculating the various π_{SV} (Figure 1C).

Computing the probability $p_{d_n/\ell_k}(\tau)$

Considering the left coordinates of τ different RP supposed to be drawn uniformly along a chromosome *K* of length ℓ_k, the maximum spread of their variable can be computed using order statistics on uniform variables. In a first time, we scale the genomic positions uniformly to the interval [0,1]. We can then, by following a classical result on the maximal range of τ variables drawn uniformly on [0,1], compute the spread of the distribution. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(\tau)}$ be a set of τ sorted independent and uniformly distributed variables.

The probability that τ RP have consistent positions (*i.e.* are at most at a distance d_n) in a chromosome of length ℓ_k writes:

$$p_{d_n/\ell_k}(\tau) = P\left(x_{(\tau)} - x_{(1)} \leq \frac{d_n}{\ell_k}\right) = \left(\frac{d_n}{\ell_k}\right)^{\tau-1} \cdot \left(\tau - (\tau-1)\frac{d_n}{\ell_k}\right)$$

Computing the probability $p_{IS}(\tau)$

In essence, the value *ISc* is not different from *d*: it is the maximum range over τ independent and identically distributed random variable. The probability that 2 RP have consistent IS with f_i being the fraction of RP of size *I* is given by:

$$p_{IS}(2) \sim \sum_{i=1}^{\max(IS)} f_i \sum_{j=i-ISc_n \vee 1}^{i+ISc_n \wedge \max(IS)} f_j$$

For τ >2, p_{IS} can be computed by dynamic programming using recurrence formulas. The IS can have any type of distribution and are in general not uniformly distributed. To be able to account for the most general case, we use a combinatorial strategy to compute the distribution of the maximum range ISc. We use a recursive formulation of the maximal IS difference between RP and then apply dynamic programming to compute this probability for any possible values of τ.

Let us introduce the recursion variable $p_{IS}(\tau, k, l)$: the probability that the range of values taken by τ RP is exactly between *k* and *l*.

Given a threshold ISc_n, the probability $p_{IS}(\tau)$ is then simply

$$p_{IS}(\tau) = \sum_{k=1}^{IS_{max}} \sum_{l=k}^{k+ISc_n \wedge IS_{max}} p_{IS}(\tau, k, l)$$

$p_{IS}(\tau, k, l)$ can easily be computed recursively:

$$p_{IS}(2, k, l) = 2f_k f_l \text{ if } k \neq l$$
$$= f_k f_l \quad \text{if } k = l$$

$$p_{IS}(\tau, k, l) = \sum_{k \leq i \leq l} p_{IS}(\tau - 1, k, l) f_i + \sum_{i>k}^{l-1} p_{IS}(\tau - 1, i, l) f_k + \sum_{i>k}^{l-1} p_{IS}(\tau - 1, k, i) f_l$$

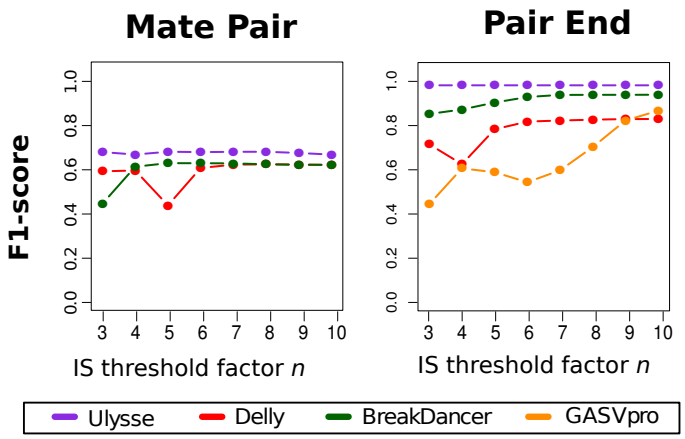
The computational cost is in $O(n.IS_{max}.ISc_n)$. To avoid long running times, we scale the IS values to the range [0-1000] which is extremely fast in practice, with no strong impact on the resulting probability.

Access to real sequencing data

Three datasets from the same individual (NA12878 family 1463) from the HapMap project were used in this study. A low coverage PE 5X dataset (μ= 233 bp, σ = 10 bp) from the 1000

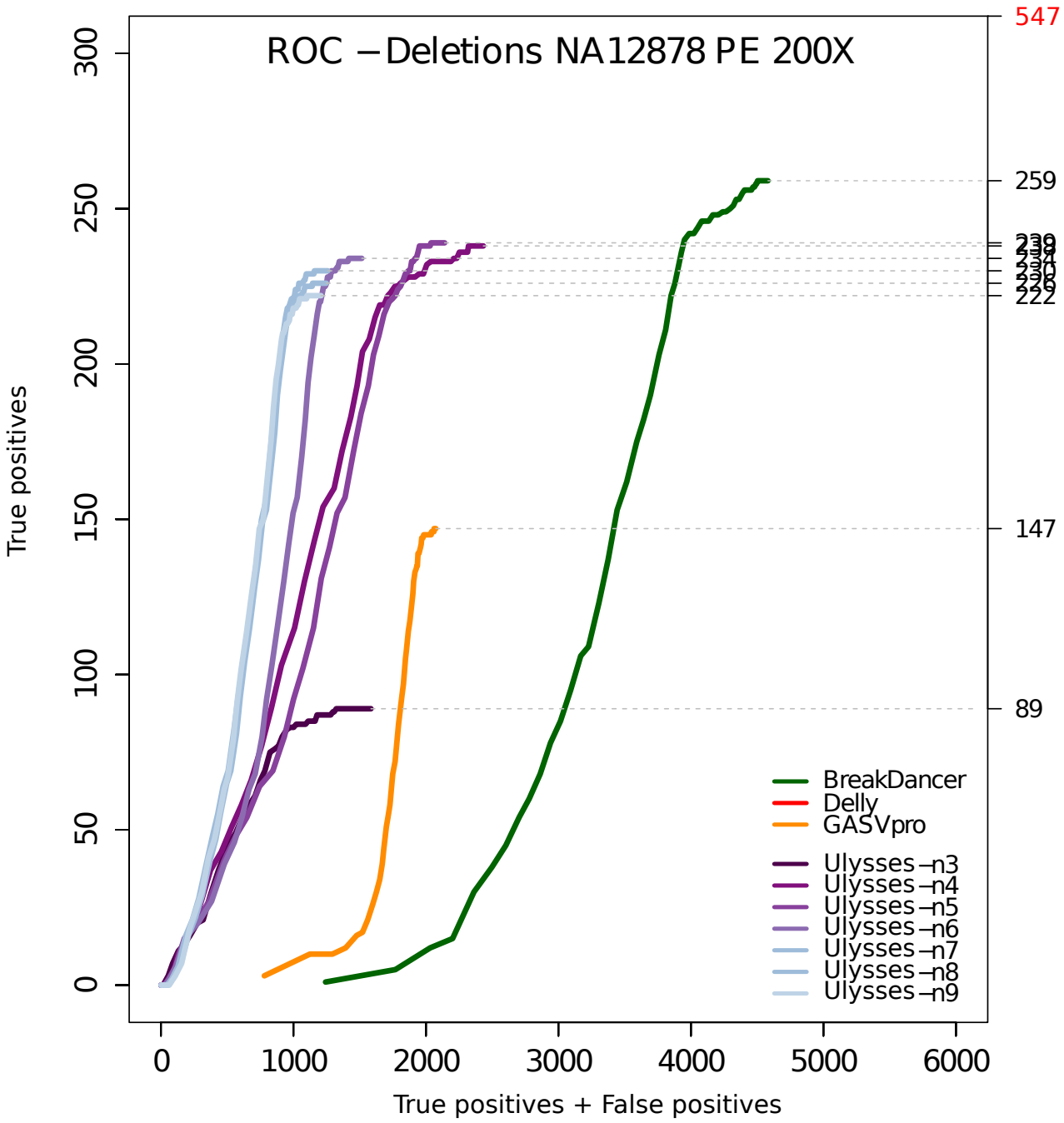
Human Genome project was downloaded from http://www.ebi.ac.uk/ena/ through sample accession number SRS000090 and the corresponding sequences were mapped to the reference genome (GRCh37/hg19) using BWA_aln 0.6.1-r104 with default parameters. A medium coverage MP 30X μ= 2.7 kb; σ = 1450 bp) and a high coverage PE 200X datasets (μ= 322 bp; σ = 75 bp) from the Illumina platinum genomes project were downloaded from http://www.ebi.ac.uk/ena/ with accession numbers ERP002490 and ERP001775, respectively. Detection of DEL was performed using the Gold Standard (GS) provided in Mills et al., 2011 for NA12878, which contained 617 deletions (hg19 liftover), and was extended by adding all deletions of the 1000 Human Genome Phase 3 variant set with an allele frequency (AF) >0.1, resulting in a total of 2209 deletions. A candidate DEL was considered to be a true positive when the difference between its coordinates and the coordinates of the GS deletion (on both sides of the deletion) was lower than 10σ. For ROC curves, true positive DEL were plotted as a function of the total number of detected DEL, with increasing relative SV coverage (in RP) as the varying threshold. Luminal A breast cancer datasets (*BT71* in Inaki et al., 2014GR) can be accessed NCBI Sequence Read Archive under accession number PRJNA253369.

Additional Figure 1



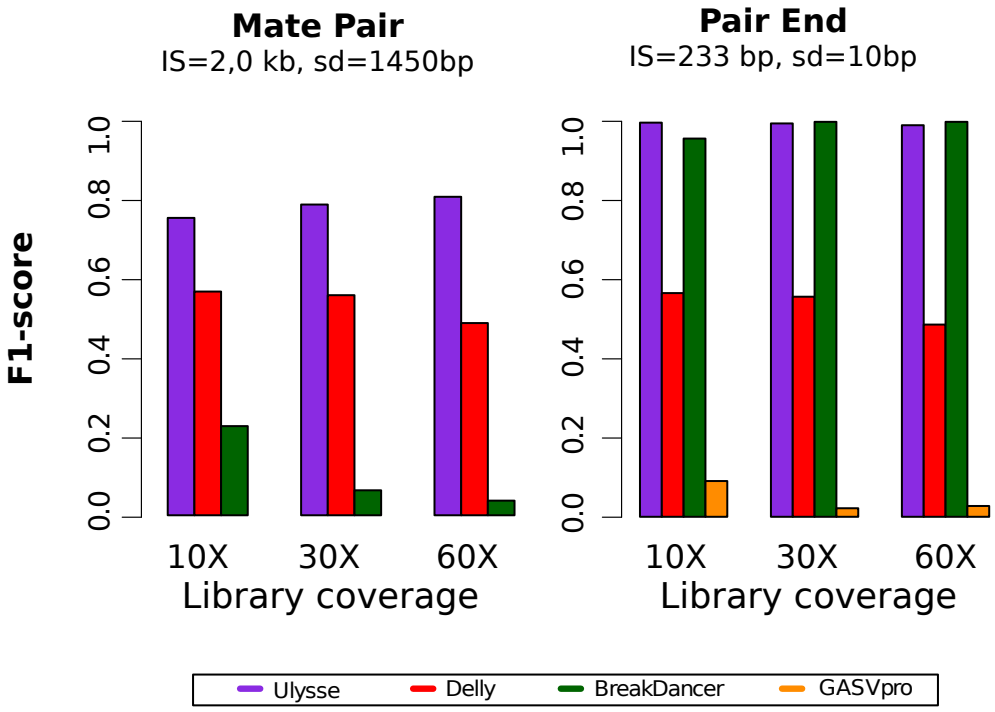
Additional Figure 1 shows, for both MP and PE 60X simulated datasets, the global F1-score value as a function of the factor n used to calculate the IS thereshold (dn = μ+ nσ). Delly, Breakdancer and GASVpro only have an equivalent of dn for DEL detection. Therefore, for these tools, changing n only affected the DEL detection while it affected all SV types in Ulysses.

Additional Figure 2



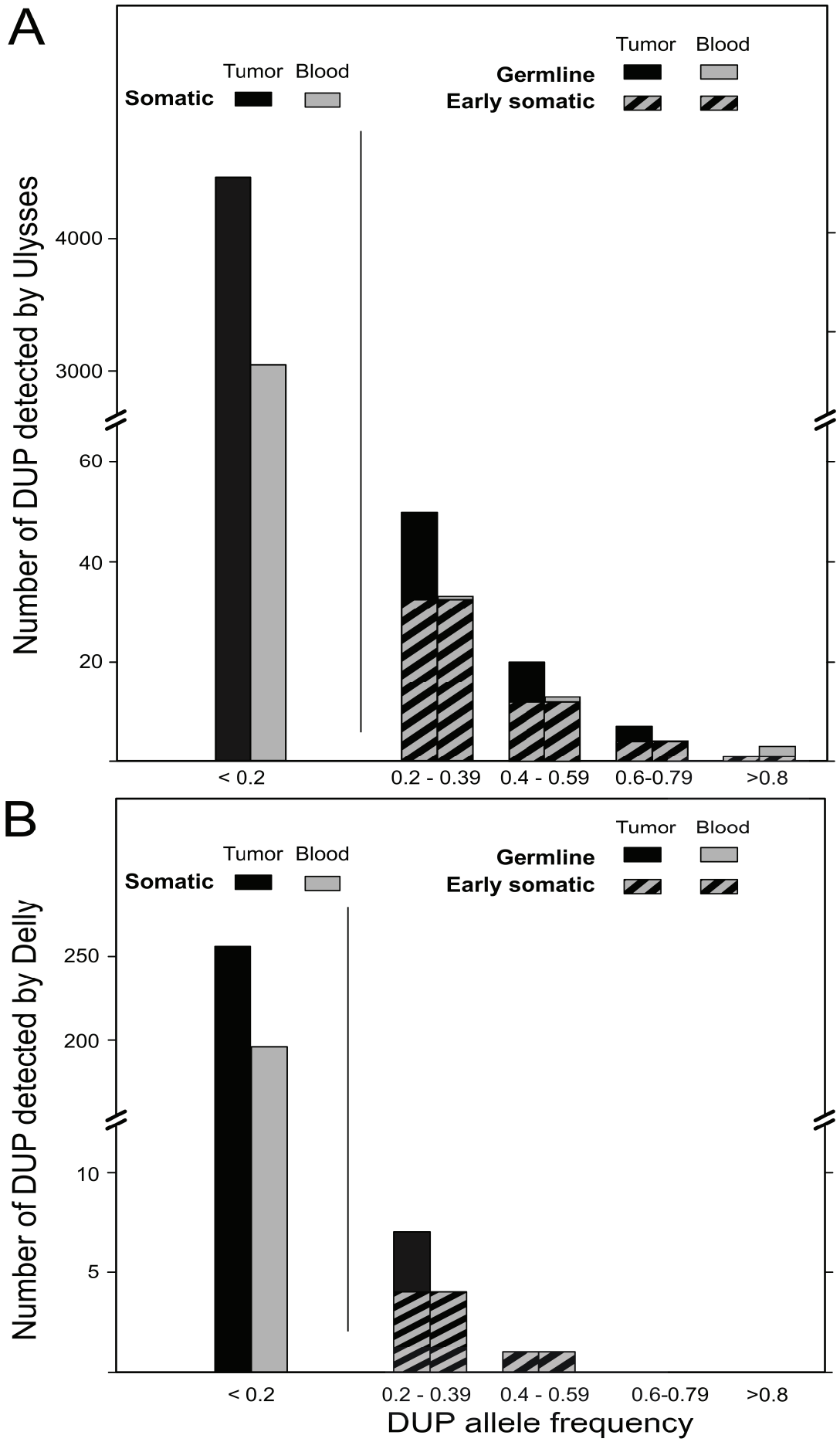
Additional Figure 2 provides ROC-like curves representing the number of true positive DEL (seeGold Standard in Methods section) as a function of the total number of predictions (truepositives + false positives) using the relative SV coverage as a varying threshold for NA12878 and PE200X. Ulysses is represented for various “n” values.

Additional Figure 3



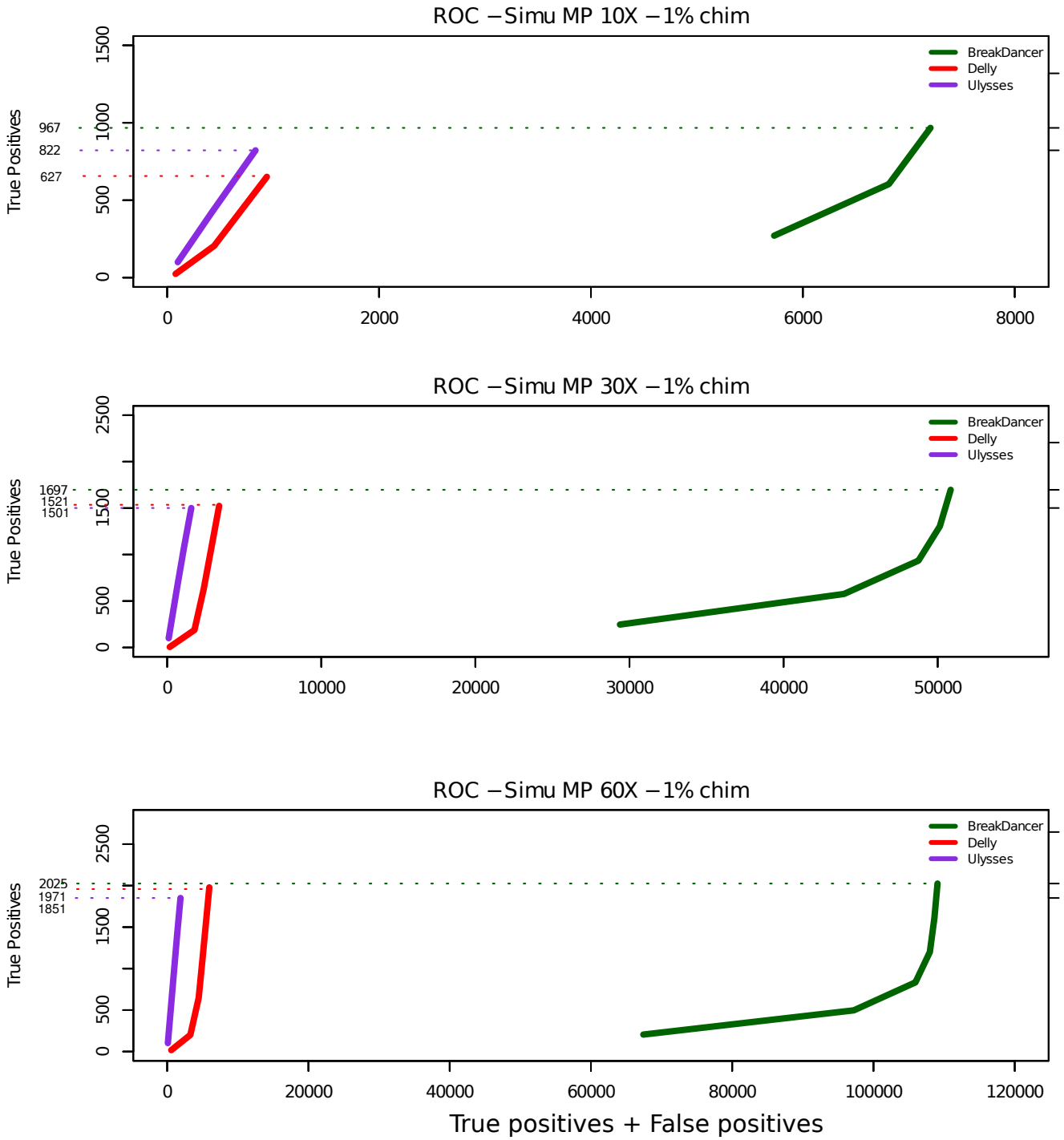
Additional Figure 3 provides the F1-scores for the four algorithms on simulated datasets (bothMP and PE datasets) at three different coverages (10X, 30X and 60X) and with 1% of chimerical PS.

Additional Figure 4



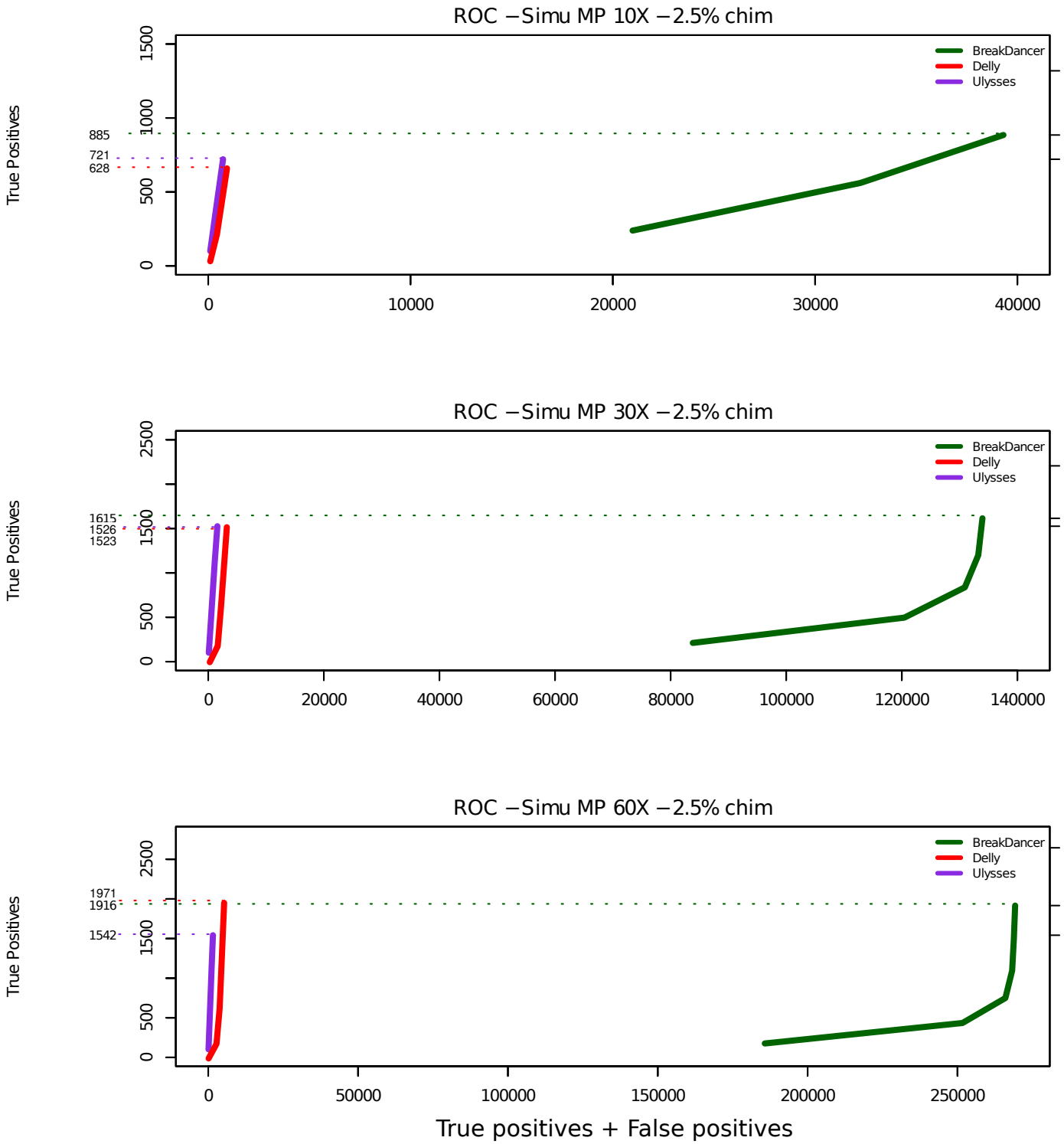
Additional Figure 4 shows the detection of somatic mosaicism in a tumor and in the peripheral blood lymphocytes (BT71 samples in *Inaki et al*, 2014). The number of DUP detected by Ulysses (A) and by Delly (B) is given for different classes of DUP Allele Frequency (AF). AF are computed for each SV as #RP/local physical coverage. All DUP with an AF>0.2 which are detected both in the blood and the tumor samples are called "germline DUP". All DUP that are only detected in one sample are called "somatic". For Ulysses and Delly, the somatic DUP in the class 'AF<0.2' have median AFs between 0.01 and 0.02 for both the tumor and the blood samples. Note that 368 out of the 4,551 somatic DUP (AF < 0.2) detected in the tumor sample by Ulysses (116 out of the 256 for Delly) were also found in the blood sample (not symbolized in the histograms). These DUP could possibly correspond to recurrent events that occurred independently in both samples.

Additional Figure 5



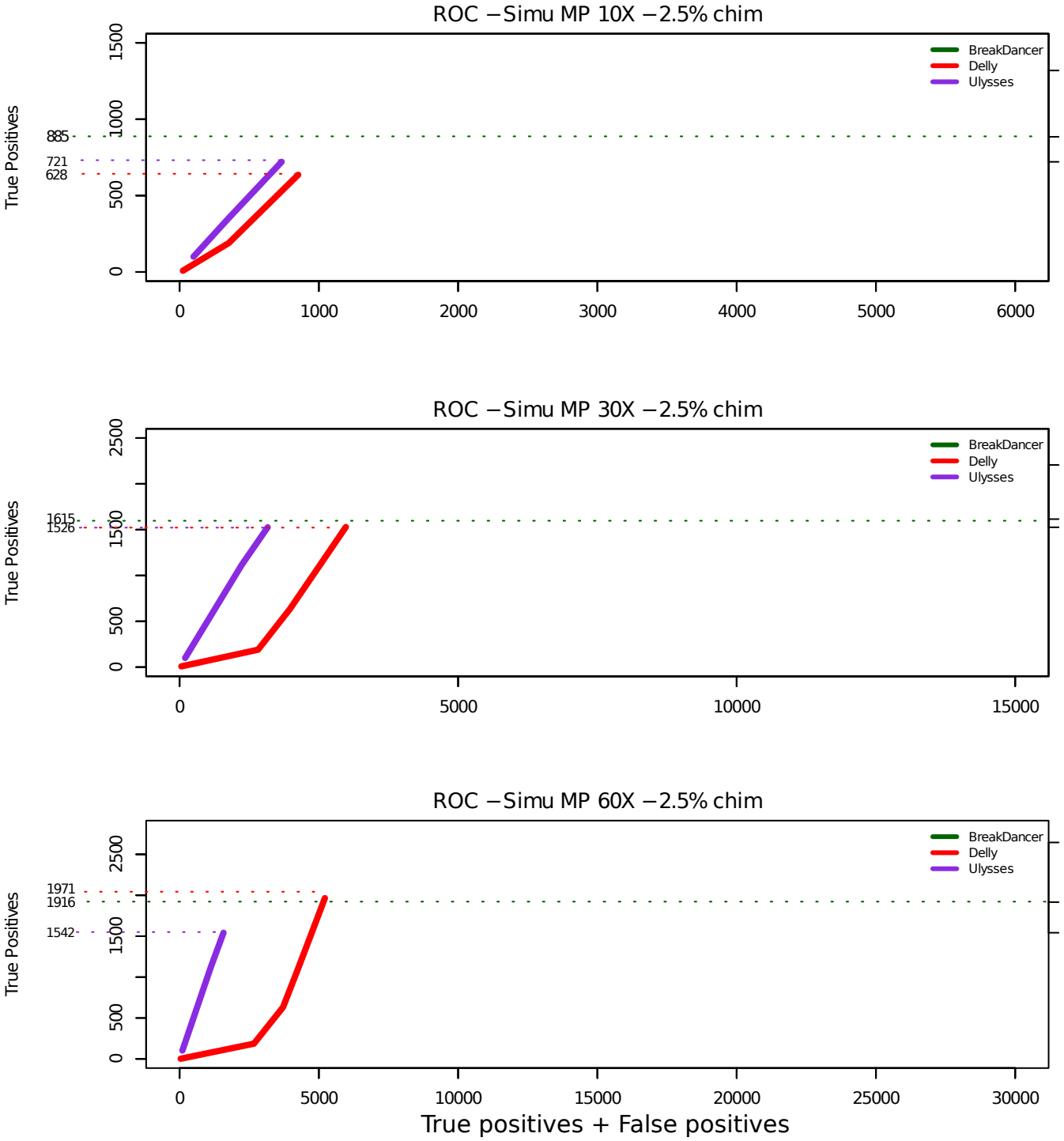
Additional Figure 5 to 9 present ROC curves showing the number of true positive SV detected as a function of the total number of SV predictions (true positives + false positives). The SV relativecoverage is used as a varying threshold. When the values of the 3 tools do not fit in the same plot area (for additional Fig. 6 and 8), another graph with a more adapted scale is provided(additional Fig. 5 and 7). Additional Fig 4: MP data with 1% of chimerical PS. Additional Fig 5: MPdata with 2.5% of chimerical PS. Additional Fig 6: same as Additional Fig. 5 with smaller x-scale.Additional Fig 7: MP data with 5% of chimerical PS. Additional Fig 8: same as Additional Fig. 7with smaller x-scale.

Additional Figure 6



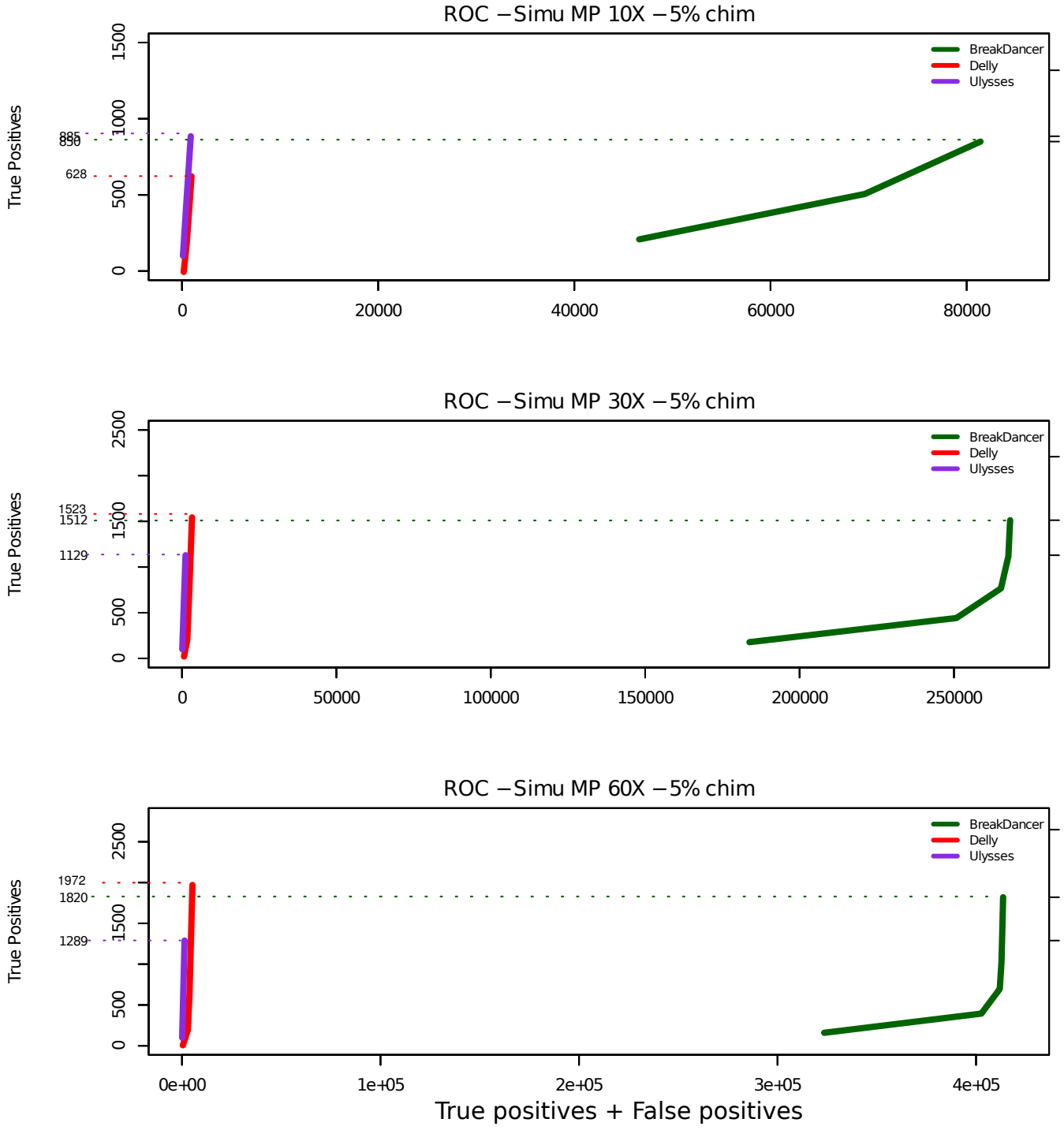
Additional Figure 5 to 8 present ROC curves showing the number of true positive SV detected as a function of the total number of SV predictions (true positives + false positives). The SV relativecoverage is used as a varying threshold. When the values of the 3 tools do not fit in the same plot area (for additional Fig. 6 and 8), another graph with a more adapted scale is provided(additional Fig. 5 and 7). Additional Fig 4: MP data with 1% of chimerical PS. Additional Fig 5: MPdata with 2.5% of chimerical PS. Additional Fig 6: same as Additional Fig. 5 with smaller x-scale.Additional Fig 7: MP data with 5% of chimerical PS. Additional Fig 8: same as Additional Fig. 7with smaller x-scale.

Additional Figure 7



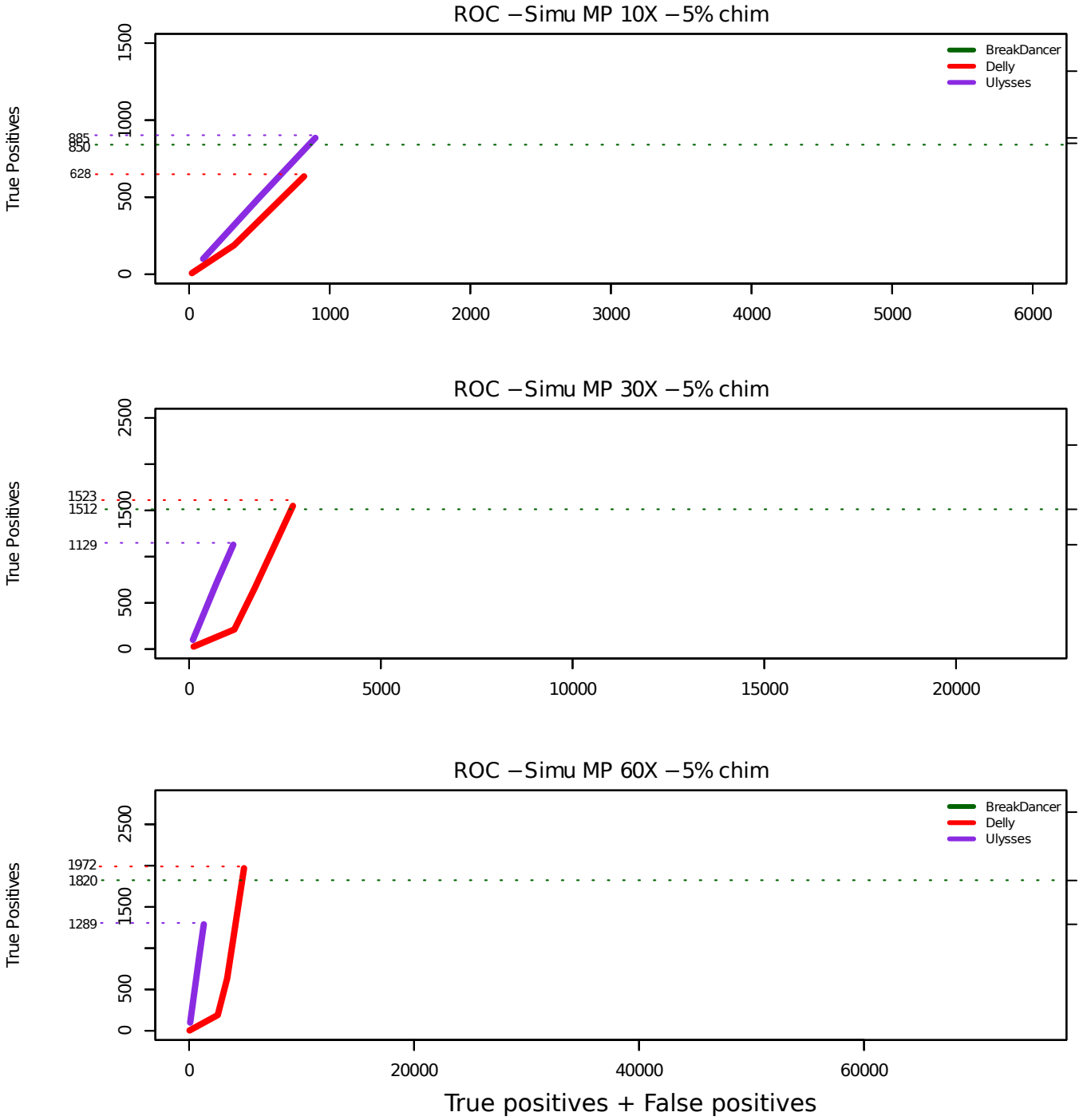
Additional Figure 5 to 8 present ROC curves showing the number of true positive SV detected as a function of the total number of SV predictions (true positives + false positives). The SV relativecoverage is used as a varying threshold. When the values of the 3 tools do not fit in the same plot area (for additional Fig. 6 and 8), another graph with a more adapted scale is provided(additional Fig. 5 and 7). Additional Fig 4: MP data with 1% of chimerical PS. Additional Fig 5: MPdata with 2.5% of chimerical PS. Additional Fig 6: same as Additional Fig. 5 with smaller x-scale.Additional Fig 7: MP data with 5% of chimerical PS. Additional Fig 8: same as Additional Fig. 7with smaller x-scale.

Additional Figure 8



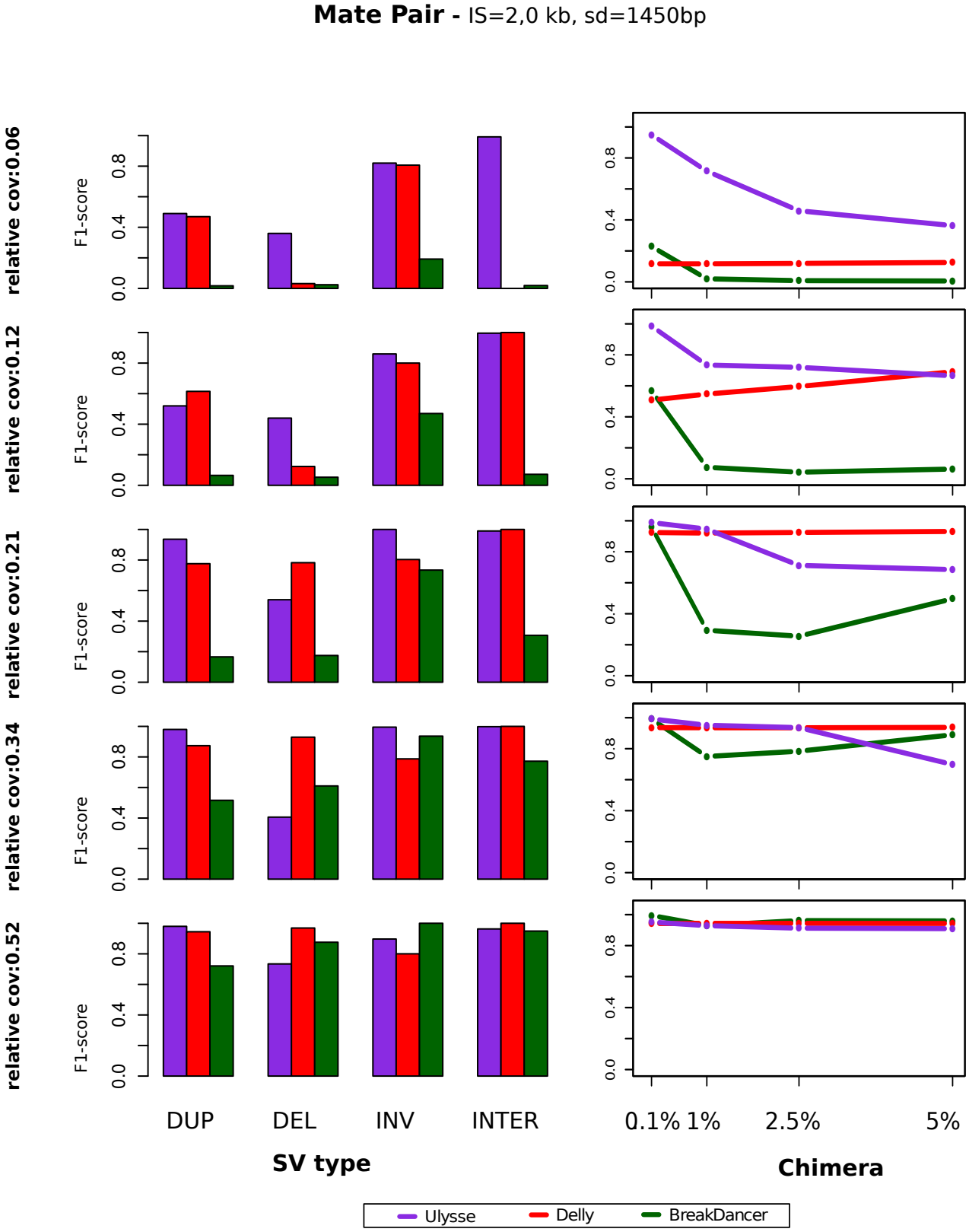
Additional Figure 5 to 8 present ROC curves showing the number of true positive SV detected as a function of the total number of SV predictions (true positives + false positives). The SV relativecoverage is used as a varying threshold. When the values of the 3 tools do not fit in the same plot area (for additional Fig. 6 and 8), another graph with a more adapted scale is provided(additional Fig. 5 and 7). Additional Fig 4: MP data with 1% of chimerical PS. Additional Fig 5: MPdata with 2.5% of chimerical PS. Additional Fig 6: same as Additional Fig. 5 with smaller x-scale.Additional Fig 7: MP data with 5% of chimerical PS. Additional Fig 8: same as Additional Fig. 7with smaller x-scale.

Additional Figure 9



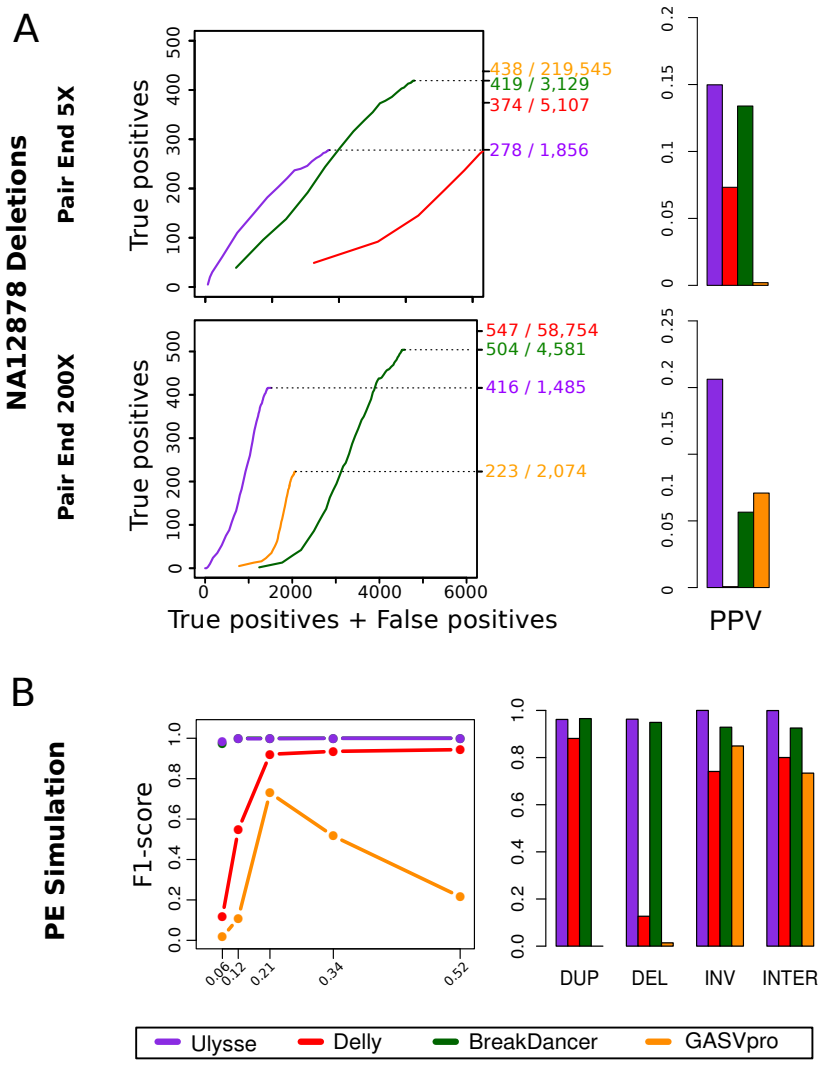
Additional Figure 5 to 8 present ROC curves showing the number of true positive SV detected as a function of the total number of SV predictions (true positives + false positives). The SV relativecoverage is used as a varying threshold. When the values of the 3 tools do not fit in the same plot area (for additional Fig. 6 and 8), another graph with a more adapted scale is provided(additional Fig. 5 and 7). Additional Fig 4: MP data with 1% of chimerical PS. Additional Fig 5: MPdata with 2.5% of chimerical PS. Additional Fig 6: same as Additional Fig. 5 with smaller x-scale.Additional Fig 7: MP data with 5% of chimerical PS. Additional Fig 8: same as Additional Fig. 7with smaller x-scale.

Additional Figure 10



Additional Figure 10 shows the detailed MP simulation results for various SV relative coverages.Refer to the legend of Figure 2 for further details.

Additional Figure 11



Additional Figure 11 (A) left panels: ROC-like curves representing the number of true positiveDEL (see Gold Standard in Methods section) as a function of the total number of predictions(true positives + false positives) using the relative SV coverage as a varying threshold for NA12878, PE5X and PE200X datasets. Right panels: histogram of the PPV of the differenttools (B) Results on a 60X PE dataset with μ=233pb, σ=10pb and 0.1% of chimerical PS. Leftpanel: F1-score as a function of SV relative coverage. Right panel: F1-score for the differenttypes of SV with a relative coverage of 0.06.

LEGENDS OF ADDITIONAL TABLES

Additional Table 1 provides sensitivity, PPV and F1-score for NA12878 PE5X, PE200X and MP30X datasets (see Methods) for the 4 detection tools.

Additional Table 2 provides sensitivity, PPV, F1-score for the MP and PE simulated datasets for each artificial SV relative coverage class for the 4 detection tools. Execution time for each dataset is also provided in the last column.

Additional Table 3 provides PPV, sensitivity, F1-score for the MP and PE simulated datasets for each type of SV for the 4 detection tools.

Additional table 1

		True Positives	Deletion Predictions	PPV
PE5X	Ulysse	278	1856	0.150
PE5X	Delly	374	5107	0.073
PE5X	Breakdancer	419	3129	0.134
PE5X	GASVpro	438	219545	0.002
PE200X	Ulysse	416	1485	0.280
PE200X	Delly	547	58754	0.009
PE200X	Breakdancer	504	4581	0.110
PE200X	GASVpro	223	2074	0.108
MP30X	Ulysse	175	367	0.476
MP30X	Delly	48	618	0.078
MP30X	Breakdancer	197	19919	0.009

Red PPV indicate the tool(s) with the best performance

Additional table 2

data type	coverage	relative coverage	nbPS	PPV				sensitivity				F1-score				Execution times (s)			
				ulysse	Delly	GASVpro	BreakDancer	ulysse	Delly	GASVpro	BreakDancer	ulysse	Delly	GASVpro	BreakDancer	Ulysse	Delly	BreakDancer	GASVpro
MP	10X	0.286	4	1	0.497	NA	0.307	0.7	0.402	0.836	0.74	0.824	0.444	NA	0.434	204	1574	136	-
MP	10X	0.444	8	0.971	0.896	NA	0.924	0.904	0.991	0.716	0.807	0.936	0.941	NA	0.862				
MP	30X	0.118	4	1	0.114	NA	0.023	0.551	0.407	0.873	0.731	0.711	0.178	NA	0.045				
MP	30X	0.211	8	1	0.736	NA	0.074	0.738	0.984	0.758	0.798	0.849	0.842	NA	0.135	990	3746	411	-
MP	30X	0.348	16	0.96	0.892	NA	0.266	0.911	0.987	0.584	0.82	0.935	0.937	NA	0.402				
MP	30X	0.516	32	0.897	0.893	NA	0.558	0.913	0.998	0.302	0.873	0.905	0.943	NA	0.681				
MP	60X	0.063	4	1	0.068	NA	0.01	0.556	0.409	0.873	0.649	0.715	0.117	NA	0.02				
MP	60X	0.118	8	1	0.378	NA	0.039	0.58	0.989	0.771	0.749	0.734	0.547	NA	0.074				
MP	60X	0.211	16	1	0.865	NA	0.179	0.898	0.984	0.627	0.813	0.946	0.921	NA	0.293	2088	6890	826	-
MP	60X	0.348	32	0.995	0.891	NA	0.637	0.911	0.984	0.36	0.911	0.951	0.935	NA	0.75				
MP	60X	0.516	64	0.916	0.899	NA	0.937	0.942	0.993	0.124	0.922	0.929	0.944	NA	0.929				
PE	10X	0.286	4	1	0.987	0.136	1	1	0.996	0.836	0.964	1	0.991	0.234	0.982	339	1432	126	872
PE	10X	0.444	8	1	0.998	0.894	1	1	0.976	0.716	1	1	0.987	0.795	1				
PE	30X	0.118	4	0.996	0.842	0.009	1	1	0.996	0.873	0.964	0.998	0.913	0.018	0.982				
PE	30X	0.211	8	1	0.971	0.111	1	1	0.98	0.758	1	1	0.975	0.194	1	885	3388	381	2058
PE	30X	0.348	16	1	0.996	0.92	1	0.998	0.984	0.584	1	0.999	0.99	0.714	1				
PE	30X	0.516	32	0.996	0.998	0.919	1	1	0.976	0.302	1	0.998	0.987	0.455	1				
PE	60X	0.063	4	0.966	0.573	0.008	0.998	1	0.996	0.873	0.951	0.983	0.727	0.016	0.974				
PE	60X	0.118	8	0.996	0.898	0.059	1	1	0.978	0.771	1	0.998	0.936	0.11	1				
PE	60X	0.211	16	0.998	0.961	0.873	1	0.998	0.984	0.627	1	0.998	0.972	0.73	1	1785	5694	771	3293
PE	60X	0.348	32	1	0.993	0.91	1	1	0.98	0.36	1	1	0.986	0.516	1				
PE	60X	0.516	64	1	0.998	0.836	1	1	0.987	0.124	1	1	0.992	0.216	1				

Red F1-scores indicate the tool(s) with the best performance

Additional table 3

data type	SV type	coverage	PPV				sensitivity				F1-score			
			ulysse	Delly	GASVpro	BreakDancer	ulysse	Delly	GASVpro	BreakDancer	ulysse	Delly	GASVpro	BreakDancer
MP	DUP	10X	1	1	0.011	0.063	0.307	0.6	0.753	0.14	0.47	0.75	NA	0.087
	DEL	10X	1	0.392	0.011	0.086	0.08	0.58	0.753	0.507	0.148	0.468	0.022	0.147
	INV	10X	1	0.667	1	1	0.553	0.667	0.79	0.643	0.712	0.667	0.883	0.783
	INTER	10X	0.98	1	0.863	0.117	0.797	0.333	0.815	0.903	0.879	0.5	NA	0.207
	DUP	30X	0.673	1	0.002	0.03	0.396	0.756	0.5	0.596	0.499	0.861	NA	0.057
	DEL	30X	1	0.109	0.002	0.024	0.168	0.732	0.5	0.54	0.288	0.19	NA	0.046
	INV	30X	0.996	0.668	1	0.864	0.544	0.796	0.828	0.532	0.704	0.726	NA	0.659
	INTER	30X	0.986	1	0.867	0.029	0.87	0.601	0.701	0.918	0.924	0.751	NA	0.056
	DUP	60X	0.993	0.996	0.002	0.016	0.473	0.79	0.413	0.577	0.641	0.881	NA	0.031
	DEL	60X	0.986	0.069	0.002	0.015	0.23	0.783	0.413	0.607	0.373	0.127	NA	0.029
	INV	60X	0.93	0.668	1	0.757	0.508	0.832	0.738	0.515	0.657	0.741	NA	0.613
	INTER	60X	0.988	1	0.866	0.016	0.89	0.667	0.637	0.907	0.936	0.8	NA	0.031
PE	DUP	10X	1	1	0.011	1	1	0.6	0.753	0.867	1	0.75	0.022	0.929
	DEL	10X	1	0.392	0.011	1	0.96	0.58	0.753	0.86	0.98	0.468	0.022	0.925
	INV	10X	1	0.667	1	1	1	0.667	0.79	0.737	1	0.667	0.883	0.849
	INTER	10X	1	1	0.863	1	1	0.333	0.815	0.908	1	0.5	0.838	0.952
	DUP	30X	0.988	1	0.002	1	0.996	0.756	0.5	0.92	0.992	0.861	0.004	0.958
	DEL	30X	0.961	0.109	0.002	0.991	0.996	0.732	0.5	0.916	0.978	0.19	0.004	0.952
	INV	30X	1	0.668	1	1	1	0.796	0.828	0.844	1	0.726	0.906	0.915
	INTER	30X	0.998	1	0.867	1	0.999	0.601	0.701	0.898	0.998	0.751	0.775	0.946
	DUP	60X	0.929	0.996	0.002	1	0.997	0.79	0.413	0.933	0.962	0.881	0.004	0.965
	DEL	60X	0.931	0.069	0.002	0.966	0.997	0.783	0.413	0.933	0.963	0.127	0.004	0.949
	INV	60X	1	0.668	1	1	1	0.832	0.738	0.867	1	0.741	0.849	0.929
	INTER	60X	1	1	0.866	1	0.999	0.667	0.637	0.861	0.999	0.8	0.734	0.925

Red F1-scores indicate the tool(s) with the best performance

2 - Mesure quantitative de la plasticité des chromosomes chez *S. cerevisiae*

Nous avons émis l'hypothèse que la plasticité des génomes pourrait être sous-estimée et qu'il devrait être possible de détecter les SV quand elles se forment dans les populations clonales de cellules. Pour tester cette hypothèse, la levure est un modèle de choix en raison de son petit génome de 12 Mb qui permet d'accéder à des couvertures de séquençage très élevées. De plus, nous venons de voir que grâce au logiciel Ulysses nous pouvons maintenant tirer parti de la forte couverture physique de librairies MP et ainsi détecter des SV en faibles fréquences dans un échantillon. Ces résultats nous ont permis de mettre en place une nouvelle approche pour quantifier le nombre, les types et la localisation de SV qui se produisent *de novo* dans des populations monoclonales de levures.

Jusqu'à présent, il n'existait que 2 grandes approches expérimentales permettant la détection de la formation de SV chez les microorganismes : la sélection de réarrangements chromosomiques avec des systèmes génétiques dédiés comme le système GCR (cf. page 75) et les MAL. Comme nous l'avons vu dans l'introduction, les systèmes de sélection des SV comme le système CYC1, le système GCR ou la mesure des CNV dans le rDNA montrent que les taux de délétions sont très hétérogènes dans le génome et vont de 10^{-2} DEL/cellule/génération à $<10^{-9}$ DEL/cellule/génération. De plus, ces approches ne sont pas nécessairement représentatives des différents types et fréquences de SV sur l'ensemble du génome puisqu'elles ne mesurent la formation que d'un seul type de SV à un locus unique. L'approche MAL permet quant à elle de fixer par dérive génétique des mutations dans une lignée de cellules. Associée au séquençage NGS, les MAL fournissent un moyen précis d'étudier des mutations ponctuelles. Cependant, leur puissance est limitée pour l'étude des SV à cause de la rareté de ces évènements et de la difficulté de les détecter dans les données de séquençage. Chez les cellules diploïdes, la possibilité qu'ont les cellules à réparer les CDB par recombinaison avec le chromosome homologue stabilise le génome et ne permet pas d'utiliser cette approche pour détecter des SV (Nishant et al., 2010; Zhu et al., 2014). Même chez les haploïdes, les MAL ne permettent de détecter au mieux qu'un petite nombre de SV par expérience (Lynch et al., 2008; Serero et al., 2014). Dans ces conditions de nombre de SV réduits, Lynch et collègues ont toutefois pu estimer les taux de DEL à $1{,}2.10^{-2}$ /cellule/génération et le taux d'INS à 2.10^{-2} /cellule/génération à partir d'une moyenne de 2.25 évènements de délétions par MAL et d'en moyenne 2.6 évènements de duplications par MAL.

Compte tenu des taux de SV mesurés avec les systèmes génétiques et les MAL, il est théoriquement possible de pouvoir détecter des SV subclonales qui se produisent *de novo* à condition d'être en mesure de détecter des sous clones présents à des fréquences inférieures à ~1/1 000 dans la population de cellules. C'est ce que propose notre approche de détection des SV subclonales (SSV pour « Subclonal Structural Variation »). Alors que les MAL nécessitent

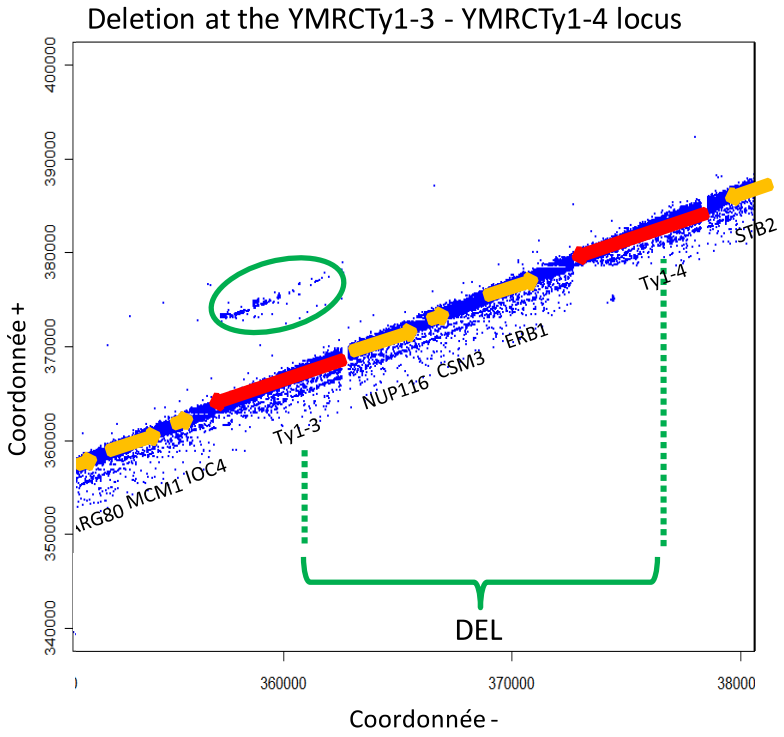


Figure 27: Illustration de l'une des premières délétions identifiée avec l'approche SSV. Le graphique représente les coordonnées de fin en fonction des coordonnées de départ de chaque RP sur une portion du chromosome XIII dans l'expérience SSV *clb5Δ*. La majorité des RP qui se trouvent sur la ligne transversale bleue correspondent aux RP concordantes avec le génome de référence. Sur ce graphique, une augmentation de la taille d'insert des RP (à cause d'une délétion qui permet de cartographier les RP à une distance plus grande qu'attendue sur le génome de référence) se visualise par un éloignement des RP vers de haut de la diagonale. Le groupe de lectures entouré en vert correspond à des RP dont la taille d'insert plus grande que la normale indique la présence d'une délétion subclonale entre Ty1.3 et Ty1.4.

de cultiver des cellules pendant des milliers de générations, notre approche consiste à n'effectuer qu'une seule culture monoclonale de ~30-35 générations suivie d'un séquençage MP avec une couverture physique de plusieurs milliers de X (Article 2, figure 1, page 131).

Ainsi, dans un premier temps, nous avons réalisé une expérience de SSV en séquençant une population de cellules sauvages sur 4 lignes de GAIIX ainsi que 4 expériences de SSV dans un fond génétique mutant pour le gène *CLB5* (cycline impliquée dans le déclenchement des origines de réplication tardives et connue pour être un gène mutateur). Ces expériences ont abouti à des couvertures physiques de l'ordre de ~5000X qui n'ont pas permis de détecter

beaucoup de SV dans le 'core genome' (c.a.d. en dehors des subtélomères) (Article 2, Tableau 1, page 131). Dans les cellules sauvages par exemple, seuls 2 DEL, 1 INS et 1 NRT ont été détectées. La figure 27 illustre l'une des délétions détectées sur le chromosome 13 dans une des expériences dans des cellules *clb5Δ*. Cet évènement s'explique par la recombinaison entre 2 Ty en orientation directe qui provoque la délétion d'un fragment d'environ 12kb et contenant 3 gènes. Cependant, 2 de ces 3 gènes sont annotés comme des gènes essentiels (*NUP116* et *ERB1*) ce qui suggère que ce fragment n'a pas été complètement délété mais qu'il a été transloqué ailleurs ou que cette délétion est associée à une aneuploïdie. Une autre explication a également été proposée par les groupes de David Bostein et Birgitte Regenberg (Møller et al., 2015). Dans cet article, les auteurs montrent grâce à une méthode d'isolement et de séquençage des ADN circulaires que la formation d'ADN circulaire extrachromomique (eccDNA) pourrait être fréquente chez la levure grâce à la recombinaison d'éléments répétés dans le génome. Les auteurs ont notamment démontré que 80% de ces eccDNA contiennent une séquence d'ARS qui pourrait expliquer le maintien de ces eccDNA lorsqu'ils déletent des gènes essentiels. Or, en plus des 3 gènes cités plus haut, ce locus contient l'ARS1312. Bien que je n'aie pas réussi à valider par PCR l'existence de cette délétion, les auteurs de ce papier ont pour leur part amplifié par PCR inverse l'eccDNA correspondant dans des cellules traitées avec de la Zeocine. D'autres SSV impliquant la délétion de gènes essentiels pourraient s'expliquer par l'existence de cellules quiescentes qui ne peuvent pas se diviser. Cependant, cette hypothèse n'est envisageable que si ces évènements sont fréquents et produisent un nombre suffisant de cellules pour que certaines d'entre elles soient détectables parmi les 10^{35} cellules sauvages. L'observation de la délétion de gènes essentiels peut également s'expliquer par l'existence de mutations compensatrices, le cas le plus simple étant celui de la co-délétion de gènes. Enfin, la dernière possibilité pouvant expliquer la détection de ces délétions de gènes essentiels est que ces délétions affectent très négativement la valeur sélective des cellules sans toutefois les rendre non viables ce qui leur permettraient de continuer à se diviser lentement.

Les expériences de SSV précédentes ont montré qu'une couverture de 5 000X permet de détecter un petit nombre de SSV. Cela suggère que cette couverture correspond à la limite inférieure de détection de SSV. Afin d'augmenter la sensibilité des expériences, nous avons donc réalisé une expérience de SSV avec des cellules sauvages dont la couverture de séquençage atteint environ 37 000X. Bien que les expériences de validations expérimentales soient toujours en cours, nous avons détecté 332 SV dans cette expérience (Article 2, figure 1 et tableau 1, page 131). À notre connaissance, c'est la première fois chez la levure qu'un tel nombre de SV *de novo* est détectée. Plus précisément, nous avons détecté dans le 'core

genome' 16 DEL, 11 INV, 15 DUP, 14 RT, 1 INS et 31 NRT. Ceci nous a permis de calculer pour la première fois des taux de mutation pour l'ensemble des types de SV dans le 'core génome'. Dans l'ordre croissant, nous estimons ces taux à: 1,5, 2,1, 3,1, 5,3, 6,3 et 10.10^{-3} /cell/division pour les INV, DEL, RT, INS, DUP et NRT respectivement (Article 2, figure 3, page 131). Notons que le taux d'insertion n'est calculé que sur la base de 1 SV et qu'il est donc probablement très imprécis. Précédemment, les taux de DEL et INS chez la levure avaient été estimés à partir d'un nombre restreint (entre 0 et 8 DUP par MAL et entre 0 et 12 DEL par MAL) de SV et atteignaient 1,2 et 2 10^{-2} /cellule/division respectivement. Nos estimations actuelles sont donc 6 et 4 fois plus faibles. Nous avons cependant vu dans la partie 1 des résultats que notre algorithme Ulysses filtre une proportion importante des vrais positifs (86% dans le cas des délétions dans les données MP30X NA12878). Ceci a notamment pu être vérifié par 2 expériences de SSV préliminaires dans lesquelles toutes les DUP (avant le filtrage par le module statistique) ont été testeés expérimentalement. Ces validations ont montré que le taux de faux négatif d'Ulysses est d'au moins 10% pour la détection des DUP. Ce pourcentage dépend également directement des seuils de détections choisis par Ulysses en fonction du niveau de bruit (la proportion de chimères) dans la librairie de séquençage. Par ailleurs, nous avons constaté que ce bruit est variable d'une expérience à l'autre et de ce fait, le taux de faux négatifs ne peut être extrapolé d'une expérience de SSV à l'autre. Les taux de SV estimés par l'approche SSV sont donc probablement conservatifs et sont donc probablement sous-estimés.

Pour la partie subtélomériques des chromosomes, nous avons calculé des taux de DEL, INV et DUP 2 à 3 fois inférieurs à ceux du 'core genome' (taux non normalisés par la taille des subtélomères). De façon intéressante, nous avons également détecté un grand nombre de translocations non réciproques (217 NRT) dont le taux de formation calculé est de 7.10^{-2} NRT/cellule/génération. Les recombinaisons interchromosomiques causées par les régions Y' à la toute fin des chromosomes sont connus depuis longtemps (Louis and Haber, 1990, 1992; Nishant et al., 2010). Cependant, les NRT détectées dans notre expérience s'étendent dans l'ensemble des régions subtélomériques qui partagent également de grands blocs d'homologies.

Dans notre expérience, environ 70% des SSV détectées présentent de grandes homologies (Ty, Long Terminal Repeats, gènes répètes, ou blocs subtélomériques dupliqués) à proximité immédiate de la position prédite des jonctions. Ceci est illustré par les *dotplots* des jonctions des SV (figure 28 et figures supplémentaires article 2). Ces données suggèrent que la majorité des SSV dans le génome se forme par des mécanismes nécessitant la recombinaison

homologue. Les autres mécanismes utilisant de la microhomologie comme le NHEJ ou le MMEJ ou les mécanismes réplicatifs de type MMIR sont en minorité mais représentent toute de même 30% des SSV. De plus, l'utilisation d'homologie pour former des SV suggère que ces dernieres sont récurrents dans les génomes contrairement aux SV formées à l'aide de microhomologie. Cette hypothèse a notamment été illustrée par la découverte d'un *hotspot* de duplication segmentale sur le chromosome IV. Quatre duplications ont en effet été trouvées dans 3 expériences de SSV indépendantes entre les *loci* des gènes dupliqués PPH21/22 et RPL35A/B. De manière intéressante, ce locus contient également le gène VMA1 qui code entre autres pour l'endonucléase PI-SceI. Cette dernière a pour rôle de couper durant la méiose les allèles VMA1 qui ne contiennent pas l'endonucléase et ainsi d'initier le processus 'homing' qui consiste à intégrer la séquence de l'endonucléase dans les allèles qui ne la contiennent pas déjà (Gimble and Thorner, 1992). L'implication potentielle de cette endonucléase en mitose dans ce *hotspot* n'a cependant pas été démontrée expérimentalement mais semble être une piste intéressante à suivre.

Au cours de ce travail, nous avons d'autres expériences de SSV dans un fond génétique mutant pour la protéine Rad27p. Cette dernière est une endonucléase impliquée dans la maturation des fragments d'Okazaki. Et bien que Rad27 soit connu comme étant l'un des gènes mutateurs les plus forts, l'augmentation du taux de SV que sa mutation provoque varie en fonction du système expérimental. Dans des expériences de GCR, le taux de SV augmente presque d'un facteur 1 000 fois chez ce mutant par rapport aux cellules sauvages (Chen and Kolodner, 1999). En revanche, des expériences dans des MAL n'ont montré qu'un accroissement d'un facteur 10 (Serero et al., 2014). Dans notre expérience SSV, nous n'avons mesuré qu'une augmentation d'un facteur 2 ou 3 de tous les types de SV par rapport aux cellules sauvages ce qui se rapproche plus des mesures faites à partir des MAL que des expériences de GCR. Une explication à cette différence entre les systèmes GCR et les MAL pourrait provenir du fait que les différents types de SV ne sont pas affectés de la même manière dans ce mutant. Pour tester cette hypothèse, j'ai construit des systèmes génétiques permettant de mesurer la formation d'INV et de RT et également utilisé un système de mesure d'une DUP construit précédemment au laboratoire (Payen et al., 2008) (les systèmes génétiques seront illustrés plus loin dans l'article correspondant à ces travaux). Mes mesures de ces taux ont indiqué que le mutant *rad27Δ* conduit à une augmentation du taux de DUP que d'un facteur 2 par rapport aux cellules sauvages alors que le taux d'INV augmente d'un facteur 26 et celui des translocations réciproques d'un facteur 160. Ces résultats illustrent la difficulté qu'il y a à comparer des taux de SV qui, contrairement aux mutations ponctuelles qui

n'affectent qu'un seul nucléotide, affectent eux un nombre de bases variable. Ceci a pour conséquence de rendre la fréquence des SV dépendante du contexte génomique et elles ne sont donc pas nécessairement comparables entre elles.

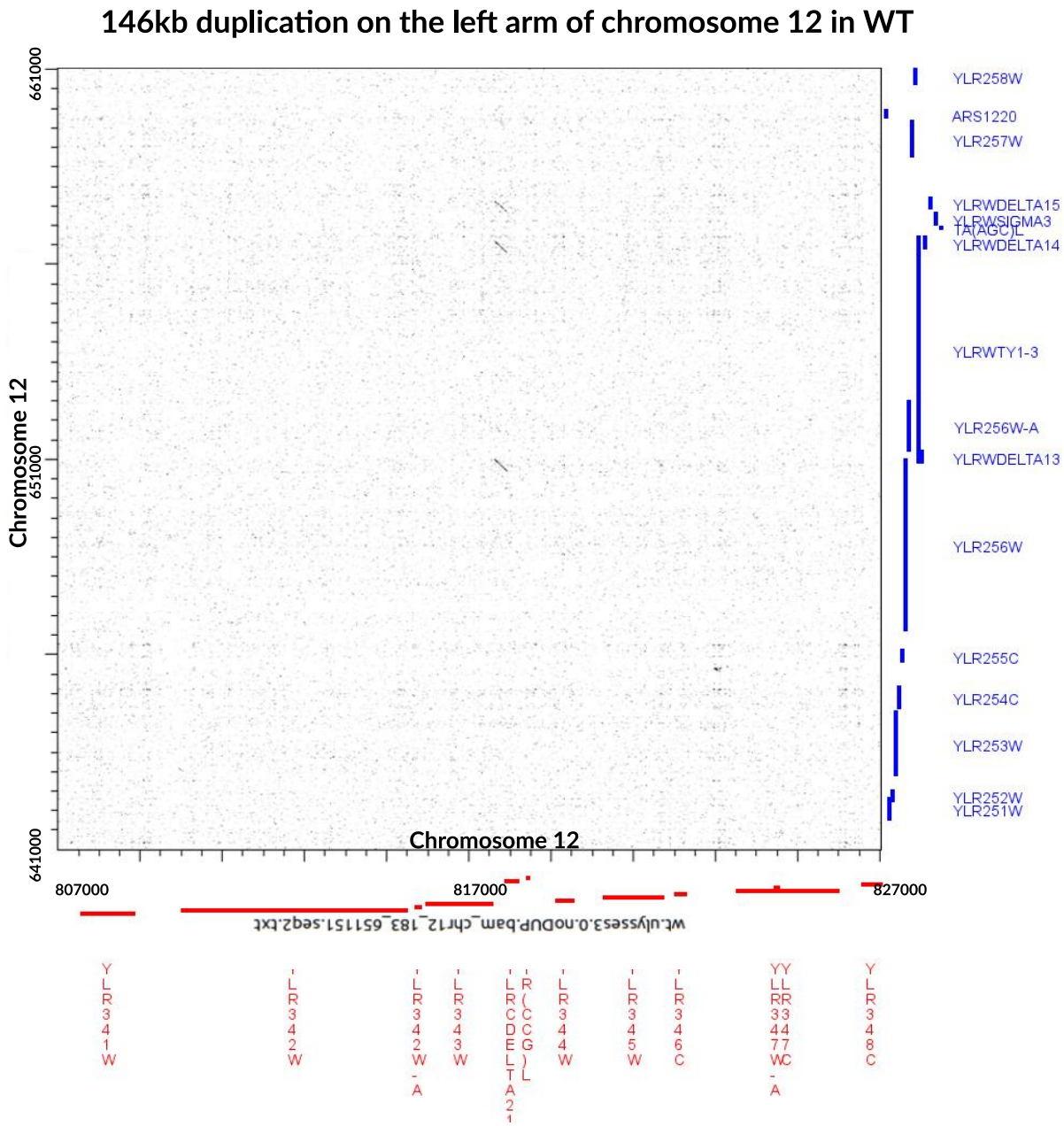


Figure 28. Dotplot de la jonction d'une duplication de 146 kb détectée sur le chromosome 12 dans l'expérience SSV WT 37000X. Le dotplot est centrés sur la position des jonctions prédites. Cette DUP est probablement formé par un mécanisme utilisant la RH comme le BIR entre le LTR YLRCDelta2-1 et un des LRT du transposon YLRWTy1-3.

Dans le contexte *rad27Δ*, nous avons d'ailleurs observé un fort enrichissement des délétions d'éléments transposables Ty qui représentent près de 50% de toutes les délétions observées (6 sur 13). Rad27p est un inhibiteur connu de la mobilité des éléments Ty chez la levure (Sundararajan et al., 2003). Cependant, cette mobilité des transposons aurait dû se caractériser par une augmentation du taux d'insertion et pas de délétion. Ces résultats suggèrent donc que Rad27p est peut être impliquée dans la stabilité des Ty dans le génome.

Pour finir, ce travail démontre pour la première fois qu'il est possible de détecter des SV subclonales dans des cultures de levures. Nous avons pu quantifier le taux et les types de SSV qui se produisent dans le génome de la levure. Nos estimations des taux des SSV se produisant naturellement dans des cellules sauvages est de l'ordre de 10^{-3} SSV /cellule/division ce qui implique que des populations réellement clonales de plus de 1 000 cellules n'existeraient pas, non seulement en termes de SNP mais également en termes de SSV. De plus, compte tenu du taux de mutation ponctuel chez la levure ($3{,}3.10^{-10}$ mutation/cellule/génération) et de la taille du génome (12,5Mb), il se produit donc en moyenne $3{,}3.10^{-10} \times 12{,}5\times10^{6} = 3{,}9.10^{-3}$ mutation/cellule/division. Les mutations ponctuelles et les SV se produisent donc à des fréquences équivalentes dans le génome. Nous avons également pu déterminer précisément les types de SV qui se produisent dans les régions subtélomériques et qui font donc varier le nombre de copie des gènes subtélomériques. Ce système présente de plus l'avantage d'être beaucoup plus rapide à effectuer que des expériences de MA. Il peut donc être facilement (au coût élevé du séquençage MP près) réalisé dans différents fonds génétiques, souches ou espèces. Le manuscrit qui suit est une version préliminaire de l'article qui présentera ce projet et qui sera soumis pour publication. La validation expérimentale des expériences SSV dans les cellules sauvages et *rad27Δ* séquencées à très haut débit (37 000 et 19 000X, respectivement) sont toujours en cours et le manuscrit ci-après ne représente donc que les résultats obtenus jusqu'à présent.

Nous avons durant ce travail étudié les SSV chez *S. cerevisiae*. Il serait toutefois intéressant de mettre en perspectives ce type de données sur la plasticité des chromosomes dans le contexte de l'évolution. A cet effet, l'approche SSV pourrait être réalisée dans plusieurs espèces dont on connaît l'histoire évolutive des génomes afin de pouvoir comparer la plasticité des chromosomes aux réarrangements chromosomiques réellement fixes au cours de l'évolution. L'exploration génomique du clade des levures *Lachancea* réalisée au laboratoire offre pour cela une excellente opportunité : au cours de ce projet, l'histoire des réarrangements chromosomiques de tous les ancêtres de 10 espèces du genre *Lachancea* a été reconstruite. Nous disposons donc d'une carte et d'une quantification précise de l'ensemble

des réarrangements fixés dans toutes les branches de l'arbre du clade. Ce travail a par exemple permis de montrer que le nombre de duplications qui se sont fixées dans ces génomes est plus élevé que le nombre de délétions (1,503 vs 1,036). Ce travail de description des génomes ancestraux du clade a fait l'objet de la soumission d'un article actuellement en revue dans le journal eLife et dont je suis cosignataire (« Reconstructing the complete genome history of the Lachancea genus uncovers a genomic clock in yeast »). Puisque l'on connaît exactement comment les génomes des espèces de clades ont évoluées il serait donc intéressant d'étudier les SSV dans chacune de ces espèces pour comparer la plasticité actuelle des chromosomes à l'histoire évolutive du clade.

Les résultats des travaux de détection des SSV chez *S. cerevisiae* sauvage ou mutante sont présentés ci-dessous sous forme d'article mais ce dernier ne représente pas la version définitive du manuscrit puisque la validation expérimentale des SSV est toujours en cours.

Article 2

A granular view of subclonal SV landscape in the yeast genome

Gillet-Markowska Alexandre & Fischer Gilles

En préparation

Les données supplémentaires de cet article sont disponibles à l'adresse :
www.lcqb.upmc.fr/ulysses/supp/supp_data.tar.gz

A granular view of subclonal SV landscape in the yeast genome

Gillet-Markowska Alexandre & Fischer Gilles

July 12, 2015

1 Introduction

Large Structural Variations (SV) of chromosomes including duplications (DUP), deletions (DEL), inversions (INV), insertions (INS), reciprocal translocations (RT) and non-reciprocal translocations (NRT) are common contributors to genomic disorders and cancers (Fanciulli *et al.*, 2007; Hollox *et al.*, 2008; de Cid *et al.*, 2009; Stephens *et al.*, 2009; Campbell *et al.*, 2010; Pinto *et al.*, 2010; Girirajan *et al.*, 2011). It is well established that SV also contribute significantly to the genetic diversity between individuals (Freeman *et al.*, 2006; Conrad and Hurles, 2007; Korbel *et al.*, 2007; Hoffmann and Rieseberg, 2008; Kidd *et al.*, 2008; Mills *et al.*, 2011; Wang *et al.*, 2012). In addition, new approaches to single-cell genome dynamics have revealed the quantitative importance of somatic SV and provided strong evidence that genomes would be much more plastic than previously thought (see Wang and Navin, 2015, for review), although the quantification of such plasticity is still controversial (Knouse *et al.*, 2014). Despite their recognized contributions to genome polymorphism, all the detected SV are likely to constitute only the 'tip of the iceberg' with many other neutral, detrimental or even beneficial SV that remain undetected because they never rise above low frequencies in cell populations. One possibility to characterize the intrinsic plasticity of genomes is to experimentally characterize *de novo* SV formation in the genome of model microorganisms.

To this end, yeast are preferential tools given their small genome size and short generation time. Currently, two experimental approaches allow to identify *de novo* SV formation in yeast: assays systems and Mutation Accumulation (MA) lines. Assay systems provide a powerful mean to identify SV but are limited to the study of one locus at a time, not giving a representative picture of the mutational landscape of the genome. In addition, most assay systems only allow to measure deletion rates. For example, *HIS4* and *CYC1* systems allowed the measurement of the first deletion rates in yeast (Fink and Styles, 1974; Sherman *et al.*, 1974).

Although deletions represented less than 1% of all identified mutations at these loci, the authors managed to measure deletion rateto be about 10^{-8} DEL/cell/division. A higher deletion rate ($\sim 10^{-5}$/cell/division) due to flanking retrotransposons at the *CYC1* locus was also observed later in specific strains (Liebman *et al.*, 1979). Exinger and Lacroute calculated the deletion at the *ura2 15-30-72* allele by selecting prototrophic revertants occurs at a rate of $\sim 10^{-9}$DEL/cell/division. Other systems like the Gross Chromosomal Rearrangements assay (Chen and Kolodner, 1999) provide a powerful mean to identify *de novo* terminal deletions and telomere replacement but are limited to the study of subtelomeric regions. In wild-type cells, GCR occur at rates of $\sim 10^{-10}$GCR/cell/division. There is a 5 order of magnitude discrepancy between the most extreme estimates, clearly showing that assay systems can be highly influenced by the presence of repetitive sequences. This is further illustrated by Szostak and Wu that proposed another approach to measure the CNV (deletion or duplication) due to unequal cross-over in rDNA arrays and derived a rate of 10^{-2} CNV/cell/division. Although assay to measure duplications have been less investigated than deletions, several approach to study duplications have been developed like the CUP1 amplification (Welch *et al.*, 1990) or the ADH2 (or ADH4) duplication selection on medium containing antimycin A (Dorsey *et al.*, 1992). More recently, Payen *et al.* designed a genetic screen that allows to measure duplication rates on the right arm of chromosome XV. Their approach showed DUP rates of $\sim 10^{-7}$ DUP/cell/division. These assay systems are not suited to infer the SV landscape at the genome-scale due to many differences are observed between the systems. Furthermore, balanced SV rates including inversions and reciprocal translocations have not been tackled by assay systems.

Alternatively, MA lines coupled to next generation genome re-sequencing yield a genome-wide insight to genomic plasticity. They do not require mutant selection, and, as such, are less biased than assay systems. Although MA lines successfully provided precise estimates of point mutations rates in yeast, they failed to identify significant numbers of SV. Lynch *et al.* identified 11 DUP and 4 DEL in four MAL after ~ 4800 generations. From these number, authors estimated the deletion rate in WT cells to 1.2×10^{-2}DEL/cell/division and the duplication rate to 2×10^{-2}DUP/cell/division but could not estimate the mutation rates of balanced SV because they did not identify RT or INV. Another study failed to identify any SV in WT MA lines after ~ 2500 generations (Serero *et al.*, 2014) showing that a decrease in the number of generation greatly cripples MA lines sensitivity. MA lines also showed the diploid yeast genome is very stable in mitosis owing to their possibility to repair DSB with the homologous chromosome: out of 145 MA lines that represent a total of $\sim 311,000$ generations, Zhu *et al.* only identified 3 CNV (> 30kb) as well as 31 aneuploidies. Similarly, Nishant *et al.* performed ~ 1700 generations MA lines in diploid cells and only identified a single retrotransposition event inside the rDNA locus as well as 15 Y' recombination. Altogether, MA lines yielded very few CNV in the core genome and no balanced SV, questioning their

relevancy for the study of *do novo* SV in yeast genome.

Here we propose a new approach to identify SV in cell populations which consists in growing a single un-selected monoclonal yeast culture and sequencing the genomic DNA of the entire culture at very high physical coverage. The detection of subclonal SV (SSV) is then performed by a dedicated algorithm, Ulysses, which achieves a highly specific SSV identification by evaluating the statistical significance of each SV against an explicit model for the generation of noise (Gillet-Markowska *et al.*, 2015).

The SSV assay applied to a single of wild-type haploid yeast cells yielded 332 SSV. We inferred mutation rates for each type of SV both in the core genome and in subtelomeric regions. We also investigated two mutator genetic backgrounds (*clb5*Δ and *rad27*Δ). Rad27p is the yeast homologue of Fen1p which is a flap endonuclease involved in DNA replication and repair (Tishkoff *et al.*, 1997). It is believed to be one of the top mutator genes in yeast with a 920 fold increase of Gross Chromosomal Rearrangement rate over WT (Chen and Kolodner, 1999). Clb5p, which is the human homologue of the tumor suppressor gene Ccnb1p, is a B-type cyclin involved in the firing of late replication origins (Donaldson *et al.*, 1998; Epstein and Cross, 1992; Schwob and Nasmyth, 1993). This gene shows a 6 fold increase of GCR rate over WT (Putnam *et al.*, 2012) and has also been reported to have a 700 fold increase of segmental duplication mutation rate (Payen *et al.*, 2008). Interestingly, SV rates in these mutants calculated with the SSV assay are only 2 to 3 times higher than in WT suggesting that locus specific assay systems might not be representative of genome-wide SV rates. Compared to MA lines, the SSV assay greatly improves the rapidity of the experiments and can therefore easily be applied to different cell populations, species or in various environmental conditions. The SSV assay is the only method capable of detecting the full spectrum of SV at the genome-scale in *S. cerevisiae* providing the first granular view of the genomic plasticity in yeast.

2 Methods

2.1 Yeast strains

All strains are derivatives of *S. cerevisiae* BY4743 (*MATa/a, his3*Δ*1/his3*Δ*1, leu2D/leu2D, met15*Δ*/MET15, lys2*Δ*/LYS2, ura3*Δ*/ ura3*Δ) (Winzeler *et al.*, 1999). Strain names and their corresponding genotypes and origins are summarized in Table S1. Gene replacements were obtained either through a classical PCR-based deletion strategy (Wach *et al.*, 1994) or from the EUROSCARF strain collection.

2.2 Subclonal populations and DNA extraction

Exponentially growing cells (OD$\sim$1) were plated on yeast YPD (glucose rich) medium and incubated 24h at 30°C. The totality of a colony was then diluted in 500mL of YPD and incubated under agitation until OD30. Total genomic DNA was then extracted with the Qiagen genomic TIP500 and the different DNA sequencing libraries were prepared by the Genoscope plateform, Eurofins MWG or GATC company (Mate-Pair or Eurofins MWG Long-Jumping Distance) and sequenced on GAIIX or Hi-Seq machines.

2.3 Raw sequences processing

Sequencing adapters were removed using Cutadapt (Martin, 2011). Read Pairs (RP) with both reads $\geq$18 nucleotides sequences were kept and reverse-complemented using Biopieces (Hansen, 1997). RP were then remapped to SGD reference sequence using bwa aln/sampe (0.7.12-r1039) using parameters -n 0 -l 0 -t 12 -o 0 and sorted with samtools (Li and Durbin, 2009). PCR and optical duplicates were then removed by using MarkDuplicates from picard tools 1.128 BroadInstitute.

2.4 SV detection and PCR validation

Structural variations were identified by clustering discordant paired sequences and filtering out false positives using statistical models as described in (Gillet-Markowska *et al.*, 2015) and `http://www.lcqb.upmc.fr/ulysses/`.

Further refinement of SV filtering were also applied:

(i) The reference S288c genome only contains 2 copies of the rDNA locus although yeast strains carry around a hundred fifty copies of rDNA arrays (Petes, 1979). As a consequence the number of rDNA sequences expected by Ulysses is under-estimated and the excess of chimerical RP not filtered led to a high amount of FP. All SV that include or overlap rDNA are therefore classified separately as 'rDNA involved' in Table 1. However, an inversion at the junction of 2 rDNA was validated with PCR and sanger sequencing suggesting that rDNA is involved in chromosomal dynamics. Note that no interchromosomal SV (RT, NRT, INS) involving chromosome number 12 (containing the rDNA) could be statistically validated with reasonable computer resources (12 cores, 128GB RAM) due to the high number of chimerical RP involving the rDNA.

(ii) Strains used for the experiments were derivatives from the BY4741 and BY4742 background (Brachmann *et al.*, 1998). They contained the standard auxotrophies (*leu2*Δ, *ura3*Δ, *his3*Δ as well as *met15*Δ for BY4741 and *lys2*Δ for

BY4742). The *leu2*Δ deletion was the only deletion detected in our SSV experiments because the other markers deletions are smaller than the insert-size of the MP libraries. The reference genome has the MATalpha mating type. Therefore, BY4741 strains also had a mating type different from the reference genome which appeared topologically identical too segmental duplication on chromosome number 3. Finally, *RAD27* gene deletion was achieved by replacing the locus with the *URA3* cassette which resulted in an insertion of a chromosome 5 fragment (*URA3*) into chromosome 11 (*RAD27*). All these events are classified as Markers in table 1.

(iii) SV were classified as mitotic gene conversion events when the range of left and/or right coordinates of the SV cluster was constrained in a region smaller than 1,000 bp. When this range was reduced to the read length (50 or 100bp), SV were classified as single nucleotide conversion events.

PCR validation of SV junctions were performed using the NEB Crimson LongAmp Taq DNA polymerase according to the manufacturer's recommendations. For each SV, one of the RP spanning the junction was chosen to design primers using ePrimer3. Primers were then blasted to check the absence of multiple hits. Another RP was chosen if at least one primer of the pair displayed multiple hits on the reference genome.

In order to estimate the false negative rate of SV detection, we performed PCR validations on all raw detected DUP in the high coverage WT and clb5Δ experiments. In the WT, out of 32 detected DUP, all were filtered out by the statistical modeule of Ulysses but two could be validated by PCR and sanger sequencing (supplementary table 2). In the clb5Δ experiment, 3 DUP were validated (PCR+sanger) out of 34 raw DUP that were also all filtered out. This data suggest our SSV approach has a false negative rate of at least 6-8% leading to an underestimation of SV frequency.

2.5 Calculation of SSV rates

During the exponential cellular growth of the clonal culture, SSV appear according to the Luria-Luria-Delbrück distribution (Zheng, 1999). If the mutation rate for a given type of SV is μ, and Nt the total number of cells produced after 35 generations, then, as the culture was started with a single cell, a deterministic estimation of the mean total number of SSV that occurred during the experiment is: $m = \mu(N-1) \sim \mu N$. However, SSV that occurred late in the culture are not detected because they are present at a too low frequency. Only SSV that appear early enough reach sufficiently high frequencies to get detected.

The detected SSV have a frequency in the culture set by the ratio τ/χ where χ is the physical coverage of the genome and τ the number of overlapping

RP defining a SV as a True Positive (TP). If a given SV is neutral, its frequency during exponential growth remains unchanged. Therefore, the minimal detectable SSV frequency τ_{mini}/χ (with τ_{mini} the minimal number of RP to detect a SSV of a given type) is equal to the frequency of the cell into which SSV appeared at generation g in the population: $\tau_{mini}/\chi \sim 1/2^g$. From this relation, we can deduce g: $g = ln(\chi/\tau_{mini})/ln(2)$. From $m = \mu N$, we deduce $\mu = m_i/N_i$ where m_i is the number of mutations and N_i the number of cells at generation i. m_i corresponds to the total number of detected SSV given the detection sensitivity at g. Given that $N_i = 2^g$, we can easily deduce:

$$\mu = \frac{m_i}{\zeta} \tag{1}$$

where $\zeta = 2^{ln\frac{\chi}{\tau_{mini}}/ln2}$. Note that this deterministic approach is an estimation of the mutation rate since SV formation are actually stochastic events.

2.6 Assay systems and fluctuation assays

Fluctuation assays were performed using BY4741 derived strains carrying 2 non-functional alleles of either *TRP1* or *URA3*. One allele is truncated in 3' and the other copy truncated in 5' leaving a 400 bp homology region repeated in the two alleles. A non-allelic homologous recombination event between the 2 hetero-alleles generates a SV that restores either tryptophan or uracil prototrophy. The recombination event between the 2 hetero-alleles generates either a reciprocal translocation when the 2 hetero-alleles are on two different chromosome, an inversion if they are on the same chromosome but not on the same strand and a segmental duplication if they are on the same chromosome in direct orientation (Payen *et al.*, 2008). The mutant cells with restored prototrophy can be easily selected by plating the cultures onto standard complete synthetic media depleted for tryptophan (CSM-TRP) or uracil (CSM-URA). Briefly, 30 parallel cultures (500 μL) were started by inoculating into rich media (Broth Yeast Extract-Peptone-Dextrose) $\sim$100 cells per well in a 2mL deepwell plate. Cells were grown without agitation at 30 ° C until they reached an optical density of $\geq$0.8 (6.10^6 cells/mL). OD was measured before making 100 to 200μL drops of each culture on dry selective plates and incubated for 4 days at 30°C before counting the number of mutants per drop. Mutation rates were calculated with the Generating Function (Hamon and Ycart, 2012) implemented in the bz-rates website `www.lcqb.upmc.fr/bzrates`.

2.7 Pulsed-field gel electrophoresis

Pulsed-field gel electrophoresis PFGE karyotypes were established in a CHEF Mapper XA (Bio-Rad) tank and transferred onto Hybond N+ membranes (Amersham) according to the manufacturer's recommendations. CHEF miggration

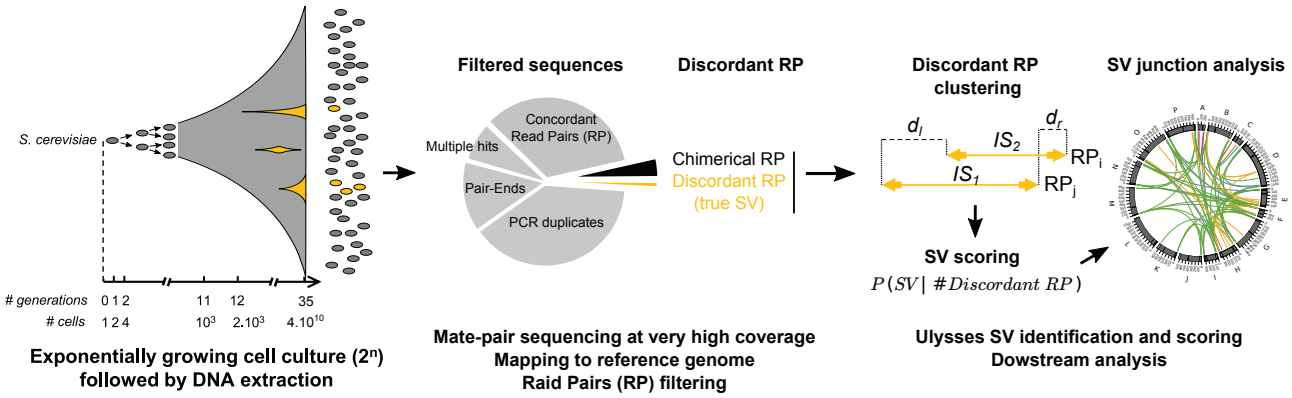


Figure 1: **Subclonal Structural Variations Detection:** A single haploid *S. cerevisiae* cell is plated onto a rich media until the colony size reaches $\sim 10^6$ cells. The colony is then transferred in rich liquid media and grown until generation 35. Genomic DNA is then extracted and 6 to 8 kb Mate-Pair sequencing is performed. RP are mapped to S288c reference genome and different steps of sequence filtering are performed before launching SV detection with Ulysses software (Gillet-Markowska *et al.*, 2015). Ulysses statistical models are used to filter out false positive SV.

program was set as follows: 10h of 60s $+60°/-60°$ 6V/cm cycles followed by 17h 90s $+60°/-60°$ 6V/cm cycles.

2.8 Growth rates

The growth rate of 3 independent mutants and wild type cells was measured by growth curve experiments in 100 μL of rich media (YPD) with a Tecan Sunrise robot in triplicates. The phenotypic lag (λ), the growth rate (μm) and the maximal cell growth (A) were then determined using the R package Grofit (Kahm *et al.*, 2010).

3 Results

3.1 The subclonal Structural Variation assay

We imagined a screen aiming at identifying the number, the types and the location of large chromosomal rearrangement occurring *de novo* in monoclonal yeast populations grown in favorable conditions. A single haploid cell was grown on a rich media plate at 30°C until it formed a colony ($\sim 10^6$ cells, 20 generations). The colony was then transferred in liquid rich media and grown at 30°C for 15 more

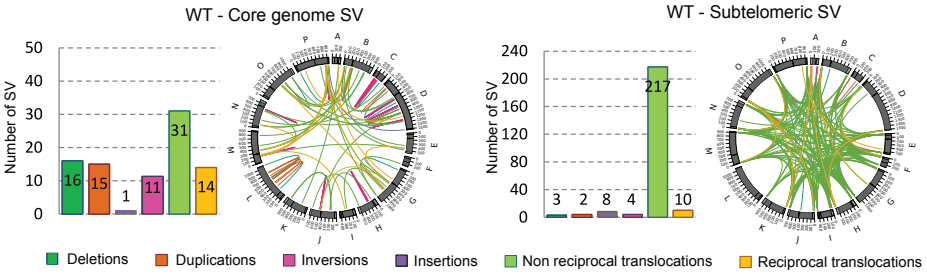


Figure 2: **Subclonal SV detection in WT cells:** SV are separated in 2 categories, namely the core genome and subtelomeric regions that correspond to regions that span from the end of the chromosomes to the first essential genes. Circos plots represent links between the DNA regions not associated in the reference genome but that form SV junctions in a fraction of the cells.

generations (35 generations in total, Fig. 1). Half of the culture was frozen for ulterior genetic analysis. Genomic DNA was prepared from the other half of the culture and mate-pair (MP) sequencing was performed. RP were mapped to the reference genome without allowing mismatches and non uniquely mapping reads were discarded (Fig. 1). PCR duplicates and contaminating paired-end reads were filtered out. Discordant RP were identified and clustered into putative SV using Ulysses (Gillet-Markowska *et al.*, 2015). MP libraries come along with elevated numbers of chimerical RP resulting from intermolecular ligations during the library construction. These chimerical RP generate a high number of FP that need to be filtered out. Therefore, SV detection was followed by a SV filtering step with Ulysses that evaluates the statistical significance of each SV against an explicit model for the generation of chimerical RP (Gillet-Markowska *et al.*, 2015) (Fig. 1). Given the absence of bottlenecks or selection in these cultures, the frequency of cells carrying a SV is expected to be extremely low. Therefore, our strategy termed 'Subclonal Structural Variation' (SSV) assay relied on sequencing the cell populations at very high physical coverage (between 5,000 and 37,000X) to be able to detect subclonal SV.

3.2 SSV experiments in WT cells

We performed a SSV experiment in WT illumina sequencing of a 8kb MP library to very high coverage (36,683X). The total number of detected SSV was 7,169 before filtering FP and 332 after filtering out FP (see methods and supplementary figure 1). As subtelomeric regions of yeast chromosomes share long regions of homology prone to the formation of SV, we classified SV occurring in these regions independently from SV occuring in the internal regions of chromosomes (called the core genome). As a result, 244 SSV fell in subtelomeric regions and 88 in the core genome (Table 1). Note that validated SV refer to statistically

Table 1: **List of detected SSV:** Markers correspond to auxotrophy markers detection, difference in mating type of reference genome. SNGC correspond to Single Nucleotide Gene Conversions. GC-tract correspond to Gene Conversion tract (< 1kb).

			SSV detection		SSV filtering								SSV		
	Physical coverage	SV type	All SSV	Validated SV	rDNA involved	Markers	Subtel. SNGC	Subtel. GC tract	Core. SNGC	Core GC tract	Total filtered SSV		**Subtelomeric regions**	**Core genome**	**Total SSV in the genome**
WT MP8kb	36,683	Deletion	814	101	82	0	0	0	0	0	82		**3**	**16**	**19**
		Inversion	4125	52	0	0	2	5	0	2	9		**4**	**11**	**15**
		Duplication	184	114	93	1	0	1	0	2	97		**2**	**15**	**17**
		Insertion	65	65	0	0	0	10	0	0	10		**8**	**1**	**9**
		Non reciprocal translocation	1154	312	0	0	2	17	0	3	22		**217**	**31**	**248**
		Reciprocal translocation	827	45	0	0	10	8	2	1	21		**10**	**14**	**24**
												Total	**244**	**88**	**332**
WT MP6kb	5,496	Deletions	297	18	0	6	1	0	3	3	13		**1**	**2**	**3**
		Inversion	2	1	0	0	1	0	0	0	1		**0**	**0**	**0**
		Duplication	32	1	0	1	0	0	0	0	1		**0**	**0**	**0**
		Insertions	11	10	0	0	0	1	1	2	4		**5**	**1**	**6**
		Non reciprocal translocation	66	3	0	0	0	0	0	0	0		**2**	**1**	**3**
		Reciprocal translocation	2	1	0	0	0	0	0	0	0		**1**	**0**	**1**
												Total	**9**	**4**	**13**
rad27Δ MP8kb	19,495	Deletion	285	45	28	1	0	0	0	1	30		**2**	**13**	**15**
		Inversion	4846	16	0	0	0	0	0	0	0		**1**	**3**	**4**
		Duplication	1206	53	23	2	0	1	2	4	32		**3**	**18**	**21**
		Insertion	941	43	0	2	0	2	0	0	4		**11**	**2**	**13**
		Non reciprocal translocation	573	106	0	0	1	11	0	0	12		**73**	**3**	**76**
		Reciprocal translocation	275	65	0	0	7	10	2	5	24		**10**	**31**	**41**
												Total	**100**	**70**	**170**
rad27Δ LJD8kb	8,721	Deletion	1000	119	3	1	11	0	56	1	72		**1**	**46**	**47**
		Inversion	7	6	0	0	0	0	0	5	5		**0**	**1**	**1**
		Duplication	967	9	0	0	0	0	0	0	0		**0**	**9**	**9**
		Insertion	1	1	0	0	0	0	0	0	0		**0**	**1**	**1**
		Non reciprocal translocation	948	125	0	0	41	1	80	3	125		**0**	**0**	**0**
		Reciprocal translocation	1	1	0	1	0	0	0	0	1		**0**	**0**	**0**
												Total	**1**	**57**	**58**
clb5Δ MP6kb	7,089	Deletions	973	9	0	7	0	0	1	0	8		**0**	**1**	**1**
		Inversion	1401	3	0	0	3	0	0	0	3		**0**	**0**	**0**
		Duplication	34	1	0	1	0	0	0	0	1		**0**	**0**	**0**
		Insertions	10	9	0	0	0	1	3	2	6		**3**	**0**	**3**
		Non reciprocal translocation	7	3	0	0	0	0	0	0	0		**3**	**0**	**3**
		Reciprocal translocation	2	1	0	0	0	1	0	0	1		**0**	**0**	**0**
												Total	**6**	**1**	**7**

significant SV, PCR validations of these SV are in progress and will be added in the future version of the manuscript. In addition, we also detected 82 deletions and 93 duplications with one endpoint lying in the rDNA.

The entire spectrum of SSV (including DUP, INV, DEL, INS, RT, NRT) was detected both in the core genome and in the subtelomeric regions (Fig. 2 and Supplementary table 3). We identified a total of 42 intrachromosomal SSV (16 DEL, 15 DUP and 11 INV) and 46 interchromosomal SSV (14 RT, 31 NRT and 1 INS) in the core genome. By contrast, very few DEL, DUP, INV were detected in subtelomeric regions (3,4 and 2 respectively) while NRT were highly enriched with 217 NRT events in subtelomeric regions. These results clearly show that chromosome dynamics differ drastically between those 2 genomic compartments. Note that another SSV experiment was performed in WT cells but to a 7 times lower coverage (5,496X, Table 1). At this coverage, only 4 SSV were detected in the core genome (2 DEL, 1 INS and 1 RT) and 9 in subtelomeric regions (1 DEL, 5 INS, 1 NRT and 1 RT) showing that the efficiency of the SSV approach strongly depends on the physical coverage of the sequencing.

3.2.1 Repeated elements enrichment at SV junctions

SV junctions location were biased towards the proximity of repeated elements. In the core genome, 71% of SV involved repeated sequences (including Long Terminal Repeats, Ty retrotransposons, tRNA, duplicated genes and subtelomeric blocks) within a 5 kb window of their predicted junction (a 5 kb window was chosen because the resolution of the experiment if roughly equal to the RP insert-size) (Supplementary table 6). In the subtelomeric regions these proportions remained steady (74%). This bias can be visualized through the dotplots of the junctions of the SV (see supplementary figures at `www.lcqb.upmc.fr/ulysses/supp/supp_data.tar.gz`). These data suggest that most detected SV occurred through NAHR mechanisms or BIR while ∼30% of SV probably involve mechanisms like NHEJ, MMEJ or MMIR. More surprisingly, these data also suggest that although subtelomeric regions have higher proportions of repeated sequences than the core genome, the frequency HR-mediated SSV is not increased in these regions.

3.2.2 Subclonal Segmental duplication landscape

In the core genome, DUP sizes ranged from 6kb to 238kb (supplementary table 2). These duplications were restricted to chromosomes 2,3,4 and 12. The number of repeated sequences detected at the junctions of DUP (73%) was nearly identical to the average (73%).

Interestingly, the *PPH21/RPL35A* and *PPH22/RPL35B* loci on chromo-

some 4 was found to be a hotspot for the formation of tandem duplications that could occur either through the recombination between the *PPH20/21* genes or the *RPL35A/B* genes. The recombination between the PPH genes produces a 96kb duplication while the recombination between the RPL genes produces a 100kb duplication. The 96 kb DUP was detected in this SSV experiment. The junction of both duplications (DUP F9up and F9down in supplementary table 7) was successfully amplified by PCR and sequenced in another SSV experiment (*clb5*Δ, see below the description of*clb5*Δ SSV experiment) suggesting these duplicated genes promote the formation of recurrent DUP. Furthermore, this locus also contains the VMA1 gene that encodes the site-specific homing-endonuclease PI-SceI. This protein is known to cleave VMA1 sequences that lack the endonuclease-coding portion to initiate homing, which introduces the endonuclease-coding sequence into the DNA (Gimble and Thorner, 1992).

In the subtelomeric regions, 1 large SSV initially identified as large DUP involving the two subtelomeric regions of the chromosome 1 is likely to represent a case of chromosomal circularization.

3.2.3 Subclonal Deletion landscape

Out of 16 detected DEL, 81% displayed repeated sequences at their junctions which is the higher proportion of all SV classes (excluding the 1 out of 1 INS that also involve repeated sequences). Out of these 16 DEL, only 1 was smaller than 10 kb and did not include essential genes. All 15 remaining DEL had sizes larger than 55 kb and spanned essential genes. As the SSV experiment was performed in haploid cells, these DEL can only be explained if they are associated with aneuploidies or if they correspond to more complex events that also involve the copy or the translocation of the deleted fragment. Another possibility would be that the deleted fragment is maintained as eccDNA due to the presence of Autonomously Replicating Sequences Mø ller *et al.* (2015). Interestingly, one of these large deletions was associated to a duplication which would have occurred between the MAT locus and HML locus (DEL id2 on chr3 and DUP id12 chr3, supplementary table 2). This DEL-DUP event can easily be explained by spontaneous mating-type switching in a subclone which would be topologically identical to a joint DUP-DEL event occurring at these loci.

3.2.4 Subclonal Inversion landscape

Inversion were mostly localized in the core genome and had sizes ranging from $\sim$5 kb to $\sim$50 kb. However, 7 INV out of 11 were smaller than 15kb (supplementary table 2 and 3). Interestingly, inversions had the lowest rate of junctions involving repeated elements (27%, supplementary table 6) suggesting that most

INV would occur independently from HR.

3.3 Subclonal translocation landscape

Interchromosomal SSV in the core genome were mainly represented by NRT and RT as only a single insertion was detected (Fig. 2, Table 1). All pairs of chromosomes (except chromosomes 3 and 12 that were excluded from the interchromosomal SSV detection due to exessive memory consumption) were involved in these events although longer chromosomes (chr2, 4, 15, 16) were more represented than short ones. RT and NRT displayed respectively 79 and 74% of junctions involving repeated sequences which is not different from intrachromosomal SSV like DUP and DEL.

3.3.1 Rates of SSV formation

We computed mutation rates (see methods) for each type of SV from the WT ∼37,000X SSV experiments. Rates were computed independently in the core genome (Fig. 3A) and subtelomeriregions (Fig. 3B) due to the long homologous segments shared between different chromosomes which might influence chromosomal plasticity.

In the core genome, the deletion rate was measured to 2.1×10^{-3} DEL/cell/division. Inversions rate was lower with 1.5×10^{-3} INV/cell/division. Insertions, reciprocal translocations and tandem duplication rate were slightly higher with mutation rates of 5.3×10^{-3} INS/cell/division, 3.1×10^{-3} RT/cell/division and 6.3×10^{-3} DUP/cell/division. Non reciprocal translocation rate was even higher with 10^{-2} NRT/cell/division.

Subtelomeric regions presented 2 to 3 times lower rates of deletions, inversion and duplications (Fig. 3B). Non reciprocal replication rate was however 7 times higher in subtelomeric regions than in the core genome.

3.4 SSV experiments in rad27Δ and clb5Δ backgrounds

We also investigated 2 mutant backgrounds by sequencing subclonal cultures of either *rad27*Δ or *clb5*Δ with Illumina 6 to 8kb mate pair libraries. Two independent *rad27*Δ experiments were performed and sequenced to either medium or high coverage (7,089X and 19,495X, Table 1). *clb5*Δ SSV experiment was only performed once and sequenced to medium coverage (8,721X). Medium coverage experiments displayed a limited number of SSV (Table 1). The *clb5*Δ experiment only displayed 1 DEL in the core genome and 3 INS and 3 NRT in subtelomeric

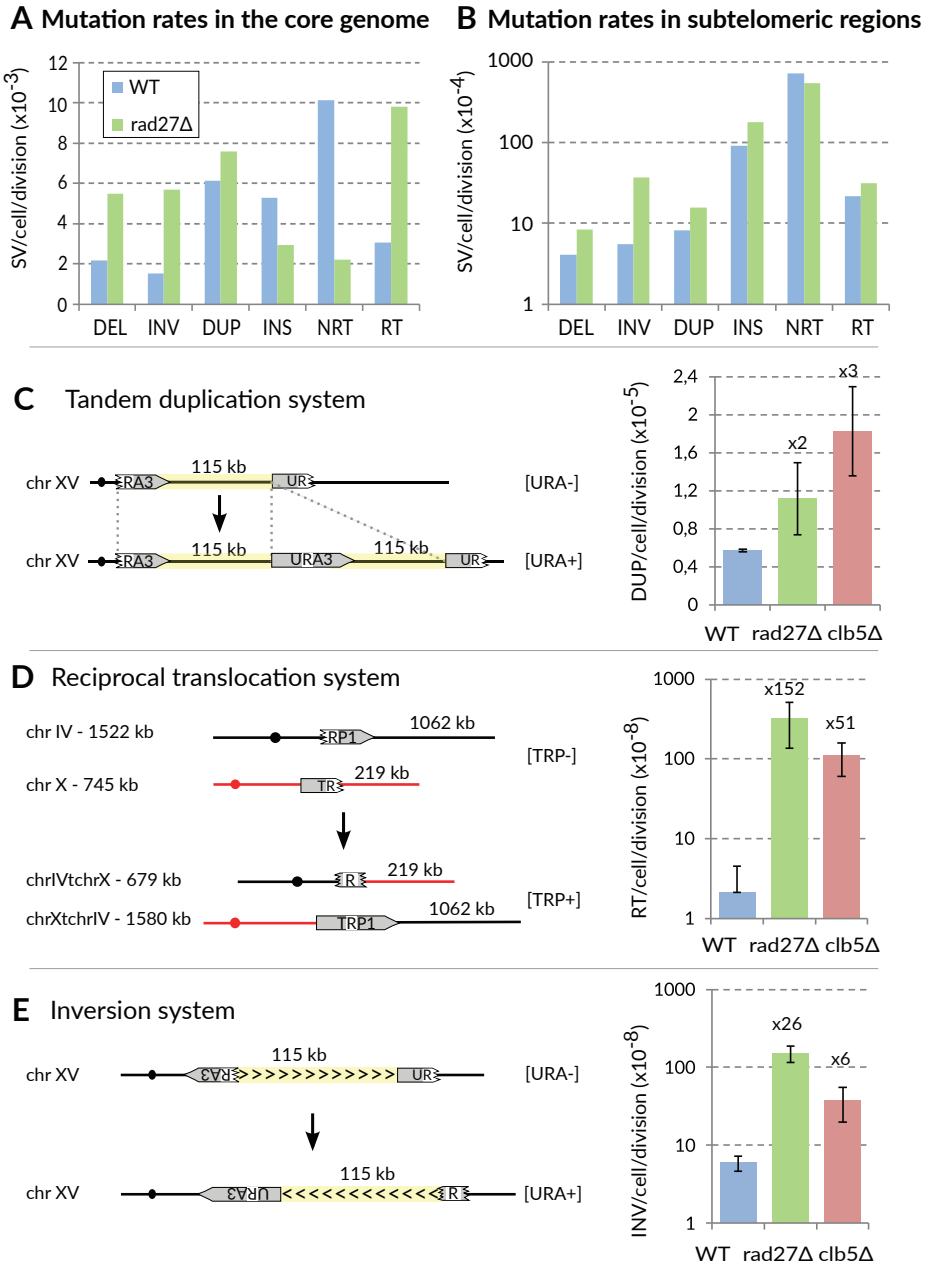


Figure 3: **SSV and Assay systems mutation rates:** (A) SSV rates in the core genome for the WT MP8 kb 36,000X SSV experiment and the *rad27*Δ MP8 kb 19,000X SSV experiment. (B) Same as A but for subtelomeric regions. (C) Left: Segmental duplication measurement system scheme. Right: histogram showing DUP rate in WT, *rad27*Δ and *clb5*Δ backgrounds. The numbers above the bars indicate the SSV rate increase fold relatively to WT cells. (D) Same as C for the reciprocal translocation assay system. (E) Same as D for the reciprocal translocation assay system.

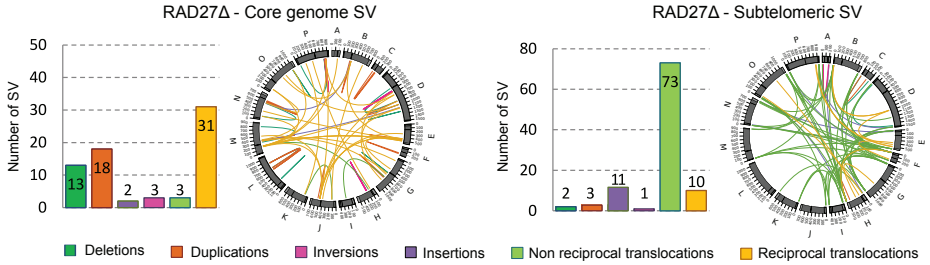


Figure 4: **Subclonal SV detection in rad27Δ SSV experiment:** same as Fig. 2

regions. Similarly, the *rad27*Δ medium coverage experiment displayed 1 INV, 9 DUP and 1 INS in the core genome and 1 DEL. Note that in this experiment, the number of DEL was elevated (46 deletions) due to frequent Ty retrotransposons deletions discussed hereafter. These data again show that the efficiency of the SSV approach strongly depends on the physical coverage of the sequencing and that a minimal physical coverage of 20,000X is required.

The *rad27*Δ high coverage SSV experiment harbored a 1.8 lower physical coverage than the WT high coverage SSV. Nonetheless, 170 SV were detected in this experiment: 70 SV in the core genome and 100 in subtelomeric regions (Table 1, Fig. 4, Supplementary table 4 and 5). In the core genome, *rad27*Δ cells displayed a 2.5 and 3.8 times higher SSV rate than the WT for DEL and INV respectively while the DUP rate was nearly identical (1.2 increase fold). Strikingly, the NRT rate was 4.5 times lower in *rad27*Δ cells than in WT while the RT rate was 3.2 times higher in *rad27*Δ than in WT cells (Fig. 3A). In the subtelomeric regions, *rad27*Δ displays SSV rates around twice those of WT cells for DEL, DUP, INS and RT. INV rates were even more increased with a 6.8 fold increase. NRT rate was nearly unchanged with 1.3 fold decrease in *rad27*Δ.

In the high coverage *rad27*Δ SSV experiment, SSV landscapes were not different from WT cells with 83% in the *rad27*Δ experiment involved repeated sequences. Interestingly, nearly 50% (6 out of 13) of the deletions detected in the core genome of the *rad27*Δ experiment are retrotransposable (Ty) elements deletions. In the high coverage experiment, out of the 6 Ty elements deletions, 4 were validated by PCR analysis while the other 2 were inconclusive due to weak PCR signals. These deletions harbored frequencies in the population (amount of RP of the SV divided by the local physical coverage) ranging from 0,6% to 6,14%. Noticeably, while 3 experimentally validated deletions could only be amplified in genomic DNA from the rad27Δ experiment and not in standard BY4741 negative control, a deletion of *YHRCTy1-1* with a different size could also be detected in the control DNA. These 2 different deletions can be explained by the presence of a Ty1 LTR near the Ty1 that could lead to a 10 kb deletion event. The independent *rad27*Δ medium coverage experiment also displayed a Ty deletion enrichment with

29 Ty deletions out of a total of 46 events with the more frequent deletions present in 43% and 11% of the population (Table 1). Note that although this medium coverage experiment had a nearly 3 times lower coverage than the high coverage replicate it displayed more than 3 times more deletions.

3.5 SSV rates compared to SV rates from assay systems

We compared the previous SSV rates with mutation rates from 3 assay systems that we designed to measure duplication, inversion and reciprocal translocation rates in the core genome (Fig. 3C,D,E). Each genetic system is composed of a 3' and 5'-deleted allele of either *URA3* (DUP, INV) or *TRP1* (RT). The 2 hetero-alleles share 400 bp of homology into which recombination events create DUP, INV or RT depending on the position of the cassettes in the genome. The recombination events produce a functional gene that restores either the uracil or the tryptophan prototrophy. Therefore, mutants cells can be selected on plated depleted for either uracil or tryptophan. PFGE was performed for 10 randomly chosen translocated mutants and 10 mutants with duplications. All strains harbored the expected chromosome size showing that this selective systems mainly produce the expected recombinants (data not shown).

Fluctuation assays were performed in both WT, *rad27*Δ and *clb5*Δ cells with the 3 genetic systems. In all 3 systems, WT cells harbored lower SV rates than the 2 mutator backgrounds (Fig. 3C,D,E, supplementary table 3). For the WT cells, we measured DUP, INV and RT rates to 5.7×10^{-6} DUP/cell/division (identical to the rates initially measured in (Payen *et al.*, 2008)), 5.7×10^{-8} INV/cell/division and 2.1×10^{-8} RT/cell/division. In the *rad27*Δ cells, the SV rates were 1.1×10^{-5} DUP/cell/division, 1.5×10^{-6} INV/cell/division and 3.3×10^{-6} RT/cell/division. Finally, in the *clb5*Δ the mutation rates were 1.8×10^{-5} DUP/cell/division, 3.7×10^{-7} INV/cell/division and 1.1×10^{-6} RT/cell/division. Both mutator genes strongly affected RT rates with 152 and 51 increase fold for *rad27*Δ and *clb5*Δ, respectively. Surprisingly, they are a very moderate effect on DUP rates with 2 and 3 times increase fold, respectively.

In the SSV assay, *rad27*Δ had the strongest effect on the RT and INV rates with 3.79 and 3.21 increase folds (Fig. 3A) and showed nearly no effect on DUP rate. Interestingly, the same trends were observed with the assay systems measuring SV rates at a single locus. However, the amplitude of those trends was much higher with assay systems than with the SSV assay.

4 Discussion

So far, the only experimental method able to detect SV at the genome-scale was MA lines. Using very high physical coverage sequencing of yeast monoclonal populations we have been able to identify a large number of SSV in WT genetic background. Our SSV detection approach allowed to detect one or 2 orders of magnitude more SV than a single ∼4800 generations MA experiments (Lynch *et al.*, 2008) or even ∼310,000 generation MA lines (Zhu *et al.*, 2014). Besides its massive improvement in the SV detection sensitivity, the SSV approach also allowed to sensibly decrease the experiment duration from nearly 2 years for ∼4800 consecutive generations to 3 days for ∼35 generations. We were able to identify 332 SSV covering the entire spectrum of rearrangements in a single WT monoclonal culture. SSV correspond to *de novo* SV occurring spontaneously at the genome scale in a cell population. We estimated SSV rates for each type of SV in the core genome: in increasing order, INV, DEL, RT, INS, DUP and NRT had rate of 1.5, 2.1, 3.1, 5.3, 6.3 and 10×10^{-3}/cell/division. To our knowledge, this is the first description of the SSV landscape at the genome scale in yeast. The deletion and insertion rate have been previously estimated in WT haploid cells to 1.2×10^{-2} DEL/cell/division and 2×10^{-2} INS/cell/division (Lynch *et al.*, 2008) which is respectively 6 and 4 times higher than our measurements. This differences could be explained by the fact that our statistical models filter out a high proportion of true positive SV (Gillet-Markowska *et al.*, 2015). Furthermore, MP libraries are limited to the detection of large deletions (larger than $\sim 6\times$ the library insert size standard deviation) which exclude all DEL that have sizes below a few kilo bases. In addition, although the usage of large insert-size libraries allows to detect all SSV that involve repeated regions smaller than the insert size, all SSV involving large repeats like in subtelomeric blocks will be filtered out. We estimated the false negative rate by testing all detected DUP by PCR in two independent SSV experiments and measured it to be at least 10%. However, this proportion could be in reality higher due to completely undetected DUP that occur in highly repetitive regions. These results suggest that although we detected several hundreds of SSV in a single experiment, the SSV rates that we inferred are conservative.

Subtelomeric regions are composed of large duplicated blocks undergoing high frequency of ectopic recombination events as measured between Y' elements (Louis and Haber, 1990, 1992; Nishant *et al.*, 2010). This high frequency of ectopic exchange, along with a nonrandom choice of subtelomeres to exchange with, could allow for amplification and diversification of the sequences in the subtelomeric regions. Interestingly, our measurement of intrachromosomal SV rates including DUP, DEL and INV showed a 2 to 3 times reduction in these regions. This difference could possibly be due to the large repeated blocks in subtelomeres, which would impair the detection of many SSV. Despite this limitation, interchromosomal SV including NRT, INS and RT had increased rates of up to 7 times. These NRT events were

not restricted to Y' elements and spanned over the whole length of subtelomeric duplicated blocks.

The analysis of the chromosomal context of each SV in SSV experiments revealed that SV preferentially formed by the ectopic recombination between homologous sequences, including retrotransposons, long terminal repeats, tRNA and repeated genes. This trend was identical for all type of SV except for INV which suggest that homologous recombination is the main source of *de novo* SV in yeast. Serero *et al.* and Lynch *et al.* found in MA lines experiments that SV junctions mostly localized within LTR or Ty elements but did not confirm this trend in WT cells as no large SV were detected in these experiments. Another observation strengthened our belief that homology is the primary vector for SV formation: we found a DUP hotspot in the right arm of chromosome 4 where 2 pairs of duplicated genes were located in direct orientation at ~ 15kb. This loci were found to be involved in DUP events in independent SSV experiments which suggests that repeated genes favor recurrent DUP formation.

rad27Δ is believed to be one of the top mutator genes in the yeast genome. As this mutant shows a ~ 1000 times GCR increase relatively to WT(Chen and Kolodner, 1999). Our genetic systems showed that *rad27*Δ does not affect all types of SV to the same extent: DUP rate was only increased twice while INV and RT rates were increased respectively 26 and 152 times relatively to WT. Such increased SV rates like in GCR or in our assay systems were not measured in SSV experiments that showed SV rates only 2 to 3 times increased relatively to the WT suggesting that chromosomal dynamics observed with selective assay systems are not representative of the true SV dynamics and that the mutator effect of *rad27*Δ would be more limited to the genome scale than what was previously thought. It is also interesting to note that an increase factor of x1,000 would not be compatible with the SSV rates that we inferred between 10^{-3} and 10^{-2}. Interestingly, we showed that the *rad27*Δ cells displayed an increased frequency of Ty deletion over WT. *RAD27* is a known inhibitor of Ty1 mobility in *cerevisiae* (Sundararajan *et al.*, 2003). In this situation we should have observed an increase of Ty elements insertions rather than deletions which was not the case. This suggests the existence of a new role for the RAD27p in stabilizing Ty elements in the genome.

More generally, the SSV strategy can be applied to any organism or strain with a moderate size genome like yeast in which very high physical coverage of the genome can be reached. As MP libraries become easier to produce we believe the SSV strategy will be useful to the community of geneticists to study chromosomal dynamics. Furthermore, given the increasing amount of evidence that SV would play a major role in phenotypic variation and adaptation, it would be of major interest to quantify the functional impact of SV on the phenotypic diversity. However, the study of the phenotypic impact of the multiple SV we detected in SSV assays would require to isolate these mutant cells. Most of the detected subclones had frequencies ¡1%

which make their screen difficult. Nonetheless, a few dozens had frequencies ranging from 1% to 6%. In this conditions, a screen of frozen cells could be considered to isolate mutant cells and analyze their effect on various growth conditions as well as on meiotic viability.

References

Brachmann, C. B., Davies, A., Cost, G. J., Caputo, E., Li, J., Hieter, P., and Boeke, J. D. (1998). Designer deletion strains derived from Saccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*, **14**(2), 115–132.

BroadInstitute (2009). Picard tools.

Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. a., Morsberger, L. a., Latimer, C., McLaren, S., Lin, M.-L., McBride, D. J., Varela, I., Nik-Zainal, S. a., Leroy, C., Jia, M., Menzies, A., Butler, A. P., Teague, J. W., Griffin, C. a., Burton, J., Swerdlow, H., Quail, M. a., Stratton, M. R., Iacobuzio-Donahue, C., and Futreal, P. A. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, **467**(7319), 1109–13.

Chen, C. and Kolodner, R. D. (1999). Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants. *Nature*, **23**(september), 81–85.

Conrad, D. F. and Hurles, M. E. (2007). The population genetics of structural variation. *Nature genetics*, **39**(7 Suppl), S30–6.

de Cid, R., Riveira-Munoz, E., Zeeuwen, P. L. J. M., Robarge, J., Liao, W., Dannhauser, E. N., Giardina, E., Stuart, P. E., Nair, R., Helms, C., Escaramís, G., Ballana, E., Martín-Ezquerra, G., den Heijer, M., Kamsteeg, M., Joosten, I., Eichler, E. E., Lázaro, C., Pujol, R. M., Armengol, L., Abecasis, G., Elder, J. T., Novelli, G., Armour, J. a. L., Kwok, P.-Y., Bowcock, A., Schalkwijk, J., and Estivill, X. (2009). Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature genetics*, **41**(2), 211–5.

Donaldson, a. D., Raghuraman, M. K., Friedman, K. L., Cross, F. R., Brewer, B. J., and Fangman, W. L. (1998). CLB5-dependent activation of late replication origins in S. cerevisiae. *Molecular cell*, **2**(2), 173–182.

Dorsey, M., Peterson, C., Bray, K., and Paqui, C. E. (1992). Spontaneous Amplification of the ADH4 gene in Saccharomyces cerevisiae. *Genetics*, **132**(4), 943–950.

Epstein, C. B. and Cross, F. R. (1992). Clb5 - a Novel B-Cyclin from Budding Yeast with a Role in S-Phase. *Genes & Development*, **6**(9), 1695–1706.

Exinger, F. and Lacroute, F. (1979). Genetic evidence for the creation of a reinitiation site by mutation inside the yeast ura 2 gene. *Molecular & general genetics : MGG*, **173**(1), 109–113.

Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J. M., Gough, S. C. L., de Smith, A., Blakemore, A. I. F., Froguel, P., Owen, C. J., Pearce, S. H. S., Teixeira, L., Guillevin, L., Graham, D. S. C., Pusey, C. D., Cook, H. T., Vyse, T. J., and Aitman, T. J. (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics*, **39**(6), 721–3.

Fink, G. R. and Styles, C. a. (1974). Gene conversion of deletions in the HIS4 region of yeast. *Genetics*, **77**(2), 231–244.

Freeman, J., Perry, G., Feuk, L., Redon, R., McCarroll, S., Altshuler, D., Aburatani, H., Jones, K., Tyler-Smith, C., Hurles, M., Carter, N., Scherer, S., and Lee, C. (2006). Copy number variation: New insights in genome diversity. *Genome research*, **16**(8), 949–961.

Gillet-Markowska, A., Richard, H., Fischer, G., and Lafontaine, I. (2015). Ulysses: Accurate detection of low-frequency structural variations in large insert-size sequencing libraries. *Bioinformatics*, **31**(6), 801–808.

Gimble, F. S. and Thorner, J. (1992). Homing of a DNA endonuclease gene by meiotic gene conversion in Saccharomyces cerevisiae. *Nature*, **357**(6376), 301–306.

Girirajan, S., Brkanac, Z., Coe, B. P., Baker, C., Vives, L., Vu, T. H., Shafer, N., Bernier, R., Ferrero, G. B., Silengo, M., Warren, S. T., Moreno, C. S., Fichera, M., Romano, C., Raskind, W. H., and Eichler, E. E. (2011). Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS genetics*, **7**(11), e1002334.

Hamon, A. and Ycart, B. (2012). Statistics for the Luria-Delbrück distribution. *Electronic Journal of Statistics*, **6**, 1251–1272.

Hansen, M. A. (1997). Biopieces.

Hoffmann, A. a. and Rieseberg, L. H. (2008). Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annual Review of Ecology, Evolution, and Systematics*, **39**(1), 21–42.

Hollox, E. J., Huffmeier, U., Zeeuwen, P. L. J. M., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P. C. M., Traupe, H., de Jongh, G., den Heijer, M., Reis, A., Armour, J. a. L., and Schalkwijk, J. (2008). Psoriasis is associated with increased beta-defensin genomic copy number. *Nature genetics*, **40**(1), 23–5.

Kahm, M., Hasenbrink, G., Lichtenberg-Frate, H., Ludwig, J., and Kschischo, M. (2010). Grofit: Fitting biological growth curves. *Journal of Statistical Software*, **33**(7).

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. a., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. a., Altshuler, D. a., Peiffer, D. a., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. a., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**(7191), 56–64.

Knouse, K. a., Wu, J., Whittaker, C. a., and Amon, A. (2014). Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proceedings of the National Academy of Sciences*, **111**(37), 13409–13414.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, a. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M., and Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, **318**(5849), 420–6.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **25**(14), 1754–1760.

Liebman, S., Singh, A., and Sherman, F. (1979). A mutator affecting the region of the iso-1-cytochrome c gene in yeast. *Genetics*, **92**(3), 783–802.

Louis, E. J. and Haber, J. E. (1990). The subtelomeric Y' repeat family in Saccharomyces cerevisiae: an experimental system for repeated sequence evolution. *Genetics*, **124**(3), 533–545.

Louis, E. J. and Haber, J. E. (1992). The Structure and Evolution of Subtelomeric Y' Repeats in Saccharomyces cerevisiae. *Genetics*, **1331**, 559–574.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S., Hartl, D. L., and Thomas, W. K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(27), 9272–7.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**(1), 10.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stütz, A. M., Urban, A. E., Walker, J. a., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. a., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. a., and Korbel, J. O. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**(7332), 59–65.

Mø ller, H. D., Parsons, L., Jø rgensen, T. S., Botstein, D., and Regenberg, B. (2015). Extrachromosomal circular DNA is common in yeast. *Proceedings of the National Academy of Sciences*, page 201508825.

Nishant, K. T., Wei, W., Mancera, E., Argueso, J. L., Schlattl, A., Delhomme, N., Ma, X., Bustamante, C. D., Korbel, J. O., Gu, Z., Steinmetz, L. M., and Alani, E. (2010). The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS genetics*, **6**(9), e1001109.

Payen, C., Koszul, R., Dujon, B., and Fischer, G. (2008). Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. *PLoS genetics*, **4**(9), e1000175.

Petes, T. D. (1979). Yeast ribosomal DNA genes are located on chromosome XII. *Proceedings of the National Academy of Sciences of the United States of America*, **76**(1), 410–414.

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., Abrahams, B. S., Almeida, J., Bacchelli, E., Bader, G. D., Bailey, A. J., Baird, G., Battaglia, A., Berney, T., Bolshakova, N., Bölte, S., Bolton, P. F., Bourgeron, T., Brennan, S., Brian, J., Bryson, S. E., Carson, A. R., Casallo, G., Casey, J., Chung, B. H. Y., Cochrane, L., Corsello, C., Crawford, E. L., Crossett, A., Cytrynbaum, C., Dawson, G., de Jonge, M., Delorme, R., Drmic, I., Duketis, E., Duque, F., Estes, A., Farrar, P., Fernandez, B. a., Folstein, S. E., Fombonne, E., Freitag, C. M., Gilbert, J., Gillberg, C., Glessner, J. T., Goldberg, J., Green, A., Green, J., Guter, S. J., Hakonarson, H., Heron, E. a., Hill, M., Holt, R., Howe, J. L., Hughes, G., Hus, V., Igliozzi, R., Kim, C., Klauck, S. M., Kolevzon, A., Korvatska, O., Kustanovich, V., Lajonchere, C. M., Lamb, J. a., Laskawiec, M., Leboyer, M., Le Couteur, A., Leventhal, B. L., Lionel, A. C., Liu, X.-Q., Lord, C., Lotspeich, L., Lund, S. C., Maestrini, E., Mahoney, W., Mantoulan, C., Marshall, C. R., McConachie, H., McDougle, C. J., McGrath, J., McMahon, W. M., Merikangas, A., Migita, O., Minshew, N. J., Mirza, G. K., Munson, J., Nelson, S. F., Noakes, C., Noor, A., Nygren, G., Oliveira, G., Papanikolaou, K., Parr, J. R., Parrini, B., Paton, T., Pickles, A., Pilorge, M., Piven, J., Ponting, C. P., Posey, D. J., Poustka, A., Poustka, F., Prasad, A., Ragoussis, J., Renshaw, K., Rickaby, J., Roberts, W., Roeder, K., Roge, B., Rutter, M. L., Bierut, L. J., Rice, J. P., Salt, J., Sansom, K., Sato, D., Segurado, R., Sequeira, A. F., Senman, L., Shah, N., Sheffield, V. C., Soorya, L., Sousa, I., Stein, O., Sykes, N., Stoppioni, V., Strawbridge, C., Tancredi, R., Tansey, K., Thiruvahindrapduram, B., Thompson, A. P., Thomson, S., Tryfon, A., Tsiantis, J., Van Engeland, H., Vincent, J. B., Volkmar, F., Wallace, S., Wang, K., Wang, Z., Wassink, T. H., Webber, C., Weksberg, R., Wing, K., Wittemeyer, K., Wood, S., Wu, J., Yaspan, B. L., Zurawiecki, D., Zwaigenbaum, L., Buxbaum, J. D., Cantor, R. M., Cook, E. H., Coon, H., Cuccaro, M. L., Devlin, B., Ennis, S., Gallagher, L., Geschwind, D. H., Gill, M., Haines, J. L., Hallmayer, J., Miller, J., Monaco, A. P., Nurnberger, J. I., Paterson, A. D., Pericak-Vance, M. a., Schellenberg, G. D., Szatmari, P., Vicente, A. M., Vieland, V. J., Wijsman, E. M., Scherer, S. W., Sutcliffe, J. S., and Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**(7304), 368–72.

Putnam, C. D., Allen-Soltero, S. R., Martinez, S. L., Chan, J. E., Hayes, T. K., and Kolodner, R. D. (2012). Bioinformatic identification of genes suppressing genome instability. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(47), E3251–9.

Schwob, E. and Nasmyth, K. (1993). CLB5 and CLB6, a new pair of B cyclins involved in DNA replication in Saccharomyces cerevisiae. *Genes and Development*, **7**(7 A), 1160–1175.

Serero, A., Jubin, C., Loeillet, S., Legoix-Né, P., and Nicolas, A. G. (2014). Mutational landscape of yeast mutator strains. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(5), 1897–902.

Sherman, F., Stewart, J. W., Jackson, M., Gilmore, R. a., and Parker, J. H. (1974). Mutants of yeast defective in iso-1-cytochrome c. *Genetics*, **77**(2), 255–284.

Stephens, P. J., McBride, D. J., Lin, M.-L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. a., Leroy, C., Edkins, S., Mudie, L. J., Greenman, C. D., Jia, M., Latimer, C., Teague, J. W., Lau, K. W., Burton, J., Quail, M. a., Swerdlow, H., Churcher, C., Natrajan, R., Sieuwerts, A. M., Martens, J. W. M., Silver, D. P., Langerø d, A., Russnes, H. E. G., Foekens, J. a., Reis-Filho, J. S., van 't Veer, L., Richardson, A. L., Bø rresen Dale, A.-L., Campbell, P. J., Futreal, P. A., and Stratton, M. R. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**(7276), 1005–1010.

Sundararajan, A., Lee, B. S., and Garfinkel, D. J. (2003). The Rad27 (Fen-1) nuclease inhibits Ty1 mobility in Saccharomyces cerevisiae. *Genetics*, **163**(1), 55–67.

Szostak, J. W. and Wu, R. (1980). Unequal crossing over in the ribosomal DNA of Saccharomyces cerevisiae. *Nature*, **284**(5755), 426–430.

Tishkoff, D. X., Filosi, N., Gaida, G. M., and Kolodner, R. D. (1997). A novel mutation avoidance mechanism dependent on S. cerevisiae RAD27 is distinct from DNA mismatch repair. *Cell*, **88**(2), 253–263.

Wach, a., Brachat, a., Pöhlmann, R., and Philippsen, P. (1994). New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. *Yeast (Chichester, England)*, **10**(13), 1793–1808.

Wang, J., Fan, H. C., Behr, B., and Quake, S. R. (2012). Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*, **150**(2), 402–412.

Wang, Y. and Navin, N. (2015). Advances and Applications of Single-Cell Sequencing Technologies. *Molecular Cell*, **58**(4), 598–609.

Welch, J. W., Maloney, D. H., and Fogel, S. (1990). Unequal crossing-over and gene conversion at the amplified CUP1 locus of yeast. *Molecular and General Genetics*, **222**(2-3), 304–310.

Winzeler, E. a., Winzeler, E. a., Shoemaker, D. D., Shoemaker, D. D., Astromoff, A., Astromoff, A., Liang, H., Liang, H., Anderson, K., Anderson, K., Andre, B., Andre, B., Bangham, R., Bangham, R., Benito, R., Benito, R., Boeke, J. D., Boeke, J. D., Bussey, H., Bussey, H., Chu, A. M., Chu, A. M., Connelly, C., Connelly, C., Davis, K., Davis, K., Dietrich, F., Dietrich, F., Dow, S. W., Dow, S. W., Bakkoury, M. E., Bakkoury, M. E., Friend, S. H., Friend, S. H., Gentalen, E., Gentalen, E., Giaever, G., Giaever, G., Hegemann, J. H., Hegemann, J. H., Jones, T., Jones, T., Laub, M., Laub, M., Liao, H., Liao, H., Liebundguth, N., Liebundguth, N., Lockhart, D. J., Lockhart, D. J., Lucau-danila, A., Lucau-danila, A., Lussier, M., Lussier, M., Menard, P., Menard, P., Mittmann, M., Mittmann, M., Pai, C., Pai, C., Rebischung, C., Rebischung, C., Revuelta, J. L., Revuelta, J. L., Riles, L., Riles, L., Roberts, C. J., Roberts, C. J., Ross-macdonald, P., Ross-macdonald, P., Scherens, B., Scherens, B., Snyder, M., Snyder, M., Sookhai-mahadeo, S., Sookhai-mahadeo, S., Storms, R. K., Storms, R. K., Ve, S., Ve, S., Voet, M., Voet, M., Volckaert, G., Volckaert, G., Ward, T. R., Ward, T. R., Wysocki, R., Wysocki, R., Yen, G. S., Yen, G. S., Yu, K., Yu, K., Zimmermann, K., Zimmermann, K., Philippsen, P., Philippsen, P., Johnston, M., Johnston, M., Davis, R. W., and Davis, R. W. (1999). Functional Characterization of the. *Science*, **285**(August), 901–906.

Zheng, Q. (1999). Progress of a half century in the study of the Luria-Delbruck distribution. *Mathematical Biosciences*, **162**, 1–32.

Zhu, Y. O., Siegal, M. L., Hall, D. W., and Petrov, D. a. (2014). Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(22), E2310–8.

3 - Identification des déterminants génétiques de la stabilité du génome chez *S. cerevisiae*

3.1 - Développement d'un site web pour mesurer les taux de mutation à partir de données de fluctuations

Dans la partie précédente, nous avons mesuré l'impact quantitatif des SV en identifiant des SV sub-clonales dans les génomes de levures sauvages, *rad27Δ* et *clb5Δ*. Nous avons également développé des systèmes génétiques permettant de sélectionner des évènements de duplications, d'inversions ou de translocations réciproques. Comme nous, l'avons évoqué dans l'introduction, les systèmes génétiques ne permettent pas de mesurer directement des taux de mutation mais seulement des fréquences de mutants dans les cultures. En effet, la détermination des taux de mutation est une problèmatique classique de génétique depuis les travaux pionniers de Luria & Delbrück (LD) ainsi que Lea & Coulson dans les années 40 qui ont développé le premier estimateur du nombre de mutations à partir de la fréquence des mutants. Par la suite et jusqu'à aujourd'hui, de nombreux autres estimateurs ont été développés.

L'un des estimateurs les plus utilisés, MSS (cf. chapitre III de l'introduction) est implémenté dans le site internet Falcor (www.keshavsingh.org/protocols/FALCOR.html). Ce site web ne permet cependant pas de prendre en compte la croissance différentielle des cellules mutantes (appelé « b ») ce qui peut conduire à une sur ou sous-estimation des taux de mutations. De plus, il arrive que lors de la réalisation de tests de fluctuation que seule une petite fraction des cultures soit étalée. Ceci constitue une perte d'information qui conduit à la sous-estimation du taux de mutation. Ce paramètre appelé « z » ou efficacité d'étalement n'est pas pris en compte par Falcor. Enfin, Falcor n'indique pas à l'utilisateur la qualité de l'ajustement entre la distribution observée du nombre de mutants par culture et la distribution théorique estimée par le modèle. Cette qualité de l'ajustement permet d'estimer si le modèle utilisé par l'estimateur du taux de mutation est adapté ou si une déviation du modèle de LD intervient dans ce contexte expérimental.

Nous disposons aujourd'hui d'estimateurs permettant de calculer des taux de mutations avec une excellente exactitude, c'est-à-dire proche de la réalité. Ces derniers prennent en compte différentes déviations du modèle de LD comme la croissance différentielle des cellules mutantes ou l'efficacité d'étalement. L'implémentation actuelle de ces algorithmes peut cependant nécessiter de disposer de licences Mathematica (Salvador 2.3, (Zheng, 1999)) ou

Matlab (Lang and Murray, 2008) et requière dans tous les cas l'utilisation de lignes de commandes qui ne facilitent pas l'expérience utilisateur.

Afin de combler ces manques, nous avons développé le site internet bz-rates (Gillet-markowska et al., 2015) (www.lcqb.upmc.fr/bzrates) qui permet de prendre en compte la croissance différentielle des mutants (b) et l'efficacité d'étalement (z) dans le calcul des taux. Il offre la possibilité de calculer rapidement les taux de mutations à partir d'expériences de fluctuation réalisées avec n'importe quels systèmes génétiques (mutations ponctuelles et SV) et dans n'importe quel organisme permettant de respecter les conditions expérimentales requises pour réaliser un test de fluctuation (cf. chapitre III de l'introduction). Bz-rates utilise l'estimateur 'Generating Function' de (Hamon and Ycart, 2012; Ycart, 2013) qui possède plusieurs avantages par rapport aux autres estimateurs. Il permet en premier lieu d'estimer conjointement le taux de mutation et la croissance différentielle des cellules mutantes ou de directement la prendre en compte lorsqu'elle a été mesurée expérimentalement. De plus, cet estimateur est beaucoup moins sensible aux jackpots (tests de fluctuation où une mutation précoce a lieu et qui produit un grand nombre de mutants dans la culture) que les estimateurs utilisant une approche par maximum de vraisemblance. Un second avantage à ne pas utiliser de maximum de vraisemblance dans la 'Generating Function' est de rendre la vitesse de calcul virtuellement nulle. Nous avons également implémenté dans bz-rates une correction du taux de mutation pour prendre en compte l'efficacité d'étalement 'z' (Jones, 1993; Stewart et al., 1990). Finalement, bz-rates fournit une représentation graphique de la qualité de l'ajustement du modèle sur les données expérimentales. Un test de chi2 de Pearson est réalisé pour établir la qualité du *fit*. L'estimateur GF fournit également les intervalles de confiances des taux de mutations et de l'estimation du fitness des mutants. L'ensemble de ces éléments doit permettre à l'utilisateur d'estimer la qualité de la prédiction et du modèle. Si le *fit* est correct mais que l'intervalle de confiance est trop grand, il suffira à l'utilisateur d'augmenter le nombre de cultures en parallèle lors du test de fluctuation afin d'obtenir un résultat plus reproductible. En revanche, si la qualité de l'ajustement n'est pas satisfaisante, cela indiquera à l'utilisateur qu'il doit probablement utiliser un estimateur prenant en compte d'autres déviations de la loi de Luria et Delbrück.

Bz-rates a été développé avec un environnement flexible (Django 1.6.6) et est distribué sous les termes de la licence « GNU General Public License ». À ce titre, des clones du site web original peuvent être facilement mis en places et d'autres estimateurs implémentés. L'article décrivant l'implémentation de ce site web a été publié dans le journal G3 : Genes | Genomes | Genetics.

Article 3

bz-rates: a web-tool to estimate mutation rates from fluctuation analysis

Alexandre Gillet-Markowska, Guillaume Louvel, and Gilles Fischer

G3 (Bethesda). 2015 Sep 2. pii: g3.115.019836. doi: 10.1534/g3.115.019836.

G3: Genes|Genomes|Genetics Early Online, published on September 2, 2015 as doi:10.1534/g3.115.019836

bz-rates: a web-tool to estimate mutation rates from fluctuation analysis

Alexandre Gillet-Markowska*, Guillaume Louvel* and Gilles Fischer*,[1]

*Sorbonne Universités, UPMC Univ. Paris 06, Institut de Biologie Paris-Seine UMR 7238, Biologie Computationnelle et Quantitative, F-75005, Paris, France , CNRS, Institut de Biologie Paris-Seine UMR7238, Biologie Computationnelle et Quantitative, F-75005, Paris, France

ABSTRACT Fluctuation analyses is the standard experimental method for measuring mutation rates in microorganisms. The appearance of mutants is classically described by a Luria-Delbrück distribution composed of two parameters: the number of mutations per culture (m) and the differential growth rate between mutant and wild-type cells (b). A precise estimation of these 2 parameters is a prerequisite to the calculation of the mutation rate. Here, we developed *bz-rates*, a web-tool to calculate mutation rates that provides three useful advances over existing web-tools. First, it allows taking into account b, the differential growth rate between mutant and wild-type cells, in the estimation of m with the Generating Function (GF). Secondly, *bz-rates* allows the user to take into account a deviation from the Luria-Delbrück distribution called z, the plating efficiency, in the estimation of m. Finally, the web-site provides a graphical visualization of the goodness-of-fit between the experimental data and the model. *bz-rates* is accessible at http://www.lcqb.upmc.fr/bzrates.

KEYWORDS
Mutation-rate; Fluctuation-Assay; Luria–Delbrück

A classical approach to calculate mutation rates (μ) in microorganisms consists in performing fluctuation analyses through multiple cultures grown in parallel under identical conditions (Luria and Delbrück 1943). Each individual culture is started with a small inoculum such that the mutational events that occur during the culture are independent. Cultures are then plated on selective media to determine the number of mutants present in each culture. Estimating the mutation rate from these experimental data is of great interest for biologists and has been the object of many mathematical developments (for review see (Foster 2006)).

Calculating mutation rates requires to first estimate the the mean number of mutations per culture (m) under the assumptions of a Luria-Delbrück distribution model (Lea and Coulson 1949). Once a value of m has been calculated, the mutation rate μ can be easily inferred by dividing m by the total number of cells in the culture (although this can lead to an underestimation of the mutation rate (Ycart and Veziris 2014)). Most of the available estimators rely on the Maximum Likelihood (ML) method which was shown to be accurate for recovering m values (Zheng 2002; Stewart 1994; Jaeger and Sarkar 1995; Sarkar *et al.* 1992; Jones 1994; Gerrish 2008). However, ML estimators can become unstable for fluctuation assays involving cultures with large numbers of mutants. In such cases, the empirical probability GF remains robust and must be preferred over ML (Hamon and Ycart 2012).

One major parameter affecting the estimation of m is b, the differential fitness between mutant and wild type cells (i.e. the ratio between the mutant and the WT growth rates). In the case of differential growth rate, several estimators that jointly calculate m and b have long been made available (Koch 1982; Jones 1994; Jaeger and Sarkar 1995; Zheng 2002, 2005; Hamon and Ycart 2012). The code of these estimators is easily accessible but requires running command lines or installing third party tools.

In addition, the estimation of m can also be affected by another parameter: the plating efficiency, z. This criteria is defined as the fraction of the cultures that is plated on selective media. This parameter accounts for the fact that not all mutants are experimentally detected when only a fraction of the cultures is plated.

Here we propose a new integrated web-tool called *bz-rates* which provides three useful advances over the only web-tool available to estimate m (Hall *et al.* 2009). First, it allows taking into account b, the differential growth rate between mutant and wild-type cells, in the estimation of m with the GF. Note that *bz-rates* does not propose new mathematical developments but fully relies on the available GF estimator. Secondly, *bz-rates* allows the user to take into account the z deviation in the estimation of m by using the formulation suggested in (Foster 2006) and initially proposed by Stewart and collaborators (Stewart *et al.* 1990). Note that more

Manuscript compiled: Wednesday 2nd September, 2015%
[1]15 rue de l'Ecole de Médecine, UMR7238 Biologie Computationnelle et Quantitative, F-75005, Paris, France, gilles.fischer@upmc.fr

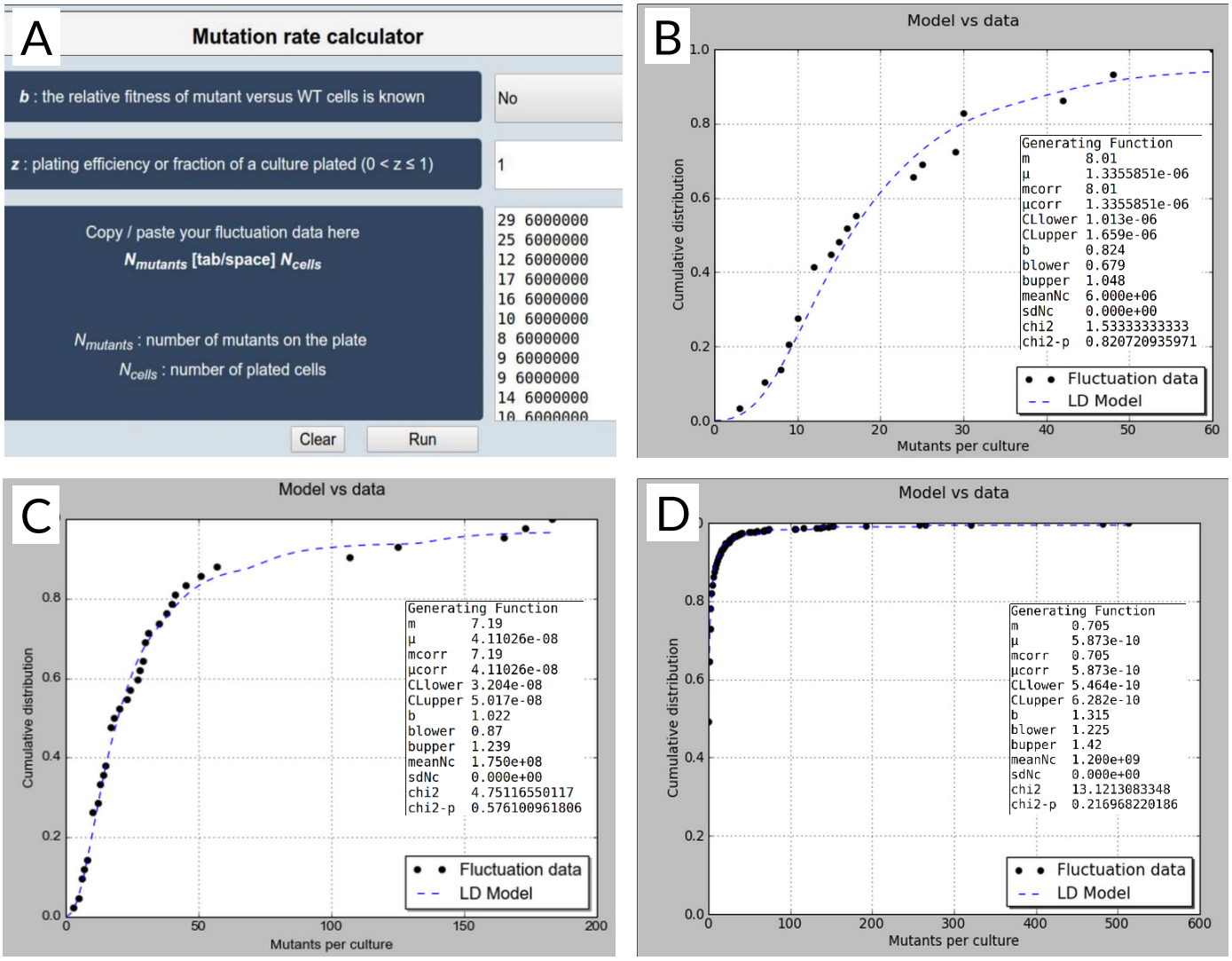


Figure 1 Screen-shots of the *bz-rates* web-site. (A) The input form is composed of 1 choice-field (for the b parameter) and 2 boxes (for the z parameter and a 2-columned data box (N_{mutant} N_{cells})). If the user chooses to manually specify a value for b, a supplementary box appears below the choice field. The z parameter is the plating efficiency which represents the fraction of a culture plated. The $N_{mutants}$ and N_{cells} box is intended to enter the number of plated mutants and plated cells in each culture, respectively. $N_{mutants}$ and N_{cells} must be spaced by a single white-space or a tabulation. Here, the $N_{mutants}$ and N_{cells} box is filled with the values from our experimental fluctuation assay described in the result section. (B to D) Each result section is composed of a numerical box (inside the plot) and a plot showing the cumulative distribution function fitted to the experimental data (B: results from our experimental fluctuation assay, (C) Results from a Luria and Delbrück fluctuation analysis of mutations conferring virus resistance in bacteria (corresponding to the pool of experiments number 1, 10, 11, 15 and 21 from table 2 in (Luria and Delbrück 1943)) (D) Results from a fluctuation experiment of mutations conferring nalidixic acid resistance in Escherichia Coli from Boe *et al.*.

recent formal mathematical treatments to this problem are also available but were not implemented here (Stewart 1991; Jones 1993; Zheng 2008a). Finally, *bz-rates* computes the goodness-of-fit – as described in (Boe *et al.* 1994) – between the experimental data and the two-parameter Luria-Delbrück model and provides the user with a graphical visualization of the fit.

METHODS

bz-rates code

bz-rates was developed in Python with the Django v1.6 framework. It is a free web-tool distributed under the terms of the GNU General Public License. *bz-rates* is accessible at http://www.lcqb.upmc.fr/bzrates. The source code, available at https://github.com/gillet/bzrates, can be easily modified to implement other estimators and clones of the tool can be set up elsewhere.

Fluctuation assay

Fluctuation assays were performed using a BY4741 yeast strain (*TRP1*Δ*5′(1-362)::natNT2*, *CYC1*Δ*::TRP1*Δ*3′*(864-958)*-hph*, *ura3*, *clb5*Δ*::KanMX4*) carrying 2 non-functional alleles of the *TRP1* gene involved in tryptophan biosynthesis on two different chromosomes. One copy is truncated in 3′ and the other copy in 5′ leaving a 400bp homology region repeated in the two alleles. A non allelic homologous recombination event between the 2 hetero-alleles generates a reciprocal translocation that restores tryptophan prototrophy. These mutant cells can therefore be easily selected by plating the cultures onto standard complete synthetic media depleted for tryptophan (CSM-TRP). Briefly, 30 parallel cultures (500μL) were started by inoculating into rich media (Broth Yeast Extract-Peptone-Dextrose) ∼100 cells per well in a 2mL deepwell plate. Cells were grown without agitation at 30°C until they reached an optical density of 0.85 (6.10^6 cells/mL) and plates were incubated for 4 days at 30°C before counting the number of mutants per plate.

Growth rates

The growth rate of 3 independent mutants and wild type cells was measured by doing growth curve experiments in 100 μL of rich media with a Tecan Sunrise robot in triplicates.

RESULTS

Implementation

bz-rates uses the empirical probability GF estimator from (Hamon and Ycart 2012; Ycart 2013). This method allows a precise estimation of m across a larger range of parameter values than the ML method. The cellular division time model chosen in *bz-rates* is not the classical exponential model but a constant division time model ('Dirac'). Although there is no universal cellular division time model as it depends on experimental conditions like the strain or the media, the 'Dirac' model is usually the most accurate for the estimation of b and as accurate as the exponential model for the estimation of m. Note that this division time model induces a positive bias in the estimation of large values of m (Ycart 2013).

When b is known, the value provided by the user is used to estimate m with the GF. However, when the mutant relative fitness b is not known, *bz-rates* estimates both m and b with the GF function.

The m_{corr} value that takes into consideration the plating efficiency z is calculated according to formula (41) in (Stewart *et al.* 1990): $m_{corr} = m \cdot (z-1)/(z \cdot ln(z))$. μ is defined as $m/\bar{N}p$ and μ_{corr} as $m_{corr}/\bar{N}t$ where $\bar{N}p$ and $\bar{N}t$ represent the mean number of cells per plate and per culture respectively. CL_{lower} and CL_{upper} provide the lower and upper confidence limits of μ_{corr} (level of confidence = 95%). σ_{Np} provides the standard deviation of the number plated cells.

To test the goodness-of-fit of the data to the model, *bz-rates* performs a Pearson's chi-squared test. The value of χ^2 gives the Pearson's chi-squared goodness of fit and the $\chi^2 - pval$ its associated p-value. The null hypothesis is rejected in the case $\chi^2 - pval < 0.01$ meaning that the cumulative distribution function does not fit with the experimental data (empirical cumulative distribution function). In this case, the user is warned that the estimation of the mutation rate is not reliable.

Interface

bz-rates is composed of a simple form (Fig. 1A). The first choice field provides the user with the possibility to indicate that b is known. In this case, the b field appears and the experimentally determined value of b can be filled in ($0 < b < \infty$). Otherwise b will be estimated computationally by the GF.

The second box allows to fill in the plating efficiency z (i.e. the proportion of cells from each culture that was plated, default value: $z = 1$).

The main field is the 'N_{mutant} N_{cells}' box that parses the fluctuation analysis counts. $N_{mutants}$ and N_{cells} are the number of plated mutants and plated cells per culture, respectively. This field is 'excel ready' thus counts can be directly copy/pasted into this box without further formatting.

The *bz-rates* result section is composed of two parts: the numerical and the graphical boxes (Fig. 1). The numerical box on the left provides the following estimates:

- m: mean number of mutations per culture not corrected by the plating efficiency (z)
- μ: mutation rate per cell per division not corrected by the plating efficiency (z)
- m_{corr}: number of mutations per culture corrected by the plating efficiency (z)
- μ_{corr}: mutation rate per cell per division corrected by the plating efficiency (z)
- CL_{lower}: lower 95% confidence limit for m_{corr}
- CL_{upper}: upper 95% confidence limit for m_{corr}
- b: mutant cells relative fitness predicted by the Generating Function (only output if b is left empty in the input field)
- b_{lower}: lower 95% confidence limit for b
- b_{upper}: upper 95% confidence limit for b
- $\overline{Nc}$: average number of plated cells per culture
- σ_{Nc}: standard deviation of the number of plated cells
- χ^2: Pearson's chi-square value
- $\chi^2 - p$: Pearson's chi-square p-value

The graphical box plots the cumulative distribution function fitted to the experimental data. It allows the user to visually judge for the correctness of the hypothesized distribution. To quantify the quality of the fit, *bz-rates* performs a Pearson's chi-squared goodness of fit as described in (Boe *et al.* 1994). If the null hypothesis is rejected (p-value<0.01), the user is advised by a red warning that the predicted and observed distributions are not in close agreement. In this case, the user should consider using another model that takes into consideration other deviations from the Luria Delbrück model such as, for instance, the post-plating growth (Lang and Murray 2008). To do so, the user should use an advanced mutation rate calculation packages to explore different models such as Salvador (Zheng 2008b).

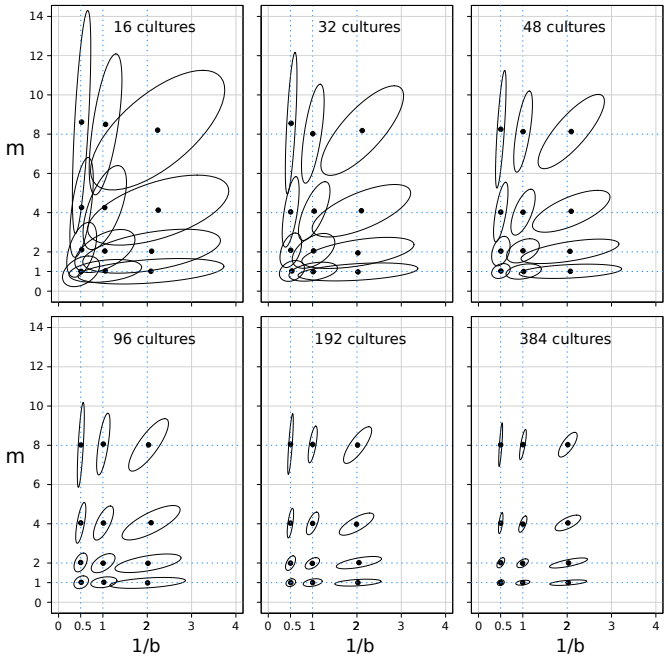


Figure 2 Performance of the *bz-rates* calculator on various simulated datasets. Each panel corresponds to simulated fluctuation datasets with either 16, 32, 48, 96, 192 or 384 independent cultures. In each panel, 200 simulations were performed for different values of m (1, 2, 4 and 8) and b (0.5, 1 and 2). The ellipses show the 95% dispersion of *bz-rates* estimations for the 200 simulations.

Experimental testing

A fluctuation assay was performed with a yeast strain carrying a genetic system that is designed to generate a functional copy of an auxotrophic gene when the cells undergo a specific chromosomal rearrangement (a reciprocal translocation, see methods). The resulting mutant cells have a strong growth defect relatively to wild type cells, probably as the result of the translocation, that was experimentally measured to 0.76 (see methods).

The form of Fig. 1 (A) is filled with the data of the 30 tubes fluctuation assay that was undertaken. We neglected to specify the mutants relative growth rate in order to compare the predicted relative growth rate of *bz-rates* to the experimental measure. Fig. 1 (B) shows *bz-rates* results. The plot indicates a good fit between the statistical distribution of the mutants and the experimental data. Pearson's chi-square goodness of fit value (1.16) and p-value (0.88) are displayed at the end of the numerical box on the left. The mutation rate (μ) is estimated to $1.33.10^{-6}$ per cell per division (95% confidence limits (CL) [$1.01.10^{-6}$ - $1.66.10^{-6}$]) and the predicted mutant relative growth rate (0.82 [0.68 - 1.05]) is in close agreement with the experimental measure (0.76).

Published datasets and simulations

In order to test our implementation and the stability of the GF estimator in *bz-rates*, we tested 2 published datasets: (i) the first dataset corresponds to a historical fluctuation assay composed of 42 parallel cultures performed by Luria and Delbrück (Fig. 1 (C)). With this dataset, *bz-rates* predicts a mutation rate ($4.11.10^{-8}$) close to the one calculated by (Luria and Delbrück 1943) ($2.48.10^{-8}$). The value of m (7.19) and b (1.022) are very close to the range of values reported in (Hamon and Ycart 2012) ([5.22-8.89] for m and [0.74-1.22] for b). We also tested one larger fluctuation dataset from (Boe *et al.* 1994) that is composed of 1102 cultures (Fig. 1 (D)). In this case, *bz-rates* reports a m value of 0.705 which is close to the one calculated in (Hamon and Ycart 2012) ([0.65-0.77]) and the one calculated in (Zheng 2005) (0.71). The mutant differential growth rate value is 1.315 which is a bit higher than the one reported by (Zheng 2005) with a maximum likelihood approach (b=1.193).

The performance of *bz-rates* was also tested on simulated datasets. We generated simulated fluctuation assays for different couples of m and b with either 16, 32, 48, 96, 192 or 384 parallel cultures (Fig. 2). As expected, the precision of the estimator increases with increasing numbers of parallel cultures. The general trend that can be inferred from these plots is that the precision on the estimation of m (and by consequence the estimation of μ) is higher for the smallest values of m. Therefore, users should not outgrow the cultures in order to limit the number of mutants that grow on selective plates.

Note that the GF estimator has also been extensively tested elsewhere (Hamon and Ycart 2012; Ycart 2013) and the reader should refer to these papers for an extensive review of the performance of this estimator.

CONCLUSION

bz-rates is a web-tool that does not require the installation of any third party tool or run any command line to estimate mutation rates. It has a minimalist design in order to provide biologists with a web-tool the most straightforward as possible. To our knowledge there was so far a single web-tool available for mutation rate calculation (Hall *et al.* 2009) but this tool does not allow to consider deviations from Luria Delbrück or to estimate the goodness of fit with the model. Therefore, *bz-rates* provides useful advances such as accounting for 2 important deviations to Luria-Delbrück distributions (b and z) as well as giving an indication of the reliability of the estimated mutation rates. We hope that *bz-rates* will reveal useful to a broad community of microbiologists and geneticists.

ACKNOWLEDGEMENT

We thank our colleagues from LCQB for fruitful discussions and particularly Nicolas Agier for invaluable tips and advice. This work was supported by a scholarship from "La Ligue contre le cancer" and by a Convergence grant (Memory) from Idex SUPER Sorbonne Université 2014.

LITERATURE CITED

Boe, L., T. Tolker-Nielsen, K. M. Eegholm, H. Spliid, and a. Vrang, 1994 Fluctuation analysis of mutations to nalidixic acid resistance in Escherichia coli. Journal of Bacteriology **176**: 2781–2787.

Foster, P. L., 2006 Methods for determining spontaneous mutation rates. Methods in enzymology **409**: 195–213.

Gerrish, P. J., 2008 A Simple Formula for Obtaining Markedly Improved Mutation Rate Estimates. Genetics **180**: 1773–1778.

Hall, B. M., C.-X. Ma, P. Liang, and K. K. Singh, 2009 Fluctuation analysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. Bioinformatics (Oxford, England) **25**: 1564–5.

Hamon, A. and B. Ycart, 2012 Statistics for the Luria-Delbrück distribution. Electronic Journal of Statistics **6**: 1251–1272.

Jaeger, G. and S. Sarkar, 1995 On the distribution of bacterial mutants: the effects of differential fitness of mutants and non-mutants. Genetica pp. 217–223.

Jones, M. E., 1993 Accounting for plating efficiency when estimating spontaneous mutation rates. Mutation Research/Environmental Mutagenesis and Related Subjects **292**: 187–189.

Jones, M. E., 1994 LB Fluctuation Experiements; Accounting Simultaneously for Plating Efficiency and Differential Growth Rate. J. theor. Biol **166**: 355–363.

Koch, A. L., 1982 Mutation and growth rates from Luria-Delbrtick fluctuation tests. Mutation Research **95**: 129–143.

Lang, G. I. and A. W. Murray, 2008 Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. Genetics **178**: 67–82.

Lea, D. and C. A. Coulson, 1949 The distribution of the numbers of mutants in bacterial populations. Journal of genetics **49**: 264–285.

Luria, E. and M. Delbrück, 1943 Mutations of bacteria from virus sensitivity to virus resistance'-'. Genetics **28**: 491–511.

Sarkar, S., Ma, and Sandri, 1992 On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. Genetica pp. 173–179.

Stewart, F. M., 1991 Fluctuation analysis: the effect of plating efficiency. Genetica **84**: 51–55.

Stewart, F. M., 1994 Tests: How Reliable Are the Estimates of Mutation Rates ? Genetics **1146**: 1139–1146.

Stewart, F. M., D. M. Gordon, and B. R. Levin, 1990 Fluctuation Analysis: The Probability Distribution of the Number Mutants Under Different Conditions. Genetics **124**: 175–185.

Ycart, B., 2013 Fluctuation analysis: can estimates be trusted? PloS one **8**: e80958.

Ycart, B. and N. Veziris, 2014 Unbiased estimation of mutation rates under fluctuating final counts. PLoS ONE **9**.

Zheng, Q., 2002 Statistical and algorithmic methods for fluctuation analysis with SALVADOR as an implementation. Mathematical Biosciences **176**: 237–252.

Zheng, Q., 2005 New algorithms for Luria-Delbrück fluctuation analysis. Mathematical biosciences **196**: 198–214.

Zheng, Q., 2008a A note on plating efficiency in fluctuation experiments. Mathematical Biosciences **216**: 150–153.

Zheng, Q., 2008b SALVADOR 2.3: A tool for studying mutation rates.

3.2 - Approche de génétique inverse pour identifier des gènes impliqués dans la formation des inversions, des translocations réciproques et des duplications

Nous avons vu dans la seconde partie de l'introduction que de nombreux mécanismes peuvent conduire à la formation de SV. Ces derniers se distinguent notamment par l'utilisation ou non de la recombinaison homologue et donc aussi par la longueur de l'homologie qu'ils nécessitent pour former une SV. D'après la partie 2 des résultats, il semble que dans des cellules de levures haploïdes, les SV se forment principalement via des mécanismes dépendants de la recombinaison homologue chez la levure haploïde. C'est notamment pour cette raison que les 3 systèmes génétiques de mesures des INV, RT et DUP, que nous avons développés afin d'analyser indépendamment la dynamique des SV, reposent sur la sélection d'évènements de recombinaison entre des hétéroallèles partageant 400 pb d'homologie. Cependant, nous avons également montré dans la partie 2 des résultats qu'environ 30% des SV détectées n'utilisent pas la recombinaison homologue puisqu'elles ne présentent pas de grandes homologies à leurs jonctions. De plus, notre laboratoire a précédemment montré que de grandes duplications segmentales peuvent se produire sur le bras droit du chromosome 15 par 2 mécanismes distincts : le BIR et le MMIR (Koszul et al., 2004; Payen et al., 2008). Ces 2 mécanismes se distinguent essentiellement par l'homologie qu'ils utilisent. Le BIR nécessite de l'homologie alors que le MMIR ne nécessite que de la microhomologie. Ainsi, bien que ces 2 mécanismes puissent aboutir à la formation de grandes duplications segmentales, les déterminants génétiques en jeu sont très différents : le BIR est rad52 (RH) dépendent alors que le MMIR est indépendant à la fois de la RH, du NHEJ et du MMEJ.

Afin de caractériser les gènes impliqués dans la formation de différents types de SV, j'ai développé 3 versions différentes de chacun des 3 systèmes (DUP, INV et RT). Dans ces 3 versions les hétéroallèles partagent 5, 58 ou 411 pb d'homologie. Pour faciliter la lecture, je donne un nom à ces système d'après le type de SV qu'ils mesurent (DUP, INV ou RT) et la longueur d'homologie qu'ils contiennent (par exemple : DUP5 fait référence au système de mesure des duplications segmentales avec 5 pb d'homologie entre les hétéroallèles). Les systèmes utilisant 5 pb d'homologie ont été créés afin d'étudier des gènes impliqués dans la formation des INV, DUP ou RT via de la microhomologie. En revanche, les systèmes utilisant 58 ou 401 pb permettent de sélectionner des SV qui se forment grâce à la RH. En effet la longueur minimale requise par la recombinaison homologue chez la levure est ~50 pb et la longueur à partir de laquelle la fréquence de recombinaison n'augmente plus exponentiellement avec l'augmentation de l'homologie est ~400pb. Les 3 systèmes sont

représentés à la figure 4 de notre article n°2 page 131 « A granular view of subclonal SV landscape in the yeast genome» que je donne dans la partie 2 des résultats ainsi qu'en annexe 3.

Notons que la longueur de l'homologie entre hétéroallèles a un impact fort sur la fréquence de formation *de novo* des SV. Ainsi, chaque système ne s'utilise pas dans les mêmes conditions. Comme nous l'avons vu dans le matériel et méthode de notre article de la partie 2 des résultats, les tests de fluctuation peuvent être réalisés en goutte pour les systèmes partageants 400 pb d'homologie. L'avantage des tests de fluctuation en gouttes par rapport à un test de fluctuation où l'on étale chaque culture sur une boîte de Petri complète est d'augmenter considérablement (16 fois) le débit des expériences puisque 16 gouttes peuvent être déposées sur une seule boîte de pétri carrée. La mesure des DUP avec les systèmes contenant 58 pb d'homologie peut également être réalisée par des tests en goutte. En revanche, des tests préliminaires ont montré que les étalements en goutte ne conviennent pas pour l'étude des systèmes INV58 et RT58. En effet, nous avons calculé que les fréquences de formation *de novo* à partir de INV58 et RT58 sont de l'ordre de 10^{-8} à 10^{-9} ce qui signifie qu'il faut étaler au moins 10^{9} cellules. Ce nombre est bien supérieur au nombre maximal de cellules que l'on peut étaler dans une goutte (~10^{7} cellules) et sur une boite de Petri (~10^{7} – 10^{8} cellules). De la même façon, les fréquences obtenues avec les systèmes contenant seulement 5 pb d'homologie sont de l'ordre de 10^{-8} à 10^{-10}. Ces systèmes ne peuvent donc pas être utilisés lors de tests en goutte.

Nous avons choisi d'utiliser ces systèmes génétiques pour étudier l'effet de mutants des voies de réparation de l'ADN. Le choix de ces gènes s'est appuyés sur le travail réalisé par Stirling et collègues (Stirling et al., 2011) dans lequel ils ont réalisé un crible chez la levure visant à identifier des gènes impliqués dans la stabilité des chromosomes. Ils ont de plus compilé leurs données avec celles d'autres cribles de l'instabilité chromosomique issus de la littérature. Ils ont ainsi obtenu un catalogue de 692 gènes impliqués dans la stabilité du génome. Brièvement, leur crible a porté sur 3 critères : le niveau de GCR, d'ALF (A-like Faker) et de CTF (Chromosome Transmission Fidelity). Le système GCR a déjà été décrit à la page 75. L'ALF consiste en fait à analyser la fréquence à laquelle des cellules haploïdes perdent leur caractère sexuel « a » ou « alpha » ce qui leur permet d'être croisées avec une souche du même type sexuel. Le CTF permet de suivre la transmission d'un chromosome surnuméraire via l'accumulation de pigments rouges dans la cellule en cas de perte du chromosome. On peut noter que ces 3 systèmes ne permettent pas de tester la formation de type de SV bien définis tels que les DUP, INV et RT. Pour notre étude, j'ai donc sélectionné dans cette liste de 692

gènes 27 dont les mutants présentent des phénotypes forts en GCR, ALF et CTF. La fonction de la majorité de ces gènes – tous impliqués soit dans la réparation de l'ADN, le point de contrôle de la phase S ou la réplication - a été évoquée dans la partie 2 de l'introduction. Un tableau descriptif de la fonction de chacun de ces 27 gènes est disponible en annexe 4 et tableau 1.

La première grande étape de ce projet a été de générer les 27 mutants. J'ai choisi d'utiliser la stratégie la plus classique chez la levure qui consiste à remplacer le *locus* de chacun des 27 gènes cibles par une cassette de résistance à une drogue comme la kanamycine ou l'hygromycine. Ces délétions ont été réalisées dans des souches de *S. cerevisiae* contenant déjà le système de mesure des duplications segmentales DUP58. Cette stratégie m'a permis de générer avec succès 25 des 27 mutants que nous souhaitions obtenir (tableau 1). L'étape suivante a été de transférer chacune de ces mutations dans les souches contenant les autres systèmes génétiques de mesure des SV. J'ai donc utilisé les mutants de la souche DUP58 comme matrice de PCR pour synthétiser des amplicons. Ceci m'a permis d'ailleurs de synthétiser des amplicons plus longs afin d'optimiser l'efficacité de transformation des autres souches. Ainsi, jusqu'à maintenant, j'ai pu générer avec succès les 25 mutations dans des souches contenant les systèmes INV411, DUP411 et RT411 (Tableau 1).

Jusqu'ici nous avons réalisé des tests préliminaires les mutants de la souche DUP58 (Figure 29). Ces tests ne correspondent pas à des tests de fluctuation mais à des tests de fréquences de mutants dans 4 cultures parallèles par mutant. Comme attendu, les mutations *mec1Δ*, *dun1Δ*, *rad27Δ* et *sgs1Δ* conduisent à une forte augmentation des fréquences de duplications par rapport au sauvage avec des augmentations allant respectivement de 8 à 12 fois. Mec1p est une protéine majeure du déclenchement du point de contrôle en phase S et est donc une protéine très importante pour l'intégrité du génome (Weinert et al., 1994). Dun1 est également un gène mutateur connu : c'est une sérine-thréonine kinase nécessaire pour l'activation transcriptionnelle de gènes de la réparation de l'ADN (Zhou and Elledge, 1993). Nous avons déjà évoqué à plusieurs reprises le rôle de Rad27p (qui code une flap-endonuclease) dans la maturation des fragments d'Okazaki et dont la non-dégradation pourrait conduire à des cassures de l'ADN au niveau des flaps (Tishkoff et al., 1997). Enfin, Sgs1 est une hélicase qui inhibe la formation des crossing-over et qui supprime les évènements de recombinaison illégitimes (Watt et al., 1996).

Tableau 1: Mutants construits dans les souches contenant le système de mesure des duplications segmentales DUP58, la souche comportant les systèmes DUP411 et RT411 ainsi que la souche avec le système INV411. Les noms de souche sur fond vert correspondent aux souches construites et les rouges sont celles qui restent à construire.

MutantΔ	Mutant ID	DUP58	DUP411+TR411	INV411
WT	wt	YKFB614	YAG142	YAG152
clb5	1	YAG7	YAG159	YAG184
fun30	2	YAG10	YAG160	YAG185
wss1	3	YAG11	YAG161	YAG186
sgs1	4	YAG12	YAG162	YAG187
rad5	5	YAG13	YAG163	YAG188
rad27	6	YAG14	YAG164	YAG189
rad18	7	YAG15	YAG165	YAG190
elg1	8	YAG16	YAG166	YAG191
chk1	9	YAG42		YAG192
rrm3	13	YAG48	YAG168	YAG193
srs2	14	YAG49	YAG169	YAG194
rad17	15	YAG50	YAG170	YAG195
sml1	16	YAG51	YAG171	YAG196
sml1, mec1	19	YAG54		
rad50	17	YAG52	YAG172	YAG197
pif1	20	YAG55	YAG173	YAG198
ku70	21	YAG56	YAG174	YAG199
dun1	22	YAG57	YAG175	YAG200
mus81	23	YAG67	YAG176	YAG201
yen1	24	YAG70	YAG177	YAG202
pol32	26	YAG112	YAG178	YAG203
rmi1	27	YAG65		YAG204
slx4	28	YAG129	YAG180	YAG205
cac3	29		YAG181	YAG206
rad52	11	YAG45	YAG182	YAG207
lig4	30			YAG208
rad24	18	YAG53	YAG219	YAG220

Deux autres mutations conduisent à une forte réduction de la fréquence de DUP (~5 fois) par rapport au sauvage : *rad52Δ* et *mrc1Δ*. Cependant, en l'absence de RH dans le mutant *rad52Δ*, on s'attend à ne plus voir de DUP alors que ce mutant ne montre qu'une diminution de la fréquence de DUP d'un facteur 5. Ces tests préliminaires n'ont toutefois été effectués qu'avec un faible nombre de réplicats sans prendre en compte la croissance différentielle des cellules recombinantes et devront être reproduits. En ce qui concerne le mutant mrc1Δ, nos résultats suggèrent une éventuelle stabilisation du génome provoquée par cette mutation. Ce

résultat est également surprenant : ce gène est un co-activateur de Rad53 (checkpoint de phase S) (Alcasabas et al., 2001) et supprime les GCR (Putnam et al., 2009). Cependant, Rad53 est à la fois un répresseur de la RH et un activateur des gènes de la réparation de l'ADN. Il sera donc intéressant de voir l'effet de ce mutant dans les systèmes de mesure des INV et des RT ainsi que dans un contexte de microhomologie pour voir si la protéine Mrc1p promeut également l'instabilité dans ces cas-là.

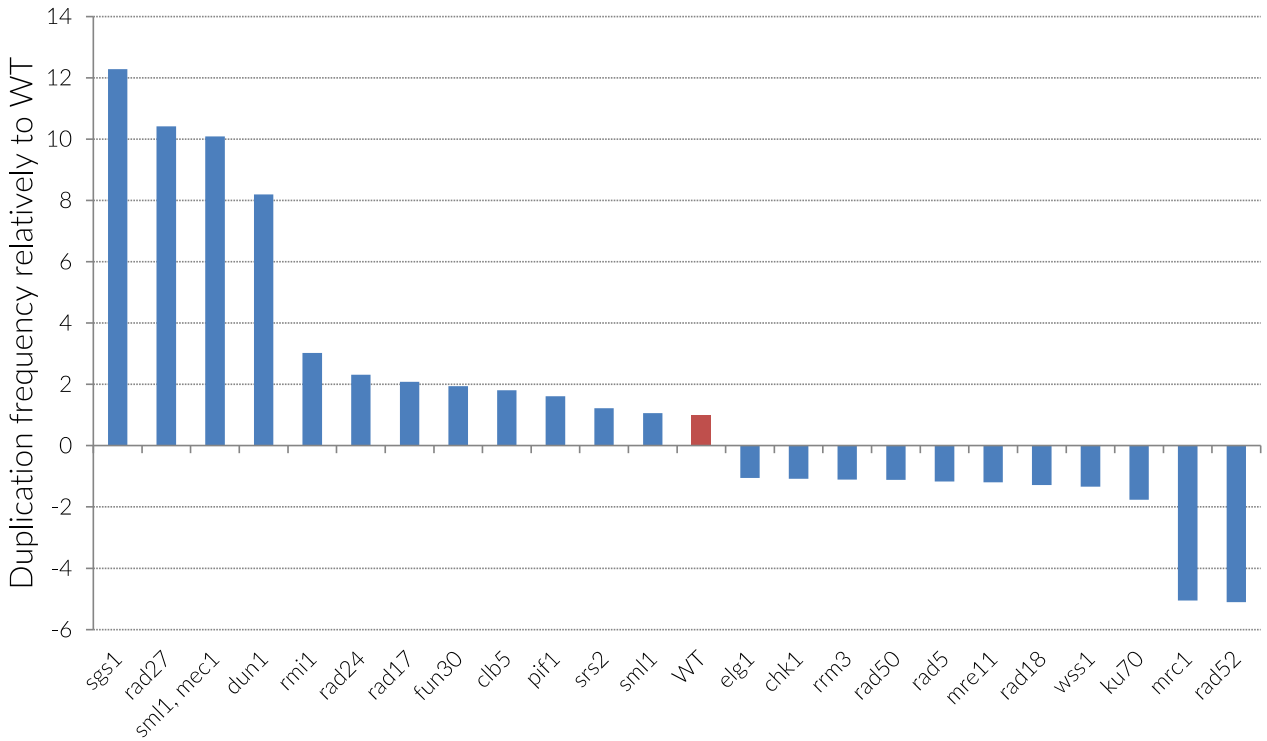


Figure 29: Fréquence de DUP58 dans des fonds génétiques mutants par rapport au sauvage. Pour les mutants dont la fréquence est plus petite que le sauvage, la fréquence tracée est -1/(fréquence par rapport au sauvage).

Dans la suite de ce projet, il sera intéressant d'effectuer tous les tests de fluctuations pour l'ensemble des mutants et des systèmes génétiques de mesures des DUP, INV et RT construits. Ces systèmes génétiques donnent probablement une idée beaucoup plus précise de la dynamique des SV que le système génétique phare de mesure de la fréquence des SV : le GCR. Le système GCR repose en effet sur la co-sélection de la perte de deux marqueurs dans une région subtélomérique. Ces régions ne possédant pas de gènes essentiels, elles ont donc moins de « contraintes » ce qui peut introduire un biais dans les mesures. Ceci expliquerait pourquoi les taux de GCR rapportés dans certains fonds génétiques mutants qui sont plusieurs milliers de fois plus élevés que ceux des cellules sauvages (Chen and Kolodner, 1999).

3.3 - Identification des facteurs génétiques contrôlant la stabilité des génomes:

Nous avons vu dans la partie précédente que la génétique inverse permet d'identifier des gènes impliqués dans le maintien de la stabilité génomique en caractérisant les phénotypes de différents mutants indépendamment les uns des autres. Bien que ce type d'approche permette de caractériser de nombreux gènes liés à l'instabilité génomique, il ne permet pas d'accéder aux relations d'épistasie entre les gènes. Dans cette partie, nous avons donc entrepris une démarche différente pour étudier l'instabilité des chromosomes. Notre idée est ici de considérer que le contrôle de la stabilité du génome est un trait quantitatif complexe contrôlé par un grand nombre de gènes. À l'intérieur d'un fond génétique donné, tous les allèles de ces gènes seraient ainsi co-adaptés pour maintenir le génome stable. Nous voulons donc mettre en évidence les variations de l'instabilité génomique chez différents fonds génétiques naturels de *S. cerevisiae*. Je présente ici les résultats préliminaires de ce projet qui a pour objectif à plus long terme d'essayer de comprendre comment certains génotypes

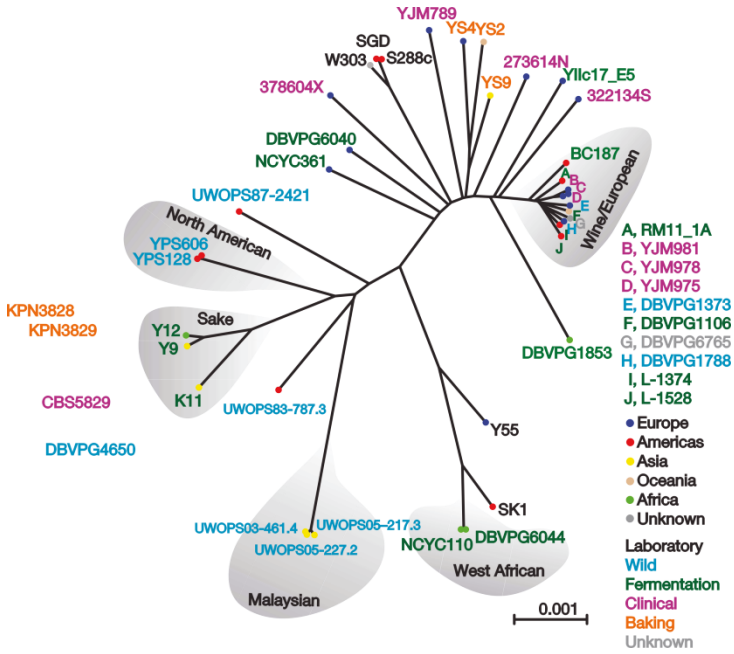


Figure 30: Arbre phylogénétique des souches de *S. cerevisiae*. Les groupes de lignées pures sont entourées en gris. La souche de référence S288c dont est dérivée la souche BY4741 que nous utilisons dans le reste de ce manuscrit se situe en haut de l'arbre. La barre d'échelle indique la fréquence des substitutions nucléotidiques. D'après Liti et al. 2009.

prédisposent à l'instabilité génomique.

Au cours de ce projet, nous avons utilisé 5 lignées de *S. cerevisiae* dites « pures » car elles possèdent des fonds génétiques distincts et représentatifs de la diversité génétique naturelle de l'espèce (Cubillos et al., 2009; Liti et al., 2009). Ainsi, contrairement aux souches

de laboratoires qui sont issues de nombreux croisements, leur génome n'est pas une mosaïque de différentes souches existantes (Figure 30). Ces 5 lignées pures étant des isolats naturels d'origines diverses, nous les distinguerons par la suite par leur origine géographique et/ou leur niche écologique : Amérique du Nord (NA), Afrique de l'Ouest (WA), Europe (WE), Malaisie (MAL) et Asie (SAKE). Le tableau 1 présente également le nom officiel de ces lignées tel qu'elles sont déposées dans les banques. Ces lignées étant sauvages, elles ne sont donc pas directement manipulables avec les outils classiques de génétique de la levure. Pour cela, nous avons construit un ensemble de souches dans lesquelles plusieurs des gènes d'auxotrophie couramment utilisés en laboratoire (*URA3*, *LYS2*, *LEU2*, *MET15*) ont été complètement délétés. Ce travail a fait l'objet d'une publication dans le journal Yeast (Louvel et al., 2013) que j'ai co-signé en second auteur (voir article 4 a l'annexe 5).

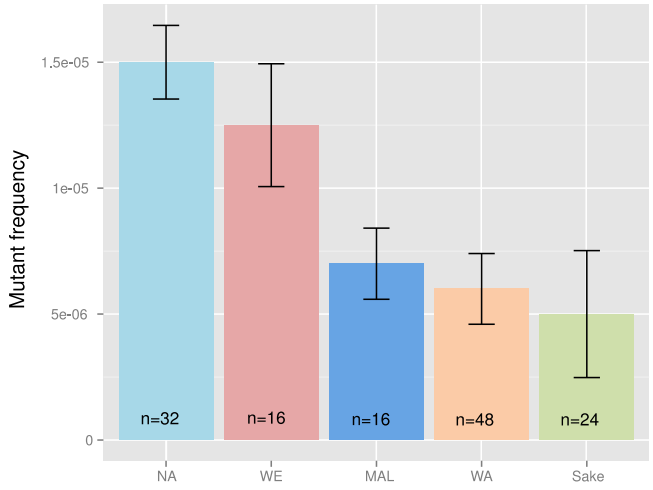


Figure 31: Mesure de la fréquence de duplication segmentale avec le système génétique DUP411 dans les 5 lignées pures. 'n' indique le nombre de cultures réalisées. NA : North American, WE : Wine European, MAL : Malaisienne, WA : West African, Sake : Asie.

Nous avons ensuite entrepris de transférer les systèmes génétiques de mesure des SV dans ces souches modèles. Ainsi, nous avons introduit dans le génome des 5 lignées le système génétique de mesure des taux de DUP411 précédemment développés (cf. partie 2 et 3 des résultats) et nous avons ensuite pu mesurer les fréquences de DUP en fonction des fonds génétiques (figure 31). Ces derniers semblent former 2 groupes : le groupe NA/WE dont les fréquences de DUP sont respectivement 1,5 et 1,25 DUP/10^{-5} cellule et le groupe MAL/WA/Sake avec des fréquences de DUP 2 à 3 fois plus faibles (~5.10^{-6}). Ces résultats suggèrent donc que la fréquence de DUP dans le génome est dépendante du fond génétique. Nous avons ensuite croisé entre elles les 5 souches possédant le système DUP411 (10 croisements possibles). Ceci a permis de produire des descendants possédants tous une combinaison unique des allèles des 2 parents générée au cours de la méiose. Parmi les 10 croisements, 4 impliquent la souche MAL qui contient 8 jonctions interchromosomiques

résultant de plusieurs évènements de translocations réciproques (Marie-Nelly et al., 2014). En effet, les ségrégants issus de ces 4 croisements présentent une viabilité de l'ordre de quelques pour cent et ces souches n'ont pas été retenues pour la suite de ce travail. Pour 4 des 6 hybrides n'impliquant pas la souche MAL, nous avons généré par dissection de tétrades 60 à 70 descendants issus de 15 à 18 tétrades (NAxSAKE, NAxWA, NAxWE et WAxWE). Pour chacun de ces croisements, la fréquence de DUP a été mesurée dans 20 descendants (Figure 32). Nous avons constaté que le niveau d'instabilité varie fortement entre les descendants (jusqu'à 200 fois plus de duplications d'un descendant à l'autre). Les descendants présentent en effet tous les phénotypes intermédiaires qu'il peut exister entre les 2 parents ainsi que des cas de transgression forts. Ces cas correspondent à des descendants ayant un phénotype plus fort ou plus faible que les parents. Or si la fréquence de DUP était contrôlée par un locus unique, on s'attendrait à n'observer que 2 classes de fréquences de DUP correspondant aux fréquences parentales. Ces résultats préliminaires suggèrent donc que des polymorphismes génétiques naturellement présents dans les populations de levures contrôlent le niveau d'instabilité intrinsèque des génomes. Le continuum de fréquences de DUP observé dans la descendance suggère lui que de nombreux QTL sont impliqués dans le maintien de la stabilité du génome.

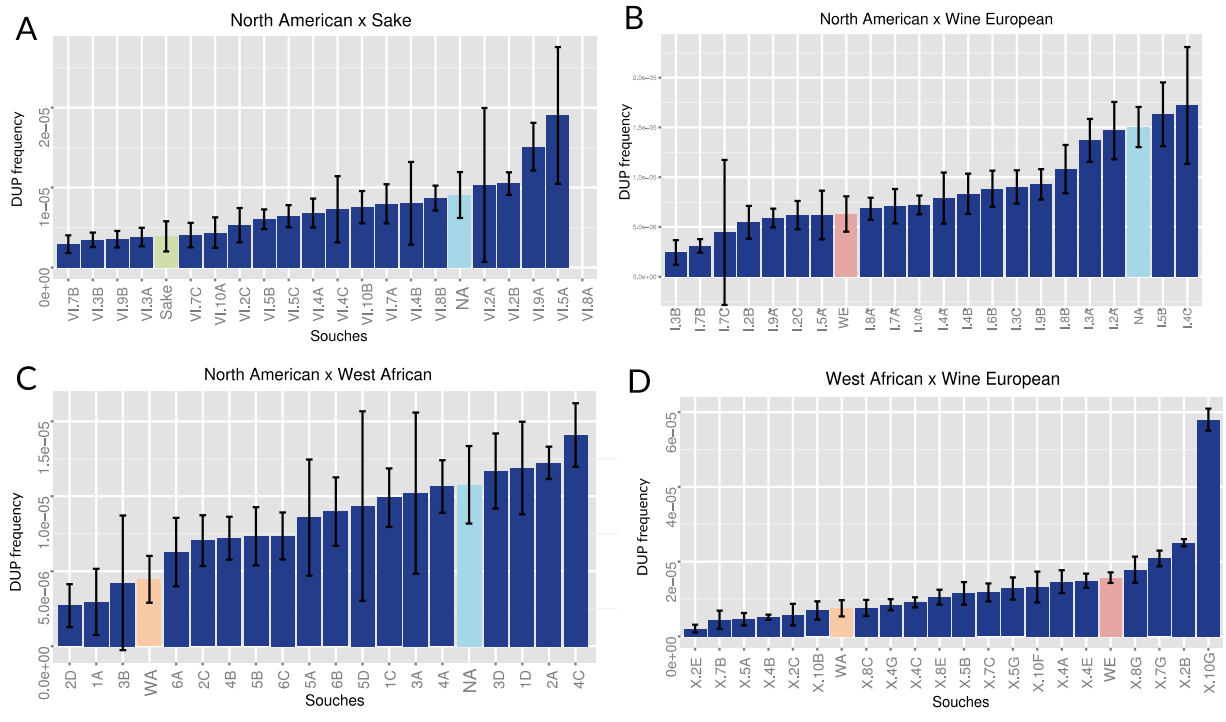


Figure 32: Fréquence de duplication segmentale dans la descendance de croisements des 5 lignées pures. Les barres colorées correspondent aux parents et les barres bleus aux ségrégants. Les barres d'erreurs correspondent à l'erreur standard.

La suite de ce projet consistera donc à identifier les déterminants génétiques responsables de l'instabilité chromosomique. Cette identification reposera sur l'analyse en masse d'un mélange de milliers de descendants. Nous faisons l'hypothèse que les descendants porteurs des fonds génétiques conférant la plus forte instabilité vont générer des duplications à plus haute fréquence et vont de ce fait envahir la culture grâce à l'avantage sélectif conféré par la duplication. Une analyse de liaison à grande échelle, consistant à séquencer des mélanges de descendants puis à rechercher des déviations significatives des fréquences alléliques nous permettra de cartographier et d'identifier précisément les polymorphismes responsables de l'instabilité. À terme, l'objectif de cette analyse est de trouver de nouveaux QTL candidats pour l'instabilité chromosomique.

L'idée de cartographier par analyse de liaison les gènes responsables de certains aspects de la stabilité du génome n'est cependant pas nouvelle chez la levure. En étudiant des populations naturelles de levures, Gatbonton et collègues ont ainsi identifié avec succès des gènes dont dépendrait en partie la variation de la longueur des télomères (Gatbonton et al., 2006). De façon analogue, Dimitrov et collègues ont identifié plusieurs allèles qui seraient responsables en quasi-totalité des différences de fréquence d'apparition des mutants *petites* qu'il existe entre différents fonds génétiques de levures (Dimitrov et al., 2009). Notre projet a lui permis d'analyser la fréquence de duplication dans la descendance d'individus hybrides. Nous avons montré que cette fréquence est probablement contrôlée par de nombreux QTL. Si l'on considère la fréquence de DUP comme un trait phénotypique alors nous pouvons envisager de cartographier les QTL associés à ce phénotype par analyse de liaison ou analyse de ségrégants en masse (X-QTL). Cette approche est potentiellement très prometteuse pour la quantification des QTL des SV. Elle devrait en effet pouvoir compléter les nombreuses approches de génétique inverse qui ont identifié des centaines de gènes de la stabilité des chromosomes mais dont les relations d'épistasie restent très peu décrites.

CONCLUSIONS ET PERSPECTIVES

Conclusions et perspectives

Au cours de ces 4 années de thèse, j'ai travaillé à l'étude d'une question dont l'etude par des généticiens comme Calvin B. Bridges, HJ Muller ou Theodosius Dobjansky remonte probablement à plus de 80 ans : quel est le niveau de plasticité des génomes eucaryotes et pourquoi le sont-ils ? Alors qu'il y a 80 ans les outils pour l'étude de la plasticité des chromosomes à la disposition des généticiens étaient peu nombreux, les possibilités pour adresser cette question sont aujourd'hui énormes. Durant ce travail de thèse, nous avons ainsi appréhendé le problème de l'instabilité chromosomique sous plusieurs angles chez la levure : de la génétique inverse à la génétique *forward* en passant par le développement d'une approche prometteuse des SV. Cette dernière permet de détecter des SV subclonales (SSV) ce qui fournit une nouvelle alternative aux MAL pour analyser les SV *de novo* dans les génomes. Cette approche SSV se caractérise premièrement par sa polyvalence comparée aux MAL. En effet, d'un point de vue pratique, les MAL nécessitent de cultiver des cellules pendant environ 2500 générations sans garantie de visualiser des SV à la fin. Or, compte tenu de la durée du cycle cellulaire chez la levure, ces 2500 générations correspondent à 1 an d'expériences incompressibles. Bien qu'il soit envisageable de réaliser plus que 2500 générations, les contraintes de temps ne permettent pas d'imaginer pouvoir augmenter ce nombre au-delà de quelques milliers. En revanche, l'approche SSV appliquée chez la levure permet de s'affranchir de cette dimension temporelle puisqu'elle nécessite seulement la réalisation d'une culture monoclonale. La polyvalence de l'approche SSV est également visible par sa sensibilité. Dans le cas des MAL, la sensibilité dépend du nombre de générations alors que pour les expériences de SSV elle est majoritairement guidée par la profondeur de séquençage. Compte tenu des progrès rapides des technologies de séquençage, il ne fait pas de doute que le coût d'obtention de couvertures physiques très importantes est amené à continuer de diminuer fortement dans le futur. Il est cependant peu probable que l'approche SSV puisse bénéficier dans un avenir proche des progrès des technologies de séquençage de troisième génération comme le PacBio RS II, Oxford Nanopore minION ou encore Life Technologies Ion Torrent. En effet, bien que ces technologies proposent de séquencer de longues molécules d'ADN uniques, elles ont pour le moment un débit inférieur de plusieurs ordres de grandeurs à celui que l'approche SSV requière.

L'application de l'approche SSV chez la levure *S. cerevisiae* nous a permis d'estimer des taux de formation de novo des différents types de SV. Nous avons aussi pu identifier différents

éléments impliqués dans la plasticité des chromosomes. Je précise que ces résultats restent encore préliminaires puisque les analyses complémentaires soutenant la localisation des SV (corrélation de la position de leurs jonctions avec des informations génétiques comme la position des origines de réplication, la position des sites fragiles ou encore la carte des nucléosomes) sont en cours. La redondance de certains *loci* pourrait indiquer que ceux sont des *hotspots* de recombinaison comme le locus *VMA1* et le locus *YMRCTy1-3/YMRCTy1-4*. Pour ce dernier, nous avons évoqué la possibilité que des ADN circulaires (eccDNA) se forment dans le génome et ceci expliquerait la délétion de 12 kb à ce locus (Møller et al., 2015). Par ailleurs, la médiane des tailles des eccDNA est d'environ 2kb et les deux plus grands eccDNA qui ont été isolés sont respectivement de 20 et 38kb. Il est cependant possible que dans la plupart des cas l'absence de grands eccDNA soit un artefact expérimental. La méthode d'isolement des eccDNA nécessite en effet l'amplification des ADN circulaires purifiés sur colonne avec la polymérase ϕ29 qui ne peut amplifier que des fragments de 10kb au maximum. Il parait donc envisageable que ce mécanisme soit responsable de la délétion de gènes essentiels dans les grandes délétions de plusieurs dizaines de kilo-bases voire centaines de kilo-bases telles que nous les avons observées.

Nous avons également observé de nombreuses translocations non réciproques dans les régions subtélomériques du génome. Compte tenu des fréquences suffisamment élevées de ces NRT que nous avons mesurées par la méthode SSV, il serait envisageable de réaliser un crible pour isoler ces mutants. Ce crible serait réalisé à partir des cellules issues des expériences de SSV (qui pourraient être congelées) et consisterait à étaler et à repiquer des centaines de colonies isolées qui en sont issues. Les clones porteurs de SV seraient alors identifiés par des PCR multiplexées permettant de tester plusieurs SV à la fois en combinant plusieurs couples d'amorces. Après l'isolement de clones mutants, il sera alors possible de quantifier l'impact phénotypique des SV en mesurant leur taux de croissance mitotique dans une grande variété de conditions environnementales (composition du milieu, gradient de température, stress environnementaux, etc.). De plus, il sera également envisageable de tester l'impact des SV sur la reproduction sexuée en testant la viabilité des spores issues de la méiose et d'identifier ainsi l'impact de différents types de SV sur l'isolement reproductif.

J'évoquais au début de cette partie la polyvalence de l'approche SSV. Cette dernière pourrait être davantage exploitée afin d'explorer la formation *de novo* des SV de manière interspécifique. L'étude de la dynamique des génomes dans d'autres espèces que la levure nécessite cependant 2 pré-requis: posséder la séquence du génome de référence de l'espèce que l'on souhaite étudier et être capable de réaliser des cultures monoclonales de cette

espèce. En tenant compte de ces contraintes j'ai sélectionné plusieurs organismes chez lesquels l'exploration de la plasticité chromosomique par une approche SSV serait potentiellement intéressante. Notons que la majorité des eucaryotes sont des microorganismes. Ceci a pour conséquence de rendre de très nombreux eucaryotes compatibles avec l'approche SSV. Les organismes que j'ai ainsi sélectionné permettraient potentiellement d'étudier la plasticité chromosomique dans les 6 grands super-groupes. Chaque espèce sélectionnée est décrite ci-dessous et est également replacée dans l'arbre des eucaryotes à la figure 33.

Super-groupe des *Excavatea*

Les *excavates* forment un super-groupe majeur d'eucaryotes unicellulaires. Ce groupe contient de nombreux organismes libres ainsi que des organismes symbiotiques incluant des parasites humains.

Dans ce super-groupe, j'ai sélectionné *Naegleria gruberi* qui est un eucaryote unicellulaire de la classe des *Heterolobosea* que l'on trouve dans le sol et dans l'eau à la fois sous forme d'amibe ou de flagellé. En culture, ce protiste se divise comme une amibe (avec un temps de génération de 1,6h) et prend sa forme flagellée en cas d'appauvrissement de son milieu nutritif (essentiellement des bactéries). Son génome de taille modeste (41Mb) possède un nombre élevé de gènes homologues avec les autres eucaryotes (Fritz-Laylin et al., 2010). Son génome de 12 chromosomes est une mosaïque de régions complètement homozygotes et d'autres hétérozygotes. De plus, il ne contient que 5.1% de séquences répétées.

Super-groupe des *Chromalveolates*

Thalassiosira pseudonana: cette diatomée a été le premier phytoplancton marin à être séquencé (Armbrust et al., 2004). Il avait été choisi car c'est un organisme modèle pour l'étude de la physiologie des diatomées et le genre auquel il appartient se retrouve dans l'ensemble des océans. De plus cet organisme possède un petit génome de 34 Mb composé de 24 paires de chromosomes. 2% du génome est composé de reliques d'éléments transposables ce qui est caractéristique de la présence et de la mobilité (passée et/ou présente) de tels éléments. Une autre particularité rend l'étude de la stabilité du génome intéressante dans cette espèce : le chromosome 23 correspond à une duplication d'une portion du chromosome 21.

Phaedactylum tricornutum : cet organisme a été le second diatomée séquencé permettant ainsi la première étude de génomique comparative chez les diatomées. Contrairement aux autres diatomées, *P. tricornutum* peut exister sous différentes formes

morphologiques qui dépendent des conditions environnementales. Son génome de 30 Mb en fait un organisme adapté a la détection de SSV (Bowler et al., 2008).

Guillardia theta : cet organisme modèle du groupe des *Hacrobia* est une algue unicellulaire dont les 3 chromosomes totalisent 87 Mb (Curtis et al., 2012).

Emiliania huxleyi: la taille du génome de cet unicellulaire phytoplanctonique du groupe des *rhizaria* est d'environ 142 Mb. Son génome est composé à plus de 64% par des séquences répétées (Read et al., 2013).

Super-groupe des *Unikontes*

Ce super-groupe compte notamment les annimaux et les champignons. Dans ce super-groupe, j'ai sélectionné *Neurospora crassa*. Cet organisme modèle est l'organisme le mieux caractérisé dans le groupe des champignons filamenteux. Son génome de 43 Mb se divise en 7 chromosomes (Galagan et al., 2003). Cette espèce présente toutefois une limitation majeure pour les l'approche SSV : ses colonies apparaissent sur une boite de Petri en 3 semaines !

Le choanoflagellés *Salpingoeca rosetta* est particulièrement intéressant à étudier car il est considéré comme le plus proche ancêtres des animaux. Son génome de 55Mb est également adapte a des expériences de détection de SSV (Fairclough et al., 2013).

Super-groupe des *Archeoplastids (Plantae)*

Ce super-groupe comprend les algues rouges, vertes, les plantes et les glaucophytes. Dans ce super-groupe, j'ai sélectionné *Chlamydomonas reinhardtii* qui est une algue verte unicellulaire mobile grâce à 2 flagelles qui lui permet de nager (Merchant et al., 2010). Son génome de 17 chromosomes atteint une taille de 112 Mb. Les caractéristiques de ce génome semblent "classiques" : distribution en gène uniforme, séquences répétées simples, éléments transposables et répétitions d'ADNr en tandem.

Dans ce super-groupe se trouve également un organisme particulier : Ostreococcus tauri. C'est une espèce d'algue verte unicellulaire de 0,8 µm de diamètre ce qui en fait le plus petit eucaryote photosynthétique connu. Son génome de 12.5 Mb est très compacte (Palenik et al., 2007).

Super-groupe des *Rhizaria*

Dans ce super-groupe principalement unicellulaires, l'étude par SSV de l'algue chlorarachniophytes (ou Chlorarachniophyta) *Bigelowiella natans* pourrait être envisagée. Son génome de 91Mb n'est en revanche pas très bien assemblé puisqu'il se présente sous la forme de 3736 contigs (Curtis et al., 2012).

Super-groupe des *Alveolates*

Les Alveolates forment un super-phylum majeur de protistes. Parmi eux, *Plasmodium falciparum* est une des espèces de Plasmodium qui sont des parasites qui causent le paludisme chez l'être humain. Son génome de 23 Mb est séquencé (Garder 2002).

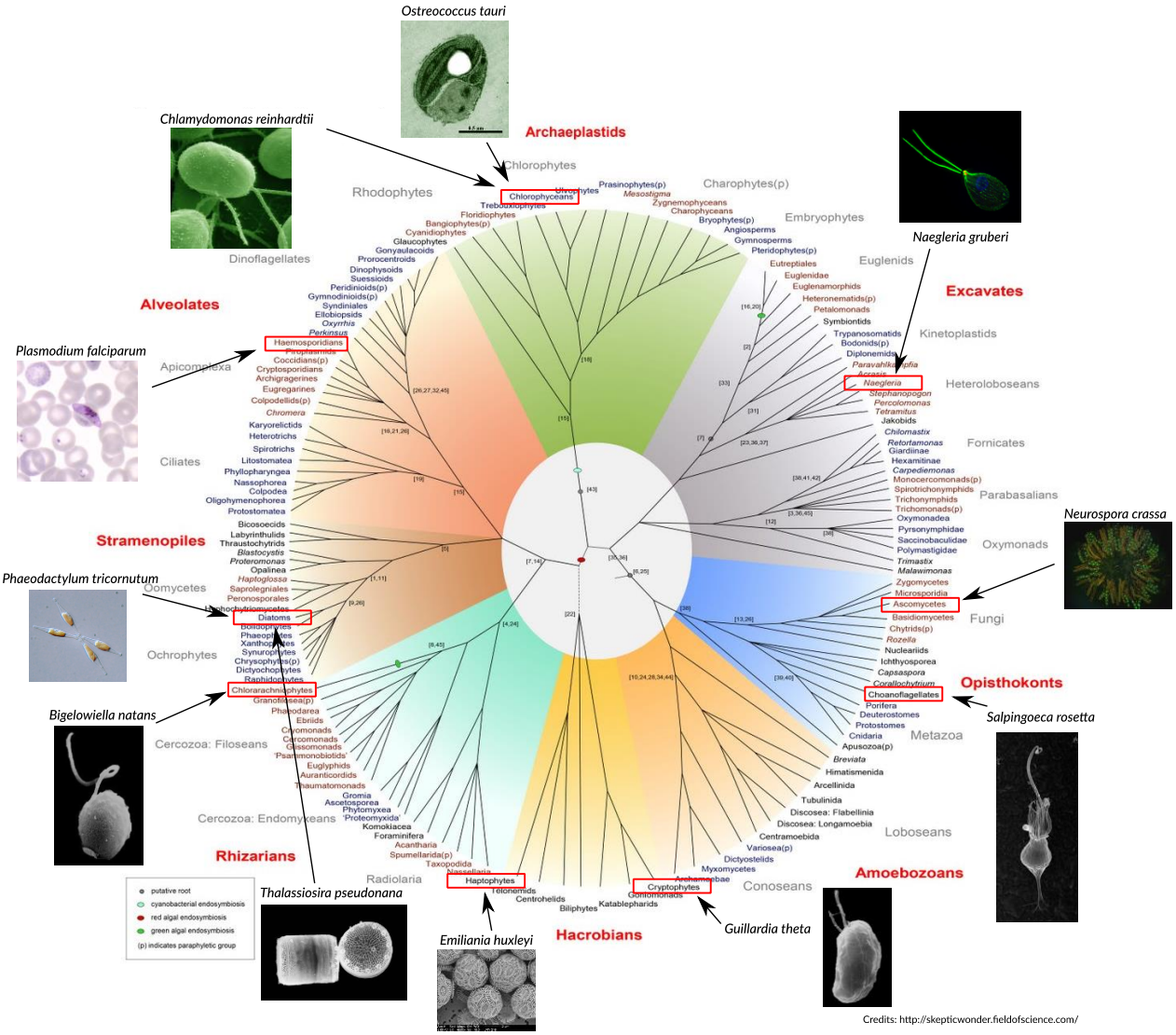


Figure 33: Arbre indicatif des eucaryotes. La longueur des branches n'est pas à l'échelle. Les organismes potentiellement éligibles pour une analyse des SSV sont repartis dans leurs règnes respectifs. Les noms de super-groupes ou infra-regnes sont indiqués en rouge.

Il y a donc 25 ans que le début du séquençage du génome humain a été lancé (en 1990). Nous avons vu au cours de ce travail comment les technologies de séquençage ont révolutionnés la génomique et plus particulièrement celui de l'étude du polymorphisme des chromosomes. Aujourd'hui les limitations d'un projet d'exploration de la plasticité des chromosomes des eucaryotes qui consiste à séquencer des dizaines d'espèces différentes est loin de sembler insurmontable. Cette évolution semble s'inscrire dans le mouvement qui a marqué la fin du « réductionnisme génétique » caractéristique de la biologie du XXe siècle et qui a été remplacé par une intégration globale de l'information sur les organismes. En effet, les techniques d'acquisition de l'information biologique à grande échelle (exome, transcriptiome, SSV…) sont monnaie courante aussi bien dans le monde académique que médicale. Un exemple symptomatique de cette boulimie d'information est aujourd'hui possible de génotype une souche du virus Ebola en moins de 30 min par séquençage. Toutefois, cet afflux d'information pose directement un autre problème : comment traiter et intégrer ces données pour décoder un signal biologique fiable ? Cette question occupe déjà les biologistes et les informaticiens et il est donc fort probable qu'elle nous occupe encore longtemps.

Références

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R. a, Hurles, M.E., and McVean, G. a (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984.

Adams, J., Puskas-Rozsa, S., Simlar, J., and Wilke, C.M. (1992). Adaptation and major chromosomal changes in populations of Saccharomyces cerevisiae. Curr. Genet. *22*, 13–19.

Agam, A., Yalcin, B., Bhomra, A., Cubin, M., Webber, C., Holmes, C., Flint, J., and Mott, R. (2010). Elusive copy number variation in the mouse genome. PLoS One *5*, 1–13.

Aguilera, A., and Gómez-González, B. (2008). Genome instability: a mechanistic view of its causes and consequences. Nat. Rev. Genet. *9*, 204–217.

Aït Yahya-Graison, E., Aubert, J., Dauphinot, L., Rivals, I., Prieur, M., Golfier, G., Rossier, J., Personnaz, L., Creau, N., Bléhaut, H., et al. (2007). Classification of human chromosome 21 gene-expression variations in Down syndrome: impact on disease phenotypes. Am. J. Hum. Genet. *81*, 475–491.

Albertini, a M., Hofer, M., Calos, M.P., and Miller, J.H. (1982). On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell *29*, 319–328.

Alcasabas, a a, Osborn, a J., Bachant, J., Hu, F., Werler, P.J., Bousset, K., Furuya, K., Diffley, J.F., Carr, a M., and Elledge, S.J. (2001). Mrc1 transduces signals of DNA replication stress to activate Rad53. Nat. Cell Biol. *3*, 958–965.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Anderson, W.W., Arnold, J., Baldwin, D.G., Beckenbach, a T., Brown, C.J., Bryant, S.H., Coyne, J. a, Harshman, L.G., Heed, W.B., and Jeffery, D.E. (1991). Four decades of inversion polymorphism in Drosophila pseudoobscura. Proc. Natl. Acad. Sci. U. S. A. *88*, 10367–10371.

Angerer, W.P. (2001). A note on the evaluation of fluctuation experiments. Mutat. Res. Mol. Mech. Mutagen. *479*, 207–224.

Armbrust, E.V., Berges, J. a, Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., et al. (2004). The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science *306*, 79–86.

Aulard, S., Monti, L., Chaminade, N., and Lemeunier, F. (2004). Mitotic and polytene chromosomes: comparisons between Drosophila melanogaster and Drosophila simulans. Genetica *120*, 137–150.

Avelar, A.T., Perfeito, L., Gordo, I., and Godinho Ferreira, M. (2013). Genome architecture is a selectable trait that can be maintained by antagonistic pleiotropy. Nat. Commun. *4*, 2235.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M.T., Perloski, M., Liberg, O., Arnemo, J.M., Hedhammar, A., and Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature *495*, 360–364.

Bartek, J., Lukas, C., and Lukas, J. (2004). Checking on DNA damage in S phase. Nat. Rev. Mol. Cell Biol. *5*, 792–804.

Batista, D.A., Pai, G.S., and Stetten, G. (1994). Molecular analysis of a complex chromosomal rearrangement and a review of familial cases. Am. J. Med. Genet. *53*, 255–263.

Bauman, J.G., Wiegant, J., Borst, P., and van Duijn, P. (1980). A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA. Exp. Cell Res. *128*, 485–490.

Bauters, M., Van Esch, H., Friez, M.J., Boespflug-Tanguy, O., Zenker, M., Vianna-Morgante, A.M., Rosenberg, C., Ignatius, J., Raynaud, M., Hollanders, K., et al. (2008). Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. Genome Res. *18*, 847–858.

Belloni, E., Muenke, M., Roessler, E., Traverso, G., Siegel-Bartelt, J., Frumkin, a, Mitchell, H.F., Donis-Keller, H., Helms, C., Hing, a V, et al. (1996). Identification of Sonic hedgehog as a candidate gene responsible for holoprosencephaly. Nat. Genet. *14*, 353–356.

Benito, C., Llorente, F., Henriques-Gil, N., Gallego, F.J., Zaragoza, C., Delibes, A., and Figueiras, A.M. (1994). A map of rye chromosome 4R with cytological and isozyme markers. Theor. Appl. Genet. *87*, 941–946.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C., et al. (2011). The genomic complexity of primary human prostate cancer. Nature *470*, 214–220.

Bergström, A., Simpson, J.T., Salinas, F., Barré, B., Parts, L., Zia, A., Nguyen Ba, A.N., Moses, A.M., Louis, E.J., Mustonen, V., et al. (2014). A high-definition view of functional genetic variation from natural yeast genomes. Mol. Biol. Evol. *31*, 872–888.

Bermejo, R., Lai, M.S., and Foiani, M. (2012). Preventing Replication Stress to Maintain Genome Stability: Resolving Conflicts between Replication and Transcription. Mol. Cell *45*, 710–718.

Bierne, H., Vilette, D., Ehrlich, S.D., and Michel, B. (1997). Isolation of a dnaE mutation which enhances RecA-independent homologous recombination in the Escherichia coli chromosome. Mol. Microbiol. *24*, 1225–1234.

Bignell, G.R., Bignell, G.R., Huang, J., Huang, J., Greshock, J., Greshock, J., Watt, S., Watt, S., Butler, A., Butler, A., et al. (2004). High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays. Genome Res. 287–295.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J.P., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics *27*, 268–269.

Bothmer, A., Robbiani, D.F., Feldhahn, N., Gazumyan, A., Nussenzweig, A., and Nussenzweig, M.C. (2010). 53BP1 regulates DNA resection and the choice between classical and alternative end joining during class switch recombination. J. Exp. Med. *207*, 855–865.

Bourque, G., Pevzner, P. a., and Tesler, G. (2004). Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes. Genome Res. *14*, 507–516.

Bourque, G., Zdobnov, E.M., Bork, P., Pevzner, P. a., and Tesler, G. (2005). Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. Genome Res. *15*, 98–110.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., et al. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature *456*, 239–244.

Bridges, C. (1936). The BAR "gene" duplication. Science (80-.). *83*, 210–211.

Brown, J., Saracoglu, K., Uhrig, S., Speicher, M.R., Eils, R., and Kearney, L. (2001). Subtelomeric chromosome rearrangements are detected using an innovative 12-color FISH assay (M-TEL). Nat. Med. *7*, 497–501.

Burt, D.W., Bruley, C., Dunn, I.C., Jones, C.T., Ramage, a, Law, a S., Morrice, D.R., Paton, I.R., Smith, J., Windsor, D., et al. (1999). The dynamics of chromosome evolution in birds and mammals. Nature *402*, 411–413.

Cahan, P., Li, Y., Izumi, M., and Graubert, T. a (2009). The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. Nat. Genet. *41*, 430–437.

Cai, X., Evrony, G.D., Lehmann, H.S., Elhosary, P.C., Mehta, B.K., Poduri, A., and Walsh, C.A. (2014). Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Rep. *8*, 1280–1289.

Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L. a, Morsberger, L. a, Latimer, C., McLaren, S., Lin, M.-L., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109–1113.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., et al. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat. Genet. *43*, 956–963.

Capdevila, J., and Belmonte, J.C.I. (2001). Patterning mechanisms controlling vertebrate limb development. Annu. Rev. Cell Dev. Biol. *17*, 87–132.

Caspersson, T., Zech, L., and Johansson, C. (1970). Analysis of human metaphase chromosome set by aid of DNA-binding fluorescent agents. Exp. Cell Res. *62*, 490–492.

Cha, R.S., and Kleckner, N. (2002). ATR homolog Mec1 promotes fork progression, thus averting breaks in replication slow zones. Science *297*, 602–606.

Chaignat, E., Yahya-graison, E.A., Henrichsen, C.N., Aı, E., Chrast, J., Pradervand, S., and Reymond, A. (2011). Copy number variation modifies expression time courses Copy number variation modifies expression time courses. 106–113.

Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.

Chan, J.E., and Kolodner, R.D. (2011). A genetic and structural study of genome rearrangements mediated by high copy repeat Ty1 elements. PLoS Genet. *7*.

Chełkowski, J., Tyrka, M., and Sobkiewicz, A. (2003). Resistance genes in barley (Hordeum vulgare L.) and their identification with molecular markers. J. Appl. Genet. *44*, 291–309.

Chen, C., and Kolodner, R.D. (1999). Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants. Nature *23*, 81–85.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681.

Ciccia, A., and Elledge, S.J. (2010). The DNA Damage Response: Making It Safe to Play with Knives. Mol. Cell *40*, 179–204.

Cobb, J. a., Schleker, T., Rojas, V., Bjergbaek, L., Tercero, J.A., and Gasser, S.M. (2005). Replisome instability, fork collapse, and gross chromosomal rearrangements arise synergistically from Mec1 kinase and RecQ helicase mutations. Genes Dev. *19*, 3055–3069.

Collins, N., Park, R., Spielmeyer, W., Ellis, J., and Pryor, a J. (2001). Resistance gene analogs in barley and their relationship to rust resistance genes. Genome *44*, 375–381.

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. Nat. Genet. *38*, 75–81.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J. V, and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127–1131.

Cremer, T., Lichter, P., Borden, J., Ward, D.C., and Manuelidis, L. (1988). Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. Hum. Genet. *80*, 235–246.

Cubillos, F. a., Louis, E.J., and Liti, G. (2009). Generation of a large set of genetically tractable haploid and diploid Saccharomyces strains. FEMS Yeast Res. *9*, 1217–1225.

Curtis, B. a, Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M.C., Ball, S.G., Gile, G.H., Hirakawa, Y., et al. (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature *492*, 59–65.

Dago, A.E., Stepansky, A., Carlsson, A., Luttgen, M., Kendall, J., Baslan, T., Kolatkar, A., Wigler, M., Bethel, K., Gross, M.E., et al. (2014). Rapid phenotypic and genomic change in response to therapeutic pressure in prostate cancer inferred by high content analysis of single Circulating Tumor Cells. PLoS One *9*.

Daley, J.M., Palmbos, P.L., Wu, D., and Wilson, T.E. (2005). Nonhomologous end joining in yeast. Annu. Rev. Genet. *39*, 431–451.

Dangl, J.L., and Jones, J.D. (2001). Plant pathogens and integrated defence responses to infection. Nature *411*, 826–833.

Dathe, K., Kjaer, K.W., Brehm, A., Meinecke, P., Nürnberg, P., Neto, J.C., Brunoni, D., Tommerup, N., Ott, C.E., Klopocki, E., et al. (2009). Duplications Involving a Conserved Regulatory Element Downstream of BMP2 Are Associated with Brachydactyly Type A2. Am. J. Hum. Genet. *84*, 483–492.

Davis, A.P., and Symington, L.S. (2004). RAD51 -Dependent Break-Induced Replication in Yeast. Mol. Cell. Biol. *24*.

Debolt, S. (2010). Copy number variation shapes genome diversity in arabidopsis over immediate family generational scales. Genome Biol. Evol. *2*, 441–453.

Debrauwère, H., Loeillet, S., Lin, W., Lopes, J., and Nicolas, a (2001). Links between replication and recombination in Saccharomyces cerevisiae: a hypersensitive requirement for homologous recombination in the absence of Rad27 activity. Proc. Natl. Acad. Sci. U. S. A. *98*, 8263–8269.

Dellinger, A.E., Saw, S.M., Goh, L.K., Seielstad, M., Young, T.L., and Li, Y.J. (2010). Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. *38*, 1–14.

Delneri, D., Colson, I., Grammenoudi, S., Roberts, I.N., Louis, E.J., and Oliver, S.G. (2003). Engineering evolution to study speciation in yeasts. Nature *422*, 68–72.

Delneri, D., Hoyle, D.C., Gkargkas, K., Cross, E.J.M., Rash, B., Zeef, L., Leong, H.-S., Davey, H.M., Hayes, A., Kell, D.B., et al. (2008). Identification and characterization of high-flux-control genes of yeast through competition analyses in continuous cultures. Nat. Genet. *40*, 113–117.

Deriano, L., and Roth, D.B. (2013). Modernizing the nonhomologous end-joining repertoire: alternative and classical NHEJ share the stage. Annu. Rev. Genet. *47*, 433–455.

Dernburg, A.F. (2001). Here, there, and everywhere: Kinetochore function on holocentric chromosomes. J. Cell Biol. *153*, 33–38.

Desai, M.M., and Fisher, D.S. (2007). Beneficial mutation-selection balance and the effect of linkage on positive selection. Genetics *176*, 1759–1798.

Deutschbauer, A.M., Jaramillo, D.F., Proctor, M., Kumm, J., Hillenmeyer, M.E., Davis, R.W., Nislow, C., and Giaever, G. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics *169*, 1915–1925.

Dhar, R., Sägesser, R., Weikert, C., Yuan, J., and Wagner, a. (2011). Adaptation of Saccharomyces cerevisiae to saline stress through laboratory evolution. J. Evol. Biol. *24*, 1135–1153.

Dimitrov, L.N., Brem, R.B., Kruglyak, L., and Gottschling, D.E. (2009). Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. Genetics *183*, 365–383.

Dobzhansky, T. (1947). Adaptive changes induced by natural selection in wild populations of Drosophila. Evolution (N. Y). *1*, 1–16.

Dobzhansky, T., and Sturtevant, a. H. (1938). Inversions in the chromosomes of Drosophila pseudoobscura. Genetics *23*, 28–64.

Dopman, E.B., and Hartl, D.L. (2007). A portrait of copy-number polymorphism in Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. *104*, 19920–19925.

Dorsey, M., Peterson, C., Bray, K., and Paqui, C.E. (1992). Spontaneous Amplification of the ADH4 gene in Saccharomyces cerevisiae. Genetics *132*, 943–950.

Drillon, G., and Fischer, G. (2011). Comparative study on synteny between yeasts and vertebrates. Comptes Rendus - Biol. *334*, 629–638.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., Montigny, J. De, Talla, E., Goffard, N., Frangeul, L., et al. (2004). Genome evolution in yeasts. Nature *430*, 35–44.

Dunham, M.J., Badrane, H., Ferea, T., Adams, J., Brown, P.O., Rosenzweig, F., and Botstein, D. (2002). Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. *99*, 16144–16149.

Durkin, S.G., and Glover, T.W. (2007). Chromosome fragile sites. Annu. Rev. Genet. *41*, 169–192.

Eitas, T.K., and Dangl, J.L. (2010). NB-LRR proteins: Pairs, pieces, perception, partners, and pathways. Curr. Opin. Plant Biol. *13*, 472–477.

Emerson, J.J., Cardoso-Moreira, M., Borevitz, J.O., and Long, M. (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science *320*, 1629–1631.

Erhart, E., and Hollenberg, C.P. (1983). The presence of a defective LEU2 gene on 2μ DNA recombinant plasmids of Saccharomyces cerevisiae is responsible for curing and high copy number. J. Bacteriol. *156*, 625–635.

Exinger, F., and Lacroute, F. (1979). Genetic evidence for the creation of a reinitiation site by mutation inside the yeast ura 2 gene. Mol. Gen. Genet. *173*, 109–113.

Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nat. Rev. Genet. *8*, 610–618.

Fairclough, S.R., Chen, Z., Kramer, E., Zeng, Q., Young, S., Robertson, H.M., Begovic, E., Richter, D.J., Russ, C., Westbrook, M.J., et al. (2013). Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. Genome Biol. *14*, R15.

Farré, a., Cuadrado, a., Lacasa-Benito, I., Cistué, L., Schubert, I., Comadran, J., Jansen, J., and Romagosa, I. (2012). Genetic characterization of a reciprocal translocation present in a widely grown barley variety. Mol. Breed. *30*, 1109–1119.

Fink, G.R., and Styles, C. a. (1974). Gene conversion of deletions in the HIS4 region of yeast. Genetics *77*, 231–244.

Fischer, G., James, S. a, Roberts, I.N., Oliver, S.G., and Louis, E.J. (2000). Chromosomal evolution in Saccharomyces. Nature *405*, 451–454.

Fischer, G., Rocha, E.P.C., Brunet, F., Vergassola, M., and Dujon, B. (2006). Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. PLoS Genet. *2*, e32.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R. a, Kirkness, E.F., Kerlavage, a R., Bult, C.J., Tomb, J.F., Dougherty, B. a, and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496–512.

Flint, J., Hill, A., Bowden, D., SJ, O., PR, S., SW, S., J, B.-K., K, B., MP, A., AJ, B., et al. (1986). High frequencies of alpha-thalassaemia are the result of natural selection by malaria. Nature *321*, 744.

Ford, C.E., and Hamerton, J.L. (1956). The chromosomes of man. Nature *177*, 1020–1023.

Ford, C.E., Jones, K.W., Polani, P.E., Almeida, J.C. de, and Briggs, J.H. (1958). A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). Lancet *273*, 711–713.

Foster, P.L. (2006). Methods for determining spontaneous mutation rates. Methods Enzymol. *409*, 195–213.

Franchitto, A. (2013). Genome Instability at Common Fragile Sites : *2013*.

Francke, U., Ochs, H.D., Martinville, B. de, Giacalone, J., Lindgren, V., Distèche, C., Pagon, R.A., Hofker, M.H., Ommen, G.-J.B. van, Pearson, P.L., et al. (1985). Minor Xp21 chromosome deletion in a male associated with expression of duchenne muscular dystrophy, chronic granulomatous disease, retinitis pigmentosa, and McLeod syndrome. Am J Hum Genet *37*, 250–267.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R. a, Fleischmann, R.D., Bult, C.J., Kerlavage, a R., Sutton, G., Kelley, J.M., et al. (1995). The minimal gene complement of Mycoplasma genitalium. Science *270*, 397–403.

Friedberg, E.C. (2005). Suffering in silence: the tolerance of DNA damage. Nat. Rev. Mol. Cell Biol. *6*, 943–953.

Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., Field, M.C., Kuo, A., Paredez, A., Chapman, J., Pham, J., et al. (2010). The Genome of Naegleria gruberi Illuminates Early Eukaryotic Versatility. Cell *140*, 631–642.

Gabor Miklos, G.L. (2005). The human cancer genome project--one more misstep in the war on cancer. Nat. Biotechnol. *23*, 535–537.

Galagan, J.E., Calvo, S.E., Borkovich, K. a, Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S., et al. (2003). The genome sequence of the filamentous fungus Neurospora crassa. Nature *422*, 859–868.

Garvin, D.F., Brown, A.H.D., Raman, H., and Read, B.J. (2000). Genetic mapping of the barley Rrs14 scald resistance gene with RFLP, isozyme and seed storage protein markers. Plant Breed. *119*, 193–196.

Gasser, S.M., and Taddei, a. (2012). Structure and Function in the Budding Yeast Nucleus. Genetics *192*, 107–129.

Gatbonton, T., Imbesi, M., Nelson, M., Akey, J.M., Ruderfer, D.M., Kruglyak, L., Simon, J. a., and Bedalov, A. (2006). Telomere length as a quantitative trait: Genome-wide survey and genetic mapping of telomere length-control genes in yeast. PLoS Genet. *2*, 0304–0315.

Gillet-Markowska, A., Richard, H., Fischer, G., and Lafontaine, I. (2015). Ulysses: Accurate detection of low-frequency structural variations in large insert-size sequencing libraries. Bioinformatics *31*, 801–808.

Gillet-markowska, A., Louvel, G., and Fischer, G. (2015). bz-rates : a web-tool to estimate mutation rates from fluctuation analysis. G3 Genes|Genomes|Genetics.

Gimble, F.S., and Thorner, J. (1992). Homing of a DNA endonuclease gene by meiotic gene conversion in Saccharomyces cerevisiae. Nature *357*, 301–306.

Girirajan, S., Campbell, C.D., and Eichler, E.E. (2011). Human Copy Number Variation and Complex Genetic Disease. Annu. Rev. Genet. *45*, 203–226.

Glessner, J.T., Connolly, J.J.M., and Hakonarson, H. (2012). Rare genomic deletions and duplications and their role in neurodevelopmental disorders. Curr. Top. Behav. Neurosci. *12*, 345–360.

Glessner, J.T., Li, J., and Hakonarson, H. (2013). ParseCNV integrative copy number variation association software with quality tracking. Nucleic Acids Res. *41*, 1–12.

Glover, T.W., Arlt, M.F., Casper, A.M., and Durkin, S.G. (2005). Mechanisms of common fragile site instability. Hum. Mol. Genet. *14*, 197–205.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U. S. A. *108*, 1513–1518.

Gobbini, E., Cesena, D., Galbiati, A., Lockhart, A., and Longhese, M.P. (2013). Interplays between ATM/Tel1 and ATR/Mec1 in sensing and signaling DNA double-strand breaks. DNA Repair (Amst). *12*, 791–799.

Goebel, S.J., Johnson, G.P., Perkus, M.E., Davis, S.W., Winslow, J.P., and Paoletti, E. (1990). The complete DNA sequence of vaccinia virus. Virology *179*, 247–266, 517–563.

Goffeau, a, Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 Genes. Science (80-.). *274*.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science *307*, 1434–1440.

Green, B.M., Finn, K.J., and Li, J.J. (2010). Loss of DNA replication control is a potent inducer of gene amplification. Science *329*, 943–946.

Haldane, J.B.S. (1935). The rate of spontaneous mutation of a human gene. J. Genet. Class. *83*, 235–244.

Halligan, D.L., and Keightley, P.D. (2009). Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. Annu. Rev. Ecol. Evol. Syst. *40*, 151–172.

Hamon, A., and Ycart, B. (2012). Statistics for the Luria-Delbrück distribution. Electron. J. Stat. *6*, 1251–1272.

Hartwell, L.H., and Smith, D. (1985). Altered fidelity of mitotic chromosome transmission in cell cycle mutants of S. cerevisiae. Genetics *110*, 381–395.

Haviv-Chesner, A., Kobayashi, Y., Gabriel, A., and Kupiec, M. (2007). Capture of linear fragments at a double-strand break in yeast. Nucleic Acids Res. *35*, 5192–5202.

Hawthorne, D.C. (1963). a Deletion in Yeast and Its Bearing on the Structure of the Mating Type. Genetics *48*, 1727–1729.

Hehir-Kwa, J.Y., Rodriguez-Santiago, B., Vissers, L.E., de Leeuw, N., Pfundt, R., Buitelaar, J.K., Perez-Jurado, L. a., and Veltman, J. a. (2011). De novo copy number variants associated with intellectual disability have a paternal origin and age bias. J. Med. Genet. *48*, 776–778.

Heng, H.H., Squire, J., and Tsui, L.C. (1992). High-resolution mapping of mammalian genes by in situ hybridization to free chromatin. Proc. Natl. Acad. Sci. U. S. A. *89*, 9509–9513.

Heng, H.H., Bremer, S.W., Stevens, J.B., Horne, S.D., Liu, G., Abdallah, B.Y., Ye, K.J., and Ye, C.J. (2013). Chromosomal instability (CIN): What it is and why it is crucial to cancer evolution. Cancer Metastasis Rev. *32*, 325–340.

Heng, H.H.Q., STEVENS, J.B., LIU, G., BREMER, S.W., YE, K.J., YE, C.J., TAINSKY, M.A., ALANWANG, Y., GENSHENGWU, and REDDY, P.-V. (2006). Stochastic Cancer Progression Driven by Non-Clonal Chromosome Aberrations. J. Cell. Physiol. *208*, 461–472.

Henrichsen, C.N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., Ruedi, M., Kaessmann, H., and Reymond, A. (2009). Segmental copy number variation shapes tissue transcriptomes. Nat. Genet. *41*, 424–429.

Herskowitz, I. (1988). The Hawthorne deletion twenty-five years later. Genetics *120*, 857–861.

Heyer, W.-D., Ehmsen, K.T., and Liu, J. (2010). Regulation of homologous recombination in eukaryotes. Annu. Rev. Genet. *44*, 113–139.

Hodgkin, J. (2005). Karyotype, ploidy, and gene dosage. WormBook 1–9.

Hoffman, E. a, Mcculley, A., Haarer, B., and Arnak, R. (2015). Break-seq reveals hydroxyurea-induced chromosome fragility as a result of unscheduled conflict between DNA replication and transcription. Genome Res. *25*, 402–412.

Hoffmann, A. a., and Rieseberg, L.H. (2008). Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? Annu. Rev. Ecol. Evol. Syst. *39*, 21–42.

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics *26*, i350–i357.

Hou, J., Friedrich, A., De Montigny, J., and Schacherer, J. (2014). Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in saccharomyces cerevisiae. Curr. Biol. *24*, 1153–1159.

Huertas, P., Cortés-Ledesma, F., Sartori, A. a, Aguilera, A., and Jackson, S.P. (2008). CDK targets Sae2 to control DNA-end resection and homologous recombination. Nature *455*, 689–692.

Hurwitz, B.L., Kudrna, D., Yu, Y., Sebastian, A., Zuccolo, A., Jackson, S. a., Ware, D., Wing, R. a., and Stein, L. (2010). Rice structural variation: A comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza. Plant J. *63*, 990–1003.

Iafrate, a J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. Nat. Genet. *36*, 949–951.

Inaki, K., Menghi, F., Woo, X.Y., Wagner, J.P., Jacques, P.-É., Lee, Y.F., Shreckengast, P.T., Soon, W.W., Malhotra, A., Teo, A.S.M., et al. (2014). Systems consequences of amplicon formation in human breast cancer. Genome Res. 1–13.

International Chicken Genome Sequencing Consortium (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*, 695–716.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat. Genet. *44*, 226–232.

Ira, G., and Haber, J.E. (2002). Characterization of RAD51-independent break-induced replication that acts preferentially with short homologous sequences. Mol. Cell. Biol. *22*, 6384–6392.

Ira, G., Malkova, A., Liberi, G., Foiani, M., and Haber, J.E. (2003). Srs2 and Sgs1 – Top3 Suppress Crossovers during Double-Strand Break Repair in Yeast. Cell *115*, 401–411.

Ireland, M.J., Reinke, S.S., and Livingston, D.M. (2000). The impact of lagging strand replication mutations on the stability of CAG repeat tracts in yeast. Genetics *155*, 1657–1665.

Ishkanian, A.S., Malloff, C. a, Watson, S.K., DeLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M. a, et al. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. Nat. Genet. *36*, 299–303.

Itsara, A., Cooper, G.M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R.M., Myers, R.M., Ridker, P.M., Chasman, D.I., et al. (2009). Population analysis of large copy number variants and hotspots of human genetic disease. Am. J. Hum. Genet. *84*, 148–161.

Jackson, A.N., McLure, C. a., Dawkins, R.L., and Keating, P.J. (2007). Mannose binding lectin (MBL) copy number polymorphism in Zebrafish (D. rerio) and identification of haplotypes resistant to L. anguillarum. Immunogenetics *59*, 861–872.

Jacobs, P. a (1981). Mutation rates of structural chromosome rearrangements in man. Am. J. Hum. Genet. *33*, 44–54.

Jacobs, P. a, and Strong, J. a (1959). A case of human intersexuality having a possible XXY sex-determining mechanism. Nature *183*, 302–303.

Jacobs, K.B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., Hutchinson, A., Deng, X., Liu, C., Horner, M.-J., et al. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nat. Genet. *44*, 651–658.

Janevski, A., Varadan, V., Kamalakaran, S., Banerjee, N., and Dimitrova, N. (2012). Effective normalization for copy number variation detection from whole genome sequencing. BMC Genomics *13 Suppl 6*, S16.

Jáuregui, B., Vicente, M.C. De, Messeguer, R., Felipe, a., Bonnet, a., Salesses, G., and Arús, P. (2001). A reciprocal translocation between'Garfi'almond and'Nemared'peach. Theor. Appl. … *102*, 1169–1176.

Jiang, Y., Wang, Y., and Brudno, M. (2012). PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. Bioinformatics *28*, 2576–2583.

Jones, M.E. (1993). Accounting for plating efficiency when estimating spontaneous mutation rates. Mutat. Res. Mutagen. Relat. Subj. *292*, 187–189.

Joron, M., Frezal, L., Jones, R.T., Chamberlain, N.L., Lee, S.F., Haag, C.R., Whibley, A., Becuwe, M., Baxter, S.W., Ferguson, L., et al. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature *477*, 203–206.

Kakeda, K., and Miyahara, S. (1995). Cytogenetical Analyses of Reciprocal Translocations in Barley. Fac. Bioresour. Mie Univ. *14*, 1–24.

Kallioniemi, A., Kallioniemi, O.-P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., and Pinkel, D. (1992). Comparative Genomic Hybridization for molecualr Cytogenetic Analysis oof solid tumors. Science (80-.). *258*, 818–820.

Kasianowicz, J.J., Brandin, E., Branton, D., and Deamer, D.W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. U. S. A. *93*, 13770–13773.

Katju, V., Farslow, J.C., and Bergthorsson, U. (2009). Variation in gene duplicates with low synonymous divergence in Saccharomyces cerevisiae relative to Caenorhabditis elegans. Genome Biol. *10*, R75.

Keane, T.M., Goodstadt, L., Danecek, P., White, M. a, Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56–64.

Kim, H.-M., Narayanan, V., Mieczkowski, P. a, Petes, T.D., Krasilnikova, M.M., Mirkin, S.M., and Lobachev, K.S. (2008). Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. EMBO J. *27*, 2896–2906.

Kim, S.-Y., Kim, J.-H., and Chung, Y.-J. (2012). Effect of Combining Multiple CNV Defining Algorithms on the reliability of CNV calls from SNP Genotyping Data. Genomics Inform. *10*, 194–199.

Kioussis, D., Vanin, E., DeLange, T., Flavell, R. a, and Grosveld, F.G. (1983). Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. Nature *306*, 662–666.

Kirkpatrick, M., and Barton, N. (2006). Chromosome inversions, local adaptation and speciation. Genetics *173*, 419–434.

Kloosterman, W.P., Francioli, L.C., Hormozdiari, F., Marschall, T., Hehir-kwa, J.Y., Abdellaoui, A., Lameijer, E., Moed, M.H., Koval, V., Renkens, I., et al. (2015). Characteristics of de novo structural changes in the human genome. Genome Res. 1–10.

Knouse, K. a., Wu, J., Whittaker, C. a., and Amon, A. (2014). Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. Proc. Natl. Acad. Sci. *111*, 13409–13414.

Koolen, D. a, Kramer, J.M., Neveling, K., Nillesen, W.M., Moore-Barton, H.L., Elmslie, F. V, Toutain, A., Amiel, J., Malan, V., Tsai, A.C.-H., et al. (2012). Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nat. Genet. *44*, 639–641.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420–426.

Koshland, D., Kent, J.C., and Hartwell, L.H. (1985). Genetic analysis of the mitotic transmission of minichromosomes. Cell *40*, 393–403.

Koskiniemi, S., Sun, S., Berg, O.G., and Andersson, D.I. (2012). Selection-driven gene loss in bacteria. PLoS Genet. *8*, 1–7.

Koszul, R., Caburet, S., Dujon, B., and Fischer, G. (2004). Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. EMBO J. *23*, 234–243.

Kraus, E., Leung, W.Y., and Haber, J.E. (2001). Break-induced replication: a review and an example in budding yeast. Proc. Natl. Acad. Sci. U. S. A. *98*, 8255–8262.

Kuo, C.H., and Ochman, H. (2010). The extinction dynamics of bacterial pseudogenes. PLoS Genet. *6*.

Lahiri, M., Gustafson, T.L., Majors, E.R., and Freudenreich, C.H. (2004). Expanded CAG repeats activate the DNA damage checkpoint pathway. Mol. Cell *15*, 287–293.

Lang, G.I., and Murray, A.W. (2008). Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. Genetics *178*, 67–82.

Langley, C.H., Stevens, K., Cardeno, C., Lee, Y.C.G., Schrider, D.R., Pool, J.E., Langley, S. a., Suarez, C., Corbett-Detig, R.B., Kolaczkowski, B., et al. (2012). Genomic variation in natural populations of Drosophila melanogaster. Genetics *192*, 533–598.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Laurie, C.C., Laurie, C. a, Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat. Genet. *44*, 642–650.

Lea, D., and Coulson, C.A. (1949). The distribution of the numbers of mutants in bacterial populations. J. Genet. *49*, 264–285.

Lederberg, B.J., and McCray, A.T. (2001). ' Ome Sweet ' Omics-- A Genealogical Treasury of Words. Sci. *15*, 8.

Lee, K., and Sang, E.L. (2007). Saccharomyces cerevisiae Sae2- and Tel1-dependent single-strand DNA formation at DNA break promotes microhomology-mediated end joining. Genetics *176*, 2003–2014.

Lee, M.C., and Marx, C.J. (2012). Repeated, selection-driven genome reduction of accessory genes in experimental populations. PLoS Genet. *8*, 2–9.

Lee, J.A., Carvalho, C.M.B., and Lupski, J.R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell *131*, 1235–1247.

Lejeune, J., Gautier, M., and Turpin, R. (1959). [Study of somatic chromosomes from 9 mongoloid children]. C. R. Hebd. Seances Acad. Sci. *248*, 1721–1722.

Lemoine, F.J., Degtyareva, N.P., Lobachev, K., and Petes, T.D. (2005). Chromosomal translocations in yeast induced by low levels of DNA polymerase: A model for chromosome fragile sites. Cell *120*, 587–598.

Lengronne, A., and Schwob, E. (2002). The yeast CDK inhibitor Sic1 prevents genomic instability by promoting replication origin licensing in late G1. Mol. Cell *9*, 1067–1078.

Letessier, A., Millot, G. a, Koundrioukoff, S., Lachagès, A.-M., Vogt, N., Hansen, R.S., Malfoy, B., Brison, O., and Debatisse, M. (2011). Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature *470*, 120–123.

Levy, S.F., Blundell, J.R., Venkataram, S., Petrov, D. a., Fisher, D.S., and Sherlock, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature *advance on*.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Liebman, S., Singh, A., and Sherman, F. (1979). A mutator affecting the region of the iso-1-cytochrome c gene in yeast. Genetics *92*, 783–802.

Liebman, S., Shalit, P., and Picologlou, S. (1981). Ty Elements Are Involved in the Formation of Deletions in DELI Strains of Saccharomyces cerevisiae. Cell *26*, 401–409.

Lindsley, D.L., Sandler, L., Baker, B.S., Carpenter, a. T., Denell, R.E., Hall, J.C., Jacobs, P. a., Miklos, G.L., Davis, B.K., Gethmann, R.C., et al. (1972). Segmental aneuploidy and the genetic gross structure of the Drosophila genome. Genetics *71*, 157–184.

Lindstrom, D.L., Leverich, C.K., Henderson, K. a., and Gottschling, D.E. (2011). Replicative age induces mitotic recombination in the ribosomal RNA gene cluster of Saccharomyces cerevisiae. PLoS Genet. *7*.

Lipinski, K.J., Farslow, J.C., Fitzpatrick, K. a., Lynch, M., Katju, V., and Bergthorsson, U. (2011). High spontaneous rate of gene duplication in Caenorhabditis elegans. Curr. Biol. *21*, 306–310.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S. a, Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337–341.

Locke, D.P., Sharp, A.J., McCarroll, S. a, McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M., et al. (2006). Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am. J. Hum. Genet. *79*, 275–290.

Loidl, J., Jin, Q.W., and Jantsch, M. (1998). Meiotic pairing and segregation of translocation quadrivalents in yeast. Chromosoma *107*, 247–254.

Long, Q., Rabanal, F. a, Meng, D., Huber, C.D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B.J., Korte, A., Nizhynska, V., et al. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nat. Genet. *45*, 884–890.

Louis, E.J., and Haber, J.E. (1990). The subtelomeric Y' repeat family in Saccharomyces cerevisiae: an experimental system for repeated sequence evolution. Genetics *124*, 533–545.

Louis, E.J., and Haber, J.E. (1992). The Structure and Evolution of Subtelomeric Y' Repeats in Saccharomyces cerevisiae. Genetics *1331*, 559–574.

Louvel, H., Gillet-Markowska, A., Liti, G., and Fischer, G. (2013). A set of genetically diverged Saccharomyces cerevisiae strains with markerless deletions of multiple auxotrophic genes. Yeast n/a – n/a.

Lovett, S.T., and Feschenko, V. V (1996). Stabilization of diverged tandem repeats by mismatch repair: evidence for deletion formation via a misaligned replication intermediate. Proc. Natl. Acad. Sci. U. S. A. *93*, 7120–7124.

Lovett, S.T., Gluckman, T.J., Simon, P.J., Sutera, V.A., and Drapkin, P.T. (1994). Recombination between repeats in Escherichia coli by a recA-independent, proximity-sensitive mechanism. Mol. Gen. Genet. *245*, 294–300.

Lu, P., Han, X., Qi, J., Yang, J., Wijeratne, A.J., Li, T., and Ma, H. (2012). Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. Genome Res. *22*, 508–518.

Lundblad, V., and Blackburn, E.H. (1993). An alternative pathway for yeast telomere maintenance rescues est1- senescence. Cell *73*, 347–360.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience *1*, 18.

Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. Nat. Genet. *39*, S43–S47.

Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., Garcia, C. a, et al. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell *66*, 219–232.

Luria, S.E., and Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. Genetics *28*, 491.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C.R., Dopman, E.B., Dickinson, W.J., Okamoto, K., Kulkarni, S., Hartl, D.L., et al. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. U. S. A. *105*, 9272–9277.

Lyon, M.F. (2003). Transmission ratio distortion in mice. Annu. Rev. Genet. *37*, 393–408.

Ma, Sandri, and Sarkar (1992). Analysis of the Luria-Delbrück Distribution Using Discrete Convolution Powers. J. Appl. Probab. *29*, 255–267.

Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, W.J., Blanchette, M., Haussler, D., and Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. Genome Res. *16*, 1557–1565.

Ma, J.-L., Kim, E.M., Haber, J.E., and Lee, S.E. (2003). Yeast Mre11 and Rad1 proteins define a Ku-independent mechanism to repair double-strand breaks lacking overlapping end sequences. Mol. Cell. Biol. *23*, 8820–8828.

Mahama, A.A., and Palmer, R.G. (2003). Translocation Breakpoints in Soybean Classical Genetic Linkage Groups 6 and 8. Crop Sci. *43*, 1602.

Manconi, A., Manca, E., Moscatelli, M., Gnocchi, M., Orro, A., Armano, G., and Milanesi, L. (2015). G-CNV: A GPU-Based Tool for Preparing Data to Detect CNVs with Read-Depth Methods. Front. Bioeng. Biotechnol. *3*, 28.

Marie-Nelly, H., Marbouty, M., Cournac, A., Flot, J.-F., Liti, G., Parodi, D.P., Syan, S., Guillén, N., Margeot, A., Zimmer, C., et al. (2014). High-quality genome (re)assembly using chromosomal contact data. Nat. Commun. *5*, 5695.

Maroni, G., Wise, J., Young, J., and Otto, E. (1987). Metallothionein gene duplications and metal tolerance in natural populations of Drosophila melanogaster. Genetics *117*, 739–744.

Marroni, F., Pinosio, S., and Morgante, M. (2014). Structural variation and genome complexity: Is dispensable really dispensable? Curr. Opin. Plant Biol. *18*, 31–36.

Marschall, T., Costa, I.G., Canzar, S., Bauer, M., Klau, G.W., Schliep, A., and Schönhuth, A. (2012). CLEVER: clique-enumerating variant finder. Bioinformatics *28*, 2875–2882.

Mathiasen, D.P., and Lisby, M. (2014). Cell cycle regulation of homologous recombination in Saccharomyces cerevisiae. FEMS Microbiol. Rev. *38*, 172–184.

Maxwell, P.H., Burhans, W.C., and Curcio, M.J. (2011). Retrotransposition is associated with genome instability during chronological aging. Proc. Natl. Acad. Sci. U. S. A. *108*, 20376–20381.

Maydan, J.S., Flibotte, S., Edgley, M.L., Lau, J., Selzer, R.R., Richmond, T. a., Pofahl, N.J., Thomas, J.H., and Moerman, D.G. (2007). Efficient high-resolution deletion discovery in Caenorhabditis elegans by array comparative genomic hybridization. Genome Res. *17*, 337–347.

Maydan, J.S., Lorch, A., Edgley, M.L., Flibotte, S., and Moerman, D.G. (2010). Copy number variation in the genomes of twelve natural isolates of Caenorhabditis elegans. BMC Genomics *11*, 62.

McBride, D.J., Etemadmoghadam, D., Cooke, S.L., Alsop, K., George, J., Butler, A., Cho, J., Galappaththige, D., Greenman, C., Howarth, K.D., et al. (2012). Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes. J. Pathol. *227*, 446–455.

McCarroll, S. a, Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I.W., Maller, J.B., Kirby, A., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. *40*, 1166–1174.

McClintock, B. (1951). Chromosome organization and genic expression. Cold Spring Harb. Symp. Quant. Biol. *16*, 13–47.

McConnell, M.J., Lindberg, M.R., Brennand, K.J., Piper, J.C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R.S., Vermeesch, J.R., Hall, I.M., et al. (2013). Mosaic copy number variation in human neurons. Science *342*, 632–637.

McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E., Hyten, D.L., Gerhardt, D.J., Jeddeloh, J. a., and Stupar, R.M. (2012). Structural Variants in the Soybean Genome Localize to Clusters of Biotic Stress-Response Genes. Plant Physiol. *159*, 1295–1308.

McMurray, M. a, and Gottschling, D.E. (2003). An age-induced switch to a hyper-recombinational state. Science *301*, 1908–1911.

McVey, M., and Lee, S.E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. *24*, 529–538.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, J., Witman, G.B., Terry, A., Salamov, A., Fritz-laylin, L.K., Maréchal-drouard, L., et al. (2010). The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. Science (80-.). *318*, 245–250.

Michalet, X., Ekong, R., Fougerousse, F., Rousseaux, S., Schurra, C., Hornigold, N., van Slegtenhorst, M., Wolfe, J., Povey, S., Beckmann, J.S., et al. (1997). Dynamic molecular combing: stretching the whole human genome for high-resolution studies. Science *277*, 1518–1523.

Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S., and Devine, S.E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. *16*, 1182–1190.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al. (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59–65.

Mimitou, E.P., and Symington, L.S. (2008). Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. Nature *455*, 770–774.

Miret, J.J., Pessoa-Brandão, L., and Lahue, R.S. (1998). Orientation-dependent and sequence-specific expansions of CTG/CAG trinucleotide repeats in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. *95*, 12438–12443.

Molenaar, J.J., Koster, J., Zwijnenburg, D. a., van Sluis, P., Valentijn, L.J., van der Ploeg, I., Hamdi, M., van Nes, J., Westerman, B. a., van Arkel, J., et al. (2012). Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature *483*, 589–593.

Møller, H.D., Parsons, L., Jørgensen, T.S., Botstein, D., and Regenberg, B. (2015). Extrachromosomal circular DNA is common in yeast. Proc. Natl. Acad. Sci. 201508825.

Montelone, B. a., Hoekstra, M.F., and Malone, R.E. (1988). Spontaneous mitotic recombination in yeast: the hyper-recombinational rem1 mutations are alleles of the RAD3 gene. Genetics *119*, 289–301.

Moriel-Carretero, M., and Aguilera, A. (2010). A Postincision-Deficient TFIIH Causes Replication Fork Breakage and Uncovers Alternative Rad51- or Pol32-Mediated Restart Mechanisms. Mol. Cell *37*, 690–701.

Morris, J.J., Lenski, R.E., and Zinser, E.R. (2012). The Black Queen Hypothesis : Evolution of Dependencies through Adaptive Gene Loss. MBio *3*, 1–7.

Mukai, T. (1964). the Genetic Structure of Natural Populations of Drosophila Melanogaster. I. Spontaneous Mutation Rate of Polygenes Controlling Viability. Genetics *50*, 1–19.

Müller, S., O'Brien, P.C.M., Ferguson-Smith, M. a., and Wienberg, J. (1998). Cross-species colour segmenting: A novel tool in human karyotype analysis. Cytometry *33*, 445–452.

Muñoz-Amatriaín, M., Eichten, S.R., Wicker, T., Richmond, T. a, Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T., et al. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol. *14*, R58.

Myung, K., and Kolodner, R.D. (2002). Suppression of genome instability by redundant S-phase checkpoint pathways in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. *99*, 4500–4507.

Myung, K., Datta, a, and Kolodner, R.D. (2001a). Suppression of spontaneous chromosomal rearrangements by S phase checkpoint functions in Saccharomyces cerevisiae. Cell *104*, 397–408.

Myung, K., Chen, C., and Kolodner, R.D. (2001b). Multiple pathways cooperate in the suppression of genome instability in Saccharomyces cerevisiae. Nature *411*, 1073–1076.

Nagai, S., Dubrana, K., Monika Tsai-Pflugfelder, M.B.D., Tania M. Roberts, Brown, G.W., Varela, E., Hediger, F., Gasser, S.M., and Krogan, N.J. (2008). Functional Targeting of DNA Damage. October *322*, 597–602.

Nair, S., Nash, D., Sudimack, D., Jaidee, A., Barends, M., Uhlemann, A.C., Krishna, S., Nosten, F., and Anderson, T.J.C. (2007). Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. Mol. Biol. Evol. *24*, 562–573.

Narayanan, V., Mieczkowski, P. a., Kim, H.M., Petes, T.D., and Lobachev, K.S. (2006). The Pattern of Gene Amplification Is Determined by the Chromosomal Location of Hairpin-Capped Breaks. Cell *125*, 1283–1296.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Newcomb, R.D., Gleeson, D.M., Yong, C.G., Russell, R.J., and Oakeshott, J.G. (2005). Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the Rop-1 locus of the sheep blowfly, Lucilia cuprina. J. Mol. Evol. *60*, 207–220.

Ng, C.K., Cooke, S.L., Howe, K., Newman, S., Xian, J., Temple, J., Batty, E.M., Pole, J.C., Langdon, S.P., AWEdwards, P., et al. (2012). The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer.pdf. J. Pathol. *226*, 703–712.

Nishant, K.T., Wei, W., Mancera, E., Argueso, J.L., Schlattl, A., Delhomme, N., Ma, X., Bustamante, C.D., Korbel, J.O., Gu, Z., et al. (2010). The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. PLoS Genet. *6*, e1001109.

Nowell, P.C., and Hungerford, D.A. (1960). A minute chromosome in human chronic granulocytic leukemia. Science (80-.). *132*, 1488–1501.

O'Huallachain, M., Karczewski, K.J., Weissman, S.M., Urban, A.E., and Snyder, M.P. (2012). Extensive genetic variation in somatic human tissues. PNAS.

Ohnishi, O. (1977). Spontaneous and ethyl methanesulfonate-induced mutations controlling viability in Drosophila melanogaster. III. Heterozygous effect of polygenic mutations. Genetics *87*, 547–556.

Van Ommen, G.-J.B. (2005). Frequency of new copy number variation in humans. Nat. Genet. *37*, 333–334.

Osborn, T.C., Butrulle, D. V., Sharpe, A.G., Pickering, K.J., Parkin, I. a P., Parker, J.S., and Lydiate, D.J. (2003). Detection and Effects of a Homeologous Reciprocal Transposition in Brassica napus. Genetics *165*, 1569–1577.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science *327*, 92–94.

Pagès, V., and Fuchs, R.P. (2003). Uncoupling of leading- and lagging-strand DNA replication during lesion bypass in vivo. Science *300*, 1300–1303.

Painter, T.S. (1934). A New Method for the Study of Chromosome Aberrations and the Plotting of Chromosome Maps in Drosophila Melanogaster. Genetics *19*, 175–188.

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S., et al. (2007). The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc. Natl. Acad. Sci. U. S. A. *104*, 7705–7710.

Pâques, F., and Haber, J.E. (1999). Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev. *63*, 349–404.

Pavelka, N., Rancati, G., Zhu, J., Bradford, W.D., Saraf, A., Florens, L., Sanderson, B.W., Hattem, G.L., and Li, R. (2010). Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. Nature *468*, 321–325.

Payen, C., Koszul, R., Dujon, B., and Fischer, G. (2008). Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. PLoS Genet. *4*, e1000175.

Payen, C., Di Rienzi, S.C., Ong, G.T., Pogachar, J.L., Sanchez, J.C., Sunshine, a. B., Raghuraman, M.K., Brewer, B.J., and Dunham, M.J. (2013). The Dynamics of Diverse Segmental Amplifications in Populations of Saccharomyces cerevisiae Adapting to Strong Selection. G3: Genes|Genomes|Genetics *4*, 399–409.

Pearson, C.E., Nichol Edamura, K., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. *6*, 729–742.

Pérez-Ortín, J.E., Querol, A., Puig, S., and Barrio, E. (2002). Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. Genome Res. *12*, 1533–1539.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F. a, Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.

Petes, T.D. (1980). Unequal meiotic recombination within tandem arrays of yeast ribosomal DNA genes. Cell *19*, 765–774.

Pevzner, P., and Tesler, G. (2003). Genome Rearrangements in Mammalian Evolution : Lessons From Human and Mouse Genomes. Genome Res. *13*, 37–45.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., Macdonald, J.R., Mills, R., et al. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. *29*, 512–520.

Piotrowski, A., Bruder, C.E.G., Andersson, R., Diaz de Ståhl, T., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., Bogdan, A., et al. (2008). Somatic mosaicism for copy number variation in differentiated human tissues. Hum. Mutat. *29*, 1118–1124.

Pirooznia, M., Goes, F.S., and Zandi, P.P. (2015). Whole-genome CNV analysis: advances in computational approaches. Front. Genet. *06*, 1–9.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.-L., Ordóñez, G.R., Bignell, G.R., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191–196.

Poli, J., Tsaponina, O., Crabbé, L., Keszthelyi, A., Pantesco, V., Chabes, A., Lengronne, A., and Pasero, P. (2012). dNTP pools determine fork progression and origin usage under replication stress. EMBO J. *31*, 883–894.

Prado, F., and Aguilera, A. (2003). Control of cross-over by single-strand DNA resection. Trends Genet. *19*, 428–431.

Prado, F., and Aguilera, A. (2005). Impairment of replication fork progression mediates RNA polII transcription-associated recombination. EMBO J. *24*, 1267–1276.

Putnam, C.D., Hayes, T.K., and Kolodner, R.D. (2009). Specific pathways prevent duplication-mediated genome rearrangements. Nature *460*, 984–989.

Quinlan, A.R., Clark, R. a, Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E., Mell, J.C., and Hall, I.M. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. *20*, 623–635.

Ranz, J.M., Casals, F., and Ruiz, A. (2001). How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. Genome Res. *11*, 230–239.

Rausch, T., Zichner, T., Schlattl, a, Stutz, a M., Benes, V., and Korbel, J.O. (2012a). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

Rausch, T., Jones, D.T.W., Zapatka, M., Stütz, A.M., Zichner, T., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Northcott, P.A., et al. (2012b). Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. Cell *148*, 59–71.

Read, B. a, Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., et al. (2013). Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature *499*, 209–213.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

Ribeyre, C., Lopes, J., Boulé, J.B., Piazza, A., Guédin, A., Zakian, V. a., Mergny, J.L., and Nicolas, A. (2009). The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. PLoS Genet. *5*.

Ricard, G., Molina, J., Chrast, J., Gu, W., Gheldof, N., Pradervand, S., Schütz, F., Young, J.I., Lupski, J.R., Reymond, A., et al. (2010). Phenotypic consequences of copy number variation: Insights from smith-magenis and Potocki-Lupski syndrome mouse models. PLoS Biol. *8*, 18–21.

Robertson, R.B. (1916). Taxonomic relationships shown in the chromosomes of Tettigidae and Acrididae. J. Morphol. *27*, 179–331.

Rothkamm, K., Krüger, I., Thompson, L.H., Kru, I., and Lo, M. (2003). Pathways of DNA Double-Strand Break Repair during the Mammalian Cell Cycle Pathways of DNA Double-Strand Break Repair during the Mammalian Cell Cycle. Mol. Cell. Biol. *23*, 5706–5715.

Rothstein, R., Helms, C., and Rosenberg, N. (1987). Concerted deletions and inversions are caused by mitotic recombination between delta sequences in Saccharomyces cerevisiae. Mol. Cell. Biol. *7*, 1198–1207.

Rowley, J.D. (1973). A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature *243*, 290–293.

Rowley, J.D. (1998). The critical role of chromosome translocations in human leukemias. Annu. Rev. Genet. *32*, 495–519.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. Nature *265*, 687–695.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. (1982). Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol. *162*, 729–773.

Sarkar, S., Ma, and Sandri (1992). On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. Genetica 173–179.

Schacherer, J., de Montigny, J., Welcker, A., Souciet, J.L., and Potier, S. (2005). Duplication processes in Saccharomyces cerevisiae haploid strains. Nucleic Acids Res. *33*, 6319–6326.

Schaeffer, S.W. (2008). Selection in Heterogeneous Environments Maintains the Gene Arrangement Polymorphism of Drosophila Pseudoobscura. Evolution (N. Y). *62*, 3082–3099.

Schaeffer, S.W., Goetting-Minesky, M.P., Kovacevic, M., Peoples, J.R., Graybill, J.L., Miller, J.M., Kim, K., Nelson, J.G., and Anderson, W.W. (2003). Evolutionary genomics of inversions in Drosophila pseudoobscura: evidence for epistasis. Proc. Natl. Acad. Sci. U. S. A. *100*, 8319–8324.

Schaeffer, S.W., Bhutkar, A., McAllister, B.F., Matsuda, M., Matzkin, L.M., O'Grady, P.M., Rohde, C., Valente, V.L.S., Aguadé, M., Anderson, W.W., et al. (2008). Polytene chromosomal maps of 11 drosophila species: The order of genomic scaffolds inferred from genetic and physical maps. Genetics *179*, 1601–1655.

Schiestl, R. (1989). Nonmutagenic carcinogens induce intrachromosomal recombination in yeast. Nature *19*, 285–288.

Schlattl, A., Anders, S., Waszak, S.M., Huber, W., and Korbel, J.O. (2011). Relating CNVs to transcriptome data at fine resolution: Assessment of the effect of variant size, type, and overlap with functional regions. Genome Res. *21*, 2004–2013.

Schlötterer, C., and Harr, B. (2001). Microsatellite Instability. Encycl. Life Sci. 1–4.

Schrider, D.R., Houle, D., Lynch, M., and Hahn, M.W. (2013). Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. Genetics *194*, 937–954.

Schrock, E., Manoir, S., Veldman, T., Schoell, B., Wienberg, J., Ferguson-Smith, M.A., Ning, Y., Ledbetter, D.H., Bar-Am, I., Soenksen, D., et al. (1996). Multicolor Spectral Karyotyping of Human Chromosomes. Science (80-.). *273*, 494–497.

Schweitzer, J.K., and Livingston, D.M. (1998). Expansions of CAG repeat tracts are frequent in a yeast mutant defective in Okazaki fragment maturation. Hum. Mol. Genet. *7*, 69–74.

Sebat, J., Sebat, J., Lakshmi, B., Lakshmi, B., Troge, J., Troge, J., Alexander, J., Alexander, J., Young, J., Young, J., et al. (2004). Large-Scale Copy Number Polymorphism in the Human Genome. *305*, 525–528.

Serero, A., Jubin, C., Loeillet, S., Legoix-Né, P., and Nicolas, A.G. (2014). Mutational landscape of yeast mutator strains. Proc. Natl. Acad. Sci. U. S. A. *111*, 1897–1902.

Shimizu, H., Yamaguchi, H., Ashizawa, Y., Kohno, Y., Asami, M., Kato, J., and Ikeda, H. (1997). Short-homology-independent illegitimate recombination in Escherichia coli: distinct mechanism from short-homology-dependent illegitimate recombination. J. Mol. Biol. *266*, 297–305.

Signon, L., Malkova, A., Naylor, M.L., Haber, J.E., and Klein, H. (2001). Genetic Requirements for RAD51 - and Replication Repair of a Chromosomal Double-Strand Break Genetic Requirements for RAD51 - and RAD54 -Independent Break-Induced Replication Repair of a Chromosomal Double-Strand Break. *21*, 2048–2056.

Sindi, S., Helman, E., Bashir, A., and Raphael, B.J. (2009). A geometric approach for classification and comparison of structural variants. Bioinformatics *25*, i222–i230.

Sindi, S., Onal, S., Peng, L., Wu, H.-T., and Raphael, B. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol. *13*, R22.

Sjögren, C., and Nasmyth, K. (2001). Sister chromatid cohesion is required for postreplicative double-strand break repair in Saccharomyces cerevisiae. Curr. Biol. *11*, 991–995.

Slack, A., Thornton, P.C., Magner, D.B., Rosenberg, S.M., and Hastings, P.J. (2006). On the mechanism of gene amplification induced under stress in Escherichia coli. PLoS Genet. *2*, 385–398.

Smith, C.E., Llorente, B., and Symington, L.S. (2007). Template switching during break-induced replication. Nature *447*, 102–105.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H., Cremer, T., and Lichter, P. (1997). Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. Genes Chromosom. Cancer *20*, 399–407.

Sonti, R. V., and Roth, J.R. (1989). Role of gene duplications in the adaptation of Salmonella typhimurium to growth on limiting carbon sources. Genetics *123*, 19–28.

Soutoglou, E., Dorn, J.F., Sengupta, K., Jasin, M., Nussenzweig, A., Ried, T., Danuser, G., and Misteli, T. (2007). Positional stability of single double-strand breaks in mammalian cells. Nat. Cell Biol. *9*, 675–682.

Speicher, M.R., Gwyn Ballard, S., and Ward, D.C. (1996). Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nat. Genet. *12*, 368–375.

Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. *5*.

Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. *61*, 437–455.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., et al. (2005). A common inversion under selection in Europeans. Nat. Genet. *37*, 129–137.

Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. (2003). The genome sequence of Caenorhabditis briggsae: A platform for comparative genomics. PLoS Biol. *1*.

Stenberg, P., Lundberg, L.E., Johansson, A.M., Rydén, P., Svensson, M.J., and Larsson, J. (2009). Buffering of segmental and chromosomal aneuploidies in Drosophila melanogaster. PLoS Genet. *5*.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L. a, et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27–40.

Stewart, F.M. (1991). Fluctuation analysis: the effect of plating efficiency. Genetica *84*, 51–55.

Stewart, F.M., Gordon, D.M., and Levin, B.R. (1990). Fluctuation Analysis: The Probability Distribution of the Number Mutants Under Different Conditions. Genetics *124*, 175–185.

Stirling, P.C., Bloom, M.S., Solanki-Patil, T., Smith, S., Sipahimalani, P., Li, Z., Kofoed, M., Ben-Aroya, S., Myung, K., and Hieter, P. (2011). The complete spectrum of yeast chromosome instability genes

identifies candidate CIN cancer genes and functional roles for ASTRA complex components. PLoS Genet. *7*, e1002057.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., Grassi, A. De, Lee, C., et al. (2007). Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science *315*, 848–853.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Sundararajan, A., Lee, B.S., and Garfinkel, D.J. (2003). The Rad27 (Fen-1) nuclease inhibits Ty1 mobility in Saccharomyces cerevisiae. Genetics *163*, 55–67.

Swanson-wagner, R. a, Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D., and Springer, N.M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. 1689–1699.

Szalay, T., and Golovchenko, J.A. (2015). A de novo DNA Sequencing and Variant Calling Algorithm for Nanopores. BiorXive 1–16.

Szostak, J.W., and Wu, R. (1980). Unequal crossing over in the ribosomal DNA of Saccharomyces cerevisiae. Nature *284*, 426–430.

Tan, C.C. (1935). Salivary gland chromosomes in the two races of Drosophila pseudoobscura. Genetics *20*, 392–402.

Tan, W.Y. (1983). On the distribution of the number of mutants at the hypoxanthine-quanine phosphoribosal transferase locus in Chinese hamster ovary cells. Math. Biosci. *67*, 175–192.

Tanaka, H., and Yao, M. (2009). Palindromic gene amplification — an evolutionarily conserved role for DNA inverted repeats in the genome. Nat Rev Cancer *9*, 216–224.

Tanaka, H., Bergstrom, D.A., Yao, M.-C., and Tapscott, S.J. (2006). Large DNA palindromes as a common form of structural chromosome aberrations in human cancers. Hum. Cell *19*, 17–23.

Tanke, H.J., Wiegant, J., van Gijlswijk, R.P., Bezrookove, V., Pattenier, H., Heetebrij, R.J., Talman, E.G., Raap, a K., and Vrolijk, J. (1999). New strategy for multi-colour fluorescence in situ hybridisation: COBRA: COmbined Binary RAtio labelling. Eur. J. Hum. Genet. *7*, 2–11.

The C. Elegans Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012–2018.

Thompson, L.H., and Schild, D. (2002). Recombinational DNA repair and human disease. Mutat. Res. *509*, 49–78.

Tishkoff, D.X., Filosi, N., Gaida, G.M., and Kolodner, R.D. (1997). A novel mutation avoidance mechanism dependent on S. cerevisiae RAD27 is distinct from DNA mismatch repair. Cell *88*, 253–263.

Torres-Rosell, J., Sunjevaric, I., De Piccoli, G., Sacher, M., Eckert-Boulet, N., Reid, R., Jentsch, S., Rothstein, R., Aragón, L., and Lisby, M. (2007). The Smc5-Smc6 complex and SUMO modification of Rad52 regulates recombinational repair at the ribosomal gene locus. Nat. Cell Biol. *9*, 923–931.

Tourrette, Y., Schacherer, J., Fritsch, E., Potier, S., Souciet, J.L., and De Montigny, J. (2007). Spontaneous deletions and reciprocal translocations in Saccharomyces cerevisiae: Influence of ploidy. Mol. Microbiol. *64*, 382–395.

Trappe, K., Emde, A.-K., Ehrlich, H.-C., and Reinert, K. (2014). Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. Bioinformatics *30*, 1–8.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. (2005). Fine-scale structural variation of the human genome. Nat. Genet.

Vallen, E. a, and Cross, F.R. (1995). Mutations in RAD27 define a potential link between G1 cyclins and DNA replication. Mol. Cell. Biol. *15*, 4291–4302.

Vazquez-Mena, O., Medina-Martinez, I., Juárez-Torres, E., Barrón, V., Espinosa, A., Villegas-Sepulveda, N., Gómez-Laguna, L., Nieto-Martínez, K., Orozco, L., Roman-Basaure, E., et al. (2012). Amplified genes may be overexpressed, unchanged, or downregulated in cervical cancer cell lines. PLoS One *7*.

Vergara, I. a, Tarailo-Graovac, M., Frech, C., Wang, J., Qin, Z., Zhang, T., She, R., Chu, J.S.C., Wang, K., and Chen, N. (2014). Genome-wide variations in a natural isolate of the nematode Caenorhabditis elegans. BMC Genomics *15*, 255.

Volik, S., Zhao, S., Chin, K., Brebner, J.H., Herndon, D.R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W.-L., et al. (2003). End-sequence profiling: sequence-based analysis of aberrant genomes. Proc. Natl. Acad. Sci. U. S. A. *100*, 7696–7701.

Walters, R.G., Jacquemont, S., Valsesia, a, de Smith, a J., Martinet, D., Andersson, J., Falchi, M., Chen, F., Andrieux, J., Lobbens, S., et al. (2010). A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature *463*, 671–675.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell *150*, 402–412.

Wang, M., Beck, C.R., English, A.C., Meng, Q., Buhay, C., Han, Y., Doddapaneni, H. V, Yu, F., Boerwinkle, E., Lupski, J.R., et al. (2015). PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. BMC Genomics *16*.

Watanabe, T., and Horiuchi, T. (2005). A novel gene amplification system in yeast based on double rolling-circle replication. EMBO J. *24*, 190–198.

Watt, P.M., Hickson, I.D., Borts, R.H., and Louis, E.J. (1996). SGS1, a homologue of the Blooms and Werner's syndrome genes, is required for maintenance of genome stability in Saccharomyces cerevisiae. Genetics *144*, 935–945.

Wei, F., Wing, R. a, and Wise, R.P. (2002). Genome dynamics and evolution of the Mla (powdery mildew) resistance locus in barley. Plant Cell *14*, 1903–1917.

Weinert, T. a, Kiset, G.L., and Hartwelp, L.H. (1994). Mitotic checkpoint eenes in budding yeast and the dependence of mitosis on DNA replication and repair. Genes Dev. 652–665.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*, 125–138.

Winkler, H. (1920). In Verbreitung und Ursache der Parthenogenesis im Pflanzen-und Tierreiche. 1–3.

Wittkopp, P.J., and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. *13*, 59–69.

Wu, D., Topper, L.M., and Wilson, T.E. (2008). Recruitment and dissociation of nonhomologous end joining proteins at a DNA double-strand break in Saccharomyces cerevisiae. Genetics *178*, 1237–1249.

Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T.M., Gan, X., Nellåker, C., Goodstadt, L., Nicod, J., Bhomra, A., et al. (2011). Sequence-based characterization of structural variation in the mouse genome. Nature *477*, 326–329.

Yamada, N. a., Rector, L.S., Tsang, P., Carr, E., Scheffer, a., Sederberg, M.C., Aston, M.E., Ach, R. a., Tsalenko, a., Sampas, N., et al. (2011). Visualization of fine-scale genomic structure by oligonucleotide-based high-resolution FISH. Cytogenet. Genome Res. *132*, 248–254.

Yates, L.R., and Campbell, P.J. (2012). Evolution of the cancer genome. Nat. Rev. Genet. *13*, 795.

Ycart, B. (2013). Fluctuation analysis: can estimates be trusted? PLoS One *8*, e80958.

Ycart, B. (2014). Fluctuation analysis with cell deaths. J. Appl. Probab. Stat. *9*, 12–28.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. *19*, 1586–1592.

Zhang, C.-Z., Spektor, A., Cornils, H., Francis, J.M., Jackson, E.K., Liu, S., Meyerson, M., and Pellman, D. (2015). Chromothripsis from DNA damage in micronuclei. Nature.

Zhang, F., Khajavi, M., Connolly, A.M., Towne, C.F., Batish, S.D., and Lupski, J.R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat. Genet. *41*, 849–853.

Zhang, H., Zeidler, A.F.B., Song, W., Puccia, C.M., Malc, E., Greenwell, P.W., Mieczkowski, P. a, Petes, T.D., and Argueso, J.L. (2013). Gene copy-number variation in haploid and diploid strains of the yeast Saccharomyces cerevisiae. Genetics *193*, 785–801.

Zhang, J., Wang, J., and Wu, Y. (2012). An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. BMC Bioinformatics *13 Suppl 6*, S6.

Zhao, W.-W., Wu, M., Chen, F., Jiang, S., Su, H., Liang, J., Deng, C., Hu, C., and Yu, S. (2015). Robertsonian Translocations: An Overview of 872 Robertsonian Translocations Identified in a Diagnostic Laboratory in China. PLoS One *10*, e0122647.

Zheng, Q. (1999). Progress of a half century in the study of the Luria-Delbruck distribution. Math. Biosci. *162*, 1–32.

Zheng, Q. (2002). Statistical and algorithmic methods for fluctuation analysis with SALVADOR as an implementation. Math. Biosci. *176*, 237–252.

Zheng, Q. (2008). A note on plating efficiency in fluctuation experiments. Math. Biosci. *216*, 150–153.

Zhou, Z., and Elledge, S.J. (1993). DUN1 encodes a protein kinase that controls the DNA damage response in yeast. Cell *75*, 1119–1127.

Zhu, Y.O., Siegal, M.L., Hall, D.W., and Petrov, D. a (2014). Precise estimates of mutation rate and spectrum in yeast. Proc. Natl. Acad. Sci. U. S. A. *111*, E2310–E2318.

Zichner, T., Garfield, D. a., Rausch, T., Stütz, A.M., Cannavo, E., Braun, M., Furlong, E.E.M., and Korbel, J.O. (2013). Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. Genome Res. *23*, 568–579.

Zieg, J., Maples, V.F., and Kushner, S.R. (1978). Recombinant levels of Escherichia coli K-12 mutants deficient in various replication, recombination, or repair genes. J. Bacteriol. *134*, 958–966.

ANNEXES

Annexe 1 :Pipeline de construction des librairies PE

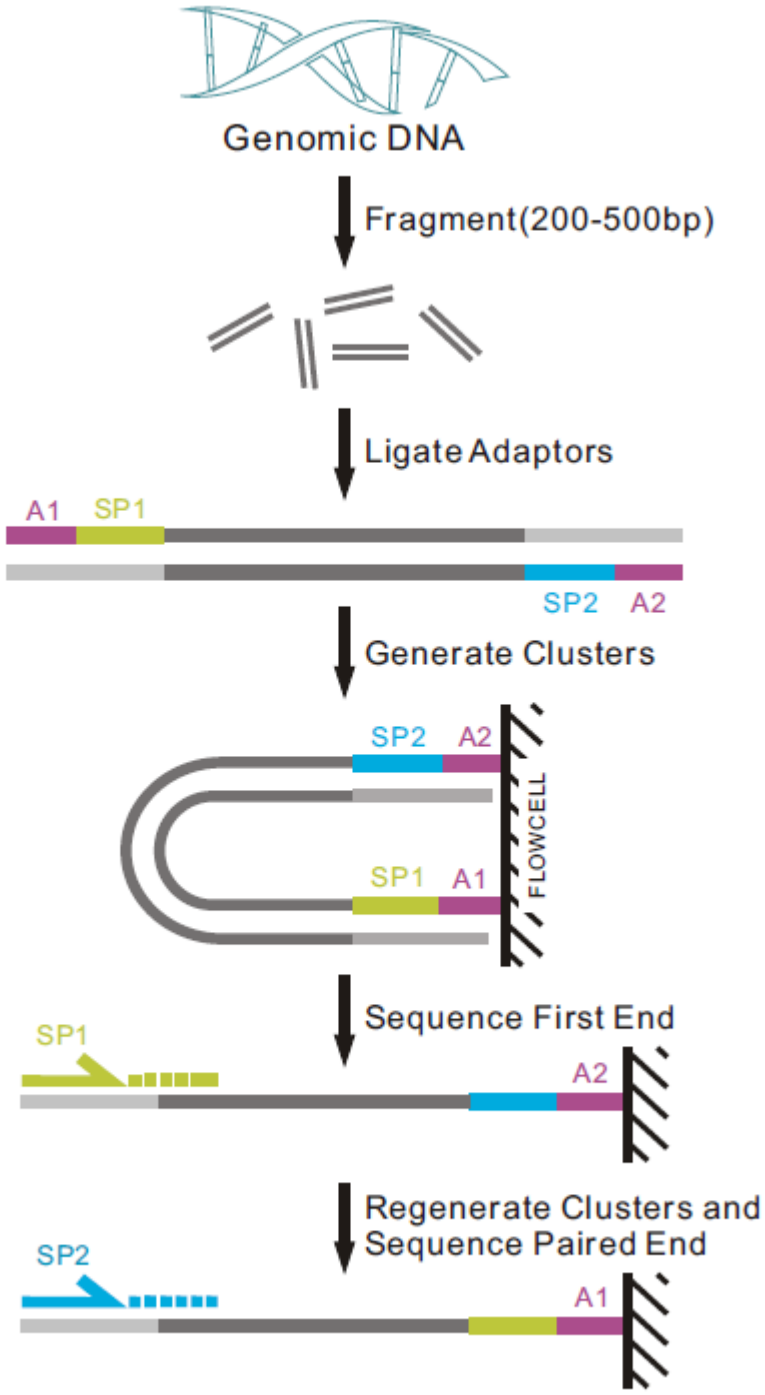
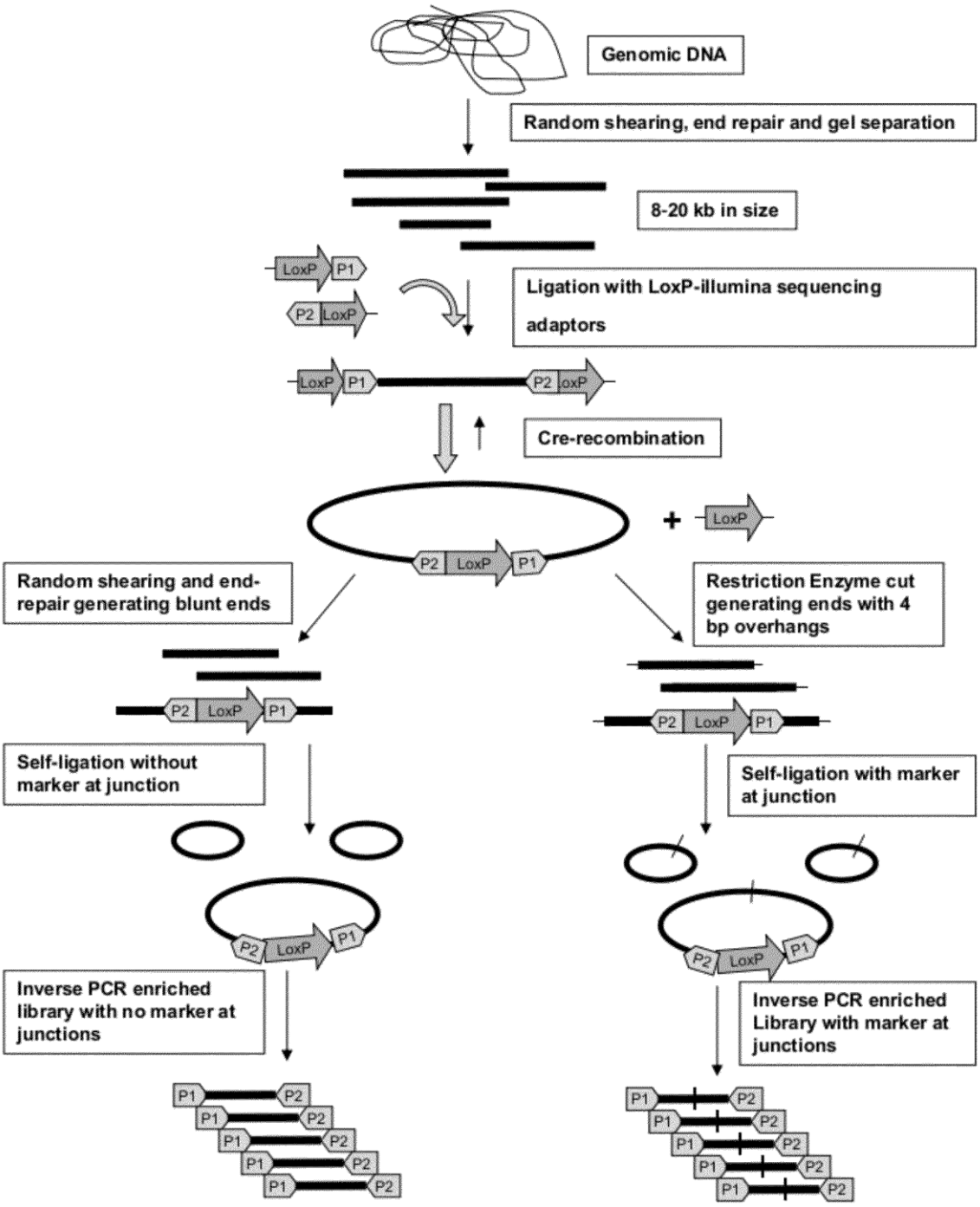


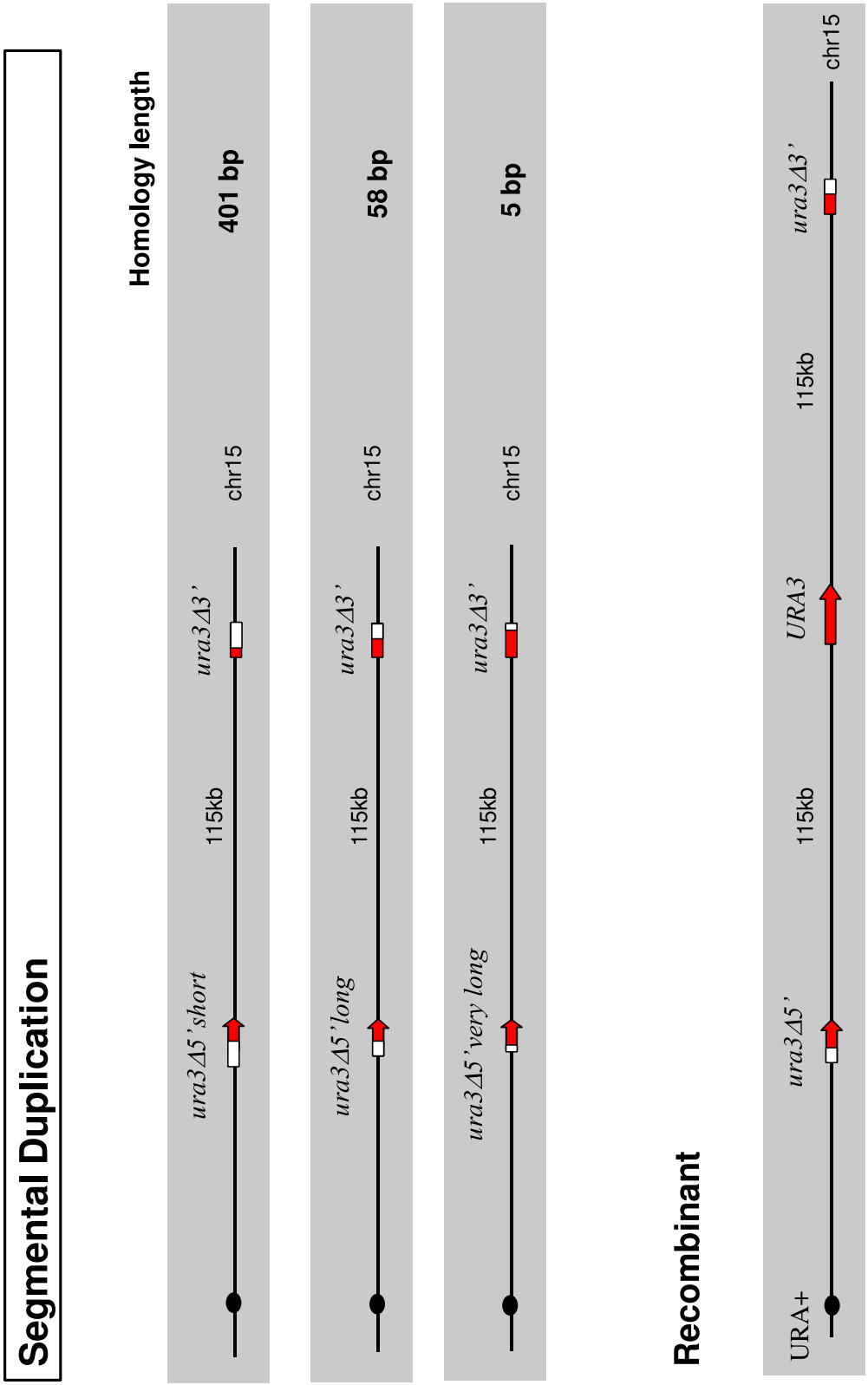
Figure 1-2-1 Pipeline of paired-end sequencing (www.illumina.com)

Annexe 2 :Pipeline de construction des librairies MP

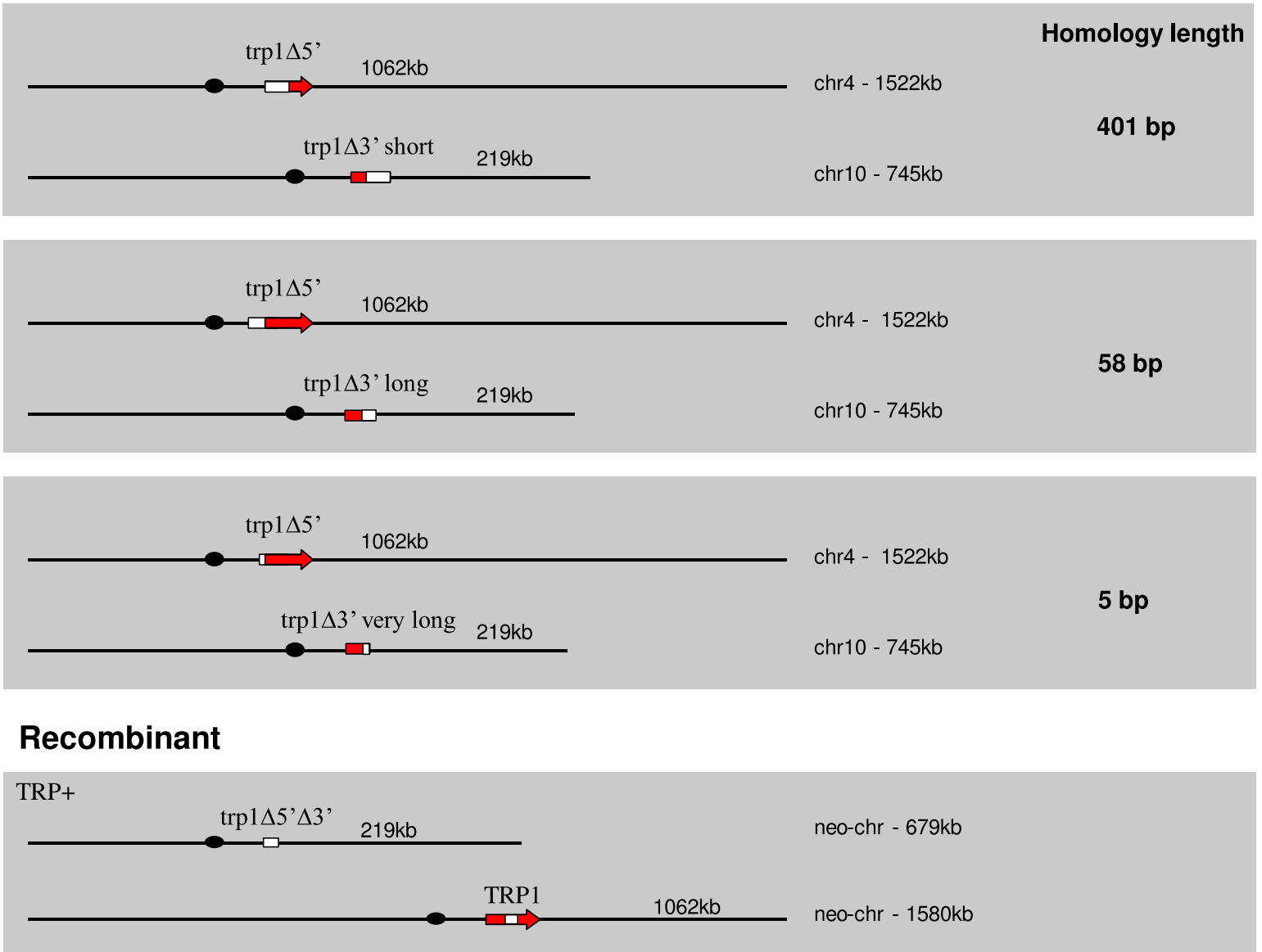


Annexe 3 : Systèmes génétiques de mesures des SV

Annexe 3.1 : DUP401, DUP58, DUP5

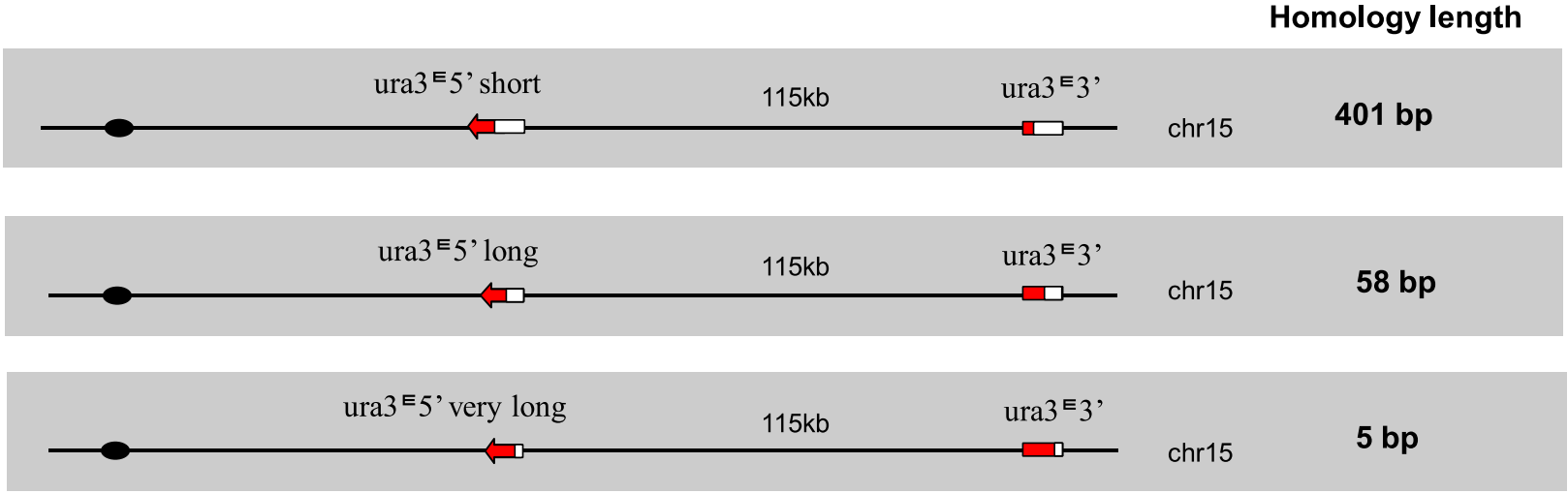


Reciprocal Translocation

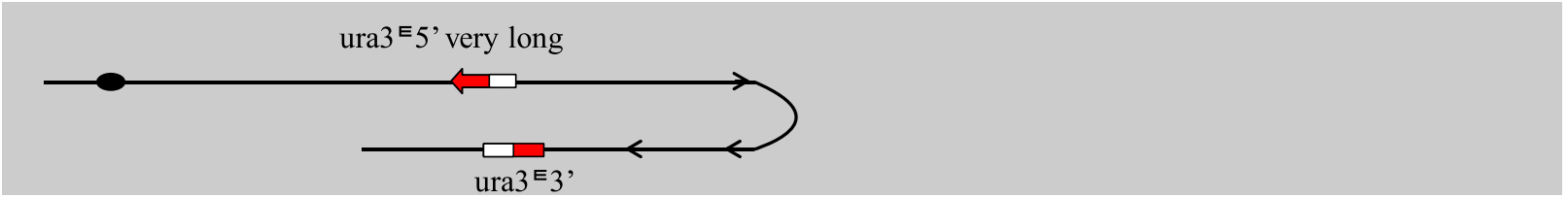


Recombinant

Inversion



Recombination



Recombinant

Annexe 4 : Mutants construits dans les systèmes génétiques DUP411, RT411, INV411

Annexe 4.1 : Partie 1

CAC3	Subunit of chromatin assembly factor I (CAF-1); chromatin assembly by CAF-1 affects multiple processes including silencing at telomeres, mating type loci, and rDNA; maintenance of kinetochore structure; deactivation of DNA damage checkpoint after DNA repair; chromatin
CHK1	Serine/threonine kinase and DNA damage checkpoint effector, mediates cell cycle arrest via phosphorylation of Pds1p; phosphorylated by checkpoint signal transducer Mec1p; homolog of S. pombe and mammalian Chk1 checkpoint kinase
CLB5	B-type cyclin involved in DNA replication during S phase; activates Cdc28p to promote initiation of DNA synthesis; functions in formation of mitotic spindles along with Clb3p and Clb4p; most abundant during late G1 phase
DUN1	Cell-cycle checkpoint serine-threonine kinase required for DNA damage-induced transcription of certain target genes, phosphorylation of Rad55p and Sml1p, and transient G2/M arrest after DNA damage; also regulates postreplicative DNA repair
ELG1	Protein required for S phase progression and telomere homeostasis, forms an alternative replication factor C complex important for DNA replication and genome integrity; involved in homologous recombination-mediated DNA repair
FUN30	Protein whose overexpression affects chromosome stability, potential Cdc28p substrate; homolog of Snf2p; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies
KU70	Subunit of the telomeric Ku complex (Yku70p-Yku80p), involved in telomere length maintenance, structure and telomere position effect; relocates to sites of double-strand cleavage to promote nonhomologous end joining during DSB repair
LIG4	DNA ligase required for nonhomologous end-joining (NHEJ); forms stable heterodimer with required cofactor Lif1p, interacts with Nej1p; involved in meiosis, not essential for vegetative growth; mutations in human ortholog lead to ligase IV syndrome and Dubowitz syndrome
MEC1	Genome integrity checkpoint protein and PI kinase superfamily member; signal transducer required for cell cycle arrest and transcriptional responses prompted by damaged or unreplicated DNA; monitors and participates in meiotic recombination
MUS81	Subunit of the structure-specific Mms4p-Mus81p endonuclease that cleaves branched DNA; involved in DNA repair, replication fork stability, and joint molecule formation/resolution during meiotic recombination; helix-hairpin-helix protein
PIF1	DNA helicase; exists in a nuclear form that acts as a catalytic inhibitor of telomerase; and as a mitochondrial form involved in repair and recombination of mitochondrial DNA; mutations affect zinc and iron homeostasis
POL32	Third subunit of DNA polymerase delta, involved in chromosomal DNA replication; required for error-prone DNA synthesis in the presence of DNA damage and processivity; interacts with Hys2p, PCNA (Pol30p), and Pol1p
RAD17	Checkpoint protein, involved in the activation of the DNA damage and meiotic pachytene checkpoints; with Mec3p and Ddc1p, forms a clamp that is loaded onto partial duplex DNA; homolog of human and S. pombe Rad1 and U. maydis Rec1 proteins
RAD18	Protein involved in postreplication repair; binds single-stranded DNA and has single-stranded DNA dependent ATPase activity; forms heterodimer with Rad6p; contains RING-finger motif
RAD24	Checkpoint protein, involved in the activation of the DNA damage and meiotic pachytene checkpoints; subunit of a clamp loader that loads Rad17p-Mec3p-Ddc1p onto DNA; homolog of human and S. pombe Rad17 protein

RAD27	5' to 3' exonuclease, 5' flap endonuclease, required for Okazaki fragment processing and maturation as well as for long-patch base-excision repair; member of the S. pombe RAD2/FEN1 family
RAD5	DNA helicase proposed to promote replication fork regression during postreplication repair by template switching; RING finger containing ubiquitin ligase; stimulates the synthesis of free and PCNA-bound polyubiquitin chains by Ubc13p-Mms2p
RAD50	Subunit of MRX complex, with Mre11p and Xrs2p, involved in processing double-strand DNA breaks in vegetative cells, initiation of meiotic DSBs, telomere maintenance, and nonhomologous end joining
RAD52	Protein that stimulates strand exchange by facilitating Rad51p binding to single-stranded DNA; anneals complementary single-stranded DNA; involved in the repair of double-strand breaks in DNA during vegetative growth and meiosis
RMI1	Subunit of the RecQ (Sgs1p) - Topo III (Top3p) complex; stimulates superhelical relaxing and ssDNA binding activities of Top3p; involved in response to DNA damage; null mutants display increased rates of recombination and delayed S phase
RRM3	DNA helicase involved in rDNA replication and Ty1 transposition; relieves replication fork pauses at telomeric regions; structurally and functionally related to Pif1p
SGS1	Nucleolar DNA helicase of the RecQ family involved in genome integrity maintenance; regulates chromosome synapsis and meiotic joint molecule/crossover formation; similar to human BLM and WRN proteins implicated in Bloom and Werner syndromes
SLX4	Endonuclease involved in processing DNA; acts during recombination and repair; promotes template switching during break-induced replication (BIR), causing non-reciprocal translocations (NRTs); cleaves branched structures in a complex with Slx1p; involved interstrand cross-link repair and in Rad1p/Rad10p-dependent removal of 3'-nonhomologous tails during DSBR via single-strand annealing; relative distribution to nuclear foci increases upon DNA replication stress
SML1	Ribonucleotide reductase inhibitor involved in regulating dNTP production; regulated by Mec1p and Rad53p during DNA damage and S phase
SRS2	DNA helicase and DNA-dependent ATPase involved in DNA repair, needed for proper timing of commitment to meiotic recombination and transition from Meiosis I to II; blocks trinucleotide repeat expansion; affects genome stability
WSS1	Sumoylated protein of unknown function, identified based on genetic interactions with SMT3; UV-sensitive mutant phenotype and genetic interactions suggest a role in the DNA damage response, processing stalled or collapsed replication forks
YEN1	Holliday junction resolvase; promotes template switching during break-induced replication (BIR), causing non-reciprocal translocations (NRTs); localization is cell-cycle dependent and regulated by Cdc28p phosphorylation; homolog of human GEN1; similar to S. cerevisiae endonuclease Rth1p

Annexe 5 : Article numéro 4

A set of genetically diverged Saccharomyces cerevisiae strains with markerless deletions of multiple auxotrophic genes

Louvel, Hélène, Gillet-Markowska, Alexandre, Liti, Gianni, Fischer, Gilles

Yeast 2013

doi: 10.1002/yea.2991

Yeast
Yeast 2013
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/yea.2991

Research Article

A set of genetically diverged *Saccharomyces cerevisiae* strains with markerless deletions of multiple auxotrophic genes

Hélène Louvel[1,2], Alexandre Gillet-Markowska[1,2], Gianni Liti[3] and Gilles Fischer[1,2]*
[1]*UPMC, UMR7238, Génomique des Microorganismes, Paris, France*
[2]*CNRS, UMR7238, Génomique des Microorganismes, Paris, France*
[3]*Institute for Research on Cancer and Ageing of Nice (IRCAN), CNRS UMR 7284, INSERM U1081, University of Nice Sophia-Antipolis, France*

**Correspondence to:*
Gilles Fischer, Université Paris 06, CNRS UMR7238, Unité de Génomique des Microorganismes, 15 Rue de l'École de Médecine, 75006 Paris, France.
E-mail: gilles.fischer@upmc.fr

Received: 22 August 2013
Accepted: 15 November 2013

Abstract

Genome analysis of over 70 *Saccharomyces* strains revealed the existence of five groups of genetically diverged *S. cerevisiae* wild-type isolates, which feature distinct genetic backgrounds and reflect the natural diversity existing among the species. The strains originated from different geographical and ecological niches (Malaysian, West African, North American, Wine/European and Sake) and represent clean, non-mosaic lineages of *S. cerevisiae*, meaning that their genomes differ essentially by monomorphic and private SNPs. In this study, one representative strain for each of the five *S. cerevisiae* clean lineages was selected and mutated for several auxotroph genes by clean markerless deletions, so that all dominant markers remained available for further genetic manipulations. A set of 50 strains was assembled, including eight haploid and two diploid strains for each lineage. These strains carry different combinations of *leu2Δ0*, *lys2Δ0*, *met15Δ0*, *ura3Δ0* and/or *ura3Δ::KanMX-barcoded* deletions with marker configurations resembling that of the BY series, which will allow large-scale crossing with existing deletion collections. This new set of genetically tractable strains provides a powerful tool kit to explore the impact of natural variation on complex biological processes. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: **yeast; auxotrophic marker; mutant strains; genetic backgrounds**

Introduction

Yeast, such as *S. cerevisiae*, has been a powerful tool in eukaryotic genetics, from the description of the information-carrying genome to the association of gene mutation with human disorders. A huge amount of the genetic information that was gained through the study of *S. cerevisiae* has resulted, more or less directly, from the development of large-scale mutant libraries (Giaever *et al.*, 2002; Mnaimneh *et al.*, 2004; Winzeler *et al.*, 1999; Yu *et al.*, 2006), which initially relied on the design of a set of laboratory strains carrying multiple markerless deletions of auxotrophic genes (Brachmann *et al.*, 1998). These strains have allowed decisive progress to be made in our understanding of many complex biological processes, including cell metabolism, the genetics of mitochondria, meiosis, replication, DNA repair, environmental adaptation and evolution (Botstein and Fink, 2011). However, the extensively studied laboratory strains are not representative of the genetic and phenotypic diversity of the species. Remarkably, extensive cell proliferation phenotyping under many environmental conditions has indicated extreme phenotypes for the laboratory strain S288c when compared to dozens of other *S. cerevisiae* strains (Warringer *et al.*, 2011). An additional striking example of the S288c peculiarity is the extremely high rate of petite formation, which can affect multiple phenotypes (Dimitrov *et al.*, 2009).

Yeast

Furthermore, recent studies have shown that the same mutation in closely related strains can lead to drastic phenotypic differences as a consequence of background-specific genetic interactions (Chin *et al*., 2012; Dowell *et al*., 2010). It is therefore essential to look at diverse backgrounds in order to pursue a comprehensive investigation of genetic polymorphism addressing, especially, how genetic combinations and contexts impact phenotypes. Furthermore, the possibility of exploiting this natural genetics variation is of extreme importance, given the emergence of *S. cerevisiae* as a model to study natural variation, ecology and population-level genetics (Cromie *et al*., in press; Hyma and Fay, 2013; Stefanini *et al*., 2012).

Genome sequencing of over 70 *Saccharomyces* strains and the study of the phylogenetic relationships within the *S. cerevisiae* population has revealed five major groups of wild-type isolates collected from diverse sources and referred to as the Malaysian, West African, North American, Wine/European and Sake groups (Liti *et al*., 2009). These groups were considered to be distinct 'clean' lineages characterized by monomorphic and private SNPs and displayed similar phylogenetic relationships across their entire genome, which contributes to forming consistent groups. The population genomics survey also revealed other strains with high sequence similarities despite distinct worldwide distribution and ecological situations (Liti *et al*., 2009). Indeed, the domestication of *S. cerevisiae* strains, e.g. selection and manipulation for human purposes, may have triggered genetic exchanges through strain crosses. Such populations with biased genomic variability, also named 'mosaic' strains, might therefore provide only a limited understanding of the relationship between the genetic background and the function, stability, dynamics and evolution of genomes. In contrast, the five groups feature distinct genetic backgrounds as a reflection of the natural diversity existing among *S. cerevisiae* species. They represent ideal tools regarding a wide variety of genetic experiments that aim to connect phenotypes to genotypes. This includes the generation of mapping populations for quantitative trait loci analysis (Cubillos *et al*., in press) and understanding trait divergence underlain by differences in expression levels (Chang *et al*., 2013; Lee *et al*., in press) and non-genetic determinants (Ziv *et al*., 2013).

Representative isolates of each of the five *S. cerevisiae* clean lineages were previously genetically engineered to make strains readily usable for laboratory purposes (Cubillos *et al*., 2009). Stable tractable derivatives were thereby developed by disrupting one *HO* allele with the hygromycin resistance cassette and replacing *URA3* with the *KanMX* cassette containing a 6 bp unique 'barcode'. Furthermore, these strains were characterized at both the genomics and the phenomics level (Liti *et al*., 2009; Warringer *et al*., 2011). However, the lack of multiple genetic markers in these new strains has limited their use. In an effort to address this issue and to extend the collection of strains that would be representative of the diversity of the species, we inactivated several of the auxotrophic marker genes commonly used in yeast genetics in the *URA3*- and *HO*-mutated derivative strains. First, *LEU2*, *LYS2* and *MET15* markerless deletions (deletions of entire ORFs without leaving any sequence from positively selectable marker at the corresponding loci) were made by the pop-in/pop-out method (Brachmann *et al*., 1998). Then, the *KanMX* cassette that was originally introduced in *URA3* for deletion (Cubillos *et al*., 2009) was removed by the same technique to allow the recycling of this dominant marker for further constructs. Versions of the *LEU2*, *LYS2* and *MET15* auxotroph strains carrying either *URA3* replaced by the *KanMX* cassette, along with a 6 bp 'barcode' (*ura3Δ::KanMX-Barcode*), or *URA3* deleted without replacement by an additional selectable marker (*ura3Δ0*), were generated. As a result, a collection of 50 strains of *S. cerevisiae* was assembled, including eight haploid and two diploid strains for each lineage, with four auxotroph markers (*LYS2*, *LEU2*, *MET15* and *URA3*) that can be used for further genetic manipulation. This set of strains will undoubtedly be very useful to explore the impact of natural diversity on phenotypic traits and to understand the genetic mechanisms underlying complex traits.

Materials and methods

Strains and media

The original wild-type yeast strains used in this study were previously reported in Liti *et al*. (2009) as part of the *Saccharomyces* Genome Resequencing Project (SGRP). An additional genomic resource was recently released providing *de novo* genome assemblies (http://www.moseslab.csb.utoronto.ca/sgrp/download.html). The five genetic backgrounds

DOI: 10.1002/yea

selected here are representatives of the major diverged lineages described in the SGRP, which were designated Malaysian, West African, North American, Wine/European and Sake. Stable haploid (*MAT***a** or *MATα*, *hoΔ::HphMX*, *ura3Δ::KanMX*) and diploid (*MAT***a**/*MATαHO/hoΔ::HphMX*, *ura3-::KanMX/ura3Δ::KanMX*) derivatives for one representative isolate of each of the five major lineages were subsequently generated and described (Cubillos *et al*., 2009). In this study, these *HO*– and *URA3*-inactivated derivatives were referred to as 'parental strains' (Table 1).

Routinely, yeast cells were grown on yeast extract, peptone and glucose medium (YPD) (Sherman *et al*., 1986). For auxotroph selections, cells were grown on complete synthetic medium (CSM) depleted for the appropriate amino acid (lysine, methionine, leucine) or uracil (Sherman *et al*., 1986). When necessary, YPD was supplemented with hygromycin B 200 μg/ml or geneticin G418 200 μg/ml and CSM was supplemented with 5-fluoro-uracil (5-FoA) 1 g/l. In order to test for petite phenotypes, the 50 auxotroph strains generated in this work were grown on glycerol-containing rich medium.

Transformations of *S. cerevisiae* strains

Transformation of the yeast cells with DNA templates was performed using the lithium acetate method (Gietz and Schiestl, 2007). Briefly, cells were grown in YPD to exponential phase, washed in water, then lithium acetate 0.1 M, and incubated with 1–2 μg DNA (previously linearized pAD plasmid or PCR product) in the presence of carrier DNA (single-stranded salmon sperm DNA) and polyethylene glycol (PEG). The cells were then heat-shocked at 42 °C for 20 min and plated on selective media.

Sporulation and haploid selection

Diploid cells were sporulated for 2–5 days at 30 °C on 2% potassium acetate agar plates. The cells were incubated for 15 min at 37 °C in a 5 mg/ml zymolyase solution. Spores were dissected using a MSM400 Singer (UK) dissection microscope on YPD agar plates and then incubated for 2 days at 30 °C. The colonies were then replica-plated onto appropriate selective media, such as CSM agar plates lacking lysine, leucine, methionine or uracil, in the case of selection for auxotroph mutant strains, or YPD agar plates supplemented with G418 for *ura3Δ::KanMX* strains. Spores from four-viable-spore tetrads with correct marker segregations were selected. Drug resistances and auxotrophies were obtained according to the expected 2:2 segregation. Phenotypes of the spores were checked by restreaking the strains on selective media and amplifying the deleted auxotroph gene loci by PCR.

Crosses between haploid strains and generation of diploids

The mating types of haploid strains were determined by crossing with *MAT***a** and *MATα* tester strains deleted for the *LYS5* gene (*lys5Δ0*). Diploids were selected on minimal medium not supplemented with amino acids or uracil. Growth following crossing with *MAT***a**, *lys5Δ0* indicated that the mating-type of the tested haploid strain

Table 1. Parental *S. cerevisiae* clean lineage strains previously described (Cubillos *et al*.., 2009) and used a starting point for further genetic manipulations

Background	Original strain	Genotype	NCYC Nos
Wine/European	DVBPG6765	(*Mat***a**/*Matα*, *HO*/*hoΔ::HYG*, *ura3Δ::KanMX*/*ura3Δ::KanMX*)	3570 [3597: (MAT**a**), 3622: (MAT*α*)]
West African	DVBPG6044	(*Mat***a**/*Matα*, *HO*/*hoΔ::HYG*, *ura3Δ::KanMX*/*ura3Δ::KanMX*)	3574 [3600: (MAT**a**), 3625: (MAT*α*)]
North American	YPS128	(*Mat***a**/*Matα*, *HO*/*hoΔ::HYG*, *ura3Δ::KanMX*/*ura3Δ::KanMX*)	3581 [3607: (MAT**a**), 3632: (MAT*α*)]
Sake	Y12	(*Mat***a**/*Matα*, *HO*/*hoΔ::HYG*, *ura3Δ::KanMX*/*ura3Δ::KanMX*)	3579 [3605: (MAT**a**), 3630: (MAT*α*)]
Malaysian	UWOPS 03–461.4	(*Mat***a**/*Matα*, *HO*/*hoΔ::HYG*, *ura3Δ::KanMX*/*ura3Δ::KanMX*)	3576 [3602: (MAT**a**), 3627: (MAT*α*)]

NCYC numbers for haploid strains are indicated between square brackets.

DOI: 10.1002/yea

was *MATα* and growth with the *MATα* tester strain characterized a *MAT***a** haploid strain.

To generate diploid strains, two haploid strains with opposite mating types were mixed on YPD agar plates, allowed to grow overnight at 30 °C and restreaked on YPD agar medium to obtain single colonies. In the case of balanced auxotrophies (selection of double- and triple-heterozygous mutant diploids), cells mixed on YPD were replica-plated onto minimal medium lacking all amino acids and supplied with uracil. Diploid colonies were further identified by testing sporulation abilities.

Pop-in/pop-out gene deletions in *S. cerevisiae* 'clean' lineages

UWOPS 03–461.4, DVBPG6044, YPS128, DVBP G6765 and Y12 *URA3*-deleted diploid strains (*ura3Δ::KanMX/ura3Δ::KanMX*) (Table 1) were transformed independently with specific *S. cerevisiae* marker deletion plasmids (ATCC-LGC) for the complete markerless deletion of auxotroph genes based on the pop-in/pop-out deletion method (Brachmann *et al*., 1998). The pAD1, pAD2 and pAD4 plasmids, respectively, carry homologous regions flanking *LEU2*, *LYS2* or *MET15* and specifically target these auxotroph genes for deletion using *URA3* as a selection marker (pAD plasmids carry the *URA3* cassette and confer uracil prototrophy). Plasmid integrants were selected on uracil-depleted CSM. The candidates were confirmed by restreaking on uracil-depleted CSM and integration of the plasmid was checked by PCR at one end of the integration site. The positive clones were heterozygous *LEU2/leu2::pAD1* or *LYS2/lys2::pAD2* or *MET15/met15::pAD4*.

Complete deletions of the targeted auxotroph genes require excision of the inserted pAD plasmids. Diploid transformed colonies were sporulated and dissected to identify four-viable-spore tetrads. Spores were selected for both hygromycin B resistance and uracil prototrophy and grown overnight in YPD liquid medium to be plated onto 5-FoA-containing CSM. After 2 days of growth at 30 °C, 5-FoA-resistant clones, i.e. uracil auxotrophs, were replica-plated onto leucine, lysine or methionine-depleted CSM, with regard to their respective mutation. Auxotrophies were checked on the appropriate corresponding selective media as well as on uracil-depleted CSM. The complete deletion of the targeted gene (and loss of the *URA3*-carrying plasmid) was further confirmed by PCR amplification of the targeted locus. According to this protocol, *leu2Δ0*, *lys2Δ0* or *met15Δ0* single mutants of mating-types *MAT***a** and *MATα* were generated in each of the five *S. cerevisiae* lineages. Triple-mutant strains were obtained by successive rounds of crosses between the single mutant strains and sporulation experiments for the five lineages.

Deletion of *URA3* and removal of the *KanMX* cassette

The *URA3* wild-type gene (1800 bp, containing the ORF and 500 bp both upstream and downstream) was amplified from the genome of the uracil prototroph *S. cerevisiae* tester strain (*lys5Δ0*). This DNA amplicon was used to transform the *LEU2*-, *LYS2*- and *MET15*-deleted mutant strains (haploids *MAT***a**, *hoΔ::HphMX*, *ura3Δ*::*KanMX*, *leu2Δ0*, *lys2Δ0*, *met15Δ0*) in order to restore a wild-type copy of *URA3*. Transformants were selected on uracil-depleted CSM plates. Their uracil prototrophy and G418 sensitivity phenotypes were retested and PCR amplification verified the restoration of the functional *URA3* gene.

The pJL164 plasmid (ATCC-LGC) was used as a template to amplify a deleted version of the *URA3* gene. This piece of DNA includes both upstream and downstream *URA3* ORF flanking regions linked together (without any *URA3* ORF sequence or additional marker). The PCR amplicon was used to transform each uracil prototroph mutant (haploids *MAT***a**, *hoΔ::HphMX*, *leu2Δ0*, *lys2Δ0*, *met15Δ0*) in order to delete *URA3*. Transformants were directly plated on YPD after transformation for overnight growth at 30 °C and replica-plated the next day onto 5-FoA plates. Cells were grown for 1–4 days at 30 °C, depending on the strain. Their uracil auxotrophy phenotype was checked on URA drop-out plates and the deletion of *URA3* was checked by PCR.

Backcross with parental strains

Mutants carrying all four auxotrophies (*MAT***a**, *hoΔ::HphMX*, *leu2Δ0*, *lys2Δ0*, *met15Δ0, ura3Δ0*) were backcrossed twice with their respective parental strain (*MATα*, *hoΔ::HphMX*, *ura3Δ::KanMX*). The first round of backcrossing led to the selection of spores carrying all four auxotrophies, which were further backcrossed once with the parental strains. After the second backcrossing, spores were selected according to the genotypes reported in Table 3.

DOI: 10.1002/yea

Selected haploid auxotroph strains were subsequently crossed to each other to generate diploid strains with balanced gene markers for the *LYS2* and *MET15* alleles (Table 3). All strains are available from the National Culture Yeast Collection (http://www.ncyc.co.uk/) and Accession Nos are listed in Table 3.

Evaluation of cell fitness

Growth assays were performed using Tecan Sunrise 96-well microplate readers and data analysis with Magellan software. First, cells were grown to stationary phase at 30 °C in liquid YPD [optical density (OD) = ~20]. Cultures were then diluted and used to inoculate 100 µl fresh YPD or CSM liquid medium at an estimated initial OD = 0.02. Cultures were grown at 30 °C with constant shaking and ODs were measured every 30 min. Sterile medium was used as control. The parental strains (both *MAT***a** and *MATαhoΔ::HphMX*, *ura3Δ::KanMX* haploids and *HO/hoΔ::HphMX*, *ura3Δ::KanMX/ura3Δ::KanMX* diploid) and the *S. cerevisiae* BY4741 strain were grown along the 50 auxotroph strains in the same conditions. Growth curves were measured in duplicate. Generation times were estimated using a dedicated program (Courbe de Croissance v. 1.4, courtesy of J. Schacherer), which automatically plots (in semi-log scale) the growth curves from the TECAN absorbance data and fits a line to the exponential phase (this fit was manually checked and adapted when necessary). Generation times are automatically calculated from the slope of this line (i.e. the growth rate) as follows: doubling time = ln2/growth rate.

For the Malaysian strains, dry weight experiments were performed in parallel in 10 ml liquid cultures (YPD or CSM). For each strain, four independent cultures were inoculated with either YPD or CSM with the same cell density [100 µl from a small colony (~10^6 cells) resuspended in 1 ml sterile water]. Each of the four cultures was grown under agitation (160 rpm) at 30 °C up to a specific time (four time points from 800 min to 2400 min for the last time point). At the four time points, cells were filtered on 0.22 µm cellulose filters (GSWP04700, Millipore), which were then dried at 65 °C for 2 h and weighed on a scale. Each experiment was performed in duplicate.

Results and discussion

Four auxotrophic marker alleles, *URA3*, *LYS2*, *LEU2* and *MET15*, commonly used in yeast genetics and genomics, were selected for complete markerless deletion. The *URA3*, *LYS2*, *LEU2* and *MET15* genes encode essential enzymes of the uracil, lysine, leucine and methionine biosynthetic pathways, respectively. The four mutations were obtained by pop-in/pop-out (Brachmann *et al*., 1998), using the pAD plasmid series, leading to complete deletion of the targeted protein coding sequences. Precisely, deletion of *LYS2* included the *LYS2* ORF, 304 bp upstream of *LYS2* ATG and 24 bp downstream of its stop codon. Deletion of *LEU2* targeted 6523 bp upstream of ATG and 63 bp downstream of the stop codon; this deletion removed a Ty2 retrotransposon from the DVBPG6765, Y12 and YPS128 strains, while this element is probably absent from the UWOPS 03–461.4 and DVBPG6044 strains, as suggested by SRGP pair-end reads spanning the insertion region (Liti *et al*., 2009). Deletion of *MET15* removed *MET15* ORF, 260 bp upstream of its ATG and 811 bp downstream of the *MET15* stop codon. *URA3* was either replaced from ATG to the stop codon by the *KanMX* cassette (Cubillos *et al*., 2009) or deleted from 222 bp upstream of its ATG to 75 bp downstream of its stop codon. The work flow of the strain construction is illustrated in Fig. 1.

Markerless auxotroph gene deletions in the *S. cerevisiae* 'clean' lineages

One representative isolate of each of the Wine/European, West African, North American, Sake and Malaysian lineages (*ura3Δ::KanMX/ura3Δ::KanMX*) (Table 1), was mutated for *LYS2*, *LEU2* and *MET15*, independently, using the pop-in/pop-out gene deletion method (Fig. 1A). This two-step process includes the integration of a linearized plasmid (pAD) into its specific chromosomal location by homologous recombination (pop-in). Overall, transformations with the integrative plasmids showed good efficiencies in all strains and for all targeted loci (Table 2), averaging 100–1000 transformants/µg pAD1 DNA, 400–1500 transformants for pAD2 and 50–1500 transformants for pAD4. The Malaysian-derived strain, however, yielded 2–30-fold fewer transformants than the other strains. This could possibly be due to the high level of aggregation of this

DOI: 10.1002/yea

strain, which could be detrimental to transformation efficiency (see supporting information, Table S1).

Afterwards, the strains that underwent plasmid excision (pop-out), leading to the deletion of the auxotroph gene, were selected, so that single mutants for either *LEU2, LYS2* or *MET15* were generated in each of the five lineages (Fig. 1A). The observed number of auxotrophic colonies obtained at this step varied greatly, depending on the genetic background and the targeted locus (Table 2). In our hands, the deletion of *LYS2* occurred more frequently than the other two deletions in the five strains (1–30%). *LEU2* and *MET15* deletions highly fluctuated between strains ($< 1\%$ to $>30\%$ and $< 1\%$ to 10%, respectively). However, this result does not seem to correlate with the size of the homology shared between the chromosome and the plasmid, since all homologous regions are similar (~1 kbp). The leucine auxotrophy phenotype was more easily obtained in the Wine/European and North American strains than in the three other strains, whereas mutants for *MET15* were more easily identified in the North American, Sake and Malaysian backgrounds. In the other cases, the efficiencies to obtain leucine and methionine auxotrophs were lower.

Furthermore, the strains did show dissimilar sporulation efficiencies on potassium acetate-containing medium (see supporting information, Table S1). NaCl- and malt-based media were also tested, but did not improve sporulation efficiencies. The North American and Malaysian strains had higher rates of sporulation, which were estimated at 80–90% after 2 days of incubation on potassium acetate medium. The other three lineages required 5–10 days on sporulation media. The West African strain gave usually 10–30% of sporulating cells. The Wine/European and Sake background strains yielded significantly fewer tetrads than the other strains, with 1–5% of sporulation. These results were consistent with the previous report (Cubillos *et al*., 2009).

Subsequently, a *leu2Δ0*, *lys2Δ0*, *met15Δ0* triple mutant in each of the five lineages was generated by successive crossings (Fig. 1B). First, two single mutants with distinct auxotrophies and opposite mating types were crossed to generate a diploid strain. From this diploid strain, spores were dissected to select colonies carrying double auxotrophies. A double-auxotroph haploid strain was then crossed with a haploid strain carrying the third single auxotrophy. Spores obtained from this diploid strain, carrying the leucine, lysine and methionine auxotrophies, were selected (*leu2Δ0*, *lys2Δ0*, *met15Δ0*).

URA3 and *KanMX* cassette as additional selectable markers

The *leu2Δ0*, *lys2Δ0*, *met15Δ0* mutant haploid strains made in each of the five lineages carried a *ura3Δ::KanMX* allele. In order to make complete markerless auxotroph strains which would also allow the recycling of the *KanMX* cassette for further genetic manipulations, we removed the *KanMX* resistance cassette that was used to inactivate the *URA3* gene (Cubillos *et al*., 2009) and generated *ura3Δ0* mutant strains. In each of the five lineages, the *MAT***a**, *ura3Δ::KanMX*, *leu2Δ0*, *lys2Δ0*, *met15Δ0* mutant strain was first transformed with a DNA fragment containing a wild-type copy of the *URA3* gene to restore the uracil prototrophy phenotype (Fig. 1C). The transformation efficiencies varied from one strain to another and reached 38, 5, 48, 189 and 6 transformants/µg DNA amplicon for the Wine/European, West African, North American, Sake and Malaysian strains, respectively (Table 2).

The subsequent uracil prototroph strains were then transformed with a DNA carrying a deleted version of *URA3* corresponding to the entire ORF from the start to the stop codon (Fig. 1C). The estimated transformation efficiency was 6 transformants/µg DNA for the Wine/European, West African and Sake strains, 2 transformants/ g DNA for the North American strain and 1 transformant/µg DNA for the Malaysian strain (Table 2). The quadruple auxotroph strains obtained here were *leu2Δ0*, *lys2Δ0*, *met15Δ0*, *ura3Δ0*.

A set of 50 auxotroph strains to explore *S. cerevisiae* natural variation

In an effort to eliminate putative secondary mutations that might have arisen during the genetic manipulations, the strain carrying all four deletions (*MAT***a**, *leu2Δ0*, *lys2Δ0*, *met15Δ0*, *ura3Δ0*) in each lineage was backcrossed twice with its respective parental strain (*MATα*, *ura3Δ::KanMX*) (Fig. 1D). Spores resulting from these backcrosses and carrying eight different genotypes were collected for each lineage (Fig. 1E, Table 3).

DOI: 10.1002/yea

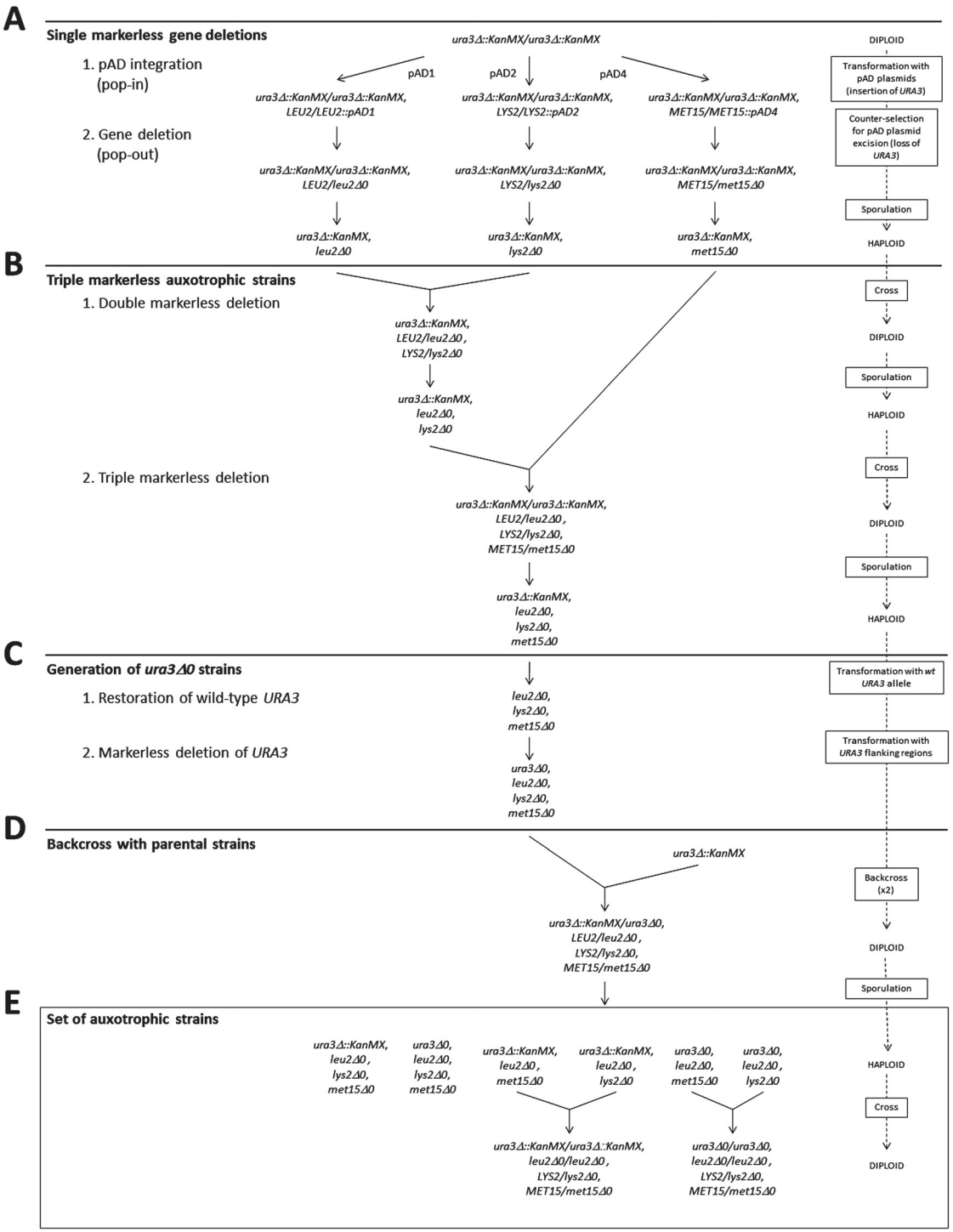


Figure 1. Genealogy of the auxotrophic strains. The successive steps followed in this study are shown. (A) Three auxotrophic genes were independently inactivated in each *S. cerevisiae* lineage without inserting selectable markers (markerless). (B) They were used to make a triple markerless auxotroph mutant strain. (C) A markerless mutated copy of *URA3* (*ura3Δ0*) was made by removing the *KanMX*-barcoded cassette, which was inserted in the *URA3* gene in the parental strains. (D) Backcrossing with the parental strains was realized twice. (E) A set of distinct auxotrophic mutant strains (haploids and diploids) was assembled as indicated

DOI: 10.1002/yea

Table 2. Semi-quantitative data obtained during the genetic manipulation of the strains

Genetic mechanism	*Carrier*	*Targeted gene*	Wine/ European	West African	North American	Sake	Malaysian
Plasmid pop-in[a]	pAD1	*LEU2*	1000	400	200	1000	100
	pAD2	*LYS2*	1500	1400	400	500	100
	pAD4	*MET15*	1500	1400	400	400	50
Plasmid pop-out (%)[b]	pAD1	*LEU2*	> 30	5	< 1	< 1	1.5–4
	pAD2	*LYS2*	1–5	2–30	3	2–10	5–20
	pAD4	*MET15*	7	10	< 1.5	< 1	1–3
Restoration of wild-type *URA3*[a]	PCR product	*URA3*	38	5	48	189	6
Deletion *ura3Δ0*[a]	PCR product	*URA3*	6	6	2	6	1

[a]Number of transformants/μg DNA.
[b]Percentage of auxotrophic cells.

Auxotroph haploid strains for leucine and lysine or leucine and methionine were selected and then crossed to produce two distinct diploid strains with balanced gene markers, i.e. heterozygotes for the *lys2Δ0* and *met15Δ0* mutations (Fig. 1E, Table 3). A collection of 50 strains was assembled, including 40 haploid strains and 10 diploid strains (Table 3). Haploid strains that are mutated for the four biosynthesis pathways of lysine, leucine, methionine and uracil are available in each of the five lineages. These quadruple auxotroph strains will allow the use of most dominant genetic markers, including *KanMX* and *NatMX* (Goldstein and McCusker, 1999), with the exception of the *HphMX* cassette, which was used to delete the *HO* gene, which greatly enhances the possibilities for developing genetic systems. Furthermore, a version of each strain is available with the *ura3Δ0* mutation and the *ura3Δ::KanMX-barcode* mutation labelled with a background-specific barcode, which is designed for identification and quantification purposes. In addition, both *MAT***a** and *MATα* strains were recovered, so that diploid derivatives could be made. The heterozygous diploid strains will also allow the generation of multiple derivatives by combining mutations as needed.

Growth characteristics of the panel of auxotrophic strains

Besides differences in sporulation efficiency (see supporting information, Table S1), the five lineages displayed distinct levels of cell aggregation. For each lineage, we observed under the microscope that cells clump together to form aggregates that cannot be dissociated (either by sonication or by treatment with EDTA or mannose to inhibit lectin-mediated cell aggregation). This was particularly evident in the Malaysian lineage (UWOPS 03–461.4 strain). However, the other backgrounds also showed some levels of aggregation, which were significantly higher for the North American (YPS128) and Wine/European (DVBPG6765) lineages than for the West African (DVBPG6044) and Sake (Y12) strains. These characteristics should be taken into account in experimental designs where cell concentration must be precisely measured.

Further, growth of the 50 auxotroph strains generated in this work was measured in rich and complete synthetic liquid media. Growth curves allowed the calculation of generation times and growth rates for each strain and for their respective parents from the North American, Wine/European, West African and Sake groups (see also supporting information, Figure S1, Table S2). However, the high level of aggregation of the Malaysian strains prevented an accurate measurement of their growth (see below and supporting information, Figure S2).

In our hands, four of the five lineages and their mutant derivatives presented a typical-like growth, so that mutations did not appear to be deleterious to these four lineages. However, it should be noted that the four backgrounds behaved distinctively from each other and that, for instance, their growth curves had different features. In rich medium, the mutant derivatives behaved similarly to their respective parents for all four genetic backgrounds and for both haploid and diploid strains (see supporting information, Figure S1, Table S2). The mutant derivatives also behaved similarly to

DOI: 10.1002/yea

Table 3. Set of 50 auxotroph strains generated in this study

Name	Background	Mating type	*ura3Δ::KanMX*-barcode	*ura3Δ0*	*leu2Δ0*	*lys2Δ0*	*met15Δ0*	NCYC No.
YLF155	Wine/European	**a**	+	–	+	+	+	3882
YLF158	Wine/European	**a**	–	+	+	+	+	3883
YLF156	Wine/European	α	+	–	+	+	+	3884
YLF159	Wine/European	α	–	+	+	+	+	3885
YLF157	Wine/European	**a**	+	–	+	–	+	3886
YLF160	Wine/European	**a**	–	+	+	–	+	3887
YLF154	Wine/European	α	+	–	+	+	–	3888
YLF152	Wine/European	α	–	+	+	+	–	3889
YLF185	Wine/European	**a**/α	–/–	+/+	+/+	+/–	+/–	3890
YLF186	Wine/European	**a**/α	+/+	–/–	+/+	+/–	+/–	3891
YLF183	West African	**a**	+	–	+	+	+	3892
YLF175	West African	**a**	–	+	+	+	+	3893
YLF163	West African	α	+	–	+	+	+	3894
YLF164	West African	α	–	+	+	+	+	3895
YLF184	West African	**a**	+	–	+	–	+	3896
YLF176	West African	**a**	–	+	+	–	+	3897
YLF162	West African	α	+	–	+	+	–	3898
YLF161	West African	α	–	+	+	+	–	3899
YLF187	West African	**a**/α	–/–	+/+	+/+	+/–	+/–	3900
YLF188	West African	**a**/α	+/+	–/–	+/+	+/–	+/–	3901
YLF130	North American	**a**	+	–	+	+	+	3902
YLF131	North American	**a**	–	+	+	+	+	3903
YLF148	North American	α	+	–	+	+	+	3904
YLF149	North American	α	–	+	+	+	+	3905
YLF133	North American	**a**	+	–	+	–	+	3906
YLF132	North American	**a**	–	+	+	–	+	3907
YLF147	North American	α	+	–	+	+	–	3908
YLF146	North American	α	–	+	+	+	–	3909
YLF190	North American	**a**/α	–/–	+/+	+/+	+/–	+/–	3910
YLF189	North American	**a**/α	+/+	–/–	+/+	+/–	+/–	3911
YLF169	Sake	**a**	+	–	+	+	+	3912
YLF170	Sake	**a**	–	+	+	+	+	3913
YLF178	Sake	α	+	–	+	+	+	3914
YLF179	Sake	α	–	+	+	+	+	3915
YLF173	Sake	**a**	+	–	+	–	+	3916
YLF171	Sake	**a**	–	+	+	–	+	3917
YLF177	Sake	α	+	–	+	+	–	3918
YLF181	Sake	α	–	+	+	+	–	3919
YLF191	Sake	**a**/α	–/–	+/+	+/+	+/–	+/–	3920
YLF192	Sake	**a**/α	+/+	–/–	+/+	+/–	+/–	3921
YLF139	Malaysian	**a**	+	–	+	+	+	3922
YLF141	Malaysian	**a**	–	+	+	+	+	3923
YLF140	Malaysian	α	+	–	+	+	+	3924
YLF142	Malaysian	α	–	+	+	+	+	3925
YLF145	Malaysian	**a**	+	–	+	–	+	3926
YLF143	Malaysian	**a**	–	+	+	–	+	3927
YLF138	Malaysian	α	+	–	+	+	–	3928
YLF136	Malaysian	α	–	+	+	+	–	3929
YLF193	Malaysian	**a**/α	–/–	+/+	+/+	+/–	+/–	3930
YLF194	Malaysian	**a**/α	+/+	–/–	+/+	+/–	+/–	3931

+, Deletion present; –, deletion absent.
The *HO* gene was replaced by the *HphMX* cassette in all strains (*hoΔ::HphMX* in haploids and *hoΔ::HphMX/hoΔ::HphMX* in diploids).

DOI: 10.1002/yea

the *S. cerevisiae* BY4741 laboratory strain (see supporting information, Table S2). In complete synthetic medium, the growth of all West African mutant derivatives was similar to that of the parental strain (in both haploid and diploid contexts). However, the growth of the Wine/European, North American and Sake mutant haploid and diploid strains was slower than that of their respective parents, with a 1.5- and 1.8-fold increase of the generation time for the Wine/European haploid and diploid mutants, respectively, and a ~2- and 3-fold increase for the North American and Sake haploid and diploid mutants, respectively. In addition, the generation time of the diploid mutants in minimal medium is significantly higher than that of the corresponding haploid mutant in the Wine/European, North American and Sake groups; being most pronounced in the North American strain (see supporting information, Figure S1, Table S2).

The high level of aggregation of the Malaysian strains prevented the use of the TECAN plate reader to assess fitness. Therefore, dry weight experiments were performed, in both rich and minimal media, to compare growth between mutants and their respective parental strains. For both haploid and diploid mutant strains, no significant growth defect was observed in rich medium, while in minimal medium the mutant strains showed important growth defects relative to their parents (see supporting information, Figure S2).

Altogether, these data emphasize the fact that multiple auxotrophies generally alter cell growth rates in stringent culture conditions, such as when limited nutrients are present (Mulleder *et al*., 2012). A similar growth defect is also observed for the BY laboratory strains (both haploid and diploid) when cultivated in minimal medium (see supporting information, Table S2). It is nonetheless interesting to note that the West African mutant strains did not show any diminished growth in complete synthetic or rich media.

In conclusion, we envision that the use of this set of markerless auxotroph strains, which originate from highly diverged genetic backgrounds, will allow the development of new approaches to experimentally measure the role of natural variation and genetic interaction upon many biological functions and adaptation processes. Individual genome analysis alone (Jelier *et al*., 2011), or integrated with additional intermediate phenotypes datasets (Skelly *et al*., 2013), can be used to develop predictive genotype–phenotype models. It is also interesting to note that a recent multi-locus sequence analysis indicates the presence of eight highly diverged lineages from China (Wang *et al*., 2012), which vastly expand the known genetic variation of *S. cerevisiae*. An approach similar to the one described in this work could also be applied to these natural isolates (and also to additional lineages described in the future) to broaden our understanding of the relationship between genotypes and phenotypes.

Acknowledgements

We thank Jaime Hughes, Conrad Nieduszynski, Edward Louis, Héloïse Muller and Guy-Franck Richard for providing us with the plasmids and parental yeast strains used in this study. We are grateful to Natacha Sertour and Christophe d'Enfert for their help in performing the growth curve experiments. We thank Joseph Schacherer for providing us with the growth curve analysis tool. This work was supported by an ATIP grant from the Centre National de la Recherche Scientifique.

References

Botstein D, Fink GR. 2011. Yeast: an experimental organism for 21st Century biology. *Genetics* **189**: 695–704.

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

Chang J, Zhou Y, Hu X, *et al*. 2013. The molecular mechanism of a *cis*-regulatory adaptation in yeast. *PLoS Genet* **9**: e1003813.

Chin BL, Ryan O, Lewitter F, *et al*. 2012. Genetic variation in *Saccharomyces cerevisiae*: circuit diversification in a signal transduction network. *Genetics* **192**: 1523–1532.

Cromie GA, Hyma KE, Ludlow CL, *et al*. in press. Genomic sequence diversity and population structure of *Saccharomyces cerevisiae* assessed by RAD-seq. *G3 (Bethesda)*.

Cubillos FA, Louis EJ, Liti G. 2009. Generation of a large set of genetically tractable haploid and diploid *Saccharomyces* strains. *FEMS Yeast Res* **9**: 1217–1225.

Cubillos FA, Parts L, Salinas F, *et al*. in press. High resolution mapping of complex traits with a four-parent advanced intercross yeast population. *Genetics*.

Dimitrov LN, Brem RB, Kruglyak L, Gottschling DE. 2009. Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of *Saccharomyces cerevisiae* S288C strains. *Genetics* **183**: 365–383.

Dowell RD, Ryan O, Jansen A, *et al*. 2010. Genotype to phenotype: a complex problem. *Science* **328**: 469.

Giaever G, Chu AM, Ni L, *et al*. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**: 387–391.

Gietz RD, Schiestl RH. 2007. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc* **2**: 31–34.

Goldstein AL, McCusker JH. 1999. Three new dominant drug resistance cassettes for gene disruption in *Saccharomyces cerevisiae*. *Yeast* **15**: 1541–1553.

DOI: 10.1002/yea

Hyma KE, Fay JC. 2013. Mixing of vineyard and oak-tree ecotypes of *Saccharomyces cerevisiae* in North American vineyards. *Mol Ecol* **22**: 2917–2930.

Jelier R, Semple JI, Garcia-Verdugo R, Lehner B. 2011. Predicting phenotypic variation in yeast from individual genome sequences. *Nat Genet* **43**: 1270–1274.

Lee HN, Mostovoy Y, Hsu TY, *et al*. in press. Divergence of iron metabolism in wild Malaysian yeast. *G3 (Bethesda)*.

Liti G, Carter DM, Moses AM, *et al*. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.

Mnaimneh S, Davierwala AP, Haynes J, *et al*. 2004. Exploration of essential gene functions via titratable promoter alleles. *Cell* **118**: 31–44.

Mulleder M, Capuano F, Pir P, *et al*. 2012. A prototrophic deletion mutant collection for yeast metabolomics and systems biology. *Nat Biotechnol* **30**: 1176–1178.

Sherman F, Fink G, Hicks J. 1986. Laboratory Course Manual for Methods in Yeast Genetics. Cold Spring Harbor Laboratory Press: New York.

Skelly DA, Merrihew GE, Riffle M, *et al*. 2013. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res* **23**: 1496–1504.

Stefanini I, Dapporto L, Legras JL, *et al*. 2012. Role of social wasps in *Saccharomyces cerevisiae* ecology and evolution. *Proc Natl Acad Sci U S A* **109**: 13398–13403.

Wang QM, Liu WQ, Liti G, *et al*. 2012. Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. *Mol Ecol* **21**: 5404–5417.

Warringer J, Zorgo E, Cubillos FA, *et al*. 2011. Trait variation in yeast is defined by population history. *PLoS Genet* **7**: e1002111.

Winzeler EA, Shoemaker DD, Astromoff A, *et al*. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**: 901–906.

Yu L, Pena Castillo L, Mnaimneh S, *et al*. 2006. A survey of essential gene function in the yeast cell division cycle. *Mol Biol Cell* **17**: 4736–4747.

Ziv N, Siegal ML, Gresham D. 2013. Genetic and nongenetic determinants of cell growth variation assessed by high-throughput microscopy. *Mol Biol Evol* **30**: 2568–2578.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Figure S1. Typical growth curves and generation times measured for the clean lineage parental and mutant strains. Data for one representative haploid mutant strain and one representative diploid mutant strain of the Wine/European, West African, North American, Sake and Malaysian lineages are shown. Growth in rich medium (YPD) or complete synthetic medium (CSM) was followed over 48 h and generation times were calculated. Growth curves for one representative haploid mutant strain (mata) of the Wine/European, West African, North American, Sake and Malaysian lineages, YLF155, YLF183, YLF130, YLF169 and YLF139 (see Table 3), respectively, are presented. Growth curves obtained from haploid or diploid mutant strains and from their respective parental strains were similar. Generation times for the haploid parental (white bar) and mutant (hatched bar) strains as well as for the diploid parental (grey bar) and mutant (grey bar with squares) are presented, except for the Malaysian lineage, because of its high level of aggregation. The error bars correspond to standard error of the mean. Representative diploid mutant strain of the Wine/European, West African, North American, and Sake lineages were YLF185, YLF187, YLF190 and YLF191 (see Table 3), respectively

Figure S2. Dry weight-based growth curves of Malaysian strains. Growth curves were obtained for haploid and diploid parental strains (UWOPS 03-461.4) as well as for one representative haploid and diploid mutant strain (YLF139 and YLF193, respectively; see Table 3). Cells were cultivated at 30°C under agitation in rich (YPD) or minimum (CSM) medium for about 40 h (2400 min) and their weight was recorded. Each experiment was performed in duplicate. The error bars correspond to the standard error of the mean.

Table S1. Estimated level of aggregation and sporulation efficiency for each clean lineage.

Table S2. Estimated generation times (G) and growth rates (μ) of parental strains and auxotrophic strains generated in this study

DOI: 10.1002/yea