Université de Montréal

Détection à grande échelle des réarrangements génomiques et élucidation de leurs mécanismes

par Samuel Tremblay-Belzile

Département de biochimie et médecine moléculaire
Faculté de médecine

Thèse présentée
en vue de l'obtention du grade de doctorat en biochimie

Avril 2018

© Samuel Tremblay-Belzile, 2018

Résumé

L'instabilité génomique se définit par l'apparition de modifications à la séquence d'un génome, qui peuvent varier de changements ponctuels d'une base à des réarrangements impliquant des grandes sections de chromosomes. Elle peut se produire suite à des dommages à l'ADN par des mécanismes de réparation sujets à l'erreur, mais également au cours des processus normaux du métabolisme de l'ADN. Certains réarrangements spécifiques sont associés à des maladies génétiques, tandis que des maladies comme le cancer affichent une instabilité génomique généralisée.

Dans le but d'étudier les mécanismes menant à l'instabilité génomique, nous avons mis au point une nouvelle approche de détection des réarrangements à partir de séquençage de nouvelle génération. Celle-ci se distingue des méthodes précédentes par sa sensibilité et sa résolution à la base près, qui permettent d'obtenir un portrait plus fidèle des modifications présentes dans le génome d'un organisme. Cette approche nous a permis de caractériser l'instabilité génomique présente dans le chloroplaste d'*Arabidopsis thaliana* et dans la mitochondrie humaine, découvrant ainsi la prépondérance des inversions à courte distance dans ces génomes. Le mécanisme menant à ce type de réarrangement, nommé demi-tour de réplication, est associé à l'interruption de la réplication par les fourches bloquées, et demeure encore peu compris. Nous avons déterminé à l'aide de lignées mutantes d'*A. thaliana* que les protéines de liaison à l'ADN simple-brin Whirly et les recombinases de la famille RecA sont toutes deux importantes pour empêcher ces réarrangements. Nous avons ensuite informatisé notre approche de façon à permettre l'analyse de plus grands génomes. Nous avons ainsi découvert la présence de demi-tours de réplication dans les génomes nucléaires de l'humain et de *Saccharomyces cerevisiae*. L'utilisation de mutants et de stress réplicatifs chez la levure nous ont permis de corréler l'apparition des demi-tours à la stabilisation des fourches de réplication bloquées.

Nos résultats ont permis de mettre en valeur l'importance des demi-tours de réplication dans l'instabilité génomique de plusieurs organismes. Bien qu'il s'agisse selon toute vraisemblance d'un mécanisme de redémarrage de la réplication sujet à l'erreur, le demi-tour de réplication n'implique aucun intermédiaire de bris double-brin. Nous proposons donc que le

demi-tour de réplication favorise également la résistance aux dommages à l'ADN lorsque la machinerie de recombinaison homologue n'est pas disponible.

Mots-clés : Réparation de l'ADN, réarrangements génomiques, fourches de réplication bloquées, bris double-brin, séquençage de nouvelle génération, chloroplastes, mitochondries, noyau.

Abstract

Genome instability is defined as a high mutation rate in an organism, ranging from single nucleotide variations to rearrangements involving large chromosome segments. These mutations can occur following the repair of DNA damage by error-prone mechanisms, or even as a result of normal DNA metabolism processes. Some rearrangements at specific loci are linked to genetic diseases, whereas diseases like cancer exhibit generalized genomic instability.

To study mechanisms leading to genome instability, we developed a new approach to detect DNA rearrangements from next-generation sequencing data. This approach differs from previous methods by its high sensitivity and base-pair resolution, which provide a more accurate portrait of the variations present in the genome of an organism. This has allowed us to characterize the genome instability present in the chloroplast of *Arabidopsis thaliana* and in the human mitochondrion. We thus discovered a high level of short-range inversions in these genomes, which can occur as a result of replication U-turns. U-turns are linked to replication fork stalling, but the steps and factors involved in this mechanism remain poorly understood. Using *A. thaliana* mutants, we have determined that the single-stranded DNA-binding Whirly proteins and the RecA family of recombinases both act to suppress replication U-turns. We then created a computer script to allow our approach to be used on larger genomes, and discovered the presence of U-turns in the nuclear genomes of humans and *Saccharomyces cerevisiae*. We compared DNA rearrangements in various yeast mutants under different conditions of replication stress, and established a correlation between stabilized stalled replication forks and replication U-turns.

Our results highlight the importance of replication U-turns as a mechanism of genome instability in the genomes of several organisms. While U-turns appear mainly as an error-prone mechanism to restart stalled replication forks, it is interesting to note that it does not involve any double-strand break intermediates. We therefore propose that U-turns may also favor resistance to DNA damage when the homologous recombination machinery is not available.

Keywords : DNA repair, genome rearrangements, stalled replication forks, double-strand breaks, next-generation sequencing, chloroplasts, mitochondria, nucleus.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

alt-EJ : Liaison d'extrémités alternative (*alternative end-joining*)

BIR : Réplication induite par les bris (*break-induced replication*)

CNV : Variation du nombre de copies (*copy number variation*)

DSBR : Réparation des bris double-brins (*double-strand break repair*)

HR : Recombinaison homologue (*homologous recombination*)

IR : Séquence inversée répétée (*inverted repeat*)

MMEJ : Liaison d'extrémités médiée par une microhomologie (*microhomology-mediated end-joining*)

NHEJ : Liaison non-homologue d'extrémités (*non-homologous end-joining*)

PARP1 : Polymérase à poly ADP-ribose 1 (*poly ADP-ribose polymerase 1*)

SDSA : Appariement dépendant de la synthèse d'ADN (*synthesis-dependent strand annealing*)

SMRT : Séquençage en temps réel de molécules uniques (*single-molecule real-time sequencing*)

SSA : Appariement simple-brin (*single-strand annealing*)

TLS : Synthèse d'ADN trans-lésion (*translesion synthesis*)

Remerciements

"The yarns of seamen have a direct simplicity, the whole meaning of which lies within the shell of a cracked nut. But Marlow was not typical (if his propensity to spin yarns be excepted), and to him the meaning of an episode was not inside like a kernel but outside, enveloping the tale which brought it out only as a glow brings out a haze, in the likeness of one of these misty halos that sometimes are made visible by the spectral illumination of moonshine."

-Joseph Conrad, *Heart of Darkness*

You sit down to write what you've already said, thoughts that you've already organised in your mind time and time again, and all that wants to come out are the things you don't already know, that you wish you could understand better. Those are the things worth writing. The things you write for yourself, and not for others. You sit down and you try to focus on what you need to write, but the need that made you sit down is relative. The need that emerges after you sit is absolute. You don't know which deserves priority, so you split your mind. You want to split it vertically to create left and right, but that doesn't work. The split happens horizontally and creates top and bottom. Top and bottom don't coexist well. Whatever gets pushed down wants to come up to the surface, and when it does, you write this. This has more meaning to you than that other thing. So you write this and wonder. You wonder if this might help give that meaning. If maybe this might be a part of that. So you decide that this *will* be part of that. This *will* give that meaning.

Those who know you will read this and see you.

It's no secret that you like to push form as far as you can make it go. When you started out, it felt like you may have been pushing substance as well, but now you see that you weren't. This is the perfect place to talk about meaning. That needs this to hold meaning to you, because that is just what you do, but this is who you are. You do what you do because you need to do something, but you are what you are no matter what you do. You do not exist in a vacuum, or what you are would hold no more meaning to you than the vacuum itself. You did not get to this

point in isolation. You got to this point because of those who helped bring out who you are, and this was the point you needed.

Those who know you will read this and see themselves.

Tradition wants you to call them by their name, but you don't like names. Names are placeholders for ideas, and you prefer to talk about the ideas without those other decorations. This may be harder to see as what it is without names, but this is not the place for clarity, because these ideas exist outside of clarity. Clarity was never the goal, and you enjoy the perfect ambiguity of it all. Ambiguity makes these ideas easier to communicate. It makes words denser. It helps break the silence that imposes itself when ideas flow faster than words usually do. It lets ideas flow even when you don't understand them. It says everything you want to say, even when you don't know what it is.

Those who know you will read this and see what it means.

1. Introduction

L'instabilité génomique peut se manifester de plusieurs façons, allant de modifications ponctuelles de bases dans la séquence d'ADN à des délétions affectant de grands segments de chromosomes. Ces changements peuvent se produire suite à des bris dans l'ADN, mais également pendant les processus essentiels au métabolisme de l'ADN, tels que la réplication et la transcription. Puisque certains de ces réarrangements de la séquence d'un génome sont associés à des maladies, l'étude de la réparation de l'ADN et des différents mécanismes qui mènent à l'instabilité génomique joue un rôle important pour notre compréhension des désordres génomiques (1). Ceci s'applique également à la stabilité du génome mitochondrial, qui possède sa propre machinerie de réplication et réparation (2, 3).

L'importance de la stabilité génomique s'étend également au-delà du noyau et de la mitochondrie, puisque le chloroplaste possède aussi un génome. Celui-ci encode entre autres plusieurs protéines qui participent à la photosynthèse, et sa stabilité est ainsi cruciale à la viabilité des plantes. En effet, plusieurs mutants chez lesquels le génome du chloroplaste présente des niveaux élevés de réarrangements montrent également un impact sur la photosynthèse. Dans certains cas, les feuilles possèdent des secteurs blancs, indiquant un défaut dans la biogénèse des chloroplastes (4–6). Puisque la photosynthèse est à la base du métabolisme de l'énergie chez les végétaux, des altérations à la fonction du chloroplaste peuvent avoir un impact négatif sur la croissance de la plante (7).

Les génomes des trois types d'organelles (noyau, mitochondrie et chloroplaste) ne sont généralement pas répliqués et réparés par les mêmes protéines, et ce au sein d'un même organisme. Cependant, les mécanismes impliqués dans le métabolisme de l'ADN suivent en grande partie les mêmes modèles, qui incluent des modes de réparation conservateurs et sujets à l'erreur (*error-prone*). Par conséquent, les méthodes utilisées pour étudier la stabilité génomique sont communes à tous les contextes biologiques. Les avantages et les limitations de chacune sont ainsi pris en considération davantage selon le mécanisme ou le réarrangement étudié que selon le génome d'intérêt.

1.1 Méthodes de détection des réarrangements génomiques

L'étude de la réparation de l'ADN se fait en grande partie par l'analyse des réarrangements génomiques qui se produisent en absence de protéines importantes au maintien de la stabilité génomique. Les méthodes permettant de détecter les séquences d'ADN ne correspondant pas au génome de type sauvage sont ainsi un facteur déterminant dans la portée des travaux sur le sujet. Quelques techniques biochimiques existent et possèdent chacune leurs limitations, mais le développement du séquençage de nouvelle génération a donné lieu à une nouvelle variété d'approches qui offrent de nombreux avantages. Les méthodes biochimiques demeurent cependant utiles pour observer des réarrangements spécifiques, confirmer des résultats obtenus par séquençage et obtenir des résultats rapidement.

1.1.1 Approches biochimiques et génétiques

1.1.1.1 Méthodes par hybridation

Le buvardage de Southern (*Southern blot*) est la méthode la plus simple et directe permettant de détecter des fragments d'ADN réarrangés. Pour cette raison, elle est utilisée autant en recherche fondamentale (8) qu'à des fins diagnostiques (9). Pour ce faire, l'ADN est digéré par des enzymes de restriction et séparé sur gel avant d'être transféré sur une membrane et hybridé à une ou plusieurs sondes permettant de suivre des fragments attendus spécifiques. L'apparition de bandes de tailles ne correspondant pas à des fragments attendus permet de déterminer qu'une altération au niveau de la séquence a eu lieu, tandis qu'une variation d'intensité peut révéler une amplification de la région du génome visée par la sonde (8).

Cette approche peut maintenant facilement être adaptée pour des études à haut débit à l'aide de micropuces, particulièrement pour observer des variations du nombre de copies de séquences génomiques (*Copy number variation*, CNV) (10). L'ajout de nouvelles sondes aux micropuces permet d'observer une quantité énorme de loci différents, couvrant le génome humain entier (Revu dans Aten et al. 2009). Bien que cette technologie fut d'abord développée avec l'objectif de détecter les sites de polymorphismes nucléotidiques (*Single-nucleotide polymorphisms*, SNP), elle est facilement adaptable à la détection de CNV. Les micropuces ne

conservent cependant pas toute l'information qui peut être obtenue par buvardage de Southern, puisqu'elles ne permettent pas de connaître précisément la taille des nouveaux fragments créés.

Combinée à la microscopie, l'hybridation *in situ* en fluorescence (*fluorescence* in situ *hybridization*, FISH) permet la détection de réarrangements chromosomiques de grands segments d'ADN, notamment lors de la détermination d'un caryotype (11). Pour cette raison, le FISH est particulièrement utilisé à des fins diagnostiques lorsqu'un réarrangement connu est associé à une pathologie (12). Il peut par contre être également utilisé en recherche, par exemple pour détecter l'amplification de séquences répétées ou d'un motif connu (13). Un des avantages majeurs de cette approche est qu'elle permet d'observer la séquence d'intérêt sur un seul chromosome à la fois, permettant d'étudier non seulement le réarrangement, mais également son contexte moléculaire. Pour cette raison, le FISH a également été adapté à l'étude de molécules d'ADN avec une plus grande résolution en étirant l'ADN en un long filament. Cette méthode est appelée FISH sur fibre (*fiber-FISH*) et permet le positionnement relatif de séquences spécifiques dans le contexte de molécules uniques (14, 15).

1.1.1.2 Méthodes par amplification

Depuis son développement, l'amplification en chaîne par polymérase (PCR) demeure la méthode la plus rapide pour détecter des réarrangements. Cet avantage majeur en fait un outil important aussi bien à des fins diagnostiques qu'en recherche (8, 9). Comme pour les méthodes par hybridation, il est cependant nécessaire de connaître au moins une partie de la séquence d'intérêt afin de produire les amorces. Par contre, une paire d'amorces peut détecter plusieurs réarrangements différents, tant que ceux-ci se trouvent à des positions génomiques assez proches. L'ADN amplifié peut ensuite être cloné et séquencé afin d'obtenir une résolution à la base près de la séquence réarrangée. L'amplification est également utile pour détecter des réarrangements rares, puisqu'une seule molécule réarrangée est suffisante pour obtenir un produit. Ceci est également un inconvénient, puisque la mesure de la fréquence d'un réarrangement dans l'échantillon ne peut être que qualitative. Certaines approches, telles que le PCR quantitatif (qPCR) et le PCR numérique (*digital PCR*) permettent de contourner cette limitation en offrant une mesure quantitative du nombre de molécules amplifiées. Le PCR

numérique a ainsi été utilisé pour la détection de translocations rares, ainsi qu'à des fins diagnostiques (16, 17).

Grâce à sa rapidité, le PCR a également été incorporé à des approches visant un plus haut débit : l'hybridation multiplex de sondes amplifiables (*multiplex amplifiable probe hybridization*, MAPH) et l'amplification multiplex de sondes ligation-dépendente (*multiplex ligation-dependent probe amplification*, MLPA). Ces approches combinent la simplicité des méthodes par hybridation et par PCR afin de cibler plusieurs sites simultanément dans un génome, tout en minimisant les coûts et le temps d'attente (18). Elles sont particulièrement adaptées à la détection de sites de CNV, mais ne confèrent aucune information additionnelle sur les réarrangements impliqués. Pour cette raison, le MAPH et le MLPA ont été particulièrement utiles à des fins diagnostiques, mais possèdent peu d'avantages par rapport aux micropuces dans le contexte de la recherche.

1.1.1.3 Systèmes rapporteurs

Contrairement à la majorité des approches mentionnées ci-haut, l'utilisation de systèmes rapporteurs ne repose pas uniquement sur la détection de réarrangements endogènes, mais sur la manipulation des conditions qui mènent à l'instabilité génomique. Pour cette raison, il s'agit d'un outil puissant en recherche pour étudier les critères spécifiques qui sont requis pour qu'un mécanisme donne lieu à un réarrangement (19, 20). Puisqu'il s'agit d'une méthode *in vivo*, il est possible d'observer l'impact de modifications au niveau de la séquence du système, mais également de différents stress qui peuvent être appliqués aux cellules. Des protéines spécifiques peuvent également être enlevées ou ajoutées pour déterminer leur importance dans le mécanisme. L'inconvénient principal de cette approche est qu'elle ne cible en général qu'un seul site du génome, rendant difficile la généralisation du mécanisme étudié à un génome entier. Il peut aussi être difficile d'être certain du mécanisme qui donne lieu au réarrangement observé dans le contexte biologique, puisque la présence du rapporteur indique seulement la présence du réarrangement final, et non son mécanisme.

1.1.2 Approches bio-informatiques

Depuis le développement du séquençage de nouvelle génération, les logiciels développés pour identifier des variations de séquence se sont rapidement multipliés. Malgré l'amélioration constante du pouvoir de détection, chaque outil présente des avantages qui lui sont propres. À cause de cela, aucun logiciel unique ne s'est établi en tant que standard dans le domaine (21). Il serait ainsi laborieux de décrire en détail le fonctionnement de tous les outils existants. Par contre, ceux-ci utilisent tous une ou plusieurs de 4 approches générales pour identifier des réarrangements: l'assemblage *de novo*, l'analyse de profondeur de séquençage, l'identification de paires aberrantes, ou l'analyse par séquences coupées.

1.1.2.1 Assemblage *de novo* de génomes

Conceptuellement, la méthode la plus simple pour identifier des variations dans un génome est d'obtenir le génome complet de deux individus et de les comparer (Figure 1A). Le séquençage de nouvelle génération, maintenant suffisamment rapide et abordable pour permettre l'assemblage de génomes entiers, a ouvert la porte à ce genre d'approches. Cependant, le temps nécessaire à l'assemblage d'un génome entier rendait initialement cette méthode difficilement envisageable pour des études de génomes aussi grands que celui de l'humain (22). Les plus petits génomes, tel que celui de la mitochondrie de souris, ont rapidement pu être étudiés par assemblage *de novo* de mutants (23). Depuis, des approches ont été développées permettant d'assembler le génome humain relativement rapidement en tirant profit des ressources informatiques disponibles pour la recherche (24). De telles méthodes peuvent également s'appliquer à un contexte clinique, particulièrement en limitant l'assemblage à des régions spécifiques du génome et en intégrant cette approche à d'autres méthodes (25, 26). Les résultats sont ainsi obtenus avec une résolution à la base près, ce qui offre de l'information à propos des mécanismes donnant lieu aux réarrangements. Par contre, l'assemblage de génomes est surtout adapté pour la détection des variations somatiques homozygotes ou hétérozygotes, mais est beaucoup moins efficace pour la détection de réarrangements rares.

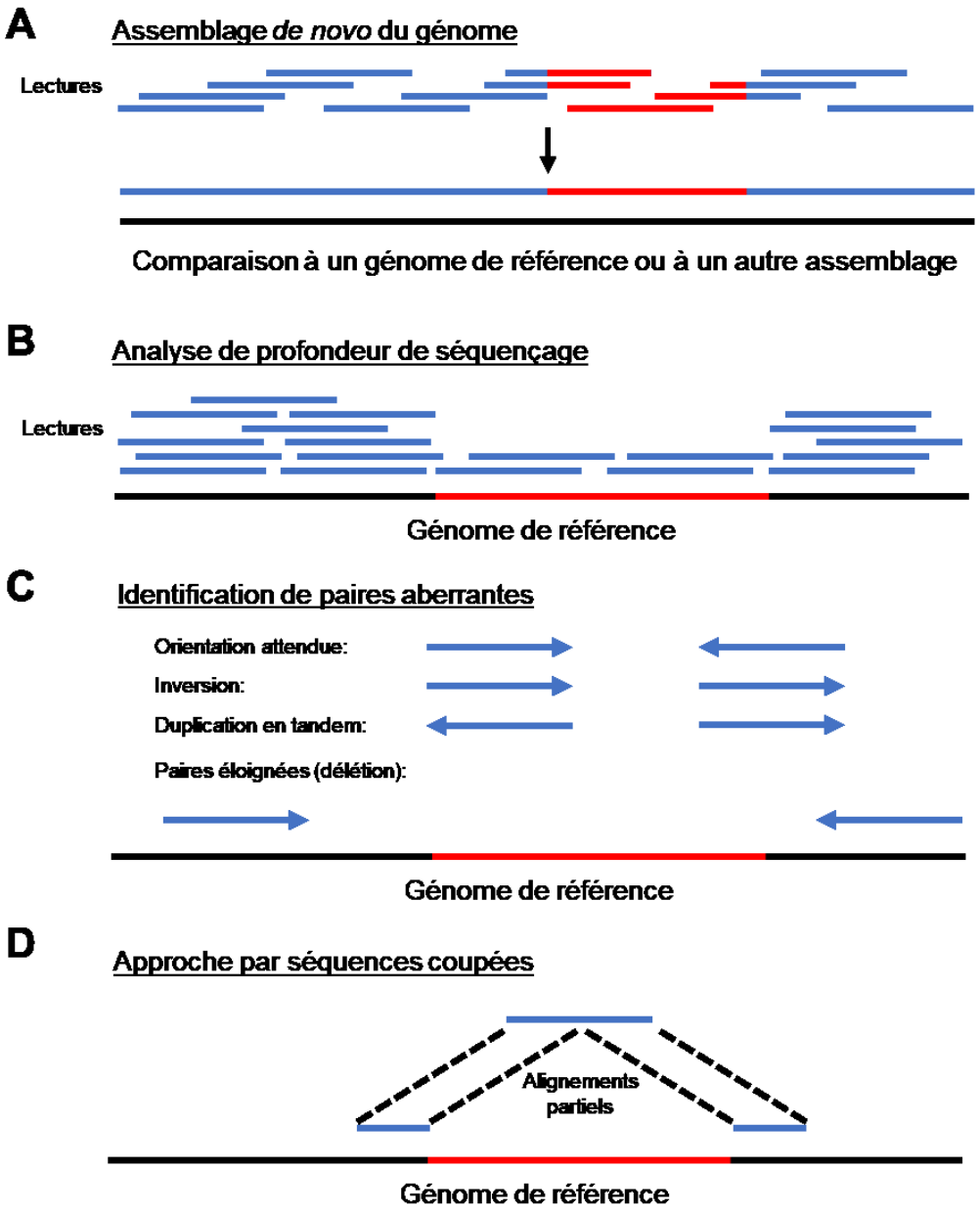


Figure 1. Approches bio-informatiques de détection des réarrangements.

Les traits bleus représentent les lectures de séquençage. Les traits rouges représentent une variation par rapport au génome de référence. Les traits noirs représentent le génome de référence.

1.1.2.2 Analyse de profondeur de séquençage

Une autre approche intuitive pour la détection de variation génomique est l'utilisation de la profondeur de séquençage (*read depth*) pour trouver des segments dupliqués ou supprimés (Figure 1B). Puisque les segments séquencés sont en principe retrouvés proportionnellement à leur fréquence dans le génome, il est ainsi possible d'associer des changements à la profondeur de séquençage à des événements de CNV (22). À cause de possibles biais dus à un trop faible taux d'échantillonnage, cette approche gagne en fiabilité quand la profondeur moyenne obtenue augmente. Un aspect intéressant de cette méthode est qu'elle est surtout efficace pour détecter des événements de CNV sur de très longs segments d'ADN, ce qui va à l'inverse des autres approches par séquençage. Pour cette raison, l'analyse de profondeur est souvent utilisée en combinaison avec d'autres approches (27–29). Cette intégration de méthodes permet également de compenser pour la faible résolution obtenue par la profondeur de séquençage, qui ne permet habituellement pas de voir un réarrangement à la base près, et confère ainsi peu d'information mécanistique. Une autre limitation importante de la profondeur de séquençage est qu'elle n'apporte aucune information supplémentaire pour les réarrangements dits balancés, dans lesquels des segments de séquence sont déplacés dans le génome, mais ne sont ni amplifiés ni perdus.

1.1.2.3 Approches par identification de paires aberrantes

Étant donné les limitations majeures de l'assemblage *de novo* et de l'analyse de profondeur de séquençage, l'une des méthodes les plus répandues depuis le développement du séquençage de nouvelle génération est l'identification de paires aberrantes (*paired-end analysis*) (Figure 1C) (22, 29–33). La préparation d'un échantillon d'ADN pour le séquençage, appelé librairie, commence dans la plupart des cas par la fragmentation du génome en courts segments d'environ 200 à 500 paires de bases. Le séquençage peut alors avoir lieu à partir d'une seule extrémité (*single-end*) ou des deux extrémités (*paired-end*). L'obtention des deux extrémités offre de nombreux avantages, puisqu'elle permet de comparer la position relative de ces extrémités sur le génome de référence. Ainsi, sans nécessairement connaître toute la séquence du fragment de départ, il est possible d'identifier la présence de réarrangements. En effet, si l'orientation relative de l'une des extrémités est contraire à ce qui est attendu, un événement

d'inversion doit avoir eu lieu dans la séquence intermédiaire non-séquencée (30). Si par contre l'orientation relative des deux séquences est inversée, cela peut indiquer la présence d'une duplication en tandem dans le génome. Similairement, si les séquences sont dans la bonne orientation, mais à une distance plus grande que la longueur attendue des fragments, la séquence intermédiaire doit contenir une délétion. Finalement, des translocations ou insertions peuvent également être détectées si les séquences des deux extrémités se retrouvent sur des chromosomes différents. Comme dans le cas de l'analyse de profondeur de séquençage, ces approches ne permettent cependant pas d'obtenir la séquence précise à la jonction du réarrangement, et ne permettent pas d'identifier le mécanisme y ayant donné lieu. L'analyse des paires aberrantes possède ainsi une utilité limitée pour identifier de nouvelles variations génétiques, mais offre néanmoins beaucoup de puissance pour identifier des régions de génomes contenant des réarrangements.

1.1.2.4 Approches par séquences coupées

La dernière approche consiste à utiliser les séquences qui ne peuvent pas être alignées sur le génome de référence afin d'identifier des jonctions de réarrangements. Il existe une grande variété d'outils et de logiciels utilisant cette méthode, communément appelée analyse par séquences coupées (*split reads*) (22). Comme avec l'assemblage *de novo*, les résultats obtenus à l'aide de séquences coupées possèdent une résolution à la base près. Ces approches sont par contre généralement moins prohibitives au niveau des ressources informatiques, puisque l'analyse se limite aux séquences qui ne peuvent pas être alignées au génome de référence. De plus, beaucoup d'outils combinent l'analyse par séquences coupées à l'une ou plusieurs des approches précédentes afin de réduire davantage la charge computationnelle.

DELLY est l'un des premiers outils à utiliser avec succès une analyse par séquences coupées, et a souvent servi de point de comparaison pour la mise au point de nouvelles méthodes (30). DELLY effectue premièrement une analyse par paires aberrantes telle que décrite ci-haut, identifiant et triant les sites de réarrangement par leur plus petite position en considérant les chromosomes en ordre alphanumérique. Les paires indiquant potentiellement le même réarrangement sont ainsi groupées avant d'être analysées par séquences coupées. Pour chaque position, DELLY identifie des paires de séquences pour lesquelles une extrémité est alignée

proche du réarrangement et l'autre n'est pas alignée. Pour simplifier l'étape d'alignement de sous-séquences, DELLY modifie la région du génome de référence correspondant au site du réarrangement de façon à créer une signature de délétion. La séquence du réarrangement est ainsi toujours initialement identifiée comme une délétion, avant d'être retransformée selon les positions véritables du génome de référence.

Tout comme DELLY, LUMPY utilise également une combinaison de paires aberrantes et de séquences coupées (32). L'analyse effectuée par LUMPY peut cependant également se limiter à l'une ou l'autre des deux approches, bien que son rendement soit grandement diminué de cette façon. LUMPY utilise presque exactement la même méthode que DELLY pour la détection de paires aberrantes, mais l'information obtenue à cette étape n'est pas utilisée pour l'analyse par séquences coupées. Cette seconde analyse est plutôt effectuée à partir des données de logiciels existants pour aligner une séquence en plusieurs fragments, tels que YAHA (34), BWA-SW et BWA-MEM (35). L'addition des résultats obtenus par les deux méthodes confère une sensibilité accrue, mais la différence de résolution entre les deux approches limite l'utilisation des données pour des études mécanistiques.

Étant donné cette limitation, certains outils optent d'utiliser uniquement une approche par séquences coupées afin d'obtenir la meilleure résolution possible pour tous les réarrangements. C'est le cas de Gustaf, qui accepte des séquences non-alignées ou avec des alignements partiels (36). Dans le cas de séquences brutes, Gustaf utilise l'aligneur Stellar pour obtenir des alignements partiels pour l'analyse. Gustaf identifie ensuite les alignements partiels considérés adjacents dans les séquences, et construit un « graphique de séquences coupées » (*split-read graph*) afin de déterminer un coût pour ce réarrangement. Cette approche considère ainsi la distance entre les séquences et le type de réarrangement avant d'identifier le réarrangement le plus probable, ce qui peut favoriser certains types de réarrangement et biaiser les résultats.

À l'opposé, l'outil Wham combine une analyse par paires aberrantes et plusieurs approches par séquences coupées pour identifier des sites de réarrangement à la base près. Wham commence par observer toutes les séquences alignées pour trouver celles pour lesquelles quelques bases ne sont pas alignées à une extrémité (*soft clipped bases*). Un site de réarrangement est identifié quand au moins 3 séquences alignées s'arrêtent à la même position

du génome. Les séquences non-alignées sont ensuite analysées par BWA-MEM pour permettre l'alignement en plusieurs fragments (35). L'information obtenue par cet alignement est quant à elle utilisée pour associer deux sites de réarrangements déjà identifiés à l'étape précédente. Wham peut ainsi identifier des allèles mutants avec plusieurs alignements pointant vers le même réarrangement, que celui-ci soit homozygote ou hétérozygote.

Une dernière approche différente par séquences coupées consiste à cibler spécifiquement un certain type de réarrangement. Puisque les approches existantes déjà décrites sont généralement moins efficaces pour détecter les inversions par rapport aux autres types de réarrangement, deux outils ont été publiés la même année pour spécifiquement identifier des inversions. Pour ce faire, SRinversion, comme plusieurs des outils décrits ci-haut, utilise les séquences non-alignées et faiblement alignées pour amorcer la recherche d'inversions (37). À la différence des autres, SRinversion considère spécifiquement la possibilité qu'une séquence contienne entièrement une inversion, de façon à ce que les deux extrémités s'alignent dans la direction opposée par rapport au milieu. Les inversions détectées sont ensuite compilées pour compter le nombre de séquences supportant chaque inversion. Micro-Inversion Detector (MID) utilise une approche similaire, commençant également avec les séquences non-alignées (38). MID aligne d'abord une extrémité de la séquence et l'utilise comme ancre dans le génome de référence pour chercher des micro-inversions avoisinantes. Les différentes possibilités obtenues sont ensuite comparées pour trouver celle possédant le meilleur score, tenant en compte la possibilité de séquences palindromiques dans le génome. SRinversion et MID visent ainsi des réarrangements qui étaient précédemment souvent manqués, mais le biais intrinsèque à ces outils les force à être très stricts dans le processus d'identification.

1.2 Les mécanismes de réparation des bris d'ADN

Les mécanismes utilisés pour réparer un bris à l'ADN ne peuvent être observés dans la séquence résultante que s'ils provoquent un changement à celle-ci. Ainsi, dans l'étude des réarrangements génomiques, les mécanismes de réparation sont habituellement séparés en deux catégories : les mécanismes conservateurs, qui recréent parfaitement la séquence de départ, et les mécanismes sujets à l'erreur, qui réparent le squelette et la structure de la molécule d'ADN, mais peuvent introduire des modifications à la séquence en cours de route. Les facteurs exacts

qui déterminent le mode de réparation utilisé ne sont pas encore bien compris, mais les évidences actuelles semblent indiquer que le choix de mécanisme est effectué par les protéines qui lient initialement l'extrémité de l'ADN (39). Bien que des erreurs puissent se produire lors de mécanismes conservateurs, les étapes par lesquelles ceux-ci procèdent permettent généralement de retrouver la séquence initiale. De la même manière, les mécanismes sujets à l'erreur peuvent à l'occasion mener à une réparation conservatrice, malgré l'absence d'étapes assurant un tel résultat. Plusieurs des études des mécanismes suivants ont été effectuées dans *Saccharomyces cerevisiae*, mais la nomenclature des protéines humaines est utilisée pour simplifier le texte, sauf en cas de mention contraire.

1.2.1 Les mécanismes conservateurs

1.2.1.1 La recombinaison homologue

Sans doute le mécanisme de réparation de l'ADN le mieux étudié, la recombinaison homologue (HR) est tellement fortement associée à la réparation conservatrice que sa forme classique est également simplement appelée « réparation des bris double-brins » (*double-strand break repair*, DSBR) (Figure 2) (39, 40). Dans cette voie, les deux extrémités double-brins autour d'un bris sont prises en charge par un complexe protéique avec une activité nucléase qui génère des extrémités 3' simple-brins (40). Des études chez la levure suggèrent que cette première étape résulte en de courtes extrémités simple-brins d'environ 300 bases, et permet ensuite une résection plus extensive (41). Chez l'humain, cette résection initiale est effectuée par le complexe MRE11-RAD50-NBS1-CtIP, dans lequel MRE11 possède l'activité endonucléase et exonucléase (42). Parmi les cofacteurs impliqués, CtIP semble avoir un rôle dans la coordination des mécanismes de réparation en fonction du cycle cellulaire (43), permettant ainsi la suppression de la HR en phase G1 alors qu'aucune chromatide sœur n'est disponible. Cette première étape de résection ouvre ensuite la porte à l'exonucléase EXO1, qui procède à la résection de l'extrémité 5' pour étendre le simple-brin 3' jusqu'à plusieurs kilobases (42). Cette étape de résection peut également être effectuée par l'exonucléase DNA2, qui est aidée par la présence de la protéine RPA sur l'ADN simple-brin (44).

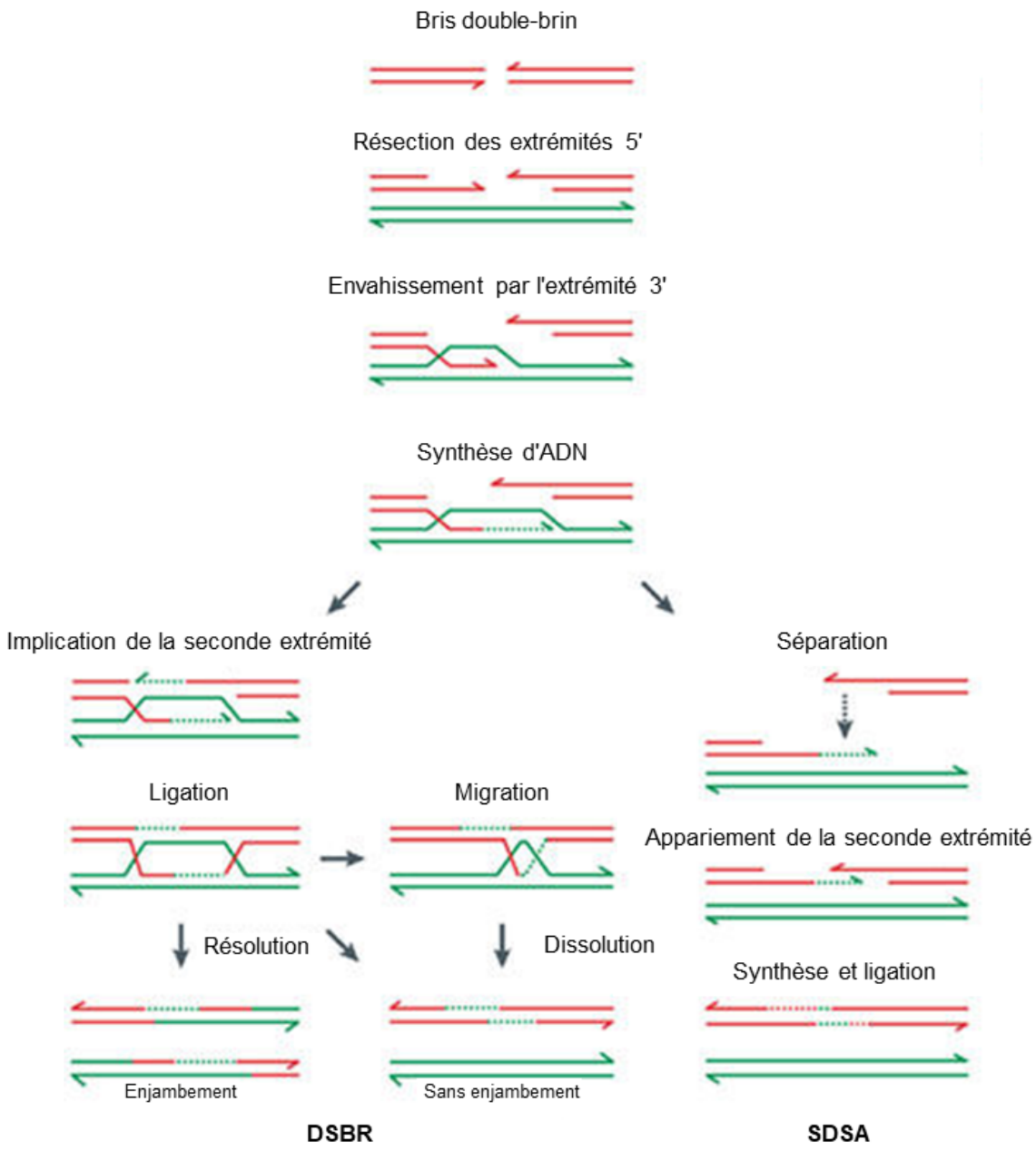


Figure 2. Étapes principales des voies DSBR et SDSA de la HR.

Les brins rouges représentent la chromatide ayant subi le bris, et les brins verts la chromatide sœur qui sert à la réparation. Les flèches représentent les extrémités 3' des brins. Figure adaptée de (59).

Suite à la résection, l'extrémité 3' simple-brin est utilisée pour rechercher une séquence homologue et procéder à la recombinaison. Ceci commence par le remplacement de RPA sur l'ADN simple-brin par un filament de RAD51 pour former le complexe présynaptique (45). La protéine RAD51 est l'homologue chez l'humain de la protéine bactérienne RecA, pour laquelle des orthologues sont impliquées dans la recombinaison homologue chez les procaryotes, archées et eucaryotes. Selon le modèle établi chez la levure, RAD51 procède de manière ATP-dépendante à la recherche d'homologie en défaisant l'appariement d'ADN double-brin pour permettre l'échange de brins (45). Malgré la caractérisation de plusieurs des acteurs impliqués, le mécanisme exact par lequel cette recherche procède est longtemps resté incompris. Il a finalement été montré, à partir de la protéine bactérienne RecA, que le complexe présynaptique effectue plusieurs contacts simultanés avec les molécules d'ADN double-brins situées à proximité (46). La vitesse à laquelle de nouveaux contacts sont observés est proportionnelle à la longueur de l'extrémité 3' simple-brin, et inversement proportionnelle à la distance entre les extrémités de la molécule cible. Ceci suggère que le complexe présynaptique utilise plusieurs contacts faibles avec l'ADN présent dans l'espace tridimensionnel avoisinant pour initier la recombinaison avec une séquence homologue (46). Ces contacts transitoires forment des complexes synaptiques, et impliquent également d'autres régulateurs positifs et négatifs qui assurent que la recombinaison a lieu avec la bonne séquence (39).

Une fois l'homologie trouvée, l'extrémité simple-brin envahit le duplex d'ADN pour former la boucle de déplacement (*displacement loop* ou *D-loop*), stabilisée par l'association de RPA au brin d'ADN déplacé (47). Cette étape de la recombinaison est également appelée complexe postsynaptique, et requiert la dissociation du filament de RAD51 par RAD54 selon un système utilisant les protéines de levure (48). L'extension de l'extrémité 3' a ensuite lieu en utilisant le brin complémentaire du duplex envahi comme matrice. Toujours selon des études utilisant les protéines de levure, la réplication implique la protéine de réplication PCNA, l'hélicase Pif1, et la polymérase δ (49, 50). Bien que la synthèse d'ADN à partir d'une chromatide sœur permette de retrouver exactement la séquence initiale du brin ayant subi un bris, la HR présente un taux de mutation plus élevé que la réplication classique (51). Dans la voie de DSBR, la seconde extrémité du bris double-brin s'apparie au brin d'ADN laissé seul pendant la formation de la boucle de déplacement. Chez la levure, cette étape est dépendante de

la protéine Rad52 (52), mais ceci ne semble pas être le cas chez l'humain avec la protéine RAD52 (53). Il est donc proposé que le DSBR chez l'humain procède par une seconde invasion du duplex d'ADN homologue par l'autre extrémité du bris (40).

Une fois la synthèse d'ADN complète suivant l'appariement des deux extrémités du bris, les deux duplex d'ADN sont croisés dans des structures appelées jonctions de Holliday, et doivent être dissociés pour donner lieu à deux molécules d'ADN indépendantes. Deux mécanismes principaux ont été observés pour ce faire : la dissolution et la résolution. La résolution consiste à couper deux des quatre brins impliqués dans le croisement des deux duplex. Chez l'humain, cette réaction est catalysée à différents moments du cycle cellulaire par la résolvase MUS81-EME1 en complexe avec SLX1-SLX4 ou par la résolvase GEN1 (54, 55). Parce que la résolvase peut cliver les brins selon deux orientations différentes, la résolution peut donner lieu à un enjambement (*crossover*) entre les deux molécules d'ADN, après lequel les duplex d'ADN de part et d'autre du site du bris sont échangés. Ceci fait en sorte qu'un côté de la molécule brisée devient associée à l'autre côté de la molécule matrice et vice-versa. La dissolution, par contre, a lieu par l'intermédiaire de l'hélicase BLM et de la topoisomérase IIIα, et ne donne jamais lieu à l'enjambement des deux molécules (56).

1.2.1.2 La réplication induite par les bris

La réplication induite par les bris (*break-induced replication*, BIR) est un mécanisme dérivé de la recombinaison homologue qui partage ses étapes de résection, recherche d'homologie, envahissement d'un duplex d'ADN homologue et de synthèse d'ADN (39). À la différence de la voie de DSBR, la BIR n'implique qu'une seule extrémité double-brin plutôt que deux (Figure 3) (57). Une fois la synthèse d'ADN amorcée dans la boucle de déplacement, elle continue jusqu'à l'extrémité du chromosome, résultant ainsi en une copie complète du chromosome à partir d'un fragment avec une extrémité double-brin. Il s'agit donc de la voie privilégiée pour réparer l'extrémité double-brin seule qui résulte quand une fourche de réplication rencontre un bris simple-brin dans la matrice. Cependant, la réplication pendant la BIR se distingue de la réplication normale par l'utilisation d'une bulle de réplication, et accumule ainsi davantage de mutations (58).

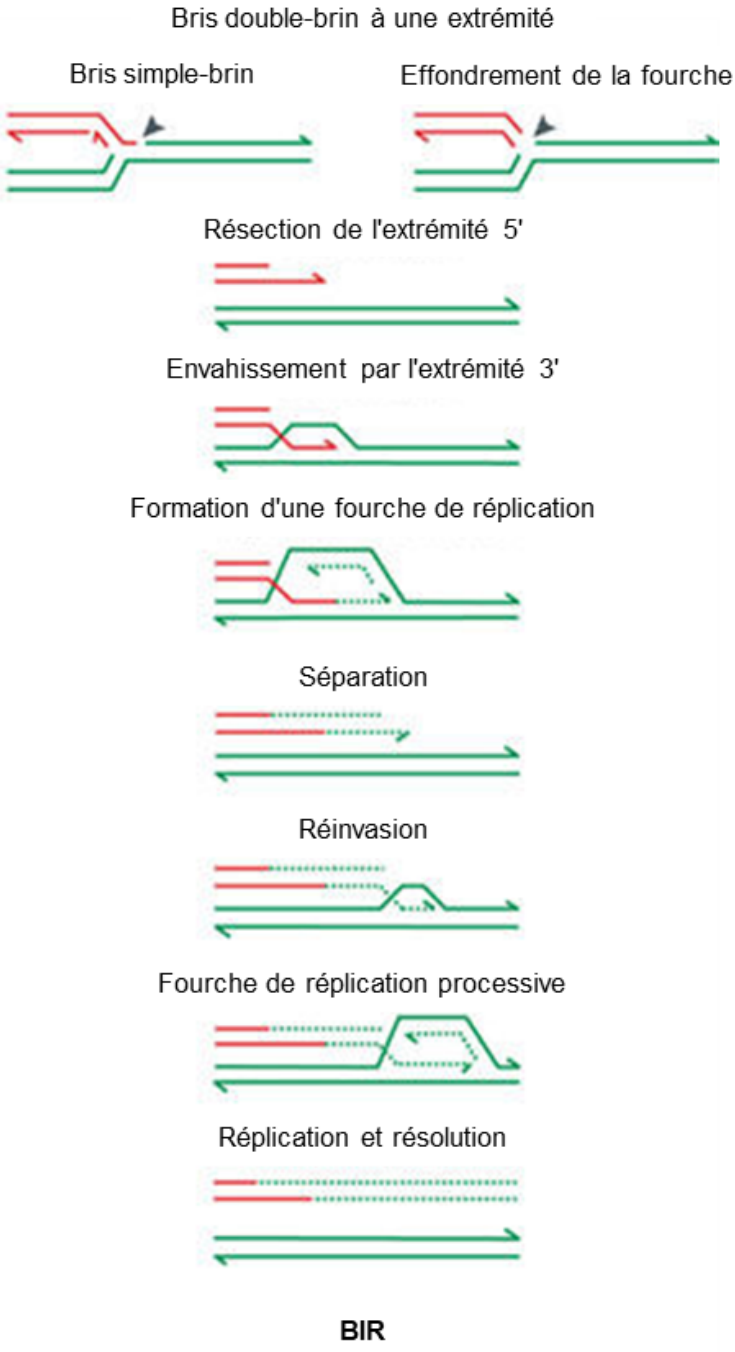


Figure 3. Étapes principales de la BIR.

Les brins rouges représentent la chromatide ayant subi le bris, et les brins verts la chromatide sœur qui sert à la réparation. Les flèches représentent les extrémités 3' des brins. Figure adaptée de (59).

1.2.1.3 L'appariement dépendant de la synthèse d'ADN

Comme la BIR, l'appariement dépendant de la synthèse d'ADN (*synthesis-dependent strand annealing*, SDSA) partage toutes ses étapes initiales avec la voie de DSBR, jusqu'à la formation de la boucle de déplacement (Figure 2) (39). Lorsque la synthèse d'ADN progresse assez loin sur le chromosome homologue, il est possible pour le brin naissant de se dissocier de la matrice et de s'apparier avec l'extrémité 3' simple-brin de la seconde partie du bris. Cet appariement permet à la synthèse d'ADN d'avoir lieu à partir des deux extrémités appariées et de reformer la molécule initiale. Puisque cette voie ne génère pas de jonctions de Holliday, elle n'implique jamais l'enjambement des deux molécules (60). Il s'agit de la différence principale entre les voies de SDSA et DSBR, qui mènent toutes les deux à la réparation conservatrice du bris. Le choix entre les deux voies est ainsi principalement déterminé par des facteurs qui stabilisent la boucle de déplacement et favorisent la DSBR ou déstabilisent la boucle pour permettre le SDSA (60). Chez la levure, les hélicases Mph1, Sgs1 et Srs2 contribuent toutes les trois à la voie de SDSA par des mécanismes différents (60). Chez l'humain, l'hélicase BLM promeut la dissolution d'intermédiaires de recombinaison, suggérant ainsi un rôle dans le choix entre SDSA et DSBR (61). Des expériences en cellules humaines ont également montré que l'hélicase RTEL1 semble être un homologue de fonction de Srs2 chez la levure, et participe également à la favorisation du SDSA (62).

1.2.2 Les mécanismes sujets à l'erreur

1.2.2.1 La recombinaison homologue non-allélique

La recombinaison homologue non-allélique (*non-allelic homologous recombination*, NAHR) n'est pas en soi un mécanisme distinct de la HR. Elle implique les mêmes acteurs protéiques et les mêmes étapes mécanistiques, mais a lieu par l'intermédiaire de séquences homologues qui ne correspondent pas au même locus du génome. Puisque la machinerie impliquée dans la HR requiert de longues séquences complémentaires, la NAHR est un mécanisme qui est relié à l'architecture des génomes plutôt qu'aux protéines qui prennent en charge la réparation (63). La NAHR est surtout favorisée par de longues séquences répétées qui se retrouvent dans les génomes, et mène surtout à des délétions lorsque les séquences sont présentes sur le même chromosome (1). Au contraire, la NAHR entre deux chromatides

différentes peut mener à des duplications ou délétions (1). Ces événements requièrent par contre des séquences d'homologie ininterrompues d'environ 200 bases, et des variations d'un seul nucléotide peuvent nuire à la NAHR (64). Étant donné cette spécificité associée à la machinerie de la HR, les réarrangements formés sont souvent récurrents entre des régions précises du génome et peuvent être associés à des pathologies (65).

1.2.2.2 Les mécanismes de recombinaison par microhomologies

Bien que l'utilisation de la machinerie de HR requiert de longs segments ininterrompus d'homologie, des événements de recombinaison sont également observés entre des locus avec une identité de séquence imparfaite (66). De tels événements sont parfois considérés comme de la recombinaison homéologue, qui suit probablement les mêmes étapes que la HR, sans que les protéines utilisées dans la recherche d'homologie soit présentes (1). Contrairement aux événements ayant lieu par NAHR, ces réarrangements sont considérés non-récurrents, puisque l'absence d'implication de la machinerie de HR fait en sorte qu'ils sont beaucoup moins dépendants d'éléments spécifiques dans l'architecture du génome.

Également, des mécanismes similaires à ceux de la HR ont été proposés à partir de l'appariement de courtes séquences homologues, appelées microhomologies (59). Une étude en levure a montré que des réarrangements utilisant des microhomologies se produisaient en l'absence de Rad52, mais pas en l'absence de la polymérase Pol32 (67). Puisque Pol32 est impliquée dans la BIR et que l'absence de Rad52 suggère qu'il ne s'agit pas d'un processus lié à la HR, un mécanisme distinct de réplication induite par un bris médiée par une microhomologie (*microhomology-mediated break-induced replication*, MMBIR) a été proposé. Bien que les modèles initiaux impliquent l'envahissement d'un duplex d'ADN par une courte séquence de microhomologie, ce modèle a été remis en doute (59). Étant donné l'absence de machinerie connue permettant de former un complexe synaptique à partir d'une microhomologie, il est plutôt proposé qu'un mécanisme de MMBIR ait lieu par l'appariement de l'extrémité 3' d'un bris ou d'un brin naissant à une microhomologie de l'ADN déjà simple-brin dans l'espace environnant (59). La synthèse d'ADN pourrait alors continuer comme dans la BIR, sans avoir impliqué de protéines pour la recherche d'homologie et la formation d'un complexe synaptique. Ce mécanisme a été proposé pour expliquer des duplications, mais il peut

également être à la base de réarrangements complexes impliquant également des inversions (59, 67).

1.2.2.3 L'appariement simple-brin

Lorsqu'un site de bris est flanqué de part et d'autre par des séquences homologues, un mécanisme a été observé permettant la réparation du bris par l'appariement des deux extrémités. Ce mécanisme, nommé appariement simple-brin (*single-strand annealing*, SSA), s'apparente à la HR, puisqu'il partage les étapes initiales de résection des extrémités (Figure 4) (39). Il s'en distingue toutefois par le fait qu'il ne restore pas la séquence initiale, et résulte plutôt en une délétion de la séquence entre les deux homologies responsables de l'appariement. L'appariement des séquences est pris en charge par Rad52 et implique également RPA (68). Les extrémités dépassant la séquence homologue sont ensuite clivées par l'endonucléase XPF/ERCC1 avec RAD52 comme cofacteur (69). La réparation est complétée par la synthèse d'ADN et la ligation des brins.

1.2.2.4 La liaison non-homologue d'extrémités classique

Les mécanismes axés sur la recombinaison sont généralement mis en opposition aux mécanismes de liaisons d'extrémités, dont la voie la plus étudiée est appelée liaison non-homologues d'extrémités (*non-homologous end-joining*, NHEJ) (Figure 4). Contrairement à la recombinaison, la liaison d'extrémités ne requiert pas la présence d'une autre molécule d'ADN pour servir de matrice lors de la réparation. Pour cette raison, les voies de liaison d'extrémités sont favorisées dans les phases du cycle cellulaire pendant lesquelles une chromatide sœur n'est pas disponible pour la recombinaison (70). Lors d'un bris double-brin, le choix entre ces deux types de réparation dépend des protéines qui lient les extrémités formées, puisque la résection des extrémités, nécessaire à la recombinaison, inhibe la NHEJ (39). Le modèle classique de NHEJ implique seulement la reconnaissance des extrémités par les protéines Ku70 et Ku80 (ensemble appelées Ku), qui recrutent la ligase d'ADN IV (*DNA ligase IV*) et son cofacteur XRCC4 pour effectuer la ligation des extrémités du bris (71). La machinerie de NHEJ est par contre flexible, et peut procéder à la réparation en l'absence de Ku, même si les extrémités ne sont pas parfaitement compatibles (72).

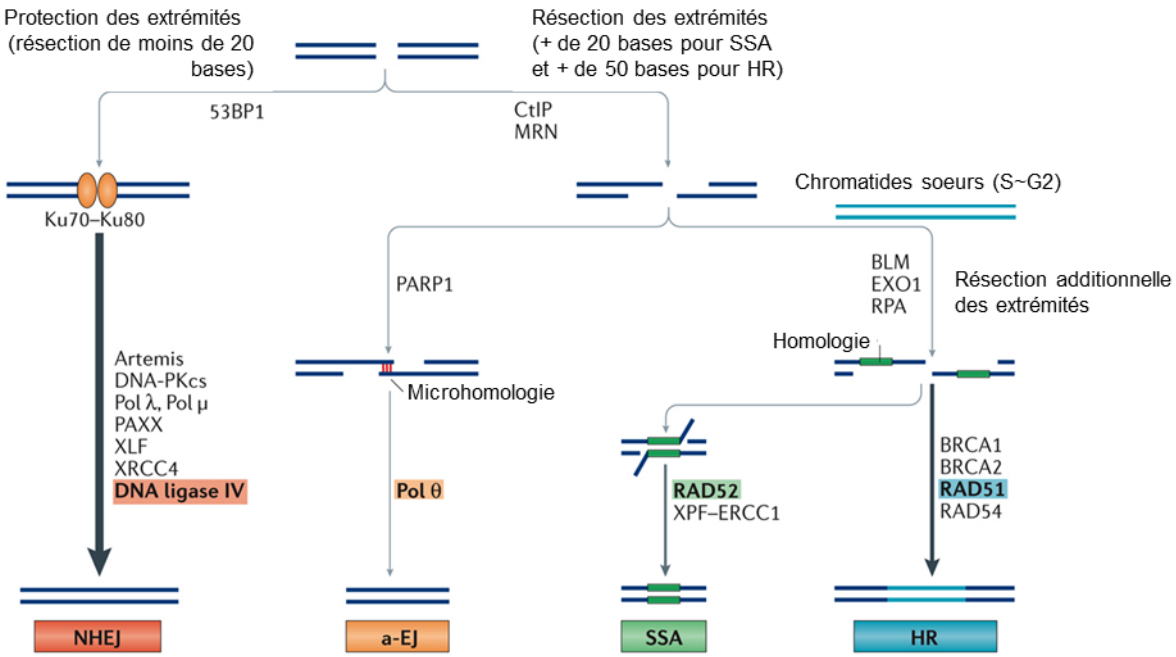


Figure 4. Principaux mécanismes de réparation d'un bris double-brin.
Les flèches représentent les principales étapes de chaque mécanisme, avec les protéines impliquées identifiées à côté. Les protéines les plus étroitement liées à chaque mécanisme sont surlignées de la couleur du mécanisme. Figure adaptée de (110).

Les protéines Ku aident cependant à la réparation efficace des extrémités, notamment en recrutant la sous-unité catalytique de la protéine kinase dépendante d'ADN (*DNA-dependent protein kinase catalytic subunit*, DNA-PKcs) pour former le complexe DNA-PK (73). Cette activité kinase stimule entre autres l'activité de l'endonucléase Artemis, qui clive les extrémités simple-brins et facilite la réparation des extrémités (74). Ku peut également recruter des polymérases au site du bris pour produire des extrémités compatibles avec le complexe XRCC4-DNA ligase IV. Pol μ est une polymérase capable d'allonger une extrémité 3' sans nécessiter de brin matrice (75). Ceci permet de créer de courtes extrémités 3' qui permettent l'appariement des deux extrémités. Pol λ effectue principalement la synthèse d'ADN à partir d'un brin matrice, permettant de compléter des gaps une fois les brins appariés (75). Le recrutement de nucléases et de polymérases au site du bris offre une grande flexibilité à la NHEJ, et contribue à son statut en tant que mécanisme de réparation sujet à l'erreur. La délétion ou l'addition de quelques bases au site du bris sont ainsi une signature qui suggère l'utilisation de la NHEJ.

1.2.2.5 La liaison d'extrémités alternative

Initialement vue comme un mécanisme de secours en cas de déficience dans la NHEJ, la liaison d'extrémités médiée par une microhomologie (*microhomology-mediated end-joining*, MMEJ) est maintenant considérée davantage comme une voie alternative de réparation (Figure 4) (76). Par conséquent, la MMEJ est souvent appelée liaison d'extrémités alternative (*alternative end-joining*, alt-EJ) (77). Contrairement à la NHEJ, en cellules humaines l'alt-EJ débute par la résection des extrémités 5' du bris afin de rendre disponible une microhomologie (78). Cette étude montre également que seulement la première étape de résection est requise, contrairement à la HR qui nécessite de longues extrémités 3'. Bien que l'alt-EJ présente des étapes similaires à la voie de SSA, les activités de RPA et Rad52 ne sont non seulement pas impliquées dans l'appariement des brins pour l'alt-EJ, mais semblent également nuire au mécanisme (79).

Les études de l'alt-EJ présentent également d'importantes différences au niveau du mécanisme chez la levure et dans les cellules de mammifères (76). Notamment, les cellules de mammifères peuvent utiliser des microhomologies aussi courtes qu'une seule base pour l'alt-EJ, tandis que la levure requiert des homologies d'environ 6-10 bases. Cette différence semble être liée à l'implication de la polymérase à poly ADP-ribose 1 (*poly ADP-ribose polymerase 1*, PARP1) dans la liaison des extrémités chez les cellules de mammifères, réduisant la nécessité de microhomologies (80). Une autre différence importante est la présence chez les mammifères de la polymérase Polθ, qui est absente chez la levure. Polθ a la particularité de faire la synthèse d'ADN selon trois mécanismes : la polymérisation sans matrice, la polymérisation à partir d'une matrice en *cis* et la polymérisation à partir d'une matrice en *trans* (77). Avec ces activités, Polθ contribue à la formation de microhomologies qui aident à l'appariement des brins, et peut également remplir le gap suivant l'appariement. Finalement, l'alt-EJ se distingue également de la NHEJ par la ligase utilisée pour terminer la réparation. Tandis que la NHEJ utilise exclusivement la ligase à ADN IV, l'alt-EJ utilise principalement la ligase à ADN III (DNA ligase III), mais peut aussi avoir recours à la ligase à ADN I (DNA ligase I) (81).

1.3 L'instabilité génomique induite par la réplication d'ADN

Les bris double-brins sont souvent considérés comme le type de dommage à l'ADN qui présente le plus grand risque à la conservation de la séquence, mais l'instabilité génomique peut également survenir en l'absence de bris au squelette de la molécule d'ADN. C'est effectivement le cas pour plusieurs événements pouvant survenir lors de la réplication du génome. Le type d'erreur le plus simple ayant lieu pendant la réplication est l'incorporation de la mauvaise base. Ceci arrive surtout lorsque la polymérase à ADN rencontre une lésion sur le brin matrice et procède à la synthèse d'ADN trans-lésion (*translesion synthesis*, TLS) (82). Puisqu'une erreur d'une base est souvent moins néfaste que l'arrêt de la réplication, certaines polymérases sont spécialisées dans la TLS. L'efficacité accrue de ces enzymes aux sites de lésions vient au coût de leur fidélité, qui est souvent inférieure à celle des polymérases réplicatives. Les modifications de bases qui surviennent à ces lésions ont néanmoins un impact limité sur la cellule, surtout lorsqu'elles ont lieu dans une séquence non-codante. Il arrive également aux polymérases à ADN d'incorporer des ribonucléotides dans un brin d'ADN, créant ainsi un site susceptible à l'hydrolyse (83). De telles incorporations doivent donc être corrigées par des ARNases qui reconnaissent les hybrides ADN-ARN. Il existe par contre d'autres mécanismes par lesquels la réplication peut induire des modifications de plus d'une base à la séquence génomique. Dans certains cas, ces mécanismes sont liés à des pathologies.

1.3.1 Le dérapage de réplication

Le dérapage de réplication (*replication slippage*) est un événement qui génère la délétion ou l'amplification en tandem d'une courte séquence, donnant l'impression que la polymérase à ADN puisse avoir glissé vers l'avant ou vers l'arrière sur le brin matrice. Contrairement à ce que son nom indique, ce mécanisme n'implique pas un déplacement de la polymérase, mais plutôt la formation de structures secondaires dans l'ADN (Figure 5). La formation d'épingles (*hairpins*) par l'appariement d'un brin à lui-même résulte en un rapprochement dans le brin matrice des nucléotides situés de part et d'autre de l'épingle (84). Face à ce type d'obstacle, la polymérase s'arrête et se dissocie, permettant au brin naissant de s'apparier à la séquence qui suit l'épingle. La polymérase peut ainsi reprendre la synthèse et créer une délétion de la séquence formant la structure secondaire (Figure 5). Ce type d'instabilité est particulièrement

associé à des courtes séquences répétées dans la même orientation, puisque les répétitions contribuent au second appariement du brin naissant. Bien que l'exemple illustré résulte en une délétion, il est également possible de générer des amplifications si des structures secondaires se forment sur le brin naissant au lieu du brin matrice. Ce type de dérapage est associé à l'amplification de triplets CAG dans le gène responsable de la maladie de Huntington (85).

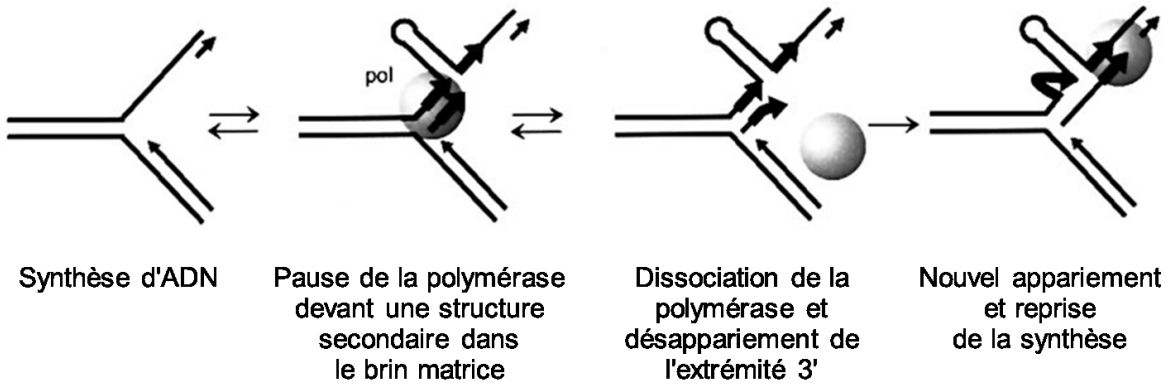


Figure 5. Schéma d'un dérapage de polymérase résultant en une délétion.
Dans cet exemple, une épingle se forme dans la matrice du brin retardé (*lagging strand*) laissé simple-brin par la progression plus rapide du brin avancé (*leading strand*). Les flèches minces représentent la direction de la synthèse d'ADN. Les flèches en gras représentent des séquences répétées. La polymérase est représentée par une sphère. Figure adaptée de (84).

1.3.2 Les fourches de réplication bloquées

Lors de la réplication, plusieurs types d'obstacles peuvent survenir et bloquer la progression de l'hélicase qui ouvre le double-brin ou de la polymérase à ADN. Il peut s'agir de structures secondaires dans l'ADN, de complexes ADN-protéines difficiles à dissocier, de modifications aux bases azotées, ou d'une déplétion de la réserve de nucléotides à incorporer. L'arrêt de la progression de la réplication mène à une structure appelée fourche bloquée (*stalled fork*), qui peut dans certains cas reprendre la synthèse normalement une fois l'obstacle surmonté. Si l'arrêt persiste, la fourche peut être prise en charge par des protéines qui la stabilisent pour permettre des mécanismes de redémarrage. Chez la levure, ces mécanismes sont dépendants de

l'activation de la kinase Rad53 par Mrc1, sans laquelle la réplication ne peut pas reprendre correctement (86). À défaut de cette stabilisation, la fourche peut s'effondrer par l'action de l'endonucléase Mus81 (87). Une fois la fourche effondrée (*collapsed fork*), la synthèse d'ADN ne peut plus continuer normalement, et la réparation doit procéder par l'intermédiaire des mécanismes de réparation des bris vus plus haut. Puisque certains de ces mécanismes sont sujets à l'erreur, il peut être préférable d'éviter l'effondrement de la fourche. Cependant, le redémarrage d'une fourche de réplication bloquée peut également donner lieu à des réarrangements.

1.3.2.1 Le régression de la fourche

Chez l'humain, le mécanisme conservateur le mieux connu pour redémarrer une fourche de réplication bloquée est appelé régression de la fourche (*fork reversal*) et possède certains éléments en commun avec la HR (Figure 6). En effet, la première étape de la régression est l'appariement des deux brins naissants en une structure ressemblant à une jonction de Holliday. Cette structure permet la synthèse d'ADN à partir du brin naissant le plus avancé. L'appariement des brins naissants implique l'activité de RAD51 et est stabilisé par PARP1 (88). Puisque la structure obtenue possède l'équivalent d'une extrémité double-brin, elle est susceptible à la dégradation par MRE11 (89). Cette dégradation est inhibée par la présence du filament de RAD51, qui est recruté à la fourche par l'intermédiaire de BRCA2 (89). Une fois la lésion réparée, la fourche de réplication peut être rétablie par l'activité de l'hélicase RECQ1 (90). Cette étape peut également avoir lieu par l'entremise de la dégradation de l'extrémité 5' de la fourche régressée par DNA2 et l'hélicase WRN (91). Bien que ces protéines soient impliquées dans la résection des extrémités double-brins, le mécanisme par lequel elles participent au redémarrage de la fourche semble indépendant. Il est également intéressant de noter que malgré l'observation de fourches régressées chez la levure, il n'est toujours pas clair si elles ont la même fonction que chez l'humain (92).

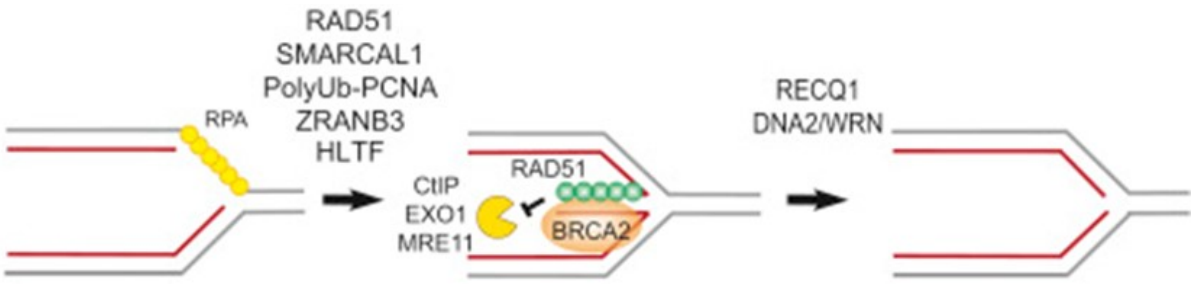


Figure 6. Schéma de la régression de la fourche et des principales protéines impliquées.

Les traits gris représentent les brins parents, et les brins rouges les brins naissants. De gauche à droite : la fourche de réplication bloquée, la fourche régressée et sa jonction de Holliday, la fourche de réplication rétablie. Figure adaptée de (93).

1.3.2.2 Le demi-tour de réplication

Un mécanisme sujet à l'erreur moins étudié pouvant survenir à des fourches bloquées est celui du demi-tour de réplication (*replication U-turn*) (Figure 7). Puisque ce mécanisme est encore peu compris, il n'existe pas de modèle unique bien établi. Par contre, les modèles proposés ont comme point commun l'utilisation de séquences répétées inversées pour permettre l'appariement du brin naissant au brin matrice opposé en orientation inverse. Une fois cet appariement effectué, la synthèse continue d'ADN mène à la duplication de la séquence précédant le site de demi-tour et à la délétion de la séquence qui le suit. Ce modèle fut proposé pour expliquer la formation de chromosomes dicentriques (contenant deux centromères) chez *S. cerevisiae* (94). Cet événement ne dépend pas de la présence de la machinerie de réparation de bris double-brins, et est ainsi considéré comme un changement de matrice (*template switch*) pendant la réplication. Un modèle similaire fut proposé chez *Schizosaccharomyces pombe*, utilisant la machinerie de recombinaison pour effectuer le changement de matrice à une fourche de réplication bloquée (19). Ce genre d'événements semble associé à un redémarrage de la réplication qui commet davantage d'erreurs que la réplication classique (20). Bien que ces modèles puissent expliquer la formation de chromosomes dicentriques, il n'est pas clair si une seconde inversion peut rétablir la réplication dans la bonne orientation. Effectivement, une comparaison des génomes de l'humain et du chimpanzé montrent plusieurs sites de variation pouvant être expliqués par un mécanisme similaire au demi-tour de réplication (95). Le

mécanisme proposé implique par contre un retour de la réplication sur le brin original après une courte élongation. En plus d'être observées entre les génomes d'humain et de chimpanzé, ces mutations existent également entre les génomes d'individus humains. Il ne s'agit donc possiblement pas d'un événement isolé au cours de l'évolution.

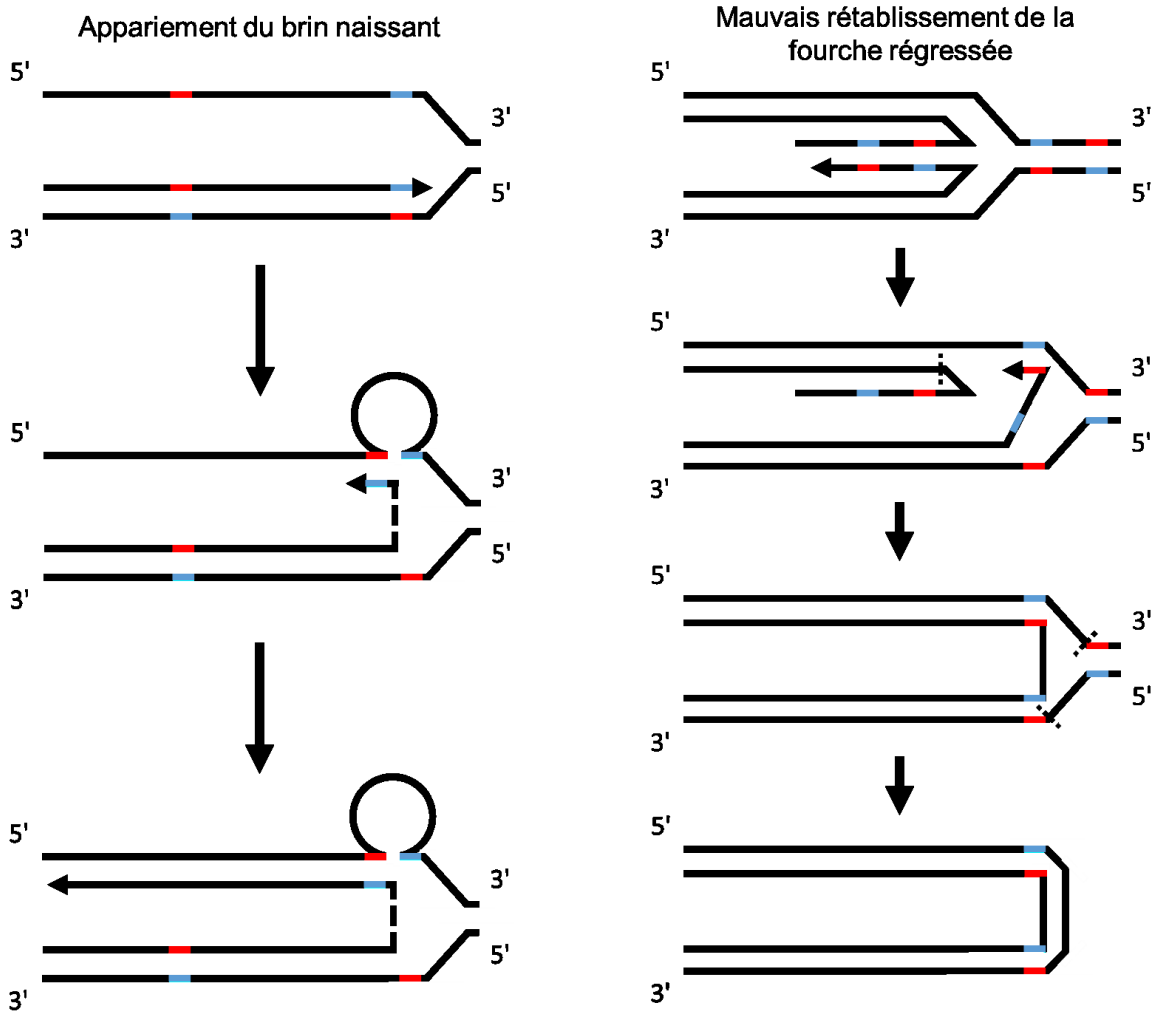


Figure 7. Deux modèles proposés pour le demi-tour de réplication.
Les flèches représentent l'extrémité du brin à laquelle a lieu la synthèse d'ADN. Les segments rouges et bleus représentent des séquences répétées en orientation inverse. Les petits pointillés représentent des sites de clivage par une endonucléase. Figure adaptée de (96).

1.4 La réparation de l'ADN dans les chloroplastes

En tant qu'organelle responsable de la photosynthèse, le chloroplaste est essentiel au développement des plantes vertes (97). La stabilité de son génome représente ainsi un sujet de grand intérêt en agriculture. Étant donné l'origine endosymbiotique du chloroplaste, un grand nombre de protéines impliquées dans la réparation de son ADN sont similaires à des homologues bactériens (98). Les mécanismes utilisés pour réparer l'ADN dans le chloroplaste sont moins bien documentés que ceux décrits dans les génomes nucléaires eucaryotes, mais plusieurs évidences indiquent que la recombinaison homologue répare également les bris double-brins dans le chloroplaste. Ceci implique entre autres des homologues des protéines bactériennes RecA, mais également des résolvases à jonctions de Holliday (4, 99). L'étude d'insertions de séquences dans les chloroplastes de *Chlamydomonas reinhardtii* suggère aussi la présence de SDSA et de SSA dans le chloroplaste (100). De plus, la BIR semble avoir lieu dans les génomes mitochondriaux de plantes, mais il n'est pas clair si c'est également le cas chez les chloroplastes (101). Certains mécanismes semblent exister pour permettre la réparation non-homologue, mais aucune évidence concluante ne montre de mécanismes de liaison des extrémités dans le chloroplaste.

1.5 Objectifs

Le développement du séquençage de nouvelle génération a eu un impact majeur sur l'étude des variations génomiques. L'assemblage d'un génome humain était jusqu'à récemment un ouvrage dispendieux qui demandait plusieurs années de travail, mais il est maintenant envisageable d'assembler plusieurs tels génomes dans le cadre d'un seul projet. Cependant, cette technologie est étrangement encore peu utilisée pour étudier les mécanismes de réparation de l'ADN. Nous nous sommes ainsi donné pour objectif d'utiliser les données obtenues en séquençage pour mieux comprendre la cause des réarrangements génomiques. Notre approche a premièrement été mise au point pour observer les réarrangements présents dans le chloroplaste de la plante modèle *Arabidopsis thaliana*. Nous nous sommes spécifiquement tournés vers des mutants présentant des phénotypes de décoloration partielle des feuilles (variégation) associés à une mauvaise biogénèse des chloroplastes. Nous avons ensuite davantage automatisé l'identification des réarrangements, de façon à rendre l'approche compatible à des génomes plus

grands, tels que le génome nucléaire humain. Ainsi, nous avons tenté d'établir un portrait global quantitatif de l'instabilité génomique chez l'humain. La même approche a ensuite été utilisée chez la levure afin de mieux comprendre le rôle de plusieurs protéines impliquées dans la stabilité génomique. Encore une fois, nous nous sommes tournés vers des mutants présentant des défauts au niveau de la réparation conservatrice de l'ADN.

2. Premier Article: Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in Arabidopsis and humans

Genome Research. 2015 May; 25(5):645-54. doi: 10.1101/gr.188573.114. Epub 2015 Mar 23.

Éric Zampini[1], Étienne Lepage[1], Samuel Tremblay-Belzile[1], Sébastien Truche, and Normand Brisson

[1]These authors contributed equally to this work.

Corresponding author: normand.brisson@umontreal.ca

Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, Québec, Canada H3C 3J7

2.1 Mise en contexte

L'étude de la stabilité génomique chez le chloroplaste d'*Arabidopsis thaliana* au laboratoire a mené à plusieurs découvertes sur le rôle des protéines de la famille Whirly dans la suppression des événements de recombinaison par microhomologies. Le double mutant *why1why3* a initialement permis la détection de tels réarrangements dans le chloroplaste par buvardage de Southern et par amplification PCR (8). Par la même approche PCR, des niveaux encore plus élevés d'instabilité génomique ont été observés chez le triple mutant *why1why3polIb*, ce qui a permis d'expliquer son phénotype de variégation jaune et de retard de croissance (5, 7). L'obtention du triple mutant *why1why3reca1* a cependant illustré les limites de l'amplification PCR de séquences réarrangées. Ce mutant possède un phénotype de variégation blanche et d'altération de la forme des feuilles qui le distingue des plantes de type sauvage et du mutant *why1why3polIb*. L'amplification PCR nous a permis d'observer un niveau élevé de réarrangements similaire entre les deux triples mutants, mais rien ne nous permettait d'expliquer la différence au niveau du phénotype. Nous avons ainsi développé une approche qui pourrait nous permettre d'observer et quantifier les réarrangements présents chez les chloroplastes de ces deux mutants à l'échelle du génome entier (102). Nous avons ainsi obtenu un portrait global de l'instabilité génomique dans chaque mutant, que nous avons comparé à celui de la mitochondrie humaine, entre autres.

Dans le travail qui a mené à cet article, j'ai participé au développement et à la mise au point de l'approche bio-informatique de détection des réarrangements à partir de données de séquençage. Éric Zampini et Normand Brisson ont conçu le projet. É.Z. a isolé l'ADN des plantes et réalisé les expériences en laboratoire. É.Z., Étienne Lepage et moi avons travaillé ensemble pour toutes les étapes d'analyse des résultats de séquençage, pour la mise au point du modèle. Sébastien Truche a participé à l'analyse et à l'interprétation des données. Tous les auteurs ont contribué à la rédaction de l'article.

2.2 Abstract

Failure to maintain organelle genome stability has been linked to numerous phenotypes, including variegation and cytosolic male sterility (CMS) in plants, as well as cancer and neurodegenerative diseases in mammals. Here we describe a next-generation sequencing approach that precisely maps and characterizes organelle DNA rearrangements in a single genome-wide experiment. In addition to displaying global portraits of genomic instability, it surprisingly unveiled an abundance of short- range rearrangements in Arabidopsis thaliana and human organelles. Among these, short-range U-turn-like inversions reach 25% of total rearrangements in wild-type Arabidopsis plastids and 60% in human mitochondria. Furthermore, we show that replication stress correlates with the accumulation of this type of rearrangement, suggesting that U-turn-like rearrangements could be the outcome of a replication-dependent mechanism. We also show that U-turn-like rearrangements are mostly generated using microhomologies and are repressed in plastids by Whirly proteins WHY1 and WHY3.Asynergistic interaction is also observed between the genes for the plastid DNA recombinase RECA1 and those encoding plastid Whirly proteins, and the triple mutant why1why3reca1 accumulates almost 60 times the WT levels of U-turn-like rearrangements. We thus propose that the process leading to U-turn-like rearrangements may constitute a RecA-independent mechanism to restart stalled forks. Our results reveal that short-range rearrangements, and especially U-turn-like rearrangements, are a major factor of genomic instability in organelles, and this raises the question of whether they could have been underestimated in diseases associated with mitochondrial dysfunction.

2.3 Introduction

The endosymbiosis events at the origin of mitochondria and plastids have been a key step in the appearance of eukaryotic cells. During evolution, most of the genes present in the genome of the endosymbionts have been lost or transferred to the nucleus, resulting in organelles with small, dense genomes that almost exclusively encode essential genes. There is therefore a strong selective pressure for the maintenance of genome stability in organelles (Wolfe et al. 1987). In plants, *Arabidopsis thaliana* has emerged as a useful model to study organelle genome rearrangements. Several actors involved in the maintenance of *Arabidopsis* plastid and mitochondrion genomes have been identified, among which are the Whirly proteins, a family of single-stranded DNA-binding proteins that guard organelles against genomic rearrangements (Maréchal et al. 2009; Cappadocia et al. 2010). It has been proposed that these proteins stabilize single-stranded DNA and guide it through conservative repair mechanisms such as homologous recombination (Maréchal et al. 2009; Cappadocia et al. 2010). Whirly proteins are found in all plant organelles and, in *Arabidopsis*, WHY1 and WHY3 are targeted to the plastids while WHY2 is targeted to the mitochondria (Krause et al. 2005). The RecA family of DNA-binding proteins also maintains organelle genome stability through their central role in homologous recombination. In *Arabidopsis*, three RecA proteins are found in organelles, with RECA1 targeted to the chloroplast, RECA3 to the mitochondrion, and RECA2 targeted to both organelles (Shedge et al. 2007). These proteins were shown to be involved in DNA double-strand break (DSB) repair and, in plastids, to maintain the structure of the genome (Rowan et al. 2010; Miller-Messmer et al. 2012). Prokaryote RecA proteins were also shown to be essential for fork reversal, a mechanism that allows the accurate restart of paused replication forks, thereby promoting fork progression in conditions of replication stress (Seigneur et al. 2000; Robu et al. 2001; Costes and Lambert 2012).

Replication represents a major challenge to plastid and mitochondrion genome stability. For example, it was shown that the mutation of the type-I polymerase POLIB in *Arabidopsis* causes replication stress at early developmental stages and increases the amount of DSBs upon genotoxic stress treatment (Parent et al. 2011). Interestingly, mutation of the mammalian mitochondrial DNA polymerase gamma has also been linked to replication stress and DSBs (Vermulst et al. 2008; Ameur et al. 2011). Replication-dependent DSBs are known to arise

when a fork collapses or encounters a nick in the matrix DNA (Zeman and Cimprich 2014). These DSBs can subsequently be repaired by homologous recombination or by error-prone mechanisms such as microhomology-mediated recombination (MHMR) and non-homologous end-joining (NHEJ) (Lundin et al. 2002; Lee et al. 2007; Hastings et al. 2009). Replication-dependent genomic instability is however not solely induced by DSB repair mechanisms (Zeman and Cimprich 2014). Indeed, stalled forks have recently been shown to produce DSB-independent fusions of nearby inverted-repeats, which lead to the formation of palindromic chromosomes (Mizuno et al. 2009; Paek et al. 2009). The mechanisms by which these fusions take place remain a subject of debate, and three distinct possibilities have been proposed: faulty template switching, tandem inversion duplications, and replication U-turns (Mizuno et al. 2009; Paek et al. 2009; Kugelberg et al. 2010; Seier et al. 2012; Mizuno et al. 2013).

To date, our understanding of how organisms deal with replication-associated genomic instability has mostly been obtained using reporter systems. A drawback of these systems is that they are impractical for the study of organelle genomes. To overcome this limitation, we developed a next-generation sequencing approach which allows the characterization of organelle DNA instability at a genome-wide level while providing information about the mechanism underlying each rearrangement formation.

2.4 Results

2.4.1 Short-range rearrangements are abundant in organelle genomes

We first evaluated overall plastid DNA (ptDNA) rearrangements in the *Arabidopsis thaliana* ecotype Col-0. To obtain a global and quantitative portrait of ptDNA instability, a heat map representation was employed in which all rearrangements are reported at the intersection of the two genomic coordinates that correspond to each side of the rearrangement junction. This analysis shows that many rearrangements occur apparently randomly in this plastid genome (Fig. 8A and Supplemental Table S1). It also reveals an over-representation of short-range rearrangements (<1,000 bp), as indicated by the higher intensity of the heat map diagonal (Fig. 8A,D). To validate our approach, the same analysis was repeated with the *Arabidopsis thaliana*

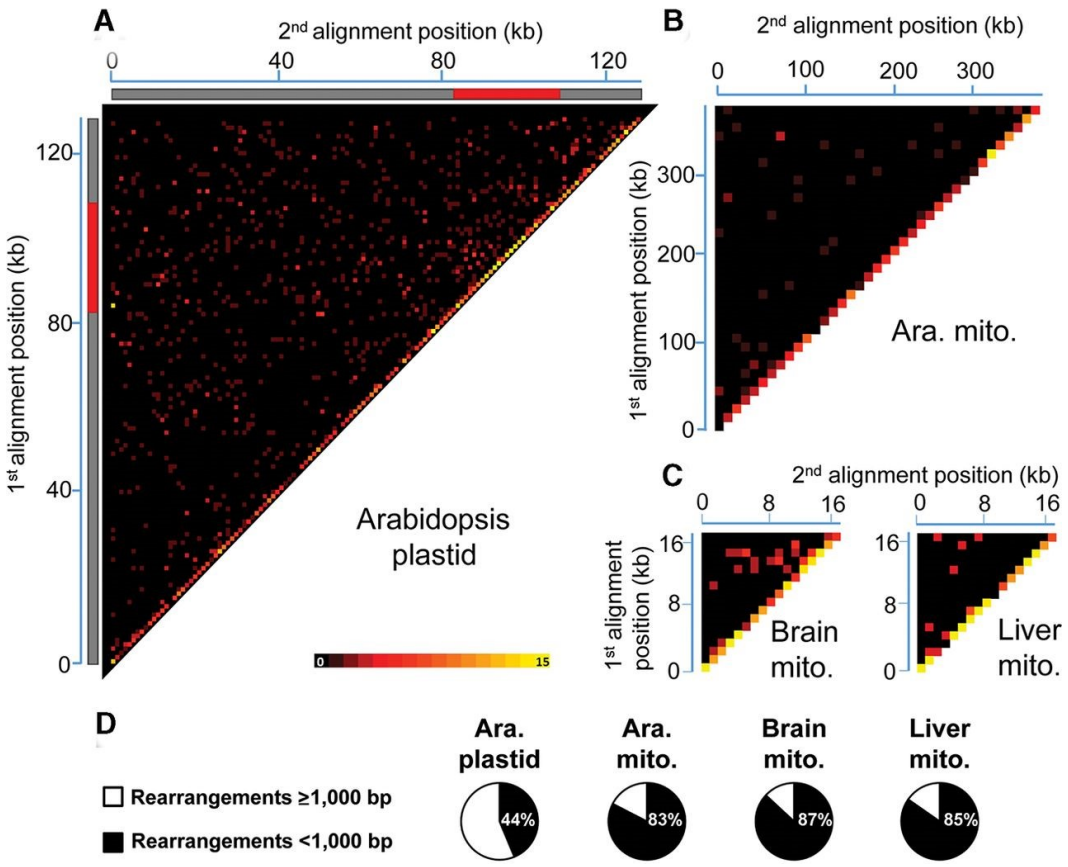


Figure 8. Global portrait of organelle genome rearrangements in *Arabidopsis* and humans.
Heatmaps depict each rearrangement as the intersection of the two genomic positions corresponding to the nucleotide on each side of the junction. (A) Rearrangement breakpoint positions of *Arabidopsis thaliana* (Col-0) plastids. Each tile represents a region spanning 1 kb along each axis. Tile intensity represents the number of rearrangements per 10,000 plastid genomes. All rearrangements mapping to the plastid large inverted repeats (IRs) were only assigned to the first IR. The plastid large single-copy region (LSC), the first IR, and the small-single copy region (SSC) are depicted as a long gray bar, a red bar, and a short gray bar, respectively. (B) Rearrangement breakpoint positions of wild-type *Arabidopsis thaliana* (Col-0) mitochondria. Each tile represents a region spanning 10 kb along each axis. Tile intensity represents the number of rearrangements per 100 mitochondrion genomes. (C) Rearrangement breakpoint positions of a representative sample for human brain (ERX385572) and liver (ERX385578) mitochondria. Each tile represents a region spanning 1 kb along each axis. Tile intensity represents the number of rearrangements per 10,000 mitochondrion genomes. (D) Proportion of short-range (breakpoint positions <1000 bp apart) and long-range (breakpoint positions at least 1000 bp apart) rearrangements of *Arabidopsis thaliana* (Col-0) plastid, mitochondrion, and of the mean of the four samples for brain and liver mitochondria. (Ara.) *Arabidopsis*, (mito.) mitochondrion.

ecotypes Ts-1 (Tossa de Mar) and Ws-2 (Wassilewskija) (Weigel and Mott 2009). Results indicate that the overall pattern and relative level of plastid DNA rearrangements are strikingly similar between ecotypes (Supplemental Fig. S1 and Supplemental Tables S2-3).

To verify if mitochondria exhibit a similar pattern of genomic instability, we generated a heat map for *Arabidopsis thaliana* (Col-0) mitochondrial genome rearrangements. Interestingly, the vast majority of rearrangements also occur at short range in this organelle (Fig. 8B,D and Supplemental Table S4). To our knowledge, this is the first time that such a high level of short-range rearrangements is reported in organelle genomes, which raises the question of whether these could also be abundant in animal mitochondria. To test this hypothesis, we subjected publicly available datasets from human brain and liver to the same analysis. This revealed a similar pattern of short-range genomic instability, reaching 86% and 84% of total rearrangements in brain and liver cells, respectively (Fig. 8C-D and Supplemental Tables S5-12). In contrast, analysis of genomic instability in *E. coli* reveals a much less striking prominence of short-range rearrangements (Supplemental Tables S13-16), indicating that the detection of high levels of these rearrangements in organelles does not occur systematically in our approach. Globally, these results indicate that short-range rearrangements are a major factor of genomic instability in both plant and animal organelles.

2.4.2 Microhomology and non-microhomology repair happen at similar rates in wild-type organelles

To get insights into the mechanisms involved in the formation of the DNA rearrangements in organelles, we further analyzed the reads corresponding to rearranged genome molecules. Because these reads are always composed of two alignments mapping distinct regions of the genome, two types of DNA rearrangements can easily be discriminated: those that possess a microhomology at their junction, most likely formed by MHMR, and those without microhomology, reminiscent of NHEJ repair (Fig. 9A). This analysis revealed that in wild-type Col-0 (WT) plastids and mitochondria, 55% and 64% respectively of genome rearrangements detected arose from microhomology-dependent pathways with microhomologies of 5 bp or more (Fig. 9B). Similarly, rearrangements dependent on

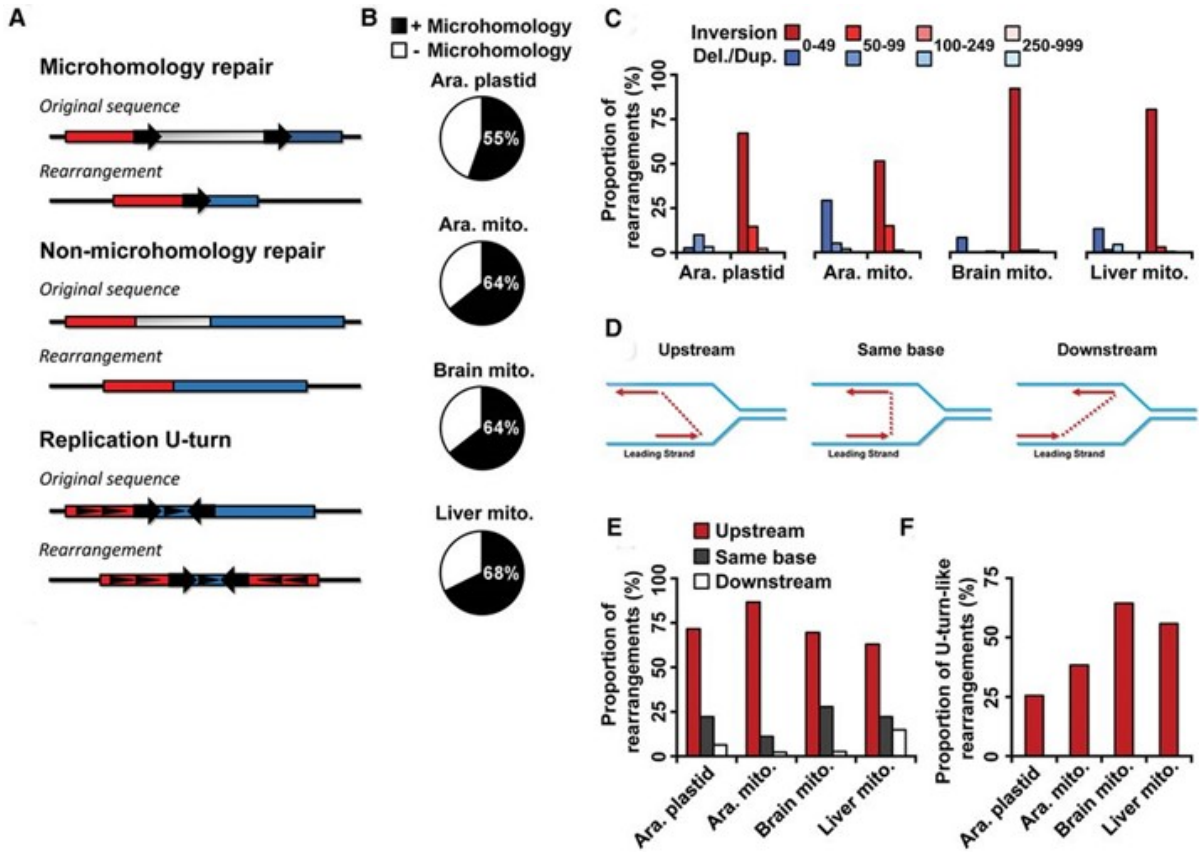


Figure 9. Analysis of organelle genome rearrangements in *Arabidopsis* and humans.

(A) Depiction of the types of rearrangement junctions observed by next-generation sequencing. Large arrows represent microhomologies and their orientation. Small triangles represent relative strand directions. (B) Proportion of rearrangements identified displaying a microhomology (≥5 bp) at their junction in each organelle. The mean of the four samples for brain and liver mitochondria is presented. (C) Proportion of short-range deletions/duplications and inversions (<1000 bp) displaying a gap of the indicated length. The y-axis represents the percentage of each type of rearrangement relative to its class size (gap length). A representative sample is shown for human brain (ERX385572) and liver (ERX385578) mitochondria. (D) Schematic representation of short-range inversions displaying a junction upstream, at the same base, or downstream on the opposite strand. The matrix strands are shown in blue, and the nascent strands in red, with the junction gap shown as a dotted line. (E) Proportion of short-range inversions (<50 bp) displaying a junction upstream, at the same base, or downstream on the opposite strand in each organelle genome. A representative sample is shown for human brain (ERX385572) and liver (ERX385578) mitochondria. (F) Proportion of total rearrangements corresponding to short-range inversions (<50 bp, U-turn-like rearrangements) in each organelle genome. The mean of the four samples for brain and liver mitochondria is presented. (Ara.) *Arabidopsis*, (mito.) mitochondrion, (del.) deletion, (dup.) duplication.

microhomologies represent 64% and 67% of total genomic instability in human brain and liver mitochondria, respectively (Fig. 9B).

2.4.3 U-turn-like rearrangements are ubiquitous among organelle short-range rearrangements

The fact that short-range rearrangements represent a large part of the total junctions identified in organelle genomes suggests that these rearrangements could arise through a specific mechanism taking place at nearby microhomologies. Analysis of rearrangement junctions at the base pair level offers the key advantage of allowing the characterization of these mechanisms. We therefore measured the proportion of rearrangements involving inversions among the short-range rearrangements (<1,000 bp) according to their gap length in comparison to other types of rearrangements (duplications and deletions) in *Arabidopsis* and human organelles. This revealed that inversions occurring over distances smaller than 50 bp are largely overrepresented (Fig. 9C). This type of genomic instability seems to be of particular importance in organelles since it represents approximately 26% and 38% of total rearrangements in *Arabidopsis* plastids and mitochondria, respectively (Fig. 9E). This proportion is even higher in human mitochondria, reaching 64% and 56% of the total brain and liver genomic instability, respectively (Fig. 9E). A closer look at the junctions of these inversions indicates that they mainly occur upstream of the junction breakpoint, on the opposite strand (Fig. 9D-F). Interestingly, although often imperfect, approximately 85% of these short-range inversions present a microhomology at their junction in plastids. The overall characteristics of the DNA junctions observed in these short range inversions can best be explained by a model in which replication executes a U-turn upstream on the opposite strand (Fig. 9D). It also suggests that the inversions take place on the opposite strand between the 5' end of the lagging strand and the site where the DNA is unwound in the replication fork.

2.4.4 Whirly proteins protect the plastid genome from MHMR but not from NHEJ

The organelle DNA polymerase POLIB and Whirly proteins have previously been shown to protect plastid genomes against long-range deletions and duplications mediated by

microhomologies. To verify whether these proteins do also offer a protection against short-range DNA rearrangements, we further characterized the Arabidopsis mutant lines *why1why3* and *polIb* by sequencing. We also included in our analysis the mutant line *reca1-1* (hereafter called *reca1*), which is mutated for plastid RECA1 (Supplemental Fig. S2). While only a small increase in ptDNA rearrangements is observed in *polIb* and *reca1* as compared to the Col-0 parent line, a much higher level of genomic instability is observed in the *why1why3* line (Fig. 10A-B and Supplemental Tables S17-19). Interestingly, the heat map diagonal region of all three mutant lines also shows a much higher intensity than the rest of the genome (Fig. 10A). It thus seems that short-range rearrangements also constitute an important part of the total ptDNA instability in these lines and that their occurrence is increased when genes involved in DNA metabolism are mutated. To get insights into the mechanisms involved in the formation of these rearrangements, we determined the prevalence of microhomology usage for each line. While approximately 0.12 DNA rearrangements per genome arose from microhomologies in WT plants, this increased to 0.8 microhomology-dependent rearrangements per genome in *why1why3* (Fig. 10B). This shift toward microhomology usage in *why1why3* is also confirmed by the analysis of the lengths of the microhomologies leading to rearrangements. Indeed, while microhomology-mediated rearrangements in WT, *reca1* and *polIb* plants are in large part produced by microhomologies of 5 to 9 bp, those generated from microhomologies of 10 to 14 bp are the most prevalent in the Whirly mutant (Fig. 10C). In contrast, the level of DNA rearrangements generated independently of microhomologies is similar in all three lines (Fig. 10B), suggesting that Whirly proteins mainly suppress the appearance of microhomology-dependent rearrangements, and do not affect microhomology-independent pathways. The *polIb*, *reca1* and *why1why3* mutations however have little effect on the accumulation of U-turn-like rearrangements (inversions occurring over distances smaller than 50 bp), which only slightly increase in all three mutant lines (Fig. 10D).

2.4.5 Whirly proteins, POLIB and RECA1 all act to maintain stability in the plastid genome

Genome maintenance is a tightly controlled process in which many proteins act in concert to repress the accumulation of DNA rearrangements. Consequently, the mutation of a

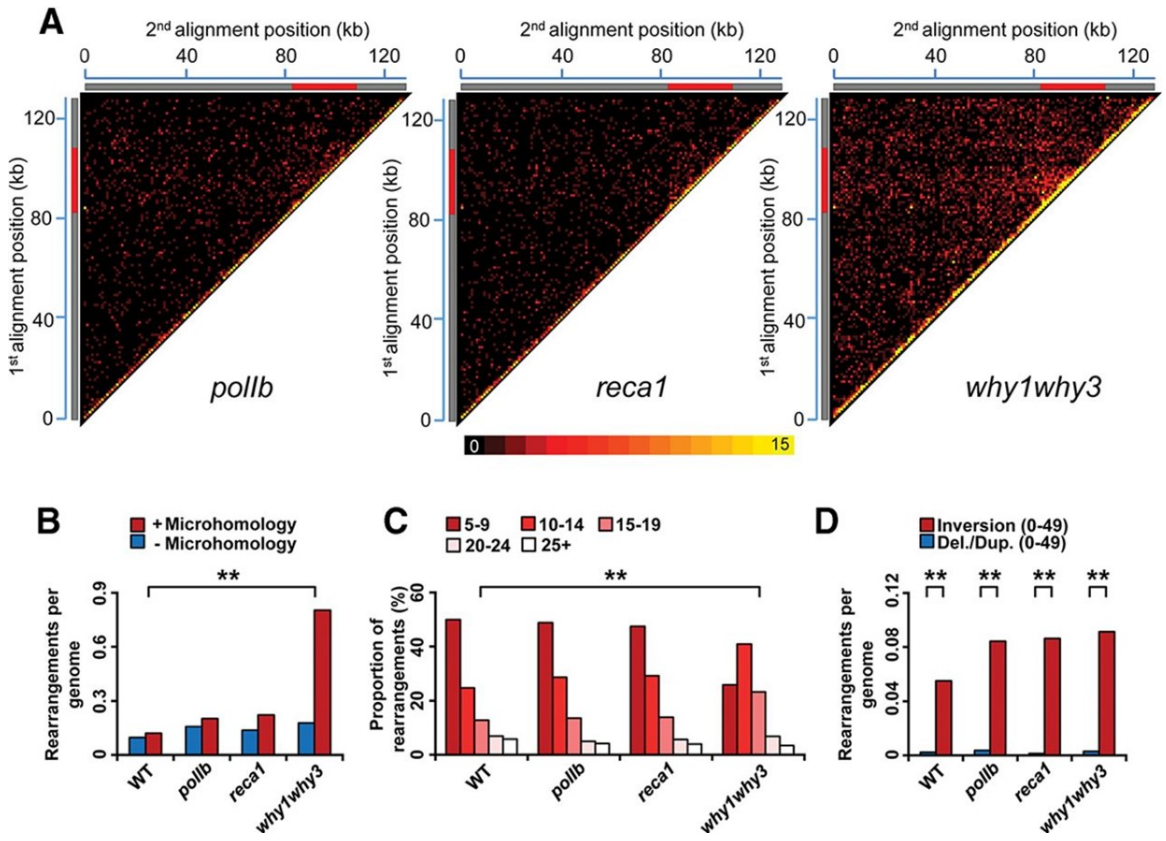


Figure 10. Global portrait of plastid genome rearrangements in *Arabidopsis* mutant lines *pollb*, *reca1*, and *why1why3*.

(A) Plastid rearrangement breakpoint positions of the indicated *Arabidopsis* mutant lines. Heatmaps depict each rearrangement as the intersection of the two genomic positions corresponding to the nucleotide on each side of the junction. Each tile represents a region spanning 1 kb along each axis. Tile intensity represents the number of rearrangements per 10,000 plastid genomes. All rearrangements mapping to the plastid large inverted repeats (IRs) were only assigned to the first IR. The plastid large single-copy region (LSC), the first IR, and the small-single copy region (SSC) are depicted as a long gray bar, a red bar, and a short gray bar, respectively. (B) Normalized amount of rearrangements per plastid genome displaying a microhomology (≥5 bp) (red) or not displaying a microhomology (blue) at their junction for each of the indicated mutant lines. (C) Proportion of total plastid rearrangements displaying a microhomology of given lengths, in bases, in each mutant line. The y-axis represents the percentage of usage of each microhomology length. (D) Normalized amount of rearrangements corresponding to short-range inversions (<50 bp, U-turn-like rearrangements, red) and short-range deletions/duplications (<50 bp, blue) in each mutant line. Data were normalized on the total number of plastid genome reads. (Del.) deletion, (dup.) duplication. Two asterisks indicate a significant difference with *P*-value < 0.00001 using a χ^2 test.

single gene may not be sufficient to significantly destabilize the genome. We verified whether *POLIB*, *RECA1* and plastid Whirly genes have additive or synergistic effects by combining mutations for these genes in *Arabidopsis*. While the phenotype of a *reca1polIb* line is indistinguishable from that of WT plants (Fig. 11A), *why1why3polIb* is characterized by a severe growth retardation phenotype, in addition to yellow-variegation (Fig 11A) (Parent et al. 2011; Lepage et al. 2013). *why1why3reca1* triple mutants show severe growth retardation as well and display white variegation and leaf distortion (Fig 11A and Supplemental Fig. S3A). Also, embryo lethality is observed in *why1why3reca1*, with 59% of seeds being unable to germinate (Supplemental Fig. S3B). The quadruple mutant *why1why3polIbreca1* could not be isolated from the progeny of *why1why3*$^{-/-}$*polIb*$^{+/-}$*reca1*$^{+/-}$ plants, nor from *why1why3*$^{-/-}$*polIb*$^{-/-}$ *reca1*$^{+/-}$ plants (Stouffer's Z test p-value <0.05), suggesting that high levels of plastid genome instability leads to embryo lethality. While the overall pattern of ptDNA rearrangements in *reca1polIb* was similar to that of WT plants, an important increase was observed in both *why1why3polIb* and why1why3reca1 (Fig. 11B and Supplemental Tables S20-22). Furthermore, a hotspot for rearrangements is present in the heat map of these two triple mutants, at the intersection of regions 86-101 kilobases (kb) and 68-100 kb. By contrast, the IR region from 101-108 kb, which encodes both plastid ribosomal RNAs, displays lower levels of instability.

Comparison of ptDNA rearrangements between *why1why3polIb* and *why1why3reca1* indicates that their amount of long-range rearrangements (≥1000 bp), arising with or without the use of microhomologies, are strikingly similar (Fig. 11D-E). This suggests that these types of rearrangements are unlikely to be responsible for the difference in phenotypes observed between the two lines. In contrast, rearrangements having occurred between regions separated by less than 1 kb were much more abundant in *why1why3reca1* than in *why1why3polIb* (Fig. 11C-D). *why1why3reca1* plastids also contain a 60-fold increase in U-turn-like rearrangements compared to WT plastids (Fig. 11F), which corresponds to approximately 60% of all short-range rearrangements present in this line. These results suggest that Whirly proteins and RECA1 both act to suppress the appearance of U-turn-like rearrangements in plastids. In addition, this high level of U-turn-like rearrangements could account for the severe phenotype observed in *why1why3reca1* plants (Fig. 11A).

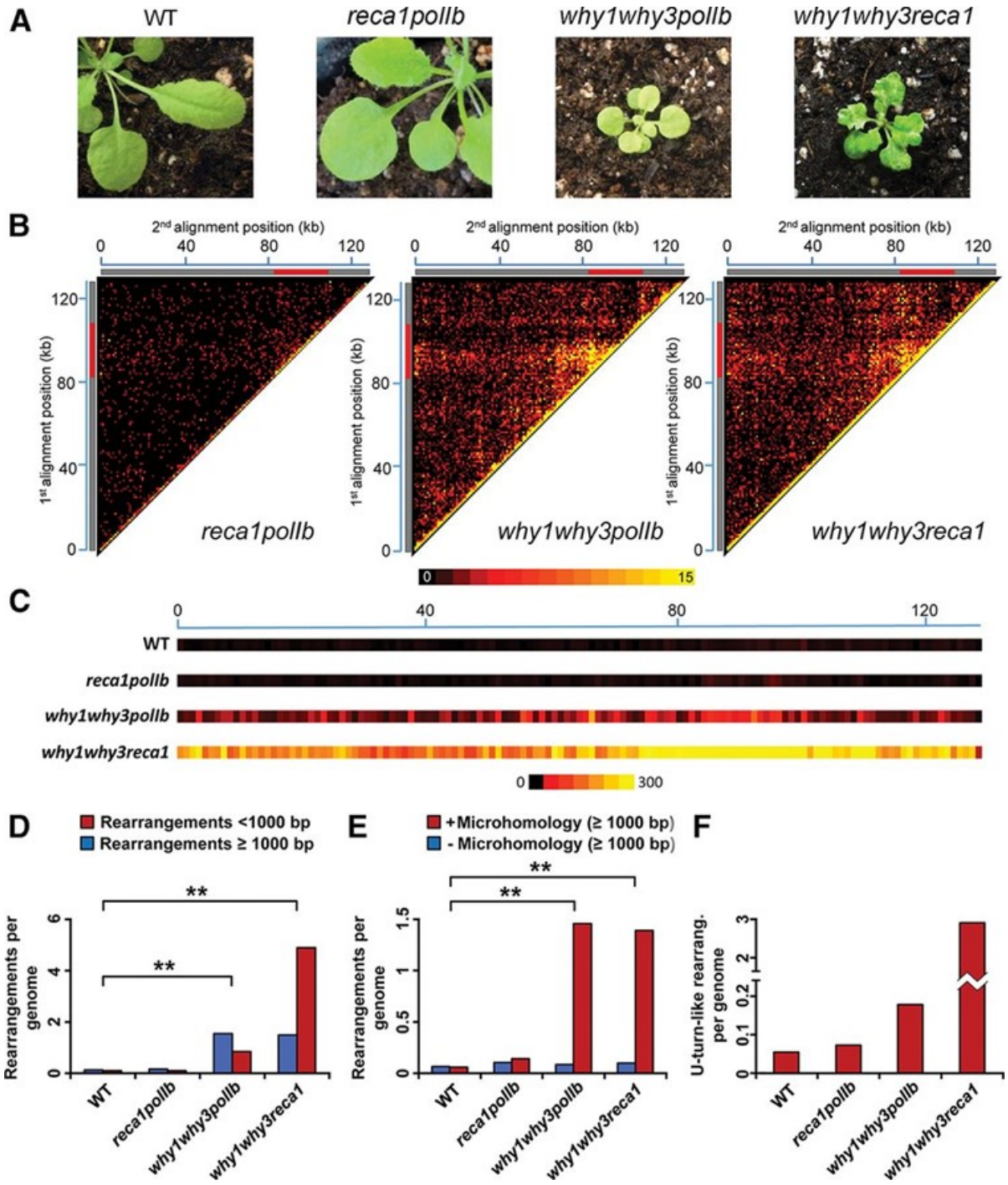


Figure 11. Global portrait of plastid genome rearrangements in Arabidopsis mutant lines *reca1polIb*, *why1why3polIb*, and *why1why3reca1*.

(A) Representative photographs of 21-d-old WT, *reca1polIb*, *why1why3polIb*, and *why1why3reca1 Arabidopsis* mutant plants. (B) Plastid rearrangement breakpoint positions of the indicated Arabidopsis mutant lines. Heatmaps depict each rearrangement as the intersection of the two genomic positions corresponding to the nucleotide on each side of the junction. Each tile represents a region spanning 1 kb

along each axis. Tile intensity represents the number of rearrangements per 10,000 plastid genomes. All rearrangements mapping to the plastid large inverted repeats (IRs) were only assigned to the first IR. The plastid large single-copy region (LSC), the first IR, and the small-single copy region (SSC) are depicted as a long gray bar, a red bar, and a short gray bar, respectively. (C) Short-range (<1000 bp) plastid rearrangement breakpoint positions of the indicated *Arabidopsis* mutant lines. Each heatmap tile represents a region spanning 1 kb along the genome and the number of rearrangements per 10,000 plastid genomes. All rearrangements mapping to the plastid large inverted repeats (IRs) were only assigned to the first IR. (D) Normalized amount of short-range (<1000 bp) and long-range (≥1000 bp) rearrangements per plastid genome for each of the indicated mutant lines. (E) Normalized amount of long-range rearrangements (≥1000 bp) per plastid genome displaying a microhomology (≥5 bp) (red) or not displaying a microhomology (blue) at their junction for each of the indicated mutant lines. (F) Normalized amount of rearrangements corresponding to short-range inversions (<50 bp, U-turn-like rearrangements) in each mutant line. Data were normalized on the total number of plastid genome reads. Two asterisks indicate a significant difference with *P*-value < 0.00001 using a χ^2 test.

2.4.6 U-turn-like rearrangements are associated to replication stress

The characteristics of the U-turn-like rearrangements identified above suggest that they might arise during replication, and that their formation could therefore be linked to replication stress. To test this hypothesis, we verified if a replication stress could be observed in *why1why3polIb* and *why1why3reca1*. Replication stress is defined here as the slowing of replication fork progression during DNA synthesis (Zeman and Cimprich 2014) which, in next-generation sequencing data, has been associated to a progressive, directional decrease in DNA copy number along the genome (Slager et al. 2014). We therefore compared ptDNA coverage curves for each mutant line to the WT. Results indicate that the coverage in WT plants has a similar pattern to those previously reported in the literature for plastid genomes (Wu et al. 2012; Ferrarini et al. 2013), and to that of all single and double mutant lines used in this study (Supplemental Fig. S4). However, the pattern for ptDNA coverage appears quite different in *why1why3polIb* and *why1why3reca1* (Fig. 5A). Regression analysis within the large single-copy region (LSC, 0-84 kb) of *why1why3polIb* reveals a steeper slope than in the WT (Fig. 12B), suggesting that replication is affected in this mutant line and is unidirectional along the LSC, going from 84 kb toward the beginning of the genome. However, yet another pattern is observed in *why1why3reca1*, with two slopes converging in the middle of the LSC in a manner consistent with bidirectional replication (Fig. 12B). Both of these patterns were also confirmed using quantitative PCR (Supplemental Fig. S5). These results therefore suggest that the *polIb* and *reca1* mutations, when combined to the *why1why3* mutations, affect replication differently. Nevertheless, the steeper slopes in the LSC observed for *why1why3polIb* and *why1why3reca1* suggest a replication stress, which could be at the origin of the increase in short-range inversions observed in the plastid genome of both of these mutant lines (Fig. 11F). Interestingly, the essential role of RecA in replication fork reversal and restart was previously shown in bacteria and could account for the replication stress observed in *why1why3reca1*. This result thus supports a role for RECA1 in plastid DNA replication.

Since replication stress is generally associated to incomplete replication of chromosomes, pulse-field gel electrophoresis (PFGE) was used to visualize the distribution of plastid genomic molecules in the previous mutant lines. This confirmed an earlier study which demonstrated that, in contrast to the WT, no monomeric form of the chloroplast genome is

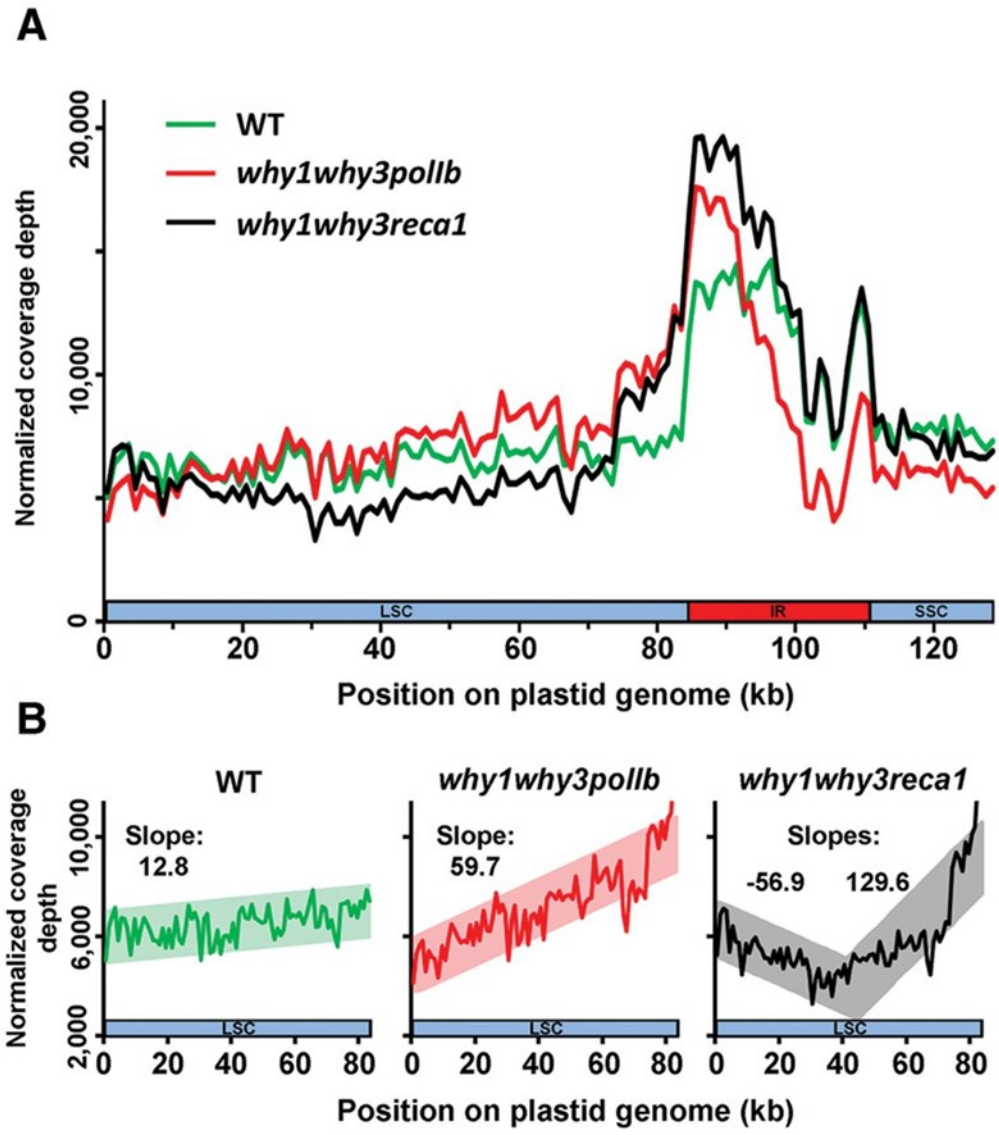


Figure 12. Plastid DNA sequencing coverage curves for *Arabidopsis* lines WT, *why1why3polIb*, and *why1why3reca1*.

(A) Plastid sequencing coverage of pools of 14-d-old *Arabidopsis* seedlings of the indicated genotypes. Positions were rounded down to 1 kb. All reads mapping to the plastid large inverted repeats (IRs) were only assigned to the first IR. The plastid large single-copy region (LSC), the first IR, and the small-single copy region (SSC) are depicted as a long blue bar, a red bar, and a short blue bar, respectively. The y-axis represents the number of reads per 1,000,000 total plastid reads. (B) Regression analysis of the plastid LSC sequencing coverage of WT, *why1why3polIb*, *why1why3reca1* seedlings. Positions were rounded down to 1 kb. The y-axis represents the number of reads per 1,000,000 total plastid reads.

observed in reca1 mutant plants (Rowan et al. 2010). Similar to reca1, disappearance of the monomeric form is observed in *why1why3polIb* and *why1why3reca1* (Supplemental Fig. S6). However, a smear of subgenomic molecules, likely associated with abortive rounds of replication, is instead observed in these mutants, supporting the hypothesis that the elevated levels of U-turn-like rearrangements in these lines could be the consequence of an ongoing replication stress. Note that subgenomic molecules are also observed in the *why1why3* mutant, suggesting that Whirly proteins are also important for replication.

2.5 Discussion

By sequencing DNA rearrangement junctions, we have been able to obtain a genome-wide portrait of DNA instability in the genomes of organelles and of a prokaryote. Unlike paired-end sequencing approaches used to detect rearrangements, our method is not affected by the distance over which the rearrangements occur. It therefore provides a view of rearrangements occurring both at short-range and long-range, while also allowing some characterization of the DNA repair mechanisms used (Supplemental Fig. S7-8). Indeed, it revealed that in *Arabidopsis* and human organelles, both microhomology-dependent and independent pathways constitute an important driving force of genome variation, with pathways using microhomologies being used slightly more often (Fig. 9B). Although illegitimate recombination has been known to occur in organelles (Small et al. 1987; Ogihara et al. 1988), there was no evidence that microhomology-independent repair, such as NHEJ, accounted for such a substantial proportion of organelle DNA rearrangements. Our approach also unveiled an unexpected pattern of genomic instability in organelles of both *Arabidopsis* and humans, with a strong propensity to generate rearrangements between closely spaced regions of the genome. Since DNA rearrangements are often associated to errors occurring during DNA repair (Lundin et al. 2002; Lee et al. 2007; Hastings et al. 2009), the high occurrence of short range rearrangements suggests that, in organelle genomes, inaccurate DNA repair takes place predominantly in the vicinity of the damaged DNA.

Our findings also reveal that, in organelles, a previously-unreported but frequent error-prone mechanism exists that most likely contributes to the restart of stalled replication forks. Indeed, we observed that U-turn-like rearrangements are particularly abundant among short-

range DNA rearrangements in both *Arabidopsis* and human organelles. The finding that an apparent DNA replication stress correlates with the appearance of U-turn-like rearrangements in plastids suggests that they are initiated in a replication-dependent manner. Paused forks are unstable structures often formed in conditions of replication stress, which leads us to hypothesize that they act as a template for U-turn-like rearrangements. The tendency of these U-turn-like rearrangements to occur upstream and the presence of short inverted repeats at most of their junctions suggest that a small inverted repeat in the 3' end of the nascent strand would misanneal to its complement upstream on the opposite strand and lead to inaccurate fork restart (Fig.13).

The fact that microhomologies favor U-turn-like rearrangements suggests that WHY1 and WHY3, which limit the accumulation of microhomology-dependent rearrangements, prevent this form of instability by repressing the misannealling of closely-spaced microhomologies. Also, the observation that single mutations of *RECA1* and *POLIB* have no visible impact on U-turn-like rearrangements but cause major changes only when combined to the mutations of *WHY1* and *WHY3* indicate the synergistic interaction between these genes and the Whirly genes. In the absence of Whirly proteins, replication stress could explain an increase in U-turn-like rearrangements, as observed in *why1why3polIb*. In contrast, we postulate that replication stress alone does not likely account for the much larger increase observed in *why1why3reca1*. We propose instead that RECA1 is involved in a conservative repair pathway that directly competes with replication U-turns. Interestingly, the recombinase RecA has been shown to be essential in bacteria to efficiently bypass lesions and restart replication by promoting replication fork reversal (Seigneur et al. 2000; Robu et al. 2001; Costes and Lambert 2012). It can thus be hypothesized that RECA1 also promotes lesion bypass and accurate fork restart in plastids, thereby limiting the accumulation of U-turn-like rearrangements. The severe phenotype observed in the *why1why3reca1* mutant also indicates that U-turn-like rearrangements have deleterious effects, and that their occurrence must be limited by multiple checkpoints.

Taken together, our results suggest a model for the generation of U-turn-like rearrangements in which progression of the leading strand polymerase is arrested during replication as a consequence of replication stress (Fig. 13). If the fork remains paused and

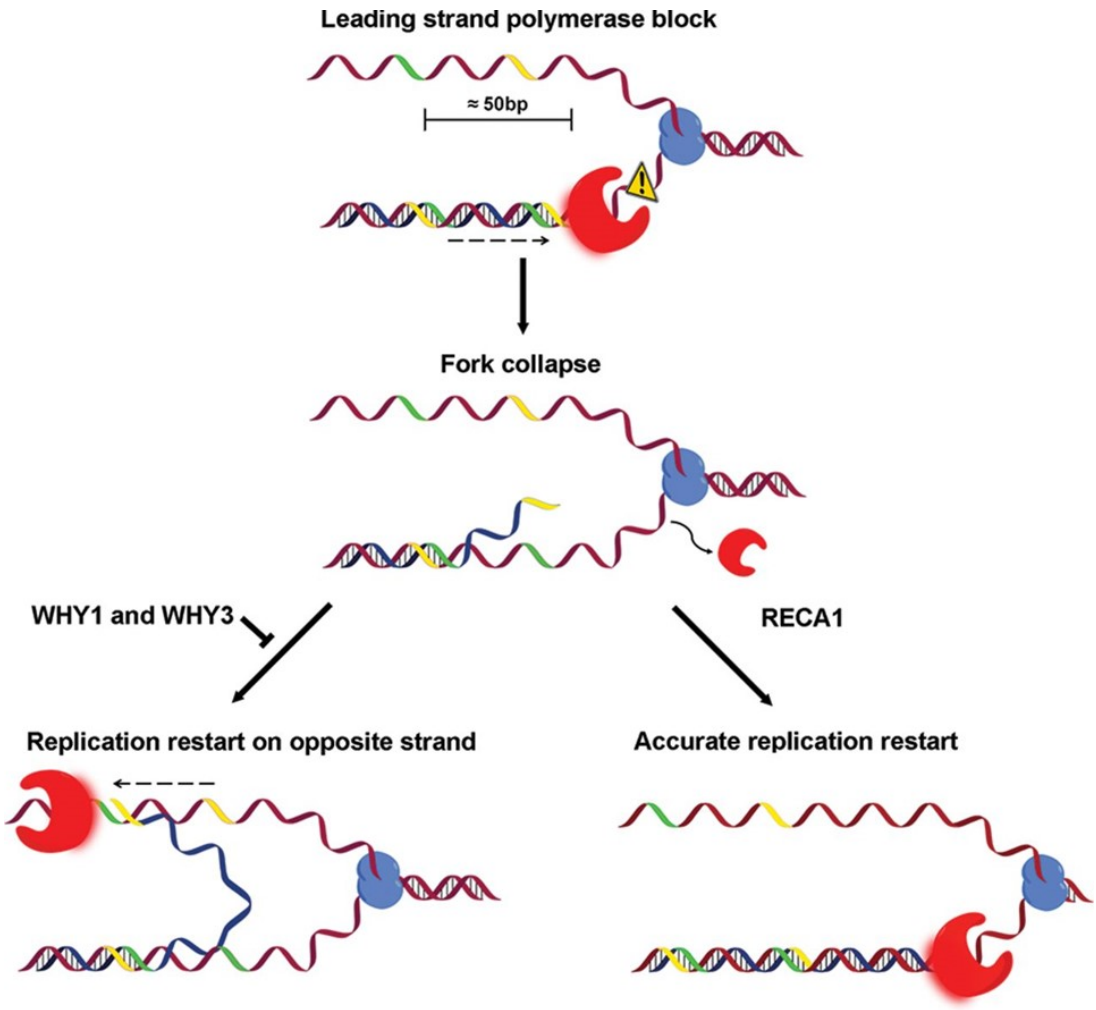


Figure 13. Model of microhomology-mediated U-turn-like inversions.

The red shape represents the DNA polymerase and the blue shape represents the DNA helicase. Yellow and green strands represent inverted microhomologies. Impediments to replication fork progression might force the leading strand polymerase to pause and eventually unload. In the presence of RecA, the impediment may be bypassed to restart replication accurately. In the absence of RecA and Whirly proteins, a microhomology located at the 3′ end of the nascent strand can anneal to its complementary inverted repeat on the opposite strand and restart the replication fork on the wrong strand. The dotted arrow represents a multistep process.

eventually collapses, mechanisms such as RecA-dependent fork-reversal can accurately restart replication. In the absence of RECA1 and Whirly proteins, forks can restart inaccurately by the misannealing of a small inverted repeat in the 3' end of the nascent strand to its complement upstream on the opposite strand. Replication would then be re-initiated on the opposite strand until it either displaces the lagging strand or is ligated to its 5' end, resulting in the formation of a palindromic chromosome. Although this mechanism generates genomic rearrangements, it might serve as an alternative mechanism to restart replication forks. Whirly proteins would avert these rearrangements by binding to the single-stranded DNA and thus preventing misannealing of small inverted repeats.

Overall, our approach reveals that short-range rearrangements, and especially U-turn-like inversions, have been largely underestimated in organelle genomes. In this study, we report that high levels of U-turn-like rearrangements correlate with the appearance of the severe phenotype observed in *why1why3reca1* and to its 59% seed lethality. Since this type of genomic instability also constitute the majority of the DNA rearrangements occurring in human mitochondria, it warrants further investigation into its link to the development of some of poorly understood mitochondrial disorders. Mitochondrial genome instability has indeed been observed in many clinical disorders including Parkinson's disease (Bender et al. 2006; Kraytsberg et al. 2006), inclusion body myositis (Moslemi et al. 1997) and cancer (Ju et al. 2014). In this regard, it will be interesting to evaluate if particular patterns of mitochondrial genomic instability are observed in the context of these disorders. Genomic rearrangements associated to inversions in the human nuclear genome have also been linked to leukemia, autism, and intellectual disability (Pui et al. 1992; Hermetz et al. 2014). It would therefore also be interesting to determine whether such rearrangements also occur through a U-turn-like mechanism and lead to the onset of such phenotypes.

2.6 Methods

Plant material and growth conditions. *Arabidopsis thaliana* (ecotype Columbia-0) mutant lines *reca1-2* (SALK_072979) (Rowan et al. 2010), *polIb-1* (SALK_134274), *why1why3*, and *why1why3polIb-1* were reported previously (Maréchal et al. 2009; Parent et al. 2011). The *reca1-1* (SALK_057982) mutant line was isolated from heterozygous seeds obtained

from ABRC (Alonso et al. 2003). Seeds were sown on soil, vernalized for 3 days at 4°C, and grown under laboratory normal light (100 mmol m^{-2} s^{-1}), at 22°C on a 16-h-day/8-h-dark cycle. Representative photographs were taken at 21 days. To determine statistical significance of *why1why3pollbreca1* lethality, chi-square tests were performed on progeny of *why1why3*$^{-/-}$ *polIb*$^{+/-}$*reca1-1*$^{+/-}$ and *why1why3*$^{-/-}$*polIb*$^{-/-}$*reca1-1*$^{+/-}$ plants. Chi-square p-values were then combined using Stouffer's Z test (Whitlock 2005).

DNA isolation and DNA-seq. For Col-0, *polIb*, *reca1*, *why1why3*, *reca1polIb*, *why1why3polIb* and *why1why3reca1*, total DNA was isolated from ≈400 mg pools of 14-day-old *Arabidopsis* plants using the cetyl trimethyl-ammonium DNA extraction protocol (Weigel and Glazebrook 2002). DNA was fragmented to ≈200-500 bp using S-Series Covaris according to Illumina's specifications. Libraries were prepared using the TruSeq DNA library preparation kit (Illumina) according to manufacturer's instructions. Efficient library generation was then assessed using a Bioanalyzer platform (Agilent) and Illumina MiSeq-QC run was performed. Sequencing was performed using an Illumina HiSeq 2000 using TruSeq SBS v3 chemistry at the Institute for Research in Immunology and Cancer's Genomics Platform (Université de Montréal). Cluster density was targeted at around 600 to 800 kilo clusters mm–2. Sequencing data is available on the NCBI Sequence Read Archive (SRA) under the BioProject Number SRP051208 (Col-0: SRX883065, *polIb*: SRX813508, *reca1*: SRX883066, *why1why3*: SRX883067, *reca1polIb*: SRX883068, *why1why3polIb*: SRX883069, *why1why3reca1*: SRX883070).

Publicly available Illumina whole-genome sequencing datasets. NCBI SRA accession numbers for paired-end Illumina whole-genome sequencing datasets for *Arabidopsis* ecotypes Ts-1 and Ws-2 are, respectively, SRX145018 and SRX145037 (submitted by the SALK Institute for Biological Studies). SRA accession numbers for paired-end Illumina whole-genome sequencing datasets for four human brain samples and four liver samples are, respectively, ERX385572, ERX385573, ERX385574, ERX385575 and ERX385576, ERX385577, ERX385578, ERX385579 (submitted by the Institute for Molecular Bioscience - The University of Queensland). SRA accession numbers for paired-end Illumina whole-genome sequencing datasets for *E. coli* are SRX154301, SRX154337, SRX154338 and SRX154342 (submitted by Indiana University).

Enrichment for reads with potential junctions. The Galaxy online software suite was used to develop a workflow that enriches reads spanning potential junctions from paired-end Illumina datasets (Supplemental Fig. S7) (Goecks et al. 2010). As part of the workflow, quality filtering was performed to keep pairs for which both reads measure at least 40 bases, have an average quality of at least 20 and display no more than 50 bases outside of the quality range (Blankenberg et al. 2010). Using Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009), both reads from each pair were mapped individually against the whole reference genome (GRCh38 with alternate loci for *Homo sapiens* and Galaxy built-in reference genomes Arabidopsis_thaliana_TAIR10 for *Arabidopsis thaliana* and eschColi_K12 for E. coli) to remove pairs of read that do not contain a junction. This also allows the removal of nuclear plastid DNA sequences (NUPTs) and nuclear mitochondrial DNA sequences (NUMTs) that would otherwise be flagged as rearrangements. The alignment was performed without using a subsequence as a seed and parameters were set to a maximum edit distance of 7, a maximum number of gap opens of 1, a maximum number of gap extensions of 3, a mismatch penalty of 3, a gap open penalty of 11 and a gap extension penalty of 4. Pairs with at least one unmapped read were then selected and the first 25 bases of both reads were mapped against the desired organelle genome (GenBank: AP000423.1 for *Arabidopsis* plastid, GenBank: Y08501.2 for Arabidopsis mitochondria, NCBI: NC_012920.1 for human mitochondria and Galaxy built-in reference genome eschColi_K12 for *E. coli*) using Bowtie for Illumina with default parameters (Langmead et al. 2009). Aligning the 25 first bases of both reads is critical, as rearrangement junctions would prevent the alignment of the whole rearranged read. The distance between each alignment position of a pair was then computed and pairs containing an unmapped read were discarded. Only pairs that fulfil one of the following conditions were then kept: 1- Each read of a pair is mapped in opposite orientation, 2- Each read of a pair is mapped in the same orientation with a distance of at least 3 bases. Finally, pairs were filtered to discard those that correspond to DNA fragments spanning junctions created by the different isoforms of the plastid genome or the circular nature of organelle DNA. Workflow statistics are listed in Supplemental Tables S23 to S28. Enriched reads spanning potential junctions are provided in Supplemental Data File 1 for all lines and organisms used in this study.

Analysis of rearranged reads. Sequences of reads containing a potential junction were aligned against the organelle genome (Positions 1 to 128214 of GenBank: AP000423.1 for *Arabidopsis* plastid, GenBank: Y08501.2 and JF729201.1 for *Arabidopsis* mitochondria, NCBI: NC_012920.1 for human mitochondria and NCBI: NC_000913.2 for E. coli) using BLAST+ (Camacho et al. 2009) and the following parameters: "blastn -query potential_junctions.fasta -db organelle_genome.fasta -out output.txt -word_size 10 -max_target_seqs 2 -evalue 0.0001 -outfmt 6". Only reads with two alignments were kept for further analysis. Among the reads that output two alignments, the longest alignment was subtracted to the length of the read, and reads with values of 5 and less were discarded. BLAST+ outputs for rearrangements reads are provided in Supplemental Data File 2 for all lines and organisms used in this study. To ensure that errors in wild-type plastid genome annotations do not generate false-positive rearrangements, the genomic sequence was confirmed using BLAST+ for each locus at which rearrangements are more abundant than 5% of the average coverage. The reference locus sequence was used as the subject sequence to align the total reads of the samples. When rearrangements were found to be in homoplasmy, these rearrangements were considered as false-positives and were discarded. For *Arabidopsis* Col-0 mitochondria, two genome annotations exist (GenBank: Y08501.2 and JF729201.1). To remove false-positives, only reads considered as rearrangements on both annotations were kept. The following algorithms were then used to analyze the rearrangement junctions according to the alignment position output by BLAST+.

Distance measurement algorithm evaluates the difference between the reference genome positions of the 3' end of the first alignment and the 5' end of the second alignment. Overlap of alignments algorithm corresponds to the subtraction of the total read length to the sum of the lengths of both alignments. Rearrangements with an overlap of at least 5 bases are considered to have occurred through the use of microhomology, while the rest are assigned to the "No Microhomology" group. This high cut-off was used to ensure the stringency of microhomology calling. Inversion algorithm corresponds to rearrangements for which the reference genome positions from the 5' end to the 3' end are in ascending order for one alignment and in descending order for the other. All remaining rearrangements are treated as deletions or duplications. U-turn-like rearrangements are defined as those which present an inversion with a distance

parameter of less than 50 bases. Downstream, same base and upstream U-turn-like rearrangements are determined by the position on the reference genome of the 5' end of the second alignment in relation to the 3’ end and direction of the first alignment. The polarity of U-turn-like rearrangements identified in R2 reads has been inverted. Local mapping for read pairs associated to a single rearrangement were selected and aligned on the *Arabidopsis thaliana* (TAIR10) genome using the Golden Helix GenomeBrowse® visualization tool (Version 2.0.7). Representative results for a duplication, a deletion and an inversion are presented in Supplemental Figure S9.

To evaluate the proportion of U-turn-like rearrangements that occurred through the use of imperfect microhomologies, reads corresponding to this type of genomic rearrangement were aligned against the plastid genome using BLAST+ and the following parameters: “blastn -query U-Turns.fasta -db plastid_genome.fasta -out output.txt -word_size 10 -max_target_seqs 2 -evalue 0.002 -outfmt 6 –penalty -1 –gap open 0 –gap extend 2”. The previous algorithms were then used to evaluate the proportion of rearrangements harboring a microhomology at the junction.

The combination of the previous workflow to enrich reads with potential junctions and these algorithms achieved 64.0% sensitivity for detection of reads containing junctions and 97.5% specificity for accurate rearrangement type calling. For sensitivity, 200 randomly chosen reads from the potential junctions obtained following the Galaxy workflow for the plastid DNA were blind-tested individually to assess if they correspond to a rearrangement or not. For specificity, 480 randomly chosen rearrangements analyzed by the algorithm were analyzed for accurate rearrangement type calling. For both sensitivity and specificity, the online BLAST interface (http://blast.st-va.ncbi.nlm.nih.gov/Blast.cgi) was used with the lowest stringency settings.

Plastid sequencing coverage analysis. Pairs with both reads fully aligned against the reference genome using BWA during enrichment for potential junctions were filtered to keep only those mapping the plastid genome. Positions of each read were rounded down to the nearest kb and all reads mapping to the plastid large inverted repeats (IRs) were only assigned to the first IR. The number of reads mapping each 1 kb range was measured and normalized relative to 1,000,000 plastid reads.

Pulse-field gel electrophoresis (PFGE) analysis of the plastid genome. Chloroplasts were isolated from 21-day-old *Arabidopsis* plants as described previously (Little 1997). Isolated chloroplasts were resuspended in homogenization buffer and then mixed 1:1 to 45oC 1% low melting point agarose in TE buffer and allowed to fix at 4oC. Isolated chloroplasts concentration were adjusted according to a low-cycle amplification of a DNA fragment of the hypothetical protein RF2 (YCF2) with the following primers: YCF2FOR, GAT CTC TGA GAG CTG TTT CCG; YCF2REV, TGT TTC GCC TCT TAC TCG GAG. Agarose plugs were then soaked overnight at 50oC in lysis buffer (0.45 M EDTA pH 8.0, 1% (w/v) sarkosyl, 1 mg/mL proteinase K) and washed in storage solution (0.45 M EDTA pH 8.0, 1% (w/v) sarkosyl). Migration was performed in 1.5% agarose gel in 0.5X TBE buffer for 46 hours at 12oC using a Bio-Rad CHEF-DR® III system. Pulses switch times were set to 120 seconds at 5 V/cm using a 120 degrees angle. The gel was then soaked successively in 0.5X TBE buffer supplemented with 1 µg/mL RNase A for 2 hours at room temperature, in 0.5X TBE buffer supplemented with 5 µg/mL ethidium bromide for thirty minutes and finally in 0.5X TBE buffer to wash the gel. Southern hybridization was then performed as described previously (Maréchal et al. 2009) using a chloroplast probe amplified using the following primers: 49741FOR, CCT TAC GTA AAG GCC ACC CTA; 54551REV, TGG GAC GCA TAA CCG GAT ATG.

Quantitative PCR analysis of ptDNA levels. Total DNA was isolated from 14-day-old Arabidopsis Col-0, *why1why3polIb* and *why1why3reca1* plants using the cetyl trimethyl-ammonium DNA extraction protocol (Weigel and Glazebrook 2002). Primers used for qPCR were calibrated to ensure the amplification of a unique PCR product and efficiency between 1.90 and 2.05. Every reaction was carried out on biological and technical triplicates relative to the amplification of nuclear DNA. Primers sequences are as follows: 7660FOR, TGA TCC AGG ACG TAA TCC GGG AC; 7802REV, CGA ATC CCT CTC TTT CCC CTT CTC C; 45345FOR, TTG GCA ATT CCT CAG GGG CAG; 45525REV, TTG ACT ATT CCT CAA GCG CGC C; 81312FOR, AGC TAC CCA ATA CTC AGG GGA TCC; 81460REV, AAA TAG AAG CAG GGC GAC GCG; nucDNA-FOR, GTT GAA GCC TCC GTT CCC TGC TA; nucDNA-REV, CTC TTC CAC CGT GCA TGG CTT GT. The Power SYBR® Green PCR Master Mix (Applied Biosystems) was used according to manufacturer's instructions. qPCR

experiments and analysis were carried out using LightCycler 480 (Roche) and the LightCycler 480 software version 1.5, respectively.

Quantitative PCR analysis of *RECA1* expression in *reca1* mutants. Quantitative PCR analysis of *RECA1* expression was performed as described previously for 21-day-old plants (Lepage et al. 2013). Every reaction was carried out on biological and technical triplicates relative to the amplification of beta tubulin. Primers used for qRT-PCR were calibrated to ensure the amplification of a unique PCR product and efficiency between 1.90 and 2.05. Primers sequences are as follows: RECA1FOR, GGT GGA GGC CTA CCA AAG GG; RECA1REV, GGT GGA GGC CTA CCA AAG GG; BetaTubFOR, TCG TTG GGA GGA GGC ACA GGT; BetaTubREV, GCT GAG TTT GAG GGT ACG GAA GCA G. The Power SYBR® Green PCR Master Mix (Applied Biosystems) was used according to manufacturer's instructions. qPCR experiments and analysis were carried out using LightCycler 480 (Roche) and the LightCycler 480 software version 1.5, respectively.

Data access. All newly generated sequencing data is freely available on the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under the BioProject Number SRP051208 (Col-0: SRX883065, *polIb*: SRX813508, *reca1*: SRX883066, *why1why3*: SRX883067, *reca1polIb*: SRX883068, *why1why3polIb*: SRX883069, *why1why3reca1*: SRX883070). The Galaxy workflow is freely available on Galaxy's Published Workflows section under the title "Rearrangement Junction Detection".

2.7 Acknowledgements

We thank M. Arteau, M. Valach and the Galaxy Team for assistance with the DNA-seq, PFGE and data processing, respectively. We also thank T. Vincent for his help in isolating mutant lines. This work was supported by scholarships from the Natural Sciences and Engineering Research Council of Canada (NSERC) to É.L., É.Z. and S.T.B., from the Fonds de Recherche du Québec – Nature et Technologies to S.T.B. and grants from NSERC to N.B.

Author contributions

É.Z. isolated plant DNA for sequencing and performed the PFGE and qPCR experiments. É.Z., É.L. and S.T.B. developed the bioinformatics approach. É.Z., É.L., S.T.B., S.T. and N.B. analyzed and interpreted the data. É.Z., É.L., S.T.B. and N.B. wrote the manuscript.

Disclosure declaration

The authors declare they have no conflict of interest.

2.8 References

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, et al. 2003. Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**: 653–657.

Ameur A, Stewart JB, Freyer C, Hagstrom E, Ingman M, Larsson NG, Gyllensten U. 2011. Ultra-deep sequencing of mouse mitochondrial DNA: mutational patterns and their origins. *PLoS Genet* **7**: e1002028.

Bender A, Krishnan KJ, Morris CM, Taylor GA, Reeve AK, Perry RH, Jaros E, Hersheson JS, Betts J, Klopstock T, et al. 2006. High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease. *Nat Genet* **38**: 515–517.

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy T. 2010. Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**: 1783–1785.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Cappadocia L, Maréchal A, Parent JS, Lepage E, Sygusch J, Brisson N. 2010. Crystal structures of DNA-Whirly complexes and their role in Arabidopsis organelle genome repair. *Plant Cell* **22**: 1849–1867.

Costes A, Lambert SA. 2012. Homologous recombination as a replication fork escort: fork-protection and recovery. *Biomolecules* **3**: 39–71.

Ferrarini M, Moretto M, Ward JA, Surbanovski N, Stevanovic V, Giongo L, Viola R, Cavalieri D, Velasco R, Cestaro A, et al. 2013. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics* **14**: 670.

Goecks J, Nekrutenko A, Taylor J, Galaxy T. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent com- putational research in the life sciences. *Genome Biol* **11**: R86.

Hastings PJ, Ira G, Lupski JR. 2009. A microhomology-mediated break-in- duced replication model for the origin of human copy number varia- tion. *PLoS Genet* **5**: e1000327.

Hermetz KE, Newman S, Conneely KN, Martin CL, Ballif BC, Shaffer LG, Cody JD, Rudd MK. 2014. Large inverted duplications in the human genome form via a fold-back mechanism. *PLoS Genet* **10**: e1004139.

Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M, Davies HR, Papaemmanuil E, Gundem G, Shlien A, et al. 2014. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**: e02935.

Krause K, Kilbienski I, Mulisch M, Rodiger A, Schafer A, Krupinska K. 2005. DNA-binding proteins of the Whirly family in Arabidopsis thaliana are targeted to the organelles. *FEBS Lett* **579**: 3707–3712.

Kraytsberg Y, Kudryavtseva E, McKee AC, Geula C, KowallNW, Khrapko K. 2006. Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. *Nat Genet* **38**: 518–520.

Kugelberg E, Kofoid E, Andersson DI, Lu Y, Mellor J, Roth FP, Roth JR. 2010. The tandem inversion duplication in *Salmonella enterica*: selection drives unstable precursors to final mutation types. *Genetics* **185**: 65–80.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**: 1235–1247.

Lepage E, Zampini E, Brisson N. 2013. Plastid genome instability leads to reactive oxygen species production and plastid-to-nucleus retrograde signaling in Arabidopsis. *Plant Physiol* **163**: 867–881.

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

Little SE. 1997. *Large- or small-scale isolation of chloroplasts using the Avanti® J series of high performance centrifuges*. Beckman Instruments, Inc., Fullerton, CA.

Lundin C, Erixon K, Arnaudeau C, Schultz N, Jenssen D, Meuth M, Helleday T. 2002. Different roles for nonhomologous end joining and homologous recombination following replication arrest in mammalian cells. *Mol Cell Biol* **22**: 5869–5878.

Maréchal A, Parent JS, Veronneau-Lafortune F, Joyeux A, Lang BF, Brisson N. 2009. Whirly proteins maintain plastid genome stability in *Arabidopsis*. *Proc Natl Acad Sci* **106**: 14693–14698.

Miller-Messmer M, Kuhn K, Bichara M, Le Ret M, Imbault P, Gualberto JM. 2012. RecA-dependent DNA repair results in increased heteroplasmy of the *Arabidopsis* mitochondrial genome. *Plant Physiol* **159**: 211–226.

Mizuno K, Lambert S, Baldacci G, Murray JM, Carr AM. 2009. Nearby inverted repeats fuse to generate acentric and dicentric palindromic chromosomes by a replication template exchange mechanism. *Genes Dev* **23**: 2876–2886.

Mizuno K, Miyabe I, Schalbetter SA, Carr AM, Murray JM. 2013. Recombination-restarted replication makes inverted chromosome fusions at inverted repeats. *Nature* **493**: 246–249.

Moslemi AR, Lindberg C, Oldfors A. 1997. Analysis of multiple mitochondrial DNA deletions in inclusion body myositis. *Hum Mutat* **10**: 381–386.

Ogihara Y, Terachi T, Sasakuma T. 1988. Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. *Proc Natl Acad Sci* **85**: 8573–8577.

Paek AL, Kaochar S, Jones H, Elezaby A, Shanks L, Weinert T. 2009. Fusion of nearby inverted repeats by a replication-based mechanism leads to formation of dicentric and acentric chromosomes that cause genome instability in budding yeast. *Genes Dev* **23**: 2861–2875.

Parent JS, Lepage E, Brisson N. 2011. Divergent roles for the two PolI-like organelle DNA polymerases of Arabidopsis. *Plant Physiol* **156**: 254– 262.

Pui CH, Carroll AJ, Raimondi SC, Schell MJ, Head DR, Shuster JJ, CristWM, Borowitz MJ, Link MP, Behm FG, et al. 1992. Isochromosomes in childhood acute lymphoblastic leukemia: a collaborative study of 83 cases. *Blood* **79**: 2384–2391.

Robu ME, Inman RB, Cox MM. 2001. RecA protein promotes the regression of stalled replication forks in vitro. *Proc Natl Acad Sci* **98**: 8211–8218.

RowanBA, Oldenburg DJ, Bendich AJ. 2010. RecA maintains the integrity of chloroplast DNA molecules in *Arabidopsis*. *J Exp Bot* **61**: 2575–2588.

Seier T, Zilberberg G, ZeigerDM, Lovett ST. 2012. Azidothymidine and other chain terminators are mutagenic for template-switch-generated genetic mutations. *Proc Natl Acad Sci* **109**: 6171–6174.

Seigneur M, Ehrlich SD, Michel B. 2000. RuvABC-dependent double-strand breaks in *dnaBts* mutants require recA. *Mol Microbiol* **38**: 565–574.

Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA. 2007. Plant mitochondrial recombination surveillance requires unusual *RecA* and *MutS* homologs. *Plant Cell* **19**: 1251–1264.

Slager J, Kjos M, Attaiech L, Veening JW. 2014. Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin. *Cell* **157**: 395–406.

Small ID, Isaac PG, Leaver CJ. 1987. Stoichiometric differences in DNA molecules containing the *atpA* gene suggest mechanisms for the generation of mitochondrial genome diversity in maize. *EMBO J* **6**: 865–869.

Vermulst M, Wanagat J, Kujoth GC, Bielas JH, Rabinovitch PS, Prolla TA, Loeb LA. 2008. DNA deletions and clonal mutations drive pre- mature aging in mitochondrial mutator mice. *Nat Genet* **40**: 392– 394.

Weigel D, Glazebrook J. 2002. *Arabidopsis: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Weigel D, Mott R. 2009. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol* **10**: 107.

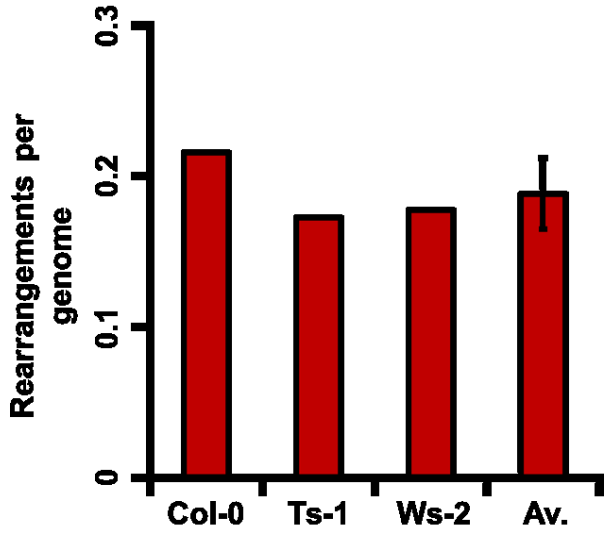
Whitlock MC. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol* **18**: 1368–1373.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc Natl Acad Sci* **84**: 9054–9058.

Wu J, Liu B, Cheng F, Ramchiary N, Choi SR, Lim YP, Wang XW. 2012. Sequencing of chloroplast genome using whole cellular DNA and solexa sequencing technology. *Front Plant Sci* **3**: 243.

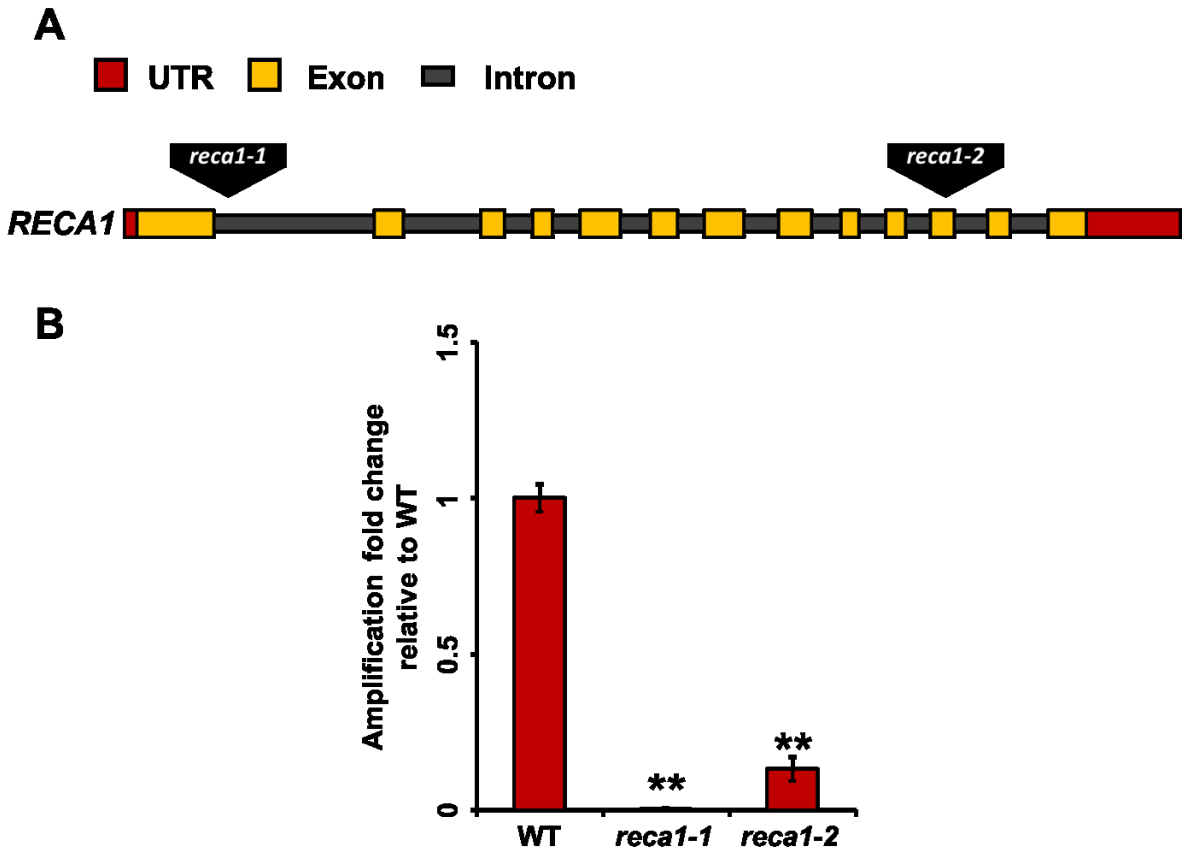
Zeman MK, Cimprich KA. 2014. Causes and consequences of replication stress. *Nat Cell Biol* **16**: 2–9.

2.9 Supplemental Material



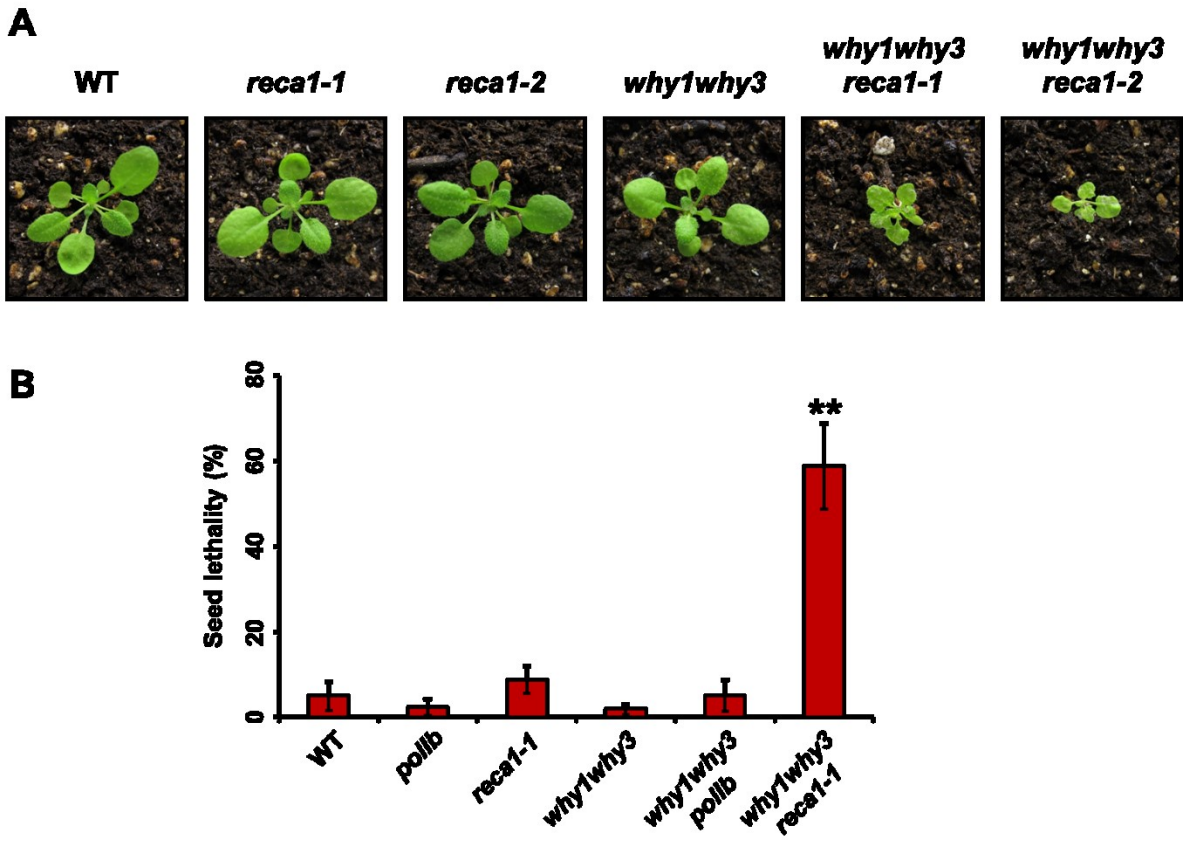
GR Supplemental Figure S1. Level of plastid DNA rearrangements in three *Arabidopsis* ecotypes.

Total number of rearrangements identified for each of the Col-0, Ts-1 and Ws-2 ecotypes. Y axis represents the number of rearrangements per 1,000,000 plastid reads. Av.: average of the three ecotypes. The error bar represents the standard deviation.



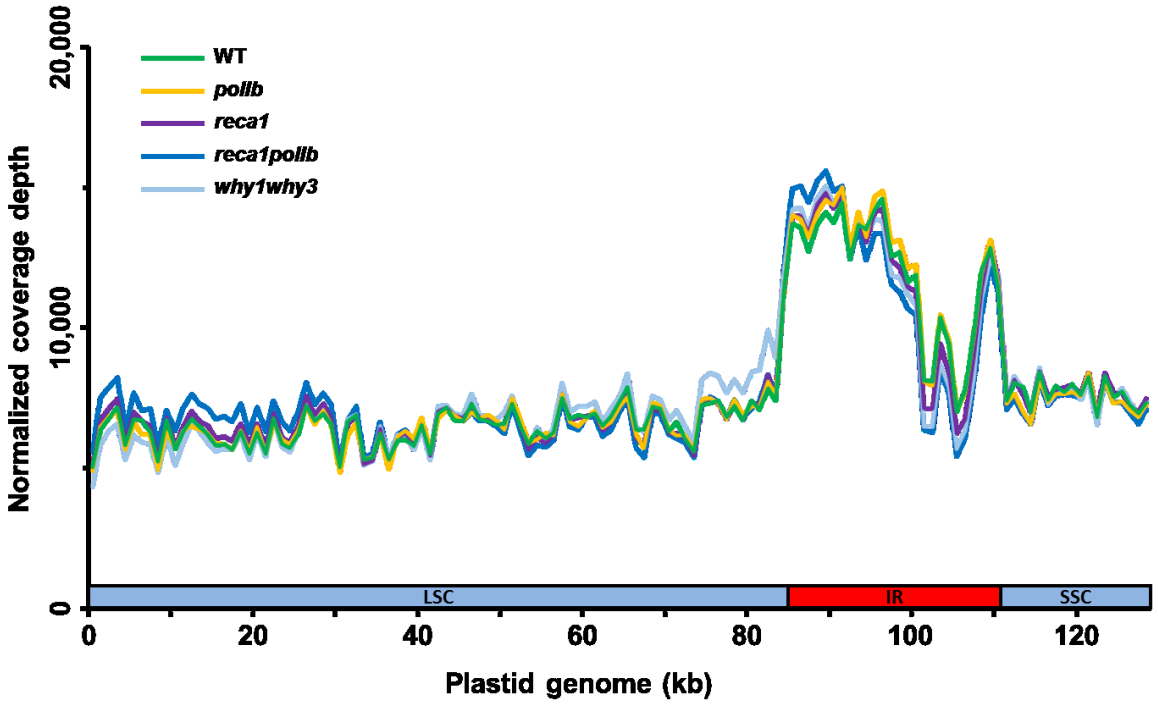
GR Supplemental Figure S2. Characterization of the *reca1-1* and *reca1-2* T-DNA insertion mutant lines.

(*A*) Schematic representation of the *RECA1* gene and insertion positions of the *reca1-1* and *reca1-2* insertions. (*B*) Quantitative PCR measurement of the *RECA1* expression levels for the *reca1-1* and *reca1-2* mutant lines relative to WT plants. Error bars represent the standard error of the mean of three biological replicates. Asterisks indicate a significant difference of a Student's *t* test *p*-value ≤ 0.01 with the WT.



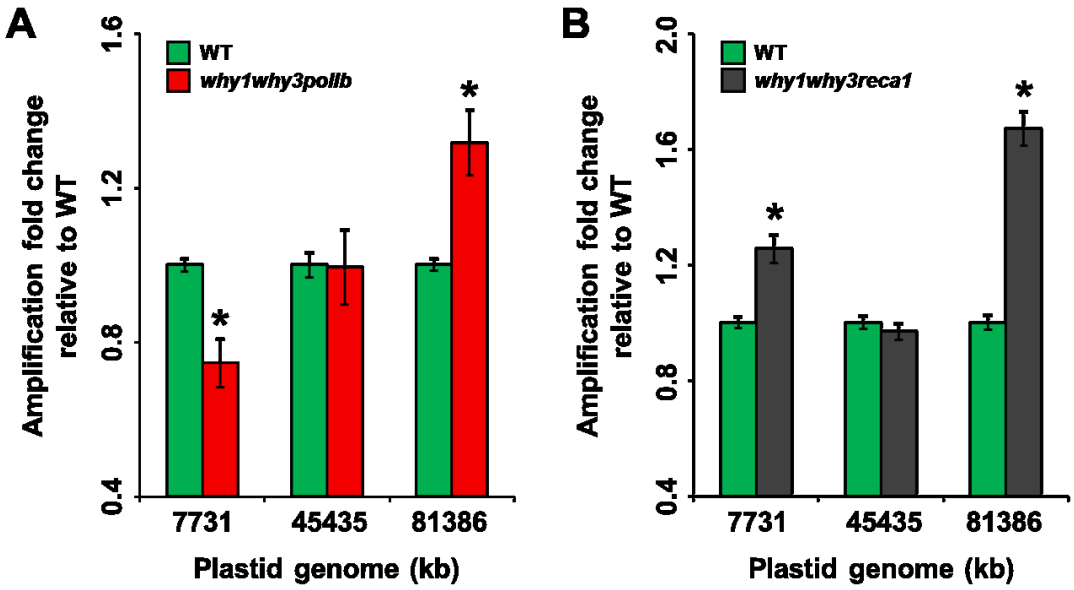
GR Supplemental Figure S3. Characterization of the *why1why3reca1* phenotype.

(*A*) Representative photographs of 21-day-old WT, *reca1-1*, *reca1-2*, *why1why3*, *why1why3reca1-1* and *why1why3reca1-2 Arabidopsis* mutant plants. (*B*) Proportion of non-germinated seeds of WT, *pollb*, *reca1-1*, *why1why3*, *why1why3polIb* and *why1why3reca1-1*, four days after vernalization. Two asterisks indicate a significant difference of a Student's *t* test *p*-value ≤ 0.01 with the WT.



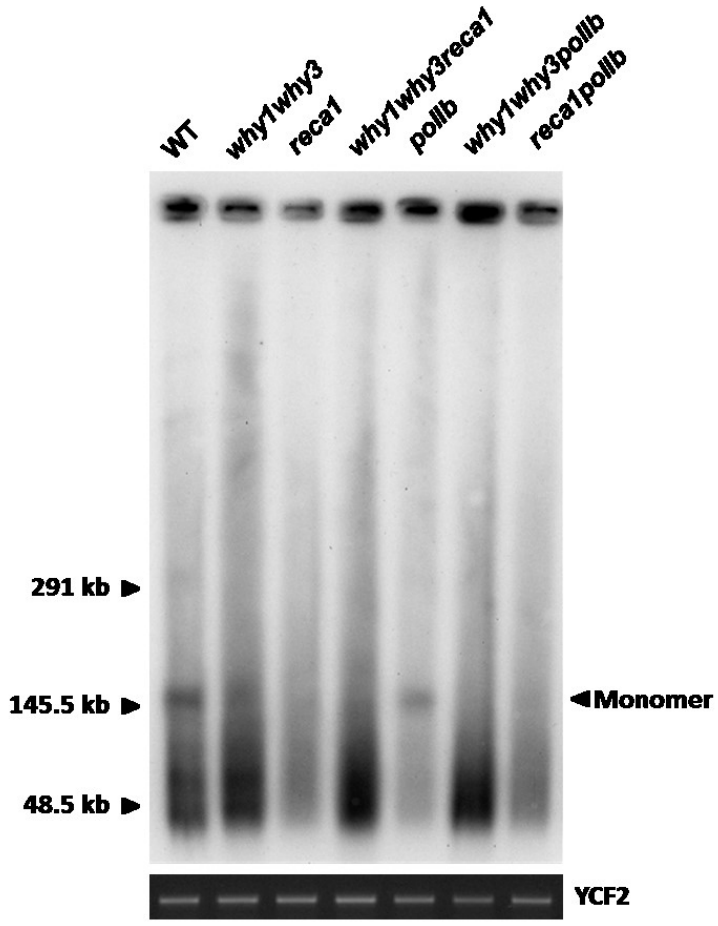
GR Supplemental Figure S4. Plastid DNA sequencing coverage curves for WT, *polIb*, *reca1*, *reca1polIb* and *why1why3 Arabidopsis* mutant plants.

Plastid sequencing coverage of pools of 14-day-old *Arabidopsis* seedlings of the indicated genotypes. Positions were rounded down to 1 kb. All reads mapping to the plastid large inverted repeats (IRs) were only assigned to the first IR. Y axis represents the number of reads per 1,000,000 total plastid reads. The plastid large-single copy region (LSC), the first IR, and the small-single copy region (SSC) are depicted as a long blue bar, a red bar and a short blue bar, respectively.



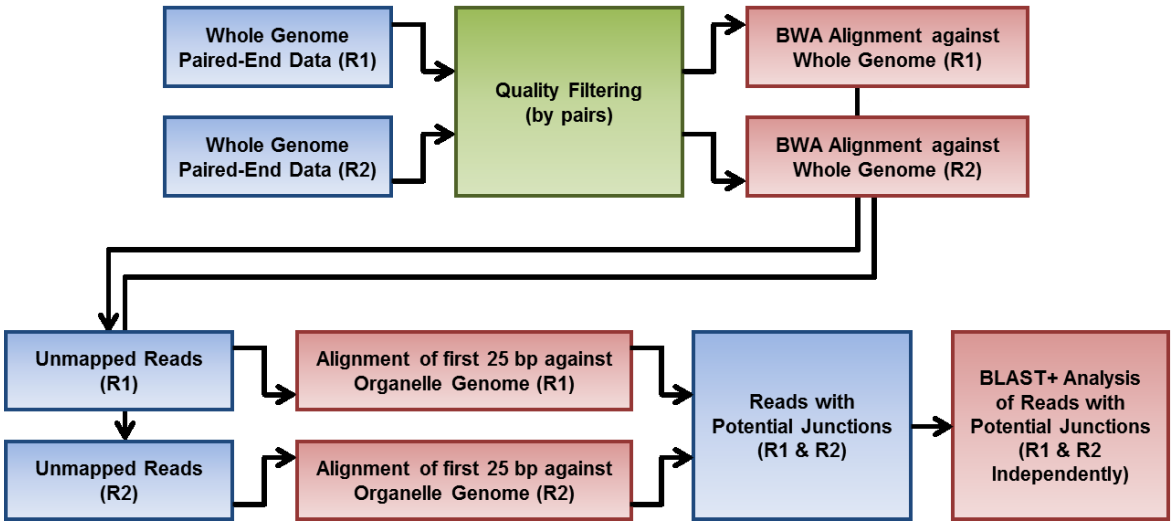
GR Supplemental Figure S5. Plastid DNA quantification at three locations of the genome in WT, *why1why3polIb* and *why1why3reca1* plants.

(*A*) Quantitative PCR measurement of ptDNA levels at three sites of the LSC in *why1why3polIb* relative to WT plants. (*B*) Quantitative PCR measurement of ptDNA levels at three sites of the LSC in *why1why3reca1* relative to WT plants. Error bars represent the standard error of the mean of three biological replicates. Asterisks indicate a significant difference of a Student's *t* test *p*-value ≤ 0.05 with the WT.



GR Supplemental Figure S6. Visualization of the distribution of the various forms of the plastid genome.

Pulsed-Field Gel Electrophoresis (PFGE) analysis of ptDNA in the indicated genotypes. The amount of DNA loaded in each well was equilibrated relative to the amplification of a *YCF2* fragment.

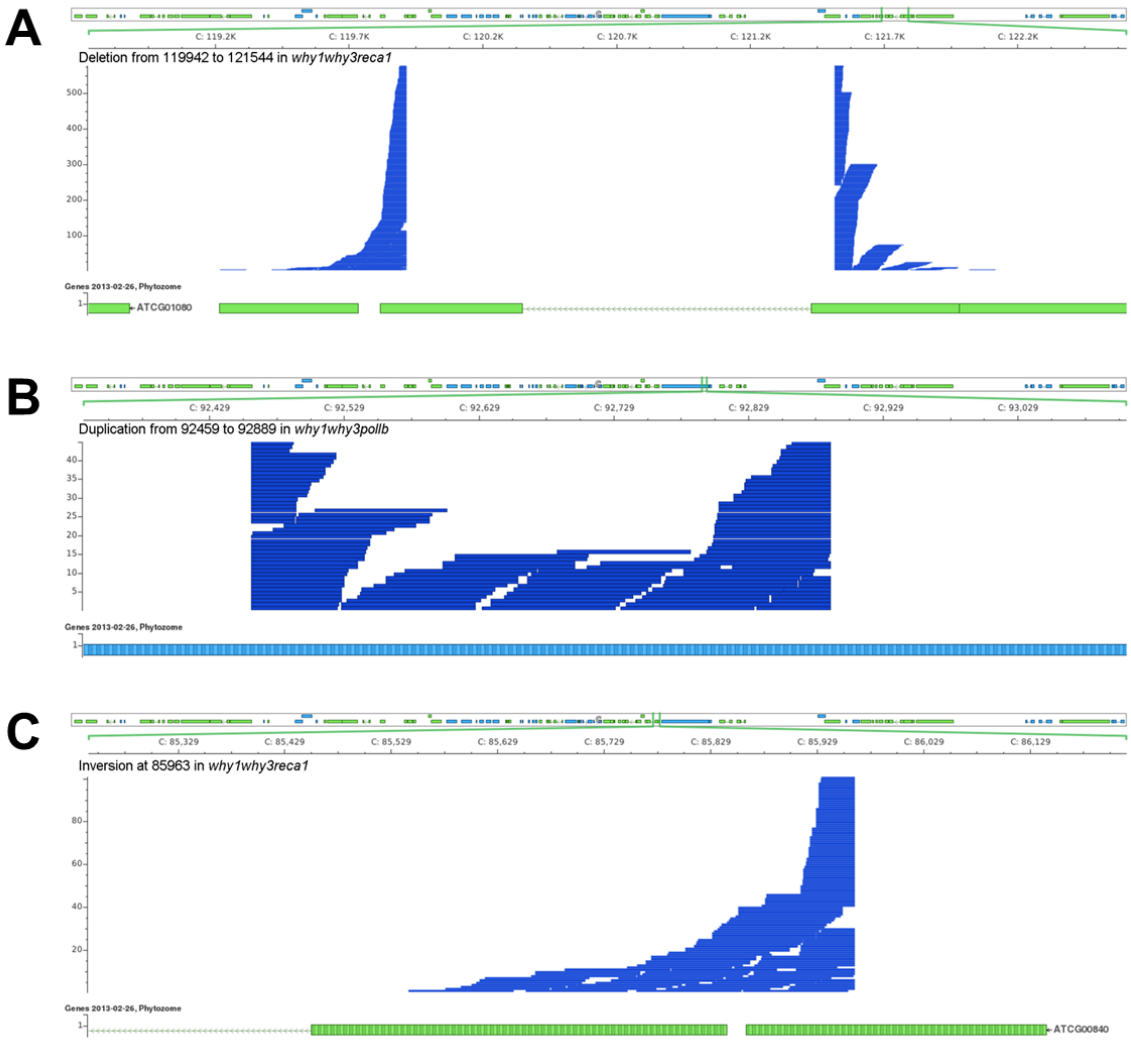


GR Supplemental Figure S7. Schematic representation of the analysis workflow.
Blue background boxes represent text manipulation steps, while the green and red backgrounds stand for quality filtering and mapping, respectively. Sequencing was performed on both ends of DNA fragments (R1 and R2). BWA: Burrows-Wheeler Aligner. R1 & R2: Paired-end sequencing read1 and read2.

	PCR and Southern Blot	Reporter Systems	Paired-End Analysis	Junction Analysis
Genome-wide	**No** *Primers and probe are specific to a single region*	**No** *Only the inserted region is observed*	**Yes**	**Yes**
Can be performed without genetic modification	**Yes**	**No** *Requires insertion of exogenous DNA into an organism's genome*	**Yes**	**Yes**
Allows comparison of samples	**Yes**	**Yes**	**No** *Detection is heavily dependent on DNA fragment length, which varies between samples*	**Yes**
Provides information about mechanism	**Yes** *Individual rearrangements need to be cloned and sequenced to obtain sequence*	**Yes** *Individual parameters can be varied to assess their importance*	**No**	**Yes** *Exact junction sequence provides some information about mechanism*
Provides a view of both short- and long-range rearrangements	**No** *Primers and probe are specific to a single region*	**No** *A single mechanism is observed at any time*	**No** *Bias toward long-range rearrangements*	**Yes**
Allows detection of short-ranged rearrangements	**Yes**	**Yes**	**No** *Detection of short-range rearrangements is limited by DNA fragment size*	**Yes**

GR Supplemental Figure S8. Comparison of the techniques used to detect genome rearrangements.

Details are provided when relevant.



GR Supplemental Figure S9. Local mapping of read pairs associated to specific genome rearrangements.

The top box in each panel shows a map of the complete plastid genome on which genes on the forward strand are presented in blue and those on the reverse strand in green. The position encompassed by the zoom on the genome are presented below. The mapping of each read is presented as a blue rectangle. The bottom box shows a map of genes in the zoom. A) Representative local mapping for a deletion. This deletion was observed 537 times in *why1why3reca1*. B) Representative local mapping for a duplication. This duplication was observed 77 times in *why1why3pollb*. C) Representative local mapping for an inversion. This inversion was observed 81 times in *why1why3reca1*.

	Col-0	*polIb*	*reca1*	*reca1polIb*	*why1why3*	*why1why3 polIb*	*why1why3 reca1*
Read Pairs	31241719	32810928	30409865	31132957	31438219	32941816	31699029
Pairs with average quality >20	28771745	31343328	28967978	29738528	29944699	31308538	30156212
Plastid Pairs (%)	17,69	16,77	19,51	13,35	20,35	19,16	20,58

GR **Supplemental Table S23**. **Workflow Statistics for *Arabidopsis* plastid DNA rearrangements.**

	Ts-1 (SRX145018)	Ws-2 (SRX145037)
Read Pairs	30012578	45521342
Pairs with average quality >20	26774195	40165795
Plastid Pairs (%)	9,89	9,54

GR Supplemental Table S24. Workflow Statistics for *Arabidopsis* ecotypes Ts-1 and Ws-2 plastid DNA rearrangements.

	Col-0
Read Pairs	31241719
Pairs with average quality >20	28771745
Mitochondria Pairs (%)	1,27

GR Supplemental Table S25. Workflow Statistics for *Arabidopsis* mitochondria DNA rearrangements.

	erx385572	erx385573	erx385574	erx385575
Read Pairs	78908272	79738231	54905866	53255234
Pairs with average quality >20	54109701	54783744	52966448	51137017
Mitochondria Pairs (%)	0,55	0,55	0,61	0,64

GR Supplemental Table S26. Workflow Statistics for human brain mitochondria DNA rearrangements.

	erx385576	erx385577	erx385578	erx385579
Read Pairs	74027979	75446145	47333976	46047382
Pairs with average quality >20	51040536	52250953	45679517	44520994
Mitochondria Pairs (%)	0,45	0,49	0,52	0,53

GR Supplemental Table S27. Workflow Statistics for human liver mitochondria DNA rearrangements.

	SRX154301	SRX154337	SRX154338	SRX154342
Read Pairs	2611112	2611112	2611112	2611112
Pairs with average quality >20	2611099	2611096	2611098	2611089
E. coli Pairs (%)	99,85	99,83	99,85	99,84

GR Supplemental Table S28. Workflow Statistics for *E. coli* DNA rearrangements.

3. Deuxième Article: Replication U-turns restart stabilized stalled replication forks in an error-prone manner

En préparation pour soumission à *Genome Research*.

Samuel Tremblay-Belzile[1], Audrey Loubert-Hudon[1], Juliana Pérez Di Giorgio[1], Alain Verreault[2], James G. Omichinski[1] and Normand Brisson[1]

Corresponding author: samuel.tremblay-belzile@umontreal.ca

[1] Department of biochemistry and molecular medicine, Université de Montréal, Montréal, Québec, H3T1J4, Canada

[2] Institute for Research in Immunology and Cancer and Department of pathology and cellular biology, Université de Montréal, Montréal, Québec, H3C 3J7, Canada

3.1 Mise en contexte

Pendant la mise au point de l'approche bio-informatique utilisée pour détecter les réarrangements au chapitre 2, nous avons testé ses limites quant à la taille des génomes analysés. Bien que notre méthode ait été efficace pour les génomes du chloroplaste (154 kb) et de la mitochondrie (367 kb) d'*Arabidopsis*, de la mitochondrie humaine (16.5 kb), et d'*E. coli* (4.6 Mb), elle n'était pas applicable aux génomes nucléaires d'*Arabidopsis* (135 Mb) ou de *S. cerevisiae* (12.5 Mb). L'enrichissement des séquences contenant des réarrangements dans Galaxy et l'obtention d'alignements partiels par BLAST+ n'étaient pas limités par la taille du génome, mais l'obtention d'une multitude d'alignements possibles rendait inefficace la détection à grande échelle de réarrangements à partir de logiciels comme Excel. Il était par contre possible d'analyser individuellement les alignements partiels de chaque séquence pour identifier le réarrangement. J'ai ainsi décidé d'automatiser l'analyse des alignements partiels, afin d'éliminer la limite de taille du génome. Après un test concluant sur le génome nucléaire d'*Arabidopsis*, j'ai optimisé le temps nécessaire à l'analyse pour permettre son utilisation sur le génome nucléaire humain (3 Gb). Les résultats obtenus ont montré une plus grande sensibilité de détection que les approches déjà publiées. Pour complémenter les données de tissus humains, je me suis tourné vers des souches de levure mutantes. La possibilité d'enlever des gènes de réparation de l'ADN et de soumettre la levure à des stress génotoxiques a permis de mieux comprendre les mécanismes impliqués dans les réarrangements détectés. La petite taille du génome a également réduit les demandes en temps et en ressources informatiques pour l'analyse des données.

Dans ce projet, j'ai conçu et écrit le programme pour la détection des réarrangements, que j'ai ensuite optimisé pour permettre son utilisation sur le génome humain. J'ai développé l'expérience sur les mutants de levure et les différentes conditions de croissance. J'ai analysé les résultats et rédigé le manuscrit. Audrey Loubert-Hudon a réalisé les expériences en laboratoire avec les souches de levure, isolé l'ADN et aidé à analyser les données. Juliana Pérez Di Giorgio a aidé à analyser les données. Alain Verreault, James G. Omichinski et Normand Brisson ont participé à l'élaboration du projet. J.P.D.G, J.G.O., N.B. et moi avons contribué à la rédaction du manuscrit.

3.2 Abstract

The identification of structural variations in genomes using next-generation sequencing data greatly facilitates the study of genetic and genomic diseases. This type of approach also provides interesting new ways to examine DNA repair, recombination, and replication to better understand sources of genomic instability. We therefore developed SCARR (Systematic Combination of Alignments to Recreate Rearrangements) to identify DNA rearrangements, and used it to examine replication U-turns in human and budding yeast genomes. SCARR exceeds the sensitivity of previous approaches, and identifies rearrangements genome-wide and with base-pair resolution, which provides insight into the mechanisms involved. We find that U-turns allow an error-prone restart of stalled replication forks without requiring any sequence homology. Our results show they are favored by the stabilization of stalled forks by Rad9 and inhibited by Exo1- and Mre11-mediated DNA degradation. U-turns may therefore contribute to chemoresistance in BRCA2-deficient cells by allowing replication fork restart without any double-strand break intermediates. U-turns result in short-range inversions that create complex structural variants, and fit a model that has been proposed to explain differences between the genomes of chimpanzees and humans. We also find that inversions occur at longer ranges through end-joining mechanisms that require Rad51 and Mre11 in G1 phase, and through snap-back replication following extensive Exo1- and Sgs1-dependent end-resection.

3.3 Introduction

Genomic instability results from the inability of DNA processes to perfectly preserve the native genome sequence. This may result in single nucleotide variants if only one base is changed, or structural variants (SVs) if longer sequences are affected. SVs include deletions, tandem duplications, inversions and translocations, and are associated to a variety of genomic diseases (Carvalho and Lupski 2016). For this reason, the study of genomic diseases not only involves the identification of SVs (Sudmant et al. 2015), but also the identification of mechanisms that lead to their formation (Carvalho and Lupski 2016). These range from non-allelic homologous recombination (NAHR), which involves long homologous sequences (Cardoso et al. 2016), to non-homologous end-joining (NHEJ), which requires no homology at all (Chang et al. 2017). Other mechanisms, such as microhomology-mediated recombination (MHMR) and alternative end-joining (alt-EJ) involve short homologous sequences termed microhomologies (Hastings et al. 2009a; Wang and Xu 2017). Although these mechanisms predominantly occur as a result of double-strand breaks (DSBs), genomic instability can also result from replication-dependent mechanisms that do not require breaks in the DNA backbone (Hastings et al. 2009b). Deletions and duplications may occur as a result of replication slippage, where the DNA polymerase jumps to a different position on the same template, whereas fork stalling and template switching (FoSTeS) can lead to more complex rearrangements (Hastings et al. 2009b). These complex events pose additional challenges for detection methods, and are therefore not as well understood as other types of SVs.

The rapidly expanding repositories of next-generation sequencing data have opened many new avenues to understanding genomic instability and its roles in evolution and genomic diseases. A landmark study in the field identified tens of thousands of SVs within the human genome (Sudmant et al. 2015), which included deletions, duplications, insertions, as well as a small number of inversions. New computational approaches to detect and investigate SVs continue to be developed to improve on detection rates and to focus on specific types of SVs. Because the detection of inversions is still lagging behind the detection of other rearrangement types, it has been the focus recently for the development of new tools (Trappe et al. 2014; Chen et al. 2016; He et al. 2016). These methods score very high on benchmark tests for sensitivity, and contribute to the detection of SVs that had been almost undetectable due to the limitations

of previous detection protocols. However, benchmark tests for inversions usually involve replacing reference sequence fragments with their reverse complement, while mechanisms known to generate inversions often lead to more complex SVs (Tremblay-Belzile et al. 2015). Detection sensitivity on simulated data may therefore not be representative of the effectiveness of an approach on real datasets.

To avoid a bias for simple inversions, we developed SCARR (Systematic Combination of Alignments to Recreate Rearrangements), which uses no *a priori* assumptions about the relative positions of DNA sequences around a breakpoint, and instead iterates through all possible genome-wide alignment combinations to find the best match. Using this approach, it is possible to identify deletions, tandem duplications and inversions at base-pair resolution. Consequently, SCARR is particularly well adapted to the detection of rare events, such as SVs that cannot be successfully replicated or result in cells that are not viable. This allows the study of DNA rearrangements based on a number of parameters, such as microhomology usage, base-pair distance between the rearranged sequences, or SV type. In addition to linking specific SVs to phenotypes, it can therefore also examine how specific factors (DNA repair proteins, replication stresses, DNA damage) impact DNA stability in greater detail. In this study, we used SCARR on datasets from healthy human tissues and several yeast mutant strains to examine patterns and mechanisms linked to short- and long-range inversions.

3.4 Results

3.4.1 SCARR sensitivity exceeds previous approaches for deletions, duplications, and inversions

To determine whether SCARR could yield more sensitive detection of DNA rearrangements than previously-published methods, we used SVsim and WGSIM to generate simulated next-generation sequencing datasets of a genome containing rearrangements, as previously described (Layer et al. 2014). SVsim generates rearrangements in a reference genome, and WGSIM simulates sequencing data from the resulting file. We then identified SVs in the datasets using SCARR to estimate its sensitivity and rate of false discovery. Even at 1X sequencing coverage, SCARR identified over 40% of deletions, duplications, and inversions,

while the false discovery rate remained around 10% (Supplementary Figure S1). This sensitivity is higher than almost all of the methods tested by Layer et al. (2014) at 2X coverage, with the exception of the sensitivity obtained with LUMPY for inversions. Since SCARR identifies all rearrangements at base-pair resolution, the data obtained can be used quantitatively. Its actual sensitivity therefore reaches close to 50% if rearrangements identified multiple times are taken into consideration. We performed the same analysis on simulated data containing no SVs as an additional control. This revealed that approximately half of the false positives are obtained as a result of genuine SVs being misidentified, whereas the other half arise as a result of mapping ambiguities in the reference genome independently of any SVs (Supplementary Figure S2).

To test SCARR on real data, we analyzed public datasets of whole genome sequencing of a healthy human brain and liver as paired-end reads of 101 bases. This produced surprisingly high amounts of rearrangements for the sequencing coverage of the initial datasets, with on average over 17000 rearrangement junctions per genome copy (Table 1 and Supplementary Files 1-2). Following detection, SCARR classifies rearrangements as deletions, duplications, inversions or translocations. For each dataset, we found that over 7000 SVs per genome copy are either deletions or duplications of less than 50 bp, and this accounts for approximately 40% of the SVs detected. Interestingly, the number of each type of rearrangement relative to coverage is very similar between the two datasets, with the exception of inversions, which are approximately 50% more common in the brain dataset.

To determine the effect of read length on rearrangement detection using SCARR, we sequenced DNA from a healthy human brain and spleen as paired-end reads of 151 bases. These datasets yielded approximately twice as many deletions, duplications, and inversions per genome copy as the shorter reads from the public datasets (Table 1 and Supplementary Files 3-4). This improvement in sensitivity is most likely the result of better sequence alignments in longer reads, which helps compensate for the higher mutation rates found at rearrangement junctions. The added sequence length also increases the chance of successful BLAST+ alignments that do not overlap, with a lesser impact on the overlapping reads that result from sequence homology. This leads to a higher proportion of junctions that share little to no homology when longer reads are used (Table 1).

	Brain 1	Liver	Brain 2	Spleen
Read length (bases)	101	101	151	151
Genome coverage	13.42	12.21	6.17	5.57
Rearrangements				
Total	236728	201932	343899	273099
Unique	170292	142117	303107	235834
Total (per genome)	17638.9	16527.9	55741.6	49042.0
Unique (per genome)	12688.7	11632.1	49129.7	42350.1
Microhomology ≥ 5 bp (%)	84.8	85.2	55.2	68.7
Total del/dup < 50 bp (per genome)	7038.0	7017.5	11573.0	11715.7
Deletions (per genome)				
Total	5900.0	5789.5	9811.9	9983.2
Unique	2996.0	2965.0	6969.7	6907.2
Duplications (per genome)				
Total	2780.0	2792.7	6289.0	6276.0
Unique	1893.2	1863.4	5201.2	5024.4
Inversions (per genome)				
Total	3084.0	2056.8	5865.3	5656.8
Unique	2866.6	1877.5	5470.8	5271.8

Table 1. Summary of coverage and rearrangements for all human datasets.

Microhomology usage represents the proportion of total rearrangements with at least 5 bases of homology at the breakpoint junction. The number of unique rearrangements is obtained by considering identical events as a single rearrangement. Total del/dup < 50 bp represents total deletions and duplications of sequences shorter than 50 bp.

3.4.2 Short-range inversions suggest replication U-turns occur in the human nuclear genome

Since the approach from which we developed SCARR proved useful in studying replication U-turns in organelle genomes (Zampini et al. 2015), we also examined the occurrence of U-turns in the human nuclear genome. U-turns involve the annealing of a nascent DNA strand to the opposing template strand, thereby creating a sequence inversion within a stalled replication fork (Mizuno et al. 2012). As a consequence of the U-turn, the new DNA strand contains a short-range inversion. We therefore calculated total inversions for each distance between 0 and 200 bases and normalized them to genome coverage (Figure 14). In all samples, we found that a large majority of short-range inversions occur at distances under 50 bases, with a maximum peak at distance 0. This supports the presence of replication U-turns as an alternate mechanism to restart stalled replication forks in the human nuclear genome.

If SCARR fails to explain a read as a combination of two shorter alignments, the script then attempts to explain it as a combination of three alignments. These SVs, which we refer to as paired rearrangements, can provide additional context to rearrangements within a single DNA molecule. Since uninterrupted replication following a U-turn results in acentric or dicentric chromosomes (Mizuno et al. 2009), they can result in large sequence alterations that would be deleterious to the cell. We therefore investigated paired rearrangements in human datasets with 151-bp reads to determine whether U-turns are followed by a second inversion event that restores replication in the original direction. Because this requires two rearrangements within one read length, the brain and spleen datasets yielded only 920 and 802 total paired rearrangements (Supplementary Files 3-4), compared to 343,899 and 273,099 single rearrangements, respectively.

15 of the paired rearrangements in the brain dataset and 14 in the spleen dataset are paired inversions, in which at least one occurred at less than 50 bp (Supplementary Files 3-4). Interestingly, 5 of the total paired inversions were identified 2 to 4 times on independent DNA fragments. Considering the sequencing coverage of approximately 6X for each dataset, these independently-identified paired inversions likely correspond to heterozygous alleles in the individuals from which the samples were obtained. In all cases except one, the longest distance

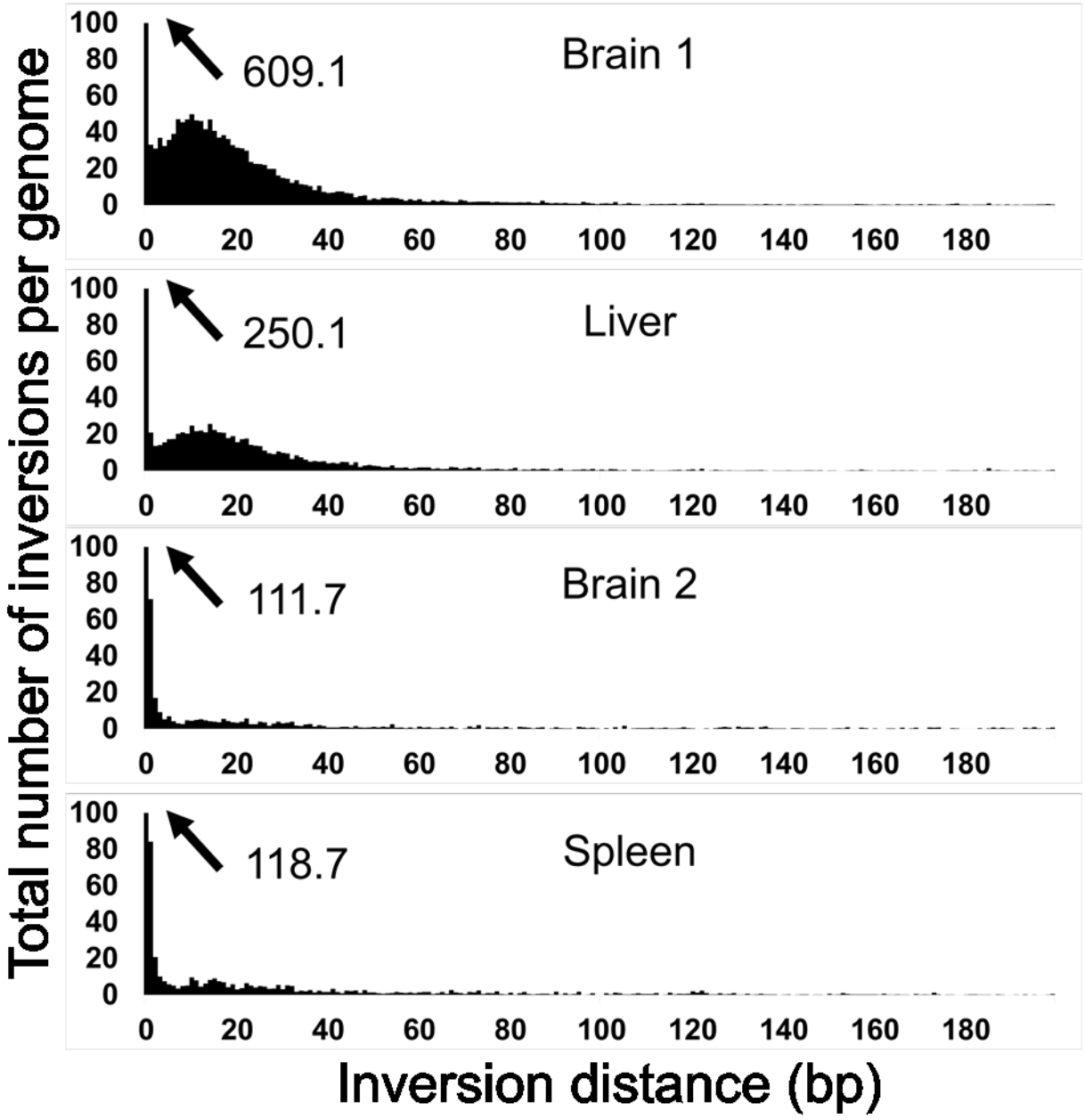


Figure 14. Patterns of short-range inversions in human datasets.

Histograms represent the total number of inversions identified for each distance value between 0 and 200 for each dataset, normalized to 1X genome coverage. Datasets are labeled according to the names in Table 1, with read lengths in parentheses for datasets derived from the same DNA samples. Values that exceed 100 are indicated with arrows and their numerical value.

of the two inversions is less than 400 bp, suggesting that either the nascent strand reanneals to its original template or that a second U-turn-like inversion occurs. The SVs produced can be complex, creating tandem inverted duplications which result in some DNA segments being triplicated (Figure 15).

3.4.3 Yeast mutant strains as a model to study inversions and replication U-turns

To further examine the proteins involved in inversions in eukaryotic nuclear genomes, we sequenced DNA from several strains of *Saccharomyces cerevisiae* with deletions in genes involved in DNA metabolism. To measure the impact of specific repair pathways on inversions, we used mutants for genes involved in end joining (*dnl4-Δ* and *yku70-Δ*), long-range end resection (*exo1-Δ* and *sgs1-Δ*), end processing (*mre11-Δ* and *sae2-Δ*), homology search and single-strand DNA binding (*rad51-Δ*, *rfa1-S373P* and *mgs1-Δ*), as well as DNA-damage sensing and replication (*srs2-Δ*, *mph1-Δ*, *rad9-Δ* and *pol32-Δ*). We grew each strain in four different conditions to test the effect of replication stress and DNA damage: YPD culture medium, camptothecin (CPT) and α-factor (αF), CPT and nocodazole (NOC), and hydroxyurea (HU). CPT forms a ternary complex with topoisomerase I and a nicked DNA duplex, whereas αF and NOC stop cells at either the G1/S or G2/M transitions, respectively. Lesions caused by CPT form DSBs when they are encountered by a replication fork. CPT treatment on cells arrested with NOC is therefore expected to have little impact on genome stability. HU depletes the nucleotide pool and induces replication stress and single-strand breaks.

We used SCARR to identify rearrangements in all datasets, and analyzed the patterns for inversions under 1 kb (Figure 16 and Supplementary File 5). As with the human datasets, we observed that inversions form a peak for distances under 50 bases, although CPT and HU stresses also showed a larger increase in long-range events. Since they occur within an open replication fork, U-turns can be identified as short-range inversions. The similarity between patterns for short-range inversions in our datasets suggests that budding yeast can serve as a model for studying replication U-turns in humans. To identify proteins involved in the U-turn pathway, we compared the number of short-range inversions (< 50 bases) and long-range inversions (≥ 50 bases) in the yeast strains under the four growth conditions. For each mutant,

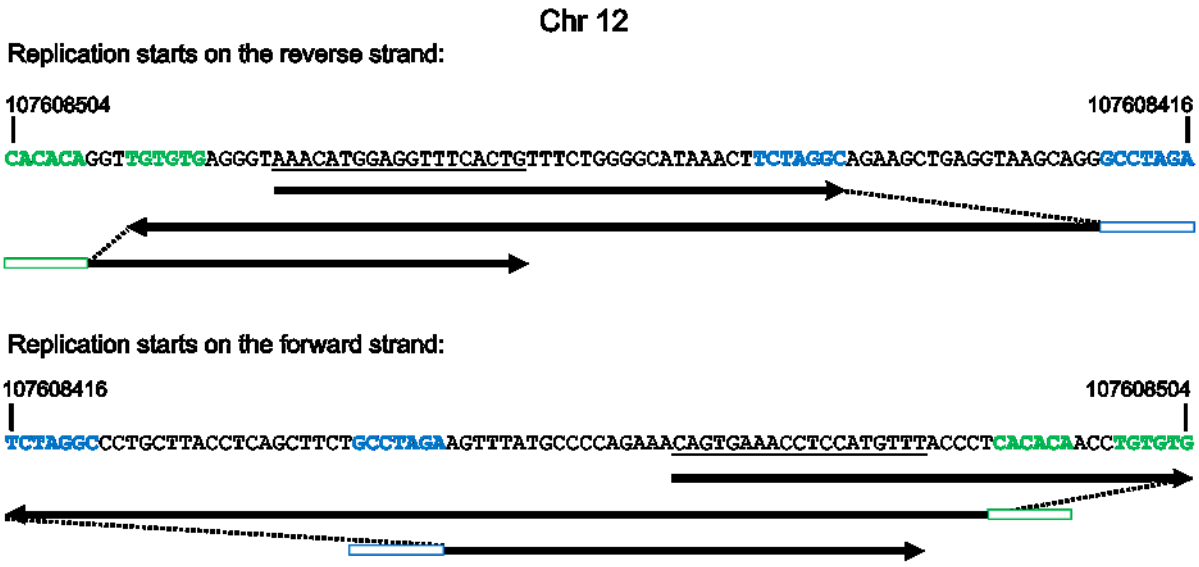


Figure 15. Example of paired inversions occurring at short range identified in the spleen dataset.

The arrows follow DNA replication, with dotted lines representing template switches and changes in orientation. The nascent strand adds the sequence presented when replication goes from left to right, and its reverse complement when replication goes from right to left. The product sequence is determined by the following the arrows, with jumps along the dotted lines. Inverted microhomologies that may contribute to the template switching are represented in blue and green, with rectangles of corresponding color where the annealing would lead to template switching. The direction of the jumps during template switches varies depending on which nascent strand gave rise to the rearrangement. The underlined sequence is triplicated as a result of the two inversions.

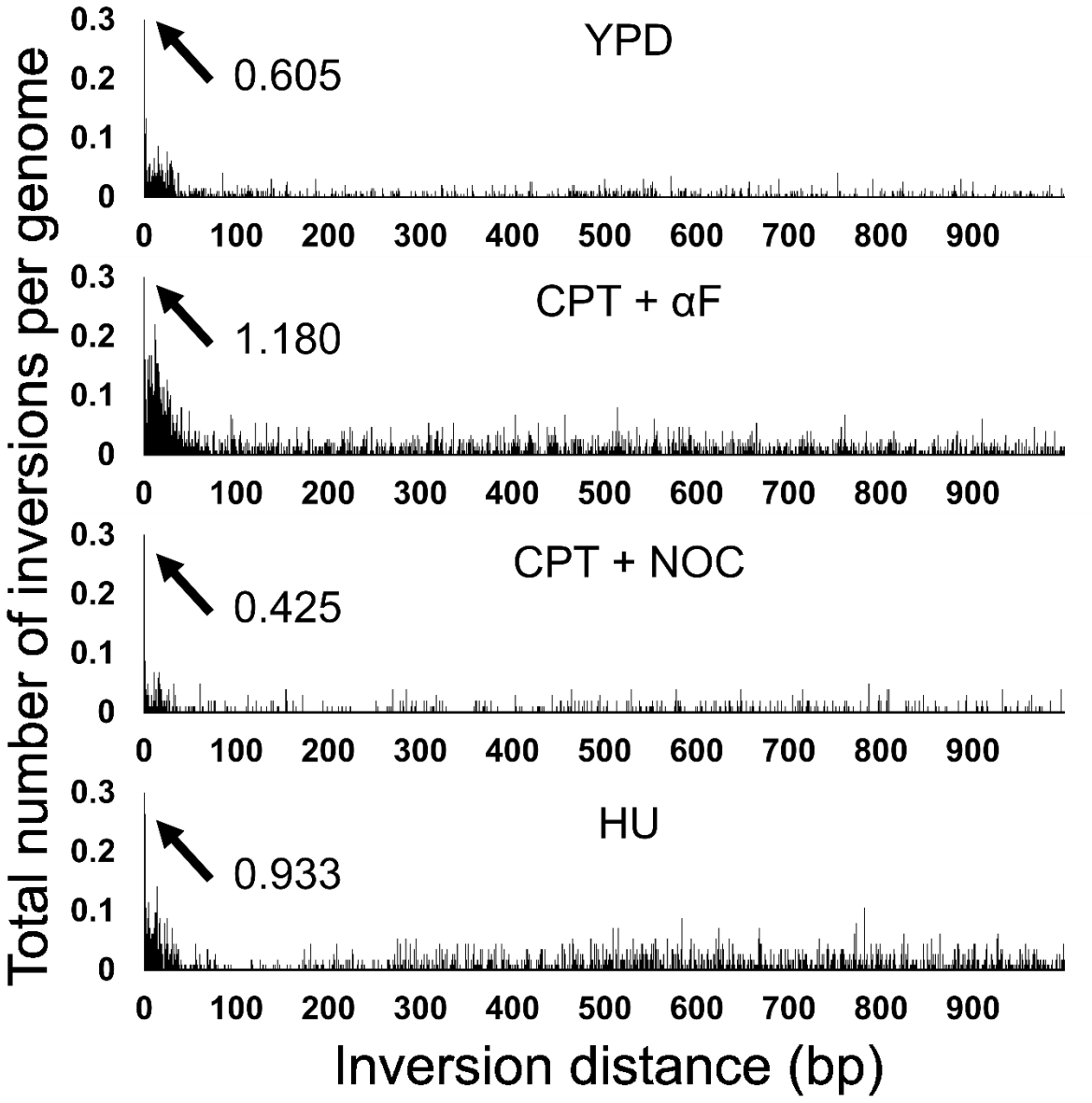


Figure 16. Patterns of short-range inversions in the yeast WT strain in the four growth conditions.

Histograms represent the total number of inversions identified for each distance value between 0 and 999 for each dataset, normalized to 1X genome coverage. YPD medium, hydroxyurea (HU), camptothecin and α-factor (CPT + αF) and camptothecin and nocodazole (CPT + NOC). Values that exceed 0.3 are indicated with arrows and their numerical value.

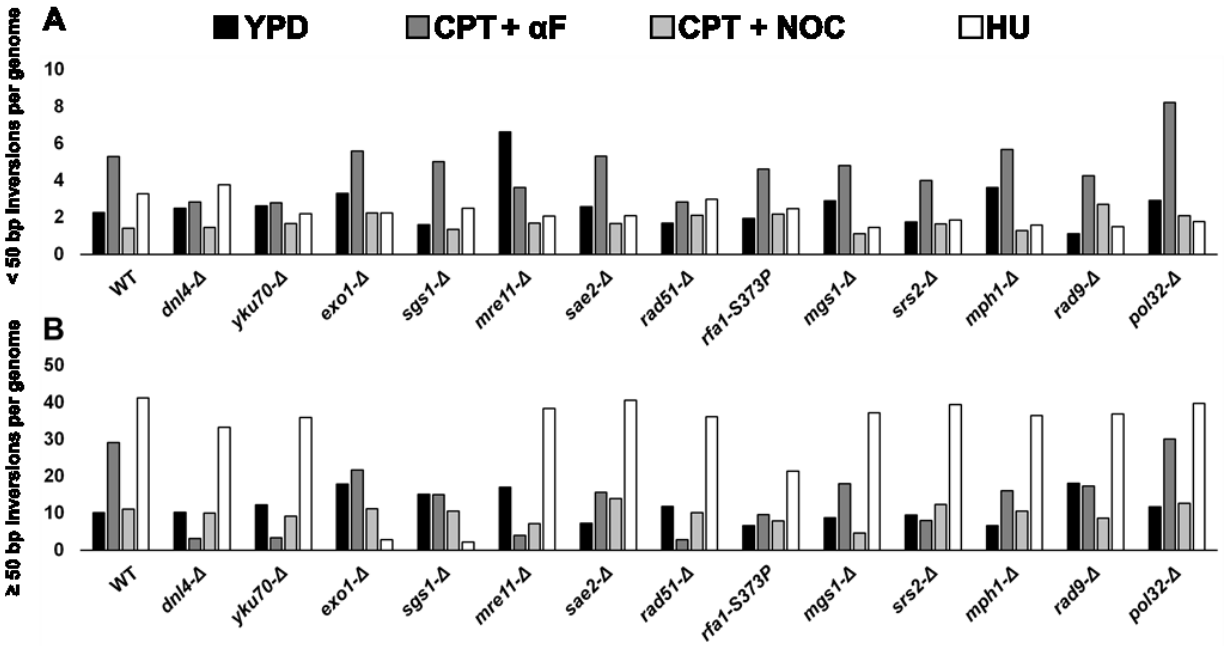


Figure 17. Comparison of short- and long-range inversions in the WT and deletion mutant strains under four different conditions.

(**A**) Total number of short-range inversions, normalized to 1X genome coverage. (**B**) Total number of long-range inversions, normalized to 1X genome coverage. YPD: YPD medium, αF: α-factor, CPT: camptothecin, HU: hydroxyurea.

we calculated the total number of these events normalized to genome coverage for each dataset (Figure 17). This result indicates that short- and long-range events are affected differently by the deletion mutants under the different growth conditions.

3.4.4 End-joining mechanisms contribute to inversions during G1 phase in yeast

Since Dnl4 (DNA ligase IV) and Yku70 (subunit of the DSB-binding Ku complex) are both closely linked to the NHEJ (Clerici et al. 2008), we observed similar results for the *dnl4-Δ* and *yku70-Δ* strains under all conditions (Figure 17). Because NHEJ is mostly active before S phase, we obtained the most striking results for these strains in the presence of CPT + αF. While these datasets showed a decrease in short-range inversions compared to the wild-type (WT) strain, they also yielded some of the lowest numbers of long-range inversions. These results suggest that most long-range inversions and some short-range inversions require end-joining machinery during G1 phase when DSB end-resection is suppressed. In contrast, all other conditions showed very little difference between the WT, *dnl4-Δ* and *yku70-Δ* strains, suggesting that HU stress mostly creates DSB ends during the S or G2 phases, when recombination pathways are active.

3.4.5 Extensive end-resection leads to long-range inversions under HU stress

Since Exo1 and Sgs1 are involved in extensive DSB end resection necessary for homologous recombination (HR) (Zhu et al. 2008), their absence may be expected to lead to an increase in rearrangements from error-prone mechanisms during G2 phase. Indeed, the *exo1-Δ* and *sgs1-Δ* strains showed increases in long-range inversions in the absence of stress. However, these strains presented almost no difference from the WT strain under CPT + NOC stress, with only a modest increase in short-range inversions in the absence of Exo1 (Figure 17). The increase in long-range inversions observed in the absence of stress therefore seems to occur outside of G2 phase. Interestingly, the *exo1-Δ* and *sgs1-Δ* strains also showed a severe decrease in long-range inversions (Figure 17 and Supplementary Figure S3) under HU stress. The distance distribution for the Exo1- or Sgs1-dependent inversions observed in the WT strain under HU stress falls well within the expected resection range of the two nucleases (Zhu et al.

2008). A modest reduction in long-range inversions compared to the WT strain was also observed in the *exo1-Δ* and *sgs1-Δ* strains under CPT + αF stress (Figure 17), but not for short-range inversions. This effect may be due to the same factors that cause a reduction in long-range inversions in these strains under HU stress.

3.4.6 Mre11 has opposite effects depending on growth conditions

The Mre11-Rad50-Xrs2 (MRX) complex and Sae2 are known to participate in the initiation of resection by Exo1 (Garcia et al. 2011; Cannavo and Cejka 2014). As such, we expected the *mre11-Δ* and *sae2-Δ* strains to show similar effects to *exo1-Δ*. However, the *sae2-Δ* strain yielded similar levels of inversions to the WT strain under the four conditions, except for a reduction in long-range inversions under CPT + αF stress, similar to the *exo1-Δ* and *sgs1-Δ* strains (Figure 17). In contrast, the *mre11-Δ* strain presented short-range inversion levels that are noticeably different from all other mutants (Figure 17). The deletion of Mre11 caused more short-range inversions than any other condition in the absence of stress, but fewer short-range inversions than the WT strain under HU and CPT + αF stresses. The *mre11-Δ* strain also presents a reduction in long-range inversions comparable to the *dnl4-Δ* and *yku70-Δ* strains in the presence of CPT + αF, but an increase in long-range inversions in the absence of stress. In YPD, the *mre11-Δ* strain showed opposite impacts on short-range inversions to those observed when the cell cycle was stopped in G1 phase or when the cells were under HU stress. This seems to indicate that Mre11 plays a role in preventing short-range inversions, specifically during the S phase.

3.4.7 Rad51 and Rfa1 contribute to long-range inversions in G1 phase

In a previous study, we found that single-stranded DNA-binding Whirly proteins and bacterial-type RecA recombinase played important roles in preventing replication U-turns in plastids of *Arabidopsis thaliana* (Zampini et al. 2015). Rfa1 (subunit of the replication protein A complex) and Rad51 both bind single-stranded DNA as part of the HR pathway (Shinohara et al. 1998), with Rfa1 also inhibiting microhomology-mediated rearrangements (Deng et al. 2014). The deletion of *RFA1* is lethal in yeast, but the *rfa1-S373P* mutation (also known as *rfa1-t33*) increases genome instability while maintaining cell viability (Deng et al. 2015). In our

datasets, the *rad51-Δ* and *rfa1-S373P* strains showed some similarities, including modest increases in short-range, but not long-range inversions compared to the WT under CPT + NOC stress (Figure 17). Under CPT + αF stress, the *rad51-Δ* strain presented the same pattern as the *dnl4-Δ*, *yku70-Δ* and *mre11-Δ* strains, suggesting roles in the same pathway for these proteins during the G1 phase. The *rfa1-S373P* strain had fewer long-range inversions in the presence of CPT + αF, and unlike for the *RAD51* deletion mutant, this decrease was also matched under HU stress. Rfa1 therefore seems to contribute to the formation of inversions outside of G2 phase by a mechanism that differs from Rad51.

Mgs1 possesses single-stand annealing activity, and the *mgs1-Δ* mutation was shown to cause an increase in recombination and genome instability during G2 phase (Hishida et al. 2001). Surprisingly, our *mgs1-Δ* strain datasets showed a decrease in long-range inversions compared to the WT under CPT + NOC and CPT + αF stresses, with little change to short-range inversions (Figure 17). The opposite was observed in the presence of HU, with long-range events maintaining a level similar to the WT, but short-range events being reduced by more than half. These effects are consistent with different roles for Mgs1 in the presence or absence of DNA-damage (Saugar et al. 2012).

3.4.8 RAD9 contributes to short-range inversions in the absence of DNA damage

The helicases Srs2 and Mph1 both favor the synthesis-dependent strand-annealing (SDSA) pathway of homologous recombination (Mitchel et al. 2013). Srs2 also inhibits recombination by dismantling Rad51 filaments, but loss of Mph1 does not reduce the efficiency of gap repair. In our results, neither the *srs2-Δ* nor the *mph1-Δ* strain showed any large variation from the WT under CPT + NOC stress (Figure 17), supporting a limited impact for the proteins on recombination. However, both deletion mutants yielded fewer long-range inversions than the WT in the presence of CPT + αF, and the *srs2-Δ* strain additionally had a smaller reduction in short-range inversions. Both mutants also presented fewer short-range inversions but no effect on long-range inversions under HU stress. Though Mph1 appears to promote inversions in the presence of DNA-damage, the deletion mutant showed the second largest increase in short-range inversions in the absence of stress.

The DNA-damage checkpoint protein Rad9 prevents DNA-degradation at stalled replication forks (Villa et al. 2018). In the absence of stress, the *rad9-Δ* strain presented the lowest level of short-range inversions, but the most long-range events (Figure 7). Compared to the WT, this strain showed fewer short-range inversions under HU stress, but more under CPT + NOC. The effect of Rad9 on short-range inversions therefore seems to be linked to its role during replication stress, rather than as a response to DNA damage.

Pol32 is an error-prone polymerase that was found to be responsible for replication during repair of broken forks (Mayle et al. 2015). In the four growth conditions, the *pol32-Δ* strain presented similar levels of long-range inversions to the WT strain. However, it showed more short-range inversions than the WT in both CPT treatments, and fewer in the presence of HU (Figure 17). Pol32 therefore appears to prevent rearrangements in the presence of DNA damage, while promoting instability during replication stress.

3.4.9 Replication U-turns require little to no sequence homology

Replication U-turns were shown in previous studies to occur at sites of inverted repeats through the annealing of short homologous sequences (Mizuno et al. 2009, 2012). The data obtained from SCARR also includes sequence homology usage patterns, which can help distinguish different rearrangement mechanisms and provide information about their homology requirements. To determine whether U-turns require inverted repeats near the site of fork stalling, we investigated the homologies involved in short-range inversions in our data. Since short-range inversions include both replication U-turns and inversions from other mechanisms, we looked in more detail at yeast deletion mutants and stress conditions that suppress long-range inversions. These conditions are also likely to suppress short-range inversions occurring through the same mechanisms, and should therefore yield a pattern of homology usage more specific to U-turns. We then compared these mutants to the WT without induced stress.

In the WT dataset, we observed two main peaks with similar profiles for both short- and long-range inversions: at 0 bases of homology, and between 4 and 15 bases of homology (Figure 18A). In both cases, the two peaks reach approximately the same maximum values. In contrast, the *dnl4-Δ*, *yku70-Δ*, *mre11-Δ* and *rad51-Δ* strains in the presence of the CPT + αF stress all yielded the same peaks, but with a higher relative proportion for short-range inversions with 0

bases of homology compared to homologies between 4 and 15 bases (Figure 18B and Supplementary Figure S4). The same pattern is also observed for the *exo1-Δ* and *sgs1-Δ* strains under HU stress (Figure 18C and Supplementary Figure S4). Interestingly, the human datasets with reads of the same length also presented the same pattern for short-range inversions, in spite of having more noticeably different patterns obtained for long-range inversions (Figure 18D). These results further support that replication U-turns occur in the human nuclear genome, and that sequence homology is not a *sine qua non* for this type of event.

3.5 Discussion

Replication U-turns were initially observed in fission yeast using a reporter system with perfect 2.6 kb inverted repeats, but were detected with repeats as short as 150 bp (Mizuno et al. 2009, 2012). From these results, Mizuno et al. (2012) suggested that shorter homologies may be sufficient for U-turns to occur at an appreciable rate. By analyzing U-turns genome-wide in budding yeast, we show that they may be favored when the nascent strand anneals to the new template using inverted repeats, but also occur frequently without any sequence homology at all. We also find that Mre11 and Exo1, which are involved in DNA degradation at stalled forks (Lemaçon et al. 2017), reduce the frequency of U-turns in the absence of stress, whereas Rad9 inhibits degradation and has the opposite effect (Villa et al. 2018). Mre11 also seems to specifically inhibit short-range inversions during the S phase, and promote their formation in G1 phase. These results are consistent with a model in which U-turns occur when replication forks remain stalled for long enough periods without being degraded for the template switch to occur.

Although it promotes inversions in the presence of CPT or HU, Mph1 also appears to prevent U-turns in the absence of DNA damage, possibly through its role in fork reversal, which enables accurate replication fork restart (Zheng et al. 2011). Though Sgs1, a RecQ helicase, has also been linked to accurate replication fork restart, it does not appear in our results to inhibit the appearance of U-turns. As the mechanisms through which repair proteins interact with stalled forks become better understood, it will be interesting to look in more detail at how they affect replication U-turns.

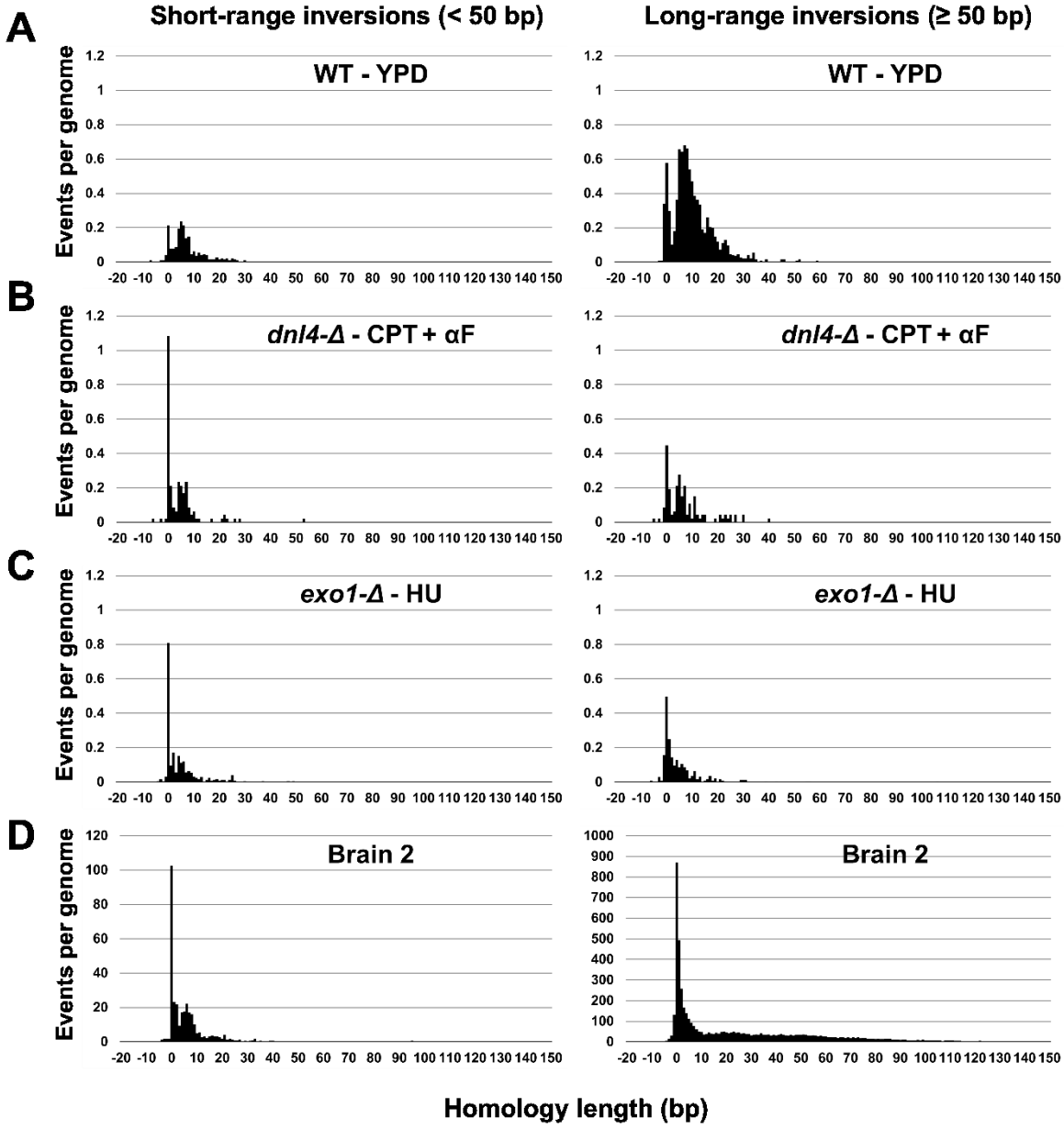


Figure 18. Homology usage for short-range (< 50 bp) and long-range inversions (≥ 50 bp).

Histograms represent total number of unique events for which a homology of given length is found at the breakpoint junction. Negative homology lengths represent base insertions at the breakpoint junction. All values are normalized to 1X genome coverage. (**A**) WT yeast strain grown in YPD. (**B**) *dnl4-Δ* strain grown in CPT (5 μg/mL) and αF. (**C**) *exo1-Δ* strain grown in HU (200 mM). (**D**) Human brain 2 dataset.

Exo1 prevents short-range inversions in the absence of stress, but its activity is responsible for the vast majority of long-range inversions under HU stress. HU has previously been shown to interfere with base-excision repair and produce single-strand breaks, which can be converted to one-ended DSBs following replication (Snyder 1984). Since both Exo1 and Sgs1 are involved in the 5' end resection of DSBs up to several kb, they likely favor snap-back replication, during which the single-stranded 3' extremity anneals back onto itself to allow self-templated DNA synthesis. Although snap-back replication and replication U-turns can both result in inversions, the distribution of long-range inversion distances in the WT strain shows a skewed distribution with a peak between 500 and 800 bases, with much lower values between 100 and 200 bases (Supplementary Figure S3). This is more consistent with a snap-back mechanism, for which the minimum inversion distance corresponds to roughly two times the length of the ssDNA that folds back onto itself.

We also report the detection of U-turns in the human nuclear genome with distance and homology usage patterns that closely resemble those observed in yeast. In some cases, U-turns are paired with a second short-range inversion to return the DNA strand to its original orientation. A similar mechanism was recently proposed to explain mutation clusters observed between chimpanzee and human reference genomes, as well as between the assembled genomes of different human individuals (Löytynoja and Goldman 2017). The rate at which we identified U-turns in healthy human tissues further supports that short-range inversions can play an important role in both the evolution of genomes and the appearance of genetic diversity within species. The limited need for sequence homology, along with the rapid drop in inversion events observed as distance increases, fits with the model that template-switching events are favored by the availability and proximity of single-stranded DNA within a replication fork (Löytynoja and Goldman 2017). Since these restarted forks are error-prone and can create new inverted repeats through inverted tandem duplications (Mizuno et al. 2012), this model also argues against the assumption that small inversions are often balanced rearrangements such as those simulated in SVsim.

The presence of U-turns in the human nuclear genome also suggests a new mechanism through which stalled replication forks may be restarted. Combined with our results in yeast, this raises the possibility that U-turns may be involved in chemoresistance in BRCA2-deficient

cells. A previous study has found that the hypersensitivity of these cells to DNA-damaging agents depends on the recruitment of MRE11 to stalled replication forks (Ray Chaudhuri et al. 2016), and our results suggest that the deletion in yeast of *MRE11* alone favors replication U-turns. Though replication U-turns result in alterations in the genome sequence, they also allow DNA synthesis to resume without creating any DSB intermediates. Since HR is impaired in BRCA2-deficient cells, an increase in U-turn rates could explain a reduced sensitivity to fork stalling. The rate at which we observe short-range inversions in healthy tissues and the proposed evolutionary model also suggest that the SVs resulting from U-turns can occur with little to no deleterious effects (Löytynoja and Goldman 2017).

When cells are arrested in G1, the *dnl4-Δ*, *yku70-Δ*, *mre11-Δ* and *rad51-Δ* strains all showed large reductions in long-range inversions, with more modest effects on short-range inversions. These results show that these rearrangements can occur through end-joining mechanisms, and confirm roles for Mre11 and Rad51 in NHEJ (Reginato et al. 2017; Konomura et al. 2017). Mre11 has previously been shown to be essential for processing of DSB extremities bound to topoisomerase complexes (Hoa et al. 2016). Given that CPT stabilizes topoisomerase at the site of the break, our results support this role for Mre11 in the NHEJ pathway. While Konomura et al. (2017) describe the role of Rad51 as enhancing ligation of DSB ends, our data suggest that the absence of Rad51 has an impact on NHEJ almost indistinguishable from Dnl4, Yku70, or Mre11. It therefore seems likely that Rad51, like Mre11, plays a role in processing obstructed DNA extremities for NHEJ.

Since datasets with longer reads improve the sensitivity of SCARR, next-generation sequencing technologies that produce longer read lengths will be an important avenue to explore. The current version of SCARR checks each read for either one or two rearrangements, but the approach can be extended to any number of rearrangements per read, provided that the sequences are long enough to successfully align enough fragments. As such, long reads will also be useful to provide context for rearrangements within single molecules. This possibility will be particularly interesting to the study of template-switching events generating inversions, since they require a second template-switching event to occur in order to resume DNA synthesis in the original direction.

3.6 Materials and Methods

Simulated datasets. SVs were randomly created using SVsim (https://github.com/GregoryFaust/SVsim) to generate 2,500 deletions, duplications, and inversions in the human reference genome (build 38) from which the mitochondrial genome has been removed. The variants each range from 100 bp to 10,000 bp, as described previously (Layer et al. 2014). Virtual sequencing was performed using WGSIM (https://github.com/lh3/wgsim) to a coverage of 1X, using paired-end reads of 150 bp and default settings for insert size and error rates, also as per Layer et al. 2014. Rearrangements with lengths that differ by more than 20 bases from the simulated SV lengths were classified as false positives. The same sequencing was also performed on the unmodified human reference genome GRCh38 (minus the mitochondrial genome), with default error and indel settings.

Public datasets. Datasets for healthy human brain (ERR419275, ERR419276, ERR419277, ERR419278) and liver (ERR419279, ERR419280, ERR419281, ERR419282) DNA were obtained from the NCBI SRA. Additional library and sequencing information relating to the datasets was described in a previous publication (Wood et al. 2015).

Human DNA samples. Human DNA was purchased commercially from BioChain, pre-extracted from the following healthy tissues: brain (Occipital lobe, catalog no. D1234062, lot B806171, male, 41 years old) and spleen (Catalog no. D1234246, lot A712149, male, 27 years old).

Yeast strains. Haploid deletion mutants for *DNL4*, *EXO1*, *MGS1*, *MPH1*, *MRE11*, *POL32*, *RAD9*, *RAD51*, *SAE2*, *SGS1*, *SRS2* and *YKU70* were obtained from the yeast Magic Marker strains from Open Biosystems, as per their protocol (Pan et al. 2004). Successful selection of the deletion mutant was confirmed by PCR using the suggested primers. To obtain *rfa1-S373P* in the same background strain, the point mutation was generated in the *RFA1* plasmid from Molecular Barcoded Yeast ORF library from Thermo Fisher Open Biosystems, and transformed into the heterozygous *rfa1-Δ* strain from the Magic Marker collection. The haploid deletion mutant containing the plasmid was then obtained by following the protocol for Magic Marker strains but subtracting uracil from the medium.

Growth conditions and DNA extraction. Cells were grown in YPD to an OD600 of 0.5 before treatment. For the CPT + αF treatment, the cells were incubated for 1 h in 5 μg/mL αF (Bachem) to stop the cell cycle. An additional 5 μg/mL αF was then added at the same time as 5 μg/mL CPT (Sigma-Aldrich). The cells were pelleted for DNA extraction 2.5 h after adding the CPT. For the CPT + NOC treatment, the cells were incubated for 2 h in 15 μg/mL NOC (Sigma-Aldrich) before adding 5 μg/mL CPT. The cells were pelleted for DNA extraction 2.5 h after adding the CPT. For the HU treatment, 200 mM HU (BioShop Canada) was added, and the cells were pelleted for DNA extraction after 3 h. For the control, 1μl/mL dimethyl sulfoxide (DMSO) was added, and the cells were pelleted for DNA extraction after 3h. DNA was extracted by phenol/chloroform and treated with RNAse A.

Illumina sequencing. Libraries were prepared from the DNA samples using the NxSeq AmpFREE Low DNA Library kit (Lucigen, Cat no. 14000-2) as per the manufacturer's protocol. This includes SPRI bead cleanup and size selection steps for a final median insert size of approximately 320 bp. Sequencing was performed on the Illumina HiSeq X Ten (302 cycles, paired-end). Library preparation and sequencing were done at Génome Québec. All new datasets were made available on NCBI SRA (Project number: SRP134058).

Rearrangement detection. Datasets were enriched for rearrangement-containing reads using a Galaxy workflow adapted from a previously published approach (Zampini et al. 2015). Adapter sequences were removed using Trim Galore! (https://github.com/FelixKrueger/TrimGalore) prior to alignment. The sequences obtained were further filtered to remove potential artifacts generated during library preparation from single-stranded DNA (Star et al. 2014). Reads were removed if the first 12 bases of paired reads were identical with 1 mismatch or less and a maximum offset of 3 bases. The remaining reads were labeled as potential junctions and aligned using BLAST+ (Camacho et al. 2009) using the following command: "blastn -query potential_junctions.fasta -db reference_genome.fasta -out output.txt -word_size 10 -evalue 0.0001 -outfmt 6". To keep file sizes manageable and accelerate the analysis, potential junctions were split into files containing 25 sequences or less and the work was parallelized. The output from BLAST+ was given as input to SCARR with default parameters. SCARR is a custom Python script available on GitHub that tests all possible combinations of BLAST+ alignments to determine whether a satisfactory match can be found

(Supplementary Figure S5). Additional custom scripts were used to sort and organize the output rearrangement files from SCARR, and are also available on GitHub.

SCARR algorithm. SCARR loads all blastn alignments from a read into an array, and determines the minimum and maximum mapped positions of the read. The script then looks for any single alignment that spans most of the read, tolerating a length difference of at most 5 bases. When any such alignment is found, the read is considered not to contain a rearrangement, and the script continues to the next. When no single alignment spans the read, SCARR tests all possible combinations of 2 alignments. For each combination, it determines what percentage of the read between the minimum and maximum mapped positions are covered by the alignments, and subtracts any gaps or mismatches in the individual alignments (weighted by default at 1.25 per mismatch and 1.5 per gap). This determines a score out of 100 for the alignment pair. The best score is kept and compared to a set threshold (by default 80), and the best alignment pair is considered as a rearrangement when it exceeds the threshold. For reads where no alignment pair passes the threshold, the process is repeated by testing all possible combinations of 3 alignments. When the best-scoring alignment triplet exceeds the threshold, it is considered as a paired rearrangement and written to a file separate from single rearrangements. When no alignment triplet passes the threshold, the read is considered as unlabeled, and the script moves on to the next.

Rearrangement analysis. For each rearrangement, microhomology length is determined by the number of bases from the original read that are aligned to both sides of the breakpoint. If some bases at the breakpoint fail to align to either side, they are counted as inserted bases. If the alignments map to different chromosomes, the rearrangement is labeled as a translocation, and no further analysis is performed. If both alignments map to the same chromosome, their relative directions are verified. If they are in opposing directions, the rearrangement is labeled as an inversion, and the distance represents the number of bases between the reference genome positions of the last base in the first alignment and the first base in the second alignment, minus any microhomology length. If both alignments are in the same direction, their relative positions are verified. If the breakpoint represents a jump forward, the rearrangement is labeled as a deletion, while a jump backwards is labeled as a duplication. If no microhomology is present, the distance is measured from the last base in the first alignment to

the first base in the second alignment. If a microhomology is present, its length is subtracted from the distance, so that distance represents the exact number of bases deleted or duplicated in the reference genome.

Availability

SCARR and associated scripts are available in the GitHub repository (https://github.com/SamTremblay/SCARR).

Accession Numbers

Sequencing datasets have been deposited at the NCBI Sequence Read Archive (SRA) under accession number PRJNA437181.

3.7 Acknowledgements

We thank G. Arseneault for her assistance with the yeast strains. The raw sequencing data was treated using Galaxy on the public server (usegalaxy.org) and on the Genetics and Genomics Analysis Platform (GenAP). Computations were made on the supercomputer Briarée, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

Funding

This work was supported by the National Sciences and Engineering Research Council of Canada (NSERC) [RGPIN 2014-05373 to N.B.] and scholarships from the NSERC and FRQ-NT to S.T.B.

Conflict of interest

The authors declare that they have no conflict of interest.

3.8 References

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Cannavo E, Cejka P. 2014. Sae2 promotes dsDNA endonuclease activity within Mre11-Rad50-Xrs2 to resect DNA breaks. *Nature* **514**: 122–5.

Cardoso AR, Oliveira M, Amorim A, Azevedo L. 2016. Major influence of repetitive elements on disease-associated copy number variants (CNVs). *Hum Genomics* **10**: 30.

Carvalho CMB, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**: 224–238.

Chang HHY, Pannunzio NR, Adachi N, Lieber MR. 2017. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol* **18**.

Chen R, Lau YL, Zhang Y, Yang W. 2016. SRinversion: a tool for detecting short inversions by splitting and re-aligning poorly mapped and unmapped sequencing reads. *Bioinformatics* **32**: btw516.

Clerici M, Mantiero D, Guerini I, Lucchini G, Longhese MP. 2008. The Yku70-Yku80 complex contributes to regulate double-strand break processing and checkpoint activation during the cell cycle. *EMBO Rep* **9**: 810–8.

Deng SK, Gibb B, de Almeida MJ, Greene EC, Symington LS. 2014. RPA antagonizes microhomology-mediated repair of DNA double-strand breaks. *Nat Struct Mol Biol* **21**: 405–12.

Deng SK, Yin Y, Petes TD, Symington LS. 2015. Mre11-Sae2 and RPA Collaborate to Prevent Palindromic Gene Amplification. *Mol Cell* **60**: 500–8.

Garcia V, Phelps SEL, Gray S, Neale MJ. 2011. Bidirectional resection of DNA double-strand breaks by Mre11 and Exo1. *Nature* **479**: 241–244.

Hastings PJ, Ira G, Lupski JR. 2009a. A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation ed. I. Matic. *PLoS Genet* **5**: e1000327.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009b. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**: 551–564.

He F, Li Y, Tang Y-H, Ma J, Zhu H. 2016. Identifying micro-inversions using high-throughput sequencing reads. *BMC Genomics* **17**: 4.

Hishida T, Iwasaki H, Ohno T, Morishita T, Shinagawa H. 2001. A yeast gene, MGS1, encoding a DNA-dependent AAA+ ATPase is required to maintain genome stability. *Proc Natl Acad Sci* **98**: 8283–8289.

Hoa NN, Shimizu T, Zhou ZW, Wang Z-Q, Deshpande RA, Paull TT, Akter S, Tsuda M, Furuta R, Tsutsui K, et al. 2016. Mre11 Is Essential for the Removal of Lethal Topoisomerase 2 Covalent Cleavage Complexes. *Mol Cell* **64**: 580–592.

Konomura N, Arai N, Shinohara T, Kobayashi J, Iwasaki W, Ikawa S, Kusano K, Shibata T. 2017. Rad51 and RecA juxtapose dsDNA ends ready for DNA ligase-catalyzed end-joining under recombinase-suppressive conditions. *Nucleic Acids Res* **45**: 337–352.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

Lemaçon D, Jackson J, Quinet A, Brickner JR, Li S, Yazinski S, You Z, Ira G, Zou L, Mosammaparast N, et al. 2017. MRE11 and EXO1 nucleases degrade reversed forks and elicit MUS81-dependent fork rescue in BRCA2-deficient cells. *Nat Commun* **8**: 860.

Löytynoja A, Goldman N. 2017. Short template switch events explain mutation clusters in the human genome. *Genome Res* **27**: 1039–1049.

Mayle R, Campbell IM, Beck CR, Yu Y, Wilson M, Shaw CA, Bjergbaek L, Lupski JR, Ira G. 2015. Mus81 and converging forks limit the mutagenicity of replication fork breakage. *Science* **349**: 742–7.

Mitchel K, Lehner K, Jinks-Robertson S. 2013. Heteroduplex DNA position defines the roles of the Sgs1, Srs2, and Mph1 helicases in promoting distinct recombination outcomes. ed. L.S. Symington. *PLoS Genet* **9**: e1003340.

Mizuno K, Lambert S, Baldacci G, Murray JM, Carr AM. 2009. Nearby inverted repeats fuse to generate acentric and dicentric palindromic chromosomes by a replication template exchange mechanism. *Genes Dev* **23**: 2876–86.

Mizuno K, Miyabe I, Schalbetter SA, Carr AM, Murray JM. 2012. Recombination-restarted replication makes inverted chromosome fusions at inverted repeats. *Nature* **493**: 246–249.

Pan X, Yuan DS, Xiang D, Wang X, Sookhai-Mahadeo S, Bader JS, Hieter P, Spencer F, Boeke JD. 2004. A robust toolkit for functional profiling of the yeast genome. *Mol Cell* **16**: 487–96.

Ray Chaudhuri A, Callen E, Ding X, Gogola E, Duarte AA, Lee J-E, Wong N, Lafarga V, Calvo JA, Panzarino NJ, et al. 2016. Replication fork stability confers chemoresistance in BRCA-deficient cells. *Nature* **535**: 382–387.

Reginato G, Cannavo E, Cejka P. 2017. Physiological protein blocks direct the Mre11–Rad50–Xrs2 and Sae2 nuclease complex to initiate DNA end resection. *Genes Dev* **31**: 2325–2330.

Saugar I, Parker JL, Zhao S, Ulrich HD. 2012. The genome maintenance factor Mgs1 is targeted to sites of replication stress by ubiquitylated PCNA. *Nucleic Acids Res* **40**: 245–257.

Shinohara A, Shinohara M, Ohta T, Matsuda S, Ogawa T. 1998. Rad52 forms ring structures and co-operates with RPA in single-strand DNA annealing. *Genes Cells* **3**: 145–56.

Snyder RD. 1984. The role of deoxynucleoside triphosphate pools in the inhibition of DNA-excision repair and replication in human cells by hydroxyurea. *Mutat Res* **131**: 163–72.

Star B, Nederbragt AJ, Hansen MHS, Skage M, Gilfillan GD, Bradbury IR, Pampoulie C, Stenseth NC, Jakobsen KS, Jentoft S. 2014. Palindromic sequence artifacts generated during next generation sequencing library preparation from historic and ancient DNA. ed. L. Orlando. *PLoS One* **9**: e89676.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Trappe K, Emde A-K, Ehrlich H-C, Reinert K. 2014. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics* **30**: 3484–90.

Tremblay-Belzile S, Lepage É, Zampini É, Brisson N. 2015. Short-range inversions: Rethinking organelle genome stability. *Bioessays* **37**: 1086–1094.

Villa M, Bonetti D, Carraro M, Longhese MP. 2018. Rad9/53BP1 protects stalled replication forks from degradation in Mec1/ATR-defective cells. *EMBO Rep* **19**: 351–367.

Wang H, Xu X. 2017. Microhomology-mediated end joining: new players join the team. *Cell Biosci* **7**: 6.

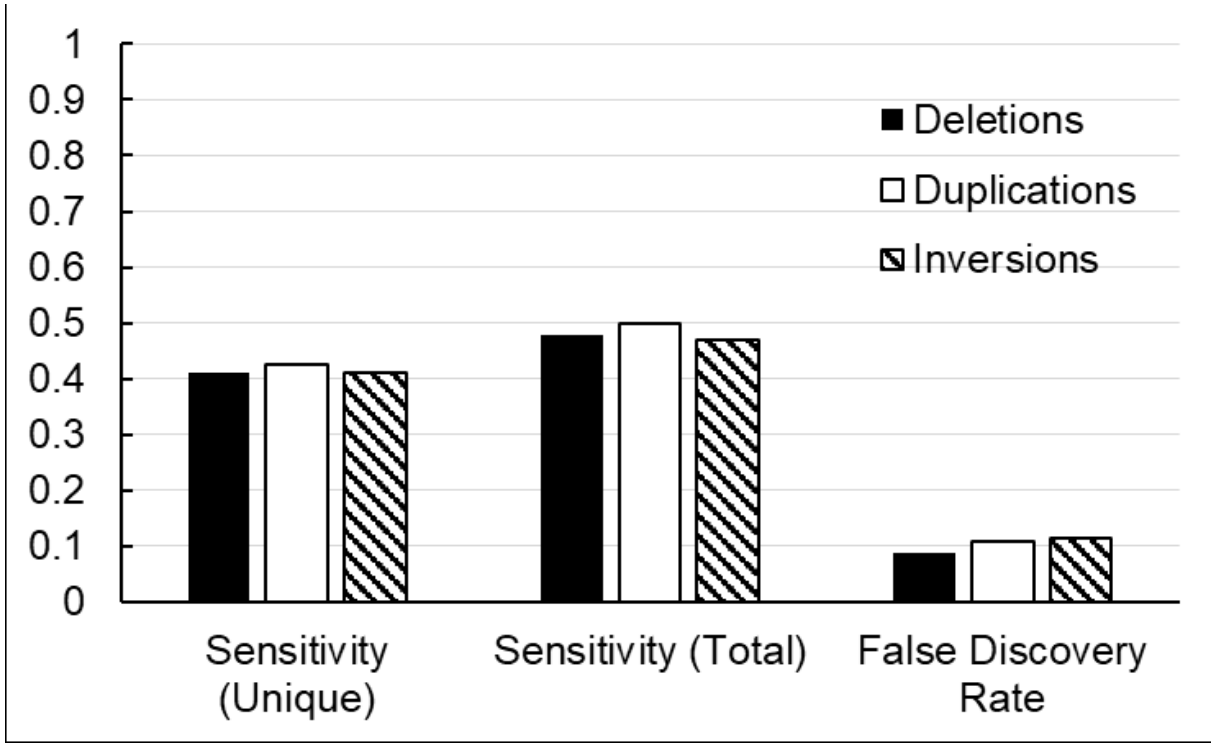
Wood DLA, Nones K, Steptoe A, Christ A, Harliwong I, Newell F, Bruxner TJC, Miller D, Cloonan N, Grimmond SM. 2015. Recommendations for Accurate Resolution of Gene and Isoform Allele-Specific Expression in RNA-Seq Data. ed. I.K. Jordan. *PLoS One* **10**: e0126911.

Zampini É, Lepage É, Tremblay-Belzile S, Truche S, Brisson N. 2015. Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in Arabidopsis and humans. *Genome Res* **25**: 645–654.

Zheng X-F, Prakash R, Saro D, Longerich S, Niu H, Sung P. 2011. Processing of DNA structures via DNA unwinding and branch migration by the S. cerevisiae Mph1 protein. *DNA Repair (Amst)* **10**: 1034–43.

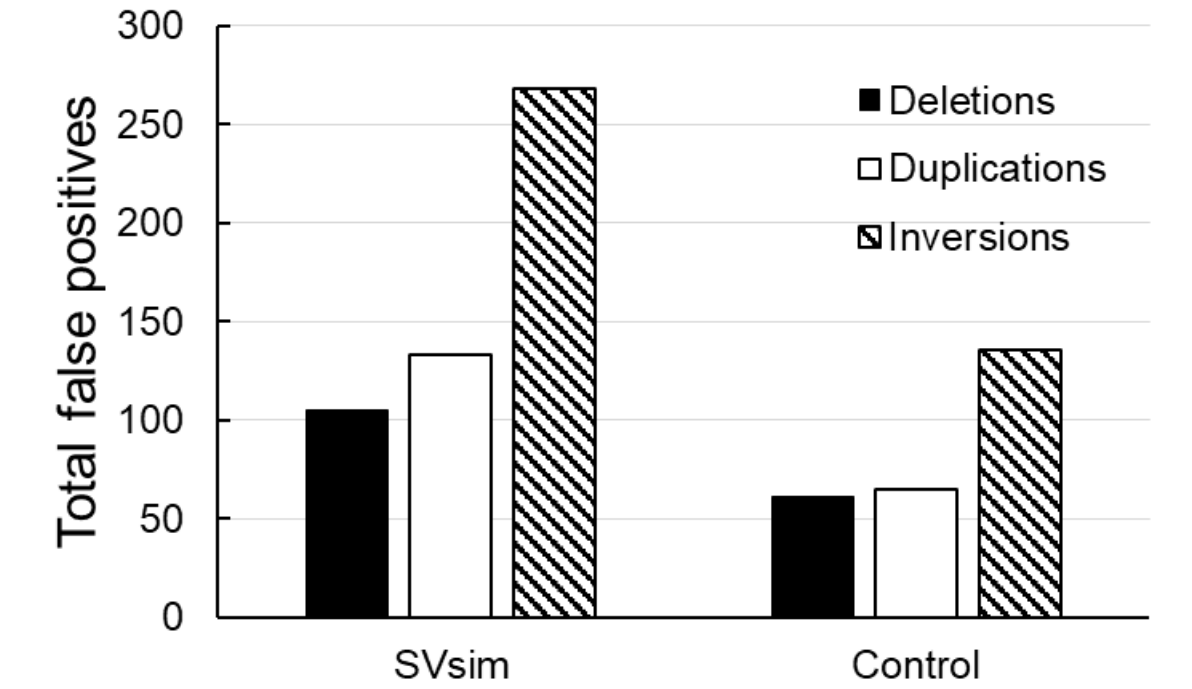
Zhu Z, Chung W-H, Shim EY, Lee SE, Ira G. 2008. Sgs1 helicase and two nucleases Dna2 and Exo1 resect DNA double-strand break ends. *Cell* **134**: 981–94.

3.9 Supplemental Material



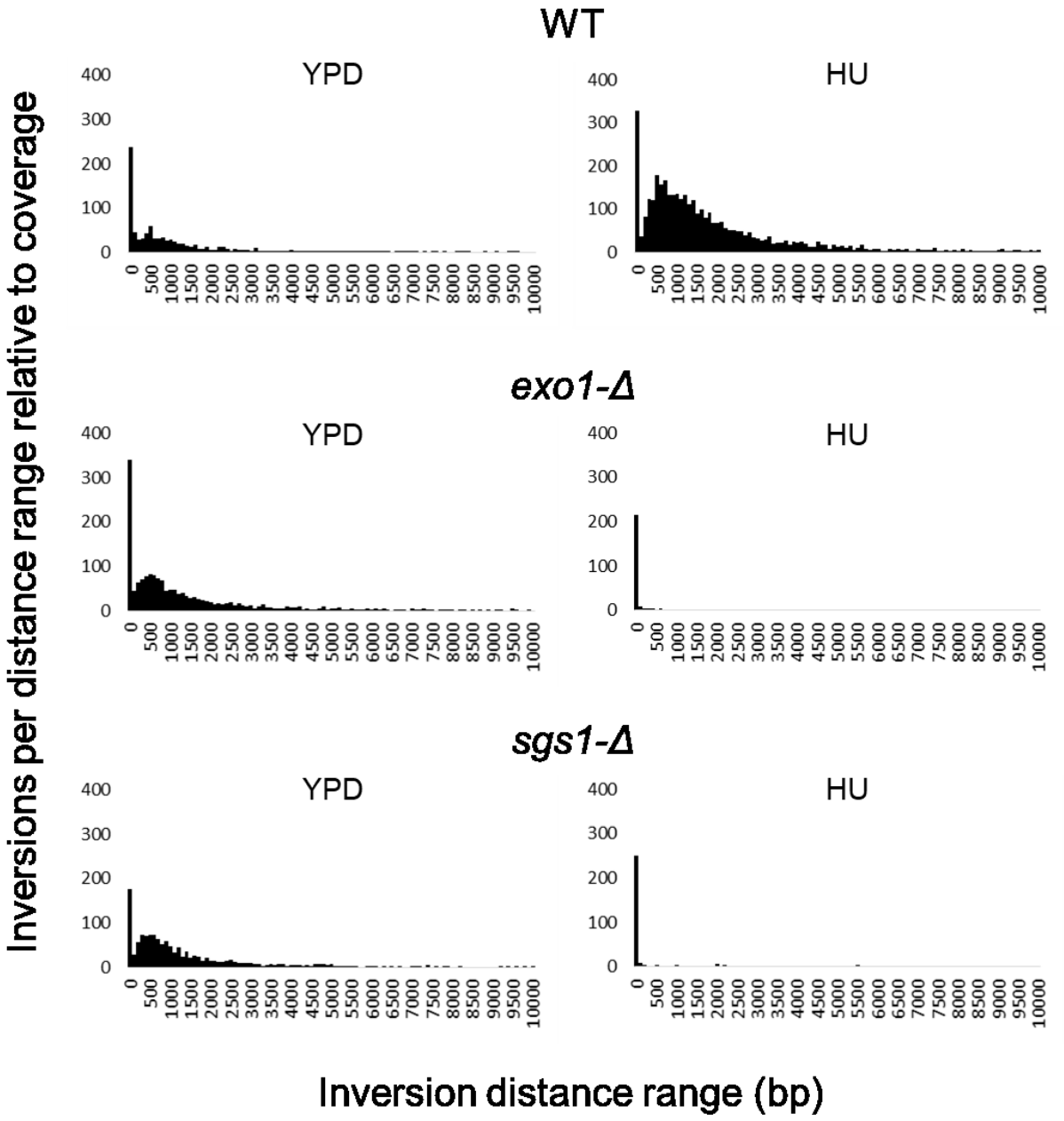
Man. Supplementary Figure S1. Sensitivity and false detection rate of SCARR based on 1X coverage of simulated data.

Sensitivity for unique rearrangements represents the ratio of unique events to the total number of different expected SVs. Each simulated deletion and duplication correspond to one expected rearrangement, and each inversion corresponds to two expected rearrangements. Total sensitivity represents the ratio of total events detected to the total number of events expected at this coverage. False discovery rate corresponds to the ratio of total false positives to the total number of events detected.

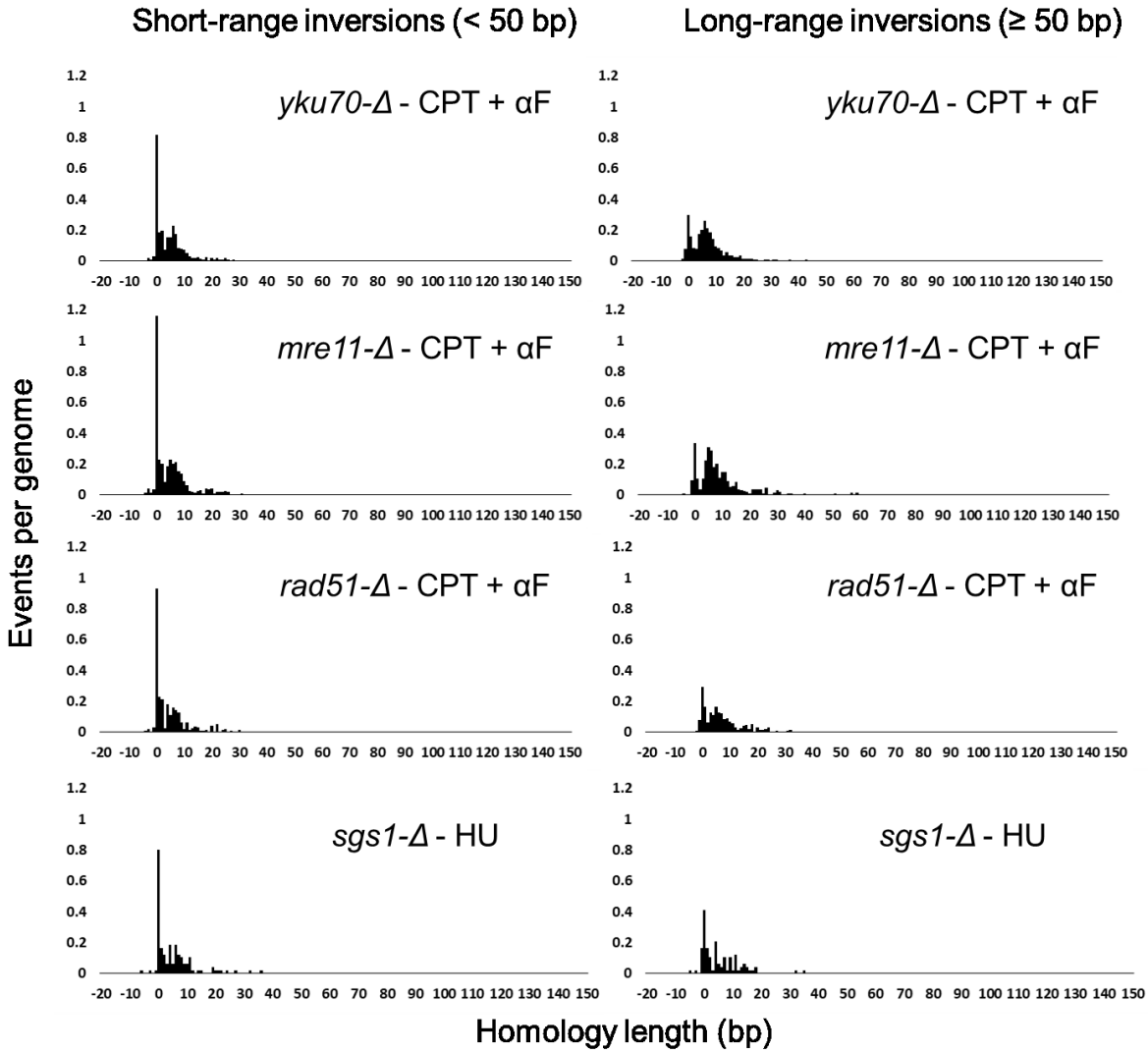


Man. Supplementary Figure S2. False discovery rate in simulated datasets with and without SVs.

Values represent the total number of each type of false positive found in 1X simulated sequencing coverage. The SVsim dataset contains 2,500 each of deletions, duplications, and inversions. The control dataset uses the unmodified reference genome.

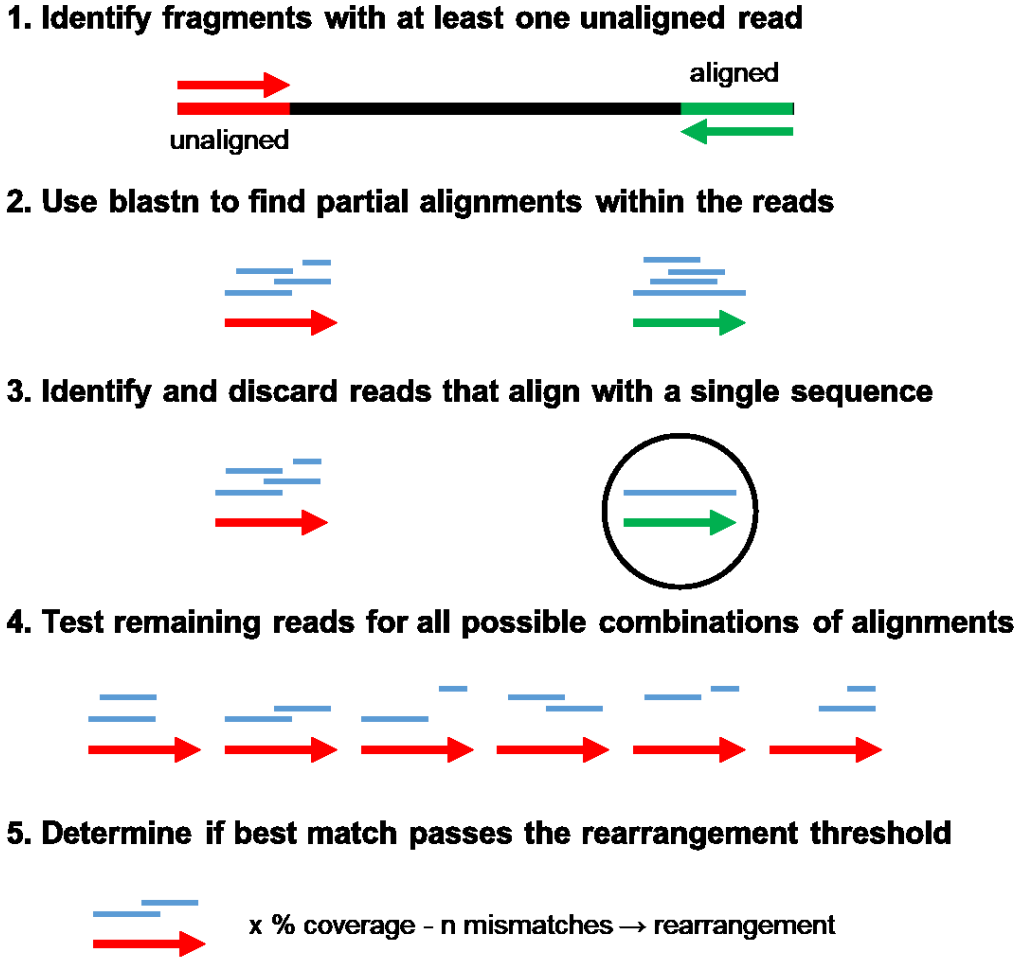


Man. Supplementary Figure S3. Short- and long-range inversions in the *exo1-Δ* and *sgs1-Δ* strains in YPD and hydroxyurea (HU) compared to the wild-type (WT). Histograms represent number of inversions per distance ranges of 100 bp. Labels represent lowest value of the range.



Man. Supplementary Figure S4. Homology usage for short-range (< 50 bp) and long-range inversions (≥ 50 bp).

Histograms represent total number of unique events for which a homology of given length is found at the breakpoint junction. Negative homology lengths represent base insertions at the breakpoint junction. All values are normalized to 1X genome coverage.



Man. Supplementary Figure S5. Summary of the approach used to identify rearrangements.

(1) Next-generation sequencing data is fed through a Galaxy workflow to obtain DNA fragments for which at least one extremity is not perfectly mapped to the reference genome, as previously described in Zampini et al. 2015. (2) blastn is used to obtain partial alignments for the remaining reads, and (3) the reads that closely match with a single alignment are discarded. (4) For the remaining reads, a rearrangement score is assigned to each possible combination of 2 partial alignments, taking into account the proportion of the read covered and the number of gaps and mismatches in the alignments. (5) The best scoring combination is called as a rearrangement only if it passes the selected score threshold.

4. Discussion

4.1 Une nouvelle approche pour répondre à un besoin précis

Le scénario idéal pour étudier l'instabilité génomique d'un organisme est de réussir à obtenir la séquence exacte de toutes ses molécules d'ADN. À partir de cette information, chaque variation dans la séquence pourrait être répertoriée et étudiée à la fois de manière individuelle et dans le contexte de la molécule d'ADN sur laquelle elle se retrouve. Bien que ce genre de résultat soit encore inatteignable, le développement du séquençage de nouvelle génération a été une étape majeure pour rendre la possibilité envisageable. Cette technologie a également eu comme effet de changer l'étape limitante dans l'utilisation du séquençage en recherche : la production de données de séquençage excède désormais notre capacité à analyser l'information contenue dans les séquences. Les méthodes développées pour traiter les données doivent donc être spécialisées pour extraire et synthétiser seulement une partie de l'information disponible. La configuration des méthodes dépend ainsi des questions biologiques auxquelles on veut répondre et de l'information qui est nécessaire pour y arriver.

Bien que de nombreuses méthodes aient été développées pour détecter presque tous les types de variation de séquence (délétions, duplications en tandem, insertions, inversions, translocations, etc.), la presque totalité d'entre elles ont pour objectif d'identifier des réarrangements spécifiques afin de les répertorier et de les associer à des phénotypes (voir section 1.1.2). Les réarrangements étudiés sont ainsi ceux qui existent dans la lignée germinale ou qui se produisent fréquemment dans les lignées somatiques. Dans le contexte d'un organisme diploïde, ceci implique que les données de séquençage contiennent l'information pour deux copies différentes ou identiques de chaque position du génome. Plusieurs études intègrent donc l'information obtenue par plusieurs approches afin d'identifier un réarrangement qui représente la moitié ou la totalité des séquences retrouvées à un locus (par exemple : (29, 103)). De cette perspective, la sensibilité de détection est beaucoup moins importante que la spécificité : les faux positifs peuvent mener à de mauvaises associations entre un réarrangement et un phénotype, tandis que la quantité de données fait en sorte qu'une variation de séquence commune peut facilement être trouvée à plusieurs reprises.

Les outils développés pour répertorier les variations de séquences communes sont ainsi mal adaptés à l'étude des mécanismes de réparation de l'ADN. Lorsque les mécanismes conservateurs ne sont pas disponibles, une proportion importante des réarrangements produits sont des événements rares qui diffèrent d'une copie du génome à l'autre. La sensibilité de détection est donc primordiale à l'obtention d'un portrait fidèle de l'instabilité génomique présente dans l'organisme. La première étape de mon projet a donc été de mettre au point une approche permettant l'étude des mécanismes menant à l'instabilité génomique à partir de séquençage de nouvelle génération. Au lieu de chercher les réarrangements, notre approche vise à expliquer toutes les données de séquençage obtenues en fonction d'un génome de référence. Un premier tri, effectué dans la suite de logiciels Galaxy, élimine les séquences correspondant presque parfaitement à une position du génome, laissant ainsi celles qui impliquent un réarrangement. L'analyse individuelle de chaque séquence favorise la grande sensibilité de l'approche, telle que présentée dans les chapitres 2 et 3. Puisque la détection dépend de la qualité des alignements, les facteurs limitants sont la proximité du réarrangement aux extrémités de la séquence, la quantité de variations de nucléotides uniques autour du réarrangement, et l'insertion de courtes séquences au site du réarrangement.

Pour éviter d'introduire un biais dans le type de réarrangements identifiés, nous avons opté pour l'alignement partiel sur le génome entier des séquences non-alignées. Chaque événement est ainsi identifié par les alignements de ses séquences partielles constituantes avant d'être classifié selon son type. Dans les petits génomes qui contiennent peu de séquences répétées, ceci occasionne très peu de faux positifs, puisque le nombre d'alignements partiels par séquence est généralement très limité. Par contre, le nombre de séquences répétées dans le génome humain fait en sorte qu'environ 10% des positions de réarrangements identifiés sont erronées (voir chapitre 2.4.1). Il en résulte qu'un réarrangement détecté une seule fois dans un jeu de données peut difficilement être considéré avec une très grande confiance. Le bruit de fond que ces réarrangements génèrent ne nuit cependant que très peu à l'étude des mécanismes de réparation de l'ADN. En effet, puisque les faux positifs sont répartis aléatoirement entre les différents types de réarrangement, leur impact dans l'analyse de mécanismes spécifiques demeure faible. Notre approche est donc parfaitement adaptée pour étudier l'instabilité

génomique dans un organisme montrant un tel phénotype, même à défaut d'hypothèse sur la source ou le type de réarrangements.

4.2 Des défis spécifiques à chaque génome

Dans les génomes de taille limitée, tels que ceux du chloroplaste et de la mitochondrie d'*Arabidopsis thaliana* et de la mitochondrie humaine, l'analyse des alignements partiels obtenus par BLAST+ peut être effectuée rapidement et efficacement en utilisant seulement un logiciel tabulateur comme Excel (sections 2.4 et 2.6). Ces génomes ne sont pas cependant dépourvus de complexités, et l'analyse de leurs réarrangements présente des défis qui leur sont propres.

Le génome mitochondrial d'*A. thaliana* a été assemblé à partir de l'écotype C24, tandis que les génomes du noyau et du chloroplaste sont dérivés de l'écotype Col-0 (104). Bien que les séquences mitochondriales des différents écotypes semblent très similaires, nos résultats ont illustré que certains blocs de plusieurs kb ont des positions différentes chez Col-0 par rapport à C24 (section 2.4.1). Grâce à la sensibilité de l'approche, ces positions ont initialement affiché des réarrangements spécifiques en quantités dans le même ordre de grandeur que la profondeur moyenne de séquençage. Ces cas particuliers ont donc nécessité un traitement manuel des résultats obtenus afin de compenser pour les différences entres les écotypes. Cependant, ceci illustre bien la capacité de notre approche à distinguer entre une variation somatique et un réarrangement rare.

Le génome mitochondrial humain, à environ 16 kb, est le plus petit génome que nous ayons étudié avec notre approche. Bien que ceci présente d'énormes avantages en termes de ressources informatiques nécessaires et de la fiabilité des alignements partiels obtenus, il s'agit également d'un désavantage pour la classification des réarrangements. En effet, les inversions sont facilement identifiables, mais il est presque impossible de distinguer entre une délétion et une duplication en tandem. La polyploïdie et la circularité du génome mitochondrial font en sorte que notre approche ne puisse pas distinguer entre la délétion d'une moitié du génome et une duplication de l'autre moitié. En posant l'hypothèse qu'une très courte délétion est plus probable qu'une duplication de la presque totalité du génome, la taille de la séquence affectée pourrait donner un indice à ce niveau. Cependant, il est difficile de déterminer une taille limite

entre les deux options, considérant que la distance maximale possible entre deux bases du génome est uniquement d'environ 8 kb. Pour contrevenir à cette limitation, il faudrait également avoir recours à une méthode qui permet l'étude de molécules uniques, telle que le FISH sur fibre (voir section 1.1.1.1). Ceci permettrait de combiner les résultats obtenus en séquençage aux différentes formes observées en microscopie pour les génomes entiers.

Un problème similaire est également survenu pendant l'étude du génome du chloroplaste d'*A. thaliana*. Les génomes de chloroplaste des plantes terrestres comprennent deux séquences inversées répétées (IR) identiques séparées de part et d'autre par deux séquences uniques (105). Pour éviter la présence d'alignements partiels identiques dans notre analyse, nous avons dû enlever la seconde IR du génome de référence utilisé. Dans tous les cas, il serait impossible de distinguer à laquelle des deux IR le réarrangement devrait être assigné. Comme pour les différences du génome mitochondrial entre les écotypes, il a ainsi été nécessaire de traiter manuellement les cas spécifiques à l'intersection des séquences uniques et des IR. La présence des IR ajoute également à la complexité de classification des réarrangements. Comme les mitochondries humaines, les chloroplastes sont polyploïdes et ne possèdent pas d'extrémités définies (106). Il est non seulement difficile de distinguer les délétions des duplications en tandem dans les séquences uniques, mais tous les types de réarrangements sont confondus lorsqu'un des alignements partiels se situe dans une IR : une inversion entre une séquence unique et une IR est identique à une délétion ou duplication entre la même séquence unique et l'autre copie de l'IR. Cette particularité rend difficile l'étude des réarrangements selon les positions relatives des séquences impliquées, et il n'existe aucune approche évidente permettant de contourner ce problème. Les technologies telles que le séquençage en temps réel de molécules uniques (SMRT, *single-molecule real-time sequencing*) de PacBio permet l'obtention de séquences de plusieurs kb provenant d'une seule molécule. Avec des séquences suffisamment longues, il serait possible de détecter les réarrangements dans les IR avec l'information nécessaire pour les placer dans leur contexte moléculaire exact. De telles approches ont déjà été développées pour l'étude de variation de séquences dans le génome humain (107).

Pour les génomes trop grands ou contenant trop de séquences similaires, l'analyse des alignements partiels à grande échelle dans un tabulateur devient impossible, car le nombre d'alignements partiels variable pour chaque séquence fait en sorte que chaque cas doit être traité

individuellement. Bien que la majorité des réarrangements soient presque triviaux lorsqu'observés un à un, le nombre de données dans un échantillon rend irréaliste l'analyse manuelle. Nous avons ainsi développé SCARR, pour permettre d'étendre notre approche à de plus grands génomes. Les résultats obtenus à partir du génome nucléaire de *Saccharomyces cerevisiae* illustrent bien l'efficacité du script, qui permet la détection de dizaines de milliers de réarrangements en quelques heures sur un ordinateur de bureau.

La transition du génome de *S. cerevisiae* à celui de l'humain, qui est environ 250 fois plus grand, a cependant compliqué la gestion des ressources informatiques. En plus de nécessiter plus de temps, l'alignement partiel des séquences par BLAST+ génère des fichiers de l'ordre de plusieurs gigaoctets pour une centaine de séquences de 100 paires de bases. L'analyse a donc été effectuée en parallèle à partir de fichiers d'environ 25 séquences afin de pouvoir effacer la sortie de BLAST+ une fois la détection par SCARR terminée. En plus de ralentir l'étude, la grande taille du génome requiert également une plus grande quantité de données pour obtenir une couverture moyenne de séquençage comparable. Il est donc plus difficile de différencier un événement rare d'une variation somatique présente dans la séquence. L'utilisation d'un modèle en levure demeure donc préférable à des cellules humaines lorsque cette option est disponible.

4.3 Tendances semblables entre les génomes

Malgré la distance évolutive qui sépare *A. thaliana*, *S. cerevisiae* et *Homo sapiens*, certains éléments communs sont observables dans les réarrangements de leurs différents génomes. Le plus flagrant de ceux-ci est la prépondérance des réarrangements à courte distance, qui suggère que l'instabilité génomique dépend en grande partie de la proximité entre les séquences réarrangées. Ceci concorde notamment avec les mécanismes proposés pour la recherche d'homologie dans les mécanismes de réparation basés sur la recombinaison (section 1.2.1.1). Bien que ceci ne soit pas particulièrement surprenant, ce résultat illustre l'importance des mécanismes de réparation qui ne requièrent pas de grandes homologies de séquences. Cette observation est possible en grande partie grâce à la sensibilité de détection des réarrangements, puisque les méthodes moins sensibles favorisent l'identification d'événements fréquents, qui sont souvent davantage favorisés par l'homologie de séquence que par la proximité.

Un des aspects les plus intéressants des réarrangements à courte distance est la fréquence relative de détection des inversions par rapport à celle des délétions ou duplications. À la fois dans le chloroplaste et dans le noyau de la levure, les courtes délétions et duplications, souvent dus au dérapage de la polymérase, sont beaucoup plus rares que les inversions à courte distance. Par contre, environ 40% des réarrangements détectés dans le noyau humain sont des délétions ou duplications de moins de 50 bases. Cette proportion est plus que deux fois plus grande que celle représentée par l'ensemble des inversions. Il est intéressant d'observer une telle différence dans le génome humain, puisque les courtes délétions et duplications et les inversions à courte distance sont des événements souvent associés à la réplication. Ceci suggère que ce processus à la base du métabolisme de l'ADN présente une source d'instabilité supplémentaire dans certains génomes. On peut toutefois difficilement identifier quels facteurs influencent ces niveaux élevés de dérapage de polymérase, puisqu'aucune condition de stress et aucun mutant n'ont été analysés dans notre étude. Puisque le génome humain est le seul dans lequel nous avons observé cette tendance, il est possible d'émettre l'hypothèse que ce genre de réarrangement est une conséquence nécessaire à la processivité de la réplication dans un grand génome.

4.4 L'étude de l'instabilité liée aux fourches bloquées

Bien que notre approche de détection des réarrangements ait été développée dans le but d'étudier les mécanismes de réparation de l'ADN en général, nos résultats indiquent que cette approche est particulièrement bien adaptée à l'étude de l'instabilité liée aux fourches de réplication bloquées. Les inversions à courte distance qui résultent des demi-tours de réplication ont été observées en bactérie et en levure par plusieurs groupes, mais les mécanismes qui sont à leur origine ont longtemps été l'objet de suppositions (96). Ces inversions se traduisent en séquences palindromiques ou quasi-palindromiques, qui présentent une tendance à former des structures secondaires. La présence d'épingles dans ces molécules les rend difficiles à détecter par les méthodes biochimiques traditionnelles telles que le séquençage Sanger, l'amplification PCR et le buvardage de Southern (108).

Les données de séquençage ont ainsi permis l'identification de nouveaux facteurs impliqués dans les demi-tours de réplication. Dans le chloroplaste, qui possède une machinerie de réparation de l'ADN similaire à celle des bactéries, l'absence combinée de la protéine de

recombinaison RECA et des protéines de liaison à l'ADN simple-brin Whirly cause une augmentation par un facteur de 60 des demi-tours. Dans le génome nucléaire de levure, la quantité de demi-tours est corrélée avec la stabilisation des fourches de réplication bloquées. Les demi-tours de réplication semblent ainsi servir de marqueur pour l'identification de protéines impliquées dans les fourches bloquées, que ce soit pour leur dégradation (Mre11 et Exo1 en levure), leur stabilisation (Rad9 en levure), pour le redémarrage de la fourche (RECA1 dans le chloroplaste et Mph1 chez la levure) ou pour empêcher les mauvais appariements (Whirly chez le chloroplaste). Les effets observés chez le chloroplaste étant très modestes chez les simples mutants, ceci suggère que l'utilisation de doubles mutants en levure permettrait également de raffiner davantage les modèles de prise en charge des fourches bloquées par les protéines de réparation.

4.5 Les conséquences possibles des demi-tours de réplication

Chez *S. cerevisiae* et *S. pombe*, les demi-tours de réplication sont associés à l'apparition d'isochromosomes, constitués de deux séquences en image miroir autour d'une position centrale (19, 94). Un isochromosome est le résultat attendu après un demi-tour sur un chromosome linéaire si la réplication procède de façon ininterrompue jusqu'à l'extrémité. Étant donné la difficulté à observer les courtes inversions par des méthodes biochimiques classiques, cette conséquence était le meilleur marqueur pour les demi-tours. La possibilité de détecter les courtes inversions à grande échelle et à travers le génome permet cependant de comparer la formation d'isochromosomes à d'autres conséquences possibles.

La comparaison des génomes de l'humain et du chimpanzé a révélé l'existence de réarrangements correspondant à un demi-tour de réplication après lequel le brin naissant retourne sur sa matrice originale après une courte élongation en utilisant le brin opposé (95). Étant donné la longueur limitée des lectures de séquençage analysées jusqu'à présent, il demeure difficile de déterminer le contexte exact des inversions à courte distance. Cependant, les résultats obtenus dans les tissus humains sains contiennent un certain nombre de doubles inversions correspondant aux variations de séquences observées entre l'humain et le chimpanzé. La possibilité de détecter une double inversion par séquençage est limitée par la longueur du brin polymérisé sur la mauvaise matrice, qui doit absolument être plus courte que la longueur des

lectures du séquenceur. Le séquençage SMRT présente ainsi une option intéressante pour déterminer la longueur moyenne des segments polymérisés en utilisant le brin opposé comme matrice. Le séquençage ne permet cependant pas de confirmer la présence d'un isochromosome, et la comparaison de ceux-ci par rapport aux doubles inversions devra être effectuée indirectement en utilisant des méthodes de détection différentes.

L'impact d'un demi-tour sur la viabilité de la cellule peut varier grandement selon le site et la longueur du segment répliqué à partir de la mauvaise matrice. Les isochromosomes possèdent soit aucun ou deux centromères, ce qui nuit à leur ségrégation pendant la division cellulaire et résulte en une monosomie ou trisomie des séquences affectées. Par contre, les doubles inversions séparées par une courte séquence intermédiaire ont contribué à l'évolution du génome humain (95), et se retrouvent également dans les tissus sains (chapitre 3), indiquant que leur impact peut être nul ou faible, particulièrement si elles ont lieu dans les séquences intergéniques. Puisque les demi-tours permettent le redémarrage de la réplication même si la lésion à l'origine du blocage n'est pas réparée, il est donc possible qu'ils agissent comme mécanisme de survie dans certaines situations. Par exemple, les cellules humaines qui ne possèdent pas la protéine BRCA2 sont hypersensibles aux agents qui endommagent l'ADN, mais cette hypersensibilité disparaît en l'absence de MRE11. Puisque MRE11 est impliquée dans la résection des extrémités lors de la réparation de l'ADN, il est surprenant que son absence permette à la cellule de mieux survivre aux dommages à l'ADN. Il semble donc probable que l'augmentation du nombre de demi-tours en l'absence de MRE11 contribue à la survie des cellules lorsque celles-ci sont autrement incapables de terminer la réplication. L'impact d'un demi-tour dépendrait ainsi des conséquences associées à l'impossibilité de redémarrer la fourche de réplication par un mécanisme conservateur.

Le génome des chloroplastes existant en une combinaison de formes circulaires, linéaires et branchées (106), la formation d'isochromosomes est plus difficile à examiner dans ce cas. Puisqu'il n'est toujours pas clair si l'ADN linéaire et les formes branchées sont uniquement le résultat de bris et de réparations aberrantes dans les génomes circulaires, il est difficile de déterminer l'impact d'un isochromosome formé à partir d'une molécule linéaire. Les chloroplastes étant polyploïdes, il est cependant fort probable que la duplication d'une moitié du génome et la délétion de l'autre moitié sur une seule molécule n'ait aucun impact.

L'augmentation des demi-tours de réplication dans le triple mutant *why1why3reca1* d'*A. thaliana* est pourtant associée à un phénotype de variégation qui indique une déficience dans la biogénèse des chloroplastes. Un demi-tour dans une molécule circulaire a comme conséquences possibles la formation d'une molécule circulaire qui contient deux copies de la séquence originale ou la formation d'un cercle extrachromosomique contenant deux copies d'une partie de la séquence (96). Dans un contexte polyploïde, il est improbable que de telles molécules aient un impact aussi grand sur la viabilité des chloroplastes.

Il est par contre intéressant de noter que les chloroplastes ne sont pas les seules organelles à présenter des génomes branchés. Le génome des mitochondries dans les cellules du cœur et du cerveau humains sont également associés en grand complexes branchés regroupant plusieurs copies du génome (109). Le changement de matrice qui se produit pendant un demi-tour de réplication résulte en une jonction en Y dans la molécule d'ADN. Si cette structure n'est pas clivée par une résolvase ou un endonucléase, elle pourrait résulter en un branchement tel que ceux observés dans les chloroplastes et dans les mitochondries de cerveau et de cœur. Puisque nos données montrent une abondance d'inversions à courte distance dans ces génomes, il serait intéressant d'investiguer la possibilité que les demi-tours soient à l'origine de certains de ces branchements. D'autre part, une trop grande augmentation du nombre de branchements, telle qu'observée dans les chloroplastes de *why1why3reca1*, pourrait occasionner des problèmes de ségrégation de l'ADN lors de la division des chloroplastes. Les chloroplastes possédant trop peu de copies du génome pourraient avoir de la difficulté à former leurs complexes photosynthétiques.

4.6 L'étude des mécanismes de réparation en cellules humaines

Bien que l'utilisation de la levure comme modèle pour l'étude des fourches bloquées chez l'humain permette d'accélérer grandement l'obtention de résultats, il n'est pas clair si tous les processus observés sont conservées entre ces deux espèces. Par exemple, la régression de la fourche de réplication est bien établie comme mécanisme conservateur pour le redémarrage des fourches bloquées chez l'humain, mais semble surtout être une source d'instabilité génomique chez la levure (92). Il sera donc éventuellement nécessaire de confirmer chez l'humain le rôle des différents facteurs identifiés en levure. Pour réduire les coûts et le temps nécessaire à l'étude

chez l'humain, il demeure cependant préférable d'explorer les fourches bloquées en levure autant que possible avant de changer d'organisme. Il sera ainsi plus facile de travailler en cellules humaines avec des hypothèses bien définies, de façon à limiter les expériences exploratoires.

Pour l'étude de certains mécanismes, comme l'alt-EJ, les différences entre la levure et l'humain sont suffisamment importantes pour justifier l'utilisation de cellules humaines. La présence de la polymérase POLθ, absente chez la levure, occasionne des différences majeures au niveau du mécanisme et de l'utilisation de l'alt-EJ chez l'humain (76). L'utilisation d'organismes modèles est également difficile pour l'étude des réarrangements présents dans les cancers et autres désordres génomiques. Il peut alors être intéressant de continuer l'optimisation de l'approche, que ce soit au niveau du premier tri sur Galaxy, de l'alignement partiel des séquences par BLAST+, ou de l'identification des réarrangements par SCARR.

4.7 Conclusion et perspectives

À partir de l'objectif de mettre au point une nouvelle méthode pour étudier les réarrangements génomiques, nous avons réalisé que les inversions à courte distance sont un type d'instabilité plus fréquent que les connaissances à ce sujet laissaient croire. L'association de ces événements à un mécanisme lié à des fourches de réplication bloquées, le demi-tour de réplication, nous confirme que l'instabilité génomique associée à des stress réplicatifs demeure en grande partie incomprise. Nos travaux ont permis de mettre au point une approche très bien adaptée à la détection de ces réarrangements, mais l'étude des protéines impliquées et de leur rôle dans ce mécanisme nécessitera également l'utilisation d'autres méthodes biochimiques et bio-informatiques. L'information obtenue à l'échelle du génome par séquençage de nouvelle génération devra être combinée à l'étude de blocages de réplication à des sites spécifiques de façon à valider les modèles obtenus.

Nos travaux ouvrent également la porte à l'étude de l'impact des demi-tours de réplication, à la fois dans les génomes nucléaires eucaryotes et dans les génomes de chloroplastes et de mitochondries. Il sera particulièrement intéressant, considérant les topologies très différentes de ces deux types de génomes, d'élucider l'impact que peuvent avoir les demi-tours sur la structure de l'ADN. La divergence proposée entre les rôles de la régression de la

fourche chez la levure et chez l'humain sera également un sujet de grand intérêt dans l'étude des fourches de réplication bloquées. Une meilleure compréhension du mécanisme de demi-tour sera nécessaire pour déterminer s'il s'agit uniquement d'une source d'instabilité génomique, ou s'il agit également comme mécanisme de survie dans des situations de stress spécifiques.

Bibliographie

1. Carvalho,C.M.B. and Lupski,J.R. (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.*, **17**, 224–238.
2. Vasileiou,P., Mourouzis,I. and Pantos,C. (2017) Principal Aspects Regarding the Maintenance of Mammalian Mitochondrial Genome Integrity. *Int. J. Mol. Sci.*, **18**, 1821.
3. Rusecka,J., Kaliszewska,M., Bartnik,E. and Tońska,K. (2018) Nuclear genes involved in mitochondrial diseases caused by instability of mitochondrial DNA. *J. Appl. Genet.*, **59**, 43–57.
4. Rowan,B.A., Oldenburg,D.J. and Bendich,A.J. (2010) RecA maintains the integrity of chloroplast DNA molecules in Arabidopsis. *J. Exp. Bot.*, **61**, 2575–88.
5. Parent,J.-S., Lepage,E. and Brisson,N. (2011) Divergent Roles for the Two PolI-Like Organelle DNA Polymerases of Arabidopsis. *PLANT Physiol.*, **156**, 254–262.
6. Xu,Y.-Z., Arrieta-Montiel,M.P., Virdi,K.S., de Paula,W.B.M., Widhalm,J.R., Basset,G.J., Davila,J.I., Elthon,T.E., Elowsky,C.G., Sato,S.J., *et al.* (2011) MutS HOMOLOG1 is a nucleoid protein that alters mitochondrial and plastid properties and plant response to high light. *Plant Cell*, **23**, 3428–41.
7. Lepage,É., Zampini,É. and Brisson,N. (2013) Plastid Genome Instability Leads to Reactive Oxygen Species Production and Plastid-to-Nucleus Retrograde Signaling in Arabidopsis. *Plant Physiol.*, **163**, 867–881.
8. Maréchal,A., Parent,J.-S., Veronneau-Lafortune,F., Joyeux,A., Lang,B.F. and Brisson,N. (2009) Whirly proteins maintain plastid genome stability in Arabidopsis. *Proc. Natl. Acad. Sci.*, **106**, 14693–14698.
9. Nikiforova,M.N., Hsi,E.D., Braziel,R.M., Gulley,M.L., Leonard,D.G.B., Nowak,J.A., Tubbs,R.R., Vance,G.H., Van Deerlin,V.M. and Molecular Pathology Resource Committee,C. of A.P. (2007) Detection of clonal IGH gene rearrangements: summary of molecular oncology surveys of the College of American Pathologists. *Arch. Pathol. Lab. Med.*, **131**, 185–9.
10. 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.

11. Aten,E., White,S.J., Kalf,M.E., Vossen,R.H.A.M., Thygesen,H.H., Ruivenkamp,C.A., Kriek,M., Breuning,M.H.B. and den Dunnen,J.T. (2009) Methods to detect CNVs in the human genome. *Cytogenet. Genome Res.*, **123**, 313–321.
12. Savic,S. and Bubendorf,L. (2016) Common Fluorescence In Situ Hybridization Applications in Cytology. *Arch. Pathol. Lab. Med.*, **140**, 1323–1330.
13. Matsubara,K., Knopp,T., Sarre,S.D., Georges,A. and Ezaz,T. (2013) Karyotypic analysis and FISH mapping of microsatellite motifs reveal highly differentiated XX/XY sex chromosomes in the pink-tailed worm-lizard (Aprasia parapulchella, Pygopodidae, Squamata). *Mol. Cytogenet.*, **6**, 60.
14. Jackson,S.A., Wang,M.L., Goodman,H.M. and Jiang,J. (1998) Application of fiber-FISH in physical mapping of *Arabidopsis thaliana*. *Genome*, **41**, 566–572.
15. Koo,D.-H., Singh,B., Jiang,J., Friebe,B., Gill,B.S., Chastain,P.D., Manne,U., Tiwari,H.K. and Singh,K.K. (2017) Single molecule mtDNA fiber FISH for analyzing numtogenesis. *Anal. Biochem.*, 10.1016/j.ab.2017.03.015.
16. Shuga,J., Zeng,Y., Novak,R., Lan,Q., Tang,X., Rothman,N., Vermeulen,R., Li,L., Hubbard,A., Zhang,L., *et al.* (2013) Single molecule quantitation and sequencing of rare translocations using microfluidic nested digital PCR. *Nucleic Acids Res.*, **41**, e159–e159.
17. Huggett,J.F., Whale,A., Lau,T.K., Leung,T.N., Wong,E.M. and Lo,Y.M. (2013) Digital PCR as a novel technology and its potential implications for molecular diagnostics. *Clin. Chem.*, **59**, 1691–3.
18. den Dunnen,J.T. and White,S.J. (2006) MLPA and MAPH: Sensitive Detection of Deletions and Duplications. In *Current Protocols in Human Genetics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, Vol. Chapter 7, p. 7.14.1-7.14.20.
19. Mizuno,K., Lambert,S., Baldacci,G., Murray,J.M. and Carr,A.M. (2009) Nearby inverted repeats fuse to generate acentric and dicentric palindromic chromosomes by a replication template exchange mechanism. *Genes Dev.*, **23**, 2876–86.
20. Mizuno,K., Miyabe,I., Schalbetter,S.A., Carr,A.M. and Murray,J.M. (2012) Recombination-restarted replication makes inverted chromosome fusions at inverted repeats. *Nature*, **493**, 246–249.
21. Lin,K., Smit,S., Bonnema,G., Sanchez-Perez,G. and de Ridder,D. (2015) Making the

difference: integrating structural variation detection tools. *Brief. Bioinform.*, **16**, 852–864.

22. Medvedev,P., Stanciu,M. and Brudno,M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**, S13–S20.
23. Williams,S.L., Huang,J., Edwards,Y.J.K., Ulloa,R.H., Dillon,L.M., Prolla,T.A., Vance,J.M., Moraes,C.T. and Züchner,S. (2010) The mtDNA Mutation Spectrum of the Progeroid Polg Mutator Mouse Includes Abundant Control Region Multimers. *Cell Metab.*, **12**, 675–682.
24. Li,H. (2015) FermiKit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics*, 10.1093/bioinformatics/btv440.
25. Rimmer,A., Phan,H., Mathieson,I., Iqbal,Z., Twigg,S.R.F., Wilkie,A.O.M., McVean,G. and Lunter,G. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, **46**, 912–918.
26. Chong,Z., Ruan,J., Gao,M., Zhou,W., Chen,T., Fan,X., Ding,L., Lee,A.Y., Boutros,P., Chen,J., *et al.* (2016) novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat. Methods*, 10.1038/nmeth.4084.
27. Campbell,P.J., Stephens,P.J., Pleasance,E.D., O'Meara,S., Li,H., Santarius,T., Stebbings,L.A., Leroy,C., Edkins,S., Hardy,C., *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–9.
28. Sindi,S.S., Onal,S., Peng,L.C., Wu,H.-T. and Raphael,B.J. (2012) An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.*, **13**, R22.
29. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Hsi-Yang Fritz,M., *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
30. Rausch,T., Zichner,T., Schlattl,A., Stutz,A.M., Benes,V. and Korbel,J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.
31. Yang,R., Chen,L., Newman,S., Gandhi,K., Doho,G., Moreno,C.S., Vertino,P.M., Bernal-Mizarchi,L., Lonial,S., Boise,L.H., *et al.* (2014) Integrated analysis of whole-genome paired-end and mate-pair sequencing data for identifying genomic structural variations in multiple myeloma. *Cancer Inform.*, **13**, 49–53.

32. Layer,R.M., Chiang,C., Quinlan,A.R. and Hall,I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.
33. Kronenberg,Z.N., Osborne,E.J., Cone,K.R., Kennedy,B.J., Domyan,E.T., Shapiro,M.D., Elde,N.C. and Yandell,M. (2015) Wham: Identifying Structural Variants of Biological Consequence. *PLOS Comput. Biol.*, **11**, e1004572.
34. Faust,G.G. and Hall,I.M. (2012) YAHA: fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics*, **28**, 2417–24.
35. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.
36. 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**, 3484–90.
37. Chen,R., Lau,Y.L., Zhang,Y. and Yang,W. (2016) SRinversion: a tool for detecting short inversions by splitting and re-aligning poorly mapped and unmapped sequencing reads. *Bioinformatics*, **32**, btw516.
38. He,F., Li,Y., Tang,Y.-H., Ma,J. and Zhu,H. (2016) Identifying micro-inversions using high-throughput sequencing reads. *BMC Genomics*, **17**, 4.
39. Ranjha,L., Howard,S.M. and Cejka,P. (2018) Main steps in DNA double-strand break repair: an introduction to homologous recombination and related processes. *Chromosoma*, 10.1007/s00412-017-0658-1.
40. Kowalczykowski,S.C. (2015) An Overview of the Molecular Mechanisms of Recombinational DNA Repair. *Cold Spring Harb. Perspect. Biol.*, **7**, a016410.
41. Zhu,Z., Chung,W.-H., Shim,E.Y., Lee,S.E. and Ira,G. (2008) Sgs1 helicase and two nucleases Dna2 and Exo1 resect DNA double-strand break ends. *Cell*, **134**, 981–94.
42. Shibata,A., Steinlage,M., Barton,O., Juhász,S., Künzel,J., Spies,J., Shibata,A., Jeggo,P.A. and Löbrich,M. (2017) DNA Double-Strand Break Resection Occurs during Non-homologous End Joining in G1 but Is Distinct from Resection during Homologous Recombination. *Mol. Cell*, **65**.
43. Anand,R., Ranjha,L., Cannavo,E. and Cejka,P. (2016) Phosphorylated CtIP Functions as a Co-factor of the MRE11-RAD50-NBS1 Endonuclease in DNA End Resection. *Mol. Cell*, **64**, 940–950.
44. Levikova,M., Pinto,C. and Cejka,P. (2017) The motor activity of DNA2 functions as an

ssDNA translocase to promote DNA end resection. *Genes Dev.*, **31**, 493–502.

45. Sugiyama,T., Zaitseva,E.M. and Kowalczykowski,S.C. (1997) A single-stranded DNA-binding protein is needed for efficient presynaptic complex formation by the Saccharomyces cerevisiae Rad51 protein. *J. Biol. Chem.*, **272**, 7940–5.
46. Forget,A.L. and Kowalczykowski,S.C. (2012) Single-molecule imaging of DNA pairing by RecA reveals a three-dimensional homology search. *Nature*, **482**, 423–7.
47. Eggler,A.L., Inman,R.B. and Cox,M.M. (2002) The Rad51-dependent pairing of long DNA substrates is stabilized by replication protein A. *J. Biol. Chem.*, **277**, 39280–8.
48. Wright,W.D. and Heyer,W.-D. (2014) Rad54 functions as a heteroduplex DNA pump modulated by its DNA substrates and Rad51 during D loop formation. *Mol. Cell*, **53**, 420–32.
49. Li,X., Stith,C.M., Burgers,P.M. and Heyer,W.-D. (2009) PCNA is required for initiation of recombination-associated DNA synthesis by DNA polymerase delta. *Mol. Cell*, **36**, 704–13.
50. Wilson,M.A., Kwon,Y., Xu,Y., Chung,W.-H., Chi,P., Niu,H., Mayle,R., Chen,X., Malkova,A., Sung,P., *et al.* (2013) Pif1 helicase and Polδ promote recombination-coupled DNA synthesis via bubble migration. *Nature*, **502**, 393–6.
51. Hicks,W.M., Kim,M. and Haber,J.E. (2010) Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. *Science*, **329**, 82–5.
52. Nimonkar,A. V, Sica,R.A. and Kowalczykowski,S.C. (2009) Rad52 promotes second-end DNA capture in double-stranded break repair to form complement-stabilized joint molecules. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 3077–82.
53. Feng,Z., Scott,S.P., Bussen,W., Sharma,G.G., Guo,G., Pandita,T.K. and Powell,S.N. (2011) Rad52 inactivation is synthetically lethal with BRCA2 deficiency. *Proc. Natl. Acad. Sci.*, **108**, 686–691.
54. Chen,X.B., Melchionna,R., Denis,C.M., Gaillard,P.H., Blasina,A., Van de Weyer,I., Boddy,M.N., Russell,P., Vialard,J. and McGowan,C.H. (2001) Human Mus81-associated endonuclease cleaves Holliday junctions in vitro. *Mol. Cell*, **8**, 1117–27.
55. Sarbajna,S., Davies,D. and West,S.C. (2014) Roles of SLX1-SLX4, MUS81-EME1, and GEN1 in avoiding genome instability and mitotic catastrophe. *Genes Dev.*, **28**, 1124–1136.
56. Wu,L. and Hickson,I.D. (2003) The Bloom's syndrome helicase suppresses crossing over

during homologous recombination. *Nature*, **426**, 870–874.

57. Sakofsky,C.J. and Malkova,A. (2017) Break induced replication in eukaryotes: mechanisms, functions, and consequences. *Crit. Rev. Biochem. Mol. Biol.*, **52**, 395–413.
58. Saini,N., Ramakrishnan,S., Elango,R., Ayyar,S., Zhang,Y., Deem,A., Ira,G., Haber,J.E., Lobachev,K.S. and Malkova,A. (2013) Migrating bubble during break-induced replication drives conservative DNA synthesis. *Nature*, **502**, 389–392.
59. Hastings,P.J., Lupski,J.R., Rosenberg,S.M. and Ira,G. (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, **10**, 551–564.
60. Mitchel,K., Lehner,K. and Jinks-Robertson,S. (2013) Heteroduplex DNA position defines the roles of the Sgs1, Srs2, and Mph1 helicases in promoting distinct recombination outcomes. *PLoS Genet.*, **9**, e1003340.
61. Bugreev,D. V., Yu,X., Egelman,E.H. and Mazin,A. V. (2007) Novel pro- and anti-recombination activities of the Bloom's syndrome helicase. *Genes Dev.*, **21**, 3085–3094.
62. Barber,L.J., Youds,J.L., Ward,J.D., McIlwraith,M.J., O'Neil,N.J., Petalcorin,M.I.R., Martin,J.S., Collis,S.J., Cantor,S.B., Auclair,M., *et al.* (2008) RTEL1 maintains genomic stability by suppressing homologous recombination. *Cell*, **135**, 261–71.
63. Stankiewicz,P. and Lupski,J.R. (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet.*, **18**, 74–82.
64. Waldman,A.S. and Liskay,R.M. (1988) Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. *Mol. Cell. Biol.*, **8**, 5350–7.
65. Zhang,F., Potocki,L., Sampson,J.B., Liu,P., Sanchez-Valle,A., Robbins-Furman,P., Navarro,A.D., Wheeler,P.G., Spence,J.E., Brasington,C.K., *et al.* (2010) Identification of uncommon recurrent Potocki-Lupski syndrome-associated duplications and the distribution of rearrangement types and mechanisms in PTLS. *Am. J. Hum. Genet.*, **86**, 462–70.
66. Boone,P.M., Liu,P., Zhang,F., Carvalho,C.M.B., Towne,C.F., Batish,S.D. and Lupski,J.R. (2011) Alu-specific microhomology-mediated deletion of the final exon of SPAST in three unrelated subjects with hereditary spastic paraplegia. *Genet. Med.*, **13**, 582–92.
67. 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.

68. Shinohara,A., Shinohara,M., Ohta,T., Matsuda,S. and Ogawa,T. (1998) Rad52 forms ring structures and co-operates with RPA in single-strand DNA annealing. *Genes Cells*, **3**, 145–56.

69. Motycka,T.A., Bessho,T., Post,S.M., Sung,P. and Tomkinson,A.E. (2004) Physical and functional interaction between the XPF/ERCC1 endonuclease and hRad52. *J. Biol. Chem.*, **279**, 13634–9.

70. Jazayeri,A., Falck,J., Lukas,C., Bartek,J., Smith,G.C.M., Lukas,J. and Jackson,S.P. (2006) ATM- and cell cycle-dependent regulation of ATR in response to DNA double-strand breaks. *Nat. Cell Biol.*, **8**, 37–45.

71. Chang,H.H.Y., Watanabe,G., Gerodimos,C.A., Ochi,T., Blundell,T.L., Jackson,S.P. and Lieber,M.R. (2016) Different DNA End Configurations Dictate Which NHEJ Components Are Most Important for Joining Efficiency. *J. Biol. Chem.*, **291**, 24377–24389.

72. Gu,J., Lu,H., Tippin,B., Shimazaki,N., Goodman,M.F. and Lieber,M.R. (2007) XRCC4:DNA ligase IV can ligate incompatible DNA ends and can ligate across gaps. *EMBO J.*, **26**, 1010–23.

73. Yaneva,M., Kowalewski,T. and Lieber,M.R. (1997) Interaction of DNA-dependent protein kinase with DNA and with Ku: biochemical and atomic-force microscopy studies. *EMBO J.*, **16**, 5098–5112.

74. Goodarzi,A.A., Yu,Y., Riballo,E., Douglas,P., Walker,S.A., Ye,R., Härer,C., Marchetti,C., Morrice,N., Jeggo,P.A., *et al.* (2006) DNA-PK autophosphorylation facilitates Artemis endonuclease activity. *EMBO J.*, **25**, 3880–3889.

75. Nick McElhinny,S.A., Havener,J.M., Garcia-Diaz,M., Juárez,R., Bebenek,K., Kee,B.L., Blanco,L., Kunkel,T.A. and Ramsden,D.A. (2005) A gradient of template dependence defines distinct biological roles for family X polymerases in nonhomologous end joining. *Mol. Cell*, **19**, 357–66.

76. Sfeir,A. and Symington,L.S. (2015) Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem. Sci.*, **40**, 701–714.

77. Kent,T., Mateos-Gomez,P.A., Sfeir,A. and Pomerantz,R.T. (2016) Polymerase θ is a robust terminal transferase that oscillates between three different mechanisms during end-joining. *Elife*, **5**.

78. Truong,L.N., Li,Y., Shi,L.Z., Hwang,P.Y.-H., He,J., Wang,H., Razavian,N., Berns,M.W.

and Wu,X. (2013) Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proc. Natl. Acad. Sci.*, **110**.

79. Deng,S.K., Gibb,B., de Almeida,M.J., Greene,E.C. and Symington,L.S. (2014) RPA antagonizes microhomology-mediated repair of DNA double-strand breaks. *Nat. Struct. Mol. Biol.*, **21**, 405–12.
80. Audebert,M., Salles,B. and Calsou,P. (2004) Involvement of poly(ADP-ribose) polymerase-1 and XRCC1/DNA ligase III in an alternative route for DNA double-strand breaks rejoining. *J. Biol. Chem.*, **279**, 55117–26.
81. Simsek,D., Brunet,E., Wong,S.Y.-W., Katyal,S., Gao,Y., McKinnon,P.J., Lou,J., Zhang,L., Li,J., Rebar,E.J., *et al.* (2011) DNA ligase III promotes alternative nonhomologous end-joining during chromosomal translocation formation. *PLoS Genet.*, **7**, e1002080.
82. Vaisman,A. and Woodgate,R. (2017) Translesion DNA polymerases in eukaryotes: what makes them tick? *Crit. Rev. Biochem. Mol. Biol.*, **52**, 274–303.
83. Reijns,M.A.M., Rabe,B., Rigby,R.E., Mill,P., Astell,K.R., Lettice,L.A., Boyle,S., Leitch,A., Keighren,M., Kilanowski,F., *et al.* (2012) Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. *Cell*, **149**, 1008–22.
84. Viguera,E., Canceill,D. and Ehrlich,S.D. (2001) Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J.*, **20**, 2587–2595.
85. Petruska,J., Hartenstine,M.J. and Goodman,M.F. (1998) Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. *J. Biol. Chem.*, **273**, 5204–10.
86. Lopes,M., Cotta-Ramusino,C., Pellicioli,A., Liberi,G., Plevani,P., Muzi-Falconi,M., Newlon,C.S. and Foiani,M. (2001) The DNA replication checkpoint response stabilizes stalled replication forks. *Nature*, **412**, 557–561.
87. Sabatinos,S.A., Green,M.D. and Forsburg,S.L. (2012) Continued DNA synthesis in replication checkpoint mutants leads to fork collapse. *Mol. Cell. Biol.*, **32**, 4986–97.
88. Zellweger,R., Dalcher,D., Mutreja,K., Berti,M., Schmid,J.A., Herrador,R., Vindigni,A. and Lopes,M. (2015) Rad51-mediated replication fork reversal is a global response to genotoxic treatments in human cells. *J. Cell Biol.*, **208**, 563–79.
89. Kolinjivadi,A.M., Sannino,V., De Antoni,A., Zadorozhny,K., Kilkenny,M., Técher,H.,

Baldi,G., Shen,R., Ciccia,A., Pellegrini,L., *et al.* (2017) Smarcal1-Mediated Fork Reversal Triggers Mre11-Dependent Degradation of Nascent DNA in the Absence of Brca2 and Stable Rad51 Nucleofilaments. *Mol. Cell*, **67**, 867–881.e7.

90. Berti,M., Ray Chaudhuri,A., Thangavel,S., Gomathinayagam,S., Kenig,S., Vujanovic,M., Odreman,F., Glatter,T., Graziano,S., Mendoza-Maldonado,R., *et al.* (2013) Human RECQ1 promotes restart of replication forks reversed by DNA topoisomerase I inhibition. *Nat. Struct. Mol. Biol.*, **20**, 347–54.
91. Thangavel,S., Berti,M., Levikova,M., Pinto,C., Gomathinayagam,S., Vujanovic,M., Zellweger,R., Moore,H., Lee,E.H., Hendrickson,E.A., *et al.* (2015) DNA2 drives processing and restart of reversed replication forks in human cells. *J. Cell Biol.*, **208**, 545–62.
92. Neelsen,K.J. and Lopes,M. (2015) Replication fork reversal in eukaryotes: from dead end to dynamic response. *Nat. Rev. Mol. Cell Biol.*, **16**, 207–20.
93. Quinet,A., Lemaçon,D. and Vindigni,A. (2017) Replication Fork Reversal: Players and Guardians. *Mol. Cell*, **68**.
94. Paek,A.L., Kaochar,S., Jones,H., Elezaby,A., Shanks,L. and Weinert,T. (2009) Fusion of nearby inverted repeats by a replication-based mechanism leads to formation of dicentric and acentric chromosomes that cause genome instability in budding yeast. *Genes Dev.*, **23**, 2861–75.
95. Löytynoja,A. and Goldman,N. (2017) Short template switch events explain mutation clusters in the human genome. *Genome Res.*, **27**, 1039–1049.
96. Tremblay-Belzile,S., Lepage,É., Zampini,É. and Brisson,N. (2015) Short-range inversions: Rethinking organelle genome stability. *Bioessays*, **37**, 1086–1094.
97. Hsu,S.-C., Belmonte,M.F., Harada,J.J. and Inoue,K. (2010) Indispensable Roles of Plastids in Arabidopsis thaliana Embryogenesis. *Curr. Genomics*, **11**, 338–49.
98. Olejniczak,S.A., Łojewska,E., Kowalczyk,T. and Sakowicz,T. (2016) Chloroplasts: state of research and practical applications of plastome sequencing. *Planta*, **244**, 517–527.
99. Kobayashi,Y., Misumi,O., Odahara,M., Ishibashi,K., Hirono,M., Hidaka,K., Endo,M., Sugiyama,H., Iwasaki,H., Kuroiwa,T., *et al.* (2017) Holliday junction resolvases mediate chloroplast nucleoid segregation. *Science*, **356**, 631–634.
100. Odom,O.W., Baek,K.-H., Dani,R.N. and Herrin,D.L. (2008) Chlamydomonas chloroplasts

can use short dispersed repeats and multiple pathways to repair a double-strand break in the genome. *Plant J.*, **53**, 842–853.

101. Christensen,A.C. (2013) Plant Mitochondrial Genome Evolution Can Be Explained by DNA Repair Mechanisms. *Genome Biol. Evol.*, **5**, 1079–1086.
102. Zampini,É., Lepage,É., Tremblay-Belzile,S., Truche,S. and Brisson,N. (2015) Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in Arabidopsis and humans. *Genome Res.*, **25**, 645–654.
103. 1000 Genomes Project Consortium,G.A., Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T., McVean,G.A., *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
104. Sloan,D.B., Wu,Z. and Sharbrough,J. (2018) Correction of Persistent Errors in Arabidopsis Reference Mitochondrial Genomes. *Plant Cell*, **30**, 525–527.
105. Yurina,N.P., Sharapova,L.S. and Odintsova,M.S. (2017) Structure of plastid genomes of photosynthetic eukaryotes. *Biochem.*, **82**, 678–691.
106. Oldenburg,D.J. and Bendich,A.J. (2004) Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. *J. Mol. Biol.*, **335**, 953–70.
107. Huddleston,J., Chaisson,M.J.P., Steinberg,K.M., Warren,W., Hoekzema,K., Gordon,D., Graves-Lindsay,T.A., Munson,K.M., Kronenberg,Z.N., Vives,L., *et al.* (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.*, **27**, 677–685.
108. Rattray,A.J. (2004) A method for cloning and sequencing long palindromic DNA junctions. *Nucleic Acids Res.*, **32**, e155.
109. Pohjoismäki,J.L.O., Goffart,S., Tyynismaa,H., Willcox,S., Ide,T., Kang,D., Suomalainen,A., Karhunen,P.J., Griffith,J.D., Holt,I.J., *et al.* (2009) Human heart mitochondrial DNA is organized in complex catenated networks containing abundant four-way junctions and replication forks. *J. Biol. Chem.*, **284**, 21446–57.
110. Chang,H.H.Y., Pannunzio,N.R., Adachi,N. and Lieber,M.R. (2017) Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell Biol.*, **18**.