Thèse de doctorat de Sorbonne Université

Complexité du Vivant, Ecole Doctorale 515
Spécialité : Génétique

Présentée et soutenue publiquement le 14 Décembre 2018 par Aubin FLEISS
pour obtenir le grade de docteur de Sorbonne Université

Impact phénotypique des réarrangements chromosomiques et évolution des génomes de levures

---

Laboratoire de biologie computationnelle et quantitative (UMR7238)
Equipe de Biologie des Génomes[1]

Devant un jury composé de :

Karine DUBRANA	Rapportrice
Marie-Claude MARSOLIER-KERGOAT	Rapportrice
Jean-Baptiste BOULÉ	Examinateur
Gianni LITI	Examinateur
Frédérique PERONNET	Examinatrice
Gilles FISCHER	Directeur de thèse

[1] Sorbonne Université, Institut de Biologie Paris Seine, CNRS UMR 7238 : Laboratoire de Biologie Computationnelle et Quantitative, Equipe Biologie des Génomes, 7-9 Quai Saint Bernard, Bâtiment C, 3ème étage, Boite 1540, 75252 Paris Cedex 05

À mes grands-mères

To a Tree

Oh, tree outside my window, we are kin,
For you ask nothing of a friend but this:
To lean against the window and peer in
And watch me move about! Sufficient bliss

For me, who stand behind its framework stout,
Full of my tiny tragedies and grotesque grieves,
To lean agains the window and peer out,
Admiring infinitesimal leaves.[2]

À un arbre

Ô, arbre à ma fenêtre, toi et moi sommes des proches,
Puisque tu ne demandes rien d'autre d'un ami :
que de se pencher contre la vitre et de regarder au dedans
Comme je m'affaire. C'est un bonheur qui me suffit,

moi qui me tiens contre son cadre fort,
plein de mes chagrins grotesques et de mes tragédies banales,
de m'appuyer contre la vitre et de regarder au dehors,
d'admirer tes feuilles infinitésimales.[3]

[2] Elizabeth Bishop, *Collected Poems*, "*Written in Youth*" (London, Chatto & Windus, 2007)
[3] J'ai risqué une traduction.

Titre :

Impact phénotypique des réarrangements chromosomiques et évolution des génomes de levures.

Résumé :

Nous avons cherché à évaluer l'impact des réarrangements chromosomiques sur l'évolution des génomes de levures selon deux approches.

La première approche a consisté à retracer les réarrangements chromosomiques au cours de l'évolution des *Saccharomycotina*. Nous avons construit un arbre phylogénétique à partir de 66 génomes issus de bases de données publiques et reconstruit la structure des génomes ancestraux des 66 espèces. La comparaison des génomes ancestraux a permi d'inférer 5150 réarrangements chromosomiques passés. Nous avons montré que selon les clades considérés, les génomes évoluent plutôt par inversion ou par translocation et que les réarrangements chromosomiques et les mutations non-synonymes s'accumulent à un rythme coordonné au cours de l'évolution.

La seconde approche a consisté à quantifier l'impact phénotypique des variations structurelles (SV) du génome en termes de taux de croissance végétative et de viabilité méiotique chez *Saccharomyces cerevisiae*. Nous avons développé une technique pour induire à façon des SV ciblés dans le génome de *S. cerevisiae*, en induisant deux coupures simultanées dans le génome de *S. cerevisiae* avec CRISPR/Cas9 et à guider la réparation des cassures par recombinaison homologue avec des oligonucléotides chimériques. Nous avons alors adapté cette technique pour induire en une étape un grand nombre de SV aléatoires. L'impact phénotypique des SV obtenus a été quantifié en méiose et en croissance végétative. Ces travaux montrent que même des réarrangements chromosomiques balancés n'affectant aucune phase codante génèrent une grande diversité phénotypique qui participe à l'adaptation des organismes à leur environnement.

Mots-clés:

Chromosome, réarrangement, evolution, ancêtre, phénotype, levure.

Title:

Phenotypic impact of chromosomal rearrangements and evolution of yeast genomes.

Summary:

The aim of this work was to assess the impact of chromosomal rearrangements on the evolution of yeast genomes with two approaches.

The first approach consisted in retracing past rearrangements during the evolution of *Saccharomycotina* yeast genomes. We have built a phylogenetic tree of 66 genomes gathered from public databases, then reconstructed the structure of all ancestral genomes of these species. By comparing the structure of reconstructed ancestral genomes, we have inferred 5150 past rearrangements. We showed that depending on the clades, genomes tend to evolve mostly by inversion or by translocation. In addition, we showed that chromosomal rearrangements and non-synonymous mutations tend to accumulate at a coordinated pace during evolution.

The second approach aimed at quantifying the phenotypic impact of structural variations of chromosomes (SVs) in terms of vegetative growth and meiotic viability in *Saccharomyces cerevisiae*. We developed a technique to induce easily targeted SVs in the genome of *S. cerevisiae* by inducing two chromosomal breaks with CRISPR/Cas9 and providing the cells with chimerical donor oligonucleotides to repair the split chromosomes by homologous recombination. We have then adapted this technique to induce multiple random SVs in a single step. The phenotypic impact of obtained variants on vegetative growth and on spore viability was quantified. These results show that even balanced chromosomal rearrangements that do not affect coding sequence generate a wide phenotypic diversity that contributes to the adaptation of organisms to their environment.

Keywords:

Chromosome, rearrangement, evolution, ancestor, phenotype, yeast.

Remerciements

J'adresse tout d'abord mes sincères remerciements aux membres de mon jury de thèse : à Karine Dubrana et Marie-Claude Marsolier-Kergoat qui m'ont fait l'honneur d'être rapportrices, ainsi qu'à Frédérique Peronnet, Jean-Baptiste Boulé et à Gianni Liti d'avoir accepté d'être examinateurs de mon travail. J'ai conscience de la chance d'avoir ces scientifiques remarquables et généreux de leur temps dans mon jury. Je remercie particulièrement Gianni Liti pour ses conseils et son expertise bienveillante tout au long de ces trois années.

Mes remerciements vont à présent à Gilles Fischer, mon directeur de thèse, pour m'avoir accepté dans son équipe, pour avoir valorisé mon travail, m'avoir encouragé à le présenter. Gilles a fait preuve d'une grande écoute, évaluant sans *a priori* mes idées parfois alambiquées, parfois naïves, parfois très audacieuses et parfois assez éloignées de ma problématique même si, j'en ai conscience, canaliser ce foisonnement a dû demander une certaine énergie. Gilles m'a également appris à appréhender la difficulté sans gaspiller mes efforts avec cette phrase « Mais tu sais Aubin, tu peux aussi résoudre les problèmes s'ils se posent et quand ils se posent ! ». Cet enseignement est précieux.

Je remercie également ma tutrice Evelyne Téoulé pour ses encouragements et son soutien au cours de cette thèse. Evelyne a eu la gentillesse de me faire part de son recul sur son propre parcours et le mien et je l'en remercie sincèrement.

J'en viens à présent à remercier mes collègues de l'équipe de Biologie des Génomes. Je souris en écrivant ce paragraphe en repensant à tous ces bons moments passés à la paillasse avec Nicolas Agier et Stéphane Delmas (Tic&Tac). Les manips n'auraient pas été aussi sympathiques sans eux. Je me dois en particulier de saluer votre sélection musicale, le laboratoire de Biologie des Génomes ne serait pas le même sans votre *playlist*… éclectique. Les ingénieurs de *Youtube* ont dû se poser quelques questions, notamment devant notre lecture hebdomadaire (voire quotidienne) de *Tata yoyo*. Je crois d'ailleurs savoir que les stagiaires du laboratoire s'en souviendront longtemps. Je remercie particulièrement Nicolas pour son aide, son humour et pour nos discussions sur l'aquariophilie, la musique et tant d'autres choses. Merci également de m'avoir fait découvrir ce *remake* magique du générique de *Maya l'abeille*. Stéphane, merci pour ta bonne humeur constante, tes remarques constructives, ton sens de l'organisation. Le fait de t'entendre répéter « 12h, cantine. 13h, verveine. 13h10, je réponds à mes mails. 13h30, dissections de tétrades, … » a fini par porter ses fruits. Un grand merci également pour avoir proposé à tes collègues enseignants de construire un TP sur la base de mes manips. Cette semaine passée à encadrer tes étudiants de Master a été une très belle expérience.

Je ne saurais oublier de remercier les autres étudiants en thèse de l'équipe : Samuel O'Donnell qui fera, j'en suis sur une thèse brillante, pour son aide, son humour et les expressions anglaises qu'il m'a apprises. Je remercie également ceux qui sont déjà partis du laboratoire : Alexandre GILLET pour son ficus et ses astuces de codage en Python, Nikolaos Vakirlis pour ses conseils en début de thèse et pour avoir animé nos séminaires Doc'n'Post-docs ainsi que Guénola Drillon qui a été d'un grand soutien quand j'avais des questions sur le fonctionnement de *Chronicle*. Un grand merci également aux stagiaires de l'équipe : Alma Chapet-Battle pour sa véritable passion pour le cinéma et son humour pince-sans-rire, Jihanne Challita pour avoir fait occasionellement une petite place à mes souches sur ses gels de PFGE, Pierre-Emmanuel Bonté et Jeanne Mattei qui m'ont aidé dans mes manips. Je salue particulièrement Pierre-Emmanuel qui a su reprendre mes scripts non-annotés comme un chef et à qui je souhaite le meilleur dans son désir de poursuivre dans le beau domaine de la bio-informatique. Un grand merci à Maëllys Born-Bony avec qui les Pokémons et les jonctions chromosomiques récalcitrantes à l'amplification par PCR n'ont qu'à bien se tenir. Je salue les stagiaires avec qui j'ai partagé mon bureau, Mariam Sissoko et Chloé Quignot, entendre le

ronronnement des ordinateurs qui font de la bio-info est d'une poésie souvent mécomprise, mais vous y êtes sensibles.

Je salue également mes collègues des autres équipes : Chloé Dequeker et Antonin Thiébaut qui ont réalisé leur thèse presque jour pour jour en même temps que moi. Antonin fait pratiquement partie de l'équipe BiG et grâce à lui, j'ai presque appris à apprécier le hard rock et le métal. Je remercie Rossella Annunziata pour l'humour et la bonne humeur qui ont éclairé les longues soirées d'analyse des données de simulations. Je ne saurais oublier les conseils avisés que j'ai reçu de mes collègues Soizic Cheminant-Navarro, Thierry Delaveau, Médine Benchouaia (quels gâteaux !), Marianne Jaubert, Frédéric Devaux, Andrès Ritter, Ingrid Lafontaine, Juliana Bernardes, Angela Falciatore, ni les discussions amicales que j'ai eues avec eux. J'adresse une pensée amicale à Marie-Laure Jérier pour son aide avec l'administration et sa participation à mon club de lecture en anglais. *Last but not least*, je remercie sincèrement Alessandra Carbone dont l'énergie et les cours fascinants m'ont encouragé à frapper à la porte du LCQB ainsi que Martin Weigt pour m'avoir accepté dans le master BIM-BMC. Mon parcours aurait été très différent sans ces deux personnes. Je salue enfin l'ensemble de mes collègues du LCQB, ce laboratoire est une belle famille.

J'adresse une pensée à Jennifer Chaumont-Sturtevant et Cédric Brudermann pour m'avoir accordé leur confiance et un rôle d'animateur à l'Espace Langues de Sorbonne Université dans le cadre de ma mission doctorale. Jennifer et Cédric m'ont encouragé à développer mes propres activités en cours de conversation anglaise et à concrétiser mon projet de club de lecture en anglais. Ces moments ont été de vraies bouffées d'oxygène.

J'en viens à présent à remercier toutes les personnes qui m'ont entouré au quotidien tout au long de cette thèse et sans qui la finalisation de ce travail aurait été impossible : mes parents, mes sœurs Aude et Marie, Jean-François, Frédéric, Franck, Hannah, Ange et Julien. Les mots me manquent pour exprimer combien leur affection, leur confiance et leurs encouragements inconditionnels ont été importants pour moi. Nous formons une équipe imbattable.

Je remercie mes amis pour les bons souvenirs qui se sont accumulés depuis trois ans : Amira et Maria, le duo inséparable du Master BIM-BMC pour ces pique-niques improvisés en bord de Seine et au jardin des plantes après les journées de manips. Je remercie Paige pour toutes ces découvertes culinaires et linguistiques, Laure, Alice et Chloé pour ces voyages en Normandie et en Sardaigne. J'adresse une pensée spéciale au groupe des petits cochons : Jean-François, Antoine, Clémence, Clément, Claire, Edouard, Florent, Ophélie et Sandrine pour tous les bons moments que nous avons passé ensemble, et ils sont nombreux. Je n'oublie en aucun cas mes amies de prépa : Camille pour ces soirées d'amitié musicale, Anne-Marie pour ses magnifiques cartes postales. J'ai une pensée pour Leslie et ses messages toujours gentils, Clémence et mes amis d'enfance qui se reconnaitront. Je remercie enfin mes amis crétois Nikos et Maria, un certain nombre de calculs ont pu être réalisés grâce à votre wifi !

J'adresse une pensée sincère à Anne-Marie Hattenberger, Christine Virlogeux, Sabine Delannoy, Patrick Fach et Paolina Gautier pour leur amitié et leur soutien. Je salue enfin amicalement Mme David pour ses encouragements et nos discussions.

SOMMAIRE

INTRODUCTION 19

1. Modèle d'étude : les levures du subphylum *Saccharomycotina* 21

1.1. Bref historique de la génomique comparative des *Saccharomycotina* 21

1.2. Ordonner la diversité des *Saccharomycotina* 24

1.3. Des génomes eucaryotes particuliers 25

1.3.1. Un faible nombre d'introns 25

1.3.2. Des centromères ponctuels 27

1.3.3. L'alternance du type sexuel 27

1.3.4. Les altérations du code génétique 27

1.3.5. La perte de l'interférence à ARN 29

1.3.6. Caractéristiques génomiques des levures *Lachancea* et de *S. cerevisiae* 29

1.3.6.1. Le genre *Lachancea* 29

1.3.6.2. *Saccharomyces cerevisiae* 30

2. Instabilité et réparation du génome 31

2.1. Les causes de l'instabilité du génome 31

2.1.1. Facteurs exogènes 31

2.1.1.1. Agents physiques 31

2.1.1.2. Agents chimiques 31

2.1.2. Facteurs Endogènes 32

2.1.2.1. La réplication et la transcription 32

2.1.2.2. Les éléments transposables 35

2.1.3. La domestication des transposons et les cassures double-brin « programmées » 37

2.1.3.1. Les éléments transposables régulateurs 37

2.1.3.2. Les éléments transposables domestiqués 37

2.2. La réparation des lésions de l'ADN et la création de réarrangements chromosomiques . 41

2.2.1. La recombinaison homologue 41

2.2.1.1. Single-strand annealing (SSA) 42

2.2.1.2. Le modèle de double jonction de Holliday (dHJ) 42

2.2.1.3. Synthesis Dependent Strand Annealing (SDSA) 42
2.2.1.4. Break-Induced Replication (BIR) 43
2.2.1.5. Multi-invasion-induced rearrangement (MIR) 44
2.2.2. La réparation par « Non-Homologous End Joining » (NHEJ) 45
2.2.2.1. Canonical NHEJ (c-NHEJ) 45
2.2.2.2. Alternative NHEJ/Microhomology Mediated End Joining (MMEJ) 45
2.2.2.3. Microhomology/Microsatellite-Induced Replication (MMIR) et Microhomology-Mediated Break-Induced Replication (MMBIR) 46
2.2.3. Mécanismes de réparations et architecture des génomes 46
3. Impact évolutif des réarrangements chromosomiques 48
3.1. Les différents types de réarrangements chromosomiques 48
3.2. La reconstruction de génomes ancestraux 49
3.2.1. Les traces de l'origine commune des génomes 49
3.2.1.1. L'homologie 49
3.2.1.2. La synténie et la parcimonie 52
3.2.2. Des traces de l'origine commune des génomes au génome ancestral 53
3.2.2.1. Techniques reposant sur les réarrangements 53
3.2.2.2. Techniques reposant sur les adjacences 56
3.2.2.3. Techniques reposant sur les arbres de gènes 58
3.2.2.4. Les avantages apportés par l'algorithme *AnChro* 59
3.2.3. Des génomes ancestraux à la dynamique des génomes 61
3.3. Valeur sélective des réarrangements chromosomiques 64
3.3.1. Réarrangements chromosomiques, méiose, isolement reproductif 64
3.3.2. Réarrangements chromosomiques et évolution du répertoire de gènes 66
3.3.3. La relation génotype-phénotype : des SNP aux SV 68
3.4. L'ingénierie des génomes et le déchiffrage de l'impact des réarrangements chromosomiques 69
3.5. La révolution CRISPR/Cas9 73

RESULTATS 79
Problématique 81
4. Réarrangements chromosomiques et évolution des génomes chez les *Saccharomycotina* 83
4.1. Évaluation des performances d'*AnChro* 83
4.1.1. Obtention des blocs de synténie 83
4.1.1.1. Détection de la synténie avec *SynChro* 84
4.1.1.2. Détection de la synténie avec *i-ADHoRe* 85
4.1.2. Reconstruction des génomes ancestraux 90
4.1.2.1. Reconstruction de génomes ancestraux avec *ANGES* 90
4.1.2.2. Reconstruction de génomes ancestraux avec *GapAdj* 90
4.1.2.3. Reconstruction de génomes ancestraux avec *MGRA* 90
4.1.2.4. Reconstruction de génomes ancestraux avec *PMAG+* 90
4.1.2.5. Reconstruction de génomes ancestraux avec *AnChro* 90
4.1.3. Validation des reconstructions des génomes ancestraux des *Lachancea* 91
4.1.3.1. Pertinence biologique des reconstructions 91
4.1.3.2. Algorithmes de détection de la synténie et qualité des reconstructions 92
4.1.4. Evaluation de la pertinence des reconstructions à partir de génomes simulés 94
4.1.4.1. Intérêt des données simulées 94
4.1.4.2. Pertinence des données simulées 94
4.1.4.3. Résultats des reconstructions 95
4.2. Reconstruction des génomes ancestraux des *Saccharomycotina* 98
4.2.1. Provenance et filtrage des génomes 99
4.2.2. Construction de l'arbre phylogénétique des 66 espèces de *Saccharomycotina* 102
4.2.2.1. Identification des homologues synténiques 103
4.2.2.2. Amélioration du signal phylogénétique et construction de trois sous-arbres 106
4.2.2.3. Construction d'arbres phylogénétiques à partir d'autres signaux 112
4.2.3. Reconstruction des génomes ancestraux 117
4.2.3.1. Choix des génomes actuels 117
4.2.3.2. Choix des meilleures reconstructions 119

4.2.4. Inférence des réarrangements chromosomiques passés 124

5. Reshuffling yeast chromosomes with CRISPR/Cas9 130

5.1. Résumé en français 130

5.2. Abstract 131

5.3. Introduction 131

5.4. Material and Methods 133

5.4.1. Strains and media 133

5.4.2. Sulphite resistance spot tests 133

5.4.3. Growth curves in stress conditions 133

5.4.4. Spore viability 134

5.4.5. Identification of CRISPR/Cas9 target sequences 134

5.4.6. Construction of CRISPR/Cas9 plasmids with one or two guides 134

5.4.7. Yeast transformation 135

5.4.8. Plasmid Stability 135

5.4.9. Estimation of CRISPR/Cas9 cutting and repair efficiencies 135

5.4.10. Pulse Field Gel Electrophoresis and colony PCR for karyotyping rearranged strains 136

5.4.11. Southern blot validation of ECM34/*SSU1* translocation 136

5.4.12. Oxford Nanopore *de-novo* genome assembly 137

5.5. Results 137

5.5.1. Rationale for chromosome reshuffling 137

5.5.2. Engineering markerless, reversible reciprocal translocations at single base-pair resolution 138

5.5.3. Engineering a deletion at the translocation breakpoint 141

5.5.4. Recapitulating a natural translocation involved in sulphite resistance in wine strains 142

5.5.5. Reshuffling chromosomes with multiple rearrangements 143

5.5.6. Exploring the phenotypic diversity of reshuffled strains 147

5.6. Discussion 148

5.7. Supplementary Tables 150

5.8. Supplementary Figures 155

Funding 158

Acknowledgements 158

Data Availability 158

CONCLUSION ET PERSPECTIVES 159

REFERENCES BIBLIOGRAPHIQUES 169

ANNEXES 187

Article: Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus (Vakirlis et al., 2016) 187

INTRODUCTION

1. Modèle d'étude : les levures du subphylum *Saccharomycotina*

1.1. Bref historique de la génomique comparative des *Saccharomycotina*

On sait aujourd'hui que le mode de vie des levures (champignons unicellulaires ne formant pas de carpophore) est apparu plusieurs fois au cours de l'évolution, chez les Ascomycètes et les Basidiomycètes en faisant notamment intervenir la diversification du Zn-cluster, un groupe de gènes régulant la formation des filaments des champignons (Nagy et al., 2014). La plupart des espèces de levures décrites appartiennent au subphylum *Saccharomycotina* (ou levures « vraies » bourgeonnantes) qui constituent donc un groupe monophylétique (Kurtzman et al., 2011). Ces dernières sont présentes sur tous les continents, dans tous types d'environnements aquatiques et terrestres et englobent un intervalle de divergence génétique supérieur à celui de l'ensemble du phylum des chordés (Dujon, 2006).

La souche de référence S288C de *S. cerevisiae* a été le premier eucaryote à être complètement séquencé, il y a maintenant vingt ans (Goffeau et al., 1996). C'était alors un effort considérable, qui a mobilisé pendant plusieurs années près de 600 personnes de par le monde. L'étape logique après ce tournant historique était de « *s'atteler à un défi encore plus grand, celui d'élucider la fonction de tous les nouveaux gènes identifiés dans cette séquence […] Ceci ferait de S. cerevisiae l'eucaryote de référence pour l'étude des fonctions communes à tous les eucaryotes et renverserait la démarche traditionnelle de la génétique, créant une démarche où l'analyse du gène (ou de la séquence d'ADN) conduit à la compréhension de la fonction biologique plutôt qu'une démarche où un changement de fonction conduit à l'identification d'un gène* » (Goffeau et al., 1996). Dans ce contexte et pour beaucoup de scientifiques à l'époque, il n'était pas évident de savoir ce que la comparaison de génomes d'espèces appartenant au même clade pouvait apporter (Dujon and Louis, 2017). Les équipes de recherche se sont mobilisées sur la caractérisation expérimentale des gènes de la levure et les efforts de séquençage se sont portés en priorité sur les modèles expérimentaux les plus utilisés pour lesquels seule une infime fraction du génome était connue : la mouche du vinaigre *Drosophila Melanogaster* (Adams et al., 2000), le nématode *Caenorhabditis elegans* (C. elegans Sequencing Consortium, 1998), la souris *Mus musculus* (Mouse Genome Sequencing Consortium et al., 2002), et l'arabette des dames *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000).

Le programme Génolevures a marqué un tournant de la génomique en cherchant pour la première fois à décrypter les mécanismes d'évolution des génomes de levures par l'analyse comparative à large échelle de données de séquençage (Souciet et al., 2000). De la masse de données jusqu'alors inégalée générée par ce projet de séquençage partiel sont apparues les premières quantifications de la divergence des orthologues et de la synténie des génomes de *Saccharomycotina* (Llorente et al., 2000; Malpertuy et al., 2000; Souciet et al., 2000), aboutissant en 2004, presque dix ans après la publication du génome de *S. cerevisiae*, à la publication de sept génomes complets supplémentaires : *Candida albicans* (Jones et al., 2004), *Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, *Yarrowia lipolytica* (Dujon et al., 2004), *Eremothecium gossypii* (Dietrich et al., 2004) et *Lachancea waltii* (Kellis et al., 2004). La comparaison de génomes d'espèces proches a permis une succession de découvertes, notamment (i) l'identification de nouvelles séquences régulatrices comme dans l'étude (Kellis et al., 2003) où les auteurs ont développé une méthode pour identifier dans les génomes partiels d'espèces de *Saccharomyces*, des régions inhabituellement conservées, prédites comme étant des régions promotrices ou des séquences régulatrices intergéniques, dont près de 50% ont pu être validées par ChIP-seq et avec des données d'expression (ii) de confirmer les événements cruciaux de l'histoire évolutive des levures comme la duplication totale du génome, proposée pour la première fois il y a maintenant une vingtaine d'années à la suite du séquençage du génome de la levure (Wolfe and Shields, 1997) et validée par la suite par deux études indépendantes

dans lesquelles les auteurs comparent le génome de *S. cerevisiae* à celui de *Lachancea waltii* (Kellis et al., 2004) et d'*Eremothecium gossypii* (Dietrich et al., 2004). Ces deux études rapportent l'identification d'une « double synténie » avec *S. cerevisiae*, c'est-à-dire des régions génomiques dont toute la longueur est conservée en synténie avec deux régions distinctes de *S. cerevisiae*. Kellis et collègues identifient plus de 250 blocs de double synténie, contenant 75% des gènes de *L. waltii* et 81% des gènes de *S. cerevisiae*. De manière intéressante, les auteurs ont montré que dans 95% des cas où les deux ohnologues étaient conservés, l'un des deux montrait une évolution accélérée, supportant l'hypothèse d'une sub-fonctionnalisation ou néo-fonctionnalisation rapide à la suite d'un événement de duplication du génome. Dietrich et collègues montrent quant à eux que 90% du génome d'*E. gossypii* est conservé en double synténie avec celui de *S. cerevisiae*.

La génétique comparative a également permis de mieux comprendre la diversification génétique à l'origine de nouveaux traits phénotypiques, notamment l'acquisition de la pathogénicité. La comparaison du génome de la bactérie pathogène *Listeria monocytogenes* à son espèce sœur non pathogène *L. inocua* a été le premier exemple d'utilisation de la génomique comparative pour identifier des ilots de gènes de virulence (Glaser et al., 2001).

En tant que pathogènes opportunistes émergents, la levure *Candida glabrata* ainsi que d'autres *Saccharomycotina* ont reçu une attention croissante depuis les années 1980. L'émergence de ces pathogènes est en partie expliquée par les progrès de la médecine, augmentant l'espérance de vie des nouveau-nés prématurés et des personnes âgées, des personnes immunodéprimées ainsi que des patients suivant un traitement immunosuppresseur. En outre, l'utilisation accrue de cathéters, d'antibiotiques et de la chirurgie favorisent la colonisation des plaies par des levures pathogènes opportunistes (Hazen, 1995; Turner and Butler, 2014). Plusieurs espèces du genre *Candida* sont des pathogènes opportunistes. Le genre *Candida* est polyphylétique et avec les récents efforts apportés pour classifier les *Saccharomycotina* selon leurs données de séquence (voir la partie suivante) et la volonté de la communauté scientifique de faire correspondre « un nom à un champignon », cette dénomination sera fort probablement amenée à évoluer dans les prochaines années. Les espèces pathogènes font essentiellement partie du clade CTG (voir la partie suivante) et dans une moindre mesure, des *Nakaseomyces* (*Saccharomycetaceae*). Les pathogènes opportunistes sont dispersés dans le clade CTG : il y a d'une part *C. albicans* and *C. dublinensis,* dans un autre clade *C. tropicalis*, un clade supplémentaire contient *C. parapsilosis*, *C. orthopsilosis* et *C. metapsilosis*. On compte également *C. guilliermondii* et dans un autre clade *Clavispora lusitaniae* et *C. auris*. Dans le clade des *Nakazeomyces*, plus proche de *S. cerevisiae* se trouvent les pathogènes *C. glabrata, C. bracarensis* et *C. nivarensis*. Cette large dispersion des pathogènes dans l'arbre des *Saccharomycotina* montre que ce caractère est apparu plusieurs fois au cours de l'évolution.

Les bases génétiques à l'origine de la mise en place de la pathogénicité ne sont encore pas très bien comprises et les avancées dans le décryptage de ce phénomène sont très liées au nombre de génomes disponibles. Une des premières études visant à décrire les variations génomiques chez les *Candida* pathogènes date de 2009. Dans l'étude (Butler et al., 2009) les auteurs rapportent des variations de taille importantes entre les génomes des espèces du clade CTG (jusqu'à 50% de variation) pour un nombre de gène quasiment identique d'environ 6000 gènes. En outre, en analysant les familles de gènes conservées entre *C. tropicalis*, *C. parapsilosis*, *L. elongisporus*, *C. guilliermondii*, *C. lusitaniae*, et *C. albicans*, les auteurs ont mis en évidence pour la première fois 21 familles de gènes enrichies dans le génome des pathogènes, codant des adhésines, des lipases, des transporteurs d'oligopeptides ainsi que des facteurs de transcription. Ces gènes sont fréquemment dupliqués en tandem, avec parfois jusqu'à six gènes voisins de la même famille présentant une grande diversité de séquence. Une succession d'études ultérieures sont venues supporter ces résultats et montrer en outre un enrichissement en gènes impliqués dans la croissance sous formes de pseudo-hyphes et hyphes vraies chez les isolats pathogènes. *C. albicans* est à ce sujet

particulièrement polymorphe (Modrzewska and Kurnatowski, 2013; Priest and Lorenz, 2015; Sudbery, 2011; Thompson et al., 2011). Il a de plus été établi que *C. albicans* exploite le fait que son codon CUG soit traduit de manière « statistique » en leucine ou en sérine du fait de l'altération de son code génétique (voir la partie suivante) pour produire des protéines de surface plus variables ce qui constitue une des manières d'échapper au système immunitaire (Miranda et al., 2013).

Butler et collègues ont également montré que les deux souches de *C. albicans* comparées avaient des génomes très similaires et colinéaires. Toutefois le génome de la souche pathogène présente une très forte perte d'hétérozygotie. Des résultats comparables ont été obtenus à partir du séquençage d'autres isolats cliniques (Ford et al., 2011; Hirakawa et al., 2015). La perte d'hétérozygotie semble être un mécanisme important dans l'acquisition de la pathogénicité en permettant la perte de sensibilité rapide aux drogues antifongiques, notamment par l'accumulation de mutations ponctuelles dans les gènes codants des pompes de la membrane plasmique (Ford et al., 2011).

Chez les *Nakazeomyces*, les mécanismes d'acquisition de la virulence font également intervenir la duplication de gènes d'adhésion. Notamment, dans l'étude (Gabaldón et al., 2013), les auteurs montrent que parmi les trois espèces pathogènes étudiées, l'espèce la plus virulente, *C. glabrata* possède le plus grand nombre de gènes *EPA* (18 copies, dont 7 nouveaux variants dans cette étude) comparativement aux deux autres espèces moins pathogènes *C. bracarensis* et *C. nivariensis* (respectivement 12 et 9 copies). L'espèce non-pathogène *N. delphensis* porte quant à elle une seule copie du gène *EPA*. En résumé, les mécanismes d'acquisition de la pathogénicité font intervenir des mécanismes de réarrangements chromosomiques qui augmentent le nombre et la diversité des protéines de surface.

En outre, l'hybridation semble également jouer un rôle dans l'émergence du caractère pathogène en permettant de rassembler dans un individu des caractéristiques complémentaires, bien que cela n'ait pas été directement démontré. La détection des hybrides n'est souvent pas facile quand les génomes parentaux n'ont pas été séquencés. En outre, des questions demeurent vis-à-vis des *Saccharomycotina* pathogènes opportunistes. Les isolats pathogènes proviennent-ils naturellement de l'environnement et si oui quelle niche écologique occupent-ils ? Est-il possible de détecter des pathogènes potentiels, « pré-adaptés » à devenir un jour pathogènes par le biais des mécanismes décrits plus haut ? Ces questions soulignent l'importance de séquencer un grand nombre d'isolats et d'espèces de levures issus de milieux différents.

Depuis l'avènement des technologies de séquençage à haut débit ou « Next Generation Sequencing » (NGS), autour de 2005, une véritable explosion de données génomiques a vu le jour. A titre indicatif, la base de données GOLD du Joint Genome Institute (https://gold.jgi.doe.gov/) compte aujourd'hui 2926 génomes d'archées, 276 232 génomes de bactéries, 25 229 génomes d'eucaryotes et 8937 génomes de virus. Pourtant, sur les >1500 espèces de levures aujourd'hui décrites (Kurtzman et al., 2011), moins d'un dixième a été séquencé et les données écologiques de ces espèces sont pour la plupart inconnues. Un ambitieux projet en cours vise à séquencer au moins un représentant de chaque espèce décrite de *Saccharomycotina* (https://y1000plus.wei.wisc.edu/). La première partie de ces génomes (332 dont 220 nouvellement séquencés) a été publiée très récemment (Shen et al., 2018b). Ces génomes constitueront une ressource unique et leur analyse promet de belles découvertes. En outre, ces données remanieront très probablement l'arbre des *Saccharomycotina* tel que nous le connaissons aujourd'hui.

1.2. Ordonner la diversité des *Saccharomycotina*

Les *Saccharomycotina* constituent un ensemble de levures aux métabolismes extrêmement diversifiés, ce qui leur a permis de coloniser tous les continents et de nombreux écosystèmes aquatiques et terrestres (Hittinger et al., 2015). La classification des levures a été d'abord établie sur des caractères métaboliques mais avec la comparaison des données génomiques, il est devenu clair que ces critères font fréquemment l'objet d'homoplasie, de convergence et de parallélisme, ce qui rend difficile leur utilisation pour établir une classification.

Il existe une succession d'exemples décrivant le gain et la perte de fonctions métaboliques dans l'arbre des *Saccharomycotina.* Notamment, dans une étude de 2004, les auteurs montrent que la voie métabolique d'utilisation du galactose (GAL) a été perdue de manière parallèle dans quatre espèces de *Saccharomycotina* : *Saccharomyces kudriavzevii, Candida glabrata, Eremothecium gossypii* et *Kluyveromyces lactis.* En particulier, chez *S. kudriavzevii* les auteurs indiquent avoir identifié sept pseudogènes conservés en synténie et similaires aux sept gènes impliqués dans cette voie métabolique chez d'autres espèces, présentant ainsi un rare exemple d'une voie métabolique entière en cours de dégénérescence (Hittinger et al., 2004). Réciproquement, une étude de 2011 montre comment la voie GAL a été cette fois-ci acquise indépendamment chez *S. cerevisiae*, *Candida albicans* et *Schizosaccharomyces pombe* (Slot and Rokas, 2010). Une étude récente a également présenté l'acquisition d'un cluster de gènes GAL chez *Torulaspora delbrueckii* (Wolfe et al., 2015). Un second exemple est celui de la prototrophie de *S. cerevisiae* pour la biotine, disparue chez son ancêtre, qui été expliquée par une combinaison d'événements de transferts horizontaux de gènes provenant de gamma-protéobactéries et alpha-protéobactéries ainsi que d'événements de duplication suivis de néo-fonctionnalisation (Hall and Dietrich, 2007). Les *clusters* de gènes impliqués dans cette voie de biosynthèse sont localisés dans les subtélomères de *S. cerevisiae*, qui constituent des régions très dynamiques (Hall and Dietrich, 2007). Un autre exemple illustre l'acquisition de voies métaboliques chez *S. cerevisiae* : il a été démontré que certaines souches de vin de *S. cerevisiae* sont capables d'utiliser le xylose comme source de carbone alors que la grande majorité des souches naturelles en sont incapables (Wenger et al., 2010). Enfin, un dernier exemple est la perte indépendante de l'enzyme HO qui initie le changement de type sexuel chez plusieurs espèces de *Saccharomycotina* (Wolfe et al., 2015).

Pendant les deux dernières décennies, l'utilisation des données moléculaires comme les séquences d'ADN ribosomique (ADNr) et l'utilisation de données du génome complet ont permis de considérablement stabiliser la classification des *Saccharomycotina* ainsi que de nombreux autres organismes. Les données moléculaires permettent de classifier les espèces en fonction du degré d'identité de leur séquence nucléotidique ou de leurs séquences protéiques. Cette approche consiste à comparer des séquences conservées chez tous les êtres vivants qu'on souhaite classer, comme par exemple l'ADN qui code pour la petite sous unité du ribosome (16S chez les procaryotes, 18S chez les eucaryotes). La comparaison de séquences permet d'augmenter de manière très importante le nombre de caractères discriminants entre les espèces, comparativement aux critères morphologiques ou métaboliques, car chaque nucléotide ou acide aminé constitue en soi un caractère. Aujourd'hui, les génomes sont séquencés et assemblés avec beaucoup plus d'aisance qu'auparavant, ce qui fournit une profusion de données pour construire des phylogénies robustes (Philippe et al., 2011; Yang and Rannala, 2012). Les données moléculaires apportent beaucoup de détails et d'éclaircissements par rapport aux méthodes de classification traditionnelles. Par exemple, sur la base des données de séquences, une étude récente revoit la taxonomie de près de 60% des 94 759 génomes de la base de données « *Genome Taxonomy Database* » (http://gtdb.ecogenomic.org/) (Parks et al., 2018).

La classification actuelle des *Saccharomycotina* comprend 12 lignées majeures (voir la topologie dans (Shen

et al., 2016a)). Toutefois, pour présenter les caractéristiques de ces espèces on peut regrouper ces différentes lignées en quatre grands groupes, sur la base de leur « architecture » génomique commune (Dujon and Louis, 2017): (i) les *Saccharomycetaceae* qui comprennent entre autres l'espèce modèle *Saccharomyces cerevisiae*, (ii) les levures du « clade CTG » qui possèdent la particularité d'utiliser le codon CUG pour spécifier la sérine et non pas la leucine, (iii) les méthylotrophes et (iv) les lignées basales que nous décrirons comme un groupe bien qu'elles soient très hétérogènes et différentes des trois groupes précédents.

Les propriétés des génomes des *Saccharomycotina* sont synthétisées sur la Figure 1, page 26 et présentées en plus amples détails dans les paragraphes suivants.

1.3. Des génomes eucaryotes particuliers

1.3.1. Un faible nombre d'introns

Pour des génomes eucaryotes, les génomes des *Saccharomycotina* contiennent pour la plupart très peu d'introns spliceosomiques. En effet environ 5% des gènes en moyenne en contiennent, sauf dans les lignées basales qui en possèdent jusqu'à 30%. Les introns sont généralement localisés en 5' des gènes et sont délimités par des séquences très conservées (Bon et al., 2003; Neuvéglise et al., 2011). Pourtant les ancêtres des espèces de ce subphyllum étaient riches en introns (Stajich et al., 2007). Dans cette étude, les auteurs analysent la position des introns dans 1161 gènes orthologues conservés chez 25 espèces d'eucaryotes (plantes, vertébrés, basidiomycètes, hémi-ascomycètes, euascomycètes) et montrent que la teneur en introns de l'ancêtre était supérieure à celle observée dans tous les génomes étudiés. Par déduction, les auteurs démontrent l'importance du phénomène de perte des introns chez les champignons. Notons que cette pauvreté en introns concerne seulement les transcrits de la Pol II, les gènes d'ARN de transfert possédant des introns dans une proportion comparable à celle d'autres eucaryotes. L'origine de la perte des introns dans les génomes de levures bourgeonnantes n'est pas totalement comprise. Une explication partielle provient du fait qu'une majorité d'introns peut être délétée chez *S. cerevisiae* avec un impact phénotypique restreint ou nul, comme démontré par l'étude (Parenteau et al., 2008). Dans cette étude, les auteurs ont systématiquement délété un tiers des introns des 283 gènes de *S. cerevisiae* qui en possèdent. Ils montrent que seule la délétion des introns dans trois gènes associés au métabolisme de l'ARN (*MTR2, YRA1* et *TAD3*) a été accompagnée d'un défaut de croissance sévère car ces introns jouent un rôle de modulation de l'expression de ces gènes. Cet effet délétère a notamment pu être compensé par l'expression de ces gènes, dont les introns avaient également été délétés, sous le contrôle d'un promoteur hétérologue fort (*ACT1*). Etonnament, les auteurs montrent en outre que la délétion des introns de 15 gènes dont la fonction est associée au cytosquelette n'affecte aucunement la croissance de *S. cerevisiae*. Quoi qu'il en soit cette perte s'est amorcée très tôt au cours de l'évolution de ces espèces, probablement en lien avec la perte de l'interférence à ARN (voir plus bas) et s'est poursuivie tout au long de l'évolution des *Saccharomycotina*. Les mécanismes à l'origine de la perte des introns ont été beaucoup étudiés en lien avec les mécanismes de réparation du génome (voir le paragraphe 2.2.3, page 46).

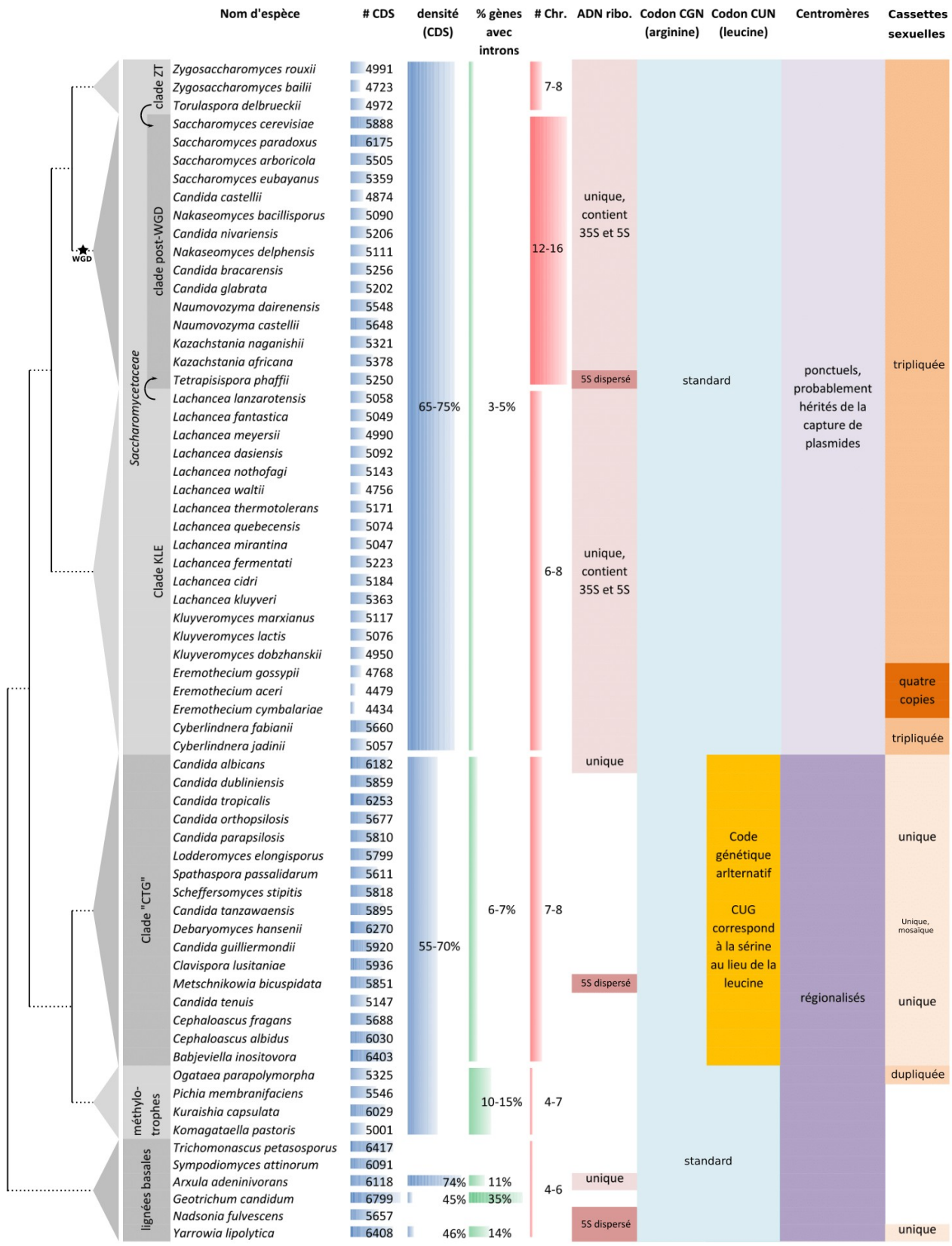


Figure 1 Caractéristiques principales des quatre groupes de *Saccharomycotina* tels que définis dans (Dujon and Louis, 2017). Les flèches courbes représentent l'hybridation à l'origine de l'événement de duplication totale du génome. Les nombres de CDS ont été extraits des génomes utilisés dans le cadre de cette thèse (voir la partie résultats).

1.3.2. Des centromères ponctuels

Par opposition aux centromères régionalisés d'environ 5kb observés chez les espèces du clade CTG, les méthylotrophes et les lignées basales des *Saccharomycotina*, qui sont des régions dans lesquelles s'accumulent les éléments transposables (pour revue (Dujon and Louis, 2017)), les *Saccharomycetaceae* présentent la particularité de posséder des centromères ponctuels. Historiquement, les centromères ponctuels ont été découverts par l'observation selon laquelle dans une collection de plasmides contenant des fragments du chromosome III de *S. cerevisiae*, un plasmide en particulier était transmis de manière très stable en mitose et en méiose. Les auteurs ont identifié dans ce plasmide la séquence centromérique *CEN3* et ont montré que les plasmides qui possèdent cette dernière ségrégaient comme des chromosomes ordinaires en mitose et en méiose (Clarke and Carbon, 1980). Depuis, on a appris que les centromères ponctuels sont constitués de trois séquences conservées dénommées CDEI, CDEII et CDEIII, disposées dans cet ordre et représentant environ 125pb au total. Il a été proposé que ces centromères aient pour origine la capture de séquences plasmidiques. En particulier, les plasmides réplicatifs 2µ de *S. cerevisiae* contiennent deux gènes *REP1* et *REP2* qui n'ont pas d'homologue dans d'autres organismes mais qui sont en revanche similaires aux gènes *Ndc10* et *Ctf13* qui codent deux protéines assurant l'interaction avec le kinétochore (ce qui forme le complexe CBF3) et le motif CDEIII avec lequel interagit l'histone CenH3 (Malik and Henikoff, 2009; Meraldi et al., 2006). Des centromères ponctuels constitués de séquences différentes des CDE de *S. cerevisiae* ont été découverts récemment chez *Naumovozyma castelli* et *Naumovozyma dairenensis* (Kobayashi et al., 2015). Chez ces deux espèces les centromères interagissent également avec le complexe CBF3.

1.3.3. L'alternance du type sexuel

L'aptitude à changer de type sexuel par conversion de l'idiomorphe de la cassette sexuelle par l'idiomorphe de l'autre type sexuel est une caractéristique qui différencie les levures des champignons multicellulaires qu'étaient leurs ancêtres. Les *Saccharomycetaceae* possèdent typiquement trois copies de leur cassette sexuelle dont une seule est active (*MAT)* et indique le type sexuel actif, les deux autres copies étant réprimées. Cette organisation est largement conservée chez les *Saccharomycetaceae* sauf chez les espèces du genre *Eremothecium* qui ont une quatrième copie de la cassette (Dietrich et al., 2013). Chez les levures du clade CTG, il n'y a par génome haploïde qu'une seule copie du locus et ces levures ne changent pas de type sexuel. Enfin, certaines espèces méthylotrophes possèdent deux copies de leur cassette sexuelle et sont capables de changer de type (Hanson et al., 2014; Maekawa and Kaneko, 2014) en intervertissant par le biais d'une inversion la copie active et la copie réprimée. Chez les *Saccharomycotina*, le remplacement de la cassette sexuelle est un processus faisant intervenir des cassures double-brin programmées de l'ADN. Nous présenterons les mécanismes de changement de type sexuel plus en détail dans le paragraphe 2.1.3, page 37.

1.3.4. Les altérations du code génétique

Chez une grande majorité d'organismes, la traduction des ARNm en protéines se fait selon le code génétique « universel » ou du moins standard, qui fait correspondre aux 64 codons possibles un des 20 acides aminés canoniques ou un codon stop. On a longtemps pensé qu'une quelconque modification du code génétique serait si délétère que ce dernier était considéré comme immuable, « gelé » (Crick, 1968). En réalité, les altérations du code génétique sont possibles, bien que rares. Elles affectent généralement les mitochondries qui possèdent leurs propres ribosomes, ARNt et un petit nombre de gènes (Ling et al., 2015). Les altérations du code génétique nucléaire sont encore plus rares et consistent dans la plupart des cas en un codon stop devenant un codon sens (Kollmar and Mühlhausen, 2017). La première altération consistant en un réassignement d'un codon sens à un autre acide aminé est celui identifié chez une levure apparentée à *Candida albicans* chez laquelle le codon CUG ne spécifie pas la leucine mais la sérine du fait de la présence

d'un ARNt additionnel propre (Kawaguchi et al., 1989). Cette altération a plus tard été identifiée comme une caractéristique partagée par les espèces proches de *Candida albicans*, désormais collectivement désignées sous le nom de « clade CTG » (Santos et al., 2011). Récemment, un autre exemple de réassignement a été publié, concernant également le codon CUG, cette fois-ci chez *Pachysolen tannophilus*, une espèce du clade méthylotrophe qui utilise ce codon pour spécifier l'alanine et non pas la leucine (Mühlhausen et al., 2016). Jusqu'alors, trois théories visaient à expliquer le réassignement du codon CUG. (i) La théorie de la « capture de codon » propose qu'un codon devient progressivement de plus en plus rare dans un génome, avant que l'ARNt qui assure sa traduction puisse être perdu et qu'un autre ARNt avec une boucle anticodon mutée assure alors sa traduction changeant ainsi sa signification. (Osawa and Jukes, 1989). Dans cette théorie, c'est la teneur en AT/GC qui serait à l'origine de la sélection de certains codons bien que les raisons n'en soient pas très claires. (ii) La théorie de l' « intermédiaire ambigu » propose qu'un ARNt muté puisse être chargé par une autre aminoacyl-ARNt synthétase que celle qui lui était préalablement assignée (Schultz and Yarus, 1994). Le décodage « statistique » du codon, c'est-à-dire soit avec l'ancien ARNt soit le nouveau, qui coexistent, conduit à la perte progressive de l'ARNt ancestral. (iii) La théorie de la « rationalisation du génome » (en anglais « *genome streamlining* ») propose que le changement d'un codon soit favorisé s'il permet de minimiser la machinerie traductionnelle (Andersson and Kurland, 1995). C'est ce modèle qui permet le mieux d'expliquer le réassignement dans le génome des mitochondries. Mühlhausen et collègues proposent un nouveau modèle, plus adapté aux réassignements du code génétique nucléaire, qu'ils ont nommé « réassignement par perte d'ARNt ». Dans ce modèle, un ARNt assigné à un codon est perdu ou muté et il est remplacé par un autre ARNt, issu de la duplication d'un ARNt possédant une boucle anticodon similaire assurant la traduction du codon avec une efficacité moindre par association wobble. Cette perte de fidélité serait accompagnée d'une diminution du nombre de codons en question dans le génome, comme observé chez les *Saccharomycotina* dont le code a changé, chez qui le codon CUG est relativement rare (Mühlhausen et al., 2016).

Dans une étude très récente (Krassowski et al., 2018), les auteurs ont utilisé les données de spectrométrie de masse de 18 espèces de *Saccharomycotina* pour investiguer l'existence d'autres exemples de réassignements. Les auteurs identifient un deuxième groupe d'espèces (baptisé Ser2) dans lequel, de manière indépendante aux espèces du clade CTG précédemment décrites (rebaptisé Ser1), c'est à nouveau le codon CUG qui est réassigné de leucine à sérine (la distinction n'est pas représentée dans la Figure 1, page 26 car cette dernière ne comprend aucune espèce de Ser2). Avec ce troisième exemple, il est apparu de manière claire aux auteurs que ces réassignements multiples du codon CUG chez les *Saccharomycotina* ne peuvent que difficilement être le fruit de coïncidences. Les auteurs proposent, de manière très convaincante que les réassignations observées chez les *Saccharomycotina* ont été provoquées par la contre sélection de l'ARNt ancestral CAG correspondant à la leucine. Cette hypothèse est notamment soutenue par le fait que l'ARNt ancestral a été perdu à plusieurs reprises au cours de l'évolution des *Saccharomycotina* : d'une part dans les trois lignées dont le codon a été réassigné, mais également dans les espèces actuelles chez lesquelles le codon CAG est toujours traduit en leucine ! En effet les auteurs montrent que les ARNt des espèces actuelles traduisant CUN en leucine portent les marques de la contre sélection (i) certains ont un intron très grand formant des structures secondaires conséquentes (ii) dans d'autres espèces, l'ARNt a été remplacé, comme chez *Lachancea thermotolerans*, probablement par transfert horizontal (iii) le gène de l'ARNt a été perdu sans remplacement, comme chez les *Saccharomyces* et la perte subséquente de la modification U_{34} de l'ARNt traduisant UAG en leucine a permis dès lors de lire les codons CUG et CUA. D'autre part, Krassowski et collègues avancent le fait que l'hypothèse du remplacement de l'ARNt défavorable est plus probable que l'hypothèse selon laquelle un nouvel ARNt est sélectionné sur la base des avantages protéomiques qu'il apporte, d'autant plus que les codons CUG sont généralement situés dans des régions non conservées des gènes. De manière très intéressante, Krassowski et collègues ont proposé qu'un plasmide « killer » également appelé « Virus-Like Element » (VLE) se trouve à l'origine de la contre sélection de l'ARNt ancestral. Les VLE sont des plasmides portant des gènes codant des ribonucléases ciblant

spécifiquement la boucle anticodon de certains ARNt. Cette hypothèse est fortement soutenue par le fait que plusieurs génomes parmi les espèces analysées contiennent des traces de séquences VLE pseudogénisées.

1.3.5. La perte de l'interférence à ARN

La machinerie d'interférence à ARN est absente chez la majorité des espèces du subphylum *Saccharomycotina* à l'exception de quelques espèces chez qui on trouve des gènes *Dicer* et *Argonaute* non-canoniques (Wolfe et al., 2015), indiquant que ce système a probablement été perdu très tôt au cours de l'évolution de ce groupe puis réacquis épisodiquement par la suite. L'ARN interférence a été découverte chez *Candida albicans*, *Naumovozyma castellii*, *Vanderwaltozyma polyspora,* mais pas chez *S. cerevisiae* (Drinnenberg et al., 2009; Wolfe et al., 2015). La perte de ce système a probablement joué un grand rôle dans l'évolution de ce clade en particulier en autorisant les cellules à héberger de longues molécules d'ARN double-brin, comme par exemple les particules « killer » qui ressemblent à des virus et dont certaines sont avantageuses. Il a notamment été démontré expérimentalement que l'ARN interférence et les particules killer ne peuvent pas coexister dans la même cellule (Drinnenberg et al., 2011).

1.3.6. Caractéristiques génomiques des levures *Lachancea* et de *S. cerevisiae*

Dans le cadre de cette thèse nous nous sommes intéressés à l'évolution des génomes des *Saccharomycotina* selon une approche *in-silico*. Cette dernière a été développée dans un premier temps à partir de l'étude des génomes des *Lachancea* puis généralisée aux *Saccharomycotina* dans leur ensemble. Nous nous sommes également intéressés à l'impact phénotypique des réarrangements du génome chez *S. cerevisiae* selon une approche *in-vivo*. C'est la raison pour laquelle nous présentons ces espèces en plus amples détails dans les deux paragraphes suivants.

1.3.6.1. Le genre *Lachancea*

Les *Lachancea* comprennent 12 génomes séquencés et annotés : *L. waltii* (Kellis et al., 2004), *L. kluyveri*, *L. thermotolerans* (Souciet et al., 2009), *L. fantastica, L. meyersii, L. dasiensis, L. nothofagi, L. mirantina, L. fermentati, L. cidri* (Vakirlis et al., 2016), *L. quebecensis* (Freel et al., 2016)*, L. lanzarotensis* (Sarilar et al., 2015). De ce fait, le genre *Lachancea* est le plus densément échantillonné de la famille des *Saccharomycetaceae*. Les génomes des *Lachancea* contiennent 8 chromosomes sauf *Lachancea fantastica* qui n'en a que 7 en raison d'une fusion de chromosomes de télomère à télomère, et autour de 5200 gènes codant pour des protéines (Vakirlis et al., 2016). Comme les autres *Saccharomycetaceae*, les *Lachancea* possèdent des centromères ponctuels, des génomes compacts codants à hauteur de 67% à 77%. Les génomes des *Lachancea* constituent un très bon échantillon pour étudier l'évolution des génomes car leur niveau de divergence protéique et le nombre de blocs de synténie observés entre ces génomes restent sous le seuil à partir duquel on ne peut plus reconstruire les génomes ancestraux de manière fiable car le signal phylogénétique est perdu (Drillon and Fischer, 2011). Les *Lachancea* constituent donc un groupe « modèle » pour l'étude de l'évolution des génomes comme nous le verrons dans la suite de cette thèse.

L. kluyveri présente une curieuse caractéristique génomique. Une région de 1 Mb, couvrant presque la totalité de la longueur du bras gauche du chromosome C (le « C-left ») de cette espèce présente un taux de GC de 53%, contrastant clairement avec la proportion de 40% observée dans le reste du génome (Payen et al., 2009). Cet enrichissement touche aussi bien les séquences non-codantes que les séquences codantes et conduit à un fort biais d'usage des codons dans les protéines codées dans le C-left. Les auteurs montrent que la synténie de cette région est bien conservée avec d'autres *Saccharomycetaceae* et montrent en outre que cette région provient probablement d'une autre espèce de *Lachancea*. De manière singulière, ce bras chromosomique est totalement dénué d'éléments transposables alors que le reste du génome en contient

en moyenne 18 par mégabase. Le profil de réplication du génome complet de *L. kluyveri* a été établi et il apparaît clairement que le *C-left* a une dynamique de réplication différente du reste du génome. Chez *L. kluyveri*, le génome est constitué d'une alternance de régions répliquées de manière précoce et de régions répliquées de manière tardive, sauf dans le *C-left* qui contient des séquences consensus d'origine de réplication différentes et qui est répliqué entièrement de manière précoce (Agier et al., 2013). Plus récemment, le séquençage de 28 souches de *L. kluyveri* a permis de montrer que cette portion de chromosome est probablement le résultat d'un événement d'introgression dans le génome de l'ancêtre de tous les isolats de *L. kluyveri*, bien que l'espèce donneuse de ce fragment de chromosome demeure aujourd'hui inconnue (Friedrich et al., 2015). En outre, l'analyse du génome de cette espèce à l'échelle de la population a permis de montrer que le C-left a un taux de recombinaison plus élevé que le reste du génome et que cette région subit un très fort taux de substitutions des nucléotides AT en GC. Ces résultats montrent que des régions subissant des régimes de recombinaison et de substitution très différents peuvent coexister au sein d'un même génome.

1.3.6.2. *Saccharomyces cerevisiae*

Le génome de *Saccharomyces cerevisiae* comprend 5771 cadres ouverts de lecture ou « *open reading frames* » (ORF). De manière surprenante pour cet organisme si étudié, près de 750 ORF de *Saccharomyces cerevisiae* n'ont pas encore été validés par des données protéomiques. Le génome de *S. cerevisiae*, comme les autres génomes des *Saccharomycotina* est environ 50 fois plus dense que le génome humain avec des gènes codants tous les 2kb environ, dont 4% seulement contiennent un ou plusieurs introns (Goffeau et al., 1996; Spingola et al., 1999). Le génome de *S. cerevisiae* comprend 16 chromosomes, dont il a été démontré qu'ils proviennent d'un événement de doublement du génome (Dietrich et al., 2004; Kellis et al., 2004; Wolfe and Shields, 1997). La réplication du génome de *S. cerevisiae* fait intervenir des origines de réplication localisées tous les 40 kpb environ sur chacun des chromosomes. Les télomères de *S. cerevisiae* sont constitués d'environ 300 pb du motif $(TG)_{1\text{-}3}$. L'ADNr, est regroupé au sein d'un locus unique, porté par le chromosome XII dont la taille est très variable. *S. cerevisiae* possède trois copies de la cassette indiquant le type sexuel dont une seule est active.

Le génome de *S. cerevisiae* contient différentes familles de rétrotransposons, appelés Ty1 à Ty5 qui sont dispersés sur les 16 chromosomes (pour revue (Lesage and Todeschini, 2005)). Ils se multiplient selon un cycle évoquant celui des rétrovirus mais n'ont pas de phase infectieuse. Tous les Ty appartiennent au groupe des *Pseudoviridae* sauf les Ty3 qui sont des *Metaviridae* (on parle aussi de « *gypsy-like elements* »). Tous les Ty sont flanqués de séquences répétées directes qu'on appelle « *Long Terminal Repeats* » (LTR). La souche de référence de *S. cerevisiae*, S288C, comprend 331 insertions impliquant des séquences de Ty et représentant environ 3% du génome. La plupart de ces insertions (85%) consistent en des séquences LTR isolées. Les Ty1 et Ty2 sont les familles les plus abondantes avec respectivement 32 et 13 éléments complets (« full-size ») et respectivement 185 et 21 LTR isolés. Les éléments Ty3, Ty4 et Ty5 sont moins abondants avec respectivement 2, 3 et 1 copies complètes et 37, 29 et 7 LTR isolés. Les Ty1, Ty2 et Ty3 sont toujours actifs dans le génome de *S. cerevisiae* ; la transcription des Ty4 produit des transcrits tronqués ne possédant pas l'extrémité nécessaire à l'amorçage de leur rétro-transcription et donc à leur multiplication, quant aux Ty5, ils sont très divergés et aujourd'hui éteints (pour revue (Lesage and Todeschini, 2005; Sandmeyer et al., 2002)).

Le génome de *S. cerevisiae* contient beaucoup de gènes conservés avec les vertébrés, notamment, près de 50% des gènes essentiels de la levure ont conservé leur fonction ancestrale et peuvent être remplacés de manière viable par leur orthologue humain (Kachroo et al., 2015). Cette propriété a permis des avancées très importantes dans l'étude des mécanismes cellulaires eucaryotes notamment le contrôle du cycle cellulaire et le transport des vésicules. Ainsi la levure a été utilisée comme outil de production de molécule

d'intérêt médical (entre autres l'insuline). La levure *S. cerevisiae* a également permis d'établir le mécanisme des chaperonnes lors du repli des protéines. Cette problématique est intimement liée à l'origine de pathologies humaines comme la maladie d'Alzheimer.

2. Instabilité et réparation du génome

Le génome contient toute l'information dont les cellules ont besoin pour se développer, survivre et se reproduire et il doit être transmis à la descendance de la manière la plus fiable possible. Cependant, tout au long de la vie cellulaire, le génome subit des lésions d'origine exogène ou endogène : altérations des bases azotées, cross-links entre nucléotides, brins d'ADN ou entre ADN et protéines. Ces altérations peuvent, en combinaison avec d'autres mécanismes biologiques, induire des cassures simple-brin (CSB) ou double-brin (CDB) dans l'ADN. Nous nous intéresserons particulièrement à ces lésions, car ce sont elles qui favorisent l'instabilité du génome et l'apparition de réarrangements chromosomiques.

2.1. Les causes de l'instabilité du génome

2.1.1. Facteurs exogènes

2.1.1.1. Agents physiques

De nombreux facteurs physiques sont capables d'introduire des cassures dans l'ADN. Notamment, les radiations ionisantes, en plus de créer d'importantes altérations des bases azotées, mènent à l'accumulation de radicaux libres qui introduisent des cassures simple-brin dans l'armature glucide-phosphate de l'ADN (Thompson, 2012; Ward, 1994). A haute fréquence, ces cassures simple-brin ou « *nicks* » peuvent avoir lieu sur les deux brins de l'ADN dans le même tour d'hélice ce qui aboutit à une cassure double brin (Milligan et al., 1995). Les radiation induisent environ dix fois plus de CSB que de CDB (Ma et al., 2012). Les extrémités de la molécule d'ADN cassée par les radiations ionisantes sont altérées chimiquement et ne peuvent pas être réparées par simple ligation des extrémités 5' et 3' à la manière des coupures induites par une endonucléase (Weinfeld and Soderlind, 1991).

2.1.1.2. Agents chimiques

De nombreux agents chimiques sont connus pour favoriser la création de cassures simple-brin (CSB) ou double-brin (CDB) notamment en fixant de manière covalente les topo-isomérases à l'ADN (topotecan, camptothecine, etoposide), favorisant ainsi la collision avec la machinerie de réplication (Koster et al., 2007). D'autres produits ralentissent la progression de la fourche de réplication en épuisant les ressources de déoxyribonucléotides disponibles à l'ADN polymérase (hydroxyurée et l'aphidicoline). Certains agents favorisant les lésions de l'ADN sont utilisés dans la lutte contre le cancer. Par exemple, les traitements chimio-thérapeutiques comprennent des agents alkylants de l'ADN (méthanosulfonate de méthyle et temozolomide), des agents favorisant les cross-links de l'ADN (mitomycine C et cisplatine) et des agents mimant l'effet de radiations ionisantes (bléomycine, phléomycine) (Chen and Stubbe, 2005; Wyrobek et al., 2005).

2.1.2. Facteurs Endogènes

2.1.2.1. La réplication et la transcription

Le fait que la majorité des DSB apparaissent de manière spontanée au cours de la réplication (Syeda et al., 2014) pourrait suggérer que l'ADN est cassé avant le passage de la fourche de réplication. On estime par exemple que l'action oxydative des espèces réactives de l'oxygène et de l'azote, en altérant les bases azotées (sites abasiques, bases oxydées, cross-links inter ou intra-brin) contribue à former près de 10^4 CSB par cellule et par jour chez l'homme (Hoeijmakers, 2009). Quand une fourche de réplication rencontre une CSB, la chromatide synthétisée à partir du brin cassé se détache. Il y a donc au final une chromatide cassée à réparer (voir le paragraphe 2.2, page 41) et une chromatide intacte.

En réalité, la réplication elle-même génère un grand nombre de cassures dans l'ADN. En plus de dissocier les deux brins d'ADN, ce qui rend ces derniers plus vulnérables aux cassures du fait de facteurs exogènes (voir plus haut) et de facteurs endogènes (nucléases, espèces réactives), les fourches de réplication peuvent être fréquemment bloquées par un obstacle comme par exemple la structure de la chromatine (Torres et al., 2004), les protéines interagissant avec l'ADN (Merrikh et al., 2012; Mirkin and Mirkin, 2007), une collision avec la machinerie de transcription (voir ci-dessous). Les fourches bloquées peuvent régresser, entrainant avec elles les deux brins néo synthétisés qui se dissocient des deux brins matrice pour s'hybrider ensemble. La structure qui en résulte, en « pied de poule » (« *chicken foot* ») est une jonction de Holliday qui est clivée par des nucléases structure-sélectives ou « HJ résolvases » comme par exemple Mus81-Mms4 ou Yen1, ce qui crée également une chromatide intacte et une chromatide cassée (Wyatt and West, 2014).

Le lien entre réplication, transcription et stabilité du génome a été très étudié afin d'expliquer pourquoi certaines régions du génome sont particulièrement fragiles. Les « sites fragiles communs » (« *Common fragile sites* », CFS) ont été identifiés dès les années 1970 en suivant l'apparition de CDB dans le génome de cellules dont les fourches de réplications ont été ralenties par l'utilisation de drogues comme l'hydroxyurée ou l'aphidicoline. Les CFS sont présents chez *S. cerevisiae* tout comme chez *M. musculus* ou dans le génome humain chez qui ils constituent des « points chauds » de translocations fréquemment observées dans les lignées de cellules cancéreuses (pour revue (Sarni and Kerem, 2016)). On a longtemps pensé que les CFS étaient constitués de séquences pouvant former des structures secondaires qui gênent la progression de la fourche de réplication. Notamment, une étude explique la fréquence de cassure des CFS à hauteur de 45% à 77%, selon le modèle statistique utilisé, par des caractéristiques locales du génome comme flexibilité de la région, la distance par rapport au centromère, ou la teneur en répétitions ALU (Fungtammasan et al., 2012). Suivant ce courant de pensée, les auteurs d'une autre étude ont cloné la séquence du site fragile FRA16D du génome humain, impliqué dans des translocations à l'origine de 25% des myélomes, dans un chromosome artificiel de levure (« *yeast artificial chromosome* », YAC) dont la cassure peut être sélectionnée positivement. Les auteurs mettent ainsi en évidence une séquence riche en AT (le site Flex1) dont la structure secondaire prédite est cruciforme (Zhang and Freudenreich, 2007), mais une autre étude contredit ces explications en montrant que la délétion de la séquence formant des structures secondaires n'empêche pas l'apparition des cassures (Finnis et al., 2005). Il a de plus été montré par la suite que les CFS ont plutôt une origine épigénétique étant donné qu'ils ne sont pas les mêmes dans tous les types cellulaires (Le Tallec et al., 2013). Cette observation est cohérente avec le fait que le profil de réplication varie selon le type cellulaire et le fait que ces régions sont pauvres en origines de réplication et donc répliquées tardivement, ce qui génère des contraintes dans la molécule d'ADN lorsque les fourches de réplications convergent (Letessier et al., 2011).

Il existe d'autres sites sensibles, les « *early replicating fragile sites* » (ERFS), identifiés initialement par ChIP-seq des fragments d'ADN interagissant avec RPA, γH2AX et BRCA1, trois protéines impliquées dans la réparation des cassures de l'ADN, après le traitement des cellules à l'hydroxyurée. Par opposition aux CFS, les ERFS sont répliqués de manière précoce et sont fortement transcrits. En revanche, les ERFS tout comme les CFS sont variables selon le type cellulaire considéré. Dans ces régions, c'est le niveau de transcription plutôt que le timing de réplication qui favorise leur instabilité (Barlow et al., 2013).

Le mécanisme moléculaire à l'origine de la fragilité des ERFS et CFS n'est pas totalement compris (pour revue (Aguilera and García-Muse, 2012; Marnef et al., 2017)). En ce qui concerne les CFS, il a été montré que ces régions codent le plus souvent de longs gènes (>300kb), ce qui favorise la collision entre les machines de réplication et de transcription (Helmrich et al., 2011) bien que la transcription des longs gènes ne permet pas toujours d'expliquer la fragilité des CFS (Le Tallec et al., 2013). Une autre explication serait l'entrée précoce de la cellule en mitose alors que ces sites sont toujours en cours de réplication, ce qui est connu pour favoriser l'apparition de CDB. La résolution des intermédiaires réplicatifs dans les chromosomes en mitose font intervenir les nucléases structure-spécifiques ERCC1 et MUS81-EME1 (Naim et al., 2013). Le mécanisme expliquant que ces régions en cours de réplication échappent aux nombreuses voies de signalisation permettant de rendre la réplication et la mitose mutuellement exclusive (pour revue récente (Zeman and Cimprich, 2014)) n'est en revanche pas encore bien compris (Le Tallec et al., 2014). En ce qui concerne les ERFS, la réplication n'est probablement pas en cause étant donné que ces régions sont répliquées de manière précoce ce qui leur laisse plus de temps avant l'entrée de la cellule en mitose.

Des travaux récents proposent un mécanisme alternatif pour expliquer la fragilité des CFS et des ERFS, faisant intervenir les R-loops et les G-quadruplex. Les R-loops sont des structures à trois brins formées lorsqu'un ARN transcrit par une ARN polymérases s'hybride au brin matrice de l'ADN, laissant le brin codant à l'état simple brin. Le mécanisme de formation des R-loops le plus généralement admis est le modèle « thread-back » dans lequel l'ARN néo-synthétisé envahit le duplex d'ADN dès sa sortie de l'ARN polymérase. Ce modèle est soutenu par la structure cristallographique de l'ARN polymérase qui montre que l'ARN et l'ADN sortent par des canaux différents (Westover et al., 2004). Le séquençage d'hybrides ADN:ARN a permis de montrer chez les cellules de mammifères que les R-loops s'accumulent dans les promoteurs de gènes fortement transcrits, enrichis en C en raison de la grande stabilité des hybrides formés par les brins d'ADN riches en C avec les brins d'ARN riches en G (Ginno et al., 2013; Ratmeyer et al., 1994; Sanz et al., 2016). Les mécanismes de contrôle des R-loops, conservés au cours de l'évolution, font intervenir les RNase H chez les procaryotes et les RNases H1 chez les eucaryotes. Ces enzymes dégradent spécifiquement l'ARN de l'hybride ADN:ARN ce qui permet de minimiser l'impact négatif des R-loops sur la stabilité du génome que nous détaillerons un peu plus loin (Helmrich et al., 2011; Lin et al., 2010; Wahba et al., 2011). Les G-quadruplex quant à eux sont des structures non-canoniques à quatre brins qui se forment essentiellement dans les promoteurs de gènes fortement transcrits, riches en G et C. Ces caractéristiques partagées laissent penser que la séparation des deux brins contribuerait à former des R-loops ainsi que des G-quadruplex de manière conjointe. Un grand nombre d'études montre que l'accumulation de R-loops/G-quadruplex favorise l'apparition de cassures double-brin, favorisant l'apparition de mutations, d'hyper-recombinaisons et de réarrangements chromosomiques (pour revue (Santos-Pereira and Aguilera, 2015)).

Notamment, trois mécanismes expliquent l'apparition de DSB au niveau des R-loops/G-quadruplex (pour revue (Sollier and Cimprich, 2015)):

i) Premièrement, l'ADN simple-brin formé par une R-loop/un G-quadruplex, pouvant mesurer plusieurs centaines de nucléotides, est particulièrement sensible aux nucléases qui s'attaquent spécifiquement aux fragments d'ADN simple brin. Notamment, l'enzyme AID de la famille APOBEC peut introduire des modifications dans les deux brins d'ADN de la R-loop en désaminant la désoxy-cytidine en désoxy-uracile, ce qui peut induire des CSB. Ces modifications de l'ADN sont par ailleurs reconnues par l'enzyme uracile-ADN glycosylase qui initie la réparation de l'ADN par excision de base, générant des CSB. Quand AID introduit des modifications dans les deux brins de la R-loop, ce processus peut induire des CDB.

ii) Chez l'humain, l'hybride ADN:ARN formant une R-loop peut être processé par deux « *Flap endonucleases* ». Les deux enzymes XPF et XPG (RAD1 et RAD2 chez la levure), excisent dans un premier temps la séquence hybride de la R-loop. La CSB qui en résulte peut être convertie en CDB soit par des endonucléases structure-spécifiques qui clivent l'ADN au niveau de la jonction simple/double brin, soit par une fourche de réplication qui perd un bras lorsqu'elle arrive au niveau de la CSB.

iii) La collision entre une R-loop/un G-quadruplex et la machinerie de réplication peuvent également générer des DSB (Tuduri et al., 2009; Wahba et al., 2011; Wellinger et al., 2006). Les R-loops peuvent constituer un gros obstacle au passage des fourches de réplications car elles sont relativement stables et sont associées à la machinerie transcriptionnelle ainsi qu'aux enzymes associées à l'ARN (Aguilera and Gaillard, 2014; Prado and Aguilera, 2005). Par conséquent, les R-loops peuvent causer des CDB sans la participation de XPF et XPG.

Les cellules peuvent éliminer les R-loops selon différents mécanismes. Comme mentionné plus haut, les R-loops peuvent être processées par des flap endonucléases. L'ARN de l'hybride ADN:ARN peut également être dégradé part les ribonucléases H. Alternativement, les R-loops peuvent être éliminées par des hélicases qui déroulent les hybrides ADN:ARN. C'est notamment le cas de l'hélicase Pif1 chez la levure. De manière intéressante, la forte affinité de Pif1 pour les d'hybrides ADN:ARN (Boulé and Zakian, 2007) lui confère un rôle régulateur dans de nombreux processus cellulaires faisant intervenir ou favorisant la création d'hybrides ADN :ARN. Notamment, Pif1 inhibe l'allongement des télomères médiés par la télomérase et empêche la formation de télomères au niveau des CDB, régule la réplication de l'ADN ribosomique et contribue à la maturation des fragments d'Okazaki (Boulé and Zakian, 2006; Mendoza et al., 2016).

Récemment, d'intéressantes observations ont été faites d'une part par rapport au rôle bénéfique des R-loops en tant que régulateur et terminateur de la transcription ou dans leur rôle potentiel dans la réparation des CDB (Costantino and Koshland, 2015; Keskin et al., 2014; Santos-Pereira and Aguilera, 2015) . D'autre part, le rôle des introns dans le contrôle de la formation des R-loops a été étudié récemment. Bonnet et collègues se sont intéressés à la formation des R-loops au niveau des gènes fortement transcrits chez la levure (>50 ARNm/h) et rapportent que, bien que la propension à former des R-loops augmente avec le niveau de transcription des gènes, le nombre d'hybrides ADN:ARN est fortement réduit parmi les gènes possédant des introns comparativement aux gènes sans introns (Bonnet et al., 2017). Les auteurs ont ensuite recherché parmi les gènes fortement transcrits, deux gènes identiques présentant naturellement un intron ou non et ont identifié *RPP1A* et *RPP1B* issus de la duplication totale du génome. De manière frappante, bien qu'étant plus transcrit que son homologue sans intron, *RPP1B* forme moins de R-loops. Les auteurs démontrent également que l'ajout artificiel d'un intron dans la région 5' de gènes formant fréquemment des R-loops limite l'apparition de celles-ci et rapportent en outre que c'est le recrutement du spliceosome et non l'épissage en tant que tel qui empêche la formation des R-loops en s'associant à l'ARN

en cours d'élongation. Poursuivant leur investigation par des analyses génomiques, Bonnet et collègues comparent le génome de différentes espèces contenant des proportions variables de gènes avec introns et indiquent que les génomes les plus riches en introns sont les mieux protégés de l'apparition des R-loops. Ces résultats suggèrent donc une explication à la conservation des introns chez les eucaryotes.

Par ailleurs Il faut compter, en plus des sources de cassures mentionnées plus haut, des mécanismes qui génèrent des CSB et des CDB de manière programmée dans les génomes à différents moments du cycle cellulaire. Un exemple emblématique est l'activité de l'endonucléase Spo11 (conservée chez les champignons, les invertébrés, les mammifères et les plantes) qui induit des CDB au niveau de centaines de sites non-aléatoires du génome (Keeney, 2008). C'est l'activité de Spo11 qui initie la recombinaison à l'origine de crossing-overs entre chromosomes homologues, cruciaux pour la ségrégation correcte des chromosomes lors de la méiose (Cole et al., 2010; Lam and Keeney, 2015a). Paradoxalement, Spo11 seule n'est pas capable d'induire des CDB *in-vivo* et son activité endonucléase est assistée par d'autres protéines qui sont très variables selon le phylum, le clade ou même l'espèce considérée (Cole et al., 2010; Keeney, 2008; Lam and Keeney, 2015b). De nombreux autres mécanismes, apparentés à l'activité des éléments transposables du génome, génèrent des CDB programmées.

2.1.2.2. Les éléments transposables

Les éléments transposables du génome ont, comme nous allons le voir, un fort potentiel pour générer des CDB et induire des réarrangements chromosomiques. Les éléments transposables composent environs 12% du génome de *C. elegans* (C. elegans Sequencing Consortium, 1998), 37% du génome de la souris (Mouse Genome Sequencing Consortium et al., 2002), 45% du génome humain (International Human Genome Sequencing Consortium, 2001), 3% du génome de la levure *S. cerevisiae* (Carr et al., 2012) et plus de 80% du génome du maïs (SanMiguel et al., 1996) chez qui ils ont été découvert en 1950 par Barbara McClintock. Les éléments transposables comprennent les éléments de classe I, qui se déplacent dans le génome sous une forme intermédiaire à ARN, et les éléments de classe II qui se déplacent sous la forme d'ADN simple ou double-brin.

L'instabilité du génome causée par les éléments transposables provient en grande partie du fait qu'ils peuvent induire eux-mêmes des CDB, et qu'ils fournissent un grand nombre de séquences répétées pouvant servir de sites de recombinaison ectopiques lors de la réparation des CDB du génome (voir la partie 2.2, page 41). Les éléments transposables partagent de grandes ressemblances fonctionnelles avec les rétrovirus (Miyake et al., 1987; Mossie et al., 1985; Varmus, 1988; Xiong and Eickbush, 1988). En raison de leur faculté à se multiplier et parce qu'on pensait auparavant que les transposons ne jouaient aucun rôle dans le génome, les éléments transposables ont longtemps été relégués au rang d' « ADN poubelle » (« *junk DNA* ») ou de « parasites génomiques » (Orgel and Crick, 1980).

Citons à titre d'exemple les *Long interspersed elements-1* (L1) qui sont des éléments de classe I (formant un intermédiaire à ARN) d'environ 6kb du génome humain. Ces rétrotransposons autonomes représentent environ 15% du génome. Une centaine de copies sont toujours actives, les autres étant tronquées (Sassaman et al., 1997). Les copies actives contribuent à la mobilité des copies inactives de L1 ainsi que d'autres rétrotransposons comme les Alu et les SINE grâce à leur endonucléase et à leur réverse-transcriptase. L'endonucléase, dénommée L1 EN, dont l'activité a été démontrée expérimentalement pour la première fois en suivant l'augmentation du nombre de foci H2AX, fortement corrélée à l'induction de CDB (Modesti and Kanaar, 2001) à la suite de la transfection de cellules HeLa par un plasmide portant une construction capable de sur-exprimer la L1 EN, ressemble à une endonucléase AP (Gasior et al., 2006), c'est-à-dire une endonucléase impliquée dans le remplacement des nucléotides abasiques (apyrimidique/apurique). La L1 EN induit des DSB dont une des extrémités 3' sortantes riche en T

peut s'hybrider avec la queue poly-A de l'ARN d'un L1 transcrit. L'extrémité 3' sortante de l'ADN sert alors d'amorce à la rétro-transcription de l'ARN du L1 qui s'intègre dans le génome (Ariumi, 2016), flanqué de deux copies du site de reconnaissance de l'endonucléase ce qui permet d'augmenter le nombre de sites de coupures potentiels de l'endonucléase (Viollet et al., 2014). Ces résultats suggèrent l'énorme potentiel de déstabilisation du génome que représentent les éléments transposables à ARN. L'étude (Gilbert et al., 2002) rapporte notamment des délétions allant jusqu'à 71kb médiées par l'intégration d'un L1 artificiel dont la rétro-transposition peut être sélectionnée positivement. La même année, une autre étude a rapporté, avec une méthode similaire, des délétions ainsi qu'une inversion de 120kb probablement induite par deux coupures introduites par la L1 EN au niveau de deux sites distants (Symer et al., 2002).

Les éléments transposables de classe II (formant un intermédiaire à ADN) sont également capables d'induire des CDB et des réarrangements chromosomiques. Chez les eucaryotes, on distingue dans cette seconde classe d'éléments transposables, trois catégories d'éléments mobiles : (i) les transposons qui s'excisent du génome qui les porte en induisant des CDB pour se réinsérer ailleurs dans le génome selon la stratégie du couper-coller (2002), (ii) les hélitrons qui sont multipliés par un processus similaire à la réplication circulaire de l'ADN (Thomas and Pritham, 2015) et (iii) les Mavericks, qui sont les plus gros éléments mobiles connus (15 à 20 kb) et qui encodent notamment deux gènes leur conférant la propriété de se répliquer et de s'intégrer par eux-mêmes : une ADN polymérase et une intégrase (Kapitonov and Jurka, 2006).

De nombreuses études ont rapporté l'implication de transposons à ADN dans l'induction de réarrangements chromosomiques chez les eucaryotes. Un des premiers exemples provient des éléments Foldback (FB) décrits pour la première fois chez la drosophile (Potter, 1982). Les FB sont de taille variable, allant de quelques centaines à quelques milliers de paires de bases. Ils sont flanqués de longues séquences répétées, qui constituent parfois la quasi-totalité de la longueur de l'élément. Il a été montré que ces éléments sont capables d'induire de grandes inversions dans le génome de *Drosophila buzzatii*. Les différentes configurations de ces inversions prédominent dans des zones géographiques définies, suggérant selon la niche écologique habitée, un avantage sélectif différent. L'étude (Betrán et al., 1998) donne un exemple d'inversions altérant de manière significative la corrélation entre taille de l'abdomen et la durée de développement. Depuis, d'autres travaux ont décrit différents types de réarrangements induits par les Foldbacks notamment des inversions, duplications, délétions, translocations chez de nombreux organismes comme la drosophile, le muflier, le maïs, ou encore le *Fusarium* (Engels and Preston, 1984; Gray, 2000; Lim and Simmons, 1994; Zhang and Peterson, 2004).

Dans son discours, à la remise du prix Nobel pour la découverte des éléments mobiles du génome chez le maïs, Barbara McClintock proposait que les éléments transposables puissent jouer, malgré leur image d'ADN égoïste et mutagène, le rôle d'éléments de régulation de l'expression des gènes (McClintock, 1984). De nombreux travaux ont permis par la suite d'étayer cette hypothèse et il est aujourd'hui admis que les éléments transposables ont non-seulement ce rôle de régulation, mais que leurs gènes ont également été fréquemment « cooptés » par le génome de leur hôte au cours de l'évolution pour accomplir une grande diversité de processus biologiques. Dans le paragraphe suivant, nous aborderons brièvement le rôle des transposons en tant que séquences régulatrices et présenterons ensuite des exemples de domestication moléculaire des transposons à l'origine de mécanismes biologiques reposant sur l'aptitude essentielle des transposons d'introduire des CDB.

2.1.3. La domestication des transposons et les cassures double-brin « programmées »

2.1.3.1. Les éléments transposables régulateurs

Premièrement, les transposons peuvent jouer un rôle de séquences régulatrices (pour revue (Biémont, 2010)). Par exemple, les Tf1 de *Schizosaccharomyces pombe*, des transposons à « Long Terminal Repeats » (LTR), s'intègrent préférentiellement en amont des gènes transcrits par la Pol II (Bowen et al., 2003), en particulier les gènes induits par un stress environnemental. Dans l'étude (Feng et al., 2013), les auteurs ont tout d'abord montré que l'intégration de Tf1 ne réduit pas l'expression des gènes adjacents. En fait, l'insertion de Tf1 active l'expression de 6 gènes parmi les 32 étudiés. Deux hypothèses permettent d'expliquer l'augmentation du niveau d'expression de ces gènes (i) la transcription initiée par le transposon se poursuit par la transcription du gène adjacent (ii) Tf1 contient des séquences *enhancer* augmentant l'activité des promoteurs des gènes préexistants. L'étude montre que la deuxième hypothèse est la bonne, c'est-à-dire que les gènes surexprimés sont transcrits à partir de leur site naturel, comme indiqué par le séquençage des cDNA et la co-migration des ARNm issus des gènes avec et sans Tf1. De manière intéressante, les auteurs montrent que l'activité promotrice des Tf1 est contrôlée par la cellule, soit par *Nonsense-mediated mRNA decay* (NMD), soit par la protéine Abp1 connue pour son rôle de répression transcriptionnelle des transposons (Daulny et al., 2016). Les auteurs démontrent en outre que le promoteur Tf1 est activé par la chaleur et rapportent que seuls les gènes naturellement impliqués dans la réponse à ce stress ont vu leur niveau de transcription augmenter du fait de l'insertion du transposon. Ces résultats suggèrent donc une synergie entre les Tf1 et la réponse au stress chez *S. pombe*.

Chez *S. cerevisiae*, les Ty5 (*Transposons of Yeast*) sont trouvés principalement dans des régions non codantes subtélomériques tandis que les Ty1 et Ty3 s'intègrent préférentiellement dans les régions situées en amont des gènes transcrits par l'enzyme *Pol III*, notamment les gènes codant pour les ARN de transfert (Bushman, 2003; Dai et al., 2007; Lesage and Todeschini, 2005; Sandmeyer, 2003) ce qui altère l'expression de ces derniers (Bolton and Boeke, 2003). De manière plus générale, en dispersant de multiples copies de séquences régulatrices dans le génome, régulant l'expression des gènes au niveau transcriptionnel, les éléments transposables contribuent de manière majeure à la formation de réseaux de régulation de l'expression des gènes. Ce modèle, proposé il y a maintenant cinquante ans (Britten and Davidson, 1969, 1971) a été illustré de nombreuses fois (pour revue (Feschotte, 2008)). Un exemple frappant en est donné par l'étude (Wang et al., 2007) qui montre que la dispersion d'éléments mobiles à LTR dans le génome humain est à l'origine de la création de près de 1500 copies du site de liaison de la protéine p53, un régulateur majeur de la suppression des tumeurs, ce qui représente près de 30% des sites de liaison de p53 validés par ChIP-seq. Une autre étude décrit comment d'anciens éléments transposables ont été transformés en éléments régulateurs hormone-dépendants, remaniant ainsi le profil d'expression des gènes des cellules de l'endomètre de manière spectaculaire. Cette étude est fascinante, car elle présente les bases moléculaires expliquant la transition de la gestation dans un œuf à la gestation des mammifères (Lynch et al., 2015).

2.1.3.2. Les éléments transposables domestiqués

Il est devenu très clair que la domestication des transposons a joué un rôle déterminant dans de nombreux mécanismes reposant sur l'induction de CDB programmées. Chez les vertébrés, la recombinaison V(D)J est un mécanisme de recombinaison site-spécifique initié par des CDB induites par les protéines RAG1 et RAG2 qui permet de créer un grand nombre de récepteurs de cellules T et d'immunoglobulines nécessaires à la reconnaissance d'antigènes diversifiés (Huang et al., 2016; Soulas-Sprauel et al., 2007). C'est un rare exemple de programme de différenciation médié par des réarrangements chromosomiques (Zufall et al.,

2005). Les gènes des immunoglobulines et des récepteurs T des vertébrés sont assemblés à partir d'un segment V pour « variable », D pour « diversity » et J pour « joining » durant le développement des lymphocytes B et T (Davis et al., 1984; Tonegawa, 1983). Ce processus d'assemblage est initié par les protéines RAG1 et RAG2, qui forment un complexe capable d'introduire des CDB au niveau des «recombination signal sequences » (RSS) du locus V(D)J. Les segments situés entre les séquences V(D)J sont éliminés et les extrémités d'ADN restantes sont refermées sous la forme d'une structure en boucle à cheveux. Ces extrémités particulières recrutent alors des facteurs de réparation impliqués dans la réparation des lésions liées aux radiations : la kinase ADN-dépendante DNA-PK, la protéine dimérique Ku, la protéine Artémis, la ligase IV qui réalisent la réparation par jonction d'extrémités non homologues (« *Non-Homologous End-Joining* », NHEJ) (Gellert, 2002). De nombreuses caractéristiques rapprochent le locus V(D)J d'un transposon. Tout d'abord, la disposition des séquences RSS évoque la disposition répétée inversée de l'extrémité des transposons. De plus, les gènes RAG1 et RAG2 sont encodés de manière compacte : très peu espacés et sans introns ce qui évoque l'organisation des gènes dans les transposons (Gellert, 2002; Lieber et al., 2004). Des études plus récentes ont montré d'une part que les séquences protéiques de RAG1 et RAG2 sont significativement similaires aux transposases de la superfamille de transposons Transib (Kapitonov and Jurka, 2006) et d'autre part que la réparation en épingle à cheveux des CDB est typique de la réparation d'autres transposases (Fugmann, 2010). Enfin, une étude a rapporté récemment la découverte d'un élément mobile « ProtoRAG » actif contenant deux gènes similaires à RAG1 et RAG2 dans le génome du lancelet (Huang et al., 2016).

Le système CRISPR/*Cas* (Clustered Regularly Interspaced Palindromic Repeats (CRISPR)/CRISPR-associated proteins (*Cas*)) est un système immunitaire adaptatif et héritable présent chez les procaryotes. Dans ce système, l'ADN exogène provenant de séquences virales ou plasmidiques est inséré dans un locus appelé « *CRISPR array* ». Ce dernier est constitué d'une répétition directe de palindromes exacts de 25 à 35 pb alternés de séquences « espaçeur » ou « *spacer* ». Le *CRISPR array* est flanqué d'un ou plusieurs opérons de gènes *Cas*. La réponse immunitaire de CRISPR/*Cas* se déroule en trois étapes : (i) Brièvement, pendant l'adaptation, un complexe de protéines Cas recherche un motif adjacent au protospacer (« Protospacer-Adjacent Motif », PAM) dans l'ADN exogène et clive un petit fragment de ce dernier, fragment qui est ensuite intégré au *CRISPR array* en tant que nouvel espaçeur entre deux répétitions. De manière intéressante, l'acquisition d'espaçeurs peut avoir lieu selon un mécanisme alternatif faisant intervenir une réverse-transcriptase (RT) (Silas et al., 2016). (ii) Pendant la phase d'expression, le *CRISPR array* est transcrit en un long ARN (pre-crRNA) qui est maturé par un complexe de protéines Cas ou par une protéine Cas unique en petits ARN nommés crRNA, constitués chacun d'une séquence espaçeur et d'une répétition. (iii) A la phase d'interférence, les crRNA, qui sont restés associés au complexe protéique ou a la protéine unique (*Cas9*) qui a permis leur maturation, servent de guide à la reconnaissance du proto-spacer et de la séquence espaçeur d'autres copies de l'ADN exogène qui sont alors clivées par ledit complexe protéique. Dans la plupart des systèmes CRISPR-*Cas*, le module adaptatif fait intervenir les protéines Cas1 et Cas2, qui forment un complexe dans lequel Cas1 est une endonucléase (intégrase) et Cas2 joue un rôle structurel. La diversité des systèmes CRISPR provient surtout des modules effecteurs sur la base desquels s'appuie la classification des différents types de CRISPR que nous ne détaillerons pas ici.

De manière très intéressante, Cas1, l'endonucléase qui permet l'intégration des séquences espaçeur dans le *CRISPR-array* n'est pas toujours codée dans le locus CRISPR-*Cas*. Dans une étude récente, les auteurs rapportent la découverte de gènes *Cas1* en dehors du locus CRISPR dans des groupes diversifiés d'Archées (Makarova et al., 2013). L'année suivante, une étude est publiée, montrant que ces copies « *solo* » homologues de *Cas1* ont un voisinage génomique évoquant fortement l'organisation d'éléments transposables avec des éléments terminaux répétés inversés et une ADN polymérase. La ressemblance de ces éléments génétiques mobiles présumés avec les Mavericks jusqu'alors décrits uniquement chez les eucaryotes, en particulier la présence de l'ADN polymérase conférant à ces éléments la capacité de se

répliquer de manière autonome ont conduit les auteurs à nommer ces éléments « casposons » (Krupovic et al., 2014). La nature « mobile » de ces éléments a en outre été validée par la démonstration expérimentale de l'activité intégrase de la casposase (Hickman and Dyda, 2015). La similarité du mécanisme d'intégration médié par la casposase avec l'acquisition des espaçeurs du système CRISPR, ainsi que la ressemblance entre les sites d'intégration des casposons avec les espaçeurs du système CRISPR indique que le module adaptatif de ce dernier est très probablement issu de la combinaison d'un casposon avec un système immunitaire d'interférence inné dont les caractéristiques exactes sont difficiles à déterminer (Béguin et al., 2016). Comme évoqué plus haut, le module effecteur du système CRISPR est très variable. Nous évoquerons ici uniquement le type II de modules effecteur car il fait intervenir l'endonucléase *Cas9* qui est à l'origine du puissant outil d'édition génomique que nous connaissons aujourd'hui (Doudna and Charpentier, 2014; Jinek et al., 2012). De manière intéressante, l'analyse structurelle de variantes de protéines Cas9 a montré que ces dernières partagent un domaine catalytique comparable à celui de *RuvC*, une endonucléase clivant les structures cruciformes de l'ADN formées par le surenroulement de séquences répétées inversées en introduisant des CSB (ou « *nicks* ») dans les brins de même polarité (Iwasaki et al., 1991), même si le reste de leur structure est très variable (Lewis and Ke, 2017), et partagent la plus grande similarité de séquence avec le gène TnpB encodant une transposase très répandue dans les transposons IS605 de bactéries et d'archées (Chylinski et al., 2014).

Il existe d'autres exemples de mécanismes biologiques dans lesquels l'introduction de CDB fait appel à des transposons domestiqués. Chez les levures bourgeonnantes et les levures à fission, le système d'alternance du type sexuel requiert l'induction de DSB « programmées » (Figure 2), (pour revue (Hanson and Wolfe, 2017)). Brièvement, chez *Saccharomyces cerevisiae*, le système d'alternance implique le remplacement unidirectionnel d'un fragment chromosomique. Le type sexuel de *Saccharomyces cerevisiae* est indiqué par le locus *MAT* dont il existe en plus deux copies inactives : *HML* qui est la copie inactive indiquant le type α et *HMR* qui est la copie inactive indiquant le type a. Les trois copies sont flanquées de régions dénommées X et Z, Y étant la région centrale. En outre, Y existe sous deux allèles que sont Yα et Ya. Le mécanisme de changement de type est induit par l'endonucléase HO qui reconnaît un site dégénéré situé à la jonction entre les éléments Y et Z (Nickoloff et al., 1986). Les régions X et Z guident alors la réparation de la cassure qui est réalisée par SDSA (voir la partie sur la réparation du génome) en utilisant préférentiellement la cassette désignant le type sexuel opposé (Haber, 2012). L'analyse phylogénétique d'HO a montré que cette endonucléase provient de la domestication d'une intéine trouvée dans certains allèles du gène *VMA1* qui code une sous-unité d'une ATPase vacuolaire. Chez *Schizosaccharomyces pombe*, le mécanisme est fonctionnellement similaire mais structurellement différent. *S. pombe* possède également trois copies du locus mat, dont deux sont inactives, flanquées de régions H1 et H2 similaires à X et Z chez *S. cerevisiae*. Chez *S. pombe*, l'alternance de type sexuel n'est pas initiée par une CDB mais par une CSB maintenue de manière stable tout au long du cycle cellulaire, convertie en CDB lors du passage de la fourche de réplication (Arcangioli, 1998; Arcangioli and de Lahondès, 2000). Cette différence avec le mécanisme de *S. cerevisiae* implique que chez *S. pombe*, seul un quart des petits-enfants mitotiques de la cellule en question change de type sexuel. L'absence d'homologie entre ces deux mécanismes indique qu'ils ont émergé indépendamment au cours de l'évolution. Chez *Kluyveromyces lactis*, une autre levure bourgeonnante, un système similaire existe mais c'est l'excision d'un transposon domestiqué qui induit la recombinaison du locus *MAT* (Barsoum et al., 2010), (Figure 2). Chez cette espèce, la conversion de *MATα* en *MATa* et la conversion de *MATa* en *MATα* fait intervenir deux transposases domestiquées, α3 et Kat1 respectivement. L'enzyme α3 est codée par un gène homologue à un gène de transposase MULE, *MATα3* alors que Kat1 est codée par un gène homologue à un gène de transposon hAT. Chez les souches *MATα* l'excision de l'élément MULE voisin de *MATα* déclenche la réparation de la CDB tandis que chez les souches *MATa*, Kat1 génère des cassures en épingles à cheveux dans les séquences résiduelles d'un transposon, voisines de *MATa*. En 2014, le mécanisme d'alternance du type sexuel a été décrit chez *Ogataea polymorpha* et *Komagataella phaffii*, des levures du clade des méthylotrophes. Ces génomes, comprennent une cassette *MATα* et une cassette *MATa*

dont la position et l'expression sont interverties par une inversion chromosomique (Hanson et al., 2014; Maekawa and Kaneko, 2014), (Figure 2). La description récente de mécanismes similaires dans le génome d'autres *Saccharomycotina* (Riley et al., 2016) et la relative conservation de la synténie autour des loci *MAT* entre les levures méthylotrophes et *Saccharomycetaceae* indique que le système à deux cassettes est probablement ancestral et qu'il s'est complexifié au cours de l'évolution (Hanson et al., 2014). De manière plus générale, la synténie autour du locus *MAT* est relativement conservée entre les trois subphylum *Saccharomycotina*, *Pezizomycotina* et *Taphrinomycotina* (Butler et al., 2004; Gordon et al., 2011a; Riley et al., 2016). Cette conservation indique que l'information du type sexuel des ascomycètes a été portée par le même locus tout au long de l'évolution de ces *phyla*. L'hypothèse selon laquelle le système d'alternance du type sexuel était initialement constitué de deux cassettes indiquant les types opposés permet notamment d'apporter une explication partielle à la ressemblance fonctionnelle frappante entre les mécanismes observés chez *S. cerevisiae* et *S. pombe* (Hanson and Wolfe, 2017).

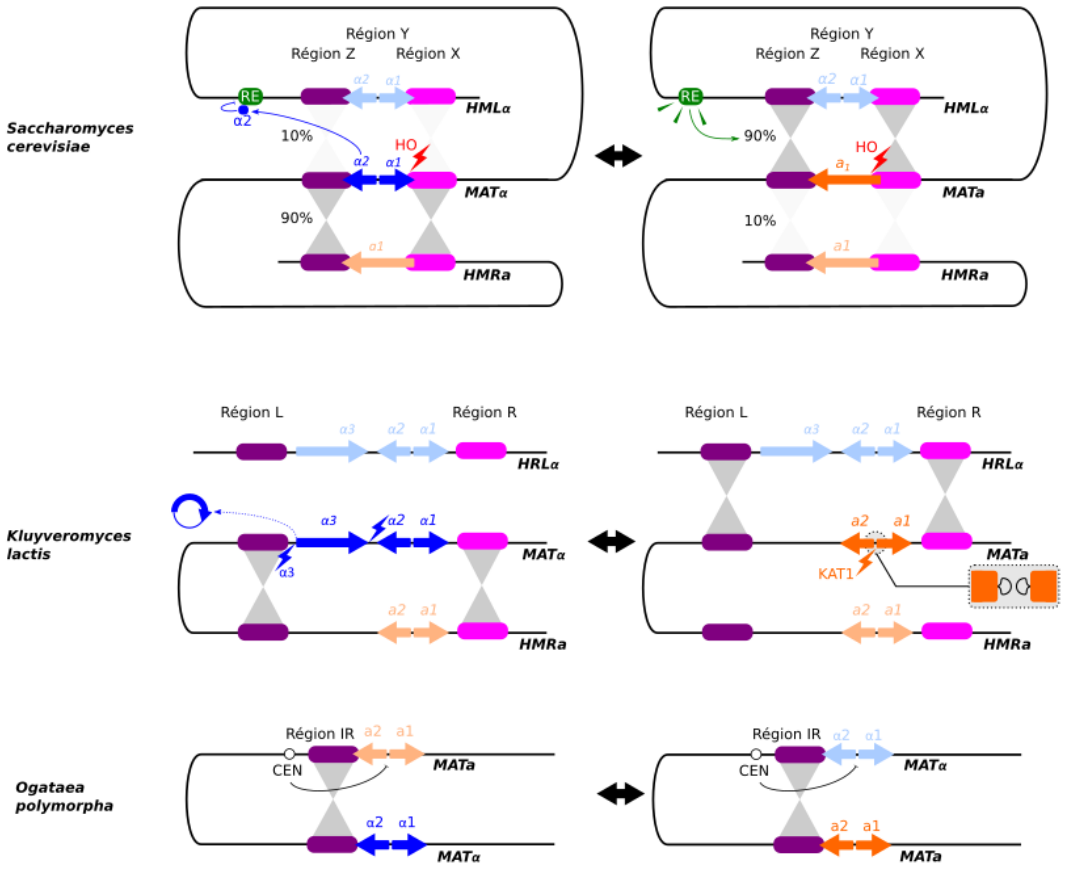


Figure 2 Organisation comparée du locus *MAT* et changement de type sexuel chez les levures *Saccharomyces cerevisiae* (*Saccharomycetaceae*), *Kluyveromyces lactis* (*Saccharomycetaceae*), et *Ogataea polymorpha* (méthylotrophe). Les gènes orange et bleu clairs sont inhibés. Les gènes oranges et bleu foncés sont actifs transcriptionnellement. (**haut**) Chez *S. cerevisiae*, l'endonucléase HO induit une CDB dans la région X. Le locus *MAT* est remplacé par conversion génique avec la cassette du type opposé dans 90% des cas. Chez les cellules *MATα* (à gauche), le complexe α2-Mcm1 recrute le répresseur Tup1 au niveau de la séquence RE, inactivant cette séquence par la formation de nucléosomes compacts. Chez les cellules *MATa* (à droite) le gène *α2* n'est pas activé. La séquence RE recrute les facteurs Fkh1 et SBF favorisant la recombinaison du locus *MAT* avec *HMLα*. (**centre**) Chez *Kluyveromyces lactis*, les cellules *MATα* expriment le gène *α3* codant pour une endonucléase. Cette dernière induit des CDB à chaque extrémité du gène *α3* qui est circularisé et rapidement éliminé car ce fragment circulaire d'ADN est non-réplicatif. Les régions d'homologie L et R guident la réparation de *MAT* par recombinaison homologue. Dans les cellules *MATa*, c'est l'endonucléase Kat1 qui induit une CDB au sein d'une séquence d'excision « fossile » d'un ancien élément mobile. Les extrémités de la cassure sont en épingle à cheveux (voir encadré), témoignant de l'origine de ce mécanisme. Notons que chez *K. lactis*, les cassettes *MAT* et *HMLα* sont situées sur le chromosome C tandis que *HMRa* se trouve sur le chromosome B. (**bas**) Chez *Ogataea polymorpha*, une inversion de 19kb intervertit les cassettes *MATa* et *MATα*. La cassette située à proximité du centromère est réprimée. Cette organisation est probablement l'organisation ancestrale qui s'est complexifiée au cours du temps pour donner les deux autres.

Dans les paragraphes précédents, nous avons présenté les différents mécanismes qui créent des cassures de l'ADN de manière accidentelle, ou programmée. Les extrémités d'ADN générées par ces mécanismes doivent être réparées par la cellule afin d'éviter la perte de matériel génétique qui serait létale. Dans la partie suivante, nous présenterons les différents mécanismes de réparation des CDB. Nous évoquerons brièvement leur efficacité relative à différents moments du cycle cellulaire ainsi que chez différents organismes. Nous nous intéresserons également à la « signature moléculaire » que ces mécanismes peuvent laisser dans les génomes.

2.2. La réparation des lésions de l'ADN et la création de réarrangements chromosomiques

De nombreux mécanismes assurent la détection et la réparation des lésions du génome. Toutefois, ces mécanismes ne sont pas sans faille et il arrive que des mutations soient introduites dans le génome. Certaines mutations sont dites ponctuelles et consistent en la modification, la délétion, ou l'insertion d'un petit nombre de nucléotides. L'impact phénotypique des mutations ponctuelles est aujourd'hui bien décrit.

D'autres mutations, fréquemment causées par les cassures de l'ADN affectent la structure du génome à large échelle : les réarrangements chromosomiques. Ces événements changent le nombre de copies et/ou la position de larges segments d'ADN. Les eucaryotes sont équipés d'un panel de mécanismes comparables, mais non identiques pour réparer les cassures de l'ADN. Nous présenterons dans les paragraphes suivants les mécanismes de réparation des CDB en prenant, sauf indication contraire, la levure *S. cerevisiae* pour modèle. Deux grands groupes de mécanismes permettent de réparer les CDB : les mécanismes de Jonction d'Extrémités Non-Homologues (NHEJ) et les mécanismes reposant sur la recombinaison homologue (HR).

2.2.1. La recombinaison homologue

La recombinaison homologue est un ensemble de voies de réparation des CDB dont les étapes sont très conservées dans l'ensemble de l'arbre du vivant. Tous les mécanismes de recombinaison homologue reposent sur une étape initiale commune : la réssection des extrémités d'ADN de 5' en 3'. Alors que pour la réparation NHEJ, les extrémités d'ADN sont réssectées sur une faible longueur, voire non-réssectées, les mécanismes de recombinaison homologue requièrent des extrémités plus longues. Le choix de réparer la CDB selon la recombinaison homologue ou la NHEJ se fait au moment de la réssection et il est fortement biaisé par le cycle cellulaire. Sur le plan moléculaire, ce choix a souvent été décrit comme une compétition entre le complexe Ku (initiant la réparation NHEJ) et le complexe MRX-Sae2 qui initie la résection (MRX est constitué de Mre11, Rad50 et Xrs2). Or, la délétion de la plupart des protéines impliquées dans la recombinaison homologue augmente l'efficacité de la réparation NHEJ (on parle ici de c-NHEJ) de manière relative et non pas absolue (Karathanasis and Wilson, 2002) ce qui indique plutôt une régulation qu'une compétition (pour revue de cette régulation, (Ceccaldi et al., 2016; Symington, 2016)). La première étape de la résection fait intervenir le complexe MRX-Sae2 qui dégrade les extrémités 5' sur environ 100 bases, puis deux machineries distinctes peuvent continuer à digérer les extrémités 5' au rythme de 4kb par heure. L'exonucléase Exo1 digère les nucléotides un par un tandis que le complexe Sgs1-Top3-Rmi1-Dna2 manifeste une activité hélicase/endonucléase et découpe le brin en fragment de quelques nucléotides (pour revue (Mehta and Haber, 2014)). A cette étape, un premier mécanisme de réparation par recombinaison homologue peut avoir lieu, le *Single-strand annealing*, qui consiste en un simple appariement de séquences simple-brin complémentaires suivi d'une synthèse et d'une ligation.

Toutes les autres voies de réparation des cassures font appel à la reconnaissance et à l'appariement d'une

extrémité coupée avec une séquence intacte de manière allélique ou ectopique (Krogh and Symington, 2004; Pâques and Haber, 1999). Après la réssection des extrémités 5', la protéine RPA stabilise le simple brin formé pour l'empêcher de former des structures secondaires. RPA est remplacé par la recombinase Rad51 (RecA chez les procaryotes). Le brin d'ADN associé aux recombinases forme un filament « recombinogène » qui recherche une séquence homologue intacte du génome. C'est l'étape limitante de la réparation homologue, qui peut durer plusieurs heures (Fabre and Zimmer, 2018). Pendant cette période, le chromosome cassé a une mobilité accrue lui permettant d'explorer un espace nucléaire dix fois supérieur à celui qu'il occupe en temps normal (2,7% du volume du noyau). De manière intéressante, les autres chromosomes participent également à cette « danse » ce qui permet aux loci homologues de co-localiser dix fois plus souvent durant cette période (Dion et al., 2012; Miné-Hattab and Rothstein, 2012). Le choix de la séquence utilisée pour la réparation est régulé de manière différente selon les espèces. Chez l'humain, la proximité spatiale des séquences homologues du génome guide fortement le choix du partenaire de recombinaison (Meaburn et al., 2007) tandis que chez la levure, les extrémités cassées semblent bénéficier d'un espace exploratoire plus grand pour trois raisons (i) le noyau est de taille plus réduite, relativement et (ii) les chromosomes sont moins contraints spatialement (iii) d'autres facteurs jouent également un rôle important dans ce choix comme le degré d'identité des séquences, la longueur des extrémités réssectées ou encore la présence de chromatine (pour revue (Bordelet and Dubrana, 2018; Fabre and Zimmer, 2018)).

2.2.1.1. Single-strand annealing (SSA)

L'hybridation simple brin est le mécanisme de recombinaison homologue le plus simple dans lequel deux séquences homologues flanquant une coupure se reconnaissent avec l'intervention des facteurs Rad52 et son homologue Rad59. Les extrémités non-homologues sont digérées par les endonucléases Rad1-Rad10 (XPF-ERCC1 chez les mammifères) dans un complexe faisant intervenir les protéines de correction de mésappariements Msh2-Msh3 et les protéines Slx4 et Saw1 qui jouent un rôle structurel (Li et al., 2008; Sugawara et al., 1997; Toh et al., 2010). Les nucléotides manquants sont alors polymérisés et les CSB sont re-liguées. La recherche des séquences d'homologie peut conduire à la délétion des séquences situées entre elles. La taille de ces délétions est restreinte par la présence de gènes essentiels. Ce mécanisme reste efficace même quand les séquences homologues sont distantes de 15kb (Pâques and Haber, 1999).

2.2.1.2. Le modèle de double jonction de Holliday (dHJ)

A la suite de la réssection d'une extrémité et l'invasion d'une séquence homologue, le nucléo-filament Rad51, hybridé à la séquence matrice (on parle aussi de *template*) forme une boucle de déplacement (D-loop). Avec l'action conjuguée d'hélicases et la synthèse d'ADN en utilisant l'extrémité résectée comme amorce, le brin néo-synthétisé est rallongé au point de présenter une séquence complémentaire de l'autre extrémité résectée de la cassure avec lequel il s'hybride, formant une double jonction de Holliday (*Figure 3*C). Une autre possibilité est que les deux extrémités issues de la résection envahissent la même séquence matrice formant une dHJ. Les jonctions de Holliday peuvent alors être dissoutes par Sgs1-Top3-RMi1 (BLM-Topo IIIα-RMI1/2 chez l'humain) ce qui ne forme pas de crossing-over. Les jonctions de Holliday peuvent également être résolues soit de manière symétrique par les résolvases YenI-CceI (Gen1 chez l'humain) ne formant pas de crossing-over, soit de manière asymétrique par Slx1-Slx4 et Mus81-Msm4 (Slx1-Slx4 et Mus81-Eme1, chez l'humain (Ciccia et al., 2003)) ce qui aboutit à un crossing-over (pour revue (Wyatt and West, 2014)). Chez la levure, la proportion de dHJ dissoutes est environ la même que la proportion de dHJ résolues (Mitchel et al., 2013). De plus, chez les eucaryotes, la résolution des jonctions de Holliday en mitose favorise fortement la réparation sans crossing-over (Pâques and Haber, 1999).

2.2.1.3. Synthesis Dependent Strand Annealing (SDSA)

La réparation SDSA (*Figure 3*D) consiste en l'invasion temporaire d'une séquence intacte par une extrémité

recombinogène. La structure formée par cette invasion est appelée boucle de déplacement ou « D-loop ». Le brin envahisseur, hybridé à la séquence homologue est rallongé par synthèse d'ADN. Il est assez étonnant de voir que dans le cadre du mécanisme de SDSA, le brin copié ne reste pas hybridé au template mais se dissocie pour revenir s'hybrider à l'autre extrémité résectée de la cassure, ce qui ne forme pas de crossing-over. Le mécanisme exact selon lequel la dissociation du brin néo-synthétisé a lieu n'est pas totalement élucidé mais les hélicases Mph1 et Srs2 semblent être impliquées (Mitchel et al., 2013). La réparation par SDSA est favorisée par rapport aux autres mécanismes de RH dans les cellules en mitose, probablement car elle minimise les risques de crossing-over. Cependant il peut arriver que la D-loop formée au cours de ce processus ne soit envahie par la deuxième extrémité recombinogène d'une CDB ce qui aboutit à la formation d'une dHJ qui, elle, peut être résolue en crossing-over. Le mécanisme de SDSA peut en outre introduire des duplications et délétions (Pâques and Haber, 1999).

2.2.1.4. Break-Induced Replication (BIR)

Nous avons mentionné précédemment comment une DSB peut être convertie en CDB à une seule extrémité au cours de la réplication. Il existe un mécanisme de réparation de l'extrémité générée par ce mécanisme. Après la réssection de l'extrémité unique de la cassure et l'invasion, dépendente des protéines Rad51 et Rad52, d'une séquence homologue, le processus de réplication induite par cassure (BIR) permet de transformer la boucle de déplacement (D-loop) en fourche de réplication unidirectionnelle (Malkova et al., 1996). Toutefois le processus BIR est très mutagène comparé au processus de réplication normal (Deem et al., 2011). Il a été démontré que la fourche de réplication mise en place dans un premier temps est très différente des fourches de réplication classiques. Elle se présente sous la forme d'une bulle avançant du fait de l'action de Pif1 avec une synthèse asynchrone du brin direct et indirect, ce qui entraîne l'accumulation de portions d'ADN simple-brin sujettes aux mutations (Donnianni and Symington, 2013; Saini et al., 2013). Il a d'autre part été montré que la fourche de réplication et la réplication des premiers kilobases par le processus BIR est très lente et change souvent de matrice (on parle de *template-switching*) avant d'être convertie en fourche de réplication classique (Smith et al., 2007). La fourche ainsi formée peut répliquer des bras entiers de chromosomes. Il a été proposé que la lenteur initiale du mécanisme soit due à des cycles d'invasion-réplication-dissociation de manière à permettre la détection de l'autre extrémité de la cassure (pour revue du BIR et son rôle, (Llorente et al., 2008)). Un premier modèle propose qu'après plusieurs essais d'invasion, si l'autre extrémité de la cassure n'est pas identifié, la réplication est amorcée et se poursuit jusqu'à l'extrémité du chromosome. Un autre modèle propose qu'à chaque cycle, l'extrémité cassée soit rallongée d'un petit fragment par réplication à partir du template puis se dissocie pour aller trouver un autre template, ce qui amorce une nouvelle synthèse, et ce jusqu'à ce qu'une séquence d'homologie appropriée permette la réparation avec la deuxième extrémité (si elle existe) par SDSA. Le processus est donc une source importante de perte d'hétérozygotie. Il a également été démontré que ce mécanisme est fortement impliqué dans l'apparition d'aberrations chromosomiques en particulier lorsque l'invasion initiale a lieu dans un élément répété du génome (Umezu et al., 2002). De manière très intéressante, une étude a montré qu'une DSB induite dans le voisinage d'un élément Ty du chromosome III de *S. cerevisiae* provoquait l'instabilité du génome et résultait en de nombreuses translocations entre Ty. Plusieurs de ces événements complexes sont cohérents avec l'implication d'un mécanisme de template-switching comme celui du BIR (VanHulle et al., 2007).

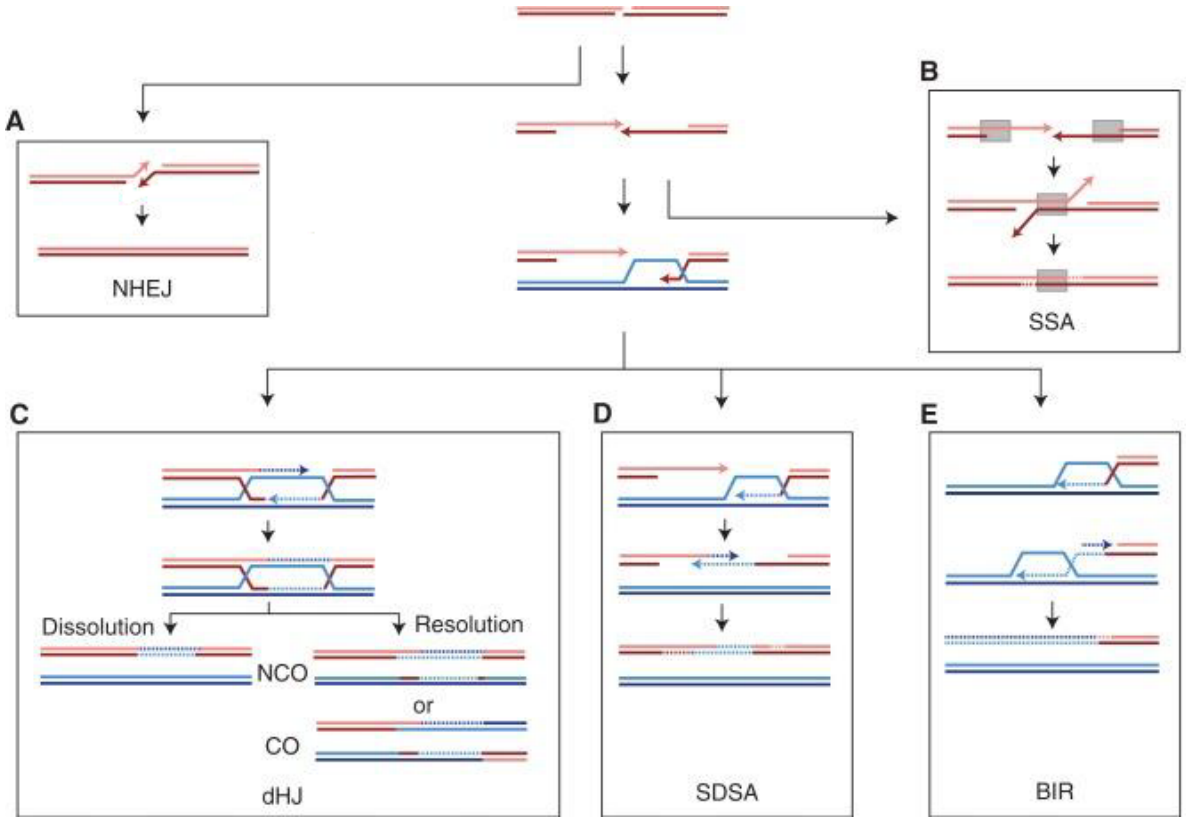


Figure 3 *Voies de réparation des CDB. (**A**) La réparation par NHEJ consiste dans un premier temps à protéger les extrémités double brin issues de la cassure. Ensuite, les extrémités présentant peu ou pas de nucléotides chevauchants sont liguées. Ce mécanisme est essentiellement fidèle, mais peut parfois introduire de petites insertions ou délétions. (**B**) La réparation par Single-strand annealing (SSA) a lieu quand la résection des extrémités de 5' en 3' met en évidence des séquences homologues de part et d'autre de la cassure. Ces séquences sont alors hybridées, les fragments de brins non hybridés sont digérés et les nucléotides manquants sont polymérisés. Les séquences initialement situées entre les régions d'homologie sont absentes du produit final. (**C**) Une double jonction de Holliday se forme par l'invasion d'une séquence homologue par les simples brins issus de la résection des extrémités de la CDB. Cette structure peut être soit démantelée, soit prise en charge par des résolvases qui donneront ou pas un crossing-over. (**D**) Dans le cas du SDSA, le brin d'ADN néo-synthétisé se dissocie de sa matrice d'homologie pour s'hybrider à l'autre extrémité simple-brin issue de la cassure. Les extrémités sont réparées sans crossing-over. (**E**) La voie de réparation par réplication induite par une cassure ou « break-induced replication » (BIR) consiste en l'invasion d'une séquence homologue par un simple brin issu de la résection qui relance la machinerie de réplication. Lorsque la séquence envahie est allélique, il y a perte d'hétérozygotie. Lorsque la séquence envahie est ectopique, cela aboutit à une translocation non réciproque. Adapté d'après (Mehta and Haber, 2014).*

2.2.1.5. Multi-invasion-induced rearrangement (MIR)

Dans une étude récente, les auteurs démontrent à la fois *in-vitro* et *in-vivo* l'existence d'un mécanisme selon lequel une extrémité d'ADN issue de la résection peut envahir successivement plusieurs séquences homologues situées sur des chromosomes différents sans que la D-loop formée à chaque invasion ne soit dissociée avant l'invasion suivante (Piazza et al., 2017). La résolution de ces D-loops fait intervenir les nucléases structures spécifiques Mus81-Mms4, Yen1 et Slx1-Slx4, formant des chromosomes transloqués à partir de chromosomes initialement intacts. De manière intéressante, il est à noter que les séquences d'homologie sur l'extrémité réssectée ne doivent pas être nécessairement voisines et que l'ADN situé entre les séquences d'homologies est retrouvé après résolution des D-loops, à la jonction des chromosomes transloqués. Les auteurs ont investigué les mécanismes de contrôle de MIR et ont identifié que Rad1 et Rad10 permettent de minimiser les invasions multiples en supprimant les extrémités hétérologues du brin recombinogène après la première invasion. Les auteurs démontrent également que les facteurs Sgs1-Top3-Rmi1, ainsi que les hélicases Srs2 et Mph1, préalablement associés à la dissociation des D-loops (pour revue (Wright et al., 2018)) inhibent la formation de translocations associées à MIR. Le contrôle de la D-loop et en particulier son démantèlement efficace sont très importants pour trois raisons (i) cela permet

de rejeter le fragment recombinogène pour que ce dernier puisse chercher une autre séquence homologue (ii) cela permet de minimiser les chances que les deux extrémités d'une cassure n'envahissent la boucle, ce qui forme une dHJ qui peut potentiellement être résolue en crossing-over, favorisant ainsi le mécanisme SDSA (Pâques and Haber, 1999; Yu et al., 2001) et (iii) cela minimise les chances que le filament recombinogène n'envahisse plusieurs séquences homologues successivement ce qui conduirait à la formation de translocations et de réarrangements en cascade (Piazza et al., 2017; Wright et al., 2018).

2.2.2. La réparation par « Non-Homologous End Joining » (NHEJ)

La jonction d'extrémités non-homologues est définie comme un processus de réparation dans lequel les extrémités d'un ADN cassé sont directement re-liguées (pour revue (Chiruvella et al., 2013)). La réparation NHEJ implique des mécanismes de modification des extrémités cassées qui partagent peu (<10bp) ou aucune homologie et qui créent parfois de petites délétions ou insertions. Le terme NHEJ a longtemps désigné une voie de réparation « canonique » (c-NHEJ) faisant intervenir Ku et la DNA ligase IV. Il existe cependant des mécanismes « alternatifs » (alt-NHEJ), parfois désignés par l'appellation de NHEJ « de secours » (« *backup NHEJ* ») faisant intervenir quelques bases d'homologie ; on parle aussi de jonction d'extrémités médiées par micro-homologie (« *micro-homology end joining* », MMEJ).

2.2.2.1. Canonical NHEJ (c-NHEJ)

La réparation NHEJ peut avoir lieu à tout moment du cycle cellulaire mais elle est 30 fois plus efficace en phase G1, quand les chromosomes n'ont pas de chromatide sœur pour réparer les CDB par recombinaison homologue et que la résection est limitée (Aylon et al., 2004; Chapman et al., 2012, 2012; Ira et al., 2004). Brièvement, les complexes Ku et MRX se fixent indépendamment au niveau de la cassure. L'hétéro-dimère Ku70/Ku80, très abondant dans le noyau se fixe à chaque extrémité d'ADN de manière séquence-indépendante, en s'hybridant à la structure phospho-glucidique de l'ADN et non aux bases azotées (Downs and Jackson, 2004). Ku empêche la résection d'avoir lieu et contribue au recrutement de la ligase Lig4. Le complexe MRX (Mre11-Rad50-Xrs2) maintient les extrémités et recrute Tel1 qui phosphoryle l'histone H2a. MRX recrute le complexe Nej1-Lif1-Lig4 via son interaction avec Xrs2. La ligase Lig4 répare alors la lésion.

2.2.2.2. Alternative NHEJ/Microhomology Mediated End Joining (MMEJ)

L'alternative-NHEJ (alt-NHEJ) regroupe les mécanismes de réparation « résiduels » observés quand la c-NHEJ n'est pas disponible dans la cellule. Les réactions enzymatiques de l'alt—NHEJ sont moins bien décrites que pour la c-NHEJ et il n'est toujours pas très clair si l'alt-NHEJ est une voie dédiée de réparation des DSB ou bien si elle simplement rendue possible par la « rencontre » de facteurs impliqués dans d'autres voies de réparation (Sfeir and Symington, 2015). En particulier, une forme d'alt-NHEJ appelée MMEJ répare les DSB en utilisant un faible nombre de bases d'homologie entre les extrémités d'ADN (<25pb) et est de plus indépendante de Rad52 (pour revue (Chiruvella et al., 2013; Daley et al., 2005)). Cette voie de réparation est indépendante du complexe Ku et de la DNA ligase IV. En revanche, elle dépend du complexe MRX-Sae2 qui initie la résection bidirectionnelle des extrémités 5' de la cassure par l'activité endonucléolytique de Mre11, puis éventuellement par l'exonucléase exo1 (Cannavo and Cejka, 2014; Garcia et al., 2011). La micro-homologie sert alors à aligner les extrémités ayant subi la réssection et les endonucléases Rad1-Rad10 digèrent les extrémités hétérologues. Les nucléotides manquants sont synthétisés et les cassures sont liguées par la ligase Cdc9 (Lig1 chez l'humain) (Sfeir and Symington, 2015).

La réparation par NHEJ est rapide, et dure environ 30 minutes chez l'humain, contre 7h pour la réparation

homologue. La réparation NHEJ est en revanche très mutagène : outre les mutations ponctuelles que ces mécanismes introduisent dans les séquences réparées, la réparation de fragments d'ADN issus de sites de coupures différents introduit des réarrangements chromosomiques parfois délétères. Chez l'humain, la majorité des CDB sont réparées par NHEJ (Ferguson and Alt, 2001). Ce mécanisme est à l'origine de translocations récurrentes (Mani and Chinnaiyan, 2011; Mitelman et al., 2007) pouvant former des oncogènes par fusion de séquence codantes (pour revue (Latysheva and Babu, 2016)) dont l'exemple le plus célèbre en est probablement la translocation t(9;22)(q34;q11) formant un chromosome très court découvert par Peter Nowell en 1956 et surnommé par la suite « chromosome de Philadelphie » causant une leucémie en plaçant le gène *ABL1* sous le contrôle du gène *BCR*. Chez la drosophile, la réparation NHEJ est le principal mécanisme à l'origine de réarrangements chromosomiques (Rothkamm et al., 2003).

2.2.2.3. Microhomology/Microsatellite-Induced Replication (MMIR) et Microhomology-Mediated Break-Induced Replication (MMBIR)

Il existe deux mécanismes permettant de relancer une fourche de réplication processive de manière comparable au BIR. Toutefois, la voie BIR est dépendante de l'invasion d'une séquence d'homologie longue et fait intervenir Rad51. Les mécanismes MMIR et MMBIR sont indépendants de Rad51. Dans une étude de 2008, les auteurs se sont intéressés aux mécanismes à l'origine de duplications de longs segments chromosomiques (Payen et al., 2008). Les auteurs ont commencé par investiguer les mécanismes moléculaires à l'origine de duplications segmentales (DS) et ont montré que le taux de formation de DS est fortement augmenté chez *S. cerevisiae* quand le gène *Clb5* est délété (*Clb5* est une cycline B connue pour activer les origines de réplication tardives) ou quand les topo-isomérases I sont inhibées par l'usage de la camptothécine. De plus, les auteurs ont trouvé que tous les mécanismes de formation des DS font intervenir la sous-unité Pol32 de l'ADN polymérase Polδ, démontrant que les DS proviennent de mécanismes réplicatifs plutôt que de mécanismes de recombinaison illégitimes. Ces résultats pourraient indiquer que la voie BIR est à l'origine des DS (Pol32 initie la synthèse de l'ADN dans le mécanisme BIR) mais les auteurs démontrent que seulement la moitié des DS peuvent être attribuées à ce mécanisme. En outre, la voie BIR nécessite l'invasion d'une séquence homologue longue alors que le mécanisme décrit ici fait intervenir de petites séquences d'homologie ou des séquences peu complexes selon un mécanisme Rad51-indépendant. Comme ce nouveau mécanisme partage avec le MMEJ la caractéristique d'utiliser séquences d'homologie courtes de manière Rad51-indépendante et que la voie de réparation est Pol32 dépendante à la manière du BIR, les auteurs ont nommé cette voie de réparation MMIR. Un mécanisme comparable a également été décrit dans les cellules humaines (Hastings et al., 2009).

2.2.3. Mécanismes de réparations et architecture des génomes

La proportion du nombre de CDB réparées par NHEJ ou RH est très dépendante du groupe phylogénétique considéré. Les levures et bactéries utilisent de manière prépondérante la recombinaison homologue tandis que les plantes et les mammifères privilégient la réparation par NHEJ (Puchta, 2005; Rocha et al., 2005; Sargent et al., 1997). Ces caractéristiques pourraient être une conséquence de la composition du génome. En effet les génomes de plantes et d'animaux contiennent une grande proportion d'éléments répétés, fournissant de nombreuses opportunités de recombinaison ectopique (Puchta, 2005; Sargent et al., 1997). Dans ce contexte, la réparation par NHEJ aurait été privilégiée afin d'éviter l'accumulation de réarrangements délétères, la contrepartie étant l'accumulation d'indels contribuant au vieillissement et à la tumorogenèse (Lieber and Karanjawala, 2004).

Cette distinction entre organismes favorisant NHEJ ou RH semble par ailleurs avoir un certain impact sur l'architecture des gènes. Par exemple, les levures *Saccharomycetaceae* se distinguent par une très faible proportion de gènes avec introns (Bon et al., 2003; Neuvéglise et al., 2011) or, de nombreux travaux englobant à la fois animaux, plantes et champignons tendent à montrer que les génomes ancestraux des eucaryotes étaient bien plus riches en introns que les génomes actuels (Carmel et al., 2007; Roy, 2006; Stajich et al., 2007). Cette observation n'est pas tout à fait expliquée toutefois la perte des introns dans les génomes des *Saccharomycotina* est cohérente avec la prévalence de la RH par rapport à la NHEJ et la présence de nombreux éléments transposables de type I encodant des réverse-transcriptases, deux éléments favorables à la recombinaison d'ARNm rétro-transcrits dans le séquences codantes des gènes, comme suggéré il y a plus de trente ans (Fink, 1987). Cette hypothèse est supportée par de rares données expérimentales (Derr, 1998) et par la localisation biaisée des introns, majoritairement dans la partie 5' de la séquence codante des gènes (Mourier and Jeffares, 2003). Cette observation est en adéquation avec le fait que les réverses transcriptases débutent leur activité à l'extrémité 3' de l'ARNm et se détachent souvent avant d'atteindre la fin du transcrit, ce qui rendrait la perte des introns par RH des gènes avec l'ADNc issu de la rétro-transcription moins probable en 5' de la séquence codante. Toutefois, les éléments transposables de type I actifs sont inégalement présents dans les génomes des *Saccharomycotina* et la contribution d'autres mécanismes expliquant la perte des introns a été explorée plus récemment.

Dans l'étude (Hu, 2006), les auteurs proposent, en plus du modèle dans lequel un gène avec introns recombine avec un ADNc issu de la rétro-transcription, un nouveau modèle dans lequel une DSB dans un intron est réparée avec l'ADNc du même locus. Dans ce mécanisme, seul l'intron cassé est perdu. Dans l'étude (Farlow et al., 2011) les auteurs émettent l'hypothèse que la création d'introns se fait par NHEJ et que la perte se fait selon une combinaison de NHEJ et de RH. Cette hypothèse est soutenue par le fait que chez les drosophiles, qui ont divergé il y a environ 40 millions d'années, le taux d'apparition et de disparition des introns sont fortement corrélés, suggérant l'implication d'un mécanisme unique. Les auteurs proposent en outre un modèle mathématique dans lequel l'efficacité relative de la NHEJ *versus* RH détermine l'augmentation ou la diminution du nombre d'introns dans les génomes dont le nombre d'introns est dynamique. Ces hypothèses sont fortement corroborées par l'étude (Hooks et al., 2014) dans laquelle les auteurs analysent près de 250 introns orthologues chez les *Saccharomycetaceae*. Ces analyses indiquent que les introns sont généralement perdus avec une grande précision, avec au maximum un ou deux codons insérés/supprimés au niveau des jonctions intron-exon. Une étude préalable ayant indiqué que toutes les espèces de levures présentent un biais dans la position des introns (en 5' des gènes), les auteurs en ont conclu que la recombinaison homologue était responsable de la plupart des pertes d'introns. Toutefois, Hooks et collègues rapportent que *L. kluyveri* a conservé légèrement plus d'introns que *L. waltii* et *L. thermotolerans*. Cette observation est intéressante au regard du fait que *L. kluyveri a* perdu les voies de réparation NHEJ/MMEJ (Gordon et al., 2011a) alors que *L. waltii* et *L. thermotolerans* ses deux plus proches parents possèdent toujours ces voies de réparation. Ceci soutient l'hypothèse que la réparation NHEJ serait aussi impliquée, bien qu'en moindre proportion que la RH, dans la perte des introns.

Comme nous l'avons vu dans cette partie, les CDB et les mécanismes qui assurent leur réparation ont un potentiel de remaniement très fort de l'information génétique. D'une part les CDB peuvent détruire des phases codantes ou des séquences régulatrices, d'autre part, les mécanismes de réparation des CDB peuvent réparer ensemble des séquences de manière ectopique, créant de nouvelles combinaisons génétiques. On imagine donc bien le potentiel délétère des réarrangements chromosomiques. Pourtant, de manière paradoxale, ces derniers se fixent dans les populations au cours de l'évolution. On peut donc se demander quel est l'impact évolutif les réarrangements chromosomiques. Afin de répondre à cette question, les réarrangements chromosomiques peuvent être considérés à deux échelles. On peut d'une part comparer les génomes de différentes espèces et essayer de retracer la contribution des différents types de réarrangements chromosomiques au cours de l'évolution. On peut également avoir une vision plus

« granulaire » et étudier l'impact phénotypique des réarrangements chromosomiques chez un organisme précis. Nous retracerons dans la partie suivante les connaissances auxquelles ont permis d'accéder ces deux grands axes de recherche.

3. Impact évolutif des réarrangements chromosomiques

3.1. Les différents types de réarrangements chromosomiques

Les réarrangements non-balancés du génome sont à l'origine de variations du nombre de copies dans le génome ou « copy number variations » (CNV). Les réarrangements balancés ne font pas varier le nombre de copies dans le génome mais déplacent des fragments de chromosomes. On distingue de plus les réarrangements intra-chromosomiques, qui n'affectent qu'un chromosome à la fois et les réarrangements inter-chromosomiques qui affectent plus d'un chromosome. Les différents types de réarrangements sont représentés sur la Figure 4.

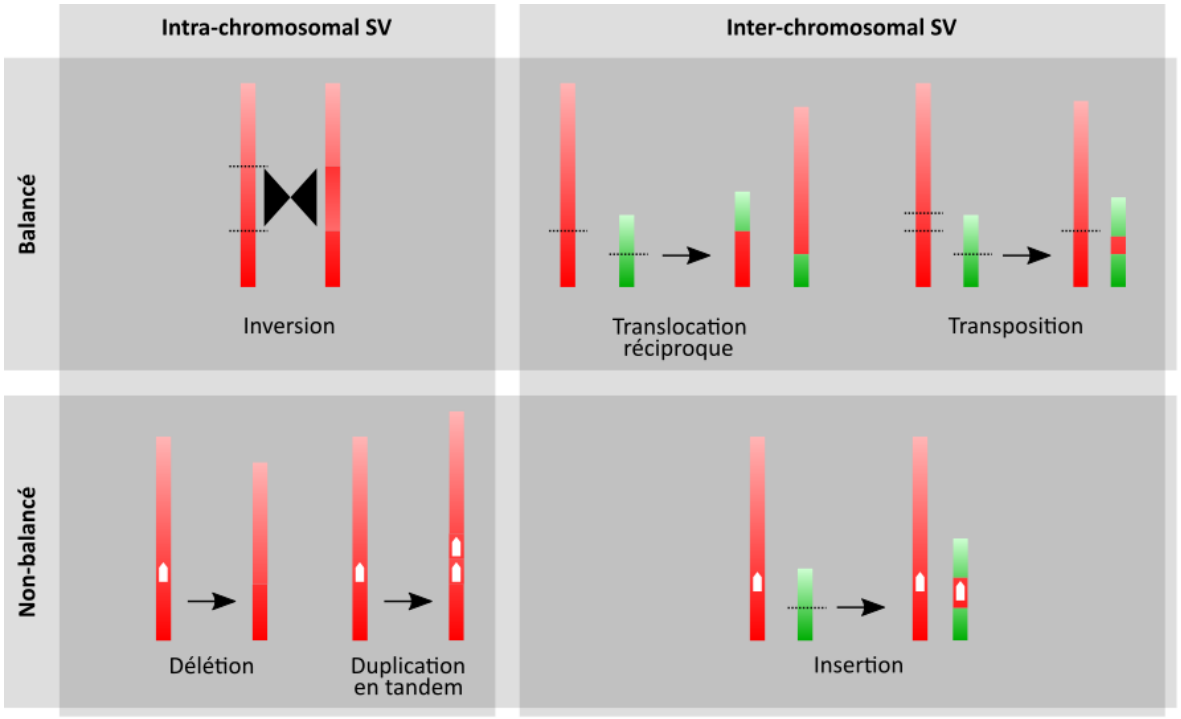


Figure 4 Exemples de réarrangements chromosomiques. **En haut à gauche**, une inversion est caractérisée par le changement d'orientation d'un fragment d'ADN au sein même d'un chromosome. Une inversion est péri-centrique si elle englobe le centromère et para-centrique si elle ne touche qu'un bras du chromosome. **En haut à droite**, la translocation réciproque consiste en l'échange de fragments de bras chromosomiques entre deux chromosomes sans gain ou perte de matériel. La transposition est le déplacement d'un fragment de chromosome qui vient s'insérer dans un autre chromosome sans gain ou perte de matériel génétique. **En bas à gauche**, une délétion est la perte d'un fragment chromosomique. **En bas à droite**, une insertion est l'acquisition de matériel génétique dupliqué à partir d'un chromosome qui s'insère dans un autre chromosome, par exemple du fait de l'activité d'un élément transposable. Adapté d'après (Harewood and Fraser, 2014)

Il existe d'autres types de mutation à large échelle que sont les événements d'introgression par hybridation et les transferts horizontaux de gènes. Ces mécanismes résultent en l'incorporation de fragments d'ADN exogène dans le génome d'une espèce. Nous en reparlerons dans le paragraphe suivant.

3.2. La reconstruction de génomes ancestraux

3.2.1. Les traces de l'origine commune des génomes

Pour comprendre comment la structure des génomes évolue, il faut pouvoir reconstruire les génomes ancestraux, c'est-à-dire identifier quels gène ils contenaient et surtout dans quel ordre ces derniers étaient disposés. Ces informations permettent alors d'inférer les événements qui ont remanié le génome ancestral dans les différentes lignées évolutives pour conduire à l'organisation des génomes tels qu'on les observe aujourd'hui. La reconstruction de génomes ancestraux repose sur l'identification de points communs entre les génomes actuels et de différences à partir desquelles on souhaite inférer un scénario réaliste.

3.2.1.1. L'homologie

La première étape de la reconstruction des génomes ancestraux est d'identifier des gènes homologues, c'est-à-dire des gènes des espèces actuelles qui ont une origine ancestrale commune. Cette tâche n'est pas aisée car tous les génomes évoluent par accumulation de mutations ponctuelles qui modifient la séquence des gènes. Ces dernières peuvent en outre inactiver des gènes (on parle de « pseudo-gènes ») ou en créer de nouveaux à partir de séquences non-codantes. Plus on compare des espèces distantes et plus les mutations ont eu le temps de s'accumuler entre les génomes. Il faut donc un moyen quantitatif pour identifier les gènes homologues avec suffisamment de sensibilité pour ne pas perdre tout signal ancestral quand on compare des espèces distantes et suffisamment de spécificité pour ne pas attribuer de relations d'homologie erronées.

La similarité est une indication du caractère homologue de deux séquences. Elle est calculée en alignant puis en comparant les séquences protéiques codées par deux gènes dont on souhaite tester le caractère homologue et en quantifiant leur ressemblance à l'aide d'une matrice de similarité qui donne un « score » pour chaque différence observée dans l'alignement. La matrice de similarité peut être construite à partir de différents critères que nous n'aborderons pas ici. Plus le score de similarité est élevé plus on peut affirmer avec confiance que les séquences sont homologues. Plusieurs mécanismes créent des séquences homologues :

Un événement de spéciation qui sépare les individus d'une espèce en deux populations isolées qui portent chacune leur propre copie de la séquence d'origine et qui divergent progressivement. On parle de séquences orthologues. L'événement de spéciation peut provenir de l'isolement physique ou de l'isolement reproductif de deux sous-populations (par exemple en raison d'un réarrangement chromosomique, voir le paragraphe 3.3, page 64).

La duplication d'un gène dans un génome forme des séquences paralogues qui peuvent diverger et se subfonctionnaliser. Les deux copies peuvent être localisées en tandem ou être dispersées dans le génome. La duplication totale du génome forme également des paralogues mais on les appelle plutôt ohnologues en référence à Susumu Ohno qui proposa le premier que la duplication de gènes ou du génome entier fournissent la « matière première » nécessaire à l'innovation évolutive (Ohno, 1970). Il a été démontré que des duplications totales du génome ont eu lieu à plusieurs reprises au cours de l'évolution des vertébrés (Coghlan et al., 2005; Huerta-Cepas et al., 2007; Vilella et al., 2009) et chez les plantes (Jiao et al., 2011). Un exemple vraiment spectaculaire de duplication totale du génome est donné par la paramécie *Paramecium tetraurelia* chez qui au moins trois duplications totales ont eu lieu successivement (Aury et al., 2006). Dans cette étude, les auteurs montrent que les gènes issus de la duplication totale du génome de *P. tetraurelia* sont conservés longtemps après l'événement de duplication non pas parce qu'ils se sub-fonctionnalisent, mais plutôt en raison des contraintes de dosage génétique.

En outre, les mécanismes d'acquisition de fragments d'ADN exogène compliquent l'identification des gènes homologues. Quand on observe des gènes similaires dans le génome de deux espèces, ces derniers n'ont pas nécessairement été transmis verticalement à partir d'un ancêtre commun. Une des deux copies peut être issue d'un événement de transfert dit « horizontal » du génome d'une espèce à un autre. De tels transferts peuvent laisser croire, si l'on considère uniquement les séquences transférées, que les deux espèces partagent une parenté proche alors qu'en fait l'histoire du fragment transféré est en désaccord avec l'arbre des espèces. C'est d'ailleurs la raison pour laquelle on combine généralement plusieurs sources d'informations pour construire un arbre phylogénétique. Le transfert de matériel génétique peut avoir lieu selon deux mécanismes différents :

Le transfert horizontal de gènes entre deux espèces forme des xénologues. Les transferts horizontaux ont lieu en proportions variables selon les lignées étudiées mais il est généralement admis que ces événements sont rares, pour plusieurs raisons. Tout d'abord, pour être hérité horizontalement, un fragment d'ADN devrait traverser des barrières mécaniques importantes, comme les membranes plasmiques et nucléaires pour atteindre les chromosomes. De plus, les gènes transférés entrent en proportion infime dans les populations, ce qui les rend sujets à la dérive génétique, rendant leur fixation improbable à moins qu'ils ne représentent un avantage sélectif énorme (tout comme pour les autres types de mutations). Enfin, chez les organismes dont les lignées somatiques et germinales sont distinctes, l'ADN hérité devrait être intégré dans le génome d'une cellule germinale pour être transmis verticalement or, si on prend l'exemple des vertébrés, les cellules des lignées germinales sont séquestrées très tôt au cours du développement embryonnaire, rendant un transfert potentiel plus difficile.

Dans l'article initial annonçant la finalisation du séquençage du génome humain, la présence de 113 gènes candidats issus de transferts horizontaux et similaires à des gènes de bactéries a été rapportée (International Human Genome Sequencing Consortium, 2001), puis commentée (Ponting, 2001). Dans ce bref article, Chris Ponting rappelle les caractéristiques de ces candidats soutenant l'hypothèse du transfert horizontal : (i) leur grande similarité avec des protéines bactériennes, plus qu'avec des protéines d'organismes non-vertébrés, (ii) la dispersion des « homologues » des gènes candidats issus de transferts horizontaux dans des groupes taxonomiques bactériens très diversifiés, tandis que leurs homologues eucaryotes sont restreints aux vertébrés, indiquant un transfert récent dans l'histoire de ce groupe, (iii) la présence d'introns dans ces gènes potentiellement issus de transferts horizontaux confirmant leur présence dans le génome humain (en plus de la validation par PCR, la possibilité d'une contamination des échantillons séquencés ayant été avancée). Cette observation exceptionnelle a déclenché un débat… exceptionnel, tant et si bien que cette « escarmouche de la guerre des génomes » a même fait l'objet d'un article du *New York Times* (Wade, 2001) ! Brièvement, Salzberg a été le premier à ré-analyser ces gènes, publiant la même année des arguments réfutant l'hypothèse du transfert horizontal (Andersson et al., 2001; Salzberg et al., 2001) Ces travaux illustrent notamment (i) l'impact du choix des génomes utilisés pour confirmer l'absence des gènes en question dans d'autres lignées d'eucaryotes, phénomène qui, en combinaison avec les pertes de gènes au cours de l'évolution peut suggérer des transferts horizontaux (ii) l'impact du choix de l'algorithme et du seuil du score d'alignement utilisé pour inférer des transferts horizontaux, plus généralement discuté dans une autre étude (Genereux and Logsdon, 2003) (iii) la concordance phylogénétique des gènes potentiellement issus de transferts horizontaux avec un scénario de transmission verticale plutôt qu'horizontale, présentée en plus ample détail dans autre étude (Stanhope et al., 2001). En résumé, Salzberg suggère un scénario dans lequel la perte des gènes au cours de l'évolution explique qu'il n'y ait pas d'homologues « reliant » ces gènes candidats aux séquences procaryotes. Dans une lettre adressée à Salzberg, DeFilippis suggère la transduction de ces séquences procaryotes aux vertébrés via des virus à ADN (DeFilippis and Villarreal, 2001; Villarreal and DeFilippis, 2000). Salzberg admet courtoisement la suggestion de DeFilippis et Villareal mais indique ne pas avoir trouvé de résultat significatif indiquant l'origine virale des gènes candidats ré-analysés. Le débat semblait clos, jusqu'à la

publication récente d'une nouvelle analyse dans laquelle les auteurs indiquent avoir identifié de nombreux gènes candidats issus de transferts horizontaux dans 10 génomes de primates, 12 génomes de mouches et 4 génomes de nématodes (Crisp et al., 2015). En outre les auteurs proposent 33 nouveaux gènes candidats dans le génome de l'humain et valident 17 candidats précédemment réfutés, affirmant par là même avoir « résolu la controverse autour des HGT dans le génome humain ». Dans une nouvelle étude, Salzberg rejette à nouveau les candidats les plus valables proposés par Crisp et réaffirme qu'en dehors de la migration, aujourd'hui bien décrite, des gènes mitochondriaux vers le génome nucléaire et la transduction de gènes médiés par les rétrovirus, aucun gène n'a été transféré horizontalement vers le génome humain (Salzberg, 2017).

Chez les levures *Saccharomycotina*, les transferts horizontaux de gènes sont plus facilement explicables par l'interaction directe de micro-organismes ou par l'acquisition de fragments d'ADN du milieu. Ainsi, des traces de séquences plasmidiques et de virus à ARN ont été détectées très tôt dans 8 parmi 20 génomes de *Saccharomycotina* (Frank and Wolfe, 2009). De plus, des gènes d'origine bactérienne ont été détectés dans pratiquement tous les génomes de levures séquencés jusqu'à présent. Par exemple, la souche de vin EC1118 de *S. cerevisiae* a hérité de plusieurs fragments chromosomiques provenant de *Zygosaccharomyces bailii* ainsi que de séquences d'origine bactérienne qui ont contribué à son adaptation à l'environnement de la fermentation du vin (Hall et al., 2005; Marsit et al., 2015; Novo et al., 2009). Chez les *Lachancea*, 24 transferts horizontaux de gènes ont été rapportés (Vakirlis et al., 2016). Ces événements correspondent au total à 85 séquences codantes dont 10 sont similaires à des gènes de *Pezizomycotina*, 3 sont similaires à d'autres eucaryotes, 11 proviennent de bactéries. De manière intéressante, *L. fantastica, L. waltii, L. thermotolerans et L. meyersii* possèdent des homologues d'une polysaccharide-lyase similaire à celle de *Botritis cinerea*. En accord avec la niche écologique dont *L. fantastica, L. waltii,* et *L. thermotolerans* ont été isolées (sur des végétaux) les homologues de ces espèces sont moins divergés que ceux de *L. meyersii* qui a été isolée de l'eau de mer. Parmi les 24 transferts horizontaux identifiés chez les *Lachancea*, 8 sont conservés chez plusieurs espèces. Selon le principe de parcimonie (voir le paragraphe suivant), qui est un principe important de la biologie évolutive, ont peut supposer qu'un transfert horizontal observé chez plusieurs espèces actuelles doit avoir eu lieu chez l'ancêtre commun de ces espèces. Un autre exemple est l'acquisition de gènes de β-lactamases par *Kuraishia capsulata* (Morales et al., 2013). *K. capsulata* possède par ailleurs la faculté d'assimiler des nitrates, une caractéristiques absente chez les *Saccharomycotina* à l'exception de quelques espèces appartenant au même clade que *K. capsulata* : *Ogataea polymorpha* et *Dekkera bruxellensis*. *K. capsulata* possède un cluster de gènes similaires aux gènes *crnA*, *niaD*, et *niiA* d'*Emericella nidulans* (aussi appelé *Aspergillus nidulans*), un champignon filamenteux du subphylum *Pezizomycotina* des Ascomycètes, chez qui ces gènes sont responsables de l'assimilation des nitrates (Johnstone et al., 1990). Ceci suggère un transfert horizontal de gènes entre *Emericella nidulans* et l'ancêtre commun de *K. capsulata*, *O. polymorpha* et *D. bruxellensis* (Morales et al., 2013).

L'introgression par hybridation peut également fausser des liens d'orthologie. Ce type d'événement consiste en le remplacement dans le génome d'une espèce, de séquences en position allélique par les séquences d'une autre espèce à la suite de la formation d'un hybride en l'absence de barrière pré-zygotique (par opposition au transfert horizontal de gènes qui désigne le transfert de matériel génétique entre deux organismes séparés par une barrière pré-zygotique). On parle également de « remplacement homologue ». Pour que le remplacement puisse être allélique, il faut que l'ADN des espèces hybridées puisse recombiner et soit donc suffisamment similaire. Chez l'humain, 2% du génome des populations non-africaines peut être attribué à l'introgression de séquences de Neandertal (Prüfer et al., 2014). Un exemple récent décrit dans le genre *Panthera* d'importantes discordances phylogénétiques le long du génome de ces espèces notamment en raison d'événements d'introgression dus à des hybridations interspécifiques. Les auteurs identifient notamment des gènes impliqués dans le développement cranio-facial, le développement des membres, le métabolisme protéique, l'hypoxie, la reproduction, la pigmentation et la perception sensorielle. De manière

intéressante, ces segments génomiques issus de l'introgression sont soumis de manière concordante à une forte pression de sélection, indiquant que les processus dans lequel ils sont impliqués sont probablement interdépendants. Ces résultats montrent que l'admixture a fortement contribué à l'adaptation évolutive des grands chats après leur spéciation (Figueiró et al., 2017).

Chez les levures du genre *Saccharomyces*, des exemples d'introgression ont été identifiés par le biais de la génomique des populations. Quand le génome complet de *S. paradoxus* a été séquencé, il est apparu clairement que les isolats européens de cette espèce possèdent une région de 23kb présentant seulement 0.1% de divergence par rapport au génome de *S. cerevisiae* alors que le reste du génome de *S. cerevisiae* et *S.* paradoxus présentent une divergence moyenne de 10 à 15% (Liti et al., 2006). Les auteurs ont privilégié l'hypothèse de l'introgression par rapport au transfert de gène car d'une part le fragment conservé est de grande taille, d'autre part le segment est localisé en position allélique et n'atteint pas les répétitions du télomère. Enfin, *S. cerevisiae* et *S. paradoxus* occupent la même niche écologique et sont naturellement capable de former des hybrides (Liti et al., 2005; Sniegowski et al., 2002). Par la suite, d'autres introgressions ont été identifiées dans les deux sens entre ces deux espèces (Muller and McCusker, 2009). Chez la levure *Lachancea kluyveri*, un fragment mesurant 1Mb et significativement plus riche en GC que le reste du génome laisse penser que cette espèce a également connu un événement d'introgression, toutefois l'espèce parentale responsable n'en est pas connue. Ces observations posent la question de la prévalence des mécanismes d'introgression dans l'évolution des levures *Saccharomycotina* et du degré de réticulation de la phylogénie de ces espèces (Dujon and Louis, 2017).

3.2.1.2. La synténie et la parcimonie

L'homologie de séquence n'est pas la seule trace de l'origine commune des génomes. Lorsque l'accumulation des données de cartographie génomique a rendu possible les comparaisons entre génomes d'espèces relativement proches, de nombreuses méthodes bioinformatiques ont vu le jour dans le but de détecter des régions de synténie conservée, qu'on appellera par la suite « régions synténiques », entre deux ou plusieurs espèces. Les régions synténiques sont des régions de différents génomes dans lesquelles l'ordre et l'orientation des gènes orthologues sont conservés. Notons que le mot synténie désignait initialement deux gènes localisés sur le même chromosome, puis a acquis la signification de voisinage direct entre deux gènes. Le terme a évolué une nouvelle fois et revêt le sens que l'on connait aujourd'hui de conservation de l'ordre des gènes orthologues dans le génome de deux espèces ou plus. Un bloc de synténie correspond donc à un groupe de gènes orthologues contigus dont l'ordre a été conservé dans deux génomes ou plus. Notons que la notion de synténie peut s'appliquer à tout marqueur génétique dont on connait la position physique ou relative. Les gènes constituent néanmoins les marqueurs les plus fréquents.

Comme nous le verrons dans la partie 3.3, page 64, les réarrangements chromosomiques sont souvent très délétères ce qui rend rares ceux qui parviennent à se fixer dans les populations au cours de l'évolution. De ce fait, si des gènes sont disposés dans le même ordre dans deux espèces différentes, il est raisonnable de penser que cette organisation est d'origine ancestrale. Il est effectivement plus facile pour une disposition donnée d'avoir été transmise à l'identique, sans réarrangement, à deux espèces plutôt que d'avoir été crée et fixée dans deux espèces par le biais de réarrangements indépendants (même si ce scénario n'est pas impossible). Ce principe, dénommé parcimonie, constitue la base de nombreux scénarios évolutifs.

Les ruptures de synténie sont donc la trace de réarrangements chromosomiques passés. Selon la divergence entre les génomes qu'on compare, les blocs de synténie peuvent comprendre quelques gènes à quelques milliers de gènes.

La synténie constitue donc une information ancestrale supplémentaire précieuse qui permet de délimiter les parties « volatiles » des génomes (comme les subtélomères) et d'accorder une plus grande confiance

quant à l'origine ancestrale de la disposition des orthologues observés dans le reste du génome. La synténie permet, en identifiant des séquences homologues, toujours actives biologiquement ou non, conservées dans un contexte synténique (i) d'identifier la localisation première de familles de gènes dispersées, (ii) de détecter l'orthologie des gènes évoluant rapidement, (iii) d'exclure de faux candidats de nouveaux gènes comme dans l'étude (Vakirlis et al., 2018) (iv) d'identifier les gènes issus de la domestication d'éléments transposables (Gordon et al., 2009; Sankoff, 2009) et (v) de reconstruire le contenu et l'ordre des gènes dans les génomes ancestraux. La synténie est un outil encore plus puissant chez les procaryotes chez qui les gènes qui interagissent fonctionnellement sont le plus souvent disposés en opérons.

3.2.2. Des traces de l'origine commune des génomes au génome ancestral

Les méthodes bioinformatiques de reconstruction de génomes ancestraux peuvent être regroupées en trois catégories : (i) les méthodes reposant sur les réarrangements, (ii) les méthodes reposant sur les adjacences et (iii) les méthodes reposant sur les arbres de gènes. De nombreux algorithmes ont été développés dans chaque catégorie, aussi nous nous intéressons aux algorithmes fondateurs de chaque type d'approche, aux logiciels que nous avons utilisés dans le cadre de cette thèse et présenterons les connaissances auxquels ont permis d'accéder la reconstruction de génomes ancestraux.

Les premières reconstructions de génomes ancestraux par des méthodes expérimentales remontent à 1938 chez les drosophiles (Dobzhansky and Sturtevant, 1938). Dans cette étude, les auteurs utilisent le fait que les chromosomes de glandes salivaires de drosophiles, polyténiques, sont bien visibles et que les chromosomes sont repliés de manière à apparier les locus homologues, pour inférer les inversions présentes dans le génome des populations américaines de *Drosophila pseudoobscura*. Les auteurs établissent en outre une « phylogénie » pour chaque chromosome reliant les arrangements chromosomiques observés dans les différentes populations par des arrêtes représentant les réarrangements. De manière amusante, les représentations schématiques du repli des chromosomes homologues en boucles (une inversion) en huit (une inversion incluse dans une inversion), etc. évoquent les graphes qui ne seront utilisés que bien plus tard dans le cadre des approches computationnelles pour reconstruire des arrangements ancestraux (voir plus bas).

Les méthodes de reconstruction ancestrales se sont d'abord intéressées à la reconstruction de l'état ancestral de séquences (Barry and Hartigan, 1987; Blanchette et al., 2004; Elias and Tuller, 2007; Fitch, 1971). Puis, la disponibilité accrue de génomes complètement assemblés et annotés a permis de tenter la reconstruction de l'ordre des gènes à l'échelle du génome entier. Reconstruire les génomes ancestraux permet d'étudier les mécanismes évolutifs des espèces actuelles. Il n'est toutefois pas facile de déterminer les configurations intermédiaires des génomes sur la seule base des espèces existantes.

3.2.2.1. Techniques reposant sur les réarrangements

Les méthodes reposant sur la distance/le nombre de réarrangements utilisent des modèles d'évolution des génomes pour rechercher les génomes ancestraux d'un groupe d'espèces en minimisant le nombre total de réarrangements (ou distance) observés sur toutes les branches d'un arbre phylogénétique. La recherche des génomes ancestraux est guidée par le principe de parcimonie qui se cache derrière le fait que la structure d'un génome reconstruit doit *a priori* provenir du scénario évolutif le moins couteux en réarrangements (le moins « long ») (*Figure 5*). Les méthodes reposant sur les réarrangements utilisent des graphes de points de cassure entre trois génomes G1 et G2 et G3 pour inférer les réarrangements qui ont fait diverger leur structure. Pour construire un tel graphe, il faut d'abord rechercher les blocs de synténie partagés entre ces génomes. Ensuite, un génome est fixé comme référence et ses blocs sont numérotés consécutivement par ordre croissant avec le signe « + ». On dit que ce génome est « identité ». L'ordre et l'orientation des blocs de synténie des autres génomes sont alors exprimés en fonction de l'identité. Ainsi si

le génome G2 possède une suite de blocs [1, 2, 3, 4, 5, 6, …] et que les blocs 2, 3 et 4 ont subi une inversion dans le génome G1, la permutation correspondante sera [1, -4, -3, -2, 5, 6]. On représente alors chaque bloc de synténie entre deux génomes (par exemple G1 et G2) par deux nœuds du graphe, chaque nœud représentant une des deux extrémités du bloc. Dans ce graphe, les arêtes représentent donc les adjacences entre les blocs. Dans la Figure 6, Les arrêtes en traits pleins représentent les adjacences dans G1 tandis que les adjacences en pointillé représentent les adjacences dans G2. Dans ce graphe, on observe que les arrêtes forment des cycles. Chacun de ces cycles peut correspondre à un ou plusieurs types de réarrangements. Un cycle de longueur 4 sous-entend l'existence d'une translocation réciproque ou d'une inversion. Pour des cycles de longueur supérieure, il devient difficile de retrouver quels types d'événements ont été mis en œuvre, en particulier lorsqu'un point de cassure est réutilisé et que deux événements ne peuvent plus être considérés indépendamment. Le principe des algorithmes reposant sur les réarrangements est de faire correspondre ces cycles à différents types de réarrangements, l'objectif étant d'identifier un « génome médian », c'est à dire la structure ancestrale qui minimise le nombre de réarrangements entre le génome ancestral reconstruit et trois espèces actuelles de référence (*Figure 5*). Même dans le cas où l'on considère seulement trois génomes, ce problème est np-complet (Avdeyev et al., 2016; Pe'er and Shamir, 1998) c'est-à-dire que la durée de résolution du problème augmente de manière exponentielle avec le volume de données à traiter.

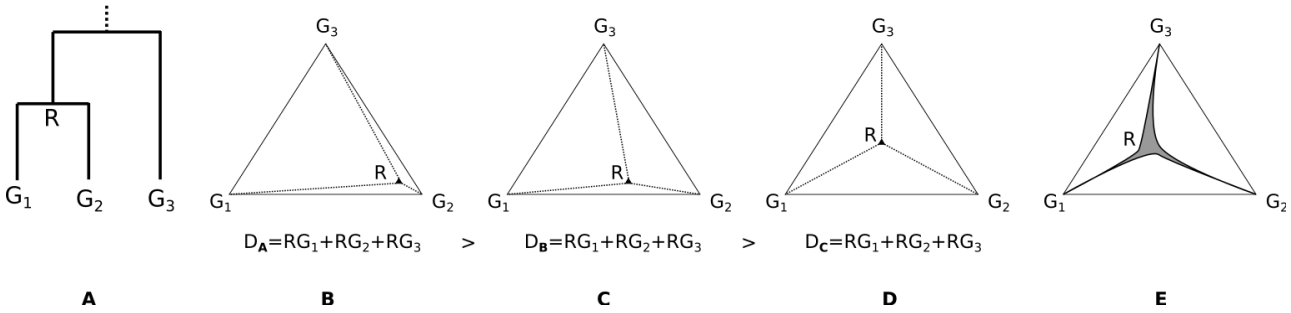


Figure 5 *Principe des méthodes de reconstruction utilisant les événements/la distance. (**A**) Les espèces actuelles G_1 et G_2 ont un ancêtre commun qu'on souhaite reconstruire R. L'espèce actuelle G_3 sert de référence externe. (**BCD**) Le logiciel recherche une structure ancestrale R telle que la distance D correspondant à la somme des distances entre la reconstruction et les génomes existants soit minimale. Ici, le scénario B est le moins parcimonieux et D est le plus parcimonieux et sera retenu. (**E**) En réalité, il est rare, voire impossible de pouvoir trouver l'agencement ancestral totalement correct. Avec les trois chemins évolutifs G_1G_2, G_2G_3 et G_3G_1 il s'agit d'inférer la structure de R qui se situe quelque part dans la zone grise. Selon la position plus ou moins excentrée de R dans cette zone, le génome Sreconstruit présentera une proportion variable d'adjacences erronées. La difficulté du problème est évidemment bien supérieure à la situation présentée ici on l'on cherche un barycentre dans un espace en 2D. La reconstruction de génomes est un problème multidimensionnel et mesurer une « distance » entre des structures de génomes est difficile.*

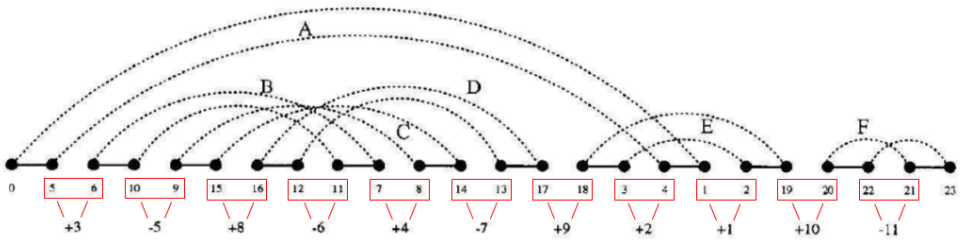


Figure 6 *Graphe de points de cassure. Chaque entité (bloc, en rouge) i de la permutation G1 = [3, −5, 8, −6, 4, −7, 9, 2, 1, 10, −11] est représenté par les nœuds 2i − 1 et 2i, le signe de i étant codé par l'ordre de 2i − 1 et 2i lorsque i est positif, et par 2i avant 2i − 1 lorsque i est négatif. Les nœuds 0 et 23 représentent les télomères. Les adjacences de la permutation G1 sont représentées par les lignes noires et celles de la permutation identité G2 par les lignes pointillées. Les différents cycles A, B, ...F représentent une ou plusieurs inversions : par exemple, F représente l'inversion du bloc 11.(Hannenhalli and Pevzner, 1999).*

Les logiciels ***BPAnalysis*** (Blanchette et al., 1997), ***GRAPPA*** (Moret et al., 2001a), ***MGR*** (Bourque and Pevzner, 2002) et ***EMRAE*** (Zhao and Bourque, 2009) ont été les premiers dans ce domaine. Toutefois ils sont extrêmement lents. Les méthodes les plus récentes d'optimisation de ce problème comprennent ***MGRA2*** (Alekseyev and Pevzner, 2009; Avdeyev et al., 2016), ***GASTS*** (Xu and Moret, 2011) et ***PATHGROUPS*** (Zheng and Sankoff, 2011).

BPAnalysis a été le premier algorithme permettant de mesurer des distances génomiques pour plus de deux génomes. En 1997, Blanchette et Sankoff définirent la notion à la base de ces approches : il y a point de cassure quand deux gènes sont adjacents dans le génome G1 mais pas dans le génome G2 et la distance entre les génomes est le nombre de points de cassure qu'ils présentent (sans considérer les réarrangements qui les ont formés). *BPAnalysis* cherche à résoudre le problème du génome médian comme un problème « du voyageur de commerce » en parcourant l'arbre dont on souhaite reconstruire les nœuds internes de manière itérative. Cette approche est très couteuse en temps de calcul puisque chaque itération est un problème np-complet à résoudre. Pour un jeu de données de 13 génomes avec 105 segments génétiques, la durée d'exécution du programme est estimée à 200 ans (Moret et al., 2001b).

MGR (Multiple Génome Réarrangements) a été développé pour répondre au manque de robustesse de *BPAnalysis* sur les génomes à plusieurs chromosomes, et pour améliorer le taux d'erreur de ce dernier. Ils ont marqué une certaine différence avec les outils précédemment présentés car ils mesurent la distance entre génomes en nombre de réarrangements plutôt qu'en nombre de points de cassure. *MGR* a été utilisé pour reconstruire les génomes ancestraux de mammifères (Murphy et al., 2005) ce qui a permis de tirer d'intéressantes conclusions sur l'évolution de ces génomes (voir le paragraphe 3.2.3, page 61). Ces résultats biologiques sont prometteurs, toutefois les reconstructions générées par *MGR* contiennent d'une part un faible nombre de gènes retracés (du fait de l'utilisation de marqueurs universels) et de plus l'algorithme utilise des approches heuristiques très « risquées » pour faire converger coûte que coûte la structure des génomes analysés en un ancêtre unique. En effet, *MGR* choisit simplement d'appliquer des événements de manière à ne pas rencontrer de « cul-de-sac » c'est à dire à générer un génome exact, même en introduisant des réarrangements peu fiables. *MGR* a été amélioré en 2008 par l'usage du modèle « *Double-cut-and-Join* » (DCJ) ce qui lui a permis de manipuler un plus grand nombre de réarrangements différents, notamment les translocations réciproques et les transpositions (Adam and Sankoff, 2008).

Le modèle DCJ (Yancopoulos et al., 2005) consiste à modéliser n'importe quel réarrangement (fusion, fission, inversion, translocation, excision et circularisation, transposition) par un seul et même mécanisme. Un événement DCJ sélectionne deux adjacences (soient deux adjacences AB et CD) qu'il rompt, les cassures étant réparées de manière à former de nouvelles adjacences AC et BD ou bien AD et BC. La nouvelle métrique associée à ce modèle de réarrangements compte simplement le nombre de DCJ ce qui simplifie beaucoup la combinatoire à manipuler pour obtenir des génomes médians et permet aux algorithmes de résolution du problème du génome médian de s'exécuter en temps linéaire, ce qui constitue un avantage majeur. En un sens, le fait d'introduire des cassures qui sont réparées de manière « accidentelle » peut sembler refléter la réalité biologique des réarrangements chromosomiques, toutefois certains aspects du modèle DCJ ne sont pas très réalistes, pour inférer la structure de génomes ancestraux comme par exemple le fait d'autoriser la circularisation de fragments (ce qui peut être le cas quand on s'intéresse aux réarrangements cancéreux impliquant des *double minute chromosomes* (Hahn, 1993)), ou le fait d'autoriser des translocations non-viables (ne tenant pas compte de la position des centromères). Une amélioration du modèle DCJ permet également de manipuler les délétions et duplications (Yancopoulos and Friedberg, 2009).

L'algorithme ***MGRA*** « *Multiple Génome Rearrangements and Ancestors* » (Alekseyev and Pevzner, 2009) dont nous reparlerons dans la partie résultats de cette thèse, repose sur une définition élargie du graphe de

points de cassure à n ≥ 2 espèces. Les nœuds du graphe peuvent donc avoir jusqu'à n arrêtes. L'algorithme répète deux étapes de façon itérative. La première consiste à identifier les réarrangements « fiables » c'est-à-dire ceux qui sont présents dans tous les scénarios parcimonieux possibles et qui ne réutilisent pas de points de cassure. La deuxième étape consiste à résoudre les réarrangements réutilisant des points de cassure à l'aide d'heuristiques pour pouvoir continuer à simplifier le graphe à l'aide de l'étape 1. L'algorithme s'arrête quand le graphe ne peut plus être simplifié. L'analyse des cycles extraits de ce graphe permet à *MGRA* de reconstruire l'arbre phylogénétique des espèces considérées, d'identifier les réarrangements sur les branches de l'arbre et enfin de générer les génomes ancestraux des espèces analysées avec une grande efficacité. Néanmoins, l'algorithme présente plusieurs limitations. Tout d'abord, comme pour les logiciels présentés précédemment, les blocs de synténie utilisés comme données d'entrée doivent être présents chez toutes les espèces analysées, ce qui restreint la proportion des génomes couverte par les blocs de synténie en particulier lorsque des espèces distantes sont analysées. D'autre part le fait que l'algorithme repose sur des heuristiques (bien que l'usage en soit restreint par rapport à MGR) pour résoudre les chemins réutilisant les points de cassure peut entrainer des erreurs. *MGRA*2 est capable de gérer des réarrangements plus élaborés que *MGRA* incluant les insertions, délétions et duplications (Avdeyev et al., 2016).

3.2.2.2. Techniques reposant sur les adjacences

Les méthodes reposant sur les adjacences/l'homologie considèrent les génomes comme un ensemble d'adjacences de marqueurs et assemblent les fragments ancestraux à partir des adjacences conservées dans un arbre phylogénétique. D'après le principe de parcimonie, les adjacences de marqueurs génomiques fortement soutenues par les différents génomes actuels sont ancestrales. L'idée en arrière plan des méthodes reposant sur les adjacences est qu'il est préférable de reconstruire des régions chromosomiques ancestrales avec certitude (même si le génome reconstruit est plus fragmenté) plutôt qu'en inférant un scénario évolutif pouvant introduire des erreurs. La première étape consiste donc à identifier les adjacences ancestrale à partir des génomes actuels et à les pondérer. La deuxième étape consiste à « agréger » avec des méthodes diverses les marqueurs avec les adjacences les plus fiables pour reconstruire la structure ancestrale (*Figure 7*). Les méthodes les plus connues de cette catégorie sont ***PMAG*** (Yang et al., 2014), ***InferCARS*** (Ma et al., 2006), ***proCARS*** (Perrin et al., 2015), ***ANGES*** (Jones et al., 2012) et ***Gapped Adjacency*** (*GapAdj*) (Gagnon et al., 2012). Le postulat de ces méthodes est que des adjacences de blocs fortement soutenues dans les différents génomes actuels sont ancestrales. De ce fait, les régions reconstruites sont appelées Régions Ancestrales Contiguës ou « *Contiguous Ancestral Regions* » (CARs). Ces approches reposent sur un autre type de graphe permettant d'inférer des réarrangements structuraux, les graphes d'adjacences (Bergeron, 2008). Ces graphes sont similaires aux graphes de points de cassure dans la mesure où ils représentent aussi des ruptures de synténie entre deux génomes réarrangés. Cependant dans ces graphes les nœuds ne symbolisent plus l'extrémité des blocs, mais une adjacence entre deux blocs. Chaque adjacence (nœud) est reliée aux adjacences avec lesquelles elle a un bloc en commun de sorte que les nœuds du graphe sont de degré au plus deux, sauf les télomères qui sont de degré 1 car ils sont situés à l'extrémité des chromosomes et n'impliquent qu'un seul bloc. Les propriétés de ces graphes permettent d'inférer une distance entre deux génomes qui renseigne sur le nombre de réarrangements qui se sont accumulés entre les génomes.

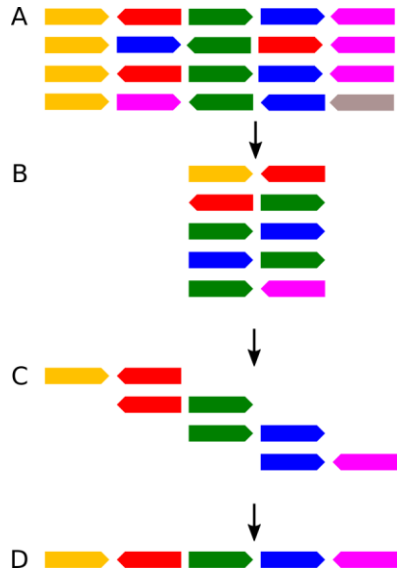


Figure 7 *Principe des méthodes de reconstruction reposant sur les adjacences. (**A**) La position de marqueurs d'homologie est recherchée dans les espèces actuelles. (**B**) Les adjacences les plus représentées sont conservées. (**C**) Selon différentes méthodes (méthode des 1 consécutifs, parcours de graphes, etc.), on tente d'agréger les marqueurs à l'aide des adjacences les plus fiables pour former les assemblages ancestraux les plus grands possibles.*

***Gapped Adjacency* (*GapAdj*)** (Gagnon et al., 2012) : Pour reconstruire un ancêtre donné, l'algorithme infère d'abord son contenu en gènes. Il réalise ensuite deux étapes de manière itérative (le nombre maximal d'itérations est un paramètre d'entrée) : lors de la première étape, les adjacences sont définies comme étant ancestrales si elles sont présentes chez suffisamment d'espèces actuelles (le seuil t est un paramètre de l'algorithme). Lors de la seconde étape, l'algorithme construit un graphe non dirigé d'adjacences, qui sont pondérées par leur score, calculé à la première étape. La reconstruction ancestrale consiste alors à trouver le cycle le plus long dans le graphe en ignorant les adjacences dont le score est inférieur à t. Pour résoudre ce problème du voyageur de commerce, *GapAdj* utilise une méthode heuristique. A chaque itération, *GapAdj* assouplit la notion d'adjacence conservée en autorisant un nombre croissant de marqueurs à s'intercaler entre des marqueurs considérés comme adjacents. Ainsi à la première itération, seules les adjacences directes entre marqueurs sont considérées. Aux itérations suivantes, on autorisera un marqueur, puis deux et ainsi de suite, à s'intercaler et on calculera les scores des adjacence. L'objectif de *GapAdj* est donc de trouver un compromis entre la fragmentation des génomes et l'exactitude des adjacences reconstruites. Les marqueurs utilisés par *GapAdj* doivent également être universels c'est-à-dire présents dans toutes les espèces analysées. Les marqueurs peuvent être dupliqués dans les génomes analysés mais *GapAdj* ne tient compte que des événements de duplication de génome total, qui doivent être indiqués dans l'arbre fourni.

ANGES (Jones et al., 2012) : De manière comparable à *GapAdj*, *ANGES* commence par établir le contenu en gènes de l'ancêtre à reconstruire. *ANGES* identifie les segments génomiques partageant une organisation similaire (en termes de marqueurs) pour chaque paire d'espèces dont le chemin évolutif passe par l'ancêtre à reconstruire. Ces ensembles de marqueurs hypothétiquement contigus dans le génome ancestral sont appelés « ACS » pour « *Ancestral Contiguous Sets* ». Chaque ACS est pondéré par son nombre d'occurrences dans les génomes actuels. *ANGES* reconstruit alors des régions chromosomiques ancestrales (CARs) en analysant les ACS par la méthode des « 1 » consécutifs.

PMAG+ (Hu, et al., 2014). L'algorithme original *PMAG* reconstruit les génomes ancestraux sur la base d'un modèle probabiliste et d'un modèle évolutif flexible. Dans un premier temps, l'algorithme calcule le contenu probable en gènes pour l'ancêtre à reconstruire. Ensuite, *PMAG* utilise la loi de Bayes pour calculer les probabilités de transitions des adjacences le long des branches de l'arbre phylogénétique fourni. L'algorithme calcule les probabilités conditionnelles de chaque adjacence puis reconstruit les génomes ancestraux en résolvant un problème de type voyageur de commerce selon la méthode heuristique Chained-Lin-Kernighan en maximisant les probabilités des adjacences utilisées. La version initiale de *PMAG* ne pouvait manipuler que les insertions et délétions. Dans une version plus récente, *PMAG+*, l'algorithme peut gérer les duplications et les duplications totales du génome. *PMAG+* a ensuite été intégré à un pipeline de reconstruction de génomes ancestraux, *MLGO* (Yang et al., 2014).

3.2.2.3. Techniques reposant sur les arbres de gènes

Ces techniques considèrent également les génomes comme des ensembles d'adjacences entre marqueurs, mais utilisent la phylogénie des familles de marqueurs en plus de la phylogénie des espèces. Brièvement, ces approches infèrent les adjacences ancestrales mais utilisent en plus les histoires de gènes pour optimiser les adjacences ancestrales. Il faut donc des histoires de gènes réconciliées avec l'arbre des espèces, c'est-à-dire que chaque événement de spéciation, duplication et transfert doit être documenté, ce qui constitue un pré-requis conséquent. Cette information permet d'inférer l'évolution des adjacences qui peuvent être gagnées, perdues, dupliquées ou transférées le long des branches de l'arbre (Figure 8). Les méthodes actuelles calculent l'histoire évolutive des adjacences soit en minimisant un critère de parcimonie, soit en maximisant une vraisemblance. La difficulté est donc d'inférer l'évolution des adjacences de manière cohérente par rapport aux histoires de gènes qui sont connues. Ces approches génèrent au final des ensembles d'adjacences pour chaque génome ancestral. Ces adjacences ne forment pas nécessairement une structure linéaire et il faut les « linéariser ». C'est l'approche suivie par *DupCar* (Ma et al., 2008) et *DeCo* (Bérard et al., 2012). Ces approches présentent l'avantage de pouvoir inférer les duplications et délétions, toutefois comme les approches auxquelles elles sont apparentées, elles génèrent des génomes fragmentés et ne permettent pas d'inférer les réarrangements passés.

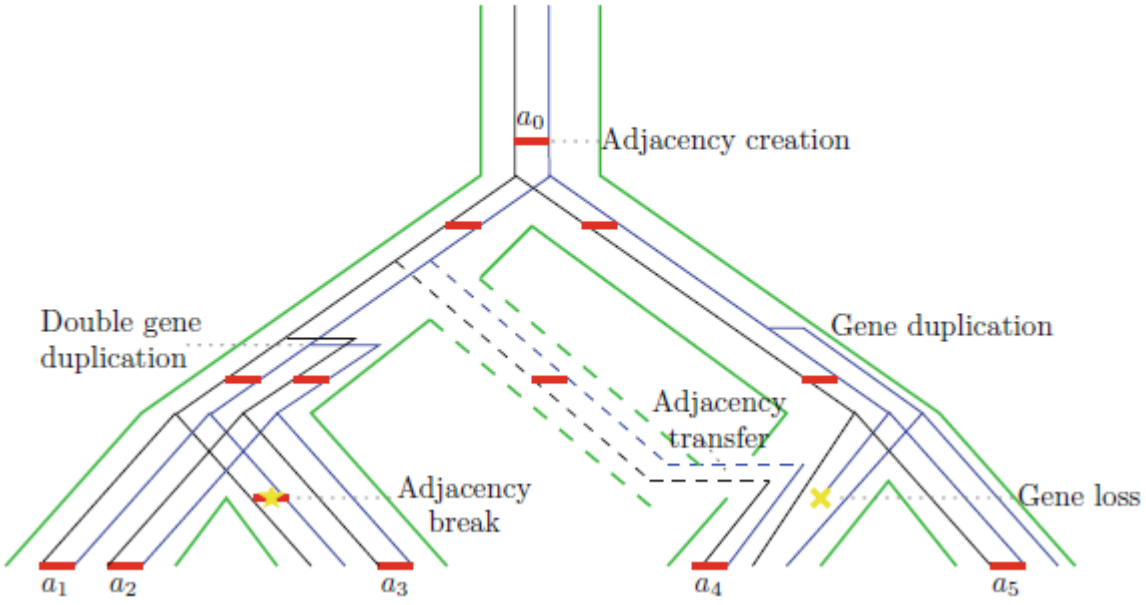


Figure 8 Histoire des adjacences (en rouge) le long des histoires de gènes (topologies noires et bleues) réconciliées avec l'arbre des espèces (en vert). Les cinq adjacences notées a_1 à a_5 partagent une origine ancestrale a_0. L'événement de duplication de plusieurs gènes (à gauche) duplique les adjacences associées. En revanche, une duplication d'un seul gène ne duplique pas d'adjacences (à droite). Les réarrangements et les pertes de gènes détruisent des adjacences. (crédit image (Anselmetti et al., 2018))

3.2.2.4. Les avantages apportés par l'algorithme *AnChro*

Les méthodes présentées dans les paragraphes précédents sont très diversifiées. Elles reposent sur des types différents d'information, les adjacences, les réarrangements, ou les arbres de gènes. Les types de marqueurs d'ancestralité sont également très diversifiés : ce peut être des gènes ou des blocs de synténie. Les marqueurs peuvent être uniques (non dupliqués dans les génomes) ou non, universaux (ou « ubiquitaires ») c'est-à-dire conservés dans toutes les espèces étudiées ou non. Chacune de ces méthodes présente des avantages et des faiblesses. Les approches reposant sur les réarrangements sont capables d'inférer les réarrangements passés mais génèrent des erreurs, les approches reposant sur l'homologie utilisent un signal plus fiable mais génèrent des génomes plus fragmentés.

AnChro développé par Guénola Drillon (Vakirlis et al., 2016) fait partie d'une suite logicielle appelée *CHROnicle*. Le premier logiciel de cette suite, *SynChro*, sert à identifier les blocs de synténie entre les paires de génomes dont on souhaite reconstruire les ancêtres (Drillon et al., 2014). Le deuxième logiciel, PhyChro, reconstruit l'arbre des espèces à partir des points de cassure identifiés en analysant les blocs de synténie générés par *SynChro*. Le troisième logiciel, ReChro, reconstruit les graphes d'adjacences entre les paires de génomes et calcule le nombre et le type de réarrangements qui ont eu lieu le long des branches de l'arbre. Enfin, *AnChro* infère l'ordre des gènes dans les génomes ancestraux. *CHROnicle* regroupe au sein d'un seul outil de nombreux avantages des approches mentionnées plus haut. Tout d'abord, *SynChro* génère des données de synténie issues de comparaisons de génomes deux à deux. C'est un avantage majeur, car en ne restreignant pas les données d'entrée à des marqueurs universels et/ou de nombre constant, on maximise l'information ancestrale exploitable pour les reconstructions. *ReChro* est capable d'utiliser d'une part les graphes d'adjacences pour valider comme ancestrales les adjacences conservées chez différentes espèces actuelles à la manière des approches reposant sur l'homologie. Il peut d'autre part, lorsque c'est possible, inférer les réarrangements passés pour mettre en évidence des adjacences des génomes ancestraux qui ne sont aujourd'hui plus observables dans les génomes actuels, à la manière des approches reposant sur les réarrangements.

AnChro reconstruit un ancêtre (noté A) en comparant dans un arbre phylogénétique binaire non-enraciné deux génomes G1 et G2, reliés par un chemin passant par A, avec le reste des espèces de l'arbre (G3, …, Gn) pour lesquels il existe un chemin rejoignant A sans croiser le chemin tracé entre G1 et G2. *AnChro* fait également appel à la liste de points de cassure définis par les blocs de G1 et G2 et la liste des blocs de synténie entre les paires G1/(G3, …, Gn) ainsi que G2/(G3, …, Gn). En réalité *AnChro* génère pour un même ancêtre A autant de reconstructions qu'il existe de paires informatives de génomes G1, G2 dans l'arbre formant un chemin passant par A.

La reconstruction d'un ancêtre est un processus en cinq étapes (*Figure 9*, page 61).

Premièrement, un score de confiance est calculé pour les adjacences entre les blocs. Supposons que la comparaison G1/G2 ait mis en évidence deux blocs A et B, adjacents dans G1 mais pas dans G2. Un score de confiance sera calculé pour déterminer la présence de l'adjacence des blocs A et B telle qu'observée dans G1, dans les génomes G3…Gn. Toutefois, l'adjacence AB de G1 ne peut pas être identifiée directement dans G3…Gn car les blocs A et B résultent d'une comparaison G1/G2. *AnChro* redescend alors à l'échelle des gènes pour rechercher dans les blocs de synténie issus de la comparaison G1/G3…G1/Gn, les RBH (reciprocal best hits) des quatre gènes de G1 situés de part et d'autre (deux en amont et deux en aval) du point de cassure AB dans G1. Cette étape est un des « piliers » d'*AnChro* car c'est elle qui permet d'utiliser toute l'information contenue dans la synténie entre les paires de génomes, sans pour autant se confiner à l'analyse des génomes deux à deux.

Les blocs de synténie sont obtenus avec *SynChro* par la comparaison de deux génomes G1 et G2 en faisant intervenir le paramètre Δ. Ce paramètre fixe le nombre de gènes consécutifs qu'on autorise à ne pas être conservés dans un bloc de synténie entre G1 et G2. Par exemple une région conservée en synténie entre G1 et G2 et contenant trois gènes consécutifs différents sera considérée comme un seul bloc avec le paramètre Δ = 3, mais sera scindée en deux blocs avec Δ ≤ 2. De même, les comparaisons de génomes G1/G3…Gn et G2/G3…Gn font intervenir le paramètre Δ'. Comme ces deux paramètres Δ et Δ' peuvent chacun varier entre 1 et 6, il existe 36 reconstructions alternatives d'un ancêtre calculé avec les mêmes génomes G1/G2/G3…Gn. Nous verrons plus loin comment ces différentes versions sont utilisées.

La deuxième étape consiste à valider les adjacences ancestrales. Pour chaque cycle du graphe d'adjacences entre G1 et G2, on attribue à chaque point de cassure le plus haut score de confiance calculé précédemment. Par exemple, si l'on reprend l'exemple de l'adjacence AB observée dans G1, on attribuera à cette adjacence le plus haut score de confiance obtenu en prenant comme référence G3, jusqu'à Gn. Si deux adjacences ancestrales contradictoires ont le même score, aucune n'est validée.

La troisième étape permet de retrouver des adjacences ancestrales supplémentaires en analysant le graphe d'adjacences tracé par ReChro. Par exemple, quand un cycle de longueur l = 2n est identifié entre deux génomes, si n-1 adjacences du cycle ont été validées, alors *AnChro* validera la $n^{ième}$ adjacence. Cette étape est également très importante car c'est elle qui donne à *AnChro* la possibilité d'inférer des adjacences ancestrales aujourd'hui disparues, ce qui s'avère impossible quand on utilise une approche reposant sur les adjacences seules.

La quatrième étape consiste à reconstruire les chromosomes ancestraux en partant des blocs télomériques validés comme ancestraux puis en progressant en ajoutant à ce début de chromosome les blocs de synténie consécutivement tant que des adjacences ancestrales sont disponibles, ou jusqu'à ce qu'un autre bloc télomérique soit atteint. Si un chromosome en cours d'élongation ne présente plus d'adjacence ancestrale validée, sa reconstruction est interrompue et il forme un *scaffold*. *AnChro* cherche alors à reconstruire un autre chromosome en partant d'un autre bloc télomérique. Si des chromosomes circulaires sont générés, l'adjacences ancestrale la moins fiable est rompue pour linéariser le *scaffold*.

Enfin, la cinquième étape vient finaliser les chromosomes ancestraux. Après la reconstruction des chromosomes (étape 4) un génome ancestral est défini comme une succession de blocs de synténie. Cette dernière étape consiste à retirer de cette reconstruction les blocs et les gènes dupliqués et à résoudre les micro-réarrangements observés entre G1 et G2 en comparant ces génomes aux génomes G3, …, Gn dans le but de produire des génomes sous la forme d'une liste de gènes ordonnés et orientés.

En conclusion, *AnChro* permet de « concilier » la précision des informations contenues dans les adjacences ancestrales tout en limitant la fragmentation des génomes reconstruits à l'aide de la résolution des chemins et des cycles des graphes d'adjacences. Cet « emprunt » algorithmique lui permet d'inférer les réarrangements chromosomiques qui ont eu lieu sur les branches de l'arbre des espèces étudiées, ce que ne permet pas une approche classique basée sur les adjacences. De plus il utilise une richesse d'information supérieure à celle retenue par les approches précédentes car il repose sur des comparaisons deux à deux de génomes, ce qui permet de limiter la perte d'information (en termes de conservation de la synténie) due à l'ajout d'une espèce plus distante dans les génomes à analyser.

Dans la partie résultats de cette thèse, nous verrons comment les performances de cet algorithme ont été évaluées et présenterons les résultats de l'application d'*AnChro* à la reconstruction des génomes de *Saccharomycotina*.

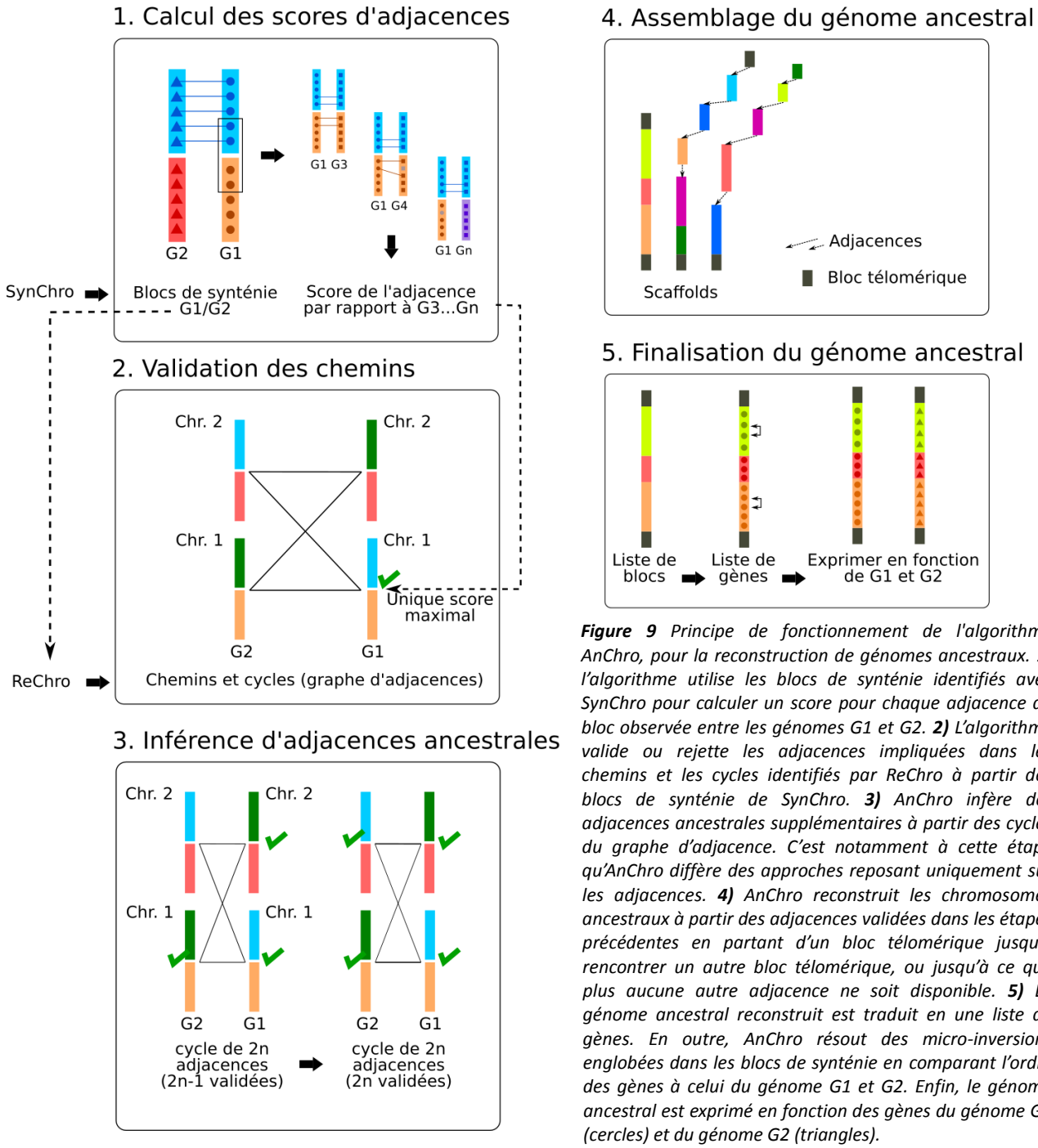


Figure 9 *Principe de fonctionnement de l'algorithme AnChro, pour la reconstruction de génomes ancestraux.* ***1)*** *l'algorithme utilise les blocs de synténie identifiés avec SynChro pour calculer un score pour chaque adjacence de bloc observée entre les génomes G1 et G2.* ***2)*** *L'algorithme valide ou rejette les adjacences impliquées dans les chemins et les cycles identifiés par ReChro à partir des blocs de synténie de SynChro.* ***3)*** *AnChro infère des adjacences ancestrales supplémentaires à partir des cycles du graphe d'adjacence. C'est notamment à cette étape qu'AnChro diffère des approches reposant uniquement sur les adjacences.* ***4)*** *AnChro reconstruit les chromosomes ancestraux à partir des adjacences validées dans les étapes précédentes en partant d'un bloc télomérique jusqu'à rencontrer un autre bloc télomérique, ou jusqu'à ce que plus aucune autre adjacence ne soit disponible.* ***5)*** *Le génome ancestral reconstruit est traduit en une liste de gènes. En outre, AnChro résout des micro-inversions englobées dans les blocs de synténie en comparant l'ordre des gènes à celui du génome G1 et G2. Enfin, le génome ancestral est exprimé en fonction des gènes du génome G1 (cercles) et du génome G2 (triangles).*

3.2.3. Des génomes ancestraux à la dynamique des génomes

Un exemple assez emblématique de problématique soulevée par les reconstructions de génomes ancestraux a été de savoir si les réarrangements chromosomiques ont des points de cassure aléatoires ou si au contraire les points de cassure sont réutilisés au cours de l'évolution. En 1984, Nadeau et Taylor ont montré que la distribution de taille des blocs de synténie au sens actuel du terme et sur la base des quelques 83 gènes homologues entre le génome humain et le génome de la souris suivait une loi de Poisson et en ont déduit que les réarrangements fixés au cours de l'évolution avaient une distribution aléatoire dans le génome. La distribution de Poisson de la taille des segments ayant été plus ou moins soutenue par des études subséquentes (Burt et al., 1999; Mural et al., 2002), une grande confiance a été accordée au fait que les génomes sont réarrangés de manière aléatoire. La première contradiction provient de l'étude (Pevzner

and Tesler, 2003) dans laquelle les auteurs, comparant le génome de la souris et le génome humain montrent que la distribution des segments conservés suit effectivement une loi de Poisson, mais que le nombre de points de cassure inféré par le logiciel GRIMM-Synteny (développé pour cette étude), est bien inférieur au nombre attendu si ces derniers étaient répartis de manière aléatoire, suggérant que près de la moitié des « *245 réarrangements pour changer les Hommes en souris* » font intervenir des points de cassure plusieurs fois. Ces résultats ont été à l'origine d'un débat animé dans lequel Sankoff et collaborateurs ont suggéré un biais méthodologique dans le fait de filtrer les petits blocs de synténie et avancé le fait que la perte de signal évolutif aussi pouvait créer des scénario évolutifs dans lesquels par « saturation » les points de cassure sont réutilisés (Sankoff and Trinh, 2005; Trinh et al., 2004).

En retour, dans l'étude (Murphy et al., 2005), les auteurs ont utilisé des blocs de synténie universels conservés entre le génome de l'Homme, du cheval, chat, chien, cochon, de la vache, du rat et de la souris pour reconstruire les génomes ancestraux de ces espèces avec l'algorithme MGR (Bourque and Pevzner, 2002). Le génome ancestral des Boréoeuthériens obtenu recouvre en nombre de gènes, environ 50% des gènes du génome humain et est très compatible avec la structure des génomes ancestraux inférés à partir de données cytogénétiques. Les auteurs se sont particulièrement intéressés au taux de réutilisation des points de cassure inférés par MGR et montrent que 20% des points de cassure identifiés ont été réutilisés au cours de l'évolution des génomes de mammifères. De façon intéressante, les auteurs montrent que les points de cassure associés aux aberrations chromosomiques cancéreuses les plus fréquentes chez l'humain co-localisent avec des points de cassure fixés au cours de l'évolution, trois fois plus souvent que dans les autres réarrangements. Des observations analogues ont été rapportées par la suite à plusieurs reprises (Darai-Ramqvist et al., 2008; Ruiz-Herrera et al., 2005). Par ailleurs, les auteurs définissent les points de cassure spécifiques aux primates comme des ruptures de synténie du génome humain par rapport à tous les génomes non-humains et montrent que 85% de ces points de cassure sont bordés de gènes dupliqués dans le même chromosome, observation soutenue par la suite (Bailey et al., 2004), suggérant une recombinaison non-allélique à l'origine des réarrangements. Enfin les auteurs démontrent que des centromères sont fréquemment formés *de-novo* au niveau des points de cassure réutilisés. Le fait que les points de cassure soient réutilisés a été largement démontré par la suite (pour revue (Faraut, 2008)).

Récemment, une étude a employé la reconstruction de génomes ancestraux avec les algorithmes *InferCARS* (Ma et al., 2006), *ANGES* (Jones et al., 2012), et une approche développée pour l'étude, pour expliquer la distribution des points de cassure des réarrangements des Boréoeuthériens (Berthelot et al., 2015). Dans cette étude, les auteurs ont cherché à modéliser à l'aide de régressions de Poisson, un modèle utilisé pour modéliser des événements rares, le taux d'apparition des points de cassure dans les intergènes du génome ancestral reconstruit en fonction des caractères intrinsèques de ces intergènes. Dans ce modèle, les gènes sont considérés comme « incassables ». Si les points de cassure sont aléatoires dans les intergènes, le nombre de points de cassure dans ces régions doit suivre de manière linéaire la longueur de ces derniers (hypothèse nulle) et suivre une loi de Poisson classique. Les auteurs montrent qu'il y a effectivement une corrélation positive très forte entre la longueur de l'intergène et le taux de cassure, mais que cette relation n'est en revanche pas linéaire. L'équation reliant le taux de cassure r à la longueur de l'intergène L obtenue est $r = 2{,}4.10^{-3} x L^{0,28}$ (Berthelot et al., 2015). Le taux de cassure croît donc bien moins rapidement qu'une équation linéaire en fonction de la longueur de l'intergène. Cette observation est intéressante car bien qu'elle ne soit pas vraiment expliquée, elle suggère néanmoins que les contraintes (quelles qu'elles soient) qui s'exercent sur un intergène court ont un effet « interactif » ou « aggravant » expliquant le nombre de cassures accru par rapport aux intergènes longs. Réciproquement on peut imaginer que les intergènes longs permettent une meilleure « relaxation » des contraintes et supportent mieux celles-ci. De plus, les auteurs montrent que le maintien de la structure du génome (de l'ordre des gènes) ne fait l'objet que d'une très faible pression de sélection pourvu que les points de cassure des réarrangements soient localisés en dehors de séquences codantes ou régulatrices. Enfin, l'étude montre que la distribution de points de cassure

évolutifs dépend fortement des probabilités de contact dans le génome ainsi que de l'accessibilité de la chromatine.

La reconstruction de génomes ancestraux permet également d'apporter la confirmation que certains événements de remaniement ont bien eu lieu au cours de l'évolution. Il est généralement admis aujourd'hui que les génomes des vertébrés se sont diversifiés à la suite de deux duplications totales du génome appelées 1R-2R, il y a environ 450 millions d'années. Toutefois, les traces d'un événement aussi ancien ne sont pas aisément détectables et d'autres scénarii ont été proposés pour expliquer l'évolution des génomes de vertébrés. Il a notamment été proposé qu'un seul événement de duplication du génome ait eu lieu et que des duplications segmentales multiples expliquent la synténie observée entre les génomes de vertébrés. Cette année, le groupe d'Hugues Roest Crollius a publié une étude dans laquelle les génomes ancestraux de la lignée menant aux vertébrés ont été reconstruits dans le but de déterminer si deux événements de duplication totale du génome ont bien eu lieu ou non (Sacerdot et al., 2018). Tout d'abord, les auteurs ont déterminé les familles de gènes présentes dans le génome ancestral des amniotes à partir des arbres de gènes du serveur Ensembl. L'algorithme AGORA (Berthelot et al., 2015) a ensuite été utilisé pour reconstruire le génome ancestral des amniotes. Le génome ainsi obtenu est composé de 470 segments dont 50% des gènes sont dans des segments de plus de 250 gènes. Partant des 56 fragments reconstruits avec plus de 50 gènes, les auteurs ont recherché les ohnologues dans le génome reconstruit en utilisant uniquement leur date de duplication ainsi qu'en s'assurant que les paires de gènes considérés appartiennent à des fragments reconstruits (CAR) différents. En comparant deux à deux les CAR issus de la reconstruction à la recherche d'ohnologues présumés, les auteurs montrent que chaque CAR reconstruit est homologue en moyenne à trois autres, supportant un scénario à deux duplications totales successives. Les auteurs ont ensuite utilisé ce génome ancestral comme « point d'appui » pour reconstruire la structure du génome ancestral tel qu'il était avant les événements 1R-2R, aboutissant à un génome à 17 chromosomes. De manière frappante, la cartographie des gènes des chromosomes ancestraux pré-1R dans le génome humain montre que la structure de ce génome ancestral est encore remarquablement visible (Sacerdot et al., 2018). En résumé, l'élégance de cette étude réside dans le fait d'avoir réussi à démontrer l'existence des événements de duplication du génome 1R-2R sans *a priori*, en utilisant le génome ancestral des amniotes reconstruit comme point d'appui pour éviter le bruit lié à la comparaison des espèces actuelles de vertébrés.

La comparaison des génomes actuels et la reconstruction de génomes ancestraux a également permis de mieux modéliser la dynamique des génomes, notamment la relation entre différents types de mutations. Le taux de mutations non-synonymes a été corrélé à plusieurs reprises avec des réarrangements chromosomiques : chez les protéobactéries et les archées, les taux de duplications et pertes de gènes corrèlent avec les taux de substitutions non-synonymes (Csűrös and Miklós, 2009; Snel et al., 2002). De manière similaire, des corrélations linéaires ont été établies chez les bactéries entre duplications et pertes, transferts horizontaux, création de gènes et taux de mutations non-synonymes (Puigbò et al., 2014). Dans cette étude, les auteurs attribuent le nom d' « horloge génomique » à cette observation par analogie à l' « horloge moléculaire » (Zuckerkandl and Pauling, 1962). Chez les levures, l'étude (Rolland and Dujon, 2011) est la première à tenter de relier la divergence de séquence avec le nombre de blocs de synténie.

La reconstruction de génomes ancestraux chez les levures a été initiée par le groupe de Kenneth Wolfe. Dans une première étude (Gordon et al., 2009), les auteurs reconstruisent manuellement le génome de l'ancêtre de *Saccharomycetaceae* précédant la duplication totale du génome, puis infèrent les réarrangements qui ont remanié le génome de cet ancêtre jusqu'au génome actuel de *S. cerevisiae*. Dans une seconde étude, l'histoire des réarrangements de *Lachancea kluyveri* depuis sa divergence de son ancêtre commun avec *S. cerevisiae* est reconstruite (Gordon et al., 2011a). Les auteurs montrent en outre que le génome de *Lachancea kluyveri* a été étonnamment peu réarrangé depuis la divergence avec l'ancêtre

commun à *S. cerevisiae* (15 réarrangements seulement). En identifiant quels gènes de l'ancêtre reconstruit sont absents chez *Lachancea kluyveri*, les auteurs ont remarqué que les gènes *DNL4*, *POL4*, *NEJ1* et *LIF1* impliqués dans la réparation NHEJ chez les levures (voir la partie 2.2.2, page 45) ont été perdus. Les auteurs émettent l'hypothèse que par conséquent, la plupart des cassures chez *Lachancea kluyveri* sont réparées par recombinaison homologue, expliquant également le mode de vie essentiellement diploïde de cette levure (de Clare et al., 2011).

3.3. Valeur sélective des réarrangements chromosomiques

3.3.1. Réarrangements chromosomiques, méiose, isolement reproductif

Les réarrangements balancés sont particulièrement délétères au bon déroulement de la méiose, qui se déroule en deux étapes : (i) la recombinaison entre les chromosomes homologues puis (ii) la réduction du nombre de chromosomes pour passer d'une cellule diploïde (2n) à une cellule haploïde (n). Cette deuxième étape dépend du bon déroulement de la recombinaison entre chromosomes homologues à la première étape. Or, les translocations réciproques à l'état hétérozygote impliquent la formation de quadrivalents méiotiques (*Figure 10*) qui impactent fortement la proportion de spores viables (Avelar et al., 2013; Liti et al., 2006; Loidl et al., 1998). En outre, la recombinaison méiotique de chromosomes dans une cellule portant une inversion paracentrique hétérozygote crée fréquemment des chromosomes di-centriques ou acentriques non viables.

De par cet impact négatif sur la viabilité des spores, les réarrangements chromosomiques contribuent au phénomène de spéciation en tant que barrière post-zygotique (Delneri et al., 2003; Hou et al., 2014). La démonstration première du fait que l'isolement reproductif ne provient pas seulement de l'accumulation de réarrangements chez les levures émane d'une étude chez les *Saccharomyces*, dans laquelle les auteurs ont testé la validité du modèle de spéciation chromosomique (Fischer et al., 2000). Dans cette étude, les auteurs ont recherché les translocations entre les génomes des espèces du genre *Saccharomyces* : *S. cerevisiae, S. paradoxus, S. bayanus, S. cariocanus, S. mikatae* et *S. kudriavzevii* en utilisant le karyotype de *S. cerevisiae* comme référence. Brièvement, les auteurs ont montré que le génome de *S. paradoxus* et *S. cariocanus*, deux espèces sœurs du clade frère de *S. cerevisiae* étaient respectivement porteuses de 0 et 4 translocations par rapport à cette dernière. Le génome de *S. kudriavzevii*, plus lointain, a été trouvé colinéaire à celui de *S. cerevisiae* (compte tenu de la résolution de la technique selon laquelle les auteurs utilisent des sondes obtenues à partir de l'amplification de gènes de *S. cerevisiae* pour identifier la position des centromères et l'extrémité des chromosomes dans les karyotypes des autres espèces), démontrant que les réarrangements chromosomiques ne sont pas la seule cause de spéciation. En outre, les auteurs ont réalisé des croisements entre *S. paradoxus* et *S. cariocanus*. En s'appuyant sur le fait que l'impact maximal d'une translocation réciproque sur la viabilité des spores issues de la méiose est de 50%, la viabilité des spores entre ces deux espèces devrait être de 6,25%. La viabilité observée de 0,66% indique clairement l'implication d'autres mécanismes contribuant à la spéciation.

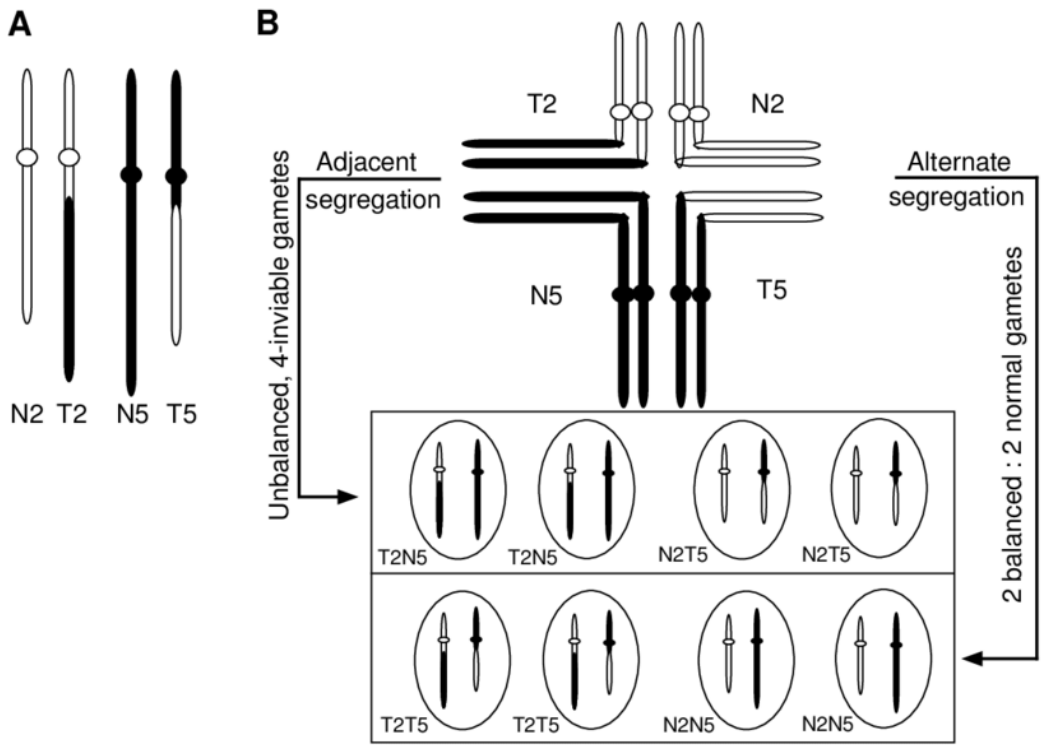


Figure 10 *Ségrégation méiotique des chromosomes dans une cellule portant une translocation hétérozygote. (**A**) Chromosomes transloqués et non-transloqués de l'hétérozygote. (**B**) Quadrivalent méiotique formé par l'appariement des segments homologues des chromosomes. La ségrégation adjacente forme des spores où une partie de l'information génétique est perdue. La ségrégation alterne donne des spores viables dont la moitié transmettra la translocation. Au final, seules 50% des spores sont théoriquement viables (Source image (Ray et al., 1997)).*

Comme nous l'avons vu dans le paragraphe précédent, les réarrangements chromosomiques ne contribuent que partiellement au phénomène de spéciation. Nous présentons dans les paragraphes suivants les autres phénomènes contribuant à la mise en place de l'isolement reproductif afin d'avoir une vision plus globale des mécanismes de spéciation.

Un premier facteur favorisant à l'isolement est l'incompatibilité de Bateson-Dobzhanski-Müller (BDM), c'est-à-dire l'incompatibilité entre des allèles fixés dans des populations indépendantes, qui ont un impact phénotypique négatif lorsqu'ils sont présents chez le même individu. Ce mécanisme a été très largement investigué, notamment chez les *Saccharomyces*. Il a été démontré que des croisements intra-spécifiques de souches de *S. bayanus, S. cariocanus, S. kudriavzevii, S. mikatae S. paradoxus, S. cerevisiae* fournissent une haute proportion de spores viables tandis que les diploïdes interspécifiques ont une viabilité extrêmement faible. L'induction de la tétraploïdie chez ces diploïdes a restauré la viabilité des spores à un niveau équivalent à celui des croisements intra-spécifiques, éliminant toute éventualité d'incompatibilité BDM dominante. De même le remplacement individuel de chromosomes de *S. cerevisiae* par des chromosomes de *S. paradoxus* n'a permis d'identifier aucune incompatibilité BDM récessive (Greig, 2007, 2009; Greig et al., 2002). Récemment, l'investigation exhaustive dans différentes conditions de croissance de la viabilité méiotique de spores obtenues de croisements issus de 27 isolats de *S. cerevisiae* provenant de niches écologiques différentes a permis d'identifier des cas d'incompatibilités environnement-spécifiques, soulignant l'impact environnemental dans la mise en place de barrières interspécifiques (Hou et al. 2015). En outre, la même étude explicite un cas d'incompatibilité BDM entre une mutation non-sens dans un gène mitochondrial porté par le génome nucléaire et un gène d'ARNt suppresseur de cette mutation. De manière analogue, une combinaison spécifique d'allèles ségrégants de Ku70 et Ku80 chez *S. paradoxus* rend le complexe Ku non fonctionnel (Liti et al., 2009). Il existe par ailleurs des exemples d'incompatibilité entre le génome nucléaire et le génome mitochondrial. Un exemple chez la souris a récemment été décrit (Ma et al., 2016). Dans cette étude, les auteurs identifient une incompatibilité nucléo-cytoplasmique en échangeant le

noyau de zygotes de la sous-espèce *Mus m. domesticus*, lignée B6 (avec l'ADN mitochondrial de B6) avec le noyau de zygotes conplastiques provenant de l'introgression de seulement 0.13% du génome de la sous-espèce *Mus m. musculus*, lignée PWD dans un fond génétique B6 et possédant le génome cytoplasmique de PWD. Les auteurs observent que le génome nucléaire de la souche conplastique est compatible avec l'ADN mitochondrial de PWD mais pas de B6, illustrant un cas d'incompatibilité nucléo-cytoplasmique. Ce choix de croisement leur a permis de restreindre les gènes candidats responsables de l'incompatibilité aux régions de PWD issues de l'introgression dans B6. Au final, les gènes *Immp2l*, *Gfm2*, *Bcl2l1* et *P2rx7*, impliqués dans des processus mitochondriaux ont été retenus comme principaux candidats (Ma et al., 2016).

Un troisième mécanisme a été avancé pour expliquer la mise en place de l'isolement reproductif impliquant la divergence de séquence, le système de réparation de mésappariement et la méiose (Liti et al., 2006; Louis, 2011). Dans ce mécanisme, les chromosomes homologues issus de parents divergés forment des intermédiaires de recombinaison au cours de la méiose pour amorcer des crossing-over, puis des chiasma indispensables au bon déroulement de la ségrégation des chromosomes, mais au lieu de recombiner, le nombre important de mésappariements conduit à la dissociation de l'appariement. Il n'y a donc pas de crossing-over et la ségrégation des chromosomes est anormale (Chambers et al., 1996; Hunter et al., 1996), ce qui rend les spores stériles. Dans l'étude (Liti et al., 2006) les auteurs tracent pour la première fois le taux de gamètes viables en fonction de la divergence de séquence, résultant de manière très intéressante en une relation monotone, décroissante et surtout graduelle, différente de la distinction franche qu'on aurait pu attendre entre inter-fertilité et inter-stérilité et bousculant le concept d'espèces biologiques chez la levure.

3.3.2. Réarrangements chromosomiques et évolution du répertoire de gènes

Les réarrangements chromosomiques non-balancés tels que les duplications, délétions, l'ajout de chromosomes entiers ou la duplication totale du génome sont d'importants contributeurs de l'évolution du répertoire de gènes. Les variations du nombre de copie (CNV) ont été très étudiées chez plusieurs espèces, grâce aux nouvelles technologies de séquençage (NGS) qui ont rendu ce type de polymorphisme très accessible. Chez l'humain, ces études ont permis d'identifier un très grand nombre de CNV dans les génomes des individus, dont beaucoup impliquent des gènes (Sudmant et al., 2015). Bien que de nombreux CNV aient été étudiés pour leur lien avec diverses pathologies (Stankiewicz and Lupski, 2010), un certain nombre d'entre eux ne sont pas associés à un effet délétère. Chez la levure, il a été montré que la duplication de segments précis du génome permet à court terme l'adaptation à des conditions environnementales limitantes ou à un déséquilibre de dosage génétique (Dunham et al., 2002; Gresham et al., 2008; Koszul et al., 2004; Payen et al., 2013; Schacherer et al., 2004) et l'expansion/la réduction du répertoire de gènes à l'échelle évolutive (Gabaldón et al., 2013, 2016; Llorente et al., 2000; Scannell et al., 2011). La duplication et la divergence des paralogues et ohnologues a été très étudiée chez les levures *Saccharomycetaceae*. En effet ce groupe d'espèces a subi une duplication totale du génome (Wolfe and Shields, 1997) suivie d'une perte rapide de certains des gènes dupliqués (Scannell et al., 2006, 2007a). Dans l'étude (Marcet-Houben and Gabaldón, 2015), les auteurs s'intéressent aux mécanismes à l'origine de la duplication totale chez les *Saccharomycetaceae*. Ils montrent que l'approche utilisée précédemment chez les vertébrés et les plantes, consistant à dater la divergence des ohnologues pour vérifier qu'ils ont bien été dupliqués dans la branche attendue de l'arbre des espèces, fournit des résultats surprenants : de nombreux paralogues semblent avoir été dupliqués avant l'événement de duplication totale du génome. Les auteurs ont évalué en détail la pertinence de leur approche méthodologique et montrent que cette anomalie est due au fait qu'une hybridation à l'origine de l'événement de doublement résulte en des profils phylogénétiques évoquant une duplication ayant eu lieu dans la lignée de l'ancêtre commun des deux espèces hybridées. Ils suggèrent que ces deux événements rares (hybridation et doublement du génome) ne sont pas dus à une simple coïncidence et proposent au final deux modèles pour expliquer leur lien de

causalité (Marcet-Houben and Gabaldón, 2015; Wolfe, 2015). Brièvement, le premier modèle propose que deux cellules diploïdes d'espèces distinctes aient formé un allotétraploïde. Les auteurs démontrent que la colinéarité des génomes des espèces parentales impliquées dans ce croisement ont rendu possible la recombinaison massive et la perte de gènes, conduisant à l'observation selon laquelle le génome final semble effectivement issu d'un simple doublement chez un diploïde. Dans le second scénario, les auteurs proposent que les cellules haploïdes de deux espèces ont formé un allo-diploïde. La duplication totale du génome aurait permis la stabilisation du génome de cet hybride et la restauration de la viabilité sexuelle.

L'impact des réarrangements balancés sur l'évolution du répertoire génétique ne saurait être négligé. Certes la plupart des réarrangements balancés fixés au cours de l'évolution ont des points de cassure dans les régions intergéniques (Berthelot et al., 2015; Peng et al., 2006; Poyatos and Hurst, 2007), les réarrangements impliquant des points de cassure dans des phases codantes étant contre-sélectionnés (Quintero-Rivera et al., 2015; Rowley, 1998). La fixation de réarrangements chromosomiques ayant détruit des phases codantes n'est pourtant pas un phénomène rare. Un exemple frappant chez les levures *Lachancea* montre que 15% des réarrangements balancés fixés au cours de l'évolution de ce clade ont détruit des phases codantes (Vakirlis et al., 2016).

Les réarrangements balancés peuvent aussi conduire à la formation de gènes chimériques par la fusion de séquence codante, ou par le remplacement de séquences régulatrices en amont des gènes, ce qui modifie leur expression. Chez l'humain, cela s'observe surtout dans les cellules cancéreuses, par exemple dans le cas du chromosome de Philadelphie, évoqué plus haut. Chez la levure, un exemple connu de translocation associée à un phénotype par altération de l'expression d'un gène est la translocation réciproque observée dans certaines souches de vin qui place le gène *SSU1* sous le contrôle du gène *ECM34*. Les souches porteuses de cette translocation présentent un phénotype de résistance accrue aux sulfites (Perez-Ortin, 2002). Les inversions jouent également un rôle adaptatif comme démontré chez la levure (Naseeb and Delneri, 2012) où des inversions dans le cluster DAL permettent une meilleure croissance sur l'allantoïne, ou encore chez la drosophile et le moustique où des inversions corrèlent avec les conditions d'habitat (Ayala et al., 2014; Tobler et al., 2014).

Un exemple particulièrement spectaculaire d'adaptation liée à des inversions chromosomiques est celui des papillons sympatriques amazoniens *Heliconius numata* et des papillons du genre *Melinaea* (Joron et al., 2011). L'espèce *Heliconius numata* présente différents motifs colorés d'ailes, chaque motif étant la copie du motif observé chez une des espèces de *Melinaea*. Les différents motifs d'*Heliconius numata* sont contrôlés par un supergène appelé Locus P, qui est un locus contenant plusieurs gènes hérités de manière intacte, c'est-à-dire ne subissant pas de recombinaison, ce qui permet de conserver des combinaisons alléliques et de réduire la production de génotypes recombinants (Joron et al., 2006). Le locus P présente neuf allèles différents respectant une hiérarchie linéaire de relations de dominance. Chaque allèle du supergène est caractérisé par une combinaison d'inversions différentes (Figure 11). Chaque allèle du locus P détermine un motif mimant les ailes d'une espèce de *Melinaea*. Cet exemple illustre comment les inversions chromosomiques peuvent se trouver au cœur de mécanismes adaptatifs. Le renforcement mutuel du même signal d'avertissement « honnête » des espèces du genre *Melinaea* et de *H. numata* (« Nous ne sommes pas bons à manger ! ») est à leur avantage : les prédateurs apprennent plus facilement à éviter les deux espèces, ce qui en fait un exemple de mimétisme Müllérien (Brown and Benson, 1974).

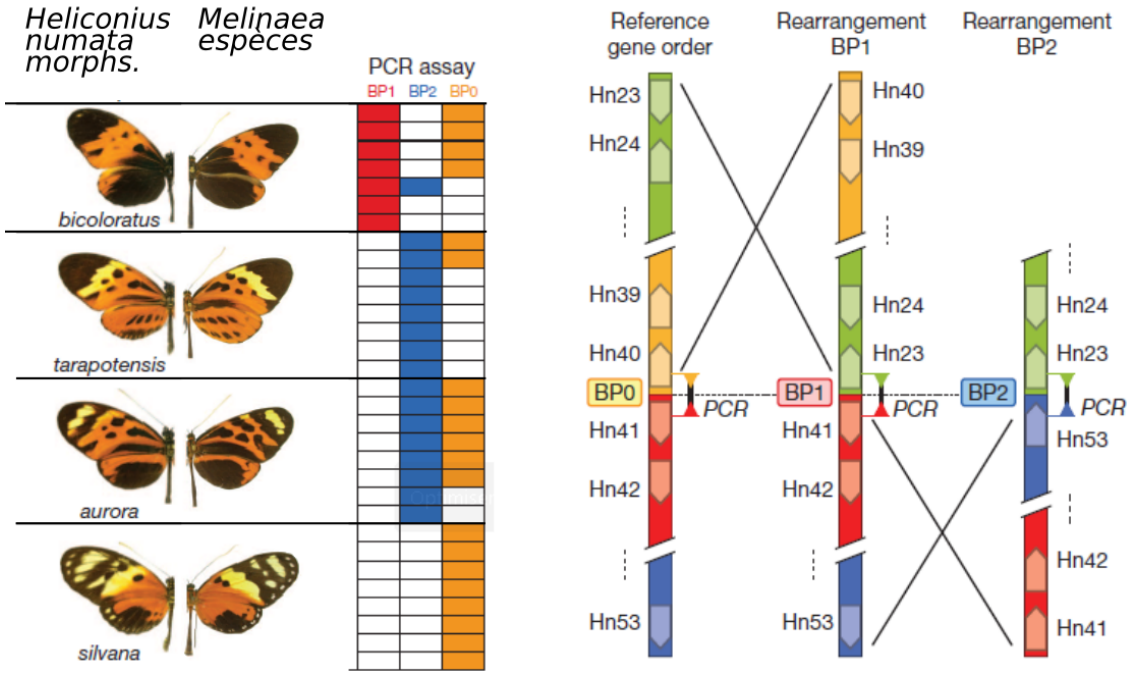


Figure 11 Réarrangements chromosomiques du locus P et polymorphisme d'*Heliconius* dans les populations naturelles. **A gauche** : polymorphisme de *H. numata* associé à différents allèles du locus P. Chaque allèle du locus P permet de mimer une espèce de *Melinaea*. **Au centre et à droite :** points de cassure détectés par PCR et réarrangements du locus P associés à chaque allèle (Adapté de (Joron et al., 2011)).

3.3.3. La relation génotype-phénotype : des SNP aux SV

La relation génotype-phénotype a d'abord été étudiée en rapport avec les mutations ponctuelles. Avec le développement des puces à ADN, puis des techniques de re-séquençage, deux techniques puissantes ont été développées afin d'identifier les régions contribuant aux traits phénotypiques : l'étude d'association à l'échelle du génome ou « Genome Wide Association Study » (GWAS) et l'analyse QTL pour « Quantitative Trait Loci mapping » (QTL mapping). Le fait que la levure ait été le premier eucaryote séquencé, ainsi que les avantages intrinsèques de ce modèle comme le fort taux de recombinaison méiotique, le temps de génération court et le génome de taille modeste ont fait de cet organisme un modèle très important dans l'étude de l'association génotype-phénotype. L'utilisation de ces approches a permis d'étudier de manière très poussée le rôle que jouent les mutations ponctuelles sur la diversité phénotypique. En particulier, la recherche des sources de l'« héritabilité manquante » a beaucoup bénéficié de l'efficacité du modèle levure. La recherche systématique de variants à l'origine de traits mendéliens dans de nombreux fonds génétiques a montré que la transmission d'un trait mono-génique, mendélienne dans un fond génétique, ne l'est pas forcément dans d'autres fonds. Ceci montre une complexité insoupçonnée dans l'interaction de la mutation avec le fond génétique (Hou et al., 2016). Une étude a démontré que l'analyse QTL permet d'identifier un grand nombre d'interactions non-additives impliquant de nombreux loci, expliquant ainsi selon le trait considéré 0 à 50% de l'héritabilité manquante, pourvu que le nombre de spores analysées soit suffisamment grand. Ceci implique que l'héritabilité manquante résiduelle est due à de nombreux autres loci dont l'effet est faible (Bloom et al., 2013). D'autre études montrent qu'une part significative de l'héritabilité manquante provient du manque d'information sur le contexte environnemental d'où proviennent les souches analysées (Wei and Zhang, 2017). Ainsi chez les organismes modèles, l'environnement est contrôlé ou connu et l'héritabilité manquante est moindre. Chez l'humain en revanche, l'environnement est difficilement contrôlable et l'héritabilité manquante est grande (Eichler et al., 2010). Malgré ces éléments de réponse sur l'origine de l'héritabilité manquante, il a été démontré que notre compréhension des traits complexes n'est pas seulement entravée par des facteurs non héritables comme l'environnement ou l'épigénétique, mais aussi par notre manque de connaissance concernant les élément génétiques des traits en question (Albert and Kruglyak, 2015; Mackay et al., 2009). Ce constat est supporté

par des études GWAS dans lesquelles les loci identifiés expliquent relativement peu l'héritabilité des traits complexes (Hindorff et al., 2009; Manolio et al., 2009). Plusieurs explications ont été suggérées pour expliquer l'héritabilité manquante, notamment l'implication d'un grand nombre de variants dont l'effet est petit, ou bien d'autres formes de polymorphisme comme les réarrangements chromosomiques, qui ont été le plus souvent exclus de ce type d'études. Il est aujourd'hui reconnu que les réarrangements chromosomiques touchent un nombre de bases du génome humain supérieur aux mutations ponctuelles (Kloosterman et al., 2015) et qu'ils contribuent de manière majeure à la diversité génétique entre les individus (MacDonald et al., 2014). Les CNV ont reçu une attention importante au cours de la dernière décennie, et leur impact phénotypique est aujourd'hui relativement bien décrit. En revanche, à ce jour, l'impact phénotypique dû aux réarrangements balancés est resté assez peu exploré.

La méthode la plus souvent entreprise pour étudier l'impact phénotypique d'un réarrangement chromosomique en l'absence d'autres formes de polymorphisme a consisté à induire des réarrangements chromosomiques de manière à disposer de souches différant uniquement par la variation structurelle d'intérêt mais par ailleurs isogéniques.

3.4. L'ingénierie des génomes et le déchiffrage de l'impact des réarrangements chromosomiques

L'approche classique pour introduire des réarrangements balancés est de générer des cassures double-brin dans le génome et de favoriser la réparation de ces cassures soit par recombinaison homologue, soit par NHEJ. Toutefois, introduire des cassures double-brins et des réarrangements est longtemps resté un défi.

Dans les études pionnières dans ce domaine, les variants structuraux étaient obtenus en induisant des cassures double-brins avec l'enzyme I-SceI dans deux moitiés de phase codantes d'un gène de sélection, préalablement introduites à des loci différents, dont on sélectionnait la réparation (Fairhead et al., 1996). Cette technique des « *split-markers* » a notamment été employée dans des cellules de souris afin d'évaluer le taux de formation de translocations réciproques à la suite de l'induction de CDB dans deux allèles du marqueur *FN2* (conférant la résistance au G418) sur deux chromosomes différents (Richardson and Jasin, 2000). Les auteurs montrent qu'une haute fréquence de colonies résistantes est obtenue par la simple induction de CDB par I-SceI dans les deux allèles de *FN2* (4.10^{-4} colonies par cellule viable) et qu'une CDB dans un seul des deux allèles de *FN2* est suffisante pour obtenir 5.10^{-6} colonies résistantes par cellule viable. Le séquençage des jonctions des souches résistantes a montré que près de 80% des colonies obtenues ne présentaient pas de translocation. Parmi ces souches, la plupart (56%) avaient réparé les cassures par conversion génique entre les allèles de *FN2*. De façon intéressante, une des deux jonctions présentait un marqueur fonctionnel, l'autre site d'I-SceI étant retrouvé intact. Dans les 9 souches présentant une translocation, le marqueur correspondait en taille et en séquence au produit attendu par la recombinaison des séquences partagées entre les deux allèles de *FN2*. Les deux autres extrémités portaient toutes la signature de la réparation par NHEJ (indels).

Plus tard, l'utilisation de l'endonucléase I-SceI a été combinée à une cassette contre-sélectionnable (« *COunter-selectable Reporter* », CORE) dans le cadre de la technique *delitto-perfetto* (de l'italien « crime parfait »), ce qui permet de générer des translocations réciproques sans laisser d'autres traces d'édition (Storici and Resnick, 2003, 2006). La technique *delitto-perfetto* consiste dans un premier temps à introduire selon une approche classique, la cassette CORE dans le génome au locus où l'on souhaite induire des CDB. La cassette CORE contient un gène permettant de sélectionner son intégration dans le génome, un gène contre-sélectionnable (le plus souvent *KlURA3* pour une utilisation chez la levure), le gène codant I-SceI sous un promoteur inductible et un site de reconnaissance d'I-SceI. Dans un second temps, l'expression

d'I-SceI est induite, ce qui provoque des CDB dans les cassettes CORE. La transformation d'oligonucléotides comme modèle de réparation permet d'introduire des mutations ponctuelles, ou des réarrangements chromosomiques. Cette technique a été utilisée pour induire une translocation entre le chromosome VII et V de *S. cerevisiae* avec une fréquence de 10^{-6} cellules transloquées (Storici and Resnick, 2006). La technique *Delitto-perfetto* a essentiellement été utilisée pour introduire des mutations ponctuelles avec une efficacité de 20%.

A la suite de l'observation selon laquelle les translocations observées entre les génomes des *Saccharomyces* ne permettaient pas d'expliquer les spéciations observées dans l'arbre des espèces (Fischer et al., 2000), une autre étude a cherché à quantifier la contribution de ces réarrangements à l'isolement reproductif chez les *Saccharomyces* (Delneri et al., 2003). Dans cette étude, la recombinaison Cre-Lox a été employée de manière à générer des souches de *Saccharomyces cerevisiae* dont le génome soit colinéaire au génome de deux souches de *S. mikatae*. Les croisements naturels de ces deux espèces fournissent des spores non-viables tandis que les souches réarrangées croisées avec *S. mikatae* voyaient la viabilité de leurs spores partiellement restaurée. Cette expérience a donc permis de montrer la contribution directe des réarrangements chromosomiques au phénomène de spéciation et a suggéré une contribution d'autres facteurs, notamment des incompatibilités, à l'isolement reproductif.

Dans l'étude (Naseeb and Delneri, 2012), les auteurs ont réarrangé le cluster de gènes DAL de *Saccharomyces cerevisiae*, codant pour la voie de dégradation de l'allantoïne, pour le rendre colinéaire à celui de *Naumovozyma castellii*, bien moins résistante à l'allantoïne et dont le cluster DAL présente deux inversions emboîtées. Ces inversions ont été reconstruites dans le génome de *Saccharomyces cerevisiae* selon deux stratégies (en inversant d'abord le petit segment, puis le grand segment du cluser DAL, ou l'inverse). Dans les résultats de la première stratégie, les auteurs montrent que l'inversion d'un gène du cluster (*DAL2*) diminue de 90% l'expression de ce gène et également l'expression de ses voisins *DAL1* et *DAL4* (de 50% et 90% respectivement). Toutefois, la contribution de l'inversion sur le niveau d'expression des gènes *DAL* n'est pas très claire car les auteurs montrent que l'expression du gène *DAL2* est diminuée de moitié dans la souche contrôle, c'est-à-dire dans la souche présentant les sites LoxPsym mais pas d'inversion, suggérant l'impact de l'insertion de cette « cicatrice » génomique sur le niveau de transcription de *DAL2*. De plus, à l'issue de la deuxième stratégie, l'impact phénotypique des inversions sur le niveau de résistance à l'allantoïne n'est pas très marqué.

D'autres études apportent une vision plus globale de la contribution phénotypique des réarrangements chromosomiques balancés sur la croissance végétative et la viabilité méiotique.

Chez les levures à fission, plusieurs souches isolées présentant peu de divergence nucléotidique (~0.5%) et très peu de différences phénotypiques ont toutes été nommées *S. pombe* du fait de ces ressemblances alors mêmes que ces souches présentent 12 inversions qui rendent la totalité des spores issues de ces croisements non-viables (Brown et al., 2011). La première explication de cette diversité caryotypique notable a été proposée par Brown et collègues, qui suggèrent que ces organismes vivent en petites communautés sans faire de croisements en dehors de leur population. Plus récemment, dans l'étude (Avelar et al., 2013), la recombinaison Cre-Lox a été utilisée pour induire dans un fond génétique unique, deux inversions et huit translocations observées dans les isolats naturels de *S. pombe*. Les auteurs ont ensuite mesuré la viabilité méiotique et le taux de croissance mitotique dans différentes conditions environnementales. Bien que les réarrangements chromosomiques présentent un effet délétère durant la reproduction sexuée, les auteurs montrent aussi que certaines souches ont un taux de croissance plus rapide en mitose dans certaines conditions environnementales, révèlant un phénomène de pléiotropie antagoniste. Ces résultats sont très intéressants car ils apportent une autre explication au maintien des variations structurelles des chromosomes au sein des populations et supportent en outre le fait qu'une simple translocation ou inversion peut sévèrement impacter la ségrégation des chromosomes en mitose

comme précédemment montré chez *S. cerevisiae* (Liti et al., 2006). Les deux explications, (Brown et al., 2011) comme (Avelar et al., 2013) permettent de concevoir comment un réarrangement chromosomique délétère en méiose parvient à se propager et potentiellement à atteindre la fixation dans la population.

Dans l'étude (Naseeb et al., 2016), les auteurs ont exploré l'impact phénotypique de grandes inversions induites à l'aide de Cre-Lox dans le génome de *S. cerevisiae* de manière à mimer des réarrangements induits par la recombinaison d'éléments Ty1. Seize souches ont été générées, contenant des inversions péri-centriques et para-centriques de tailles variables (respectivement 26 à 567kb et 167 à 800kb). De manière surprenante, les auteurs rapportent une faible corrélation négative entre la viabilité des spores obtenues par méiose et la taille des inversions. Une corrélation négative plus forte a été observée entre le nombre de *hotspots* de recombinaison méiotique et le pourcentage de spores viables. D'autre part, les auteurs montrent que les inversions affectent non seulement le niveau de transcription des gènes présents dans le segment inversé mais également d'un grand nombre de gènes dispersés dans l'ensemble du génome. Pour autant, le taux de croissance mitotique des souches réarrangées n'est impacté que dans 12 conditions nutritives sur les 94 testées, suggérant un fort effet « tampon » de l'expression des gènes sur l'impact potentiel des variations structurelles induites par les Ty1.

Dans la lignée de la recombinaison Cre-Lox, une nouvelle approche portant le nom de SCRaMbLE-ing (SCRaMbLE signifie « Synthetic Chromosome Rearrangement and Modification by LoxP-mediated Evolution ») utilise la recombinaison Cre-Lox dans des souches de levures possédant des chromosomes synthétiques. Dans ces nouveaux chromosomes, les codons stop TAG ont été remplacés par des codons stop TAA, les régions subtélomériques ont été délétées, ainsi que les introns, les ARNt (relocalisés sur un chromosome circulaire indépendant), les transposons, les cassettes *MAT* silencieuses. En outre, des sites LoxPsym ont été insérés en 5' et en 3' (dans les régions 3' UTR) des gènes non-essentiels (Figure 12). L'induction de *Cre* dans les souches contenant des chromosomes synthétiques génère de très nombreux réarrangements chromosomiques (Annaluru et al., 2014; Hochrein et al., 2018; Jia et al., 2018; Shen et al., 2016b).

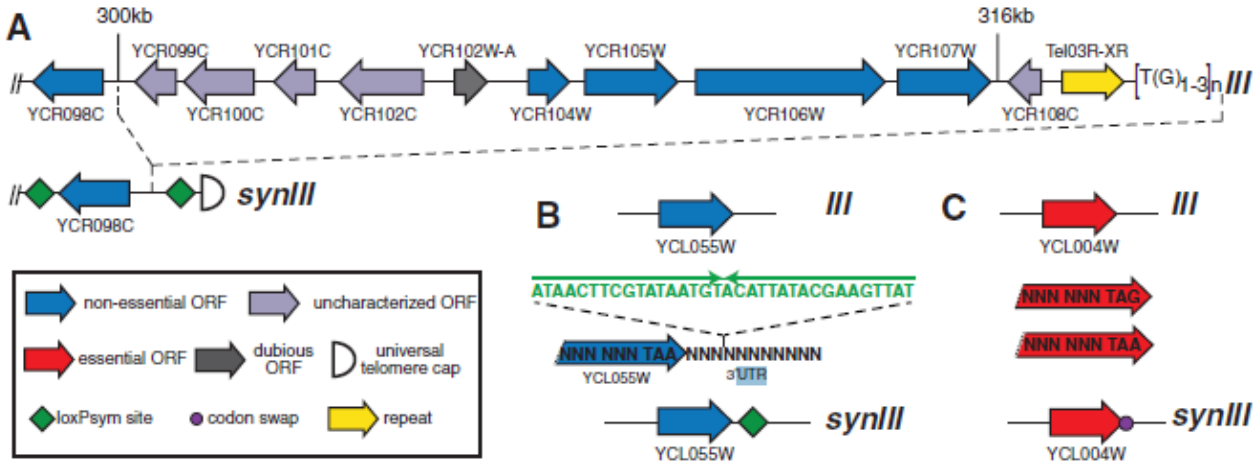


Figure 12 Conception des chromosomes synthétiques utilisés dans l'approche SCRaMbLE-ing. (**A**) exemple d'un subtélomère du chromosome synthétique III. (**B**) Les sites LoxPsym rajoutés en 5' et en 3' UTR des gènes non-essentiels sont indiqués par des losanges verts. (**C**) Les changements de codons sont indiqués par des disques violets (crédit image (Annaluru et al., 2014)).

Cette approche s'est avérée particulièrement efficace pour générer des souches avec des aptitudes métaboliques améliorées (Blount et al., 2018; Jia et al., 2018; Luo et al., 2018a; Shen et al., 2018a). Citons à titre illustratif l'étude (Blount et al., 2018) qui rapporte que la recombinaison Cre-Lox de chromosome synthétique V avec un plasmide portant les gènes de biosynthèse de la violacéine (un pigment) provenant de *Chromobacterium violaceum* a permis d'obtenir une souche 2,3 fois plus productive que la souche

initiale simplement transformée avec le plasmide. Selon une approche similaire, les auteurs ont obtenu après SCRaMbLE-ing de la souche synthétique contenant un plasmide portant les gènes de biosynthèse de la pénicilline, une souche deux fois plus productive que la souche initiale. Enfin, le même processus répété avec un plasmide portant les gènes *XYL1* et *XYL2* indispensables à la production d'éthanol à partir du xylose chez *Scheffersomyces stipitis* a permis d'obtenir une souche synthétique de *S. cerevisiae* avec un taux de croissance moyen de 0.2 h^{-1} ce qui est comparable au taux de croissance de 0,18h^{-1} précédemment rapporté dans une autre souche de *S. cerevisiae* ayant fait l'objet d'intenses recherches pour optimiser cette voie métabolique par mutagénèse dirigée et surexpression.

Une autre technique de restructuration du génome reposant sur une endonucléase température-dépendante (TaqI) a récemment été développée par Muramoto et collaborateurs pour introduire de manière conditionnelle de multiples réarrangements dans le génome d'*Arabidopsis thaliana* et de *S. cerevisiae* (Muramoto et al., 2018). Chez la levure, les auteurs rapportent que chez les diploïdes, 28% des colonies obtenues présentaient des génomes réarrangés après l'induction de TaqI (contre 3% chez les contrôles), alors que 0% des haploïdes sont réarrangés, suggérant que ces derniers sont moins tolérants aux réarrangements chromosomiques du fait de la plus faible redondance de leur matériel génétique. Les auteurs rapportent l'observation de 1,3 SNV/souche parmi les 13 mutants séquencés alors qu'aucun SNV n'est observé chez les souches contrôles. Les auteurs ont de plus évalué le nombre de colonies dont le gène *CAN1* a muté du fait de l'induction de *Taq1* chez une souche *rev3Δ* (*rev3* code pour une ADN polymérase *error-prone*). Le nombre de colonies est significativement moindre dans la souche *rev3Δ* par rapport à la souche sauvage, suggérant qu'une partie des cassures est réparée selon un mécanisme de réparation *error-prone* comme un mécanisme de HR avec de longues extrémités résectées ou la voie MMEJ/alt-NHEJ. De manière intéressante, 40 conversions géniques courtes ont été détectées (7kb en moyenne) confirmant la prompte réparation des CDB par recombinaison homologue. Six translocations non-homologues ont été découvertes au niveau de sites de reconnaissance de TaqI sans mutations additionnelles indiquant que les extrémités sortantes de 2 nucléotides de ces CDB ont probablement pu être directement réparées par c-NHEJ. Deux souches présentaient des translocations médiées par des séquences d'éléments transposables. Cette approche a notamment généré des souches possédant des traits phénotypiques d'intérêt biotechnologique comme l'augmentation de la biomasse (*A. thaliana*) ou la production accrue d'éthanol à partir du xylose (*S. cerevisiae*).

Des méthodes utilisant les protéines doigts de zinc (ZFN) et les nucléases effectrices de type activateur de transcription ou « Transcription Activator-Like Effector Nucleases » (TALEN) ont également été développées pour introduire des réarrangements ciblés dans le génome de la levure, du poisson zèbre et dans les cellules de mammifères. Dans l'étude (Brunet et al., 2009) notamment, les auteurs ont utilisé I-SceI et une enzyme ZFN pour induire deux cassures simultanées dans le génome de cellules humaines, fournissant des cellules portant la translocation d'intérêt avec une fréquence de 2.10^{-3}. Le séquençage de 57 jonctions des souches transloquées a révélé que 25% des jonctions avaient été réparées par jointure sans autre altération, les jonctions restantes portaient des délétions à leurs deux extrémités. En outre, 44% des jonctions présentent des séquences de micro-homologie. Dans l'étude (Piganeau et al., 2013), les auteurs ont reconstitué à l'aide de ZFN et TALEN deux translocations oncogéniques dans des cellules humaines, confirmant l'expression des gènes chimériques qui en résultent. Les jonctions des translocations obtenues étaient identiques aux jonctions observées dans les tumeurs de patients. En utilisant la même approche, les auteurs ont réverté la translocation dans ces cellules de patients. Une approche similaire a été utilisée chez le poisson zèbre pour induire des délétions et des inversions (Xiao et al., 2013).

Bien que toutes ces technologies aient fourni de précieux résultats, elles présentent certaines limitations. Certaines reposent sur l'utilisation de marqueurs génétiques qu'il faut introduire au préalable dans le génome à éditer et qu'il faut pouvoir recycler, ou bien elles laissent des « cicatrices » génomiques

(split-markers, delitto-perfetto, Cre-Lox). D'autres sont difficiles à mettre en œuvre et même si elles permettent en principe d'éditer le génome sans introduire de marqueurs ou de cicatrices, leur rendement reste extrêmement faible (ZFN, TALEN). D'autres encore imposent de travailler dans un fond génétique défini (SCRaMbLE). Ce sont les raisons pour lesquelles le développement du système CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated 9) a apporté une véritable plus-value pour le domaine de l'ingénierie des génomes (Alexander, 2018; Doudna and Charpentier, 2014; Fraczek et al., 2018).

3.5. La révolution CRISPR/Cas9

Le système CRISPR/Cas9 initialement adapté des bactéries consiste en une endonucléase encodée par le gène Cas9 de *Streptococcus pyogenes* et d'un court ARN, appelé ARNg, guidant la nucléase vers sa cible génomique. L'ARNg peut être facilement conçu pour cibler tout locus à proximité d'un motif adjacent au protospacer (PAM) c'est-à-dire de la séquence NGG. Cette technologie est désormais utilisée en routine pour induire des cassures double-brin dans une grande diversité d'espèces (Wang and Qi, 2016).

Un descriptif plus détaillé du fonctionnement du système CRISPR/Cas9 est donné dans le paragraphe 2.1.3.2, page 37.

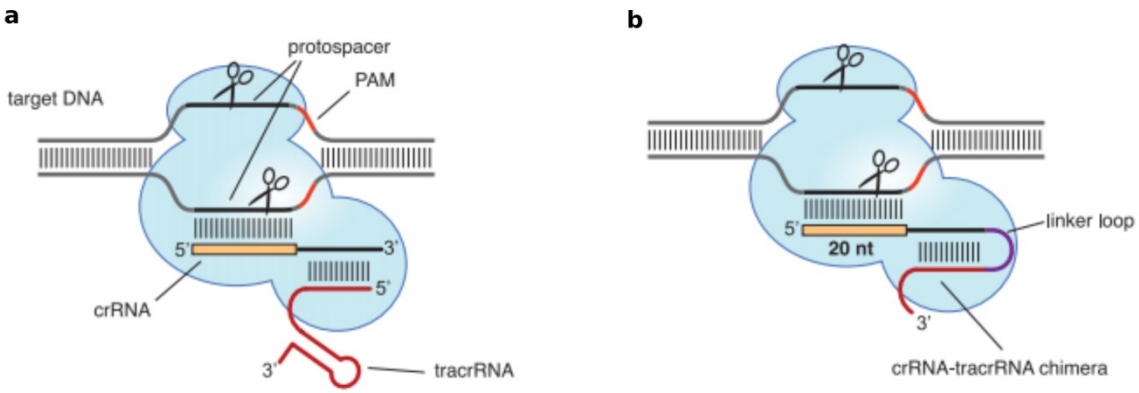


Figure 13 Cas9 est dirigée dans le génome par un ou deux ARN simple brin. Cas9 est représentée en bleu. La séquence PAM, indispensable au fonctionnement de Cas9 est indiquée en orange dans l'ADN cible. (**a**) Dans le système CRISPR-*Cas9* de type II, comme celui de *Streptococcus pyogenes*, Cas9 est guidée vers sa cible génomique par le complexe formé du tracrRNA qui joue un rôle structurel (en rouge) et par le crRNA qui donne la spécificité de la cible (en jaune). (**b**) Dans le système artificiel utilisé le plus souvent, le tracrRNA et le crRNA sont fusionnés en un seul ARN chimérique (crédit images (Doudna and Charpentier, 2014)).

Chez la levure, les CDB induites par CRISPR/Cas9 peuvent être réparées avec une grande efficacité en fournissant des fragments d'ADN homologues aux sites de cassure aux cellules transformées. Sur ce principe, des travaux ont montré qu'il était possible d'éditer les génomes avec une grande précision et une grande efficacité. Dans l'étude (DiCarlo et al., 2013), les auteurs démontrent qu'il est possible d'induire simultanément des mutations ponctuelles avec près de 99% d'efficacité dans deux loci différents du génome de la levure en transformant des cellules exprimant *Cas9* de manière constitutive avec des oligonucléotides guidant la réparation des CDB par RH et un plasmide portant des cassettes d'expression d'ARNg. En outre, les éditions ont été réalisées sans introduire de marqueur génétique dans le génome, la sélection des transformants se faisant uniquement sur l'acquisition du plasmide transformé.

Le nombre d'éditions génomiques simultanées a rapidement augmenté : en 2015 l'utilisation d'un site de

clonage USER a été rapportée pour cloner en une étape jusqu'à 5 ARNg dans un plasmide (Jakočiūnas et al., 2015). La même année, CRISPR-Cas9 a été utilisé pour introduire des délétions multiples et pour intégrer en une étape jusqu'à six gènes dans le génome de *S. cerevisiae* (Mans et al., 2015). Dans l'étude (Bao et al., 2015), les auteurs ont employé un système assez proche du système original (voir Figure 13a) pour muter plusieurs gènes. En résumé, un premier plasmide porte Cas9 et une cassette d'expression de la partie structurelle de l'ARNg permettant son interaction avec Cas9 (le tracrRNA). Un deuxième plasmide contient un locus CRISPR constitué d'une alternance de spacers et de repeats. Chaque spacer est constitué d'une séquence cible qui sert à guider une CDB fusionnée à une séquence donneur qui servira à réparer la cassure en question. La transcription de ce locus produit un pré-crRNA qui est maturé en crRNA matures par les RNAses de *S. cerevisiae*. Chaque crRNA peut alors former un complexe avec un tracrRNA et Cas9 et induire les CDB ciblées. Les séquences « donneur » portées par le locus CRISPR servent alors de modèle de réparation aux CDB. La technique permet de muter jusqu'à trois loci en une étape, mais guère plus car il y a un fort effet de position des crRNA dans le locus CRISPR ainsi conçu : le cinquième crRNA du locus a une efficacité proche de zéro.

Depuis l'étude (DiCarlo et al., 2013), illustrant la grande performance du système CRISPR/Cas9 pour induire des mutations ponctuelles dans le génome de *S. cerevisiae*, le système CRISPR/Cas9 a été adapté au format haut débit pour l'édition parallèle massive du génome. Le principe de cette approche, représenté Figure 14, consiste à transformer une population de cellules exprimant *Cas9* avec un mélange de plasmides portant chacun une combinaison unique d'un ARNg et d'une séquence servant de modèle aux CDB induites par l'expression de l'ARNg. En plaçant le mélange de cellules ainsi transformées dans des conditions de croissance compétitives et en suivant au cours du temps par séquençage la fraction de la population que représente chaque variant qu'on a induit, on peut déterminer les variants avantageux et les variants défavorables.

Ces approches sont extrêmement puissantes car elles permettent d'analyser l'impact phénotypique relatif de milliers de variants en parallèle. On peut ainsi tester l'impact phénotypique d'un type de mutations ponctuelles à l'échelle du génome entier comme dans l'étude (Sadhu et al., 2017), focalisée sur l'impact phénotypique des codons stop prématurés (« premature-termination codons», PTC), ou dans l'étude (Sharon et al., 2018) dans laquelle les auteurs testent un total de 32000 combinaisons ARNg-donneur pour induire en parallèle dans le génome de la souche de *S. cerevisiae* de référence (S288C), 16000 SNP et indels observés dans une souche de vignoble (RM). Avec cette approche, on peut également choisir de concentrer les mutations ponctuelles au niveau d'un locus d'intérêt comme dans l'étude (Roy et al., 2018) dans laquelle les auteurs ont saturé le gène essentiel *SEC14* en mutations non-synonymes. Ce gène constitue une cible potentielle de traitement antifongique. Ces quelques études ont d'ores et déjà fourni d'impressionnants résultats qui sont brièvement résumés ci-dessous.

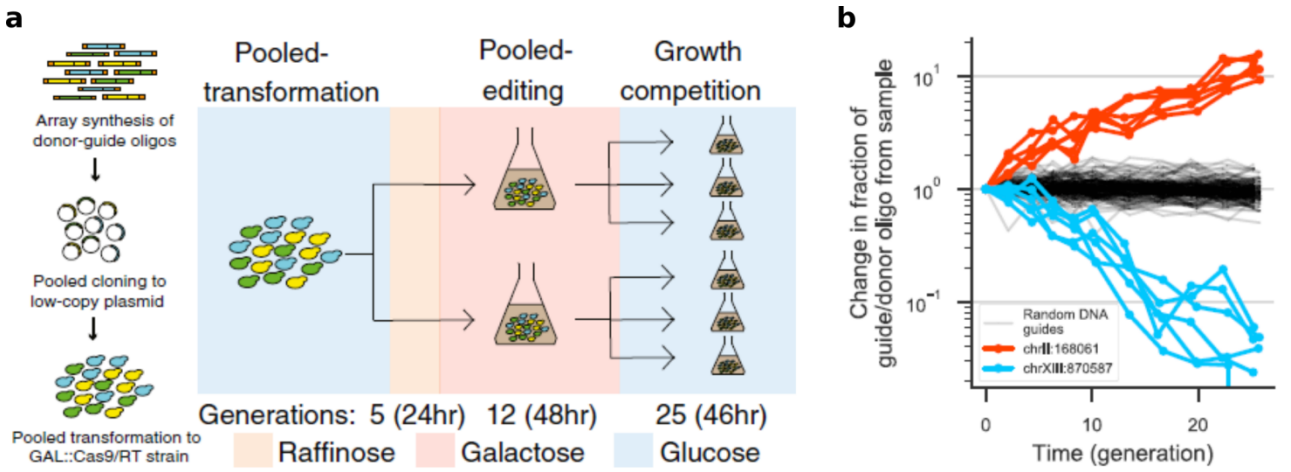


Figure 14 Principe de l'édition CRISPR « en pool » pour mesurer l'impact phénotypique de milliers de variants génétiques indépendants. (**a**) un ensemble de fragments d'ADN synthétiques contenant une combinaison de cassettes d'expression d'ARNg et de séquences « donneur » est cloné dans des plasmides. Les plasmides sont transformés en pool dans la levure. L'expression de *Cas9* est induite (dans le cas présent par la présence de galactose) puis les cellules sont mises en compétition dans un ou plusieurs milieux différents. (**b**) en séquençant les codes-barres associés à chaque combinaison d'ARNg-donneur et/ou en séquençant directement ce construit, on peut mesurer la proportion relative des différents variants dans la population. Les réplicats d'un variant ayant un effet phénotypique avantageux sont en orange. Inversement, un variant délétère est en bleu. Les variants neutres sont en noir (Crédits images (Sharon et al., 2018)).

L'induction systématique de PTC dans 1030 gènes considérés comme essentiels dans la souche de laboratoire de *S. cerevisiae* et une souche NMD-déficiente a permis de montrer que c'est bien la présence de PTC qui est délétère et non la dégradation des transcrits présentant un codon stop par NMD. Il a en outre été montré que les PTC sont moins délétères quand ils sont situés à moins de 37 codons de la fin de la phase codante et lorsqu'ils se trouvent en dehors de domaines annotés ou de domaines conservés. Près de la moitié des gènes, codant pour la plupart des enzymes, ne semblent tolérer aucun PTC situés même très proches de la fin de la phase codante tandis que les gènes associés à l'épissage des ARNm sont plus tolérants. Les auteurs de l'étude (Sadhu et al., 2017) se sont intéressés aux 16 gènes les plus tolérants en PTC et ont montré que pour la plupart, ces gènes ont été annotés comme essentiels du fait de leur caractère indispensable dans la souche dans laquelle ils ont été annotés initialement (auxotrophe pour la leucine et l'uracile) alors que leur délétion dans une souche prototrophe n'est pas létale. Un exemple assez marquant concerne *CWC24*, un gène très conservé du spliceosome, dans lequel un domaine protéique entier peut être perdu sans effet phénotypique (Sadhu et al., 2017).

Dans l'étude (Sharon et al., 2018), les auteurs ont reproduit un grand nombre de SNP d'une souche de vignoble dans la souche de référence. Plus de 170 variants présentaient une différence de *fitness* supérieure à 1% et 17 d'entre eux avaient un impact supérieur à 5%. Ces résultats suggèrent que ces variants présentent un avantage dans certains environnements en étant délétères dans d'autres car des variants avec de tels impacts seraient rapidement fixés dans les populations naturelles s'ils n'étaient pas contrebalancés par des effets négatifs. De manière surprenante, les 23 variants présentant les effets phénotypiques les plus marqués sont localisés dans des régions promotrices, alors que plusieurs études précédentes rapportaient essentiellement des variants dans des phases codantes. Globalement, les variants significatifs sont essentiellement localisés dans des promoteurs divergents et dans une moindre mesure dans les promoteurs unidirectionnels. En outre 33% des variants ayant un effet phénotypique marqué affectent des sites d'interaction de l'ADN avec des facteurs de transcription connus. Les auteurs identifient d'autre part 156 variants synonymes et 95 variants faux-sens significatifs bien que les variants significatifs affectant des phases codantes sont globalement sous-représentés par rapport à l'ensemble des variants testés. Ces résultats remettent en question l'assomption générale selon laquelle les mutations synonymes sont neutres. Sharon et collègues se sont ensuite demandés si l'adaptation des souches étudiées (S288C et RM) à leur niche respective serait visible dans le type de séquences présentant des variants significatifs. De

manière frappante, les gènes associés au GO-term « traduction cytoplasmique » (essentiellement des gènes ribosomaux) et les gènes impliqués dans la croissance sous forme de pseudo-hyphes sont fortement enrichis en variants significatifs au niveau de leur promoteur. Enfin, les auteurs ont cherché à déterminer si des variants proches tendent à avoir un impact phénotypique positif ou négatif. De manière frappante, les auteurs montrent que 98% des variants présents dans un intervalle de 50pb les uns des autres tendent à influencer un seul et même allèle.

Récemment les auteurs de l'étude (Roy et al., 2018) ont développé une approche appelée MAGESTIC permettant de saturer en substitutions non-synonymes des gènes d'intérêt. Les auteurs se sont intéressés en particulier au gène essentiel *SEC14* qui constitue une cible potentielle de traitement antifongique. La difficulté à saturer un gène essentiel en mutations non-synonymes provient du fait qu'une mutation absente du pool de cellules transformé peut soit être une édition qui n'a pas fonctionné, soit qui a fonctionné mais qui est délétère. La létalité de la délétion de *SEC14* étant supprimée par la délétion du gène *KES1* ou *CKI1*, les auteurs ont appliqué MAGESTIC dans une souche *KES1*-déficiente afin de pouvoir également obtenir des mutants détrimentaux. Un total de 98,5% des variants prévus (sur un total de 1382) a pu être détecté en séquençant le pool de souches transformées et moins de 0,5% des variants présentaient des indels liés à la réparation par NHEJ dans le locus d'intérêt. En cultivant le pool de cellules en présence de doses sublétales de NPPM (un inhibiteur de *SEC14*), les auteurs ont pu identifier des résidus fonctionnellement importants dans le contact NPPM-SEC14. Ces résidus ont pu être validés par rapport à des études antérieures. En outre, MAGESTIC a permis d'identifier d'autres substitutions d'intérêt. Les auteurs ont recréé manuellement les substitutions identifiées par MAGESTIC et démontré une grande cohérence des résultats phénotypiques de ces souches par rapport aux résultats fournis par le séquençage des codes-barres dans le pool de cellules. En résumé, MAGESTIC permet de tester des centaines voire des milliers de variants génétiques et d'estimer avec une grande précision leur impact phénotypique en utilisant comme proxy la fréquence des codes-barres associé à chacune des combinaisons ARNg-donneur.

La puissance des techniques haut-débit présentées plus haut, reposant sur la grande efficacité du système CRISPR-Cas9 et sur l'aptitude naturelle de la levure à réparer les CDB par recombinaison homologue laisse entrevoir des avancées extrêmement rapides des connaissances sur la relation phénotype-génotype dans un futur proche. Toutefois, ces études se sont intéressées au polymorphisme nucléotidique et à ce jour aucune étude à large échelle n'a tenté de disséquer l'impact phénotypique des réarrangements chromosomiques balancés.

Des techniques reposant sur CRISPR/Cas9 ont néanmoins été conçues pour remanier la structure des génomes. Notamment, Sasano et collègues ont combiné l'utilisation de CRISPR-Cas9 afin d'augmenter l'efficacité de la technique PCS visant à fragmenter un chromosome en chromosomes plus petits (Sasano et al., 2016). Dans la technique PCS, des cellules sont transformées avec deux cassettes d'ADN présentant respectivement une séquence d'homologie avec une des moitiés du locus au niveau duquel on souhaite fragmenter un chromosome ainsi que des séquences « seed » de télomères. Une des cassettes porte en plus un centromère pour le fragment de chromosome qui n'en porte pas. Sasano et collègues montrent que le fait d'utiliser CRISPR-Cas9 pour induire une CDB dans le locus à fragmenter augmente d'un facteur 200 le taux d'obtention de cellules présentant les fissions attendues et permet d'introduire jusqu'à trois fissions simultanées avec une efficacité de 16%.

Inversement, le système CRISPR-Cas9 a été utilisé pour concaténer des chromosomes. Dans l'étude (Shao et al., 2018), les auteurs ont fusionné en 15 étapes les 16 chromosomes de la souche BY4742 de *S. cerevisiae*. A chaque étape, un centromère et deux télomères ont été délétés. Les auteurs ont ensuite établi le profil tridimensionnel des chromosomes des souches BY4741 et de la souche SY14 dont tous les chromosomes sont concaténés, ainsi que de deux souches intermédiaires, SY6 et SY13 présentant respectivement 9 et 2 chromosomes à partir de données de Capture de Conformation Chromosomique (3C) (Figure 15).

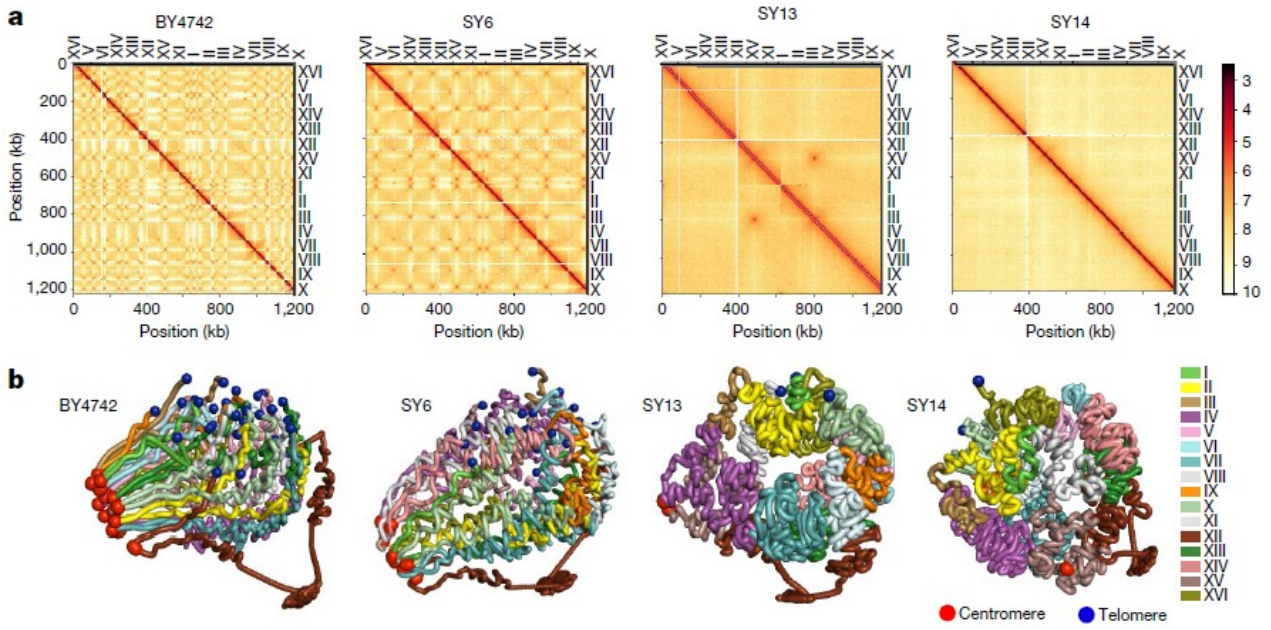


Figure 15 Interactions chromosomiques et structure tridimensionnelle du génome des souches BY4741, SY6, SY13 et SY14. (**a**) Carte normalisée des points de contacts des quatre génomes à une résolution de 10-kb. Les fréquences de contacts sont indiquées par le gradient de couleur : blanc pour les fréquences faibles, rouge pour les fréquences hautes. (**b**) Conformation 3D des quatre génomes (crédits image (Shao et al., 2018)).

Comme on peut le voir, la délétion des centromères et des télomères s'accompagne d'importants changements dans le repli des chromosomes. Les interactions entre centromères restants (en rouge) se renforcent quand le nombre de centromères diminue, tandis que les interactions entre les régions dont les centromères ont été délétés diminuent. Chez BY4742 et SY6, l'organisation du génome est étonnamment proche : les centromères sont regroupés en périphérie du spindle pole body (SBP) qui est le centre organisateur des microtubules chez la levure, tandis que les télomères interagissent avec l'enveloppe nucléaire. Les bras chromosomiques sont tendus entre ces deux points d'ancrage. Chez SY13 et SY14, le génome adopte une conformation très repliée et globulaire. Les auteurs ont ensuite comparé le transcriptome de SY14 à celui de BY4742. De manière très surprenante, seulement 28 gènes présentent une expression différentielle. Sept gènes adjacents aux télomères délétés ont notamment été surexprimés chez SY14, ce qui est cohérent avec la perte de la répression liée à la position de ces gènes et cinq gènes situés à proximité des télomères restants ont été sous-exprimés, confirmant l'effet de l'identité « renforcée » des télomères. De façon intéressante, huit gènes impliqués dans la réponse aux stress de l'ADN ont été surexprimés, suggérant des contraintes sur l'ADN apparues du fait de la concaténation des chromosomes. De plus, le taux de croissance de SY14 est remarquablement comparable à celui de BY4742 quant à l'utilisation de différentes sources de carbone. La souche SY14 présente cependant une petite réduction du taux de croissance à partir de certaines sources d'azote. Les auteurs ont ensuite généré l'équivalent *MATα* des souches BY4742, SY6, SY13 et SY14 par remplacement du locus *MATα* pour évaluer la viabilité méiotique des souches avec des chromosomes concaténés. La viabilité des spores est décroissante avec le nombre de chromosomes fusionnés mais décroit assez lentement : elle est de 98% pour le croisement BY4742/BY4742[a] et de 87,5% pour la souche SY14. Une autre étude rapporte des observations très similaires quant à l'impact de la concaténation des chromosomes sur le transcriptome et sur la viabilité des spores issues de souches concaténés (Luo et al., 2018a). En outre, les auteurs de cet article ont réalisé des croisements entre souches issues de différentes étapes de concaténation. Le nombre de spores et la proportion de spores viables diminuent rapidement quand l'écart augmente entre le nombre de chromosomes des deux parents. Ainsi, les croisements entre la souche à 16 chromosomes et une souche à 12 et 8 chromosomes aboutissent respectivement à un taux de spores viables de 10 et 1%. Ces deux études

illustrent la remarquable robustesse des génomes aux changements de structure chez la levure.

L'utilisation de CRISPR/Cas9 en combinaison avec Cre-Lox dans les cellules de mammifères a également été rapportée pour induire des translocations réciproques (Brunet and Jasin, 2018). Le principe est d'introduire deux CDB dans deux chromosomes distincts, puis de réparer les extrémités des cassures par recombinaison homologue avec un ADN donneur portant un marqueur de sélection. Le marqueur est ensuite perdu par recombinaison Cre-Lox, ne laissant qu'un site LoxP à la jonction chromosomique (Vanoli et al., 2017).

RESULTATS

Problématique

Au cours de cette thèse, nous avons cherché à évaluer l'impact des réarrangements chromosomiques sur l'évolution des génomes de levures selon deux approches.

Nous avons d'une part mis en œuvre une approche *in-silico* visant à reconstruire les génomes ancestraux des levures *Saccharomycotina* pour pouvoir inférer les réarrangements chromosomiques balancés qui ont réarrangé ces génomes depuis leur divergence avec les *Pezizomycotina* il y a 798 à 1166 millions d'années (Hedges et al., 2004). La première étape de ce travail a été d'évaluer les performances du logiciel de reconstruction de génomes ancestraux *AnChro*, développé au laboratoire par Guénola Drillon. Pour ce faire, nous avons comparé la pertinence des reconstructions générées par *AnChro* à celles obtenues en utilisant quatre autres logiciels de reconstruction de génomes ancestraux à la fois sur des données réelles (les génomes des *Lachancea*) et sur des données simulées. Ce travail, présenté dans la partie 4.1, fait partie intégrante de l'article (Vakirlis et al., 2016), disponible en annexe.

Ayant quantifié la pertinence des reconstructions générées par *AnChro*, nous avons cherché à reconstruire les génomes ancestraux des levures *Saccharomycotina* à partir d'un nombre maximal de génomes issus des bases de données publiques, dans le but d'identifier quels réarrangements ont contribué à l'évolution des génomes de ce subphylum. Comme nous l'avons vu dans l'introduction, les *Saccharomycotina* recouvrent un intervalle évolutif très vaste, comparable en termes de divergence protéique à celui des Chordés. Quel signal phylogénétique utiliser pour reconstruire un arbre phylogénétique sur un intervalle évolutif aussi vaste ? Quels critères utiliser pour reconstruire des génomes ancestraux de qualité ? Faut-il utiliser toute l'information disponible ? Combien d'inversions et de translocations ont remanié ces génomes au cours de l'évolution ? A quel rythme s'accumulent les réarrangements dans les différentes lignées ? Les différents types de réarrangements ont-ils tous la même prévalence selon les lignées ? Voilà les questions auxquelles nous tenterons de répondre dans la sous-partie 4.2.

Dans un second pan de ce travail de thèse, nous avons cherché à disséquer, en l'absence d'autres formes de polymorphisme, l'impact phénotypique des réarrangements chromosomiques chez la levure *S. cerevisiae*. Comme nous l'avons vu dans l'introduction, un grand nombre d'études se sont déjà attelées à cette tâche, rendue ardue par le fait que les outils moléculaires ne permettaient pas, jusqu'à récemment, d'induire des réarrangements chromosomiques de manière ciblée et propre dans les génomes. Nous avons donc dans un premier temps développé une nouvelle approche reposant sur le système CRISPR/Cas9 nous permettant d'induire des réarrangements ciblés dans le génome de *S. cerevisiae* avec une efficacité proche de 100% et sans laisser d'autres traces dans le génome, que le réarrangement souhaité.

Cette technique constitue un outil puissant pour étudier l'impact phénotypique des réarrangements chromosomiques dans des fonds génétiques contrôlés. Nous avons utilisé cette technique pour induire et pour défaire des réarrangements ciblés de manière réversible dans le génome de *S. cerevisiae*. Dans un second temps, nous avons utilisé CRISPR/Cas9 pour induire simultanément plusieurs coupures dans le génome de *S. cerevisiae* au niveau de séquences de Ty3-LTR afin d'obtenir une grande diversité de souches présentant des réarrangements multiples. Par le biais de ces expériences, nous avons cherché à répondre aux questions suivantes : l'impact phénotypique d'un réarrangement chromosomique est-il le même dans différents fonds génétiques ? Le génome est il capable de se réparer à la suite de cassures multiples ? Quel est l'impact phénotypique des réarrangements chromosomiques sur la croissance végétative et la viabilité des spores dans des souches dont aucun gène n'a été détruit ou remanié ?

Ces résultats font l'objet d'un article intitulé « *Reshuffling yeast chromosomes with CRISPR/Cas9* », actuellement en revue auprès du journal *Genome Research* et dont je suis premier auteur. Le texte de cet article a été incorporé à cette thèse dans la partie 5, à partir de la page 130.

4. Réarrangements chromosomiques et évolution des génomes chez les *Saccharomycotina*

4.1. Évaluation des performances d'*AnChro*

AnChro a été développé afin de répondre à plusieurs limitations des approches de reconstruction de génomes pré-existantes. *AnChro* repose sur l'utilisation de comparaison de génomes deux à deux, ce qui permet de maximiser la couverture des génomes reconstruits. Pour reconstruire un ancêtre donné, *AnChro* utilise à la fois les adjacences ancestrales validées du fait de leur présence dans plusieurs espèces actuelles et également les adjacences ancestrales aujourd'hui disparues, mais qu'on peut inférer à partir de graphes d'adjacences. *AnChro* est le premier logiciel à combiner ces deux approches.

Dans ce chapitre, nous verrons comment les performances d'*AnChro* ont été évaluées à la fois sur des données réelles que sont les génomes des *Lachancea* et sur des données simulées. Nous avons de plus comparé *AnChro* à quatre autres logiciels de reconstruction : *ANGES*, *GapAdj*, *MGRA* et *PMAG+*. Pour réaliser ces reconstructions, nous utiliserons comme données d'entrée les blocs de synténie.

4.1.1. Obtention des blocs de synténie

AnChro est conçu pour utiliser les données de synténie générées par *SynChro*. De plus, selon les espèces actuelles choisies pour reconstruire un génome ancestral, celui-ci peut être reconstruit de différentes manières. Pour obtenir la meilleure reconstruction possible on sélectionne alors celle qui respecte le mieux des critères objectifs : nombre de *scaffolds*, nombre de gènes retracés, nombre de contradictions structurelles. Cette approche est celle qui a été suivie dans l'étude (Vakirlis et al., 2016).

Cependant, afin de comparer équitablement les performances d'*AnChro* avec d'autres logiciels de reconstruction ne partageant pas ces caractéristiques, nous avons également choisi de reconstruire les ancêtres du clade *Lachancea* en s'affranchissant de l'utilisation de *SynChro* et en ne reconstruisant qu'une seule version des génomes ancestraux selon deux approches. Premièrement, (i) en utilisant des blocs de synténie générés avec *SynChro* et utilisés soit directement par *AnChro*, soit convertis en marqueurs universels pour les quatre autres logiciels de reconstruction de génomes ancestraux. Deuxièmement, (ii) en utilisant des blocs de synténie générés par le logiciel *i-ADHoRe* par la comparaison de deux génomes ou plus (Simillion et al., 2008). Dans ce cas nous avons d'une part identifié la synténie entre paires de génomes pour *AnChro* et d'autre part généré des blocs de synténie universels pour *ANGES*, *GapAdj*, *MGRA* et *PMAG+*. Un schéma récapitulatif est proposé Figure 16.

Avec *AnChro*, nous avons en outre décidé de reconstruire chaque ancêtre selon une seule combinaison de génomes G1 et G2. Nous avons pour cela choisi G1 et G2 partageant le plus faible nombre de blocs de synténie. Notons que ces conditions sont désavantageuses pour *AnChro* qui est conçu pour reconstruire un ancêtre de différentes manières selon les paires d'espèces utilisées comme point de comparaison.

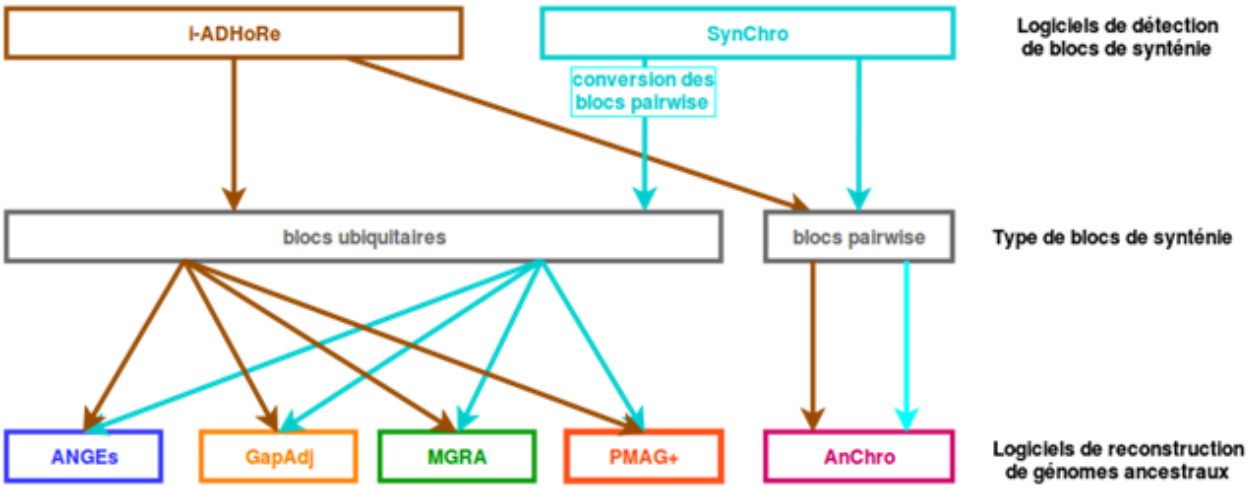


Figure 16 Obtention de blocs de synténie avec *SynChro* et *i-ADHoRe* et conversion des blocs de synténie pour les 5 logiciels de reconstruction *AnChro*, *ANGES*, *GapAdj*, *MGRA*, *PMAG+*.

Nous avons choisi de comparer les performances des logiciels *AnChro, ANGES, GapAdj, MGRA* et *PMAG+* à partir des reconstructions des neuf génomes ancestraux des levures du genre *Lachancea* (notés A1 à A9) en utilisant comme groupe externe le génome de *Zygosaccharomyces rouxii* (Figure 17).

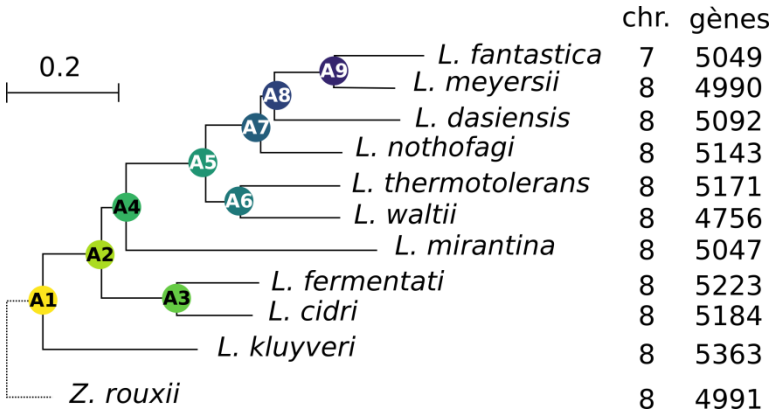


Figure 17 Arbre phylogénétique des *Lachancea*. Les ancêtres sont nommés *A1* à *A9*. Toutes les espèces possèdent huit chromosomes sauf *L. fantastica*. Le nombre de gènes codant des protéines varie entre 4756 (*L. waltii*) et 5363 (*L. kluyveri*) (adapté d'après (Vakirlis et al., 2016)).

4.1.1.1. Détection de la synténie avec *SynChro*

Pour chaque génome à analyser, *SynChro* a besoin de trois sources d'information : (i) les séquences protéiques correspondant à toutes les séquences codantes des génomes à comparer ; (ii) les coordonnées génomiques de chaque élément du génome et leur type (séquence codante, séquence non-codante, centromère, etc.) ; (iii) un résumé des caractéristiques de chaque chromosome (nom, nombre d'éléments génétiques). Ces informations peuvent être extraites des fichiers EMBL de chaque espèce.

SynChro procède alors en plusieurs étapes pour identifier les blocs de synténie. Dans un premier temps, il identifie les meilleurs homologues réciproques ou « *Reciprocal Best Hits* » (RBH) pour chaque paire de génome à analyser. Soient deux protéines p1 et p2 codées respectivement par deux gènes g1 et g2 appartenant respectivement à deux génomes G1 et G2 ; les gènes g1 et g2 sont RBH si la protéine p1 est la plus similaire à p2 parmi toutes les protéines codées dans le génome G2, et réciproquement. A la fin du calcul des scores de similarité, ces derniers sont normalisés entre 0 et 100 et seuls les RBH dont le score est supérieur à 40 sont conservés. De plus, *SynChro* ne conserve que les RBH dont le rapport des longueurs est

strictement inférieur à 1, 3.

Dans un second temps, *SynChro* identifie les RBH conservés en synténie. Le seul paramètre à fixer est Δ', qui correspond à la distance maximale que l'on autorise (en nombre de RBH) entre les ancres, c'est-à-dire la distance maximale à partir de laquelle des RBH ne sont plus considérés comme appartenant au même bloc. Quand Δ est petit, les blocs obtenus ont tendance à être fragmentés. Au contraire, quand Δ augmente des blocs peu espacés ont tendance à fusionner. *AnChro* autorise des valeurs de Δ entre 1 et 6. Les blocs de synténie entre les paires de génomes des *Lachancea* ont été calculés pour toutes les valeurs de Δ.

Les blocs de synténie issus de *SynChro* peuvent être utilisés tel quels avec *AnChro*. En revanche, les autres logiciels de reconstruction utilisent des marqueurs ubiquitaires, c'est-à-dire les blocs de synténie conservés dans tous les génomes comparés. Les blocs de synténie issus de comparaisons de génomes deux à deux générés avec *SynChro* ont donc été transformés en blocs ubiquitaires par transitivité : si un génome G1 et un génome G2 partagent un bloc de synténie et qu'un segment de G2 appartenant à ce bloc est également conservé en synténie avec une région d'un génome G3, alors la région de G1 et de G3 correspondant au segment dans G2 sont elles-mêmes conservées en synténie. Si plusieurs régions d'un même génome sont candidates pour faire partie d'un même bloc, la région partageant le maximum d'ancres avec les régions génomiques déjà choisies est conservée. Les blocs candidats ainsi détectés ont été filtrés : un bloc n'est considéré valide que s'il comprend au moins deux ancres dans chaque génome. De plus les bordures des blocs sont ajustées : on ne conserve que la portion du bloc qui est comprise entre la première et la dernière ancre retrouvées chez toutes les espèces.

La couverture des génomes des *Lachancea* et de *Z. rouxii* par les blocs de synténie obtenus à l'aide de *SynChro* (et comparée avec la couverture de *i-ADHoRe*, voir paragraphe suivant) est représentée Figure 20, page 89.

4.1.1.2. Détection de la synténie avec *i-ADHoRe*

Le logiciel *i-ADHoRe* utilise en données d'entrée des « listes de gènes » récapitulant la position et le brin où se trouvent les gènes sur chaque chromosome de chaque génome ainsi qu'un tableau récapitulant entre chaque paire de génomes l'ensemble des gènes homologues identifiés. Cette liste de gènes homologues peut être obtenue en réalisant un alignement de toutes les séquences codantes d'un génome contre toutes les séquences codantes d'un autre génome, puis en appliquant un filtre pour éliminer les alignements non significatifs. Or, la première étape de *SynChro* consiste à identifier les RBH entre les paires de génome. Nous avons donc utilisé les RBH identifiés par *SynChro* pour générer la table d'homologie.

L'algorithme *i-ADHoRe* identifie tout d'abord des régions homologues colinéaires entre deux génomes en traçant des « gene homology matrix », ou GHM (Figure 18). Cette matrice est obtenue en traçant un tableau à 2 dimensions dans lequel l'entrée des lignes et des colonnes correspondent à l'ordre des gènes dans deux génomes à comparer. Pour chaque cellule du tableau, si l'entrée de la ligne et de la colonne correspondent à des gènes orthologues, le résultat de la cellule vaut 1. Si les gènes des deux génomes ne sont pas orthologues, la cellule vaut 0. Le logiciel i-*ADHoRe* recherche des clusters de gènes alignés dans cette matrice qui correspondent à des régions homologues dans les génomes analysés.

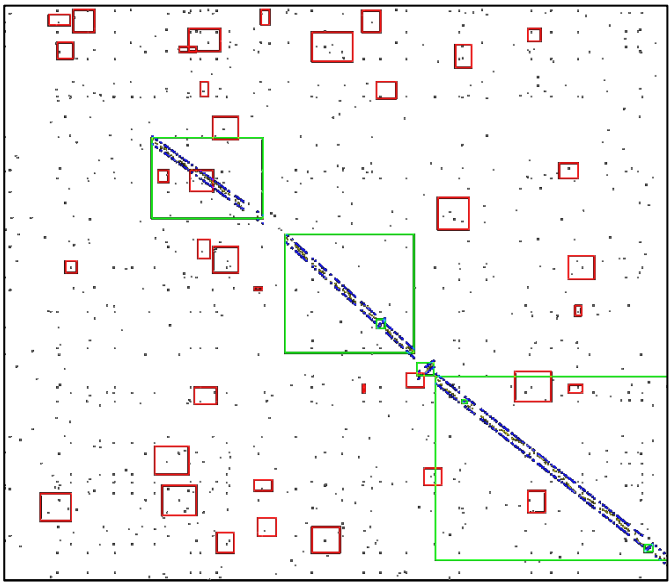


Figure 18 Exemple de « Gene Homology Matrix » entre deux chromosomes d'*Arabidopsis Thaliana* (Simillion et al., 2008). On identifie trois grandes régions homologues (encadrées en vert). Les points bleus représentent l'intervalle de confiance de la régression linéaire utilisée pour valider ces multiplicons. Les cadres rouges indiquent des multiplicons qui n'ont pas été validés.

Les paires des gènes homologues issus de clusters linéaires de la matrice sont ensuite alignés. L'utilisateur peut choisir entre deux algorithmes : soit l'algorithme de Needleman et Wunsch (Needleman and Wunsch, 1970), soit un aligneur progressif développé pour *i-ADHoRe* (Fostier et al., 2011) et permettant de corriger avec l'ajout de nouvelles « séquences » les erreurs d'alignement créées aux étapes antérieures de l'algorithme. Les alignements sont convertis en un profil combinant l'ordre et l'orientation des ancres. Ensuite, d'autres segments homologues sont recherchés en comparant les autres génomes au profil établi précédemment (et non plus à de simples segments de génomes). Si de nouveaux segments s'alignent avec le profil, ils sont incorporés à ce dernier. Le processus est répété de manière itérative jusqu'à ce que plus aucun nouveau segment homologue ne soit détecté à l'aide du profil. L'utilisation d'un profil permet de détecter une conservation de la synténie même « lointaine », par exemple dans le cas où des gènes homologues auraient été perdus dans certains des génomes analysés.

Les paramètres suivants ont été utilisés : cluster_type=colinear (on recherche des régions homologues colinéaires et non pas simplement avec un contenu en gènes similaires dont l'ordre ne serait pas conservé), alignment_method=gg2 (la dernière méthode d'alignement des profils implémentée dans *i-ADHoRe* 3.0 qui corrige les erreurs précoces de l'alignement de séquences multiples), gap_size=30 (pseudo-distance maximale autorisée entre les points d'un cluster de gènes), cluster_gap=35 (pseudo-distance maximale entre de petits blocs de synténie qui seront fusionnés en un bloc plus grand). Ces deux valeurs ont été choisies assez grandes afin de permettre la détection de blocs de synténie entre des génomes très divergés afin de limiter la perte d'information due au fait que l'on cherche des blocs ubiquitaires. Dans le cas où l'on cherche des blocs de synténie en comparant des génomes deux à deux, nous avons fixé gap_size=6 et cluster_gap=6. Nous avons fixé q_value=0.75 (valeur comprise entre 0 et 1 indiquant la mesure de linéarité minimale d'un groupe d'ancres de la GHM considérées comme appartenant à un bloc), prob_cutoff=0.01 (seuil de probabilité déterminant qu'un cluster de gènes n'apparait pas par chance dans un génome), anchor_points=3 (nombre minimal d'ancres nécessaire pour générer un bloc), level_2_only=false (false pour identifier des blocs ubiquitaires, true si l'on cherche des blocs issus de comparaisons de génomes deux à deux, voir paragraphe b), number_of_threads=4 (nombre de processeurs affectés au calcul).

Les résultats d'*i-ADHoRe* sont organisés sous la forme d'un graphe orienté dans lesquels les nœuds sont appelés « multiplicons » et représentent des blocs de synténie à une itération donnée du programme. Les multiplicons contiennent des « segments » homologues appartenant à différents génomes. On appelle « niveau » le nombre de segments que possède un multiplicon. Ainsi tout multiplicon est au moins de niveau 2. A chaque itération, un nœud fils est ajouté à chaque multiplicon pour lequel un nouveau segment est détecté. Les multiplicons correspondant à des blocs ubiquitaires dans n espèces sont donc de niveau n. A la dernière itération de l'algorithme, quand plus aucun nouveau segment n'est détecté, les feuilles du graphe correspondent aux blocs de synténie « définitifs » identifiés par *i-ADHoRe*. Ces feuilles ont été

extraites, et seuls les multiplicons possédant exactement un segment dans chaque génome ont été conservés.

Chaque multiplicon identifié par *i-ADHoRe* a été transformé en une liste de coordonnées génomiques décrivant la localisation des segments du bloc de synténie dans les différents génomes analysés. Chaque segment est localisé par le nom du génome dans lequel il apparait, le nom du chromosome qui le porte, sa position de début et de fin (en paires de bases) sur ce chromosome et enfin par son orientation, déduite de son alignement avec le profil généré par *i-ADHoRe*. Ces marqueurs sont bien ubiquitaires (présents dans tous les génomes) et à plus forte raison, uniques (présents exactement une fois dans chaque génome). Toutefois certains se chevauchent ou se recouvrent complètement. *ANGES* n'accepte pas de tels marqueurs. De plus, *GapAdj*, *MGRA* et *PMAG+* prennent en données d'entrée des permutations signées de blocs de synténie, et ne supportent donc pas le chevauchement ou le recouvrement de marqueurs. C'est pourquoi les marqueurs ont été filtrés selon les règles suivantes :

- Si un grand marqueur en recouvre un plus petit, le marqueur le plus grand est conservé.

- Si un marqueur est recouvert par plusieurs autres, il est éliminé. Dans les marqueurs restants après ce filtre, si plusieurs marqueurs sont tous recouverts par un même grand marqueur, on détermine s'il est préférable de retirer le grand marqueur ou les petits en cherchant à maximiser la couverture moyenne (en paires de bases) dans les différents génomes.

- Si deux marqueurs se chevauchent partiellement, le marqueur le plus grand est raccourci pour que le chevauchement soit nul. Si le marqueur ainsi raccourci est plus court que la longueur équivalente à trois gènes du génome analysé (en paires de bases), il est éliminé.

Au terme de ces étapes, nous avons obtenu 342 marqueurs recouvrant en moyenne 75% (en paires de bases) des génomes des *Lachancea* (Figure 19, page 88). Les blocs obtenus ont été écrits dans les différents formats attendus par *ANGES* (coordonnées génomiques des segments de chaque bloc), *GapAdj*, *MGRA* et *PMAG+* (qui prennent tous les trois des permutations signées de marqueurs en données d'entrée).

Pour reconstruire les génomes des *Lachancea* avec *AnChro* à partir des blocs de synténie identifiés par *i-ADHoRe*, nous avons détecté les blocs de synténie entre les différentes paires de génomes avec les paramètres *pairwise* mentionnés précédemment : cluster_gap=6, gap_size=6. Toutefois nous avons conservé anchor_points=3 afin d'éviter que trop de petits blocs ne soient générés. Nous avons également fixé level_2_only=true car les blocs sont alors recherchés entre des paires de génomes. Nous avons converti les multiplicons d'iADHoRe au format attendu par *AnChro* (blocs de synténie et ancres). Le nombre et la taille des blocs obtenus avec ces paramètres sont très similaires à ceux obtenus avec *SynChro*.

Nous avons représenté Figure 20, la couverture des génomes des *Lachancea* par les blocs ubiquitaires obtenus avec *SynChro* et *i-ADHoRe*. Au total nous avons identifié avec *SynChro* 339 blocs ubiquitaires, recouvrant en moyenne 77% (en paires de bases) des génomes des *Lachancea*. On remarque que cette proportion est légèrement plus importante que celle qu'on a obtenue avec *i-ADHoRe* (Figure 20). Cette différence peut être attribuée au fait que l'on autorise les blocs de deux ancres seulement dans les blocs reconstruits par *SynChro* alors que le nombre minimal d'ancres par bloc recommandé pour détecter les blocs de synténie avec *i-ADHoRe* est de 3.

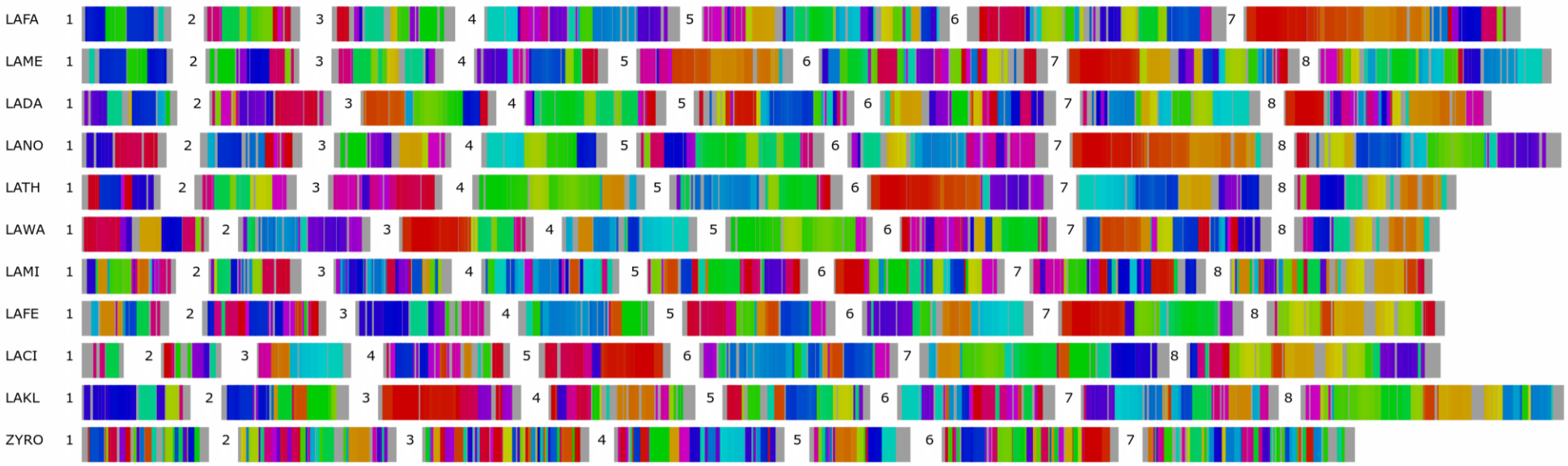


Figure 19 Carte chromosomique représentant la localisation des marqueurs ubiquitaires obtenus avec *i-ADHoRe* en analysant les génomes des *Lachancea* et de *Zygosaccharomyces rouxii*. Chaque ligne représente un génome, chaque bloc numéroté symbolise un chromosome. La couleur grise indique les régions non-couvertes par un bloc de synténie. En moyenne, 75% des génomes (en paires de bases) sont recouverts par un bloc de synténie ubiquitaire. On remarque que les télomères, régions dynamiques dans lesquels s'accumulent des séquences répétées, ne sont pas couverts.

LACI : *Lachancea cidrii*, LADA : *Lachancea dasiensis*, LAFA : *Lachancea fantastica*, LAFE : *Lachancea fermentati*, LAKL : *Lachancea kluyveri*, LAME : *Lachancea meryersii*, LAMI : *Lachancea mirantina*, LANO : *Lachancea nothofagi*, LATH : *Lachancea thermotolerans*, LAWA : *Lachancea waltii*, ZYRO : *Zygosaccharomyces rouxii*

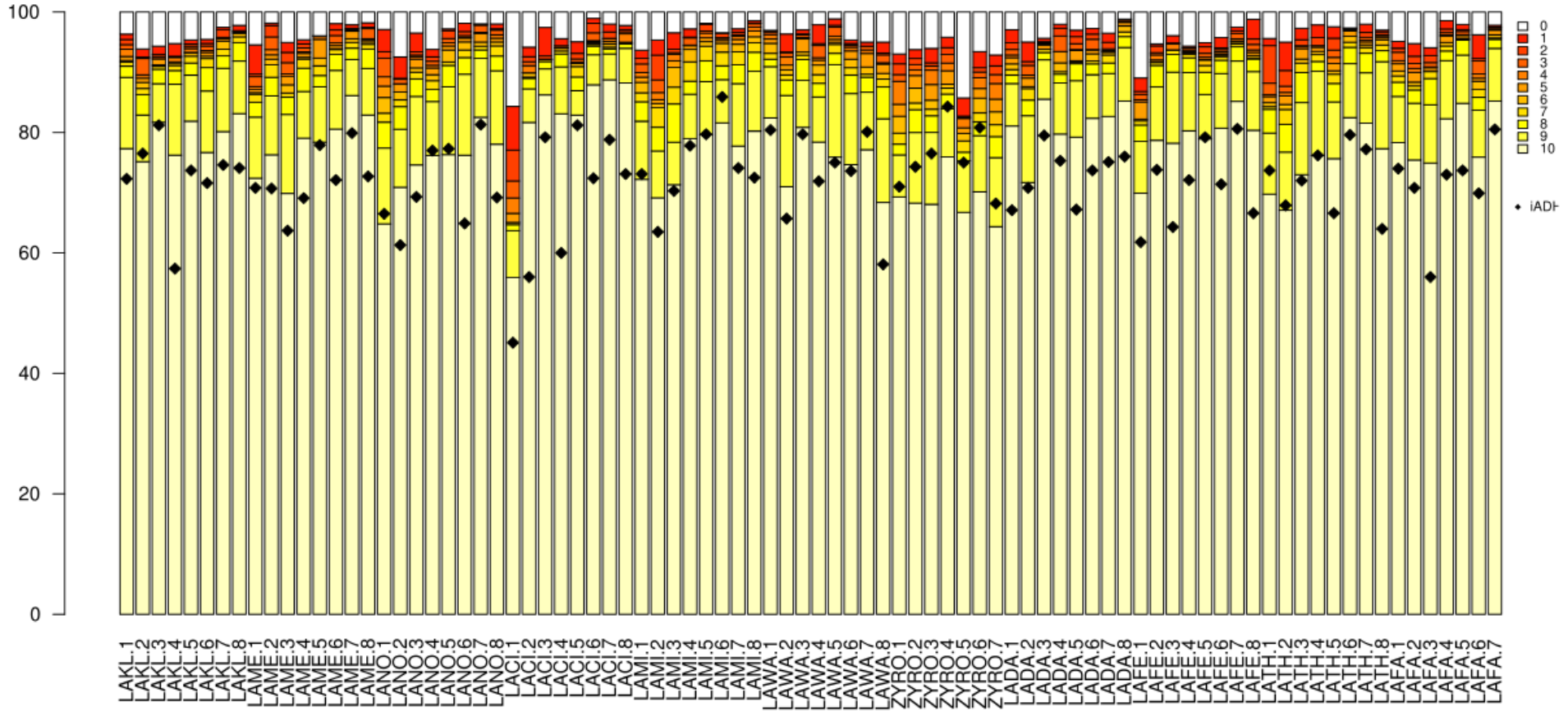


Figure 20 Proportion (en paires de bases) des chromosomes des espèces de *Lachancea* et de *Z. rouxii* couverts par les blocs de synténie détectés par *SynChro et i-ADHoRe*. Chaque colonne représente un chromosome. La couleur indique le nombre d'espèces avec lequel une région considérée est retrouvée en synténie. On remarque que les blocs ubiquitaires (retrouvés chez 10 espèces en plus du génome considéré) et représentés en beige couvrent environ 77% des génomes. Pour la majorité des chromosomes, les blocs de synténie ubiquitaires identifiés par *SynChro* recouvrent un plus grand nombre de paires de bases que les blocs identifiés par *i-ADHoRe* (losanges noir). De plus on remarque qu'environ 20% des génomes sont couverts par des blocs de synténie non-ubiquitaires. Contrairement aux autres logiciels de reconstruction de génomes ancestraux, *AnChro* est en mesure d'exploiter l'information de ces blocs.

LACI : *Lachancea cidrii*, LADA : *Lachancea dasiensis*, LAFA : *Lachancea fantastica*, LAFE : *Lachancea fermentati*, LAKL : *Lachancea kluyveri*, LAME : *Lachancea meryersii*, LAMI : *Lachancea mirantina*, LANO : *Lachancea nothofagi*, LATH : *Lachancea thermotolerans*, LAWA : *Lachancea waltii*, ZYRO : *Zygosaccharomyces rouxii*

4.1.2. Reconstruction des génomes ancestraux

4.1.2.1. Reconstruction de génomes ancestraux avec *ANGES*

Les reconstructions ont été réalisées à partir des blocs ubiquitaires générés d'une part par *i-ADHoRe* et d'autre part par *SynChro*. Nous avons adapté les paramètres de reconstruction à nos marqueurs ubiquitaires et uniques : markers_unique=2 (les marqueurs sont présents exactement une fois), markers_universal=2 (dans tous les génomes analysés), markers_doubled=1 (cette option permet à *ANGES* d'inférer l'orientation des blocs dans le génome ancestral reconstruit et pas seulement leur ordre relatif). Les autres paramètres ont été fixés d'après les configurations recommandées pour la reconstruction de génomes ancestraux de levures (Chauve et al., 2010; Jones et al., 2012).

4.1.2.2. Reconstruction de génomes ancestraux avec *GapAdj*

Les génomes actuels écrits sous la forme de permutations signées sont fournis à *GapAdj*, et s'accompagnent de l'arbre phylogénétique des espèces analysées. Les seuls paramètres requis par l'algorithme sont Δ (fixé à 1) correspondant à la pseudo-distance maximale autorisée pour définir deux blocs comme adjacents et un seuil (fixé à 0.6) définissant une adjacence comme ancestrale si elle est retrouvée chez un certain nombre d'espèces. Ces paramètres ont également été utilisés pour reconstruire les génomes ancestraux des génomes simulés.

4.1.2.3. Reconstruction de génomes ancestraux avec *MGRA*

Cet algorithme utilise lui aussi des permutations signées de blocs de synténie ubiquitaires et uniques. Le seul paramètre à fixer est le nombre d'itérations que le programme doit réaliser. Nous avons fixé ce paramètre à 3.

4.1.2.4. Reconstruction de génomes ancestraux avec *PMAG+*

PMAG+ n'est pas téléchargeable mais peut être utilisé en ligne (Hu et al., 2014). Aucun paramètre n'est fixé par l'utilisateur. Afin de faciliter la reconstruction des nombreux ancêtres des génomes simulés, un script python a été créé pour écrire et soumettre automatiquement les requêtes au site web de *PMAG+* (http://www.geneorder.org/).

4.1.2.5. Reconstruction de génomes ancestraux avec *AnChro*

AnChro requiert trois paramètres pour reconstruire chaque génome ancestral (noté Anc) d'un arbre phylogénétique donné. (i) Le premier correspond à la valeur de Δ utilisée par *SynChro* pour calculer les blocs de synténie entre les génomes G1 et G2, appartenant à deux espèces actuelles reliées par un chemin passant par Anc dans l'arbre phylogénétique. (ii) De même, le deuxième paramètre, Δ' définit la valeur de Δ à utiliser lors de la détection par *SynChro* des blocs de synténie partagés entre les génomes G1/(G3, …, Gn) et G2/(G3, …, Gn). (iii) Le troisième paramètre définit la paire de génomes (G1, G2) ainsi que la liste de génomes (G3, …, Gn) à utiliser pour effectuer la reconstruction. Il existe donc en principe ΔxΔ'

reconstructions possibles d'un ancêtre (soit 36 car Δ et Δ' sont compris entre 1 et 6), nombre auquel il faut multiplier le nombre de paires G1, G2 possibles.

Dans le cadre de la reconstruction de l'histoire évolutive des *Lachancea* (Vakirlis et al., 2016), la version de chaque génome ancestral des *Lachancea* (A1, …, A9) reconstruit avec *AnChro* a été sélectionnée selon des paramètres optimaux permettant d'obtenir des reconstructions maximisant le nombre de gènes des chromosomes reconstruits tout en minimisant le nombre de *scaffolds* et le nombre de contradictions inter-chromosomiques c'est-à-dire d'erreurs de reconstruction. Toutefois, nous nous intéressons ici aux performances d'*AnChro* en comparaison aux autres logiciels de reconstruction, c'est pourquoi nous analyserons une reconstruction « par défaut » d'*AnChro*. Cette reconstruction par défaut est obtenue en choisissant comme ancêtres G1 et G2 les deux génomes partageant un nombre de blocs de synténie minimal (G1 et G2 sont les deux génomes ayant accumulé le moins de réarrangements depuis leur divergence de l'ancêtre à reconstruire).

Dans le cas de la reconstruction de génomes ancestraux simulés, nous avons fixé Δ'= Δ''=6 (en adéquation avec les paramètres gap_size=6 et cluster_gap=6 utilisés par *i-ADHoRe* pour détecter les blocs de synténie entre les paires de génomes simulés). Pour chaque ancêtre à reconstruire, toutes les combinaisons de G1, G2, (G3, …, Gn) ont été calculées.

4.1.3. Validation des reconstructions des génomes ancestraux des *Lachancea*

4.1.3.1. Pertinence biologique des reconstructions

La pertinence des reconstructions d'un ancêtre a été évaluée selon trois critères : (i) le nombre de chromosomes, qui d'après le principe de parcimonie doit être proche de celui des espèces actuelles, c'est-à-dire 8 dans le cas des *Lachancea* ; (ii) le nombre de gènes retracés qui doit être maximal, soit proche de 5000 ; (iii) le nombre de centromères par chromosomes, un chromosome ne pouvant avoir qu'un seul centromère. L'utilisation de ce critère est rendue possible par la grande conservation en synténie des centromères ponctuels des *Saccharomycetaceae* (Gordon et al., 2011a). Dans les génomes actuels des *Lachancea*, 76 centromères parmi les 79 au total sont flanqués par des gènes orthologues. La probabilité de reconstruire par chance les ancêtres des *Lachancea* avec un seul centromère par chromosomes a été calculée par ancêtre pour *AnChro*. Ces probabilités s'étendent de 2.0×10^{-3} à 3.5×10^{-4} (Vakirlis et al., 2016) ce qui fait de ce critère une importante validation de la qualité des reconstructions.

Pour replacer les centromères, la position de ces derniers est recherchée dans les reconstructions de la manière suivante : dans les génomes reconstruits à partir de marqueurs ubiquitaires, on recherche les centromères qui sont dans un bloc de synténie. Pour retrouver les centromères restants, on examine dans les génomes actuels le numéro des blocs de synténie ubiquitaires situés en amont et en aval de ces centromères. On recherche ensuite ces blocs dans les génomes reconstruits et on considère qu'un centromère peut être replacé avec précision si les deux blocs qui l'encadrent dans un génome actuel sont retrouvés côte à côte dans le génome reconstruit.

4.1.3.2. Algorithmes de détection de la synténie et qualité des reconstructions

Nous avons comparé les résultats de reconstruction des cinq logiciels à partir des blocs de synténie générés par *i-ADHoRe* (Figure 21) et par *SynChro* (Figure 22).

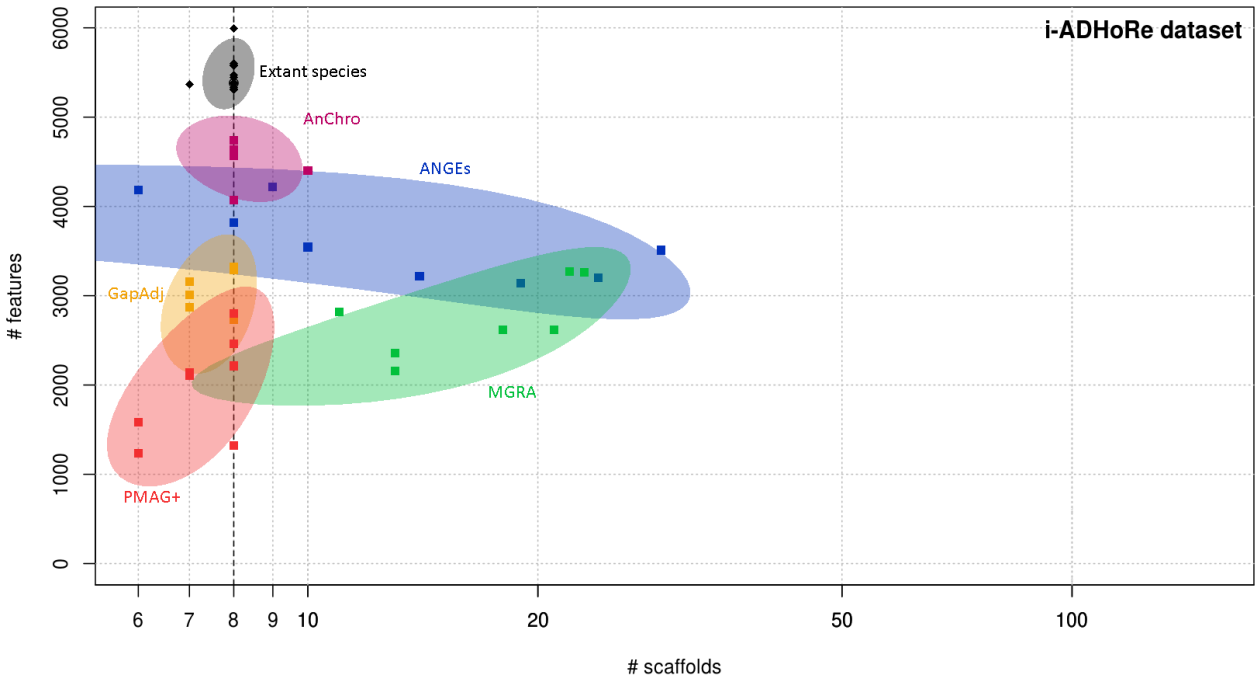


Figure 21 Nombre de *scaffolds* et d'éléments génétiques dans chaque reconstruction de génome ancestral en fonction du logiciel de reconstruction utilisé avec les données de synténie générées par *i-ADHoRe*. Ces résultats de reconstruction correspondent aux modalités décrites par les flèches marron de la Figure 16, page 84.

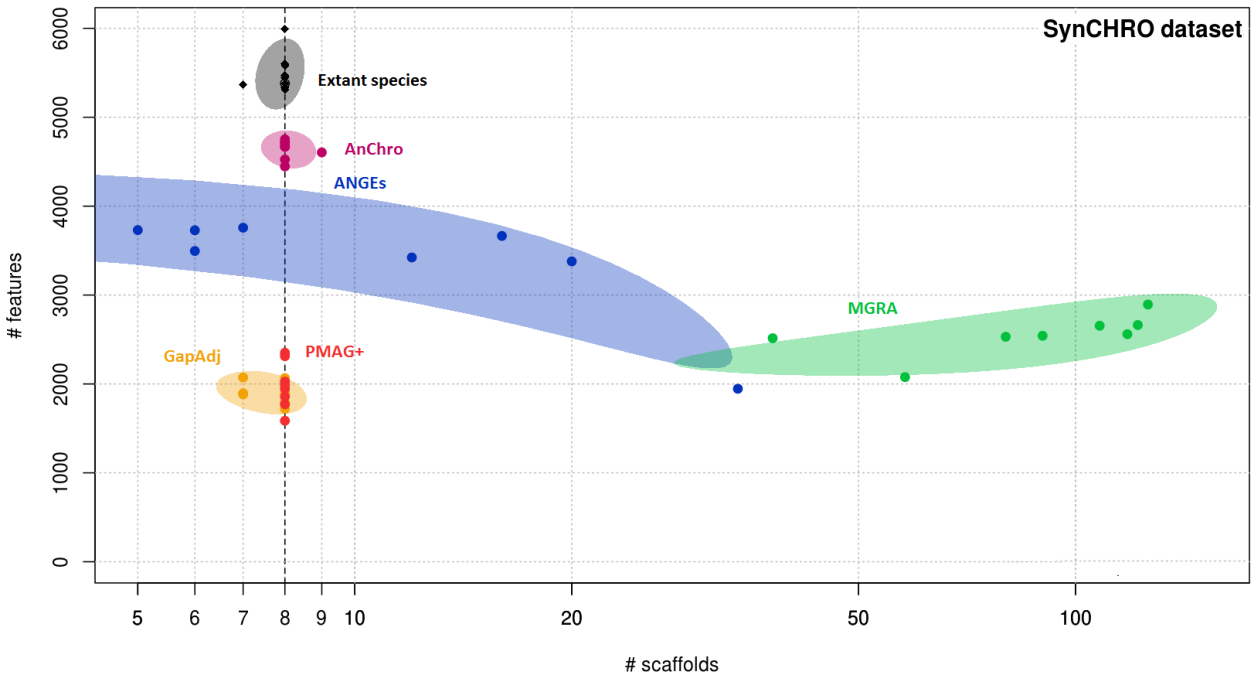


Figure 22 Nombre de *scaffolds* et d'éléments génétiques dans chaque reconstruction de génome ancestral en fonction du logiciel de reconstruction utilisé avec les données de synténie générées par *SynChro*. Ces résultats de reconstruction correspondent aux modalités décrites par les flèches cyan de la Figure 16, page 84.

Pour le logiciel *AnChro*, on observe que les reconstructions des génomes ancestraux des *Lachancea* obtenues à partir des données de *SynChro* ont bien 8 chromosomes (sauf un ancêtre), contrairement aux reconstructions obtenues à partir des données d'*i-ADHoRe* où deux génomes ont plus de 8 chromosomes et un nombre de gènes plus restreint. Le nombre de gènes retracés est néanmoins similaire pour les deux jeux de données.

A l'inverse, pour *ANGES*, *GapAdj*, *MGRA* et *PMAG+*, les reconstructions obtenues à partir du jeu de données d'*i-ADHoRe* sont meilleures que les données des blocs de synténie générées par *SynChro* en termes de fragmentation des génomes, le nombre de gènes retracés étant comparable entre les ancêtres obtenus à partir des deux jeux de données.

Nous avons ensuite retracé la position des centromères dans les génomes reconstruits à partir des blocs de synténie générés par *i-ADHoRe* (Figure 23). On observe qu'*AnChro* a reconstruit les génomes les plus contigus et les plus complets. De plus, les centromères sont uniques au sein de chaque chromosome.

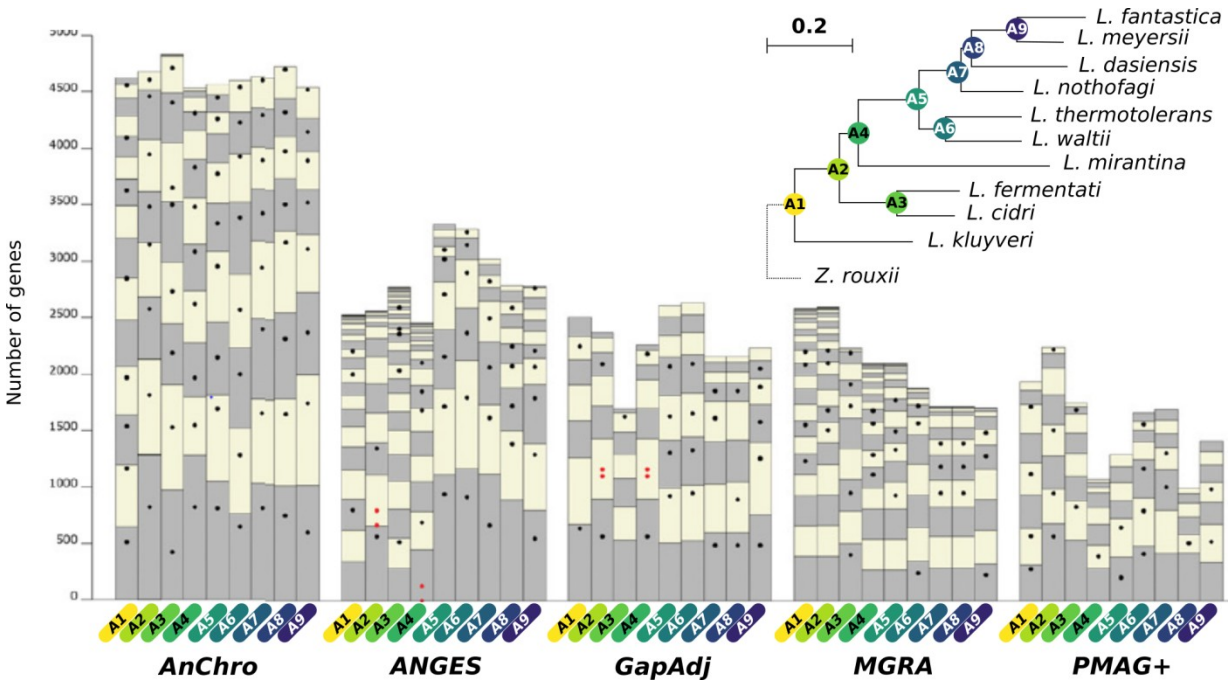


Figure 23 Comparaison des neuf ancêtres des *Lachancea* (A1 à A9) reconstruits par *AnChro*, *ANGES*, *GapAdj*, *MGRA*, et *PMAG+*. Les blocs de synténie ont été identifiés avec *i-ADHoRe* pour tous les logiciels de reconstruction. Pour *AnChro*, une seule reconstruction par défaut est représentée. Chaque colonne représente le génome d'un ancêtre donné comme une alternance de boîtes beiges et grises dont la taille est proportionnelle au nombre de gènes retracés. Les points noirs indiquent les centromères uniques par chromosome. Les points rouges indiquent les centromères quand ils sont surnuméraires pour un chromosome. (Vakirlis et al., 2016)

En définitive, sur ce jeu de données réelles, *AnChro* a permis de générer les reconstructions les plus pertinentes.

4.1.4. Evaluation de la pertinence des reconstructions à partir de génomes simulés

4.1.4.1. Intérêt des données simulées

Les critères biologiques présentés précédemment permettent d'apprécier la qualité des génomes reconstruits dans leur globalité, mais ne donnent pas accès aux différents types d'erreurs des outils de reconstruction utilisés. Par exemple, on pourrait avoir l'impression qu'un génome reconstruit est pertinent sur la base de son contenu en gènes alors qu'il introduit un grand nombre d'adjacences fausses dans les reconstructions. Comme on ne dispose malheureusement pas d'ADN ancestral de levures *Lachancea* pour évaluer la pertinence des génomes reconstruits par comparaison des espèces actuelles, il faut trouver un autre moyen pour mesurer la performance des outils de reconstruction.

Afin de quantifier les taux d'erreurs des différents outils de reconstruction (*ANGES*, *GapAdj*, *MGRA* et *AnChro*) nous avons simulé de manière réaliste l'évolution de génomes comparables en taille à ceux des *Lachancea*. L'objectif de ces simulations est de présenter aux logiciels de reconstruction un grand nombre de problèmes similaires à celui que représente la reconstruction des génomes réels de *Lachancea* afin de quantifier avec précision le taux d'erreur de chaque méthode (puisqu'on connait la structure des génomes ancestraux simulés) et ainsi de valider indirectement la pertinence des reconstructions à partir de données réelles.

Chaque simulation démarre avec un génome de 5000 gènes, distribués entre 8 chromosomes, contenant chacun un centromère. Notons que les génomes sont ici représentés par des listes d'identifiants de gènes, aucune donnée protéique ou nucléique ne leur est associée. Nous avons ensuite simulé 100 arbres binaires à 11 feuilles et 9 nœuds, similaires à l'arbre des *Lachancea* enraciné avec une espèce externe au clade. Ces arbres ont été générés selon le processus décrit dans (Kuhner and Felsenstein, 1994). Dans chaque arbre, des réarrangements chromosomiques balancés (inversion et translocations réciproques) ont été simulés le long des branches de l'arbre avec un minimum de 10 événements par branche et de manière à ce que le nombre de blocs et la taille des blocs de synténie dans les génomes simulés suivent la distribution observée dans les génomes des *Lachancea*. Etonnamment, nous avons dû induire selon les simulations entre 3 et 7 fois le nombre de réarrangements balancés estimés à partir des génomes de *Lachancea*, probablement parce que le simulateur utilisé n'introduit pas de duplications, délétions ou transpositions. Pour chaque réarrangement, nous avons fixé comme équiprobables les translocations réciproques et les inversions. La taille de ces dernières suit une loi de Poisson d'une moyenne de 5 gènes.

4.1.4.2. Pertinence des données simulées

Nous avons utilisé *i-ADHoRe* pour détecter les blocs de synténie *pairwise* entre les génomes simulés et avons comparé le nombre et la taille des blocs avec la distribution observée dans les génomes des *Lachancea* avec *SynChro* (Figure 24). La pertinence de cette comparaison mérite quelques explications. Premièrement, les blocs de synténie *pairwise* identifiés avec *SynChro* et *i-ADHoRe* sont très similaires en taille et en nombre, ce qui autorise leur comparaison. Deuxièmement, les données simulées n'étant associées à aucune donnée protéique, seul le logiciel *i-ADHoRe* permet de détecter la synténie entre les génomes simulés car *SynChro* requiert l'utilisation de séquences protéiques pour la reconstruction des blocs

de synténie. Nous avons donc décidé d'utiliser des blocs issus d'*i-ADHoRe* pour reconstruire les génomes ancestraux des simulations. Ce choix est plutôt en faveur des logiciels *ANGES, GapAdj, MGRA* et *PMAG+*, qui génèrent de meilleures reconstructions à partir de ces données et en défaveur d'*AnChro* qui comme nous l'avons vu, reconstruit des génomes ancestraux un peu plus fragmentés avec ces blocs. Au final, on observe que la distribution du nombre de gènes par blocs suit bien la distribution observée dans les génomes réels actuels des *Lachancea* (Figure 24). Ce jeu de données simulées a donc été utilisé par la suite pour évaluer la performance des logiciels de reconstruction *AnChro, ANGES, GapAdj, MGRA* et *PMAG+*.

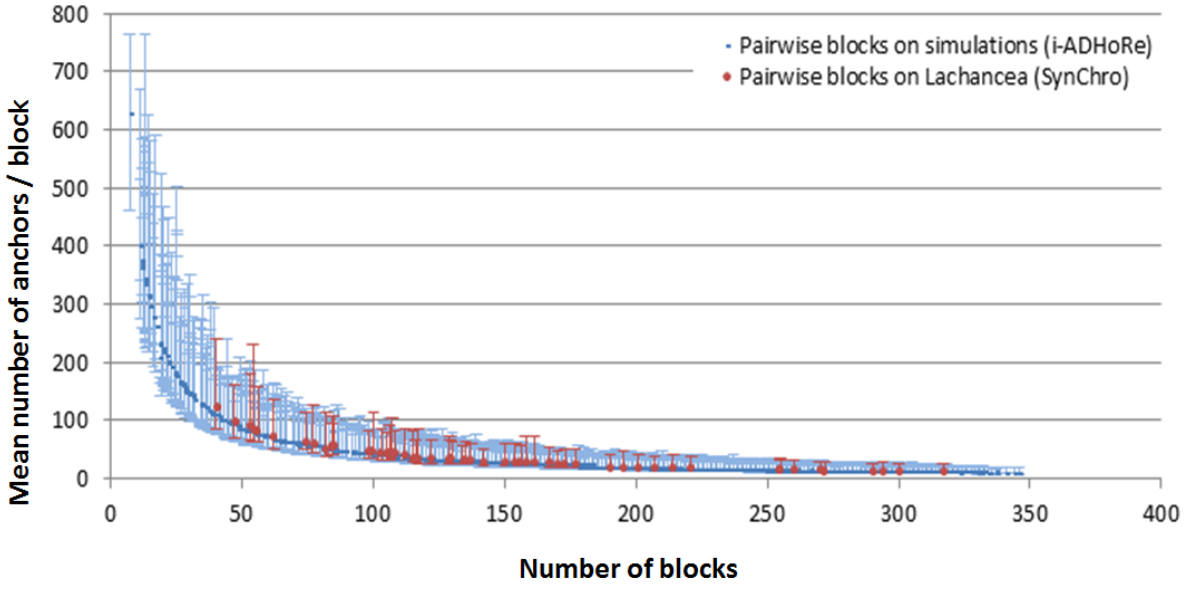


Figure 24 Nombre de blocs de synténie et d'ancres obtenus à partir des comparaisons de génomes deux à deux avec *SynChro* sur les génomes des *Lachancea*, en rouge et *i-ADHoRe* sur les génomes simulés en bleu. Chaque point représente une comparaison de deux génomes. Les intervalles de confiance représentent la dispersion du nombre d'ancres dans les blocs.

4.1.4.3. Résultats des reconstructions

Les 900 génomes ancestraux (9 ancêtres x 100 arbres), correspondants aux nœuds de tous les arbres simulés ont été reconstruits avec *ANGES*, *GapAdj*, *MGRA*, *PMAG+* et *AnChro*. Notons que pour *AnChro*, nous n'avons pas optimisé la reconstruction des ancêtres, une seule version ayant été calculée par ancêtre, en utilisant la comparaison entre les génomes G1 et G2 les moins réarrangés c'est-à-dire les génomes dont le nombre de blocs de synténie est minimum.

La qualité des reconstructions en termes de nombre de *scaffolds*, nombre de gènes et en pourcentage de *scaffolds* possédant un seul centromère est représentée (Figure 25). Les reconstructions par défaut d'*AnChro* comprennent en moyenne 4602 gènes tandis que les autres outils ont une moyenne d'environ 1500 gènes. Le nombre d'ancêtres reconstruits avec huit chromosomes est variable : 213 avec *AnChro*, 217 avec *PMAG+*, 46 avec *GapAdj*, 3 avec *ANGES* et seulement 1 avec *MGRA*. Les reconstructions d'*AnChro*, *PMAG+*, *GapAdj*, *ANGES* et *MGRA* ont un pourcentage de *scaffolds* avec un seul centromère de 42, 0%, 22, 1%, 13%, 3, 5% et 3, 1% respectivement (Figure 25).

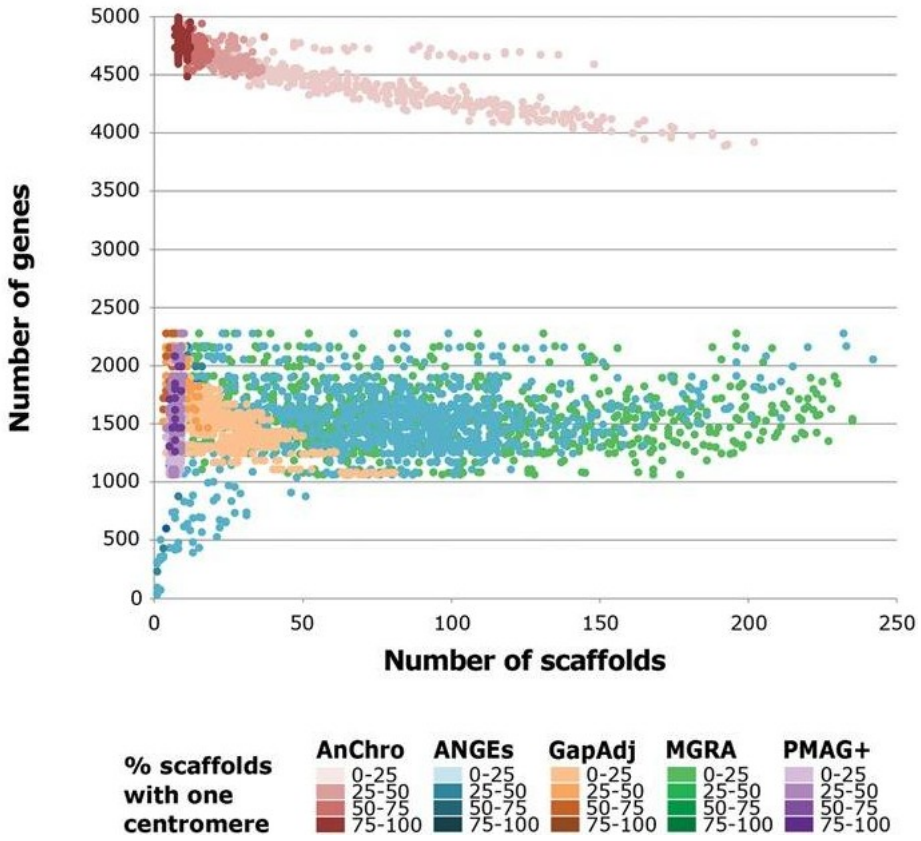


Figure 25 Reconstructions de génomes ancestraux à partir des espèces actuelles simulées. La figure représente 900 ancêtres reconstruits correspondant à 9 ancêtres par simulation et 100 simulations, pour chacun des logiciels : *AnChro* (reconstruction par défaut), *ANGES*, *GapAdj*, *MGRA* et *PMAG+*. Chaque génome ancestral est représenté par un point. La qualité des reconstructions est représentée par le nombre de gènes retracés (idéalement 5000), le nombre de *scaffolds* (idéalement 8) et la proportion de *scaffolds* de chaque reconstruction possédant un unique centromère (idéalement 100%). (Vakirlis et al., 2016)

De plus, nous avons calculé la proportion d'adjacences reconstruites de manière correcte et erronée entre les blocs de synténie des 900 génomes reconstruits comparativement à leur équivalent simulé (Figure 26). Pour ce faire, nous avons appliqué la méthode expliquée ci-après sur tous les génomes reconstruits. Pour toute adjacence de blocs dans l'ancêtre reconstruit (noté R), nous avons déjà déterminé si cette adjacence est télomérique, c'est-à-dire si cette adjacence est constituée d'un bloc de synténie « réel » et d'un bloc artificiel représentant un télomère. Si oui, on vérifie que cette adjacence est également présente dans la simulation auquel cas l'adjacence correspond à un télomère. En résumé, elle est correcte et on la représente en vert. Si cette adjacence ne correspond pas à un télomère dans l'ancêtre simulé, c'est que l'adjacence avec le bloc voisin observé dans le génome simulé n'a pas pu être reconstruite, on représente cette adjacence en gris. Dans le cas où l'adjacence de R considérée n'est pas située à l'extrémité d'un *scaffold*, on cherche si cette adjacence existe dans le génome simulé correspondant. Si c'est le cas, l'adjacence est directement catégorisée comme correcte (en vert). Dans le cas contraire, l'adjacence a été reconstruite de manière erronée. Nous avons fait la distinction entre différents types d'erreur. Si les deux blocs qui forment cette adjacence sont tous deux localisés sur le même chromosome simulé, on parle d'erreur intra-chromosomique (en orange) et plus spécifiquement, s'ils sont voisins c'est une inversion d'un seul bloc : les deux blocs sont bien voisins dans la reconstruction mais leurs extrémités en contact ne sont

pas les bonnes. Ces erreurs sont représentées en bleu. Si les deux blocs de l'adjacence reconstruite ne sont pas situés sur le même chromosome simulé, l'erreur est plus grave car elle mélange des chromosomes simulés. On parle d'erreur inter-chromosomique (en rouge). On peut en principe identifier les cas où deux télomères simulés ont été rejoints dans la reconstruction mais cette distinction n'est pas représentée dans la Figure 26.

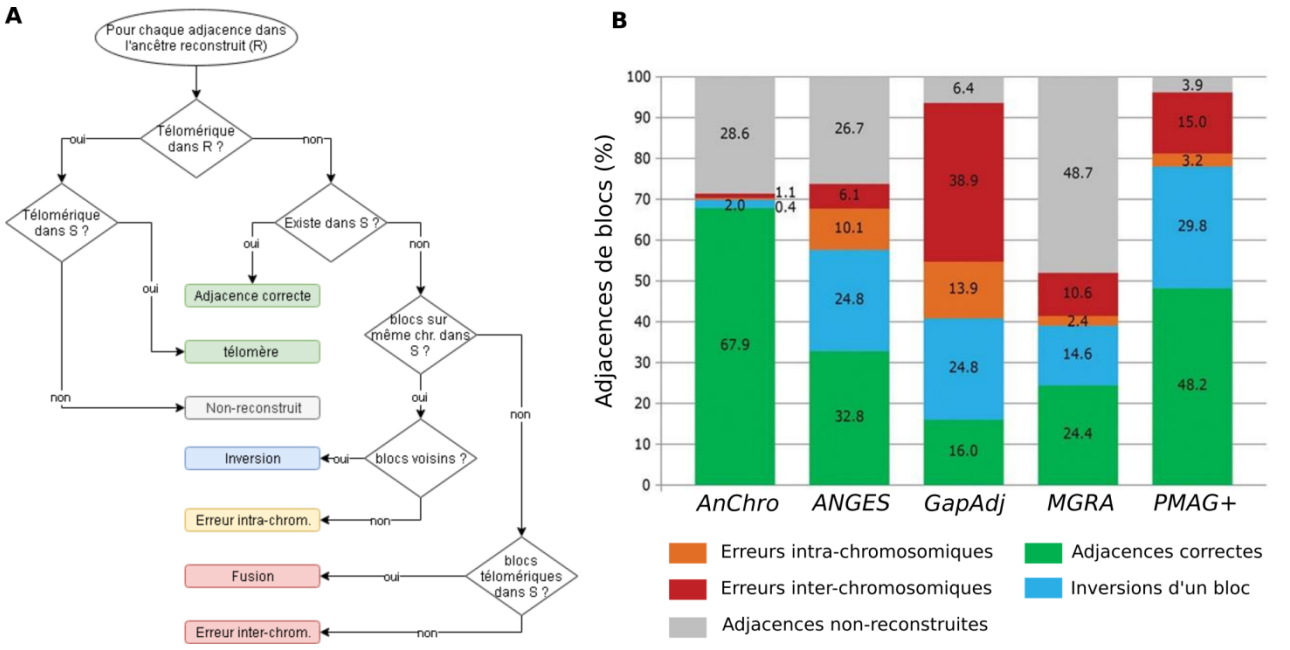


Figure 26. Quantification des erreurs de reconstructions à partir des données simulées. Les adjacences correctes sont en vert. Les erreurs intra-chromosomiques ont été décomposées en deux catégories : les inversions d'un seul bloc sont en bleu et les autres erreurs sont en orange. Les erreurs inter-chromosomiques sont en rouge. La proportion d'adjacences non reconstruites est également indiquée en gris. (**A**) méthode de quantification des erreurs. Le génome ancestral reconstruit est noté R. Le génome ancestral simulé est noté S. (**B**) Proportion moyenne d'adjacences reconstruites de manière correcte et incorrecte dans les 900 reconstructions obtenues avec les cinq programmes.

Cette analyse a montré qu'*AnChro* reconstruit la plus haute proportion d'adjacences correctes (67,9%) comparativement à *PMAG+*, ANGEs, *MGRA* et *GapAdj* (48.2%, 32.8%, 24.4% and 16%, respectivement). De plus, *AnChro* reconstruit la plus petite fraction d'adjacences fausses (3,5%) comparativement à *MGRA*, ANGEs, *PMAG+* et *GapAdj* (27.6%, 41%, 48% and 77.6%, respectivement). Ces chiffres donnent une valeur prédictive positive de 0.95, 0.50, 0.47, 0.44 et 0.17 pour *AnChro*, *PMAG+*, *MGRA*, ANGEs and *GapAdj*, respectivement (Équation 1).

Équation 1 : Calcul de la valeur prédictive positive (VPP)

$$\begin{cases} VPP = \dfrac{VP}{VP + FP} \\ VP : vrais\ positifs \\ FP : faux\ positifs \end{cases}$$

Si l'on considère l'ordre relatif des blocs indépendamment de leur orientation, les adjacences dues aux inversions d'un seul bloc peuvent être considérées comme correctes. Par conséquent, le taux d'erreurs dans les reconstructions des génomes des *Lachancea* est de 3,5% à 1,5% si l'on considère les blocs inversés comme correctement positionnés. Notons que ces proportions peuvent avoir été légèrement surestimées car ces évaluations reposent sur les reconstructions par défaut d'*AnChro* (blocs issus d'*i-ADHoRe* et pas d'optimisation *a posteriori*) alors que les reconstructions des génomes des *Lachancea* utilisées pour inférer l'histoire évolutive du clade ont été optimisées. Dans ce cas, les inversions d'un seul bloc sont considérées comme correctes et les VPP deviennent 0.98, 0.81, 0.78, 0.75 et 0.44 pour *AnChro*, *PMAG+*, ANGEs, *MGRA* et *GapAdj*, respectivement.

4.2. Reconstruction des génomes ancestraux des *Saccharomycotina*

Comme nous l'avons vu dans la sous-partie précédente, le logiciel *AnChro* permet d'obtenir les meilleures reconstructions de génomes ancestraux des *Lachancea*. Il est également très performant sur des données simulées. Etant donné la grande performance de cet outil, nous nous sommes demandé s'il était possible d'utiliser *AnChro* pour reconstruire les génomes ancestraux d'autres espèces, dont la dynamique des génomes serait potentiellement différente, ou pour reconstruire des génomes ancestraux plus anciens que ceux des *Lachancea*. Cela permettrait d'étudier les réarrangements qui ont façonné ces génomes et de comparer les dynamiques évolutives de différents groupes de génomes. Nous nous sommes donc intéressés à reconstruire les génomes ancestraux à l'échelle de l'ensemble du subphylum des *Saccharomycotina* en utilisant les génomes disponibles dans les bases de données publiques.

Cela représente un défi, car comme nous l'avons vu dans l'introduction, ces espèces recouvrent un large intervalle évolutif et leur génome est très divergé. Comparativement, les *Lachancea* forment un clade particulièrement approprié à la reconstruction de génomes ancestraux étant donné le nombre d'espèces actuelles séquencées et les taux de divergence « graduel » entre ces différentes espèces. En outre, les génomes des *Lachancea* utilisés dans la sous-partie précédente sont de haute qualité : les chromosomes sont assemblés de télomère à télomère et un effort particulier a été réalisé pour annoter ces génomes de la manière la plus précise possible, tandis que beaucoup de génomes publiés sont fragmentaires ou sont mal annotés.

Nous décrirons dans cette sous-partie comment nous avons reconstruit les génomes ancestraux de 66 espèces de *Saccharomycotina* pour inférer les réarrangements chromosomiques qui ont contribué à remodeler ces génomes au cours de l'évolution. Nous verrons que les nombres et les types de réarrangements chromosomiques sont variables selon les lignées. Néanmoins, nous verrons qu'à l'échelle de l'ensemble des *Saccharomycotina* étudiées, le nombre total de réarrangements chromosomiques balancés est bien corrélé au pourcentage de divergence protéique, suggérant la synchronisation de l'accumulation des mutations non-synonymes et des réarrangements chromosomiques.

4.2.1. Provenance et filtrage des génomes

Nous avons cherché à rassembler le plus grand nombre possible de génomes de *Saccharomycotina* à partir des bases de données publiques. Afin de pouvoir reconstruire des génomes ancestraux cohérents, il faut exploiter des génomes actuels de qualité. Des génomes trop fragmentés augmentent artificiellement le nombre d'adjacences non-reconstruites, des génomes mal annotés, contenant trop peu de gènes codants (nous nous intéressons ici à l'ordre ancestral des gènes codants des protéines) diminuent la couverture des génomes reconstruits. Aussi nous avons filtré les génomes de manière à ne conserver que les génomes suffisamment bien assemblés et annotés.

Nous avons rassemblé au total 79 génomes de *Saccharomycotina* à partir des bases de données publiques (Tableau 1).

Tableau 1 Liste des 79 génomes de *Saccharomycotina* rassemblés à partir des bases de données publiques.

Espèce	**Code**	**Souche**	**Référence**
Arxula adeninivorans	ARAD	LS3	(Kunze et al., 2014)
Ascoidea rubescens	ASRU	DSM 1968	(Riley et al., 2016)
Babjeviella inositovora	BAIN	NRRL Y-12698	(Riley et al., 2016)
Candida albicans	CAAL	WO-1	(Jones et al., 2004)
Candida auris	CAAU	6684	(Chatterjee et al., 2015)
Candida bracarensis	CABR	CBS 10154	JGI : Gs0004734
Candida castellii	CACA	CBS 4332	JGI : Gp0041973
Candida caseinolytica	CACS	NRRL Y-17796	(Riley et al., 2016)
Candida dubliniensis	CADU	CD36	(Jackson et al., 2009)
Candida glabrata	CAGL	CBS138	(Dujon et al., 2004)
Candida guilliermondii	CAGU	ATCC 6260	(Butler et al., 2009)
Candida nivariensis	CANI	CBS 9983	NCBI : ASM104691v1
Candida orthopsilosis	CAOR	Co 90-125	(Riccombeni et al., 2012)
Candida parapsilosis	CAPA	CDC317	JGI : Gp0002580
Candida tanzawaensis	CATA	NRRL Y-17324	(Riley et al., 2016)
Candida tenuis	CATE	ATCC 10573	(Wohlbach et al., 2011)
Candida tropicalis	CATR	MYA-3404	(Butler et al., 2009)
Cephaloascus albidus	CEAL	ATCC 66658	JGI : *Cephaloascus albidus* ATCC 66658 v1.0
Cephaloascus fragans	CEFR	12-1022	JGI : Cephaloascus fragrans 12-1022 v1.0
Clavispora lusitaniae	CLLU	ATCC 42720	(Butler et al., 2009)
Cyberlindnera fabianii	CYFA	YJS4271	(Freel et al., 2014)
Cyberlindnera jadinii	CYJA	CBS1600	(Rupp et al., 2015)
Dekkera bruxellensis	DEBR	AWRI1499	(Curtin et al., 2012)
Debaryomyces hansenii	DEHA	CBS767	ENA : GCA_000006445.2
Eremothecium aceri	ERAC		(Dietrich et al., 2013)
Eremothecium cymbalariae	ERCY	DBVPG#7215	(Wendland and Walther, 2011)
Eremothecium gossypii	ERGO	ATCC 10895	(Dietrich et al., 2004)
Geotrichum candidum	GECA	CLIB 918 (ATCC 204307)	(Morel et al., 2015)
Hanseniaspora uvarum	HAUV	DSM 2768	ENA : GCA_000968475.1
Hanseniaspora valbyensis	HAVA	NRRL Y-1626	(Riley et al., 2016)
Hyphopichia burtonii	HYBU	NRRL Y-1933	(Riley et al., 2016)
Kazachstania africana	KAAF	CBS 2517	(Gordon et al., 2011b)
Kazachstania naganishii	KANA	CBS 8797	(Gordon et al., 2011b)
Kluyveromyces dobzhanskii	KLDO	CBS 2104	ENA : GCA_000820885.1

Kluyveromyces lactis	KLLA	NRRL Y-1140	(Dujon et al., 2004)
Kluyveromyces marxianus	KLMA	DMKU3-1042	(Lertwattanasakul et al., 2015)
Komagataella pastoris	KOPA	CBS 7435	(Küberl et al., 2011)
Kuraishia capsulata	KUCA	CBS 1993	(Morales et al., 2013)
Lachancea cidri	LACI		(Vakirlis et al., 2016)
Lachancea dasiensis	LADA		(Vakirlis et al., 2016)
Lachancea fantastica	LAFA		(Vakirlis et al., 2016)
Lachancea fermentati	LAFE		(Vakirlis et al., 2016)
Lachancea kluyveri	LAKL		(Vakirlis et al., 2016)
Lachancea lanzarotensis	LALA	CBS 12615	(Sarilar et al., 2015)
Lachancea meyersii	LAME		(Vakirlis et al., 2016)
Lachancea mirantina	LAMI		(Vakirlis et al., 2016)
Lachancea nothofagi	LANO		(Vakirlis et al., 2016)
Lachancea quebecensis	LAQU	CBS 14088	(Freel et al., 2016)
Lachancea thermotolerans	LATH		(Vakirlis et al., 2016)
Lachancea waltii	LAWA		(Vakirlis et al., 2016)
Lodderomyces elongisporus	LOEL	NRRL YB-4239	(Butler et al., 2009)
Metschnikowia bicuspidata	MEBI	NRRL YB-4993	(Riley et al., 2016)
Millerozyma farinosa	MIFA	CBS 7064	(Louis et al., 2012)
Nakaseomyces bacillisporus	NABA	CBS 7720	(Gabaldón et al., 2013)
Naumovozyma castellii	NACA	CBS 4309	(Gordon et al., 2011b)
Naumovozyma dairenensis	NADA	CBS 421	(Gordon et al., 2011b)
Nakaseomyces delphensis	NADE	CBS 2170	(Gabaldón et al., 2013)
Nadsonia fulvescens	NAFU	DSM 6958	(Riley et al., 2016)
Ogataea parapolymorpha	OGPA	DL-1	(Ravin et al., 2013)
Pichia kudriavzevii	PIKU	SD108	NCBI : ASM198332v1
Pichia membranifaciens	PIME	NRRL Y-2026	(Riley et al., 2016)
Saccharomyces arboricola	SAAR	H-6	(Liti et al., 2013)
Saccharomyces cerevisiae	SACE	S288c	(Goffeau et al., 1996) (GCF_000146045.2)
Saccharomyces eubayanus	SAEU	FM1318	(Baker et al., 2015)
Saccharomyces kudriavzevii	SAKU		(Scannell et al., 2011)
Saccharomyces mikatae	SAMI	IFO 1815	(Scannell et al., 2011)
Saccharomyces paradoxus	SAPA		(Scannell et al., 2011)
Saccharomyces uvarum	SAUV		saccharomycessensustricto.org
Scheffersomyces stipitis	SCST	CBS 6054	(Jeffries et al., 2007)
Spathaspora passalidarum	SPPA	NRRL Y-27907	(Wohlbach et al., 2011)
Sympodiomyces attinorum	SYAT	NRRL Y-27639	JGI : *Sympodiomyces attinorum* NRRL Y-27639 v1.0
Tetrapisispora phaffii	TEPH	CBS 4417	(Gordon et al., 2011b)
Torulaspora delbrueckii	TODE	CBS 1146	(Gordon et al., 2011b)
Trichomonascus petasosporus	TRPE	NRRL YB-2093	JGI : *Trichomonascus petasosporus* NRRL YB-2093 v1.0
Vanderwaltozyma polyspora	VAPO	DSM 70294	(Scannell et al., 2007b)
Wickerhamomyces ciferrii	WICI	NRRL Y-1031 F-60-10	(Schneider et al., 2012)
Yarrowia lipolytica	YALI	CLIB122	(Dujon et al., 2004)
Zygosaccharomyces bailii	ZYBA	CLIB 213	(Galeote et al., 2013)
Zygosaccharomyces rouxii	ZYRO	CBS732	(Souciet et al., 2009)

Nous avons dans un second temps comparé le nombre de contigs ainsi que le contenu en gènes dans chacun des génomes rassemblés (Figure 27A) et exclu du jeu de données les génomes dont le nombre de contigs constitue une valeur extrême haute de la distribution afin de limiter selon un critère objectif le niveau de fragmentation des génomes. Nous avons également éliminé du jeu de données les génomes dont le nombre de gènes sont des extrêmes hauts et bas afin d'éliminer les génomes mal annotés. Enfin, nous avons calculé les meilleurs hits réciproques (« *Best Reciprocal Hits* », RBH) entre toutes les paires de génomes (Figure 27B). Les génomes partageant en moyenne moins de RBH avec tous les autres génomes ont été éliminés car un faible nombre de RBH impacterait la détection de la synténie et la qualité finale des génomes reconstruits. Il est intéressant de remarquer que plusieurs génomes éliminés (*Hanseniaspora uvarum*, *Hanseniaspora valbyensis*, *Dekkera bruxellensis*) appartiennent à des lignées comportant très peu d'espèces séquencées (*Saccharomycodaceae*, *Pichiaceae*), suggérant qu'un échantillonnage plus important de la diversité de ces lignées est nécessaire pour pouvoir exploiter ces génomes. On observe également des « *clusters* » riches en RBH conservés dans la matrice, supportant très bien la partition proposée dans l'étude (Dujon and Louis, 2017) et utilisée dans l'introduction pour présenter les *Saccharomycotina* : *Saccharomycetaceae,* Clade CTG (*Debaryomycetaceae*, *Metschnikowiaceae*), méthylotrophes (*Pichiaceae*, *Komagataella*), et lignées basales (clade *Yarrowia*).

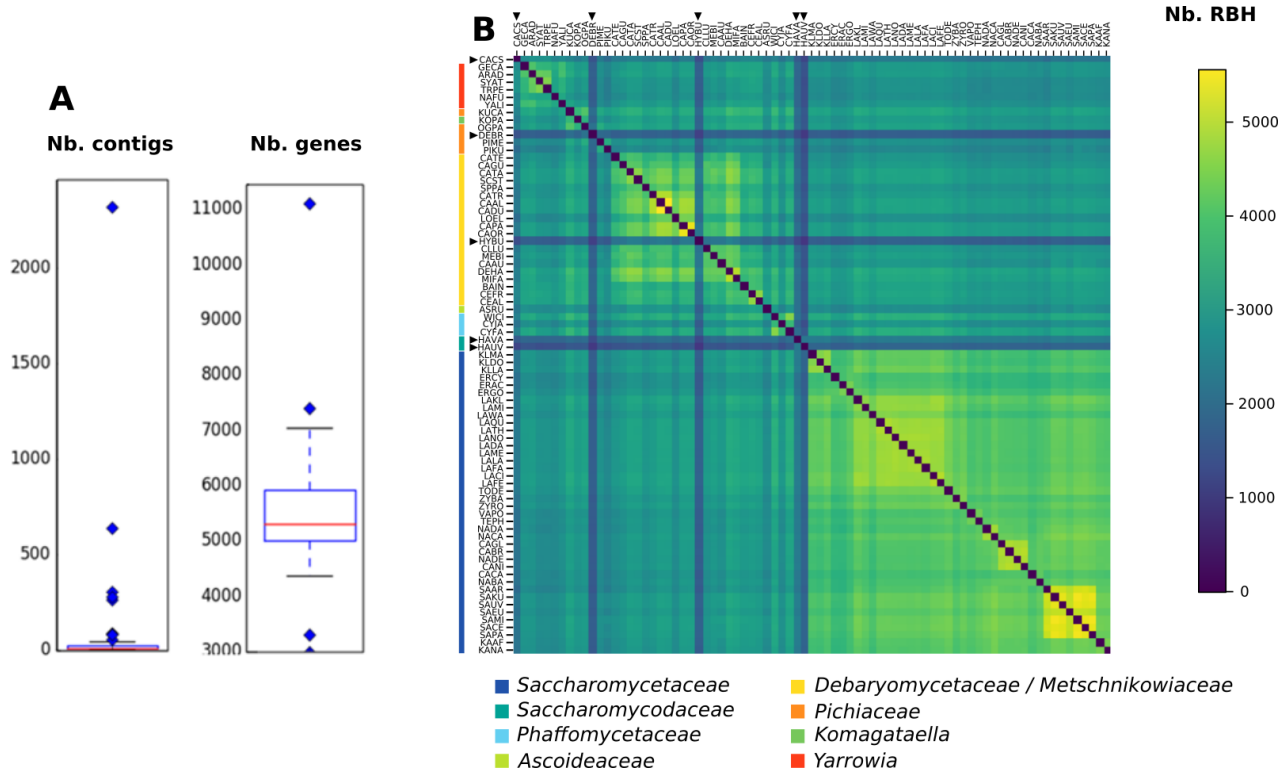


Figure 27 Filtrage des génomes du jeu de données. (**A**) Filtrage des génomes trop fragmentés et des génomes avec des nombres de gènes extrêmes. (**B**) Chaque point de la matrice représente le nombre de RBH partagés entre deux génomes selon un indice coloré allant du jaune (5600) au violet (0). La diagonale a été représentée en violet pour une meilleure visibilité. La grande lignée à laquelle appartient chaque espèce est représentée par la barre colorée sur le coté gauche de la matrice et par la légende en position inférieure (voir (Shen et al., 2016a)). Les génomes éliminés du jeu de données sont indiqués par des flèches noires. Les noms d'espèces sont indiqués par un code à quatre lettres (voir le Tableau 1).

Sur la base des trois critères définis précédemment, les génomes répertoriés dans le Tableau 2 ont été exclus du jeu de données à l'exception du génome de *Millerozyma farinosa*, qui a été conservé par erreur. Ce génome a été utilisé à l'étape suivante, c'est-à-dire la construction d'un arbre phylogénétique, mais n'a en revanche plus été utilisé ensuite, ni comme source d'information lors de la reconstruction des génomes ancestraux ni pour l'inférence des réarrangements passés. Le jeu de données final est donc constitué de 66 génomes.

Tableau 2 Génomes filtrés selon le nombre de contigs, de gènes, de RBH.

Espèce	Nb. contigs	Nb. gènes	Nb. RBH
Ascoidea rubescens	Elevé		
Candida auris	Elevé	Elevé	
Dekkera bruxellensis	Elevé		Faible
Hanseniaspora uvarum	Elevé	Faible	Faible
Hanseniaspora valbyensis	Elevé		Faible
Pichia kudriavzevii	Elevé		
Saccharomyces kudriavzevii	Elevé		
Saccharomyces mikatae	Elevé		
Saccharomyces uvarum	Elevé		
Vanderwaltozyma polyspora	Elevé		
Wickerhamomyces ciferrii	Elevé		
Hyphopichia burtonii		Faible	Faible
*Millerozyma farinosa**		Elevé	
Candida caseinolytica			Faible

Cet ensemble de génomes constitue donc un jeu de données de qualité suffisemment homogène pour reconstruire les génomes ancestraux des *Saccharomycotina.* Toutefois, avant d'entreprendre cette tâche, il nous a fallu construire un arbre phylogénétique.

4.2.2. Construction de l'arbre phylogénétique des 66 espèces de *Saccharomycotina*

Construire un arbre robuste est la base de toute approche de reconstruction évolutive. En particulier, lors de la reconstruction des génomes ancestraux, l'arbre va guider le choix des espèces de référence et les nœuds internes de l'arbre représentent les ancêtres qu'on souhaite reconstruire. Il est donc primordial de disposer d'une topologie correcte pour ne pas reconstruire des génomes ancestraux qui n'auraient jamais existé. Comme le montre l'étude (Shen et al., 2016a), parue après que notre arbre eut été finalisé, la topologie inférée à partir de données de génomes complets est remarquablement cohérente avec les topologies précédemment publiées reposant sur les séquences ribosomiques. Nous aurions donc pu utiliser une topologie préexistante pour guider la reconstruction des génomes ancestraux. Toutefois, des jeux de génomes différents sont susceptibles de faire varier les longueurs de branches de l'arbre et comme nous le verrons plus loin, nous aurons besoin des longueurs de branches sur notre propre jeu de données. Aussi nous avons reconstruit un arbre phylogénétique sur les 66 génomes de *Saccharomycotina* rassemblés que nous avons validé *a posteriori* avec l'arbre de l'étude (Shen et al., 2016a).

4.2.2.1. Identification des homologues synténiques

Pour construire l'arbre phylogénétique des 66 génomes, nous avons choisi d'utiliser les homologues synténiques non-dupliqués comme orthologues. Nous avons d'abord identifié les blocs de synténie avec *SynChro* (Drillon et al., 2014) avec le paramètre Δ=3. Pour rappel, Δ correspond au seuil de fragmentation maximum autorisé pour décider si deux régions synténiques proches entre deux espèces forment un seul bloc de synténie ou deux blocs distincts.

Nous avons recherché par transitivité les blocs de synténie ubiquitaires entre les 66 espèces à partir des blocs de synténie entre paires de génomes (Figure 28). Cela signifie que si des gènes du génome 1 sont conservés en synténie avec le génome 2 et que ces gènes du génome sont eux-mêmes conservés en synténie avec le génome 3, alors ont considère que les régions du génome 1 et 3 sont conservées en synténie. Nous avons ensuite extrait les gènes homologues conservés dans l'intersection de ces blocs ubiquitaires. Dans chaque région, les gènes conservés à travers tous les génomes sont des homologues synténiques. Au total, 28 homologues synténiques uniques sont partagés par les 66 espèces considérées (Tableau 3).

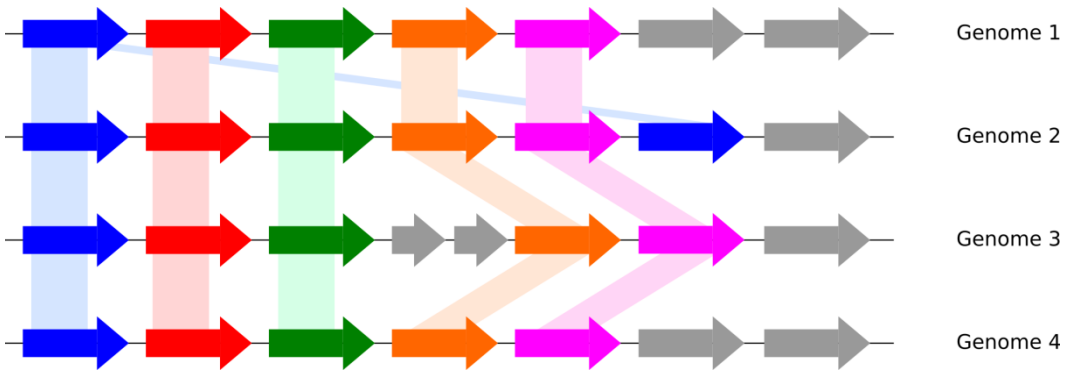


Figure 28 Détection des homologues synténiques uniques. Dans cet exemple, les gènes bleus, rouges, verts, orange et roses forment un bloc conservé d'homologues synténiques avec Δ=3. Le gène bleu présente un paralogue (à droite) mais celui-ci ne fait pas partie d'un bloc de synténie partagé par toutes les espèces et est donc exclu. Notons que l'insertion des deux gènes gris entre le gène vert et orange dans le génome 3 pourrait, si nous avions choisi Δ<2, conduire à la perte des gènes oranges et roses.

Tableau 3 Orthologues synténiques uniques des *Saccharomycotina* partageant 75% de similarité ou plus.

#	Identifiant	Symbole	Nom
1	YCR047C	BUD23	BUD site selection
2	YDL097C	RPN6	Regulatory Particle Non-ATPase
3	YDR148C	KGD2	alpha-KetoGlutarate Dehydrogenase
4	YDR375C	BCS1	ubiquinol-cytochrome c reductase (bc1) Synthesis
5	YDR454C	GUK1	GUanylate Kinase
6	YEL051W	VMA8	Vacuolar Membrane Atpase
7	YER025W	GCD11	General Control Derepressed
8	YFL038C	YPT1	Yeast Protein Two
9	YGL058W	RAD6	RADiation sensitive 519 YGL058W
10	YHR088W	RPF1	Ribosome Production Factor
11	YIL078W	THS1	THreonyl tRNA Synthetase
12	YJR064W	CCT5	Chaperonin Containing TCP-1
13	YJR065C	ARP3	Actin-Related Protein
14	YKL013C	ARC19	ARp2/3 Complex subunit
15	YKL056C	TMA19	Translation Machinery Associated
16	YKL170W	MRPL38	Mitochondrial Ribosomal Protein, Large subunit
17	YKR081C	RPF2	Ribosome Production Factor
18	YLR197W	NOP56	NucleOlar Protein of 56.8 kDa
19	YLR289W	GUF1	Gtpase of Unknown Function
20	YLR355C	ILV5	IsoLeucine-plus-Valine requiring
21	YOR136W	IDH2	Isocitrate DeHydrogenase
22	YOR157C	PUP1	PUtative Proteasome subunit
23	YOR261C	RPN8	Regulatory Particle Non-ATPase
24	YPL028W	ERG10	ERGosterol biosynthesis
25	YPL211W	NIP7	Nuclear ImPort 546 YPL211W
26	YPL235W	RVB2	RuVB-like 1416 YPL235W
27	YPR016C	TIF6	Translation Initiation Factor
28	YPR132W	RPS23B	Ribosomal Protein of the Small subunit

Nous avons ensuite recherché un enrichissement des « Gene Ontology terms » (GO-terms) pour les ces 28 gènes sur la base des annotations de *S. cerevisiae* à l'aide de l'outil « *SGD Gene Ontology Slim Mapper*» (https://www.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl) mis à disposition par le site web Saccharomyces Genome Database (SGD) (Tableau 4). Seuls les GO-terms dont la p-valeur est inférieure à 10^{-2} sont indiqués. On observe que les termes rapportés peuvent globalement être regroupés autour de deux grandes catégories que sont la catégorie « mitochondrie » et la catégorie « ribosome ». Il est intéressant de constater que ces gènes, conservés en synténie tout au long de l'évolution des *Saccharomycotina* sont impliqués dans ces processus cellulaires centraux de respiration et de traduction. Ceci suggère peut être l'importance de la localisation de ces gènes dans le génome.

Tableau 4 Enrichissement en GO-slim (Process) à partir des 28 homologues synténiques identifiés chez les *Saccharomycotina*. Les Go-slim se rapportant à la respiration/la mitochondrie sont indiqués en bleu. Les GO-slim se rapportant à la traduction/au ribosome sont indiqués en vert. Les GO-slim se rapportant au métabolisme protéique, ici la dégradation des protéines, sont indiqués en rouge.

GO term	Gene(s)	p-value
ribosomal subunit export from nucleus	BUD23, RPF1, TIF6	1, 1E-10
rRNA processing	BUD23, RPF1, RPF2, NOP56, NIP7, RVB2, TIF6, RPS23B	1, 9E-09
ribosomal large subunit biogenesis	RPF1, RPF2, NIP7, TIF6	4, 3E-08
regulation;translation	GCD11, GUF1, RPS23B	6, 2E-05
mitochondrion organization	KGD2, BCS1, ARP3, ARC19, ILV5	1, 2E-04
ribosome assembly	RPF1, RPF2	2, 6E-04
organelle inheritance	ARP3, ARC19	2, 6E-04
proteolysis involved in cellular protein catabolic process	RPN6, RAD6, PUP1, RPN8	7, 3E-04
cellular respiration	KGD2, IDH2	3, 4E-03
nuclear transport	BUD23, RPF1, TIF6	5, 6E-03
protein complex biogenesis	RPN6, BCS1, YPT1, ARP3	7, 1E-03

Nous avons ensuite inféré l'arbre phylogénétique des *Saccharomycotina* avec la méthode du maximum de vraisemblance sur le concaténât des 28 familles d'homologues synténiques présentés plus haut. L'alignement multiple des séquences protéiques a été réalisé avec l'algorithme Muscle 3.8.31 et les paramètres par défaut (Edgar, 2004), les parties de l'alignement de piètre qualité ont été nettoyées avec *Gblocks* 0.91 (Castresana, 2000), et l'arbre phylogénétique a été inféré avec *PhyML* 3.0 (Guindon et al., 2010) avec le modèle LG, s paramètres « -s best » et 100 *bootstraps*.

L'arbre phylogénétique obtenu est représenté Figure 29. Cet arbre est globalement bien soutenu par les valeurs de Bootstrap (Figure 29, insert). De plus, les espèces des quatre grands groupes présentés dans l'introduction : (i) *Saccharomycetaceae*, (ii) CTG clade, (iii) méthylotrophes et (iv) lignées basales sont correctement séparées. La topologie de cet arbre est très cohérente avec les topologies précédemment publiées, cependant, elle présente une erreur de branchement, indiquée en rouge. Afin de résoudre cette erreur, nous avons subdivisé l'arbre au niveau de son nœud supérieur de manière à former trois grands groupes d'espèces : (i) *Saccharomycetaceae*, (ii) le clade CTG et les méthylotrophes, (iii) les lignées basales. Sur la base des clusters d'espèces enrichis en RBH présentés Figure 27, cette partition devrait permettre de considérablement intensifier le signal phylogénétique et de reconstruire des sous-arbres mieux résolus.

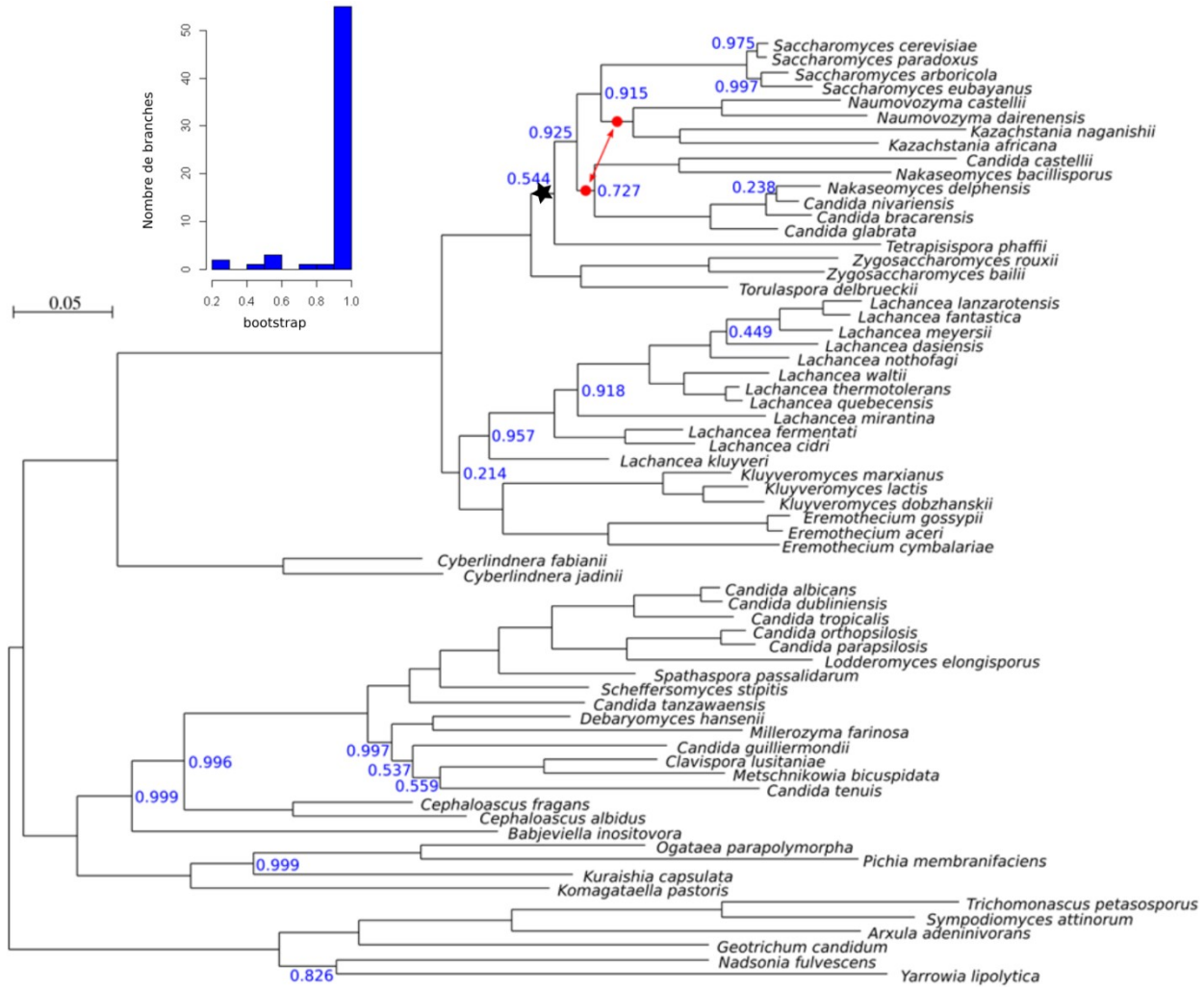


Figure 29 Arbre phylogénétique des *Saccharomycotina* obtenu à partir de 18 homologues synténiques uniques. L'étoile noire indique la duplication totale du génome dans le clade des *Saccharomycetaceae*. L'erreur de branchement des clades *Naumovozyma/Kazachstania* et *Candida/Nakaseomyces* est indiquée en rouge. Les valeurs de bootstrap différentes de 100% sont indiquées en bleu.

4.2.2.2. Amélioration du signal phylogénétique et construction de trois sous-arbres

Les blocs de synténie ayant été précédemment identifiés avec *SynChro*, les homologues synténiques uniques ont directement été identifiés par transitivité dans chaque sous-arbre. Les *Saccharomycocetaceae* partagent 354 homologues synténiques, les espèces du CTG clade et méthylotrophes 134 et les espèces des lignées basales 348. Les homologues synténiques cartographiés sur le génome de *S. cerevisiae* pour l'arbre 1, sur le génome de *C. albicans* pour l'arbre 2 et sur le génome de *Y. lipolytica* pour l'arbre 3 sont uniformément répartis entre les chromosomes (Figure 30, Figure 31, Figure 32) et au sein des chromosomes (non-représenté).

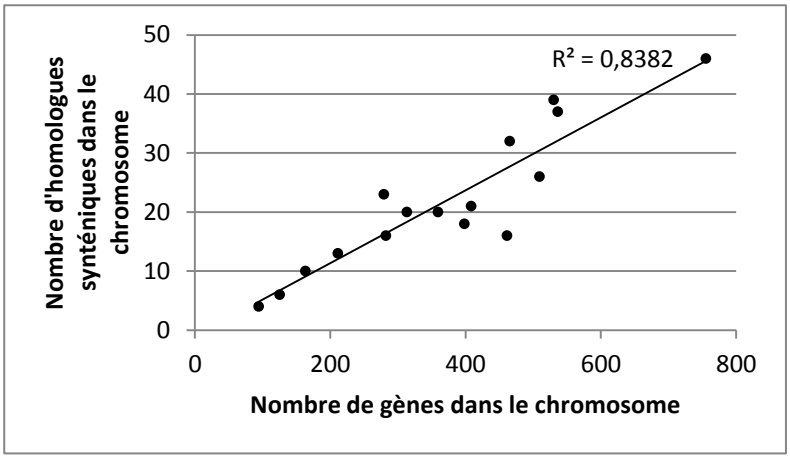


Figure 30 Nombre d'homologues synténiques partagés chez les *Saccharomycetaceae* cartographiés sur le génome de *S. cerevisiae*. Les abscisses représentent la taille des chromosomes (en nombre de gènes), les ordonnées le nombre d'homologues synténiques correspondant. *La p-valeur est de $6{,}5.10^{-7}$.

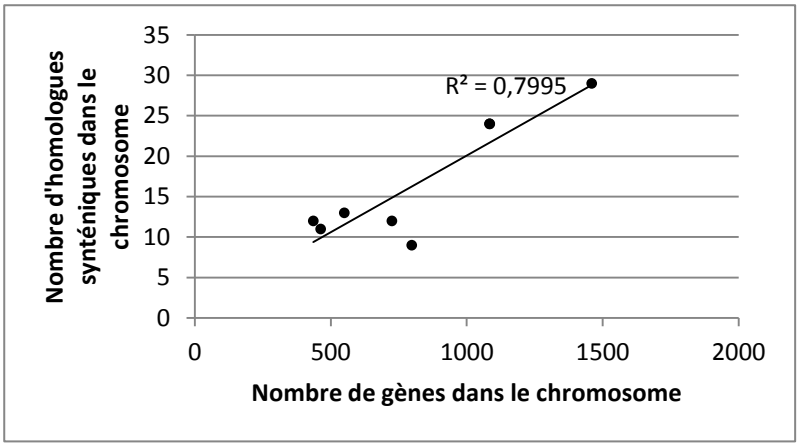


Figure 31 Nombre d'homologues synténiques cartographiés chez les espèces du clade CTG et méthylotrophes mappés sur le génome de *C. albicans*. Les abscisses représentent la taille des chromosomes (en nombre de gènes), les ordonnées le nombre d'homologues synténiques correspondant. *La p-valeur est de 3.10^{-3}.

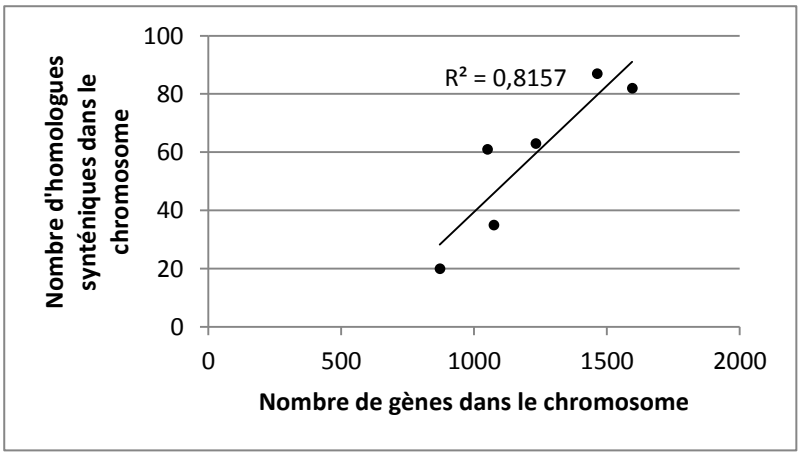


Figure 32 Nombre d'homologues synténiques partagés chez les espèces des lignées basales cartographiés sur le génome de *Y. lipolytica* Les abscisses représentent la taille des chromosomes (en nombre de gènes), les ordonnées le nombre d'homologues synténiques correspondant. *La p-valeur est de 1.10^{-2}.

L'enrichissement en GO-slim associés aux homologues synténiques des *Saccharomycocetaceae* a été recherché comme précédemment avec *SGD Gene Ontology Slim Mapper* (Tableau 5). Comme on peut le voir, les catégories ribosomes/transcription, mitochondrie/respiration et métabolisme des protéines sont à nouveau observés. En outre, de grandes catégories fonctionnelles nouvelles sont apparues comme la transcription et la modification des ARN (notamment des ARNt), l'adressage cellulaire et le transport nucléaire. En outre, des termes englobant diverses activités métaboliques (en gris) sont apparus. La diversification des GO-terms avec l'augmentation du nombre d'homologues synténiques était un phénomène attendu. Toutefois il est intéressant d'observer que des gènes homologues représentatifs de la diversité des espèces qu'on compare sont impliqués dans des processus cellulaires également diversifiés.

Tableau 5 Enrichissement en GO-slim (Process) à partir des 354 homologues synténiques identifiés chez les *Saccharomycotina.* Les Go-slim se rapportant à la respiration/la mitochondrie sont indiqués en bleu. Les GO-slim se rapportant à la traduction/au ribosome sont indiqués en vert. Les GO-slim se rapportant au métabolisme protéique sont indiqués en rouge. La modification de l'ARN est en violet, la transcription est en orange, l'adressage cellulaire est en jaune et le métabolisme des cofacteurs (et autres métabolismes) est en gris.

GO-slim	gènes	p-value
cellular amino acid metabolic process	CYS3, LYS2, SHM1, LEU2, IDP1, LYS4, PRO1, SAH1, ARG5, 6, TRP2, MET6, FRS2, VAS1, CYS4, YHR020W, PUT2, THS1, MET3, CPA2, MET14, TRP3, FRS1, ILV5, IDH1, FPR1, MET2, WRS1, IDH2, GLN4, SER1, CPA1, PRO2, HTS1, GLN1	8, 55427E-13
cellular respiration	ETR1, KGD2, RIP1, COX4, COX13, QCR9, LSC2, COX6, KGD1, MDH1, ACO1, NDI1, IDH1, CYT1, IDH2, LSC1, COX11	3, 05857E-09
ribosomal large subunit biogenesis	MAK16, SPB1, RSA4, RLI1, NSA2, PRP43, NOP7, RPF1, RPF2, RIX7, RLP24, AFG2, NOP2, RPL3, NOC2, RRS1, RPL5, NIP7, TIF6	6, 05861E-09
nuclear transport	BUD23, YRB1, RLI1, MSN5, ARB1, GLE2, KAP123, SRM1, SEH1, RPS2, RPS26A, CSE1, CRM1, RPF1, NMD3, SSL2, RPB4, NPA3, RIX7, GPN3, KAP95, PSE1, YDJ1, RPS15, GPN2, RRS1, TIF6	2, 19589E-08
nucleobase-containing small molecule metabolic process	CDC19, ADE1, ATP1, ATP3, ATP16, TPI1, GUK1, URA3, RIP1, SAH1, COX4, ADE5, 7, ADE6, QCR9, ADE3, PFK1, COX6, QNS1, GUT2, RPE1, ATP2, ADE13, NDI1, ADE4, CYT1, RKI1, SER1	5, 16943E-07
ribosomal subunit export from nucleus	BUD23, RLI1, ARB1, SRM1, CRM1, RPF1, NMD3, RIX7, RRS1, TIF6	1, 07327E-06
transcription from RNA polymerase III promoter	RPB5, RPC11, SPT15, CKB1, DST1, RPC10, CKB2, RPB10, RPB8	1, 83725E-06
mitochondrion organization	ATG8, RIM2, NFS1, MIC10, GGC1, KGD2, BCS1, TIM9, FMP32, RPN11, COX13, MSP1, PHB1, QCR9, TIM10, MAS2, ARC15, FIS1, TIM17, PET191, ARP3, MAS1, ACO1, ILV5, TOM40, YDJ1, TIM23, MGM1, MGR2, CBP3, NAT3	4, 10518E-06
rRNA processing	MAK16, SPB1, KRR1, BUD23, PWP2, FAP7, NHP2, RRP45, SNU13, NSA2, CGR1, PRP43, NOP7, RPS20, SSZ1, RRP4, RPF1, IMP3, RPF2, CBF5, PWP1, NOP56, UTP21, UTP15, NOP2, IMP4, RCL1, PNO1, RRS1, NIP7, RVB2, DIM1, TIF6, RPS23B	1, 19589E-05
tRNA aminoacylation for protein translation	FRS2, VAS1, YHR020W, THS1, FRS1, WRS1, GLN4, HTS1	1, 72348E-05
ribosomal small subunit biogenesis	KRR1, BUD23, PWP2, FAP7, SNU13, PRP43, NOP7, RPS20, IMP3, UTP21, UTP15, IMP4, KRE33, RCL1, PNO1, RRS1, DIM1, RPS23B	4, 13212E-05
generation of precursor metabolites and energy	CDC19, ETR1, TPI1, KGD2, RIP1, COX4, COX13, QCR9, PFK1, LSC2, COX6, KGD1, MDH1, ACO1, NDI1, IDH1, CYT1, IDH2, LSC1, COX11	4, 17928E-05
translational initiation	RLI1, GCD6, GCD11, PAB1, RPB4, GCN3, TIF11, CDC33, SUI3, TIF5	7, 67701E-05
transcription from RNA polymerase I promoter	RPB5, SPT15, CKB1, DST1, RPC10, UTP15, RPA49, CKB2, RPB10, RPB8, RPA135	0, 000154428
regulation of translation	RLI1, GCD6, SKP1, GCD11, PAB1, GCN20, RPS2, SSZ1, RPB4, GCN3, GUF1, CBP3, TIF5, RPS23B	0, 000159857
proteolysis involved in cellular protein catabolic process	RPT2, RPN6, RPN5, YRB1, UBC1, SKP1, PRE1, RPN11, PRE4, RAD6, PRE9, PUP2, RPN1, VPS24, RPT1, UBC7, YDJ1, PRE6, HRT1, PUP1, CDC31, RPT4, RPN8, NAT3	0, 000207921
DNA-templated transcription, termination	SUP45, RPC11, DST1, RAD6, RPB3, SWD2, YSH1	0, 001674533
ribosome assembly	RSA4, RPS26A, RPF1, RPF2, SDO1, RPL3, RPL5, TIF5	0, 002642344
protein targeting	ATG8, GET3, YRB1, TIM9, KAP123, YPT1, MSP1, TIM10, MAS2, SEC11, TIM17, NPA3, VPS1, YPT52, DID2, MAS1, GPN3, KAP95, TOM40, PSE1, YDJ1, TIM23, VPS68, GPN2, ARL3, MGR2, SRP54, BET2	0, 003181138

cellular ion homeostasis	VMA2, NFS1, GGC1, SKP1, NHX1, VMA8, VMA7, PFK1, VMA16, VMA5, SMF3, VMA6, ATX1, VMA11	0, 00359578
cofactor metabolic process	CDC19, COQ1, PDX3, NFS1, HEM3, TPI1, ADE3, PFK1, LSC2, QNS1, CFD1, GUT2, RPE1, NDI1, RKI1, CAT5, LSC1, SPE3	0, 004356494
nucleobase-containing compound transport	RIM2, GGC1, YRB1, MSN5, GLE2, SRM1, RPS2, RPS26A, CRM1, SSL2, RPB4, VPS24, PSE1, YDJ1, RPS15	0, 005153217
DNA-templated transcription, initiation	TAF10, SPT15, DST1, TFG2, SSL2, RPB4, RPT1, TAF9, RPT4	0, 005777855
RNA modification	NFS1, SPB1, BUD23, TRM8, NHP2, TRM1, THG1, CFD1, GCD14, TCD2, CBF5, NOP56, IKI3, NOP2, TRM11, ELP3, DIM1	0, 007390246

Pour le second sous-arbre, regroupant le CUG clade et les méthylotrophes, nous avons recherché un enrichissement en termes GO-slim en utilisant la liste des homologue synténiques identifiés chez *Candida albicans* à l'aide de l'outil *CGD Gene Ontology Slim Mapper* rendu disponible par *Candida Genome Database* (http://www.candidagenome.org/cgi-bin/GO/goTermMapper) (Tableau 6). De manière surprenante d'autres termes apparaissent, en particulier les termes relatifs au cytosquelette et au transport par vésicules ainsi que les termes « response to chemical » et « response to drug » potentiellement en lien avec le caractère pathogène de plusieurs espèces de ce groupe.

Tableau 6 Enrichissement en GO-slim (Process) à partir des 134 homologues synténiques identifiés chez les espèces du CUG clade et méthylotrophes.

GO term	Gene(s)	p-value
ribosome biogenesis	RVB2, C1_09710C_A, RPL11, RPF2, RPF1, SSB1, C3_07230W_A, RPS23A, BUD23, C1_12350W_A, RCL1, CR_07080W_A, RPL12, CR_00460C_A, SIK1, UTP15, RRS1, C4_06210C_A, NIP7, RPS18, RPS1	3,60729E-09
translation	GCD11, MES1, RPL16A, DED81, RPL11, RPP1A, C5_04900C_A, SSB1, RPS23A, MRPL19, GUS1, ACO2, RPL28, RPS12, RPL13, RPL12, TMA19, RPL10A, GCN3, RPS18, DRG1, RPS1, SES1	2,67628E-07
organelle organization	YPT31, RVB2, ARC19, CR_05970C_A, YKT6, HSP104, PFY1, C6_02190C_A, RPL11, RPF2, YPT1, RPF1, C3_07230W_A, ARP2, TUB1, TIM44, TIM17, RAD6, TIM9, CYP1, RPT6, ARC18, CDC3, TOP1, ACO1, ARP3, CDC48, CR_08560C_A, ADA2, CR_07080W_A, RIM2, RPL12, SIT4, SAR1, RRS1, C4_04920W_A, VPS2, NIP7, ACT1, SEC4, PR26, ILV5, HSP90, TIM22	6,1998E-07
response to chemical	YPT31, ARC19, SOD2, HSP104, RPF2, PHO85, YPT1, C3_07230W_A, ARP2, RPB7, RPS23A, TUB1, RAD6, ARC18, ARP3, FUR1, CDC48, ADA2, TPK2, FAS1, TRR1, RPN8, TMA19, SIT4, CKA2, UTP15, RRS1, C4_06210C_A, VPS2, YMC1, ACT1, MKC1, SEC4, HSP90	3,25397E-06
RNA metabolic process	MES1, RVB2, C1_09710C_A, DED81, RPF2, YPT1, RPF1, C2_07200W_A, SSB1, RPB7, RPS23A, GUS1, RAD6, BUD23, RPT6, TOP1, CDC48, C1_12350W_A, RCL1, CR_07080W_A, CR_00460C_A, SIT4, SIK1, RPA12, UTP15, RRS1, C4_04920W_A, C4_06210C_A, NIP7, RPS18, MKC1, RPC40, RPS1, SES1	4,56868E-06
response to drug	YPT31, ARC19, RPF2, YPT1, ARP2, RPB7, TUB1, ARC18, ARP3, FUR1, ADA2, FAS1, RPN8, CKA2, UTP15, RRS1, C4_06210C_A, VPS2, YMC1, MKC1, SEC4, HSP90	5,4143E-06
cell development	C2_08270C_A, C6_02190C_A, PHO85, ARP2, C2_00400C_A, CYP1, TPK2, ACT1, MKC1, SEC4, CR_10140W_A	5,59051E-06
transport	YPT31, SEC21, YKT6, C2_08270C_A, PFY1, FEN12, YPT1, RPF1, SSB1, C3_07230W_A, ARP2, RPS23A, TUB1, TIM44, NHX1, ACC1, TIM17, TIM9, BUD23, ARC18, C5_00150C_A, C1_03370W_A, TOP1, C4_02460W_A, C2_09970C_A, CDC48, C1_12350W_A, CR_07080W_A, VMA11, RIM2, SAR1, MSN5, RRS1, C4_06210C_A, VPS2, RPS18, YMC1, ACT1, SEC4, TIM22	1,74397E-05
cytoskeleton organization	YPT31, ARC19, PFY1, ARP2, TUB1, ARC18, CDC3, ARP3, CDC48, CR_08560C_A, SIT4, ACT1	4,51899E-05
protein catabolic process	RPN6, CDC34, C6_02190C_A, RAD6, RPT6, CDC48, PRE10, C4_02470C_A, PRE2, PUP1, VPS2, PR26	0,000330965
vesicle-mediated transport	YPT31, SEC21, YKT6, C2_08270C_A, FEN12, YPT1, C3_07230W_A, ARP2, ARC18, C4_02460W_A, CDC48, SAR1, VPS2, ACT1, SEC4	0,000757728

Pour le troisième sous-arbre, seul un très faible nombre de gènes a pu être cartographié sur le génome de *Yarrowia lipolytica* avec le serveur PANTHER (http://www.pantherdb.org/) et les résultats ne sont pas présentés.

Pour chaque groupe, le concaténât des homologues synténiques a été utilisé pour inférer un arbre phylogénétique avec PhyML selon la méthode décrite précédemment. Les trois sous-arbres sont représentés Figure 33. On remarque que l'erreur de branchement dans la phylogénie des *Saccharomycetaceae* a été corrigée. Ces trois sous-arbres constituent une base solide pour la reconstruction des génomes ancestraux qui correspondent aux nœuds internes de l'arbre.

Par conséquent nous avons 36 ancêtres à reconstruire dans le sous-arbre des *Saccharomycetaceae* (« arbre 1 »), 20 ancêtres dans l'arbre du clade CTG et méthylotrophes (« arbre 2 ») et 4 ancêtres dans l'arbre des lignées basales (« arbre 3 »). Ces ancêtres sont numérotés sur la Figure 33.

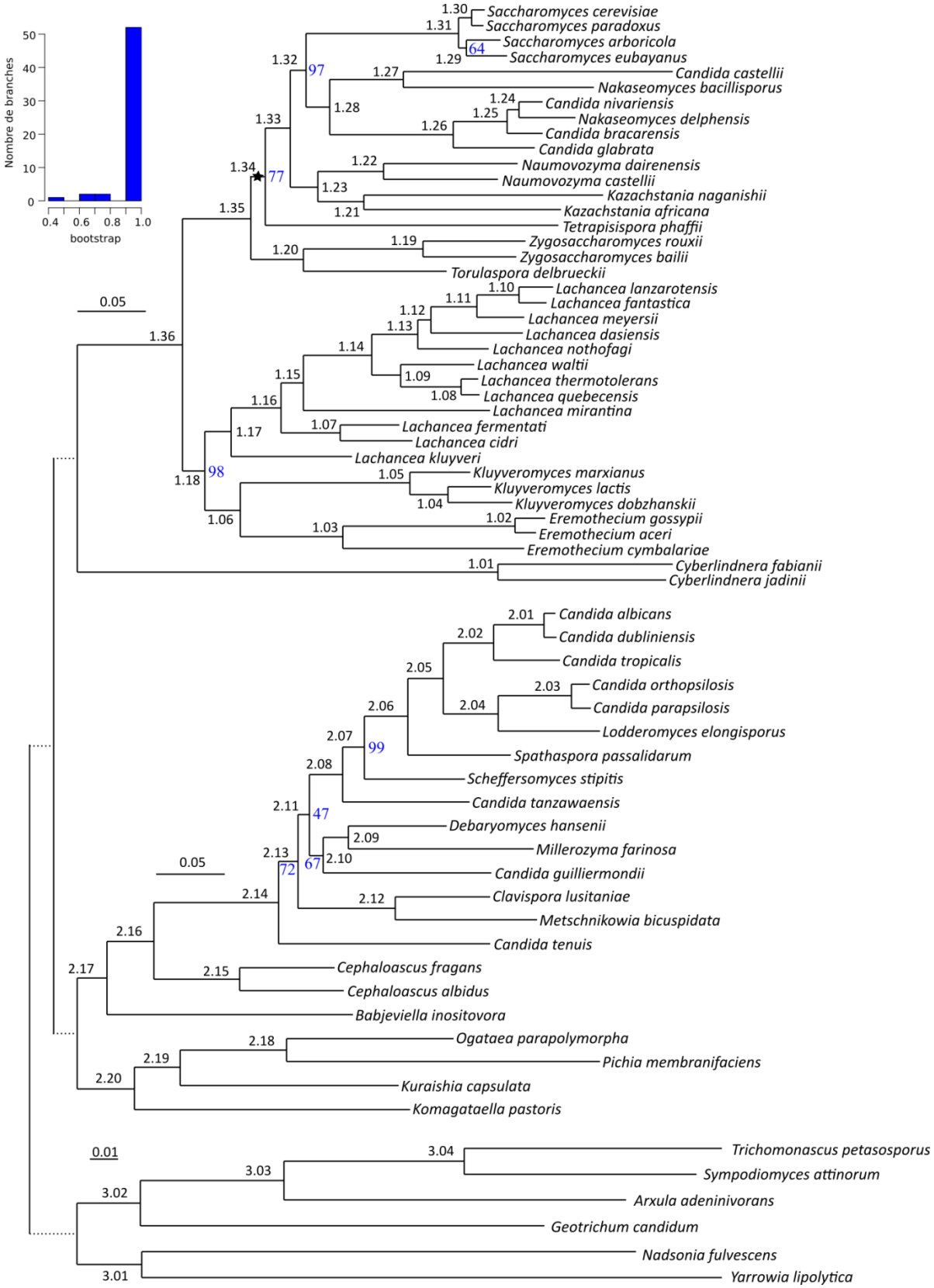


Figure 33 Arbre phylogénétique des *Saccharomycotina*. La séparation entre les trois sous-arbres est indiquée par des branches en pointillés. La duplication totale du génome est indiquée par l'étoile noire. Les valeurs de bootstrap inférieures à 100 sont indiquées en bleu. Les ancêtres de l'arbre 1, 2 et 3 sont respectivement numérotés 1.01 à 1.30, 2.01 à 2.20 et 3.01 à 3.04.

4.2.2.3. Construction d'arbres phylogénétiques à partir d'autres signaux

L'analyse des séquences protéiques des homologues synténiques nous a permis d'obtenir un arbre phylogénétique robuste et cohérent pour reconstruire les génomes ancestraux. Toutefois, comme nous avons remarqué dans la matrice Figure 27, le nombre de RBH entre les paires de génomes de notre jeu de données forment des clusters suggérant la présence d'un signal phylogénétique exploitable pour reconstruire un arbre phylogénétique. Nous avons donc calculé une matrice de distance à partir de la matrice Figure 27 et reconstruit un arbre phylogénétique avec les paramètres par défaut de l'algorithme Fitch de PhyML. L'arbre obtenu est présenté Figure 34.

Comme on peut le voir, la topologie de l'arbre obtenu est assez cohérente, avec 60% de « *splits* » corrects (en vert dans la Figure 34). Un split est une bipartition des espèces actuelles obtenue en scindant une branche de l'arbre. L'arbre des *Saccharomycetaceae* est reconstruit avec seulement 10 erreurs. La plupart des erreurs se trouvent dans l'arbre des CUG et méthylotrophes dans lequel seulement 7 splits ont été inférés correctement, probablement en raison des nombreuses branches internes courtes, qui sont mal résolues. L'arbre des lignées basales est correct à une erreur de branchement près. On remarque *a posteriori* que le nombre de RBH dans la matrice Figure 27, page 101 parait plus résolutif (plus « contrasté ») et globalement plus élevé entre les *Saccharomycetaceae* que dans les autres sous-arbres, expliquant la différence de qualité entre les trois sous-arbres. Il est intéressant de voir que le répertoire de gènes en lui-même permet déjà de générer un arbre qui ait du sens. Cela témoigne de la dynamique importante du répertoire de gènes dans l'arbre des *Saccharomycotina*.

Comme nous l'avons expliqué dans l'introduction, les réarrangements s'accumulent au cours du temps en faisant progressivement diverger la structure des génomes. Nous avons cherché à reconstruire un arbre phylogénétique à partir de ce signal. Pour cela, nous avons analysé les blocs de synténie identifiés par *SynChro* (Δ=3) entre les espèces actuelles avec le logiciel *PhyChro*. Le principe de cet outil est de séparer de manière itérative les génomes actuels en deux groupes sur la base d'adjacences de gènes incompatibles. Un arbre est ensuite reconstruit selon une approche *bottom-up* en regroupant les génomes qui minimisent le nombre d'adjacences incompatibles. *PhyChro* estime les longueurs de branches à partir du nombre de réarrangements inférés entre les génomes des espèces actuelles. L'arbre phylogénétique obtenu avec *PhyChro* est présenté Figure 35. On voit que les trois grands groupes d'espèces (sous-arbre 1, 2 et 3) sont correctement séparés, de même que les *Saccharomycetaceae* issues de la duplication totale du génome et les espèces n'ayant pas été dupliquées. On observe cependant un grand nombre de branches très courtes et des embranchements multiples, indiquant que l'identification des réarrangements n'est pas toujours très résolutive pour construire un arbre. Néanmoins, nous avons comparé les splits de l'arbre généré par *PhyChro* aux splits de l'arbre obtenu à partir de l'analyse des homologues synténiques. Les résultats sont représentés sous forme d'un cladogramme, pour plus de lisibilité (Figure 36). On observe, par rapport à l'arbre obtenu à partir des homologues synténiques, seulement 15 différences dans l'ensemble de l'arbre : 9 splits incorrects et 6 embranchements multiples. Les embranchements multiples sont liés à l'absence d'information plutôt qu'à de véritables erreurs, le taux de bipartitions correctes est donc de 86%.

Les arbres inférés à partir de ces deux signaux indépendants, issus de l'accumulation des mutations non-synonymes ou des réarrangements chromosomiques, fournissent donc des topologies remarquablement concordantes. Cela renforce la validité de notre topologie. Cependant, on voit que les longueurs de branches relatives entre ces deux arbres ne sont pas corrélées (insert dans la Figure 36).

Toutefois, l'analyse directe de la synténie sur les génomes actuels fournit souvent une approximation peu précise du nombre de réarrangements passés, notamment en raison de la réutilisation des points de cassure au cours de l'évolution des génomes (Guénola Drillon, thèse). Afin d'inférer les réarrangements passés de manière plus fiable, il faut d'abord reconstruire les génomes ancestraux, puis les comparer entre eux.

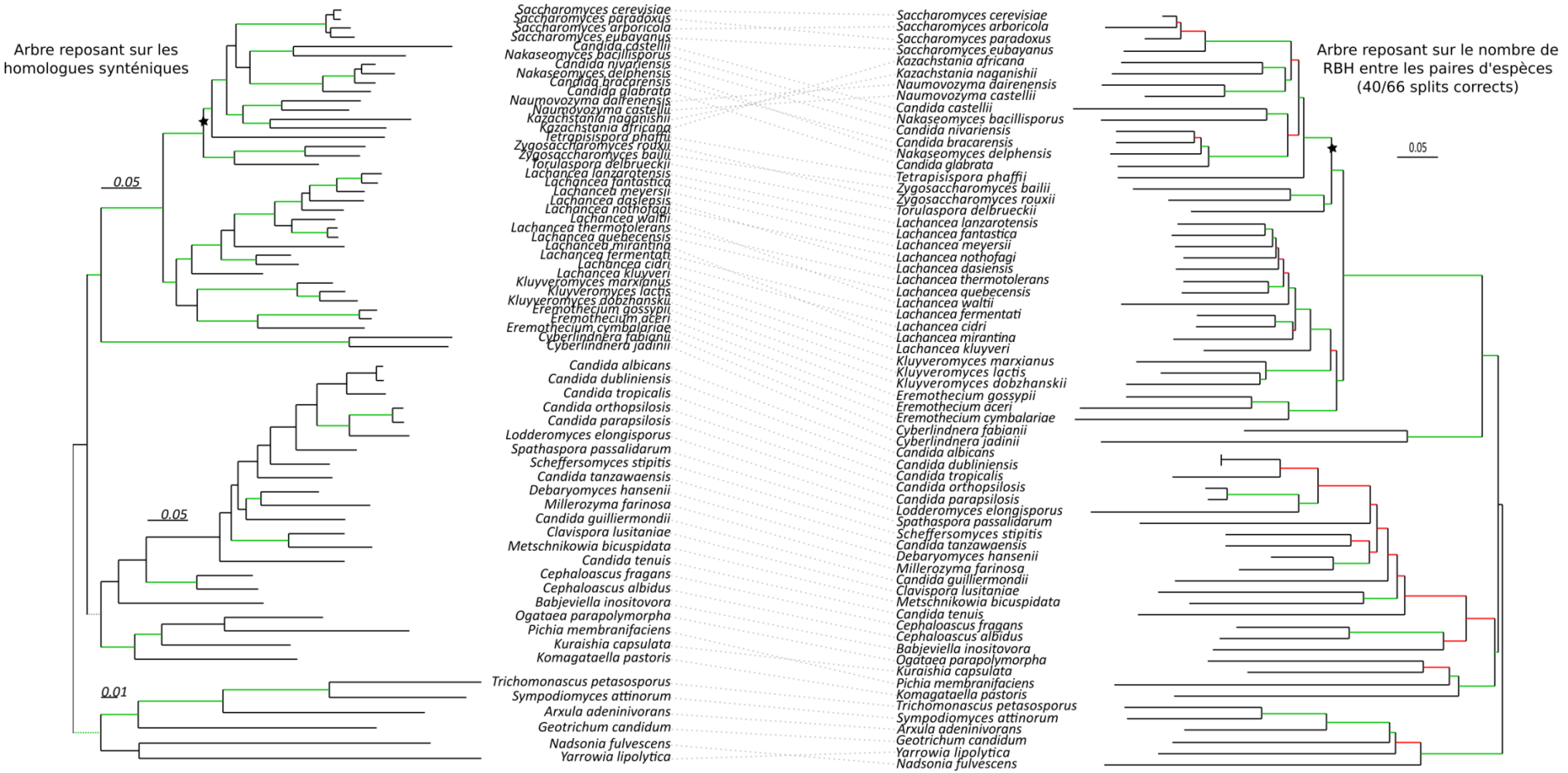


Figure 34 Comparaison entre l'arbre phylogénétique obtenu à partir de l'analyse des séquences protéiques des homologues synténiques (à gauche) et l'arbre obtenu en calculant une matrice de distance à partir du nombre de RBH entre les paires d'espèces, sans analyser leur séquence (à droite). La duplication totale du génome est indiquée par l'étoile noire. Les espèces actuelles ont été reliées en gris pour faciliter la comparaison des deux arbres. L'arbre reconstruit à partir du nombre de RBH présente 40 splits corrects et donc partagés avec l'arbre de gauche (en vert). Les splits faux sont indiqués en rouge. Les branches menant aux espèces actuelles n'entrent pas dans la comparaison des splits et sont représentées en noir.

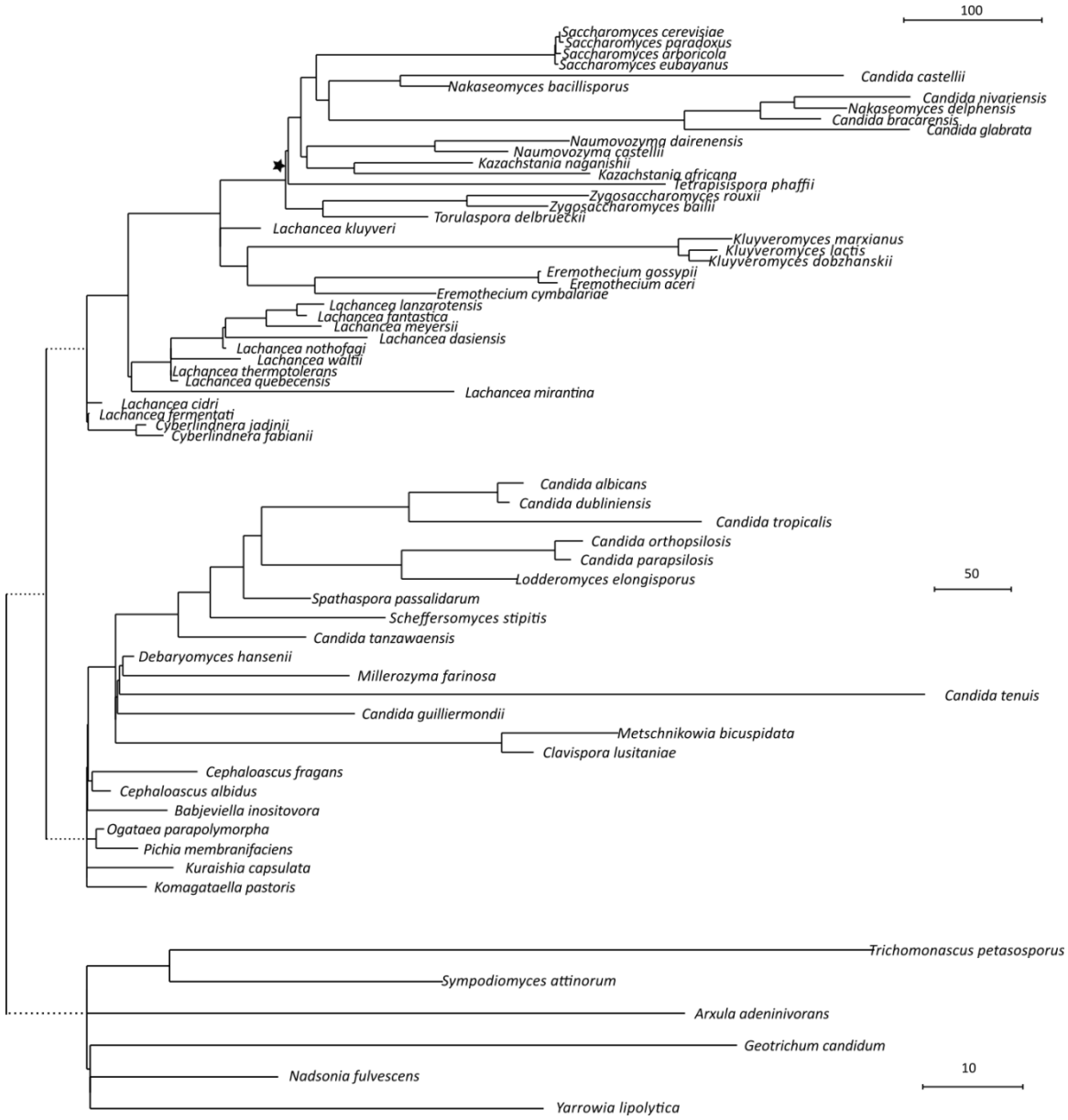


Figure 35 Arbre phylogénétique des 66 espèces généré par *PhyChro*. Les longueurs de branches correspondent à des nombres de réarrangements inférés entre les paires d'espèces actuelles. La duplication totale du génome est indiquée par l'étoile noire.

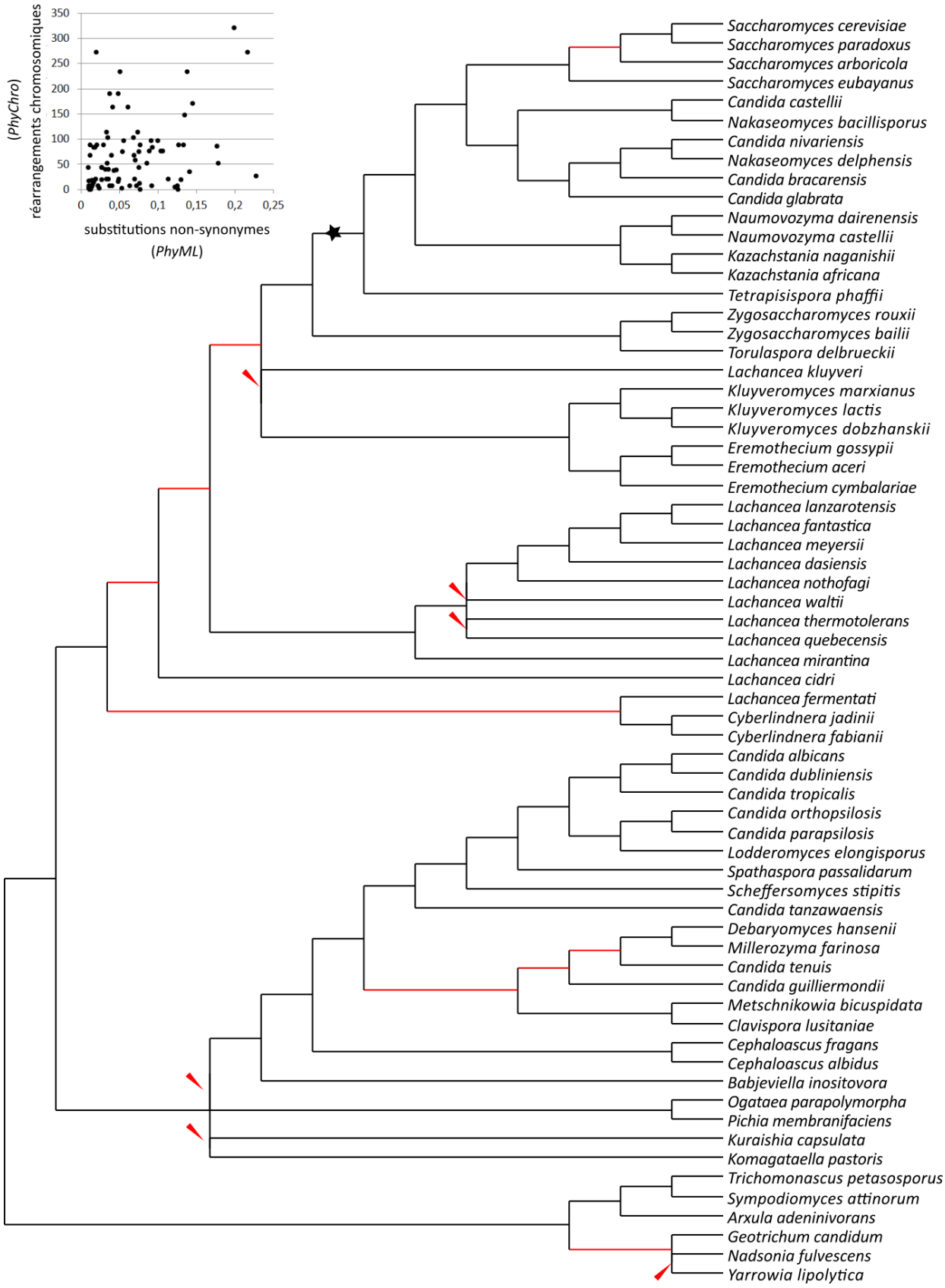


Figure 36 Splits corrects et incorrects dans la topologie de l'arbre généré par *PhyChro.* Les splits faux sont représentés par les branches en rouge, les embranchements multiples sont indiqués par des flèches rouges. L'insert représente la comparaison entre les longueurs de branches de l'arbre inféré à partir des homologues synténiques (substitutions non-synonymes) et les longueurs de branches inférées par *PhyChro* (nombre de réarrangements inférés directement entre les génomes des espèces actuelles). Seules les branches correspondant à des splits communs aux deux arbres sont comparées.

4.2.3. Reconstruction des génomes ancestraux

4.2.3.1. Choix des génomes actuels

Nous utiliserons l'arbre obtenu à partir de l'analyse des homologues synténiques pour reconstruire les génomes ancestraux des 66 génomes de notre jeu de données. *AnChro* peut reconstruire plusieurs versions d'un même génome ancestral selon les espèces actuelles qu'on compare (Figure 37).

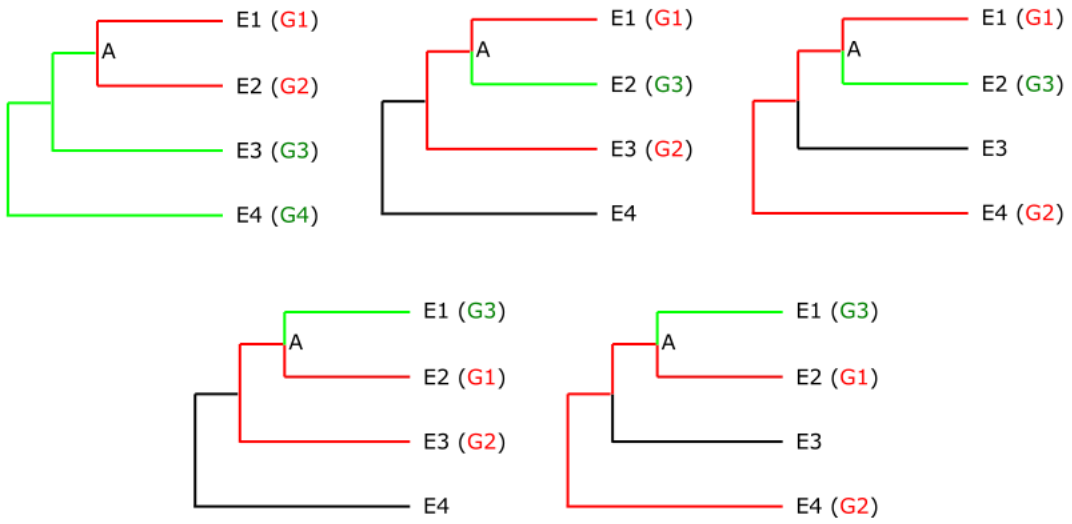


Figure 37 Choix des génomes actuels pour reconstruire un génome ancestral A. On peut effectuer autant de reconstructions de A qu'il existe de couples d'espèces reliées par un chemin évolutif passant par A, ici 5. Les génomes G1 et G2 sont les génomes comparés et les génomes G3, …, Gn sont les génomes référents servant de points de comparaison externe. Un chemin évolutif passant par A implique nécessairement qu'un des deux génomes G1 ou G2 soit un descendant de A.

De manière générale, soit A un ancêtre dont sont issues trois branches menant respectivement à N1, N2 et N3 espèces actuelles, le nombre de combinaisons d'espèces (G1, G2) possibles pour reconstruire l'ancêtre A vaut N1*N2 + N2*N3 + N1*N3. En décidant de travailler dans trois sous-arbres indépendants, on réduit déjà beaucoup le nombre de combinaisons d'espèces pour reconstruire chaque génome ancestral. Mais il en reste tout de même un grand nombre. Rien que pour l'ancêtre 1.01 dans l'arbre des *Saccharomycetaceae* (voir Figure 33) cela représente 73 combinaisons, sachant de plus qu'avec une même combinaison de génomes G1/G2/G3…Gn, *AnChro* reconstruit 36 versions d'un même ancêtre selon les valeurs de Δ et Δ' utilisées. Il y a donc théoriquement 6153 reconstructions à calculer pour l'ensemble des 36 génomes ancestraux des *Saccharomycetaceae*, 1623 reconstructions pour les 20 ancêtres du clade CTG/méthylotrophes et 40 reconstructions pour les 4 ancêtres des lignées basales. A raison de 5 minutes par reconstruction, cela représente pratiquement un mois de calcul. Nous avons donc cherché un moyen de réduire le nombre de reconstructions.

Comme la reconstruction des génomes ancestraux repose sur l'interprétation du graphe d'adjacences associé à la comparaison G1/G2, la comparaison entre un génome dupliqué (issu d'un événement de duplication totale) et un génome non-dupliqué rend le graphe aberrant car ces graphes ne peuvent en effet rendre compte que des inversions, translocations, fusions et fissions mais pas des événements de duplication totale. Il faut donc choisir G1 et G2 soit tous deux dupliqués, soit tous deux non-dupliqués. Dans l'arbre des *Saccharomycotina*, seuls les *Saccharomycetaceae* ont connu un événement de duplication totale

du génome. En appliquant la contrainte de séparation des génomes dupliqués et non-dupliqués au sous-arbre 1, on diminue dans ce groupe d'espèces le nombre de combinaisons possibles de 6153 à 2322.

Cependant, il n'est pas nécessaire de toutes les calculer car la qualité de la reconstruction de chaque génome ancestral dépend de la proximité des espèces comparées par *ReChro* qui résout les chemins et les cycles du graphe d'adjacences. De ce fait, plus G1 et G2 sont divergés, plus les réarrangements se sont accumulés entre eux et plus le génome est difficile à reconstruire. Pour optimiser *a priori* les reconstructions des génomes ancestraux des *Saccharomycotina*, nous avons choisi les espèces G1 et G2 à comparer pour reconstruire chaque ancêtre.

Les blocs de synténie étant la trace des réarrangements chromosomiques passés, on peut être tenté de sélectionner pour reconstruire un ancêtre, G1 et G2 comme les deux génomes partageant le moins de blocs de synténie. Or si l'on représente pour chaque paire de génomes des *Saccharomycotina* le nombre de blocs de synténie (calculés avec Δ=3) en fonction de la divergence protéique moyenne des deux génomes (Figure 38, page 119) on observe deux tendances. Jusqu'à 38% de divergence environ, les génomes accumulent des mutations non-synonymes (la divergence protéique augmente) et le nombre de réarrangements croissants augmente le nombre de blocs de synténie. A partir de 38% de divergence, les réarrangements chromosomiques s'accumulent tellement que la synténie disparait, alors que l'homologie des séquences protéiques est toujours identifiable (Drillon et al., 2014). On souhaite donc choisir des paires de génomes G1 G2 qui divergent de moins de 38% et si possible avec un nombre de blocs de synténie minimum pour reconstruire les génomes ancestraux. Selon ce critère, nous avons réussi à sélectionner une combinaison optimale pour chaque génome ancestral à reconstruire (Figure 38, page 119).

A partir du moment où G1 et G2 sont fixés, les génomes G3…Gn appartiennent forcément à la troisième branche issue de l'ancêtre qu'on souhaite reconstruire. A ce stade, nous avons utilisé *SynChro* avec toutes les valeurs de Δ (1 à 6) pour identifier les blocs de synténie entre toutes les paires informatives de génomes (G1/G2, G1/G3…Gn et G2/G3…Gn) définies par ces combinaisons. Cela représente au total 783 paires de génomes, multipliées par 6 valeurs de Δ soit 4698 comparaisons, ce qui réduit un peu le nombre de combinaisons par rapport aux $\left[\binom{38}{2}+\binom{22}{2}+\binom{6}{2}\right].6 = 5694$ paires de génomes au total dans les trois sous-arbres. Nous nous sommes alors demandé si, pour reconstruire les génomes ancestraux, toutes les espèces G3…Gn étaient utiles. On peut inclure tous ces génomes ou bien n'en prendre qu'un sous ensemble, ou même qu'un seul. *AnChro* est conçu de telle manière que le fait de rajouter des génomes au jeu de données ajoute en principe de l'information sans pertes (Guénola Drillon, thèse). Cependant, l'ajout de tous les génomes G3…Gn, en particulier lorsqu'ils sont très divergés, rend parfois plus difficile et plus lente la validation des adjacences ancestrales au point que la reconstruction de certains génomes ancestraux n'aboutit pas. Nous n'avons donc conservé pour chaque combinaison G1/G2/G3…Gn les paires de génomes G1/G3…Gn et G2/G3…Gn dont la divergence moyenne est inférieure à 38% (en bleu dans la Figure 38). Quand aucun génome de G3…Gn ne permettait de satisfaire cette condition, nous avons conservé tous les génomes G3…Gn disponibles.

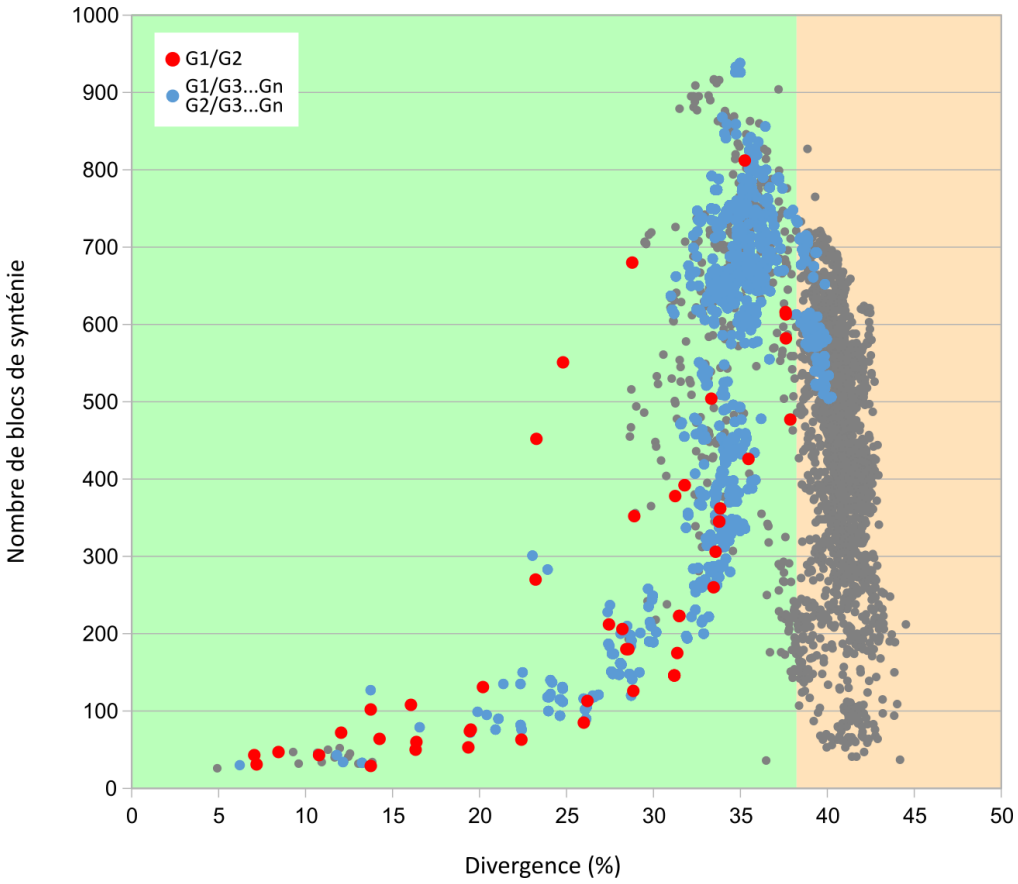


Figure 38 Nombre de blocs de synténie en fonction de la divergence protéique moyenne pour chaque paire de génomes des *Saccharomycotina*. Chaque point représente une paire de génomes. La zone verte indique les paires d'espèces divergeant de moins de 38%. La zone orange indique les paires d'espèces qui divergent de 38% ou plus. Les points rouges indiquent les paires de génomes G1/G2 comparés pour reconstruire les ancêtres des *Saccharomycotina*. Les points bleus indiquent les comparaisons de G1 et G2 avec les génomes référents G3…Gn. Adapté d'après (Drillon et al., 2014).

4.2.3.2. Choix des meilleures reconstructions

Après la reconstruction des génomes ancestraux, nous disposons de 36 versions de chaque génome ancestral. On compare alors l'ordre des gènes retracés dans toutes ces versions, deux à deux. Ces comparaisons permettent d'identifier différents types de contradictions : intra-chromosomiques, inter-chromosomique, fragmentation. On choisit alors la reconstruction qui minimise le nombre de contradictions par rapport à toutes les autres. Un exemple est présenté Figure 39. Nous avons appliqué au préalable cette approche aux génomes des *Lachancea* et avons remarqué que les reconstructions les moins contradictoires sont également celles qui sont reconstruites en respectant le mieux les critères biologiques définis précédemment (nombre de *scaffolds*, nombre de gènes retracés, nombre de centromères par chromosome), ce qui est cohérent.

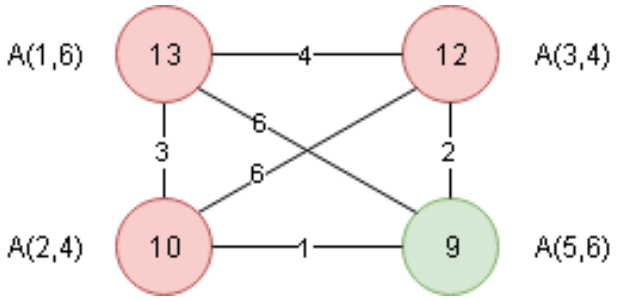


Figure 39 Choix de la reconstruction ancestrale optimale pour un ancêtre (A) parmi les versions (Δ, Δ') possibles. Chaque nœud représente une version de l'ancêtre A obtenue avec les valeurs de (Δ, Δ') entre parenthèses. Les arrêtes représentent le nombre de contradictions entre les différentes versions. Le chiffre dans chaque nœud représente le nombre total de contradictions de la version de l'ancêtre avec toutes les autres. La version en vert est la moins contradictoire, c'est celle qu'on conserve.

Nous avons alors tracé le nombre de *scaffolds* et le nombre de gènes retracés dans les ancêtres reconstruits en fonction des caractéristiques (pourcentage de divergence, nombre de blocs de synténie, nombres de gènes conservés en synténie) des génomes G1 et G2 utilisés pour calculer les reconstructions (Figure 40). On remarque que le niveau de fragmentation des génomes ancestraux augmente de manière exponentielle avec la divergence protéique et le nombre de blocs de synténie et diminue avec un nombre croissant de gènes conservés en synténie entre G1 et G2. De manière assez symétrique, le nombre de gènes retracés diminue avec l'augmentation du pourcentage de divergence et l'augmentation du nombre de blocs de synténie et augmente avec le nombre de gènes conservés en synténie. En résumé, ces résultats valident *a posteriori* la pertinence de l'approche employée pour choisir les espèces G1/G2 (Figure 38) et démontrent la « valeur prédictive » du pourcentage de divergence et du nombre de blocs de synténie sur la qualité du génome ancestral obtenu. Notons que sept génomes ancestraux se distinguent clairement des autres génomes reconstruits (en rouge dans la Figure 40). Il s'agit des génomes 2.16, 2.17, 2.18, 2.19, 2.20, 3.01 et 3.02.

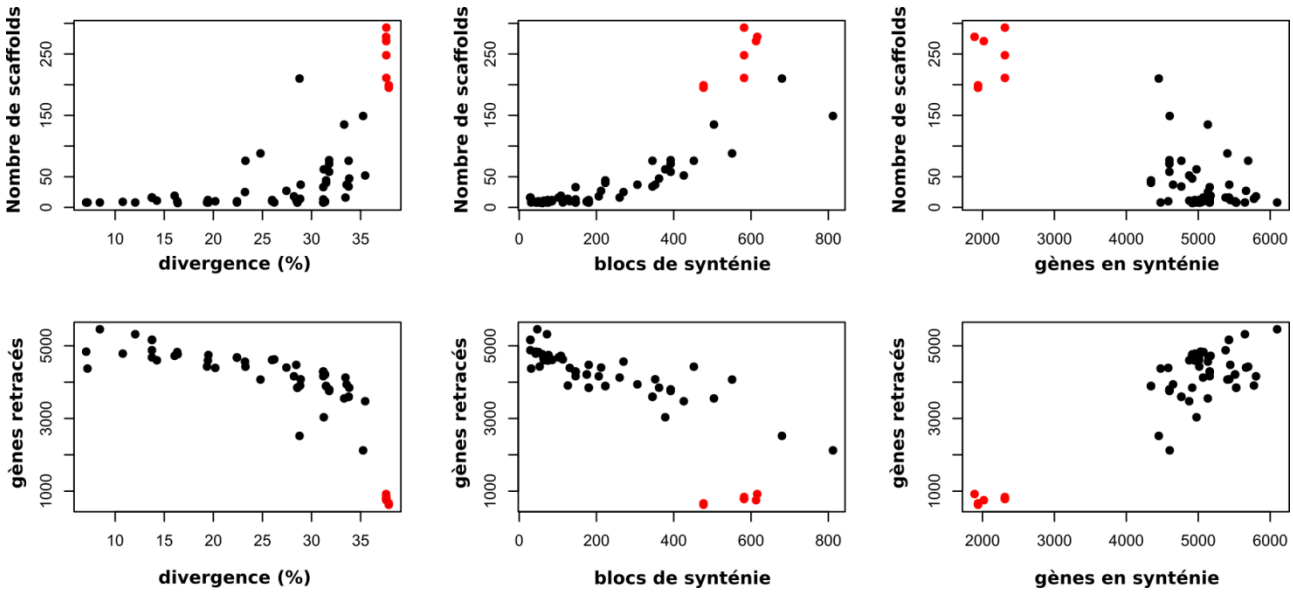


Figure 40 Nombre de *scaffolds* et nombre de gènes retracés dans les génomes ancestraux reconstruits en fonction du pourcentage de divergence, du nombre de blocs de synténie et du nombre de gènes conservés en synténie entre les génomes G1 et G2 utilisés pour les reconstruire. Les sept reconstructions exclues sont représentés en rouge.

Les résultats des reconstructions des génomes ancestraux ainsi que la paire d'espèces actuelles G1/G2 utilisées pour chaque reconstruction sont représentés sur la Figure 41. On observe que les 60 génomes ancestraux ont pu être reconstruits. On voit également les sept génomes ancestraux qui se distinguent du reste des reconstructions en termes de nombre de gènes et de niveau de fragmentation (2.16, 2.17, 2.18, 2.19, 2.20, 3.02 et 3.01) ont été reconstruits en utilisant comme espèces G1 et G2, les génomes *Babjeviella inositovora* avec *Lodderomyces elongisporus* ainsi que *Babjeviella inositovora* avec *Ogataea parapolymorpha* et *Geotricum candidum* avec *Nadsonia fulvescens*, trois couples d'espèces définissant des chemins évolutifs très longs (Figure 41). Dans le cas de 3.01, les deux espèces G1 et G2 comparées, *Geotricum candidum* et *Nadsonia fulvescens* sont trop divergées, le signal synténique est saturé mais on ne peut pas choisir d'espèces plus proches pour obtenir de meilleure reconstruction. Pour les autres génomes (2.16, 2.17, 2.18, 2.19, 2.20, 3.02), les paires de génomes G1/G2 utilisées se trouvent à la limite de la région verte dans la Figure 38 (page 119), indiquant qu'un seuil de 37% voire 36% au lieu de 38% de divergence serait plus adapté pour choisir les paires de génomes G1/G2. Par conséquent, ces génomes ancestraux ont été exclus des analyses ultérieures.

Nous nous sommes alors intéressés aux nombres de gènes retracés ainsi qu'au nombre de *scaffolds* dans les 53 autres reconstructions (Figure 42, page 123). Comme on peut le voir, le nombre de gènes retracés est centré autour de 4000 ce qui représente environ 70% de la taille des génomes actuels (Figure 42C). De plus, les ancêtres reconstruits sont très contigus comme on peut le voir Figure 42B : 34/53 (64%) des génomes sont reconstruits en moins de 25 *scaffolds*. On remarque avec intérêt qu'*AnChro* est capable de reconstruire des génomes ancestraux moins fragmentés que les génomes G1/G2 utilisés (Figure 42A, flèches noires). C'est particulièrement visible dans le cas de 1.04 reconstruit à partir du génome de *K. dobzhanskii*, 1.08 reconstruit à partir de *L. quebecensis* et 1.10 reconstruit à partir de *L. lanzarotensis*. Cette aptitude d'*AnChro* provient du fait que ce dernier peut inférer des adjacences ancestrales qui ont disparu. Dans la plupart des cas, on observe qu'un des deux génomes G1/G2 est peu fragmenté. Cela peut laisser penser que l'ancêtre reconstruit est de bonne qualité grâce à ce génome, mais ce n'est pas toujours le cas. Par exemple les ancêtres 1 .11, 1.24 et 1.25 sont moins fragmentés que les deux génomes actuels qui ont servi à leur reconstruction.

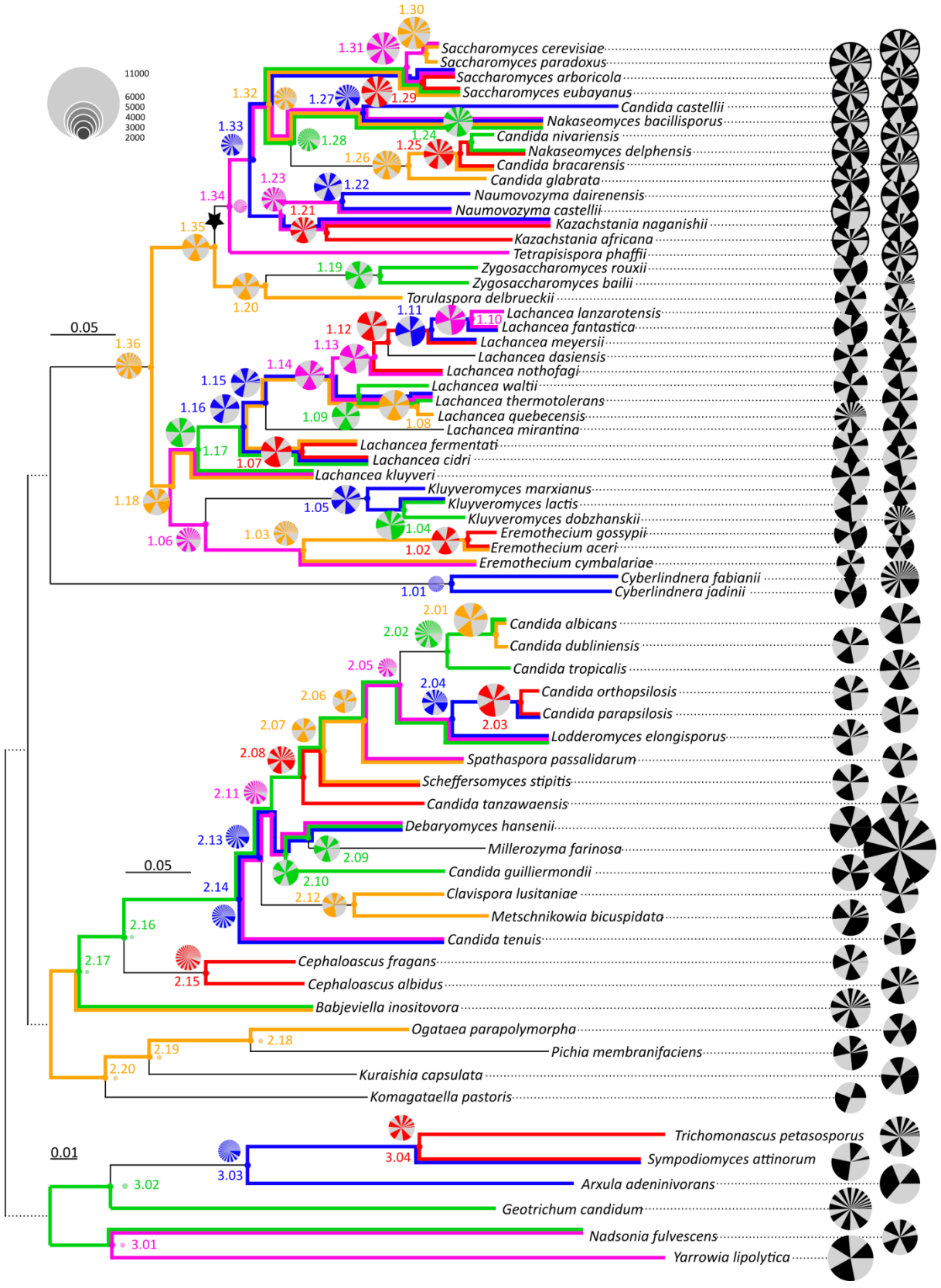


Figure 41 Reconstructions des génomes ancestraux des *Saccharomycotina*. Les génomes sont représentés sous la forme de diagramme circulaires dont la taille est proportionnelle au nombre de gènes retracés. La taille des secteurs est proportionnelle au nombre de gènes dans les *scaffolds* reconstruits. Les génomes actuels sont représentés en niveaux de gris. La duplication totale du génome est indiquée par une étoile noire et les génomes actuels issus de cet événement sont entourés en noir. Les génomes ancestraux sont représentés en couleur. En outre, la paire d'espèces G1/G2 utilisée pour reconstruire chaque ancêtre est représentée par un chemin coloré dans l'arbre. La couleur du chemin est la même que celle de l'ancêtre. Les branches fines, en noir ne font partie d'aucun chemin évolutif G1/G2.

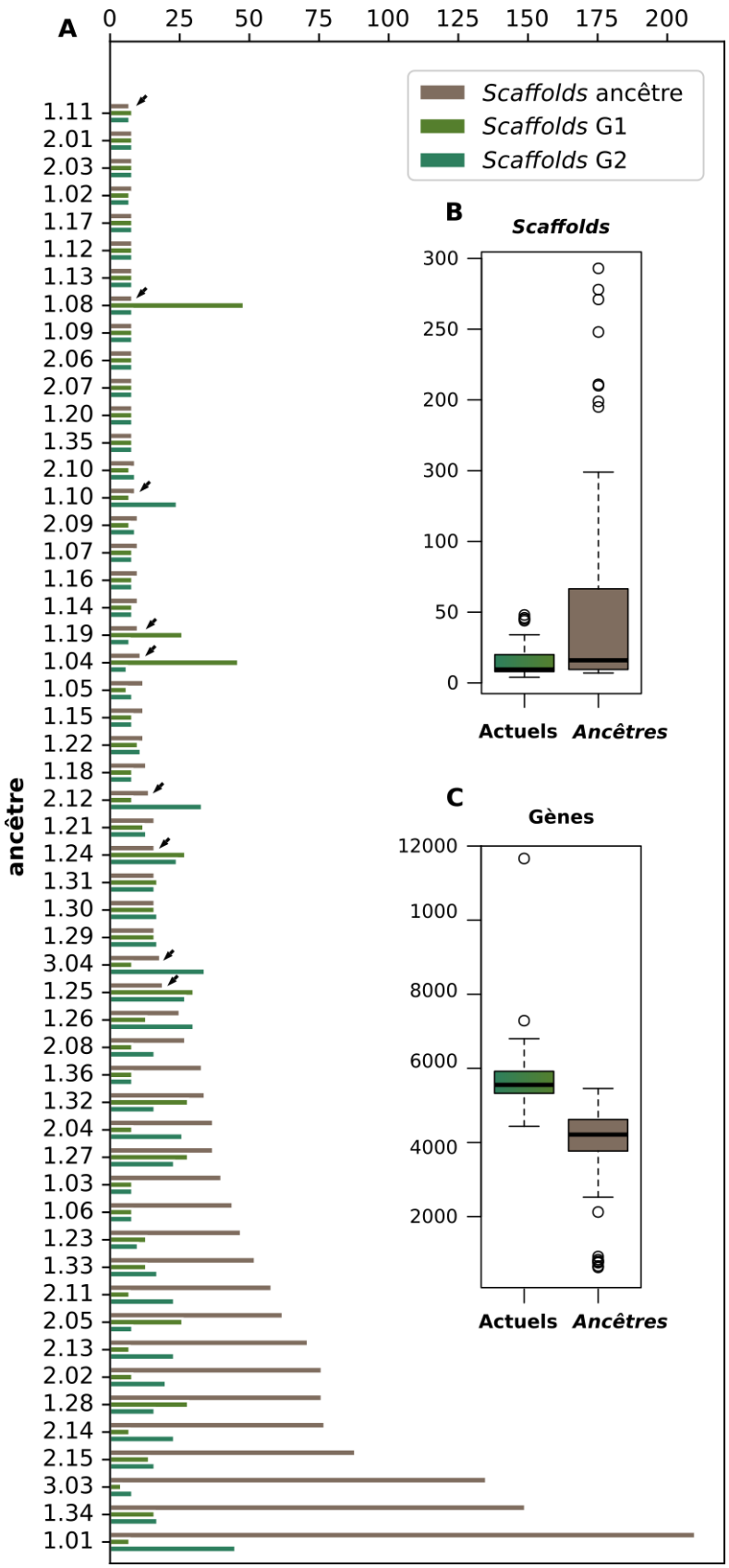


Figure 42 Nombre de *scaffolds* et contenu en gènes des génomes ancestraux reconstruits (en gris) et des génomes actuels (en vert). (**A**) Comparaison du nombre de chromosomes dans les reconstructions comparativement aux génomes G1 et G2 qui ont servi à les reconstruire. Les flèches indiquent des génomes ancestraux moins fragmentés que les espèces référentes actuelles. (**B**) Distribution du nombre de *scaffolds* dans les génomes reconstruits et actuels. (**C**) Distribution du nombre de gènes dans les génomes reconstruits et actuels.

4.2.4. Inférence des réarrangements chromosomiques passés

Pour inférer les réarrangements chromosomiques passés, on compare la structure des génomes situés aux nœuds consécutifs de l'arbre. On compare donc soit deux nœuds internes pour inférer les réarrangements sur une branche interne soit une espèce actuelle avec son ancêtre commun le plus récent pour inférer les réarrangements sur les branches terminales de l'arbre. A cet effet, nous avons calculé les blocs de synténie entre les génomes des nœuds consécutifs de l'arbre avec *SynChro* en utilisant le paramètre Δ=0. Nous avons ensuite utilisé *ReChro* pour détecter à partir des blocs de synténie, les réarrangements qui ont eu lieu sur chaque branche de l'arbre. Notons que le paramètre Δ=0 de *SynChro* permet à *ReChro* d'identifier les inversions d'un seul gène en plus des réarrangements décrits à partir de blocs de synténie plus grands (Vakirlis et al., 2016).

Remarquons qu'*AnChro* exprime chaque génome ancestral de deux manières : en fonction des gènes du génome G1 et en fonction des gènes du génome G2. Par conséquent, il existe au plus quatre manières différentes de comparer la structure de deux génomes consécutifs de l'arbre. Cela pose problème, car en exprimant un génome ancestral en fonction des gènes d'espèces actuelles différentes, on peut en principe observer des différences dans les blocs de synténie inférés à l'étape *SynChro* (Δ=0). Ceci peut engendrer des différences dans les réarrangements inférés par *ReChro*. Nous avons donc inféré les réarrangements balancés en comparant toutes les « versions » disponibles d'une même branche, c'est-à-dire en effectuant toutes les comparaisons des nœuds consécutifs exprimés de différentes manières.

Nous avons ensuite calculé pour chaque branche de l'arbre, la différence absolue entre le nombre de réarrangements balancés inférés pour une « version de branche » par rapport au nombre moyen de réarrangements sur toutes les versions de la branche. La distribution de ces écarts est représentée sur la Figure 43.

On observe que la distribution des écarts est très fortement centrée sur 0, ce qui signifie que la détection des réarrangements est robuste quelles que soient les versions des génomes ancestraux qu'on compare. On observe une valeur extrême de 250 qui correspond à la branche de l'arbre située entre *Millerozyma farinosa* et son ancêtre le plus récent. Cet écart est dû au fait que *Millerozyma farinosa* est un hybride interspécifique dont la structure du génome perturbe fortement la détection des réarrangements selon la version de l'ancêtre qu'on considère : exprimée en fonction des gènes de *Candida guilliermondii* ou des gènes de *Debaryomyces hansenii*. Cette observation était donc prévisible. Nous avons exclu les résultats de cette branche conformément à notre décision préalable de ne pas utiliser le génome de *Millerozyma farinosa* comme source d'information. A l'échelle de l'ensemble de l'arbre, l'écart au nombre de réarrangements moyen entre les différentes versions d'une même branche est de 3,2%. Dans la mesure du possible nous avons tout de même choisi de comparer pour chaque branche, des génomes consécutifs exprimés en fonction des gènes de la même espèce actuelle. Par exemple, les ancêtres 1.10 et 1.11 peuvent tous deux êtres exprimés en fonction des gènes de *L. fantastica*. Pour les autres branches, nous avons comparé les versions des génomes consécutifs dont la divergence protéique est la plus faible. Ces deux mesures visent à minimiser le bruit lié à la détection de la synténie à l'étape *SynChro* (Δ=0).

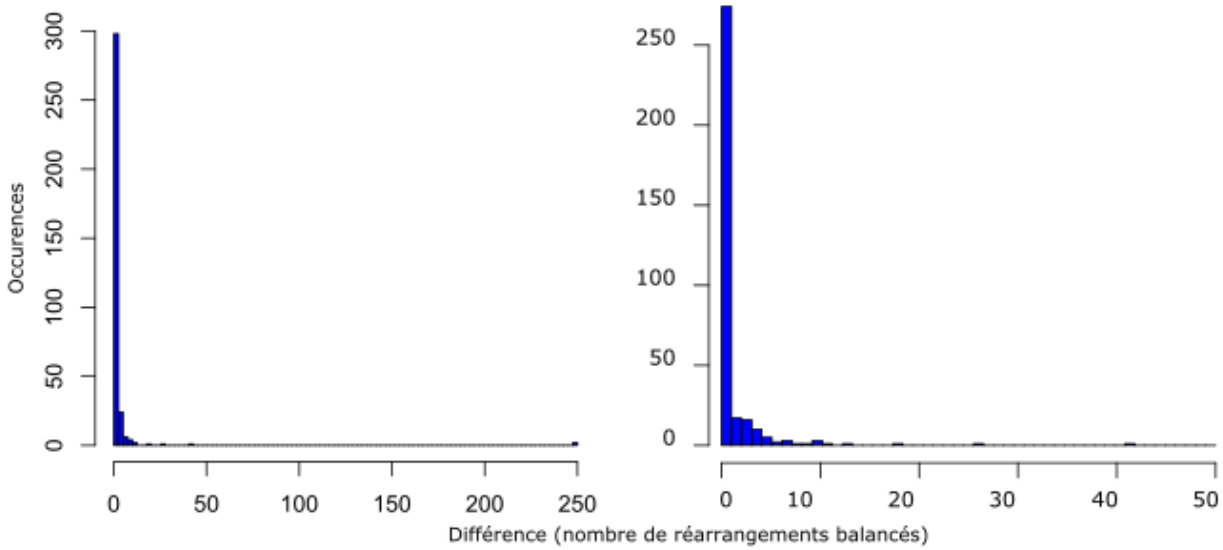


Figure 43 Distribution des différences entre le nombre de réarrangements de chaque version de chaque branche de l'arbre des *Saccharomycotina* et la moyenne de la branche calculée à partir des différentes versions des génomes ancestraux. **A gauche** : distribution pour toutes les versions de toutes les branches de l'arbre**. A droite**, la valeur extrême à 250 a été exclue car elle correspond à la branche menant à l'espèce actuelle *Millerozyma farinosa*, un hybride dont la structure du génome perturbe fortement la détection des réarrangements.

Le résultat de l'inférence des réarrangements chromosomiques balancés est représenté sur la Figure 44.

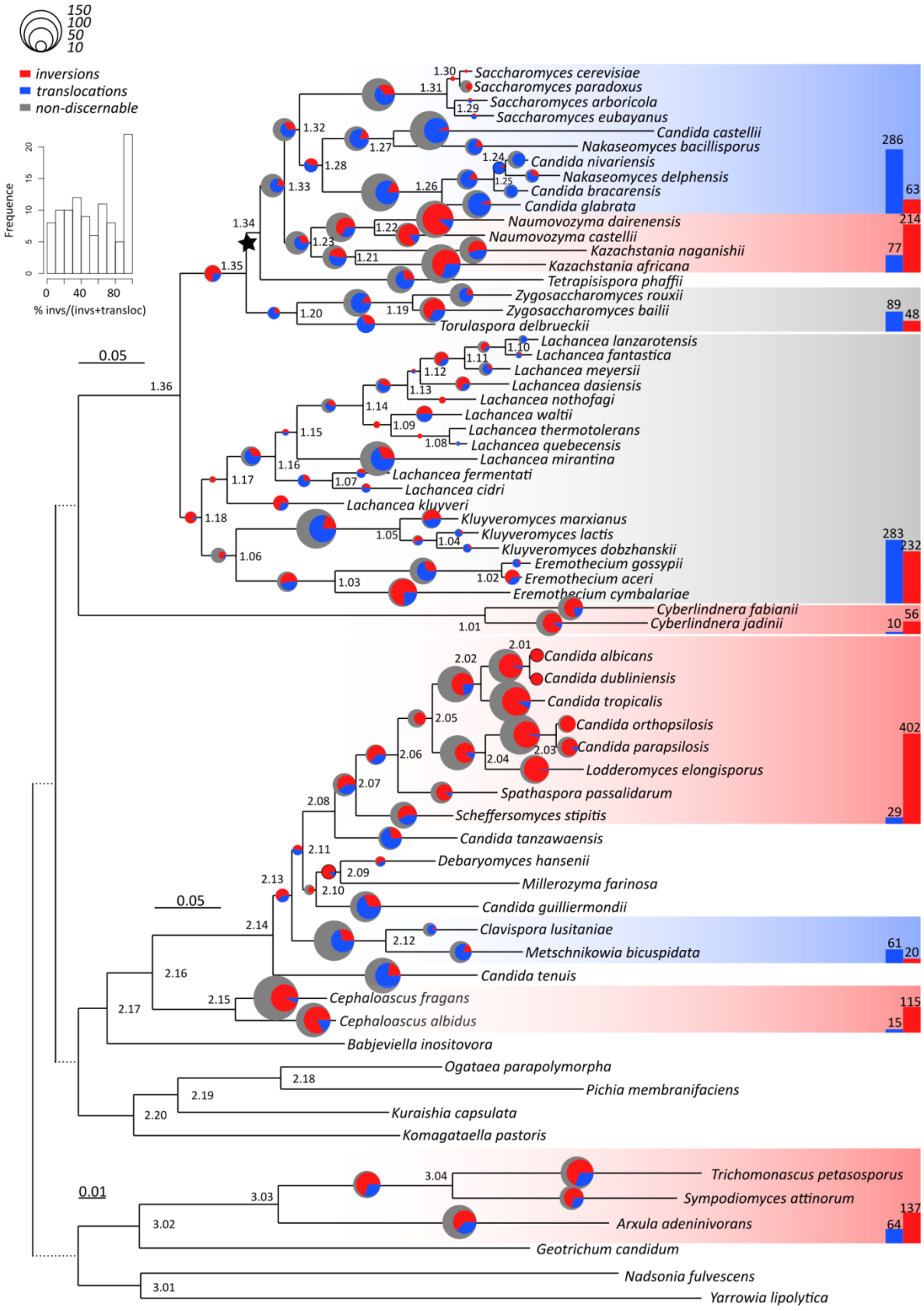


Figure 44 Réarrangements chromosomiques balancés dans l'arbre des *Saccharomycotina*. Pour chaque branche, le nombre d'inversions et de translocations est figuré par un diagramme à secteurs avec les inversions en rouge et les translocations en bleu. Les événements dont on ne peut pas distinguer s'ils sont une inversion ou une translocation sont représentés en gris. Les clades encadrés en rouge contiennent au moins trois fois plus d'inversions que de translocation. Les clades dans lesquels les translocations sont majoritaires sont en bleu. Les clades dans lesquels les inversions et translocations sont en proportions équivalentes sont en gris. Pour chaque clade encadré, les barres rouges et bleues indiquent le nombre global de translocations et d'inversions.

Au total, 5150 événements ont pu être inférés dans l'ensemble de l'arbre des *Saccharomycotina* (Figure 44). On remarque que selon les groupes d'espèces considérés, les réarrangements ne semblent pas s'accumuler au même rythme : pour des branches de longueur comparable, les génomes protoploïdes semblent avoir accumulé moins de réarrangements que les génomes issus de la duplication totale du génome. On pourrait penser que le nombre supérieur de réarrangements observé provient du fait que les génomes issus de la duplication ont massivement perdu les ohnologues, fragmentant la synténie. Or, comme nous l'avons vu dans l'introduction, il a été démontré que cette perte a été très rapide et s'est probablement déroulée entre l'événement de duplication totale et la divergence de *Tetrapisispora phaffii.* Ces génomes sont donc probablement devenus plus instables à la suite de l'événement de duplication. Les génomes du clade CTG sont plus réarrangés que les génomes issus de la duplication totale du génome (voire plus bas et Figure 45). Ces observations selon lesquelles les réarrangements n'ont pas le même taux d'apparition selon les clades considérés sont cohérentes avec d'autres études rapportant des nombres de blocs de synténie (comme estimation du nombre de réarrangements) différents dans différents clades de levures (Fischer et al., 2006).

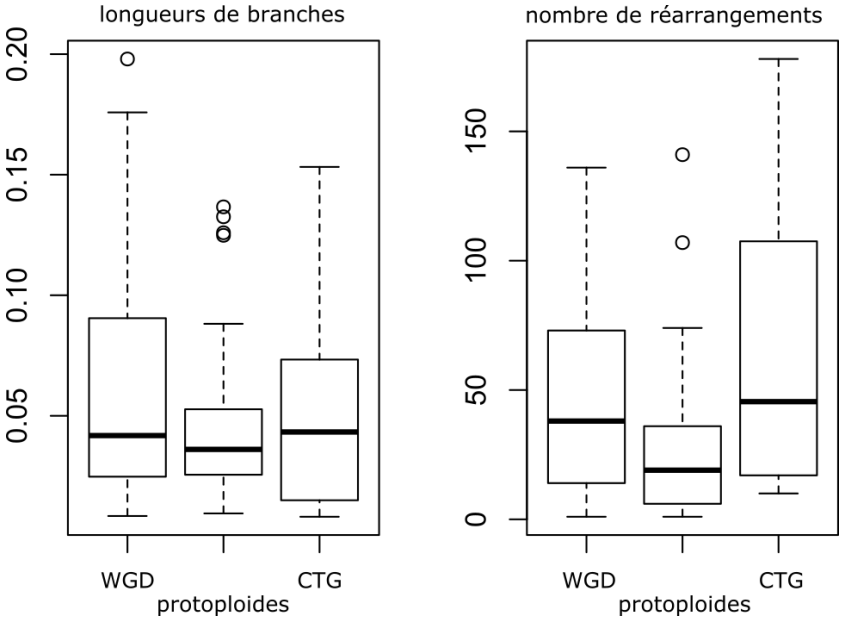


Figure 45 Distribution de la longueur des branches et nombre total de réarrangements (inversions, translocations et non-discernables) dans le clade des espèces issues de la duplication totale du génome (WGD), des *Saccharomycetaceae* qui n'ont pas la duplication totale du génome (protoploïdes) et des espèces du clade CTG (CTG).

Intéressons nous à présent aux types de réarrangements inférés. Parmi les 5150 réarrangements balancés identifiés, 1513 (29,4%) sont des inversions, 1174 (22,8%) sont des translocations et 2463 (47,8%) sont des événements « non-discernables » dont on ne peut pas déterminer s'il s'agit d'inversions ou de translocations. Le pourcentage global inversions/inversions+translocations dans l'ensemble de l'arbre des *Saccharomycotina* est de 56,3%, mais les pourcentages locaux varient beaucoup (histogramme dans la Figure 44). De plus on observe dans l'arbre que de manière assez cohérente, certaines lignées évoluent majoritairement par translocations et d'autres par inversion. Par exemple, chez les espèces issues de la duplication totale du génome, les génomes descendants de l'ancêtre 1.32 (dont les espèces actuelles sont *S. cerevisiae* à *C. glabrata*) ont évolué essentiellement par translocation. Le clade frère des espèces descendant de 1.23 (*N. dairenensis* à *K. africana*) évolue principalement par inversions. On peut imaginer que la présence de séquences dupliquées favorise les recombinaisons ectopiques et explique l'instabilité

des génomes issus de la duplication totale du génome. Les génomes des espèces du clade CTG évoluent presque exclusivement par inversions. La prévalence des inversions de petite taille chez ces espèces a déjà été rapporté chez les *Saccharomycotina* (Butler et al., 2009; Seoighe et al., 2000). Un mode d'évolution comparable a été rapportées chez les champignons filamenteux Ascomycètes (Hane et al., 2011). Dans cette étude, les auteurs identifient une nouvelle forme d'évolution chromosomique selon laquelle les gènes homologues sont maintenus sur un même chromosome, même si l'ordre et l'orientation des gènes est considérablement remanié. Les auteurs appellent ce mode d'évolution « méso-synténie » par analogie avec la micro-synténie désignant la conservation de la synténie sous la forme de petits segments de quelques gènes entre deux espèces et avec la macro-synténie désignant la conservation de grands segments conservés en synténie entre différents chromosomes. Le fait que les inversions soient prévalentes dans les clades issus des embranchements supérieurs de l'arbre (en particulier dans les lignées basales et le clade CTG) et que l'ensemble de l'arbre soit globalement enrichi en inversions (histogramme de la Figure 44) laisse penser que le mode de réarrangement par inversion est ancestral. Il existe d'autres exemples de génomes eucaryotes qui évoluent majoritairement par inversion comme par exemple celui de la drosophile (Bhutkar et al., 2008; Ranz et al., 2007), et du poulet (Burt et al., 1999). Les mécanismes expliquant le type d'évolution chromosomique, par inversions (conduisant à l'observation de méso-synténie) ou par translocation (conduisant à l'observation de macro-synténie) selon les lignées n'est pas claire. Ces différences proviennent peut être de mécanismes cellulaires différents, ou de mécanismes dont l'efficacité relative est différente. L'organisation spatiale du génome dans le noyau ainsi que la proportion et les types d'éléments répétés dans le génome de ces espèces pourrait surement apporter d'intéressants éléments de compréhension.

On observe que le nombre de réarrangements non-discernables est plus important dans les clades dont le taux d'apparition des réarrangements est grand. Cela est explicable par le fait que l'accumulation de points de cassure entre deux génome, ainsi que la réutilisation des points de cassure au cours de l'évolution des génomes crée des cycles dans le graphe d'adjacences dont la résolution ne permet pas de dire quel types d'événements ont eu lieu.

Une autre observation frappante est que le nombre de réarrangements balancés inférés dans l'arbre est corrélé à la longueur des branches de l'arbre (Figure 46). Notamment, la somme des inversions et des translocations est significativement corrélée aux longueurs de branches chez les *Saccharomycotina* (R^2=0,40, P-valeur<10^{-10}). Cette corrélation reste observable quand on décompose les données en *Saccharomycetaceae* d'une part (R^2=0,55, P-valeur<10^{-10}) et clade CTG/méthylotrophes d'autre part (R^2=0,26, P-valeur<10^{-2}). Les translocations seules restent bien corrélées dans l'arbre complet mais également quand on distingue Saccharomycetaceae et CTG/méthylotrophes contrairement aux inversions qui ne semblent corréler aux longueurs de branches que chez les *Saccharomycetaceae*. Ces résultats montrent que la fixation des mutations ponctuelles non-synonymes et des réarrangements balancés se fait de manière coordonnée. Cette observation est en contraste avec les résultats présentés dans la Figure 36 dans laquelle la longueur des branches de l'arbre généré par *PhyChro* ne semblait pas corrélée à la longueur des branches de l'arbre obtenu avec *PhyML* à partir des homologues synténiques. Cela illustre le fait que la détection des réarrangements passés directement à partir des génomes actuels est fortement bruitée. Ce même constat a également été rapportée dans l'étude (Sacerdot et al., 2018).

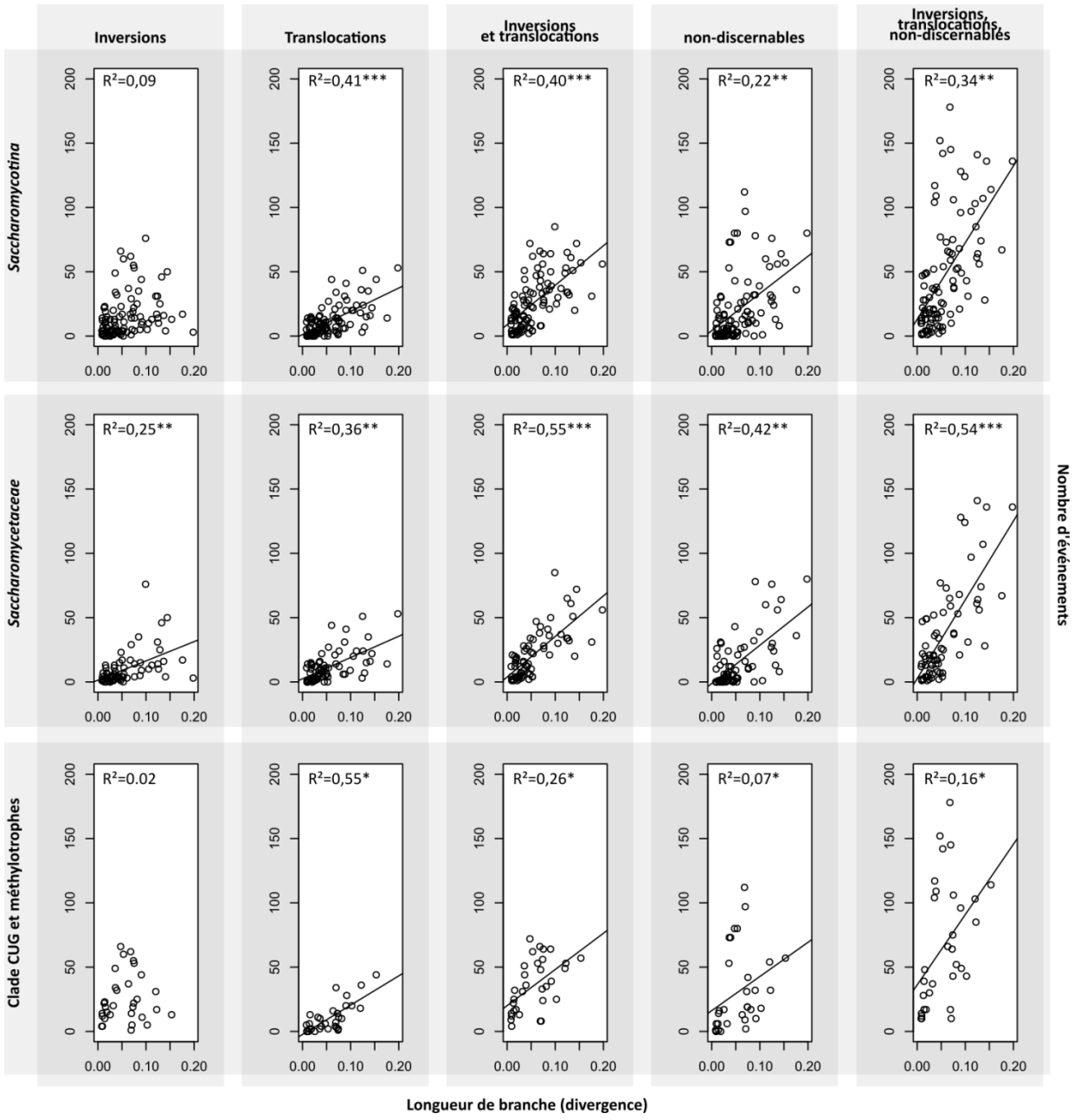


Figure 46 Corrélations entre le nombre de réarrangements chromosomiques balancés et la longueur des branches de l'arbre (divergence protéique). La première ligne représente l'ensemble de l'arbre des *Saccharomycotina*. La deuxième et troisième ligne représentent respectivement **l**es *Saccharomycetaceae* et les espèces du clade CTG et des méthylotrophes. (*, ** et *** indiquent respectivement des p-valeurs inférieures à 10^{-2}, 10^{-5} et 10^{-10}).

Cette « horloge génomique » avait été rapportée chez les levures *Lachancea* (Vakirlis et al., 2016) mais il n'était pas évident que cette observation soit valide dans un groupe d'espèces plus divergées. La généralisation de cette horloge moléculaire des *Lachancea* (Vakirlis et al., 2016) à l'ensemble du subphylum des *Saccharomycotina*, du moins aux 66 espèces que nous avons pu analyser, laisse peut être entrevoir une règle générale de l'évolution des génomes eucaryotes.

5. Reshuffling yeast chromosomes with CRISPR/Cas9

Aubin Fleiss[1], Samuel O'Donnell[1], Nicolas Agier[1], Stéphane Delmas[1] and Gilles Fischer[1,#]

[1]: Sorbonne Université, CNRS, Institut de Biologie Paris-Seine, Laboratory of Computational and Quantitative Biology, F-75005, Paris, France.

[#] Corresponding author: Gilles Fischer gilles.fischer@sorbonne-universite.fr

5.1. Résumé en français

Dissocier l'impact phénotypique des réarrangements de la contribution phénotypique du fond génétique requiert une procédure versatile pour générer de manière efficace et propre des variations structurelles. Nous avons développé une méthode reposant sur le système CRISPR/Cas9 pour réarranger le génome de la levure *S. cerevisiae* en induisant soit des translocations ciblées, soit des réarrangements multiples du génome. En induisant des CDB sur des chromosomes différents et en forçant la réparation des chromosomes en trans par recombinaison homologue avec des oligonucléotides chimériques, nous avons pu induire des translocations réciproques avec une précision de l'ordre de la paire de bases. Ces translocations peuvent être introduites et défaites de manière réversible selon le même principe. Afin de générer des réarrangements multiples, nous avons dirigé des coupures dans des séquences présentes en plusieurs copies dans le génome et laissé le génome réparer ces lésions en utilisant les copies intactes comme modèle. Comme nous le verrons, cette stratégie génère un grand nombre de caryotypes différents avec à la fois des réarrangements chromosomiques balancés mais également des réarrangements non-balancés. Les translocations réciproques ainsi que les réarrangements multiples ont été validés avec différentes techniques : PCR, Southern blot, séquençage et assemblage de génomes complets à partir de *long-read.* Nous avons dans un premier temps récapitulé dans le génome de la souche de référence, la translocation réciproque *SSU1/ECM34*, connue pour conférer une résistance accrue aux sulfites aux souches de vin. Etonnamment, la souche de laboratoire ainsi réarrangée ne présente pas le phénotype d'intérêt, suggérant que la translocation seule ne confère pas ce trait. Dans un second temps, nous avons montré que la viabilité des spores issues de la méiose dans des souches présentant des réarrangements multiples à l'état hétérozygote est très fortement impactée négativement. Ces mêmes souches, testées dans différentes conditions de croissance végétative stressantes montrent une grande diversité phénotypique et dans certains cas des avantages significatifs en termes de taux de croissance, alors même qu'aucune phase codante n'a été altérée par des réarrangements.

5.2. Abstract

Untangling the phenotypic impact of chromosomal rearrangements from the contribution of the genetic background requires a versatile procedure to efficiency generate scarless and markerless structural variations. We developed a CRISPR/Cas9-based method to reshuffle the yeast genome by engineering either precisely targeted reciprocal translocations or multiple structural variations. Generating two double-strand breaks on different chromosomes and forcing the trans-chromosomal repair by homologous recombination result in reciprocal translocations at the base-pair resolution. We made these translocations either irreversible by deleting a small sequence at the junction or reversible to the original chromosomal configuration by inducing the backward translocation. Generating multiple DSBs by targeting repeated sequences in the genome and letting the uncut copies of the repeats being used as template for trans-chromosomal repair resulted in multiple rearrangements. This strategy yields a large diversity of karyotypes including both balanced and unbalanced rearrangements. We validated the targeted translocations and characterized the multiple rearrangements by long-read *de novo* genome assemblies. To test the phenotypic impact of rearranged chromosomes we first recapitulated in a lab strain the *SSU1/ECM34* translocation known to provide increased sulphite resistance to wine isolates. Surprisingly, this resulted in decreased sulphite resistance in the reference strain showing that the sole translocation is not the driver of increased resistance. Secondly, we showed that all strains carrying multiple rearrangements had severely impaired spore viability but showed large phenotypic diversity in various stressful conditions leading in some instances to a strong fitness advantage, although no coding region was altered by the rearrangements.

5.3. Introduction

Genetic polymorphisms are not restricted to base substitutions and indels but also include large-scale Structural Variations (SVs) of chromosomes. SVs comprise both unbalanced events, often designated as copy number variations (CNVs) including deletions and duplications, and balanced events that are copy number neutral and include inversions and translocations. Both have a phenotypic impact, however the prevalence and the fitness effect of balanced SVs has been less documented than CNVs, partly because they are much more challenging to map than CNVs and also because quantifying their fitness contribution independently from the confounding effect of base substitutions remains challenging. Natural balanced chromosomal rearrangements result from the exchange of DNA ends during the repair of Double Strand Breaks (DSBs) either through Homologous Directed Repair (HDR) between dispersed repeats or intact chromosomes carrying internal repeat sequences homologous to the DNA ends (Piazza et al., 2017), or through Non-Homologous End Joining (NHEJ) (Branzei and Foiani, 2008). Artificial balanced rearrangements are classically engineered by inducing targeted DSBs and promoting repair through both HDR and NHEJ. However, inducing targeted DSBs and engineering scar-less chromosomal rearrangements has remained challenging. In early studies structural variants were obtained through I-SceI-induced DSB repair between split alleles of a selection marker (Fairhead et al., 1996; Richardson and Jasin, 2000). In later developments, the use of the I-SceI endonuclease was combined to a “COunter-selectable REporter” or CORE cassette in the frame of the delitto-perfetto technique, allowing the generation of a reciprocal translocation in a

scar-less fashion (Storici and Resnick, 2003, 2006). Other techniques based on Cre-Lox recombination were used to make the genomes of *Saccharomyces cerevisiae* and *Saccharomyces mikatae* colinear and generated interspecific hybrids that produced a large proportion of viable but extensively aneuploid spores (Delneri et al., 2003). Cre/Lox recombination was also used to assess the impact of balanced rearrangements in vegetative growth and meiotic viability (Avelar et al., 2013; Naseeb and Delneri, 2012; Naseeb et al., 2016). A novel approach using yeast strains with synthetic chromosomes allowed extensive genome reorganization through CreLox-mediated chromosome scrambling (Annaluru et al., 2014; Hochrein et al., 2018; Jia et al., 2018; Shen et al., 2016b). This approach proved to be efficient to generate strains with a wide variety of improved metabolic capacities (Blount et al., 2018; Jia et al., 2018; Luo et al., 2018b; Shen et al., 2018a). Muramoto and collaborators recently developed a genome restructuring technology relying on a temperature-dependent endonuclease to conditionally introduce multiple rearrangements in the genome of *Arabidopsis thaliana* and *S. cerevisiae,* thus generating strains with marked phenotypes such as increased plant biomass or ethanol production from xylose (Muramoto et al., 2018). Methods using Zinc Finger Nucleases (ZFNs) and Transcription Activator-Like Effector Nucleases (TALENs) were also developed to generate targeted rearrangement in yeast, mammalian and zebrafish cells (Brunet et al., 2009; Piganeau et al., 2013; Richard et al., 2014; Xiao et al., 2013). Although these technologies provide very useful insights, they are often difficult to implement and/or rely on the use of genetic markers. For this reason, the development of the CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated) system has boosted the field of genome engineering (Alexander, 2018; Doudna and Charpentier, 2014; Fraczek et al., 2018). This system, initially derived from immune systems of bacteria, consists of an endonuclease encoded by the Cas9 gene of *Streptococcus pyogenes* and a short RNA that guides the endonuclease at the targeted genomic locus. The gRNA can be easily designed to target any genomic locus proximal to a “NGG” Promoter Adjacent Motif (PAM). This technology is now routinely used to introduce targeted DSBs in genomes from a wide variety of species (Wang and Qi, 2016). In yeast, CRISPR/Cas9 induced DSBs can be repaired with high efficiency by providing homologous repair DNA cassettes, allowing a variety of genome editions. Previous studies achieved the introduction of point mutations, single and multiple gene deletions and multiplexed genome modifications at different loci by transforming cells with plasmids bearing single or multiple gRNAs and linear DNA repair templates (DiCarlo et al., 2013; Jakočiūnas et al., 2015; Mans et al., 2015, 2018). CRISPR-based approaches have also been developed to add centromeres and telomeres to chromosome fragments (Sasano et al., 2016) concatenating chromosomes (Luo et al., 2018a; Shao et al., 2018) and for massively parallel genome editing to generate large libraries of genetic variants (Bao et al., 2015; Roy et al., 2018; Sadhu et al., 2017).

CRISPR/Cas9 also opened new avenues in the study of genome structure especially with the engineering of translocations in mammalian cells with high efficiency. The principle is to introduce two DSB in two distinct chromosome with CRISPR, then repair the DNA ends in trans by HDR with donor DNA carrying a selection marker, lost in a second step by Cre/Lox recombination leaving a single loxP element at the chromosomal junction (Vanoli et al., 2017).

In this study, we developed a CRISPR-Cas9 multiplexed genome editing strategy to generate markerless reversible and non-reversible reciprocal translocations in yeast with base-pair precision and high efficiency. We also induce multiple DSBs by targeting scattered Ty3-LTRs, thereby generating patchwork

chromosomes resulting from multiple translocations and complex rearrangements. Finally, we quantified the phenotypic impacts of both targeted and multiple rearrangements on meiotic fertility and mitotic growth in various stress conditions.

5.4. Material and Methods

5.4.1. Strains and media

The strains of *Saccharomyces cerevisiae* BY4741, (*MATa*, *his3Δ1*, *leu2Δ0*, *ura3Δ0*, *met15Δ0*) and BY4742 (*MATα*, *his3Δ1*, *leu2Δ0*, *ura3Δ0*, *lys2Δ0*) were used for generating the translocations between the *ADE2* and *CAN1*, *ECM34* and *SSU1* genes, and for the multiple rearrangements. All pre-cultures were performed in YPD. After transformation, cells were selected on complete synthetic medium depleted in leucine (CSM-Leu). Plasmid cloning steps were performed in chemically competent *Escherichia coli* DH5α. Ampicillin resistant bacteria were selected on LB medium supplemented with ampicillin at 100 µg/ml. The SK1 strains 1513 (*MATα, ho::LYS2*, *ura3*, *leu2::HISG*, *lys2*, *arg4*(Nnde1)-Nsp, *thr1-A*, *SPO11*-HA3-His6::*KanMX4*) and 1708 (*MATa*, *ho::LYS*2, *ura3*, *leu2::HISG, lys2*, *arg4*-Bgl::NdeI-site1°, *CEN8::URA3*) provided by Bertrand Llorente (CRCM, Marseille) were used for their high sporulation efficiency compared to BY strains to perform crosses and to quantify spore viability of the rearranged strains. The sulphite resistance phenotype of our engineered strains was compared to that of wine strains Y9J_1b and DBVPG6765 provided by Joseph Schacherer (Université de Strasbourg).

5.4.2. Sulphite resistance spot tests

Sulphite resistance of two wine strains and two engineered lab strains (YAF082, YAF083) was assessed on YPD plates buffered at pH = 3.2 and supplemented with sulphites according to an established procedure (Park et al., 1999) with minor modifications: buffered plates were prepared by adding tartaric acid / potassium, sodium tartrate buffer to a final concentration of 38 mM to freshly autoclaved agar YPD. Sulphites plates were prepared by spreading a 1M stock solution of Na_2SO_3 on 75 ml buffered square plates to a final concentration of 2 mM or 4 mM. Sulphites were let to diffuse overnight at room temperature. The next day, for each experiment, we dispensed in triplicate 10^5, 10^4, 10^3, 10^2 cells of each strain in 50 µL sterile water drops. The position of each strain in the three replicates was different to minimize neighbouring and edge effects. The plates were incubated for 4 days at 30 °C before being scanned. The whole experiment was repeated 3 times.

5.4.3. Growth curves in stress conditions

Generation times of strains were measured under the following conditions: rich medium (YPD), salt stress (YPD + 1.5 M NaCl), inhibition of nucleic acid synthesis (YPD + hydroxyurea 20 mg/ml), presence of caffeine known as a toxic purine analogue (YPD + caffeine 16 mM). These drug concentrations impeding but not stopping cell growth were determined as the concentration required to generate an approximately three-fold increase in the generation time of BY4741 as compared to the YPD growth condition. Compared to the observed generation time of 80 minutes in YPD, there was a fold increase of 3.7, 3.85 and 2.8 in

NaCl, Caffeine and HU conditions respectively. Cells were diluted from a saturated YPD culture to 2.10^6 cells/ml and 10 µl corresponding to 2.10^4 cells were used to inoculate a new plate with 90µl of fresh medium per well. OD_{600} measurement was performed in precision mode every 15 minutes for 3 days using the TECAN Sunrise micro-plate reader and Magellan v7.2 software. Constant high intensity "interior mode" agitation and 30°C temperature were maintained during all experiments. In addition, a few Ø 500 µm glass beads (Sigma G8772) were added to each well to prevent cell aggregation. For each strain and condition, growth curves were carried out in triplicate and the generation time was extracted using the R growthcurver package and averaged on the three measurements.

5.4.4. Spore viability

Colonies of rearranged strains originating from the BY background and SK1 strains of the opposed mating type were mixed and spread on the same YPD plate and left overnight at 30°C. The next day, the cells were re-suspended in distilled water and single cells were picked using the Sanger MSM 400 micro-manipulator and left to grow on YPD for 2 days at 30°C until a colony appeared. Colonies originating from single cells were replicated on sporulation medium and left for 4 to 7 days at 30°C until tetrad appeared. For each cross, 10 tetrads from two diploid strains were dissected on YPD and left to grow for 3 days before counting viable spores.

5.4.5. Identification of CRISPR/Cas9 target sequences

For the translocation between *ADE2* and *CAN1*, the target sequences CAN.Y and ADE2.Y found in the literature (DiCarlo et al., 2013) were re-used. The new targets formed by the first translocation were targeted to "reverse" the translocation and restore the original junctions. For the *ECM34*/*SSU1* translocation, specific CRISPR/Cas9 target sequences with minimal off-targets were chosen as close as possible to the natural recombination site of wine strains with the CRISPOR v4.3 website (http://crispor.tefor.net) using the reference genome of *Saccharomyces cerevisiae* (UCSC Apr. 2011 SacCer_Apr2011/sacCer3) and the NGG protospacer adjacent motif. For multiple rearrangements, we identified 39 occurrences of Ty3-LTRs in the latest version of the genome of *Saccharomyces cerevisiae* S288C (accession number GCF_000146045.2), which we have aligned using MUSCLE. Four sequences were incomplete and excluded from further analysis. We then manually selected a suitable gRNA sequence targeting five Ty3-LTR elements and looked for off-targets to this target sequence with CRISPOR. Predicted off-targets had either mismatches with the chosen guide and/or were devoid of PAM indicating that they would not be recognized by CRISPR/Cas9.

5.4.6. Construction of CRISPR/Cas9 plasmids with one or two guides

All plasmids used in this study were obtained by cloning either 20bp, corresponding to the target sequence of a single gRNA, or a 460 bp synthetic DNA fragment allowing to reconstitute two gRNA expression cassettes in pGZ110 (Figure 47, page 138). All oligonucleotides are listed in Supplementary Table 1, page 150. The plasmid pGZ110 was kindly provided by Gang Zhao and Bruce Futcher (Stony Brook University). We linearised pGZ110 with the enzyme LguI (ThermoFischer FD1934) and gel purified the backbone. In

order to clone one target sequence we first annealed two oligonucleotides of 23 bases with adequate 5' overhangs of 3 bases to obtain the insert. Annealing was performed by mixing equimolar amounts of forward and reverse oligonucleotides at 100pmol/µl with NEBuffer 4 (New England Biolabs), heating 5 minutes at 95°C and allowing the mix to cool down slowly to room temperature. We then mixed 100ng of backbone with 20pmol of double stranded insert and performed the ligation with the Thermo Fischer Rapid DNA ligation Kit (K1422) according to manufacturer instructions. In order to obtain plasmids with two gRNA expression cassettes, we first digested and gel-purified the 460bp synthetic DNA fragments with LguI to obtain inserts with adequate 5' overhangs of 3 bases for ligation in pGZ110. We then mixed 100ng of backbone and 20ng of insert and performed ligation as explained previously. Plasmids were amplified in *E. coli*. Oligonucleotides and synthetic DNA fragments were ordered from Eurofins. Refer to supplementary material for oligonucleotides and plasmid sequences.

5.4.7. Yeast transformation

Yeast cells were transformed using the standard lithium acetate method (Gietz and Woods 2002) with modest modifications : 10^8 cells per transformation were washed twice in 1mL of double distilled water, then washed twice in 1mL of 0.1x LiAc,TE before adding the T-mix and either donor-DNA or the equivalent volume of water and vortexing. Double-stranded donors for repair of CRISPR-induced DSBs were prepared beforehand by mixing equimolar amounts of forward and reverse oligonucleotides at 100pmol/µl with NEBuffer 4 (New England Biolabs), heating 5 minutes at 95°C and allowing the mix to cool down slowly to room temperature. Cells underwent a heat-shock of 25 minutes at 42°C. Centrifugation steps were performed at 9000 rpm with a tabletop centrifuge. After transformation, we checked for each transformation tube that the number of viable cells on YPD was comparable. Cells were re-suspended in distilled water, plated on leucine depleted medium and incubated 4 days at 30°C.

5.4.8. Plasmid Stability

The pGZ110 plasmid is highly unstable when selection is removed. To cure the plasmid cell were grown overnight in YPD at 30°C. Ten individual cells were micromanipulated on YPD plates with the MSM400 micromanipulator (Singer) and grown at 30°C for 2 days. Colonies were then serially replicated on CSM-Leu and YPD. All 10 colonies lost the ability to grow on CSM-Leu.

5.4.9. Estimation of CRISPR/Cas9 cutting and repair efficiencies

The *ADE2*/*CAN1* experiment was used to quantify the translocation yields. We performed 16 independent transformation experiments. The transformations of 10^8 cells with the plasmid bearing no gRNA resulted in a median number of 6055 transformants, indicating high transformation efficiency. Transformation with the plasmid bearing two gRNAs but without providing any donor DNA for repair produced a median number of 12 transformants, indicating high cutting efficiency of 99.8% (Figure 48, page 140). Co-transformation with the same plasmid and either PM or RT-donors provided median numbers of 315 and 358 transformants respectively, indicating high repair efficiencies both in cis and in trans. In total, PM-donor transformations provided 5360 colonies, 91.3% of which were pink [*ade2*]. We replicated 498 pink colonies on canavanine

medium and 97,6% were resistant [*can1*]. In comparison only 53.9% of the 102 white transformants replicated on canavanine were resistant. RT-donor transformations provided in total 6735 colonies, 95,6% of which had acquired the pink phenotype. We then replicated 545 pink strains onto canavanine medium and 542 of them (99.5%) were resistant to canavanine. Also 20.5% of 78 white colonies replicated on canavanine medium were resistant (Figure 48, page 140).

For the backward translocation restoring the ADE2/CAN1 original chromosomal configuration the transformation of two translocated strains with the plasmid bearing no gRNA produced on average 4010 colonies while cells transformed with the gRNAs plasmid but without donor DNAs produced on average 13 colonies, indicating efficient cutting of the chimerical junctions. Co-transformation with the gRNAs plasmid and donor fragments provided on average 207 colonies, 94.2% of which had the restored white phenotype. We replicated 16 white transformants on canavanine-supplemented medium and all strains were sensitive to canavanine.

5.4.10. Pulse Field Gel Electrophoresis and colony PCR for karyotyping rearranged strains

Whole yeast chromosomes agarose plugs were prepared according to a standard method (Török et al., 1993) and sealed in a 1% Seakem GTC agarose and 0.5x TBE gel. Pulse Field Gel Electrophoresis (PFGE) was conducted in with the CHEF-DRII (BioRad) system with the following program: 6V/cm for 10 hours with a switching time of 60 seconds followed by 6V/cm for 17h with switching time of 90 seconds. The included angle was 120° for the whole duration of the run. We compared observed karyotypes with expected chromosome sizes and tested the chromosomal junctions by colony PCR with ThermoFischer DreamTaq DNA polymerase.

5.4.11. Southern blot validation of ECM34/*SSU1* translocation

Southern blot was used to validate the translocation between *ECM34* (ch. VIII) and *SSU1* (ch. XVI). Genomic DNA was extracted from BY4741 and the engineered strain YAF082 using the Qiagen DNA buffer set (19060) and Genomic-tip 100/G (10243) according to manufacturer instructions and further purified and concentrated by isopropanol precipitation. Digestion of 10µg of genomic DNA per strain/probe assay was carried out using FastDigest EcoRI (ThermoFischer FD0274). Electrophoresis, denaturation and neutralisation of the gel were performed according to established procedure (Sambrook et al., 1989). Transfer on nylon membrane (Amersham Hybond XL) was performed using the capillarity setup (Khandjian, 1987; Southern, 1975). The membrane was UV-crosslinked with the Stratalinker 1800 device in automatic mode. Probes targeting the genes YHL044W and *ARN1* located upstream and downstream of the cutting site in the *ECM34* promoter respectively and the genes *NOG1* and *SSU1* located upstream and downstream of the cutting site in the *SSU1* promoter respectively were amplified with ThermoFischer DreamTaq, DIG-11-dUTP deoxyribonucleotides (Roche 11 175 033 910) and gel purified. Blotting and revelation were conducted using the Roche DIG High Prime DNA Labelling and Detection Starter kit II (11 585 614 910) according to manufacturer instructions. Imaging was performed using the G:BOX Chemi XT4 (Syngene) with CSPD chemiluminescence mode. All oligonucleotides are described in Supplementary Table 1, page 150.

5.4.12. Oxford Nanopore *de-novo* genome assembly

DNA from strains YAF019 and YAF064 was extracted using QIAGEN Genomic-tip 20/G columns and sheared using covaris g-TUBEs for average reads lengths of 8kb and 15kb respectively. DNA was repaired and dA-tailed using PreCR and FFPE kits (New England Biolabs) and cleaned with Ampure XP beads (Beckman Coulter). SQK-LSK108 adapters were ligated and libraries run on FLO-MIN107 R9.5 flowcells. Raw signals were basecalled locally using Albacore v2.0.2 with default quality filtering. Flowcell outputs are shown in Supplementary Table 2, page 154.

YAF019 was assembled using the LRSDAY v1 pipeline (Yue and Liti, Nature Protocols, 2018), including nanopolish v0.8.5 correction and excluding pilon polishing due to lack of illumina data. Due to only 19x coverage, the corrected ErrorRate for Canu assembly was increased to 0.16. Assembly data are shown in Supplementary Table 3, page 154.

YAF064 was processed using the LRSDAY v1 pipeline. Linear chromosomes were assembled by SMARTdenovo v1 using 40x coverage of the longest Canu-corrected reads and combined with a Canu assembled mitochondrial genome. For canu correction, due to 200x coverage, correctedErrorRate was set at 0.75. All corrected reads were aligned against a reference-quality S288C genome, assembled with PacBio reads (Yue et al., 2017) using LAST-921. Split reads were used to highlight structural variations not apparent in the SMARTdenovo assembly. Read coverage was used to calculate an increase in the number of copies of particular regions within the rearranged genome. Evidence of rearrangements defined by overlapping reads and changes in copy number were used to manually adjust the assembly prior to nanopolish v0.8.5 and Pilon v1.22 error correction.

5.5. Results

5.5.1. Rationale for chromosome reshuffling

We used a single-vector, pGZ110 (Bruce Futcher, personal communication), which encodes both the Cas9 nuclease gene and a gRNA expression cassette (Figure 47). This cassette allows cloning of either a unique 20 bp fragment corresponding to the target sequence of the gRNA or a synthetic DNA fragment of 460 base-pairs reconstituting two different gRNAs in tandem (Figure 47). This system is highly versatile as it allows cloning of any pair of gRNA in a single ligation step. Such pairs of gRNAs can induce two concomitant DSBs which, upon recombination with chimerical donor DNA, generate reciprocal translocations with a single-nucleotide precision when the DSBs are introduced in different chromosomes. Furthermore, a single gRNA that targets Long Terminal Repeats (LTRs) scattered on different chromosomes can induce a higher number of DSBs, eventually resulting in multiple rearrangements upon repair with the uncut LTR copies. All rearrangements are engineered in a scar-less fashion and without integrating any genetic marker in the genome.

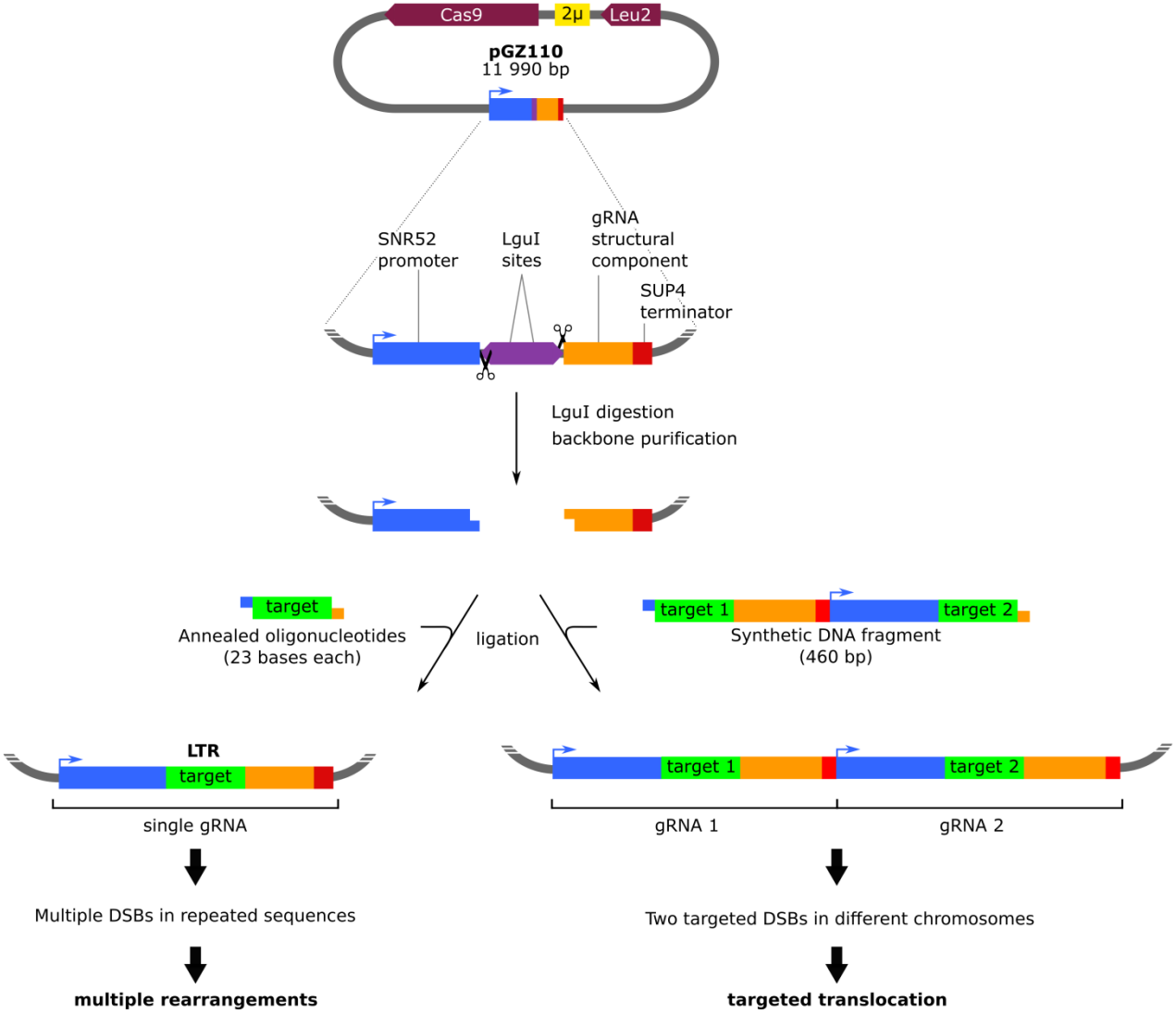


Figure 47 Experimental strategy used to clone in one step a single or a pair of gRNA target sequences. The pGZ110 plasmid encodes the Cas9 nuclease under constitutive promoter TEF1p, the gRNA expression cassette and the LEU2 selection marker. The cassette consists of an SNR52 promoter (blue) separated from the structural component of the gRNA (orange) and a SUP4 terminator (red) by two divergently oriented LguI sites (purple) allowing to clone the gRNA target sequence (green). Upon digestion, the two LguI sites generate non-complementary single strand overhangs of 3 bases. This system is highly versatile as it allows to clone in a single step either a 20 bp oligonucleotide corresponding to the target sequence of a unique gRNA (left) or a synthetic DNA fragment of 460 base-pairs reconstituting two different gRNA expression cassettes in tandem (right).

5.5.2. Engineering markerless, reversible reciprocal translocations at single base-pair resolution

We first engineered a reciprocal translocation between two reporter genes leading to phenotypes easy to observe upon disruption. Mutation in the *ADE2* gene involved in purine nucleotide biosynthesis results in the accumulation of a red pigment while mutating the *CAN1* gene which encodes an arginine permease confers canavanine resistance to the cells. We made this translocation reversible such that we can control the alternation between the two phenotypes [*ade2*, *can1*] and [*ADE2, CAN1*]. To generate two concomitant DSBs on chromosomes V and XV carrying *CAN1* and *ADE2*, respectively, we cloned two previously described gRNA target sequences, namely CAN1.Y and ADE2.Y (DiCarlo et al., 2013). To repair the DSBs, we

co-transformed the cells with the plasmid bearing two gRNAs and “donor” DNA fragments of 90 base-pairs each composed of two homology regions of 45 bp identical to the sequences flanking CRISPR cutting sites (Figure 48A upper). Two combinations of DNA repair donor fragments were used. As a control, we used donors called Point Mutation-donors (PM-donors), promoting the intra-chromosomal repair of DSBs in cis and mutating PAM sequences into stop codons, thus preventing further Cas9 activity. Besides, the donors promoting inter-chromosomal repair in Trans, thus leading to a Reciprocal Translocation were called RT-donors. No point mutation needed to be introduced into the PAM sequence in this case because the translocation generates chimerical target sequences not complementary to the original gRNAs (Figure 48A upper). Transformation with the plasmid bearing two gRNAs and either PM or RT-donors resulted in 89.1% and 95.1% of the colonies showing both the pink and resistance to canavanine phenotypes [*ade2*, *can1*], respectively (methods, Figure 48B). We then tested by PCR the chromosomal junctions of 16 [*ade2*, *can1*] strains recovered from the RT experiment. The expected chimerical junctions were validated in all strains. We further validated the translocation by karyotyping two [*ade2*, *can1*] strains by PFGE. No other visible chromosomal rearrangements could be observed apart from the expected translocated chromosomes VtXV and XVtV (Figure 48C). Sanger sequencing of 250 bp around the chimerical junctions of these two strains confirmed that the translocation occurred right at the position defined by the sequence of the RT-donors with no additional mutation (Supplementary Figure 1, page 155). Finally, we cured the plasmid from the two strains before performing the reverse translocation to restore *ADE2* and *CAN1* (5.4 Material and Methods).

We cloned a pair of gRNAs that target the chimerical junctions formed by the *ADE2-CAN1* translocation in the Cas9 plasmid and designed repair donors to restore *ADE2* and *CAN1* to their original configuration (Figure 48A, lower). Co-transformation with the gRNAs plasmid and donor fragments resulted in 94.2% of colonies which restored the white and canavanie sensitive phenotypes [*ADE2, CAN1*]. We performed pulse field electrophoresis karyotyping of 2 [*ADE2, CAN1*] strains. No difference could be observed between the karyotype of these two strains and that of the original BY4741 strain (Figure 48C). The chromosomal junctions of the 2 de-translocated strains were Sanger-sequenced and were found identical to BY4741 natural junctions (Supplementary Figure 1, page 155). These results demonstrate that reversible chromosomal translocations can be engineered at base-pair resolution with high efficiency.

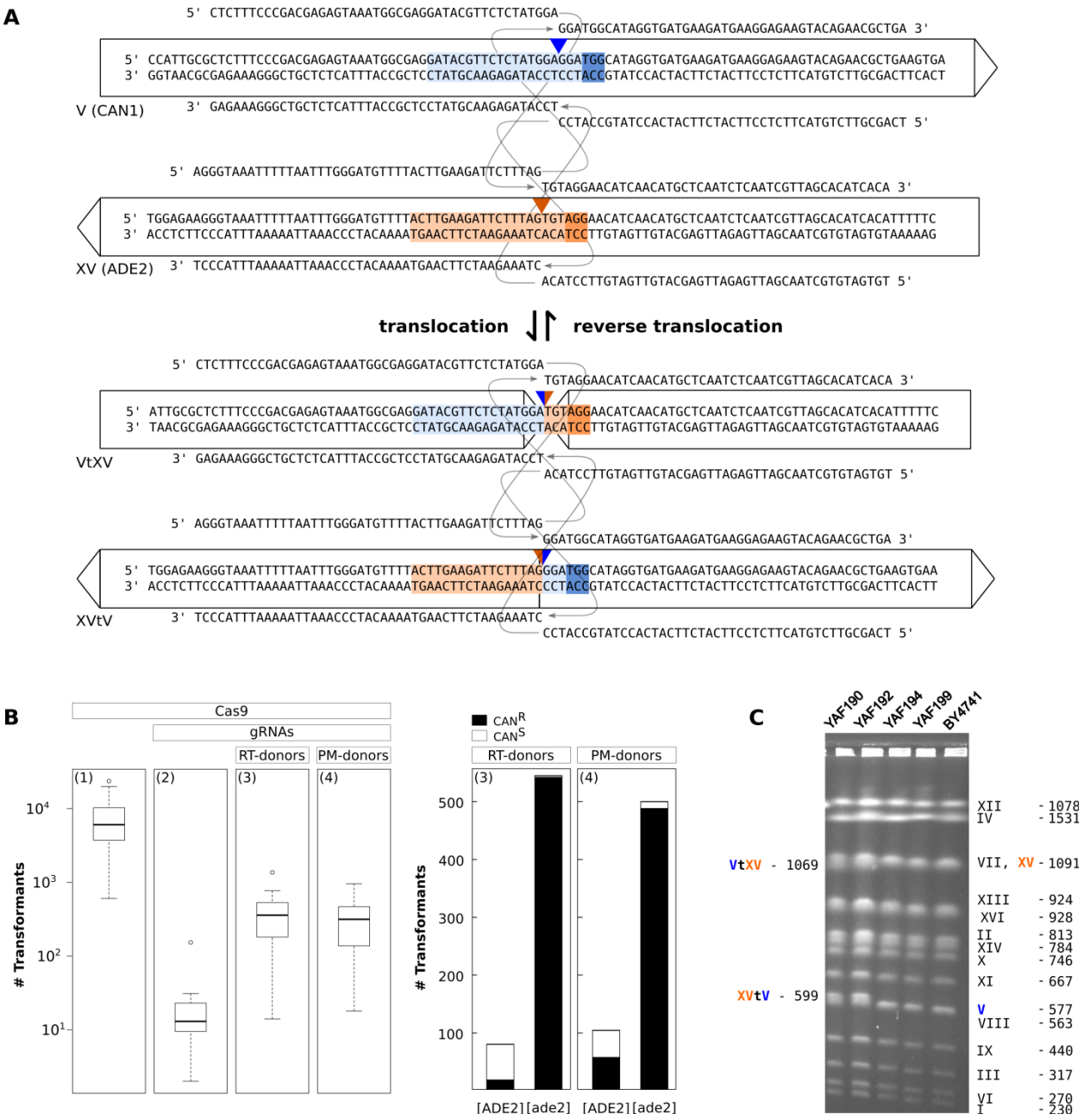


Figure 48 A Reversible markerless translocation. (**A**) The two gRNA target sequences are highlighted in light blue and orange. PAM sequences are highlighted in dark blue and orange. Triangles indicate DSBs sites. Arrows framing the sequences indicate the orientation of coding phases. Donor nucleotides are represented above and below the frames by sequences linked by thin arrows to indicate their homology with the two different chromosomes. Top part: gRNAs and donors used to engineer the translocation between chromosomes V and XV at the ADE2 and CAN1 loci in BY4741, respectively; bottom part: gRNAs and donors used to revert the translocation between chromosomes VtVX and XVtV, thus restoring the natural junctions of BY4741. (**B**) Left: boxplots indicating the total number of transformants obtained in 16 independent transformation experiments. Panels (1) and (2) indicate efficient transformation with the Cas9 plasmid and high cutting efficiency of the gRNAs, respectively. Panels (3) and (4) show high mutation efficiency of the ADE2 locus by both the Reciprocal Translocation (RT) and Point Mutation (PM) donor DNAs used to repair the DSBs. Right: proportion of canavanine resistant (CANR) and sensitive (CANS) clones obtained in a total of 16 experiments with PM-donors (3) and RT-donors (4), showing that in both cases, more than 95.6% of [ade2] transformants are also mutated at the CAN1 locus. (**C**) PFGE karyotypes of two strains carrying the ADE2-CAN1 translocation (YAF190, YAF192) and two strains with restored chromosomes V and XV (YAF194 and YAF199 originating from YAF190 and YAF192, respectively). Chimerical chromosomes are denoted VtXV and XVtV. Original chromosome XV and V from the reference strain BY4741 are indicated in orange and blue respectively.

5.5.3. Engineering a deletion at the translocation breakpoint

In some instances, reciprocal translocations occur between repeated sequences (micro-homology or homologous sequences) that can also be used as template for the reverse translocation event restoring the WT configuration. Stabilizing a phenotype associated to an engineered translocation would therefore require the concomitant deletion of the homology regions to avoid such reversion events. We used our system to simultaneously generate a deletion of a few nucleotides and a reciprocal translocation. We used the same CAN1.Y and ADE2.Y gRNA target sequences as above and designed new donor fragments inducing deletions of 27 and 23 bp, including the PAM sequences, on chromosome V and XV, respectively (Supplementary Figure 2A, page 155). As above, we obtained a high proportion of [*ade2*,*can1*] transformants (96%). PCR of the junctions and karyotyping of 4 strains showed that chromosome V and XV underwent the expected reciprocal translocation (Figure 49A). The genome of one translocated strain was sequenced with Oxford Nanopore MinIon and *de-novo* assembled (Material and Methods, Supplementary Table 2, page 154, Supplementary Table 3, page 154). The translocated and reference genomes are collinear, except for chromosomes V and XV (Figure 49B). No other rearrangement is observed in the translocated strain, suggesting no major off target activity of the Cas9 nuclease. In addition, the junction sequences are identical to the sequences of the chimerical donor fragments ((Supplementary Figure 2B, page 155). These experiments demonstrate that a deletion and reciprocal translocation can be concomitantly engineered at base-pair resolution with CRISPR/Cas9 in the yeast genome.

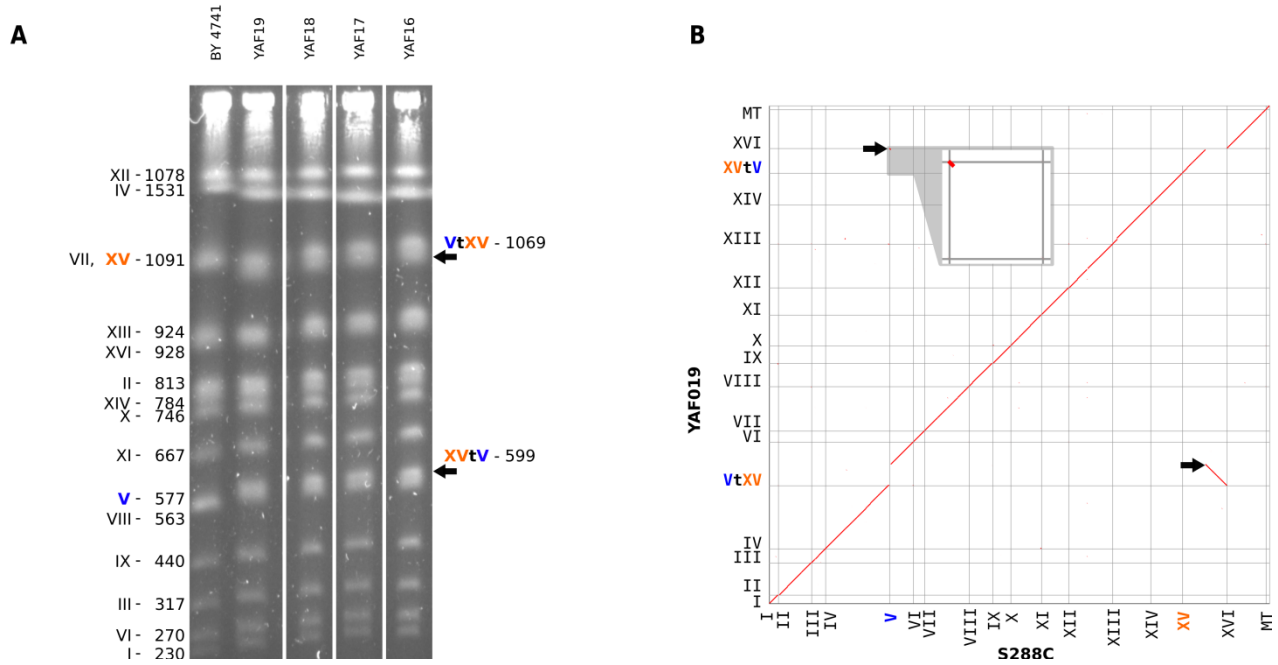


Figure 49 A non-reversible translocation between ADE2 and CAN1 genes. (**A**) PFGE karyotypes of 4 independent strains carrying the translocation (YAF16 to YAF19). All symbols are identical to Fig. 1C. (**B**) Homology matrix of de-novo assembled strain YAF019 vs S288c reference genome. Translocated fragments are indicated by black arrows.

5.5.4. Recapitulating a natural translocation involved in sulphite resistance in wine strains

It was previously reported that a reciprocal translocation between the promoters of *ECM34* and *SSU1*, a sulphite resistance gene, created a chimerical *SSU1-R* allele with enhanced expression resulting in increased resistance to sulphite in the wine strain Y9 (Perez-Ortin, 2002). This translocation resulted from a recombination event between 4 base-pair micro-homology regions on chromosomes VIII and XVI. We engineered the same translocation into the BY4741 background and tested sulphite resistance.

We first designed the two gRNA target sequences as close as possible to the micro-homology regions (Supplementary Figure 3A, page 156). The first gRNA targets the *SSU1* promoter region, 115 base-pairs upstream of the start codon. The second gRNA targets the promoter region of *ECM34*, 24 base-pairs upstream of the start codon (Supplementary Figure 3A, page 156). To mimic the translocated junctions present in the wine strains, we designed double stranded synthetic DNA donors of 90 base-pairs centered on the micro-homology regions but not on the cutting sites. In addition, each donor also contained a point mutation in the PAM sequences to prevent subsequent CRISPR recognition (Supplementary Figure 3A, page 156).

The transformation with the Cas9 plasmid containing the two gRNAs and the donor DNA yielded on average 202 transformants. We tested natural and chimerical junction by colony PCR for 16 transformants and found the expected chimerical junctions in 15 of them (Supplementary Figure 3B, page 156). This translocation was not visible by PFGE karyotyping because the size of the translocated chromosomes were too close to the size of the original chromosomes. To further validate the rearrangement, we checked the junctions by southern blot for one translocated strain with probes flanking the two cutting sites on chromosome VIII and XVI (Figure 50A). This experiment clearly shows the presence of the chimerical junction fragments in the rearranged strain as compared to the WT. Finally, we sequenced the junctions of two translocated strains and found that the rearrangement occurred within the expected micro-homology regions. In addition, the system was designed such that both mutated PAMs ended up into the promoter of *ECM34*, therefore likely to have no impact on the expression of the sulphite resistance gene *SSU1* (Supplementary Figure 3C, page 156).

We then compared sulphite resistance between the lab strain in which we engineered the translocation, the non-translocated parental strain and wine isolates that either carry or are devoid of the translocation of interest. Surprisingly, we found that the engineered strain was the least resistant of all strains, including the non-translocated reference strain (Figure 50B). This suggests that the promoter of *ECM34* in the BY background is weaker than the *SSU1* promoter. The wine isolate with the translocation was the most resistant, followed by the wine isolate without the translocation. It is interesting to note that the most resistant wine strain has tandem repeats of roughly 100bp in the promoter region of *SSU1* gene brought by the translocation with the original *ECM34* locus (Perez-Ortin, 2002). These repeats are absent from the reference strain.

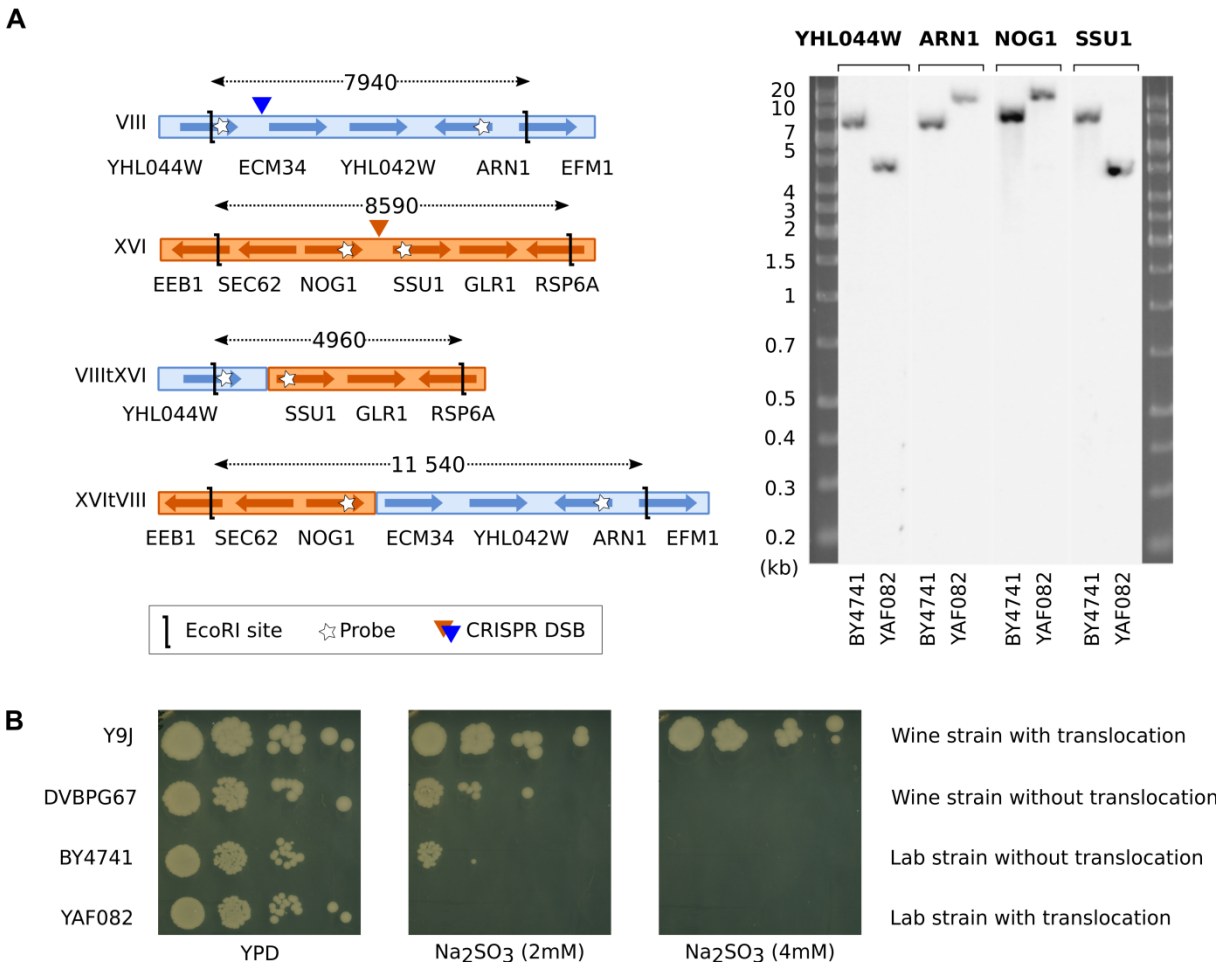


Figure 50 A reciprocal translocation involved in sulphite resistance. (**A**) Left: schematic view of the chromosomal regions surrounding the translocation breakpoints. The double arrows indicate the length of the EcoRI restriction fragments. Right: Southern blot on the translocated strain YAF082 and parental strain BY4741. The probes are indicated on the top of the lanes. (**B**) Quantification of sulphite resistance of the strains Y9J, DVBPG6765, BY4741, YAF082 in the presence of 2 and 4 mM of Na2SO3.

5.5.5. Reshuffling chromosomes with multiple rearrangements

We introduced multiple DSBs in a single step using a unique gRNA targeting repeated Ty3 LTR sequences. There are 35 complete copies of Ty3 LTRs dispersed throughout the genome and these sequences are polymorphic (Figure 51A, Supplementary Figure 4, page 157). Four of them comprise a region identical to the gRNA target sequence and also contain a PAM. A fifth copy, also flanked by a PAM, differs from the target sequence by a single mismatch at its 5' end and therefore might also be recognized by the gRNA (Supplementary Figure 4A, page 157). These 5 copies are located in chromosomes IV, VII, XV and XVI (Figure 51A). All the other Ty3 LTRs contain several mismatches or indels and/or are devoid of PAM, suggesting that Cas9 will not cut at these sites (Supplementary Figure 4A, page 157). Therefore, we expect from 1 to 5 sites to be concomitantly cut upon transformation with the Cas9/gRNA plasmid. In contrast to all experiments described above, we do not provide any donor DNA assuming that DSBs will be repaired by using uncut homologous LTR copies as template.

We transformed BY4741 and BY4742 cells with the Cas9/gRNA plasmid and recovered a total 211 and 159 transformants, respectively. We PFGE karyotyped 42 BY4741 and 36 BY4742 derived strains. Out of these, 39 strains showed clear chromosomal rearrangements on the gels with 29 different karyotypes (Figure

51B). This result demonstrates that genomes are efficiently reshuffled by our strategy. Considering that all transformants must have repaired all DSBs and that all repaired chromosomes must be monocentric, we can predict 23 rearranged karyotypes (types B to X in Supplementary Figure 4B, page 157) as well as a WT-like karyotype that would result from DSB repair without any rearrangement (type A in Supplementary Figure 4B, page 157). In total 20 strains showed such predicted rearrangements, representing 11 distinct types (B, C, D, F, G, H, J, K, R, T and V in Figure 51C). We validated the presence of all expected junctions by colony-PCR in 10 out of the 20 strains and on average 3 junctions out of 5 were also validated by PCR in the remaining 10 strains. We sequenced all chromosomal junctions in 2 strains that show the most frequently observed rearranged karyotype (type J in Supplementary Figure 4B, page 157). We found that all junctions, from both chimerical and un-rearranged chromosomes, were mutated in their PAM compared to the original target sequence (Supplementary Figure 4C, page 157). This shows that during shuffling, all targeted sites were cut and repaired using as donor other Ty3 LTRs that had no PAM. Additional mutations in the region corresponding to the gRNA target sequence were also observed for 3 junctions, but were too few to identify which copy of Ty3 LTR was used as donor (Supplementary Figure 4C, page 157). We compared the theoretical frequency of each possible junction in the predicted types, considering types B to X being equiprobable, with the observed frequency of the junctions in the predicted strains whose type was assigned by PFGE (Figure 51D). We found very similar distributions, suggesting that DSBs were repaired in a random way. Moreover, we obtained 19 strains with 17 distinct unexpected karyotypes involving chromosomes other than the four targeted ones (Figure 51B and Figure 51C). For instance, chromosomes XI and XIV that have no PAM sequence associated with their Ty3 LTRs (Supplementary Figure 4A, page 157) are absent from several karyotypes suggesting that they can be rearranged in the absence of DSB (Figure 51B). In addition, some of these unexpected karyotypes showed an apparent genome size increase suggesting the presence of large duplications (Figure 51B). Using Oxford Nanopore MinION, we sequenced and *de-novo* assembled the genome of a strain showing a DNA content increase in PFGE (black star in Figure 51B). We characterized in this strain an unequal reciprocal translocation between chromosomes VII and XV (Figure 52A). The junctions corresponded to the targeted CRISPR cut site on chromosome XV but not on chromosome VII. In the chimerical chromosome XVtVII, the junction occurred away from the expected site resulting in a 30kb increase in DNA content (Figure 52A). In the chimerical chromosome VIItXV, the junction also occurred away from the expected site but was accompanied by a truncated triplication of a 110 kb region. The missing part of the triplication corresponds to the 30 kb region found in the reciprocal chromosome XVtVII. The sequencing coverage relatively to the reference genome clearly confirmed the triplication of the complete 110 kb region (Figure 52B). The displaced 30kb segment is referred to as region *a* and the other 80 kb segment as region *b*. In summary, one copy of region *a* lies at the chimerical junction of chromosome XVtVII, whereas the remaining two and three copies of region *a* and *b,* respectively, are found in tandem at the chimerical junction of chromosome VIItXV (Figure 52C). Interestingly the 110kb segment, composed of regions *a* and *b*, is flanked by two Ty3 LTRs. Moreover, two full length Ty2 retrotransposons are found, one in chromosome XV directly flanking the targeted Ty3 LTR and the other one in chromosome VII at the junction between regions *a* and *b* (Figure 52C).

Finally, we also recovered 39 type A strains that had a WT karyotype (Figure 51C). For all of them we validated the presence of the un-rearranged junctions by colony PCR. We sequenced the junctions in 4 independent clones. Surprisingly, we found that these clones gather transformants that underwent two

different paths within the same experiment. Firstly, 3 clones had junctions identical to the original LTR sequences with intact PAM, suggesting that Cas9 did not cut the target sites. Secondly, in the fourth clone, the PAM sequences of three junctions were mutated, showing that the corresponding chromosomes were cut and repaired yet without any rearrangement (Supplementary Figure 4C, page 157). In this clone, one junction could not be amplified, possibly because of a small-sized indel that could not be observed on the PFGE profile.

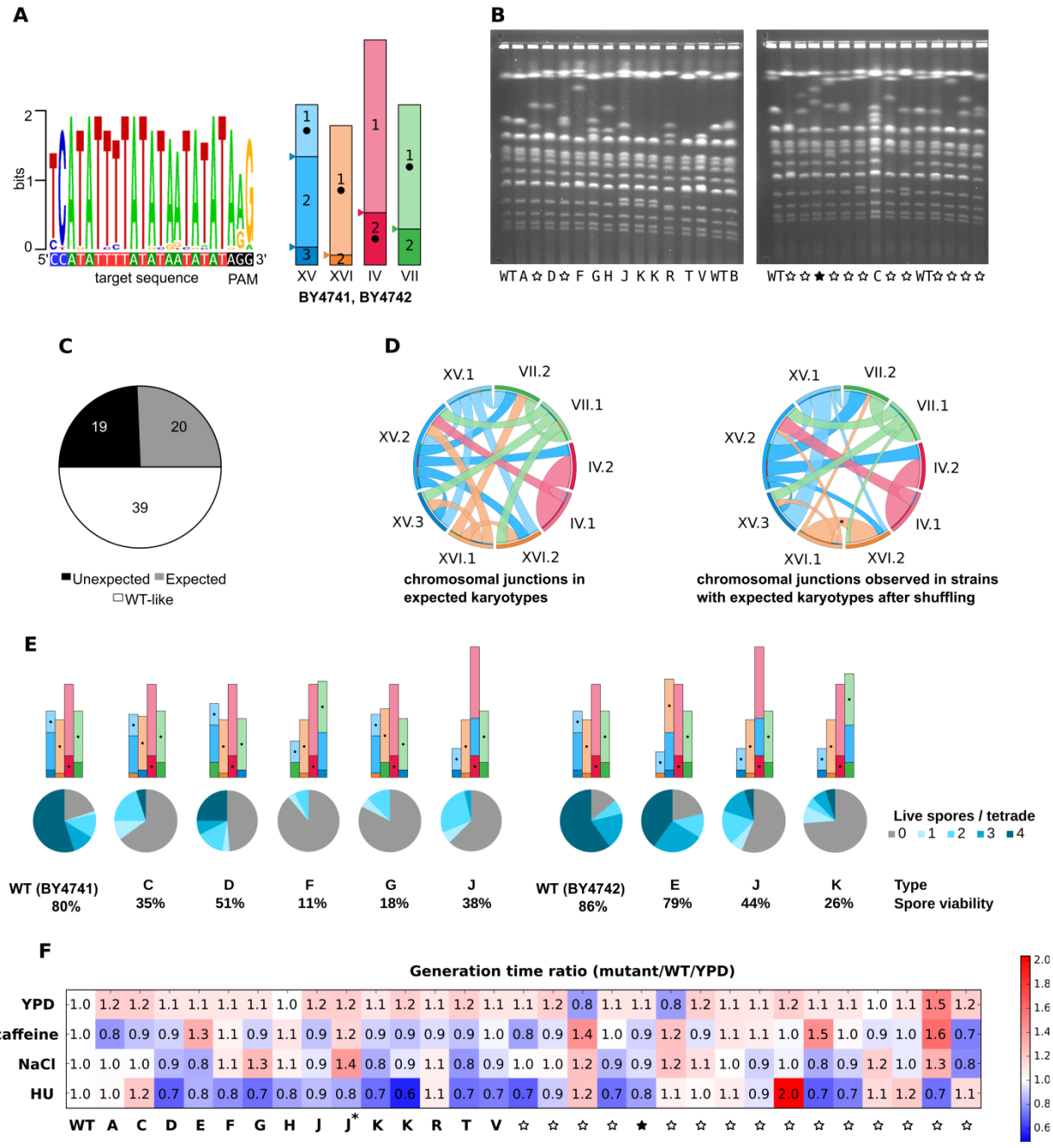


Figure 51 Induction of multiple rearrangements and phenotypic diversity of reshuffled strains. (**A**) Left: logo of the Ty3-LTR regions targeted by CRISPR/Cas9. The chosen target sequence is indicated below the logo with its associated PAM sequence in black. Right: Genomic location of the five LTR sequences that best match the gRNA target sequence. Cutting sites are indicated by triangles. Chromosomes are represented proportionally to their size in kb. Centromeres are represented by black dots, not at their actual positions for readability (**B**) Twenty-six distinct karyotypes obtained by PFGE after after inducing multiple rearrangements, including type A, the wild-type-like. Letters indicate expected karyotypes (see supp fig 4B), that is, karyotypes explained by balanced rearrangements without change in DNA contents, when applicable. Unexpected karyotypes are indicated by a star. The strain with the black star has been sequenced (Figure 52). (**C**) Relative proportion of BY4741 and BY4742 derived strains with a WT like,

predicted or unexpected karyotype. (**D**) Proportions of expected and observed chromosomal junctions. Left: theoretical frequency of chromosomal junctions in expected rearranged karyotypes (types B to X obtained by simulating the balanced repair of the 5 DSBs induced by CRISPR/Cas9 in Ty3-LTRs). Right: observed frequency of chromosomal junctions in type B to X as assigned by PFGE. The junction XVI.1-XVI.2, denoted by * is significantly enriched compared to the prediction (Chi2 conformity test with pval=0.03). (**E**) Percentage of viable spores and proportion of tetrads with 0, 1, 2, 3, 4 viable spores obtained from crosses between rearranged BY and WT SK1 strains. (**F**) Ratio between the generation times of rearranged strains and WT (columns) in various stress conditions (rows). Each ratio is the mean of 3 replicates. When applicable, the type of predicted strains is indicated at the bottom of the matrix, unexpected karyotypes are indicated by a star. J* denotes a type J strain where all junctions could not be validated by PCR. All rows have been normalized by the YPD value except the first row. All columns have been normalized by the WT value.

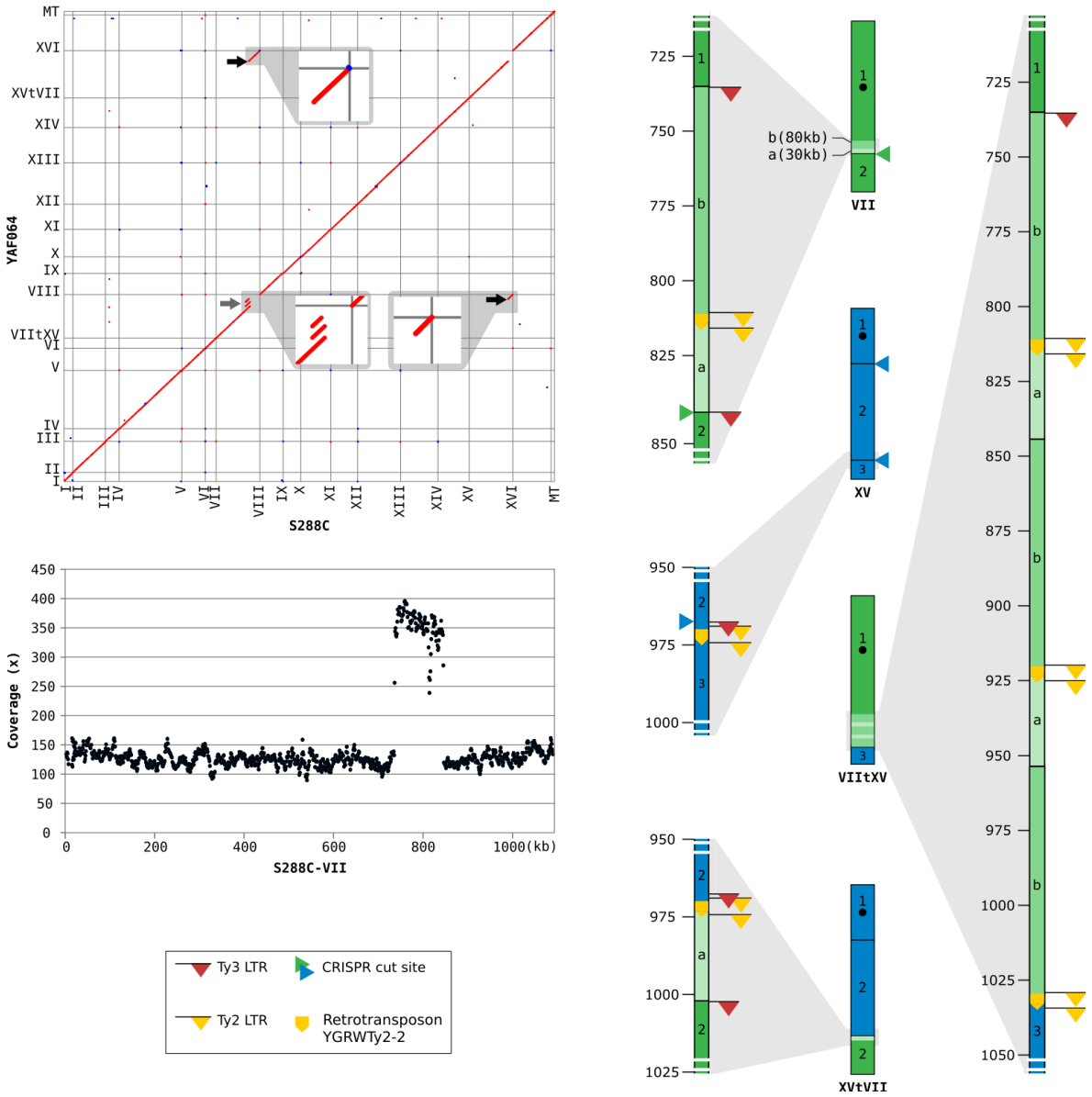


Figure 52 Genome with complex rearrangement. (**A**) Homology matrix between the genomes of the strain showing an increase in global DNA content in PFGE (YAF064) and S288c. Translocated fragments are indicated by black arrows. The tandem triplication at the junction of chromosome VIItXV is indicated by the grey arrow. (**B**) Coverage of the YAF064 reads remapped on the reference chromosome VII. Each dot represents a window of 1kb. (**C**) Architecture of chromosomes VII (in green) and XV (in blue) of the reference strain and chimerical chromosomes VIItXV and XVtVII of the rearranged strain YAF064. Light grey triangles represent zoom-ins on chromosomal junctions. Ty3-LTRs and Ty2 LTRs elements are represented by red and yellow flags respectively. Full-length Ty2 elements are represented by yellow boxes. The Ty3 LTR copies targeted by CRISPR/Cas9 are indicated by green and blue triangles next to chromosomes VII and XV, respectively. Regions a and b, triplicated in the shuffled strain, are represented in lighter green shades.

5.5.6. Exploring the phenotypic diversity of reshuffled strains

Firstly, we tested the meiotic fertility of diploid strains heterozygous for the chromosomal rearrangements by measuring spore viability in their offspring (Figure 51E). We tested 8 different strains with predicted karyotypes and PCR-validated junctions. These strains comprise various rearrangements including reciprocal translocations between 2 and 3 chromosomes (types C, D, E, F and G, respectively) and transpositions (types J and K). For each cross we dissected 10 tetrads from two independent diploid strains (20 tetrads in total). The control strains without any rearrangement shows 80 and 86% of viable spores with most tetrads harbouring 3 to 4 viable spores (Figure 51E). By contrast, all heterozygous diploids had a impaired spore viability ranging from 11 to 51% with predominantly only 1 to 2 viable spores per tetrad, except one that shows 79% of spore viability (type E in Figure 51E). We observed no clear correlation between the type of rearrangement and the impact on fertility. One strain with a single reciprocal translocation between two chromosomes (type F) shows the lowest viability, comparable to that of a more rearranged strain with 2 translocations between 3 chromosomes (type G). By contrast, the strain showing the highest spore viability also has a single translocation between two chromosomes (type E). These results show that knowing the type of rearrangements is not sufficient to predict their quantitative impact on meiotic fertility, although most of them have a drastic negative effect.

Secondly, we measured the mitotic growth of rearranged strains in rich medium and under stress conditions including a variety of abiotic factors or drugs interfering with cell division. NaCl exerts an osmotic stress (Blomberg 1997), hydroxyurea (HU), is an inhibitor of nucleic acid synthesis (Koç et al., 2004), and caffeine is a cell-growth inhibitor known to interfere with the TOR pathway (Loewith and Hall, 2011). We tested representative strains for 13 distinct predicted karyotype categories and 16 unexpected karyotype categories and indicated the generation time ratio of the rearranged mutants relatively to the WT strain and to YPD condition (Figure 51F). Note that the type A strain that has a WT-like karyotype showed either a generation time identical to the WT strain as expected or a slightly increased one (1.2 times, Figure 51F). To be conservative, we decided to consider as significant only the values smaller than 0.8 and greater than 1.2. In rich medium, all but one rearranged strain show no significant growth difference than the WT, showing that chromosome reshuffling has little impact on mitotic fitness in this growth condition (from 1.0 to 1.2 times longer generation times) although in one case the rearrangements, involving extensive aneuploidy, proved highly disadvantageous (second strain on the right with a generation time ratio of 1.5). In both caffeine and NaCl conditions, 4 and 3 rearranged strains were negatively affected respectively, with the slow grower in the rich medium being also strongly impaired in these two conditions (Figure 51F). By contrast, in HU, the numbers of mutants positively and negatively impacted are very imbalanced with 11 and 1 strains growing slightly faster and much slower than the WT, respectively. Interestingly, we found both predicted and unexpected karyotypes showing significant growth differences in the various environmental conditions with both positively and negatively impacted strains. In addition, several strains showed alternatively increased and decrease mitotic fitness depending on the conditions showing genotype by environment interactions.

5.6. Discussion

In this study, we developed a versatile CRISPR/Cas9-based method allowing to engineer, with the same efficiency as point mutations, both precisely targeted reciprocal translocations and multiple rearrangements. Cloning in a single step any two pair of gRNAs into a CRISPR/Cas9 vector allows generating any translocation at base-pair resolution on-demand. After the translocation, the new chromosomal junctions are not recognized by Cas9 which makes the incorporation of point mutation in the PAM sequence unnecessary. The preservation of an intact PAM sequence allows re-targeting the chimerical junctions by Cas9 and inducing the backward translocation restoring the original chromosomal configuration. Alternatively, introducing a small deletion at the translocation breakpoint ensures the stability of the rearranged configuration. Cloning in a single step a single gRNA that targets dispersed repeat sequences allows generating multiple chromosomal rearrangements simultaneously leading to a large diversity of karyotypes. Only one half of them could be explained by the expected repair in trans between the targeted DSBs (Figure 51C). Considering that these predictable rearrangements were only represented by a small number of strains, some being not represented at all (Supplementary Figure 4B, page 157) and that the unexpected karyotypes were all nearly unique, the number of possible rearrangements accessible by our method is probably very large.

One reason that could explain some of the unexpected karyotypes would be that CRISPR/Cas9 may recognize and cut the two Ty3 LTRs on chromosome XII that contain a PAM despite small indels in the gRNA target sequence (Supplementary Figure 4A, page 157). Such types would be difficult to identify by PFGE because of the large size of chromosome XII and its frequent size variations due to the rDNA cluster dynamics. One cannot exclude that additional LTRs were cut by CRISPR/Cas9 even if this seems very unlikely because all other LTRs are devoid of PAM. A more plausible explanation would be multipartite ectopic recombination between cut and uncut chromosomes, as in (Piazza et al., 2017). In this study, the authors report that the single end of a DSB can invade multiple sequences located on intact chromosomes during its search for homology. This leads to translocations between intact chromosomes with insertions of fragments of the triggering single-end sequence in the translocation junctions. Consistently with this possibility we found that one unexpected chimerical junction resulted from Ty3-LTRs being “outcompeted” during repair by the full-length Ty2 elements found in the vicinity of the CRISPR-induced DSBs (Figure 52). We also found several karyotypes with rearrangements in untargeted chromosomes suggesting that rearrangements can occur in intact chromosome in the absence of DSB.

Our procedure of genome reshuffling is reminiscent of the restructuring technology developed by Muramoto and collaborators (2018) that uses a temperature-dependent endonuclease to conditionally induce multiple DSBs in the genome of yeast and *A.thaliana*. However, our reshuffling procedure is potentially more versatile because we can target different types of repeated sequences with variable locations (within or between genes, subtelomeric or internal, etc) and variable copy numbers. By analyzing the genome of *S. cerevisiae* for repeated gRNA target sequences we identified 39,374 repeated targets with varying number of occurrences in the genome (from 2 to 59). Repeated targets were mostly found within Ty transposable elements (11,503) and protein coding genes (10,368 in 479 genes located in all 16

chromosomes). Interestingly, we also found 3,438 repeated target gRNA located in intergenic regions which should allow to rearrange chromosomes without disrupting coding sequences. In addition, our CRISPR/Cas9 shuffling procedure is doable in any genetic background and therefore potentially more versatile than the Cre/Lox based SCRaMbLE techniques that can be used only in yeast strains with synthetic chromosomes.
Our approach provides the possibility to quantify the role of structural variants in organismal phenotypes. The translocation between the *SSU1* and *ECM34* genes provides increased sulphite resistance to wine isolates and is believed to be a consequence of sulphite supplementation in wine-making (Perez-Ortin, 2002). Surprisingly, the translocation we engineered in the reference strain resulted in decreased sulphite resistance. We hypothesize that this phenotype is linked to the absence of repeats in the promoter of *ECM34* in the BY4741 background in contrast to wine strains (Perez-Ortin, 2002), implying that the promoter of *ECM34* is weaker than the *SSU1* promoter in the BY background. Therefore, the translocation itself would not be sufficient to explain increased sulphite resistance. These findings provide a striking example of the advantages brought by our technique to untangle the phenotypic impact of SVs from that of the genetic background.

We also measured the phenotypes of strains carrying multiple rearrangements both in mitotic and meiotic conditions. We showed that crosses between rearranged and parental strains produced very few viable spores, supporting earlier studies revealing that meiotic pairing and segregation is impaired in diploid strains bearing heterozygous translocations (Avelar et al., 2013; Liti et al., 2006; Loidl et al., 1998). In vegetative growth, reshuffling provided a great phenotypic diversity with strains showing fitness advantages under specific environmental conditions. Similar findings were previously described both in *S. cerevisiae* (Colson et al., 2004; Naseeb et al., 2016) and *Schizosaccharomyces pombe* (Avelar et al., 2013). However, in our study all rearrangements are completely markerless and scarless (no CRE/Lox site) and no gene was disrupted nor duplicated (at least for the predicted karyotypes) suggesting that balanced rearrangements between repeated sequences simply reconfiguring the chromosome architecture are sufficient to create fitness diversity. Therefore, we believe that CRISPR/Cas9 chromosome reshuffling is a powerful new tool complementing the features of pre-existing genome restructuring technologies and will prove useful in the study of genotype to phenotype relationship and for isolating new chromosomal combinations of biotechnological interests.

5.7. Supplementary Tables

Supplementary Table 1 : Oligonucleotides used in this study

Name	Sequence	Type	Description
Reciprocal translocation of ADE2/CAN1			
synth1	GTGAAAGCTCTTCAATCACTTGAAGATTCTTTAGTGTGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG GCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGTTTTTTTGTTTTTTATGTCTGCGGCCGCG GTACCCAATTCGCTCTTTGAAAAGATAATGTATGATTATGCTTTCACTCATATTTATACAGAAACTTGATGT TTTCTTTCGAGTATATACAAGGTGATTACATGTACGTTTGAAGTACAACTCTAGATTTTGTAGTGCCCTCTT GGGCTAGCGGTAAAGGTGCGCATTTTTTCACACCCTACAATGTTCTGTTCAAAAGATTTTGGTCAAACGCT GTAGAAGTGAAAGTTGGTGCGCATGTTTCGGCGTTCGAAACTTCTCCGCAGTGAAAGATAAATGATCGA TACGTTCTCTATGGAGGAGTTTGAAGAGCAGAAAT	double gRNA	synth DNA ready for cloning in pGZ110 for translocating CAN1 and ADE2 genes
AF117	CTCTTTCCCGACGAGAGTAAATGGCGAGGATACGTTCTCTATGGA TGTAGGAACATCAACATGCTCAATCTCAATCGTTAGCACATCACA	donor	translocation ADE2/CAN1
AF118	TGTGATGTGCTAACGATTGAGATTGAGCATGTTGATGTTCCTACA TCCATAGAGAACGTATCCTCGCCATTTACTCTCGTCGGGAAAGAG	donor	reverse complement of AF117
AF119	AGGGTAAATTTTTAATTTGGGATGTTTTACTTGAAGATTCTTTAG GGATGGCATAGGTGATGAAGATGAAGGAGAAGTACAGAACGCTGA	donor	translocation ADE2/CAN1
AF120	TCAGCGTTCTGTACTTCTCCTTCATCTTCATCACCTATGCCATCC CTAAAGAATCTTCAAGTAAAACATCCCAAATTAAAAATTTACCCT	donor	reverse complement of AF119
Point-mutation of ADE2/CAN1			
AF121	AGGGTAAATTTTTAATTTGGGATGTTTTACTTGAAGATTCTTTAGTGT cta AACATCAACATGCTCAATCTCAATCGTTAGCACATCACA	donor	ADE2 repair cassette with STOP codon
AF122	TGTGATGTGCTAACGATTGAGATTGAGCATGTTGATGTT tag ACACTAAAGAATCTTCAAGTAAAACATCCCAAATTAAAAATTTACCCT	donor	reverse complement of AF121
AF8	CCCGACGAGAGTAAATGGCGAGGATACGTTCTCTATGGAGGAT tag	donor	CAN1 repair cassette with STOP codon

	ATAGGTGATGAAGATGAAGGAGAAGTACAGAACGCTGAAGTGAA		
AF9	TTCACTTCAGCGTTCTGTACTTCTCCTTCATCTTCATCACCTAT cta ATCCTCCATAGAGAACGTATCCTCGCCATTTACTCTCGTCGGG	donor	reverse complement of AF8
Translocation with deletion of ADE2/CAN1			
AF25	TGCTTAAGCTCTCTCTTCACTTCAGCGTTCTGTACTTCTCCTTCATCTTC ATCTTCAAGTAAAACATCCCAAATTAAAAATTTACCCTTCTCCAGAAACA	donor	translocation CAN1/ADE2 with deletion
AF26	TGTTTCTGGAGAAGGGTAAATTTTTAATTTGGGATGTTTTACTTGAAGAT GAAGATGAAGGAGAAGTACAGAACGCTGAAGTGAAGAGAGAGCTTAAGCA	donor	reverse complement of AF25
AF27	AATTGTATCCATTGCGCTCTTTCCCGACGAGAGTAAATGGCGAGGATACG TGCTCAATCTCAATCGTTAGCACATCACATTTTTCAGCTAGTTTTTCGAT	donor	translocation CAN1/ADE2 with deletion
AF28	ATCGAAAAACTAGCTGAAAAATGTGATGTGCTAACGATTGAGATTGAGCA CGTATCCTCGCCATTTACTCTCGTCGGGAAAGAGCGCAATGGATACAATT	donor	reverse complement of AF27
Reverse translocation of ADE2/CAN1			
synth5	gtgaaaGCTCTTCaATCgatacgttctctatggatgtgttttagagctagaaatagcaagttaaaataaggctagtccgttatcaacttg aaaaagtggcaccgagtcggtgtttttttgttttttatgtctgcggccgcggtacccaattcgctctttgaaaagataatgtatgattatgcttt cactcatatttatacagaaacttgatgttttctttcgagtatatacaaggtgattacatgtacgtttgaagtacaactctagattttgtagtgc cctcttgggctagcggtaaaggtgcgcatttttttcacaccctacaatgttctgttcaaaagattttggtcaaacgctgtagaagtgaaagttg gtgcgcatgtttcggcgttcgaaacttctccgcagtgaaagataaatgatcacttgaagattctttagggagtttgaagagcagaaat	double gRNA	synth DNA ready for cloning in pGZ110 for restoring CAN1 and ADE2 genes
AF123	CCCGACGAGAGTAAATGGCGAGGATACGTTCTCTATGGAGGATGGCATAGGTGATGAAGATGAAGGAG AAGTACAGAACGCTGAAGTGAA	donor	Reverse translocation CAN1/ADE2 for chV
AF124	TTCACTTCAGCGTTCTGTACTTCTCCTTCATCTTCATCACCTATGCCATCCTCCATAGAGAACGTATCCTCGC CATTTACTCTCGTCGGG	donor	reverse complement of AF123
AF125	AGGGTAAATTTTTAATTTGGGATGTTTTACTTGAAGATTCTTTAGTGTAGGAACATCAACATGCTCAATCT CAATCGTTAGCACATCACA	donor	Reverse transloc TP CAN1/ADE2 for chXV
AF126	TGTGATGTGCTAACGATTGAGATTGAGCATGTTGATGTTCCTACACTAAAGAATCTTCAAGTAAAACATCC CAAATTAAAAATTTACCCT	donor	reverse complement of AF125

PCR validation at ADE2/CAN1 loci			
AF32	TTTCACGACGTTGAAGCTTCACAAA	PCR oligo	validation of ADE2/CAN1 translocations
AF33	ACAAATAAGCAACTCCAATGACCAC	PCR oligo	validation of ADE2/CAN1 translocations
AF34	AGGAACACTTTGGGTAACTGCTATA	PCR oligo	validation of ADE2/CAN1 translocations
AF35	AAAGACCTGTACCAATAGTACCACC	PCR oligo	validation of ADE2/CAN1 translocations
SSU1/ECM34 translocation involved in sulphite resistance in wine strains			
Name	**Sequence**	**Type**	**Description**
synth4	gtgaaaGCTCTTCaATCTAACTGCAAAAAAATGTCACgttttagagctagaaatagcaagttaaaataaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtggtgctttttttgttttttatgtctgcggccgcggtacccaattcgctctttgaaaagataatgtatgattatgctttcactcatatttatacagaaacttgatgttttctttcgagtatatacaaggtgattacatgtacgtttgaagtacaactctagattttgtagtgccctcttgggctagcggtaaaggtgcgcatttttttcacaccctacaatgttctgttcaaaagattttggtcaaacgctgtagaagtgaaagttggtgcgcatgtttcggcgttcgaaacttctccgcagtgaaagataaatgATCATTTGTGTGTTGTCTAAATGgtttgaagagcagaaat	double gRNA	double gRNA synthetic fragment for SSU1/ECM34 translocation
AF74	TGAATTTACGAGCTGTATAAAAGAACTACAAGGAAGTTGTAACTG CAAATTACAGCTTTCCCCTAGTAACGATTGTTGATTGAGCTCAGA	donor	translocation SSU1/ECM34 promoters
AF77	TCTGAGCTCAATCAACAATCGTTACTAGGGGAAAGCTGTAATTTG CAGTTACAACTTCCTTGTAGTTCTTTTATACAGCTCGTAAATTCA	donor	reverse complement of AF74
AF75	ACTTGTGATATTGGCTGAACAAATTCTC t GCATTTAGACAACACA CAAAAAAATGTCACC a GGACGATATCATCACAAATATGGAGGGCC	donor	translocation SSU1/ECM34 promoters
AF76	GGCCCTCCATATTTGTGATGATATCGTCC t GGTGACATTTTTTTG TGTGTTGTCTAAATGC a GAGAATTTGTTCAGCCAATATCACAAGT	donor	reverse complement of AF75
PCR validation of SSU1/ECM34 translocation			
AF78v2	GGAGGATAAGTATCCAGAGATGG	PCR oligo	validation of translocation SSU1/ECM34 promoters
AF79	ACCGATGCCTGAAAAAACC	PCR oligo	validation of translocation SSU1/ECM34 promoters
AF80	GTCTGTAGCCGAAAATCTGA	PCR oligo	validation of translocation SSU1/ECM34 promoters

AF81v2	GCTATTACTAGACAGACTGCGAG	PCR oligo	validation of translocation SSU1/ECM34 promoters
Southern blot validation of SSU1/ECM34 translocation			
AF127	gctgatctggagatgagaagtaatgg	PCR oligo	amplify probe YHL044W for
AF128	tggaggacagcttgaacctccg	PCR oligo	amplify probe YHL044W rev
AF129	aggatgacagcactaatgaaaaggt	PCR oligo	amplify probe ARN1 for
AF130	cagtagttcaagccactgcatagc	PCR oligo	amplify probe ARN1 rev
AF131	gggttctacaattctgacgatgagg	PCR oligo	amplify probe NOG1 for
AF132	ctgtcttaccgacaccacgc	PCR oligo	amplify probe NOG1 rev
AF133	acttgctcttacgaggcagttt	PCR oligo	amplify probe SSU1 for
AF134	attcccaatcccttccaaaggg	PCR oligo	amplify probe SSU1 rev
Genome shuffling			
AF88	atcccatattttatataatatat	gRNA	Ty3 LTR guide forward
AF89	aacatatattatataaaatatgg	gRNA	Ty3 LTR guide reverse
AF92	acgtactcaaacaactacactt	PCR oligo	multiple cuts Ty3 LTR 15.2 end forward
AF93v2	TCAATTAGGAAGGGCAGCTTG	PCR oligo	multiple cuts Ty3 LTR 7.2 begin reverse version 2
AF94	aaactccgccttctatagctg	PCR oligo	multiple cuts Ty3 LTR 4.1 end forward
AF95v2	gtaaccagttgtcagaacgga	PCR oligo	multiple cuts Ty3 LTR 15,2 begin reverse
AF96v2	CCTTAGCTAACCTACTCTAACG	PCR oligo	multiple cuts Ty3 LTR 7.1 end forward
AF97	ggaatcttagcgccttaaatgc	PCR oligo	multiple cuts Ty3 LTR 4,2 begin reverse
AF98v2	tctcagagtggtacacgtcag	PCR oligo	multiple cuts Ty3 LTR 15,1 end start
AF99	gtccctattccgataatcttagcag	PCR oligo	multiple cuts Ty3 LTR 15,3 start reverse
AF113	ctagctgaacaactcaaacg	PCR oligo	multiple cuts Ty3 LTR 16.1 start
AF114	tctcgacgattactttagatc	PCR oligo	multiple cuts Ty3 LTR 16.2 end

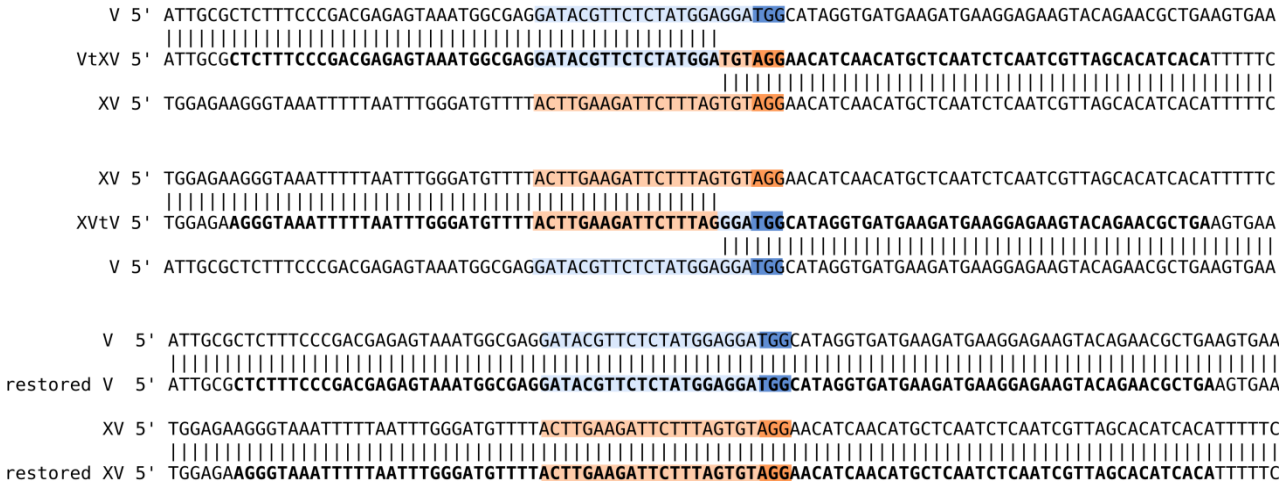
Supplementary Table 2 : Flowcell outputs

	YAF019		YAF064	
	All reads	Reads with Q ≥ 7	All reads	Reads with Q ≥ 7
Total gigabases	0.2864264	0.2795613	2.7346126	2.6395977
Total reads	38229	33380	785220	548331
N50 length	10505.0	10587.0	7031.0	7135.0
Mean length	7492.4	8375.1	3482.6	4813.9
Median length	7403.0	8175.0	2311.0	3857.0

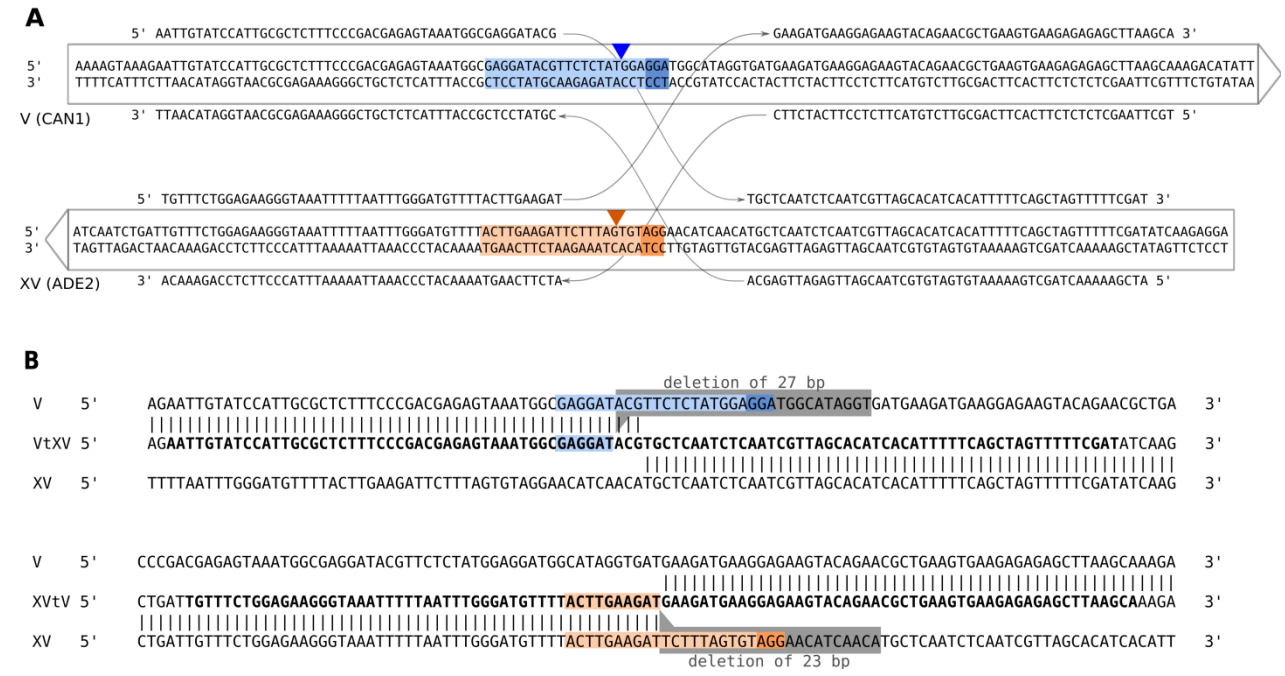
Supplementary Table 3 : SMARTdenovo assemblies

	YAF019	YAF064
Total sequence count	35	24
Total sequence length	12002666	12102300
Min sequence length	28808	18399
Max sequence length	1050679	1520788
Mean sequence length	342933.31	504262.50
Median sequence length	252028.00	503804.50
N50	589863	770201
L50	8	6
N90	194937	434841
L90	23	14
A%	30.89	30.86
C%	19.08	19.14
G%	19.18	19.15
T%	30.86	30.86
AT%	61.74	61.71
GC%	38.26	38.29
N%	0.00	0.00

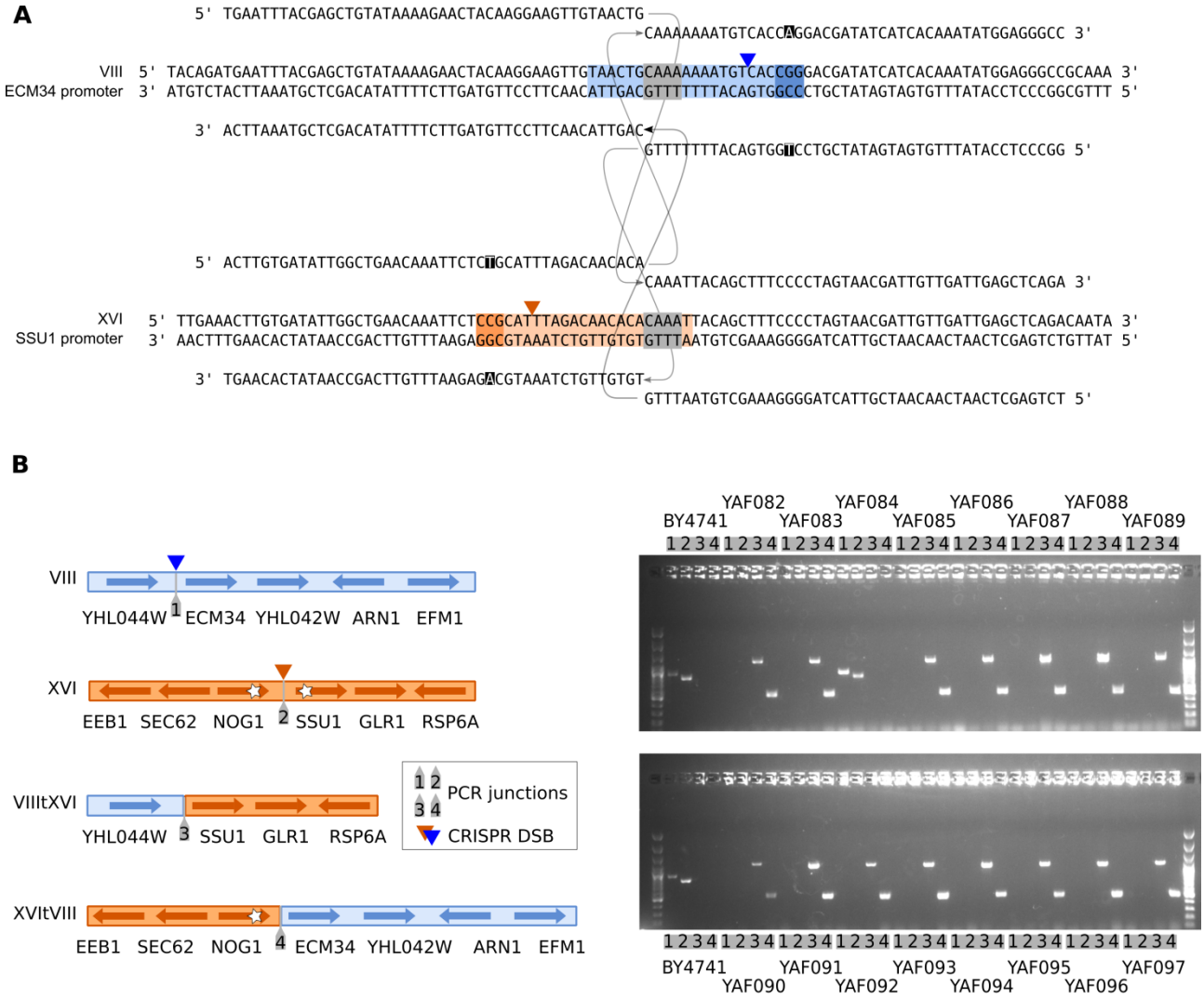
5.8. Supplementary Figures



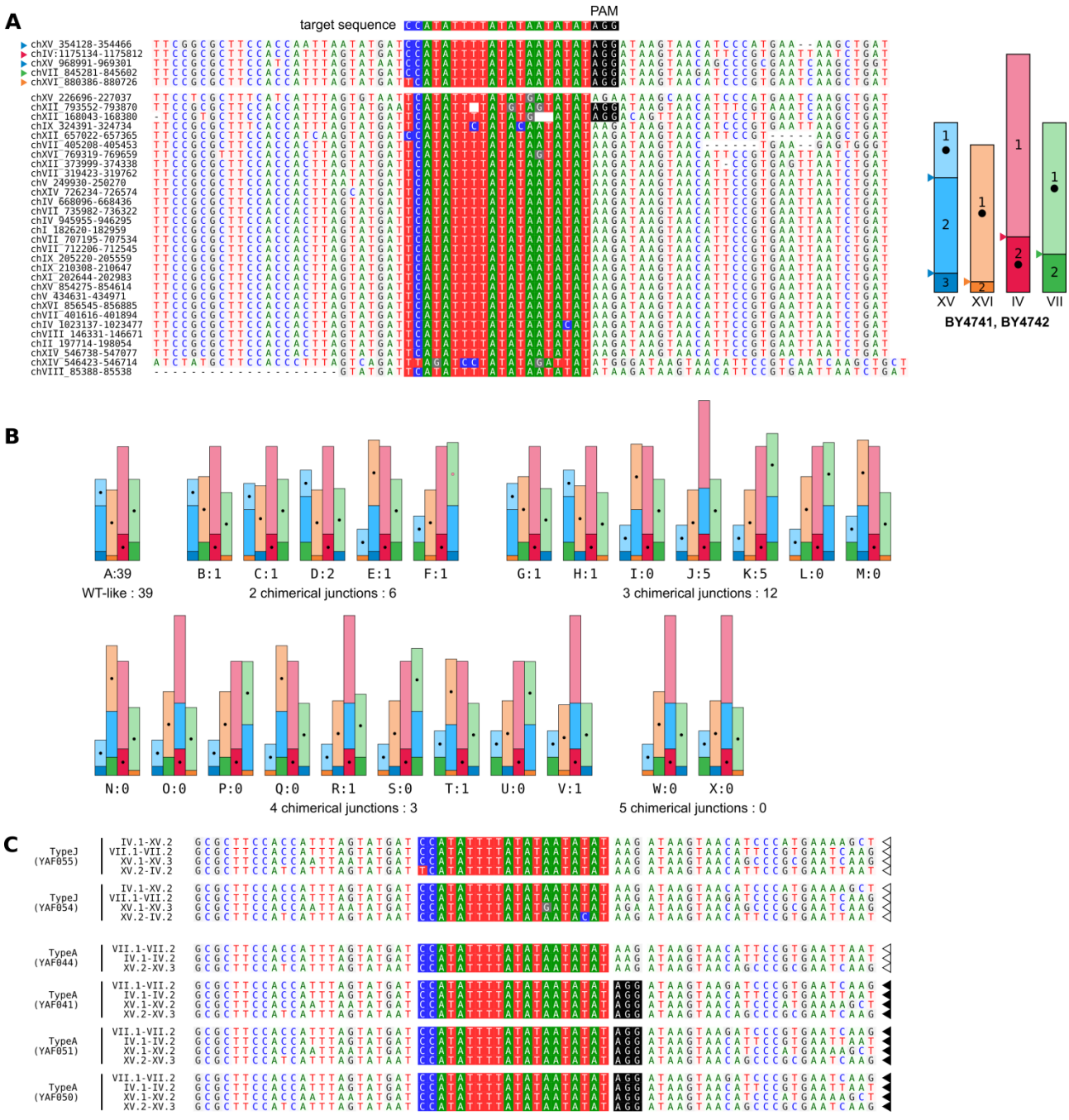
Supplementary Figure 1 Sanger sequencing of the translocated junctions in YAF190, YAF192 and de-translocated junctions in strains YAF194, YAF199. The sequences corresponding to donor oligonucleotides are shown in bold. The two gRNA target sequences are highlighted in light blue and orange. PAM sequences are highlighted in dark blue and orange.



Supplementary Figure 2 (**A**) Targeted sequences and donors used to engineer the translocation with a small deletion. The two gRNA target sequences are highlighted in light blue and orange. PAM sequences are highlighted in dark blue and orange. Triangles indicate DSBs sites. Arrows framing the sequences indicate the orientation of coding phases. Donor nucleotides are represented above and below the frames by sequences linked by thin arrows to indicate their homology with the two different chromosomes. (**B**) Alignments of the *de-novo* assembled chimerical junctions on reference chromosomes V and XV. Donor sequences used to direct the translocation are in bold. Deleted sequences in chromosomes V and XV are highlighted in grey. The translocation occurred at the targeted position with a base-pair resolution.



Supplementary Figure 3 (**A**) gRNAs and donors used to perform the translocation between the promoters of SSU1 and ECM34. The micro-homology regions between the two chromosomes are highlighted in grey. The two gRNA target sequences are highlighted in light blue and orange. PAM sequences are highlighted in dark blue and orange. Triangles indicate DSBs sites. Mismatches brought by the donors to inactivate the PAMs are highlighted in black. (**B**) Colony PCR on 16 transformants. The 2 natural and chimerical junctions are denoted 1, 2 and 3, 4, respectively. (**C**) Sanger sequencing of YAF82 chimerical junctions. The two mismatches brought by the donors to prevent CRISPR/Cas9 activity after repair (indicated by stars) are present in chromosome XVItVIII.



Supplementary Figure 4 (**A**) Left: Multiple alignment of Ty3-LTR sequences. The gRNA target sequence is indicated on the top and the corresponding homologous regions in the LTRs are highlighted below. The associated PAM sequence is highlighted in black. The five LTR sequences at the top are the best matches to the gRNA target sequence. Right: location of the five best matches of the chosen gRNA. (**B**) Predicted karyotypes (types A to X) achievable by the repair of the DSBs induced in 5 Ty3-LTRs. Numbers indicate the total of strains obtained from BY4741 and BY4742 that carry the corresponding karyotype, as observed by PFGE. Chromosomes are represented proportionally to their size in kb. For readability centromeres are represented in the middle of their carrying fragment. Karyotypes with 2 chimerical junctions have undergone a single translocation. Karyotypes with 3 chimerical junctions have undergone either two translocations between three chromosomes (types G, H, L, M) or one transposition (types I, J, K). Karyotypes with 4 chimerical junctions have undergone a combination of translocations and transpositions. Karyotypes with 5 chimerical junctions have lost all natural junctions. (**C**) Sanger sequencing of the junctions of two type J strains (YAF055 and YAF054) and four type A strains (YAF044, YAF041, YAF051, YAF050). Junctions where the PAM was mutated are indicated by white triangles. Sequences where the PAM is intact are indicated by black triangles and PAM is highlighted in black.

Funding

This work was supported by the Agence Nationale de la Recherche [ANR-16-CE12-0019].

Acknowledgements

We thank our colleagues, Gianni Liti (IRCAN, France) and Joseph Schacherer (Université de Strasbourg, France) for fruitful discussions and constructive suggestions. We are grateful to Bruce Futcher and Gang Zhao (Stony Brook University, USA) for providing the pGZ110 plasmid and to Bertrand Llorente (CRCM, France) and Joseph Schacherer for the SK1 and wine strains, respectively. We are also grateful to Maëlys Born-Bony for her experimental help and to the 'Biologie Moléculaire et Cellulaire' Master students from Sorbonne Université who performed transformation experiments during their practical in the frame of the teaching module called 'Biologie synthetique et ingenierie des genomes' (4V150).

Data Availability

The Oxford Nanopore sequencing data are deposited in the Sequence Read Archive under the project number (accession number pending).

CONCLUSION ET PERSPECTIVES

« Quel signal phylogénétique utiliser pour reconstruire un arbre phylogénétique des *Saccharomycotina* ? Quels critères utiliser pour reconstruire des génomes ancestraux de qualité ? Faut-il utiliser toute l'information disponible ? Combien d'inversions et de translocations ont remanié ces génomes au cours de l'évolution ? A quel rythme s'accumulent les réarrangements dans les différentes lignées ? Les différents types de réarrangements ont-ils tous la même prévalence selon les lignées ? » Ces questions sont celles que nous avons tenté d'approfondir dans cette thèse.

Nous avons tout d'abord rassemblé autant de génomes de *Saccharomycotina* que possible à partir bases de données publiques. Après avoir filtré les génomes de manière à disposer d'un jeu de données exploitable, nous nous sommes attelés à la construction d'un arbre phylogénétique des 66 génomes rassemblés. Nous avons construit un arbre sur la base des taux de mutations non-synonymes dans les homologues synténiques. Le fait de combiner la similarité de séquence à la conservation de la synténie permet d'identifier des gènes dont le caractère homologue est très probable. L'arbre obtenu (Figure 33, page 111) est cohérent avec les topologies pré-existantes reposant sur la séquence de l'ARN ribosomique 16S et avec des topologies parues ultérieurement, générées à partir de l'information complète du génome (Shen et al., 2016a). Nous accordons donc une grande confiance à l'arbre que nous avons obtenu. Ce dernier a servi de référence tout au long de ce travail de thèse.

Nous avons ensuite exploré des signaux phylogénétiques moins « classiques ». Dans un premier temps nous avons construit un arbre phylogénétique à partir du nombre de « *Reciprocal Best Hits* » (RBH) entre paires d'espèces sans utiliser la séquence de ces derniers, ce qui est inhabituel. En effet, pour construire une phylogénie, on exploite en principe les données de séquences des gènes appartenant à des familles. Nous voulions néanmoins connaitre la résolution du signal contenue dans la comparaison des protéomes complets. Il ressort de cette analyse que la topologie obtenue est assez cohérente (60%) avec celle de l'arbre de référence (Figure 34, page 114). En revanche, dans l'arbre construit à partir des nombres de RBH, les branches menant aux espèces actuelles sont plus longues que les branches internes comparativement à l'arbre de référence. Ces résultat suggèrent que l'évolution du répertoire de gènes (du moins son approximation assez grossière par le nombre de RBH) reflète assez bien l'histoire des espèces bien qu'étant moins résolutive. Ces observations font écho à la contribution de la divergence de séquence sur l'isolement reproductif et la spéciation, évoquée dans la partie introductive.

Dans un second temps, nous avons généré un arbre phylogénétique à partir des nombres de réarrangements chromosomiques inférés par *PhyChro* entre les paires de génomes actuels des *Saccharomycotina*. Comme nous l'avons vu, la topologie de cet arbre partage 86% de bipartitions correctes avec l'arbre de référence, indiquant que l'évolution de la structure des génomes est cohérente avec l'histoire des espèces (Figure 35, page 115 et Figure 36, page 116). Nous avons vu cependant que les longueurs de branches inférées par *PhyChro* ne sont pas corrélées aux longueurs de branches de l'arbre généré par *PhyML*. Ce résultat était assez surprenant, car dans le clade des *Lachancea*, qui nous a servi de modèle pour la reconstitution de l'histoire évolutive des génomes, les branches de l'arbre généré par *PhyChro* et par *PhyML* corrèlent (non représenté). Cela indique qu'à une grande échelle évolutive, le signal des réarrangements chromosomiques est saturé et on observe de ce fait un « tassement » de la structure interne de l'arbre. Notons d'ailleurs que les erreurs dans la topologie générée par *PhyChro* concernent

principalement des nœuds anciens. Nous pourrions à l'avenir quantifier plus précisément l'intervalle évolutif dans lequel il est pertinent d'utiliser des génomes actuels pour inférer des réarrangements chromosomiques passés, mais nous prédisons que la puissance de cette approche est très rapidement bruitée (voir l'insert de la Figure 36). Cependant, cette approche est peut être utilisable sur des ensembles de génomes de qualité comme ceux des *Lachancea*, comparativement à certains génomes que nous avons utilisé (voir plus bas).

Afin d'inférer des réarrangements chromosomiques de manière plus fiable, nous avons décidé de comparer la structure de génomes ancestraux reconstruits avec le logiciel *AnChro*. Nous avons dans un premier temps évalué la pertinence de ce logiciel sur des données réelles et des données simulées. Il ressort de ces analyses qu'*AnChro*, en étant capable d'exploiter l'information contenue dans les blocs de synténie entre paires de génomes actuels, génère des génomes ancestraux à la fois très contigus, denses en nombre de gènes retracés et avec un faible nombre d'adjacences erronées (Figure 25, page 96 et Figure 26, page 97). *AnChro* nous a permis de reconstruire la structure des génomes ancestraux de 66 levures du subphylum des *Saccharomycotina* (Figure 41, page 122), qui ont divergé des *Pezizomycotina* il y a 798 à 1166 millions d'années. Comme nous l'avons montré, *AnChro* est assez robuste vis-à-vis de l'utilisation de génomes fragmentés et capable de générer des reconstructions ancestrales plus contigües que les génomes actuels (Figure 42, page 123). A ce titre, il serait intéressant de quantifier la pertinence des reconstructions qu'on obtiendrait selon une utilisation « progressive » d'*AnChro* pour reconstruire des génomes ancestraux, c'est-à-dire en utilisant des génomes ancestraux reconstruits comme base pour reconstruire des génomes ancestraux plus lointains. Ces derniers bénéficieraient du fait que des adjacences sont déjà reconstruites, minimisant ainsi le niveau de fragmentation des ancêtres lointains qu'on reconstruirait. En revanche, dans cette approche la propagation d'erreurs est un risque. Cela souligne l'importance d'utiliser des génomes actuels de qualité. Le fait qu'*AnChro* soit capable de reconstruire des ancêtres plus contigus que les génomes qui ont servi à les reconstruire suggère que les reconstructions ancestrales générées par *AnChro* pourraient dans une certaine mesure être utilisées pour ordonner les *scaffolds* dans des génomes actuels mal assemblés. La similitude entre les algorithmes d'assemblage de données de séquençage reposant sur des graphes de de Bruijn et les outils de reconstruction de génomes ancestraux reposant sur les graphes de points de cassure a d'ailleurs été remarquée (Lin et al., 2014) et plusieurs approches « d'assemblage guidé par référence » ont été développées sur la base de cette idée (Aganezov et al., 2015; Muñoz et al., 2010).

Nous sommes parvenus à inférer au total 5150 réarrangements balancés dans l'arbre des *Saccharomycotina* dont 1513 (29,4%) sont des inversions, 1174 (22,8%) sont des translocations et 2463 (47,8%) sont des événements non-discernables entre ces deux catégories (Figure 44, page 126). Nous avons en outre montré que le nombre de réarrangements chromosomiques balancés inféré par la comparaison des génomes des nœuds consécutifs de l'arbre des 66 génomes (deux génomes ancestraux consécutifs ou une espèce actuelle avec son ancêtre le plus récent) corrèle avec la longueur de branche inférée par *PhyML* sur la base des taux de mutations non-synonymes (Figure 46, page 129). Ce résultat avait été rapporté dans le clade des *Lachancea* (Vakirlis et al., 2016) et il est intéressant de voir qu'il se généralise à l'ensemble des 66 *Saccharomycotina* étudiées. Cela montre que l'accumulation des réarrangements chromosomiques et des mutations non-synonymes se fait de manière coordonnée au cours de l'évolution. La base mécanistique de cette horloge moléculaire n'est pas identifiée mais on peut imaginer la contribution de mécanismes

mutagènes de réparation des CDB, comme par exemple le mécanisme BIR qui peut induire des translocations, des pertes d'hétérozygotie sur de longues portions chromosomiques ainsi qu'un grand nombre de mutations lors des phases de recherche d'homologie par *template-switching.* Une étude récente rapporte qu'une des caractéristiques génomiques des souches pathogènes de *C. albicans* est la présence dans le génomes de longues régions de pertes d'hétérozygotie encadrées par des régions dans lesquelles le taux de mutation *de-novo* est 800 fois supérieur. Cette observation laisse imaginer comment les taux de réarrangements chromosomiques et de mutations non-synonymes peuvent être coordonnés. Quoi qu'il en soit, la cohérence de ces deux signaux renforce la pertinence de la comparaison de génomes ancestraux pour inférer des réarrangements chromosomiques et montre que cette approche permet de s'affranchir au moins en partie du bruit lié à la comparaison de génomes actuels, comme le montre également une étude récente (Sacerdot et al., 2018).

Au-delà des proportions globales des types de réarrangements inférés (29,4% d'inversions, 22,8% de translocations et 47,8% d'événements non-discernables) nous avons vu que la proportion de ces différents types de réarrangements est variable selon les lignées considérées, certaines évoluant principalement par inversion tandis que d'autres évoluent majoritairement par translocations (Figure 44, page 126). Il serait intéressant de rechercher les propriétés des génomes à l'origine de ces deux modes évolutifs.

S'agit-il de caractéristiques « mécanistiques » de ces génomes ? En comparant les familles de gènes dans les génomes « à inversions » ou les génomes « à translocations » il est peut-être possible de déterminer des familles de gènes candidates expliquant ces modes d'évolution. Ces familles pourraient être mutuellement exclusives entre ces deux catégories de génomes ou bien comprendre un nombre variable de membres.

Pourrait-il s'agir de ressemblances en termes d'architecture des génomes ? Des études précédentes ont rapporté que la dispersion des points de cassure n'est pas aléatoire dans les génomes mais dépend notamment de la longueur des intergènes (Berthelot et al., 2015). Cette caractéristique semble peu explicative du fait que certains clades évoluent principalement par inversion ou par translocation. En effet, les génomes des espèces du clade CTG qui évoluent principalement par inversion, ont des génomes de taille assez variables (jusqu'à 50% d'écart) alors que leur nombre de gènes est à peu près comparable.

S'agirait-il alors de ressemblances entre le nombre et la position des séquences répétées du génome, notamment des éléments transposables ? Nous avons vu dans l'introduction que les éléments transposables sont capables d'induire de nombreux réarrangements chromosomiques en générant des CDB, et/ou en fournissant un substrat ectopique pour la réparation de ces lésions. Nous avons d'autre part montré expérimentalement que des cassures multiples dans des éléments transposables génèrent un grand nombre réarrangements différents. De plus, de nombreux caryotypes obtenus suggèrent que des chromosomes initialement intacts ont été réarrangés. On peut donc s'interroger sur la prévalence des réarrangements dus aux transposons dans les génomes des *Saccharomycotina*. Il serait d'ailleurs très intéressant de comparer la localisation des éléments transposables avec celle des points de cassure des événements non-discernables (47,8% dans l'arbre des *Saccharomycotina*) car, rappelons-le, ces événements font intervenir des points de cassure réutilisés au cours de l'évolution et/ou témoignent de réarrangements complexes. Malheureusement, la dynamique des éléments transposables n'est pas bien décrite pour

beaucoup d'espèces de notre jeu de données. D'ailleurs, dans 60% des génomes que nous avons rassemblés, les éléments mobiles du génome ne sont pas annotés et aucun gène d'élément transposable n'a été annoté à première vue (nous avons recherché dans ces annotations la co-occurrence ainsi que les occurrences isolées des termes suivants : « mobile », « element », « transpos », « repeat », « Ty », « like »). Les génomes les plus anciennement séquencés que nous ayons utilisés remontent aux années 2000, époque à laquelle les génomes n'étaient déjà plus considérés comme de simples « collections » de gènes reliés entre eux par des séquences inutiles. Aussi, les éléments transposables ne sont probablement pas annotés en raison du fait que ces éléments sont assez difficiles à détecter de manière automatisée dans les génomes des *Saccharomycotina*, chez qui ils représentent le plus souvent autour de 5% du génome (Bleykasten-Grosshans and Neuvéglise, 2011). Dans un grand nombre de génomes pour lesquels on considère que les éléments transposables sont annotés (soit parce que les éléments mobiles sont annotés en tant que tels, soit parce qu'on identifie des clusters de gènes de transposons), c'est-à-dire 26 des 66 génomes analysés, les éléments mobiles sont souvent annotés comme « Ty-like », témoignant d'une approche comparative pour identifier ces éléments et non d'une recherche *de-novo*. En outre, les éléments transposables représentent une difficulté à l'assemblage des contigs en chromosomes complets, du moins pour les techniques de séquençage « Sanger » et les technologies NGS. Ces difficultés tendent à disparaitre avec le développement du séquençage en longue molécule unique.

Il est certain que des génomes actuels mal assemblés sont susceptibles de propager des erreurs dans les reconstructions. Nous souhaitions donc pour poursuivre cette thèse, corriger les assemblages préexistants en collaboration avec l'équipe de Romain Koszul, à l'Institut Pasteur. Ce laboratoire a publié une approche algorithmique performante et d'une grande élégance pour corriger des assemblages erronés à partir de données 3C, c'est-à-dire de capture de conformation chromosomique (Marbouty et al.; Marie-Nelly et al., 2014). A cet effet, nous avons d'ores et déjà rassemblé la quasi-totalité des souches de levures utilisées dans cette étude. Avec ces données, il serait possible de comparer la localisation des points de cassure au cours de l'évolution des *Saccharomycotina* (identifié à partir de la comparaison de génomes ancestraux de haute qualité) avec des données de contacts chromosomiques. Comparativement aux génomes de mammifères chez qui les interactions 3D de la chromatine expliquent en grande partie la distribution des points de cassure dans les génomes (Berthelot et al., 2015), chez les quelques levures pour lesquelles des données de capture de conformation chromosomique existent (des *Saccharomycetaceae*), on observe assez peu de contacts entre les différents chromosomes. Il est cependant possible que des comportements différents soient observés chez des levures plus divergées. En outre, ces données permettraient d'étudier la conservation des points de contact au cours de l'évolution des génomes, ce qui n'a jamais été entrepris sur un intervalle évolutif aussi large.

Par ailleurs, plusieurs améliorations techniques ont été et pourraient encore être apportées au pipe-line d'analyse actuel. Tout d'abord, en termes de temps de calcul, l'étape la plus longue de *SynChro* est l'identification des RBH car cette étape consiste à calculer les scores de similarité entre toutes les protéines de chaque paire de génomes du jeu de données. Par comparaison, le temps de calcul est d'environ 5 minutes. C'est une étape incontournable qui ne peut pas être réduite. Pour cette raison, la seule amélioration que nous avons pu apporter pour diminuer le temps nécessaire à l'identification des RBH consiste à paralléliser le code, ce qui réduit le temps de calcul par le nombre de processeurs disponibles. La

même amélioration a été apportée à *AnChro*.

En outre, ce dernier peut générer autant de versions d'un génome ancestral A qu'il existe de paires informatives de génomes G1/G2 reliées par un chemin évolutif passant par A (Figure 37, page 117), auquel on peut encore multiplier le nombre de combinaisons possibles de génomes G3…Gn qu'on utilise comme référence. Cette propriété est intéressante quand on peut se permettre de calculer toutes les combinaisons, toutefois quand le nombre de génomes ancestraux à reconstruire est grand, il devient souhaitable d'exclure *a priori* les combinaisons G1/G2. C'est ce que nous avons fait en implémentant le choix des paires d'espèces G1/G2 tel que décrit à la Figure 38, page 119. Cette stratégie a permis d'obtenir des combinaisons uniques par ancêtre et des reconstructions très pertinentes en nombre de gènes retracés et en nombre de *scaffolds*. Cependant nous avons également vu que le paramètre de 38% de divergence n'était pas suffisamment strict, une valeur de 37 ou 36% étant préférable. On pourrait également remplacer complètement cette stratégie pour choisir les génomes G1/G2 dont la couverture en synténie est maximale. De cette manière, on maximise le nombre de gènes qui entrent dans la comparaison G1/G2 et le contenu en gènes de l'ancêtre reconstruit. De plus cette approche semble plus facilement transposable à d'autres types de génomes pour lesquels la distribution du nombre de blocs de synténie en fonction du pourcentage de divergence serait différente de celle observée chez les *Saccharomycotina*.

Comme nous l'avons vu, *AnChro* génère 36 versions de chaque reconstruction ancestrale, quelle que soit la combinaison de génomes actuels utilisés. En effet, Δ et Δ' pouvant prendre les valeurs 1 à 6, il y a 6x6 reconstructions par paires de génomes informatives G1/G2 dans l'arbre phylogénétique. Plusieurs méthodes sont envisageables pour obtenir une version unique de l'ancêtre c'est-à-dire un couple (Δ, Δ') unique. La stratégie actuelle consiste à choisir la meilleure reconstruction selon une optimisation hiérarchisée en minimisant d'abord la somme du nombre *scaffolds* et de contradictions inter-chromosomiques, puis en identifiant la reconstruction possédant le plus de gènes (Vakirlis et al., 2016). Une autre méthode consisterait à extraire un « consensus » à partir des reconstructions obtenues des différents couples (Δ, Δ'). Il faudrait tout d'abord construire un graphe dans lequel les nœuds représentant les gènes retracés seraient reliés par des arrêtes dont le poids représenterait le nombre de fois où cette arrête est observée dans les reconstructions issue des différents couples (Δ, Δ'). Il s'agirait alors de trouver un compromis entre un chemin de haut poids et de longueur maximale. Enfin, une interface graphique permettant d'utiliser CHROnicle de manière plus ergonomique est presque finalisée.

Au cours de cette thèse, nous avons également étudié l'impact phénotypique des réarrangements chromosomiques chez *S. cerevisiae* dans le but de répondre aux questions suivantes : L'impact phénotypique d'un réarrangement chromosomique est-il le même dans différents fonds génétiques ? Le génome est-il capable de se réparer à la suite de cassures multiples ? Quel est l'impact phénotypique des réarrangements chromosomiques sur la croissance végétative et la viabilité des spores dans des souches dont aucun gène n'a été détruit ou remanié ?

Afin de répondre à ces questions nous avons dans un premier temps développé une méthode reposant sur le système CRISPR/Cas9 afin de pouvoir introduire à façon des réarrangements chromosomiques dans le génome de *S. cerevisiae* (Figure 47, page 138). Nous avons ensuite montré que l'utilisation de cette

méthode permet d'induire et de défaire des translocations réciproques avec 98% d'efficacité dans le génome de *S. cerevisiae* sans introduire de mutations additionnelles (Figure 48, page 140). Cette approche a ensuite été utilisée pour induire une translocation réciproque entre le promoteur du gène *SSU1* et du gène *ECM34*. Cette translocation avait été préalablement rapportée comme conférant une résistance accrue aux sulfites dans les souches de vin de *S. cerevisiae* (Perez-Ortin, 2002). Etonnamment, la souche transloquée, dont la jonction chimérique en amont du gène *SSU1* correspondait en tout point à la séquence de la jonction observée dans les souches de vin, ne possède pas le phénotype de résistance aux sulfites. Nous en avons déduit que la translocation seule ne confère pas le phénotype. Les souches de vin présentent en plus de cette translocation, des répétitions de 80 paires de bases dans leur promoteur chimérique, répétitions que nous prévoyons à l'avenir d'introduire dans le génome de notre souche transloquée afin de déterminer si ces derniers restaureront le phénotype attendu. Si ce n'était pas le cas, cela signifierait que le phénotype lié à cette translocation dépend d'interactions avec le fond génétique.

Le contexte technologique actuel est propice pour répondre à ce genre de questions. Comme nous l'avons vu dans l'introduction, il est aujourd'hui possible de tester un grand nombre de variants génétiques en parallèle (Roy et al., 2018; Sadhu et al., 2018; Sharon et al., 2018). A ce jour, ces techniques se sont focalisées sur les mutations ponctuelles. Toutefois, compte tenu de l'efficacité de la méthode expérimentale développée au cours de ce travail de thèse, il est tout à fait envisageable d'appliquer le même genre d'approche haut-débit pour générer et tester l'impact de réarrangements chromosomiques. On pourrait pour cela transformer en parallèle des cultures indépendantes de *S. cerevisiae* avec une combinaison de plasmides-CRISPR/oligonucléotides de réparation et sélectionner les cellules transformées par la simple acquisition du plasmide pour récupérer des souches dont le génome est remanié. On pourrait en outre réaliser ce genre d'expériences dans différents fonds génétiques (nous avons à cet effet cloné un plasmide contenant un gène de résistance à l'hygromycine plutôt que le gène *LEU2*). On pourrait alors valider les jonctions chromosomiques chimériques des variants obtenus par PCR. Alternativement, on pourrait envisager d'intégrer les sequences de réparation au plasmide et transformer directement un mélange de plasmides dans des cellules pour quantifier l'avantage relatif que procure chaque variant.

Nous avons également utilisé CRISPR/Cas9 pour induire des CDB multiples dans les Ty3-LTR de *S. cerevisiae*. Dans cette expérience, que nous avons baptisée « *reshufflling* », les Ty3-LTR reconnus par le système CRISPR/Cas9 sont coupés tandis que les autres copies polymorphes ne sont pas coupées et servent de modèle de réparation aux CDB (Figure 51, page 145). Nous avons ainsi obtenu 50% de transformants réarrangés. Nous avons remarqué avec intérêt que seulement la moitié des caryotypes réarrangés correspondaient à des remaniements prévisibles par la simple jonction des extrémités des CDB. Les autres transformants réarrangés présentent des remaniements chromosomiques impliquant des chromosomes que nous n'avons par ciblés, probablement selon un mécanisme récemment décrit (Piazza et al., 2017), ou bien présentent des duplications ou des degrés variables d'aneuploïdie (Figure 51, page 145 et Figure 52, page146). Nous en avons déduit que comme les caryotypes prédits sont représentés par de petits effectifs, ou sont même pour certains totalement absents, la diversité des variants générés est très importante.

Nous avons ensuite mesuré l'impact phénotypique des réarrangements générés avec cette technique dans différentes conditions. Nous avons notamment montré que les souches diploïdes obtenues par croisement

d'une souche non-réarrangée avec une souche réarrangée produisent une très faible proportion de spores viables (Figure 51, page 145). Ces résultat sont cohérents avec d'autres études antérieures (Avelar et al., 2013; Liti et al., 2006; Loidl et al., 1998). Toutefois nous ne sommes pas parvenus à relier le type ou le nombre de réarrangements, ni la taille des fragments transloqués avec les taux de spores viables obtenus. Cela suggère que d'autres facteurs contribuent à la viabilité des spores issues de souches réarrangées à l'état hétérozygote. On sait d'une part que la recombinaison des chromosomes est indispensable à la formation de crossing-overs et à la bonne ségrégation des chromosomes en méiose. D'autre part, on sait que la fréquence de recombinaison le long des chromosomes n'est pas uniforme. On peut donc imaginer que la localisation des hotspots de recombinaison joue un rôle important sur la viabilité des spores issues de souches hétérozygotes pour des réarrangements chromosomiques.

Le *reshuffling* des chromosomes génère des souches réarrangées avec une grande diversité phénotypique en croissance végétative. On remarque avec intérêt que certaines souches obtenues présentent des avantages de croissance significatifs dans certaines conditions environnementales (Figure 51, page 145). Des résultats comparables ont été rapportés dans d'autres études chez *S. cerevisiae* (Colson et al., 2004; Naseeb et al., 2016) et *Schizosaccharomyces pombe* (Avelar et al., 2013). Même des souches dans lesquelles aucune phase codante n'a été détruite ou dupliquée présentent des phénotypes significativement différents de la souche sauvage, soutenant que l'architecture du génome peut avoir un impact phénotypique. En somme, cette nouvelle technique de *reshuffling* des chromosomes est une nouvelle façon de générer de la diversité caryotypique et phénotypique. Cette technique présente plusieurs avantages. Elle est utilisable dans n'importe quelle souche. Nous avons ainsi *reshufflé* la souche SK1 et obtenu des proportions comparables de souches réarrangées. Elle permet également de cibler différents types de séquences en nombre variable. Combinée à une approche de sélection des transformants dans des conditions de croissance spécifiques, cette technique possède donc un fort potentiel pour générer des souches d'intérêt biotechnologique, à la manière des techniques SCRaMbLE reposant sur Cre-Lox.

Les techniques d'édition génomique reposant sur le système CRISPR/Cas9, englobant également la technique développée au cours de cette thèse, bénéficient d'un grand potentiel pour disséquer l'interaction phénotype-génotype. On peut aujourd'hui créer rapidement et efficacement presque n'importe quel variant génétique sans marques de l'édition du génome dans un grand nombre d'organismes différents. A titre d'exemple, des souches de *S. cerevisiae* ont été transloquées à plusieurs reprises dans le cadre de cette thèse. Pourtant, une fois les plasmides CRISPR/Cas9 chassés des cellules éditées, rien ne permet de distinguer la souche de laboratoire originale de la souche ayant ce passé d'édition génomique. Dans cette optique, les questions les plus évidentes qu'on peut se poser sont : un organisme édité est-il considéré comme un organisme génétiquement modifié (dont le génome contient par définition un fragment d'ADN exogène) ? Comment contrôler l'utilisation de CRISPR/Cas9 alors que l'utilisation en est rendue toujours plus facile ? Quelle sera la réelle portée de l'utilisation de CRISPR/Cas9 alors que les applications reposant sur ce système ou ses variantes sont toujours plus nombreuses ?

Il est certain que ce système repousse les limites actuelles de la biologie et permettra de répondre à autant de problématiques scientifiques qu'il apporte de questions sociétales et éthiques.

REFERENCES BIBLIOGRAPHIQUES

Adam, Z., and Sankoff, D. (2008). The ABCs of MGR with DCJ. Evol. Bioinforma. Online *4*, 69–74.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185–2195.

Aganezov, S., Sitdykova, N., AGC Consortium, and Alekseyev, M.A. (2015). Scaffold assembly based on genome rearrangement analysis. Comput. Biol. Chem. *57*, 46–53.

Agier, N., Romano, O.M., Touzain, F., Cosentino Lagomarsino, M., and Fischer, G. (2013). The Spatiotemporal Program of Replication in the Genome of Lachancea kluyveri. Genome Biol. Evol. *5*, 370–388.

Aguilera, A., and Gaillard, H. (2014). Transcription and Recombination: When RNA Meets DNA. Cold Spring Harb. Perspect. Biol. *6*.

Aguilera, A., and García-Muse, T. (2012). R Loops: From Transcription Byproducts to Threats to Genome Stability. Mol. Cell *46*, 115–124.

Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. *16*, 197–212.

Alekseyev, M.A., and Pevzner, P.A. (2009). Breakpoint graphs and ancestral genome reconstructions. Genome Res. *19*, 943–957.

Alexander, W.G. (2018). A history of genome editing in Saccharomyces cerevisiae. Yeast Chichester Engl. *35*, 355–360.

Andersson, S.G., and Kurland, C.G. (1995). Genomic evolution drives the evolution of the translation system. Biochem. Cell Biol. Biochim. Biol. Cell. *73*, 775–787.

Andersson, J.O., Doolittle, W.F., and Nesbø, C.L. (2001). Are There Bugs in Our Genome? Science *292*, 1848–1850.

Annaluru, N., Muller, H., Mitchell, L.A., Ramalingam, S., Stracquadanio, G., Richardson, S.M., Dymond, J.S., Kuang, Z., Scheifele, L.Z., Cooper, E.M., et al. (2014). Total Synthesis of a Functional Designer Eukaryotic Chromosome. Science *344*, 55–58.

Anselmetti, Y., Luhmann, N., Bérard, S., Tannier, E., and Chauve, C. (2018). Comparative Methods for Reconstructing Ancient Genome Organization. Methods Mol. Biol. Clifton NJ *1704*, 343–362.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796–815.

Arcangioli, B. (1998). A site- and strand-specific DNA break confers asymmetric switching potential in fission yeast. EMBO J. *17*, 4503–4510.

Arcangioli, B., and de Lahondès, R. (2000). Fission yeast switches mating type by a replication–recombination coupled process. EMBO J. *19*, 1389–1396.

Ariumi, Y. (2016). Guardian of the Human Genome: Host Defense Mechanisms against LINE-1 Retrotransposition. Front. Chem. *4*.

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature *444*, 171–178.

Avdeyev, P., Jiang, S., Aganezov, S., Hu, F., and Alekseyev, M.A. (2016). Reconstruction of Ancestral Genomes in Presence of Gene Gain and Loss. J. Comput. Biol. J. Comput. Mol. Cell Biol. *23*, 150–164.

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

Ayala, D., Ullastres, A., and González, J. (2014). Adaptation through chromosomal inversions in Anopheles. Front. Genet. *5*.

Aylon, Y., Liefshitz, B., and Kupiec, M. (2004). The CDK regulates repair of double-strand breaks by homologous recombination during the cell cycle. EMBO J. *23*, 4868–4875.

Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D., and Eichler, E.E. (2004). Hotspots of mammalian chromosomal evolution. Genome Biol. *5*, R23.

Baker, E., Wang, B., Bellora, N., Peris, D., Hulfachor, A.B., Koshalek, J.A., Adams, M., Libkind, D., and Hittinger, C.T. (2015). The Genome Sequence of Saccharomyces eubayanus and the Domestication of Lager-Brewing Yeasts. Mol. Biol. Evol. *32*, 2818–2831.

Bao, Z., Xiao, H., Liang, J., Zhang, L., Xiong, X., Sun, N., Si, T., and Zhao, H. (2015). Homology-Integrated CRISPR–Cas (HI-CRISPR) System for One-Step Multigene Disruption in *Saccharomyces cerevisiae*. ACS Synth. Biol. *4*, 585–594.

Barlow, J.H., Faryabi, R.B., Callén, E., Wong, N., Malhowski, A., Chen, H.T., Gutierrez-Cruz, G., Sun, H.-W., McKinnon, P., Wright, G., et al. (2013). Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell *152*, 620–632.

Barry, D., and Hartigan, J.A. (1987). [Statistical Analysis of Hominoid Molecular Evolution]: Rejoinder. Stat. Sci. *2*, 209–210.

Barsoum, E., Martinez, P., and Åström, S.U. (2010). α3, a transposable element that promotes host sexual reproduction. Genes Dev. *24*, 33–44.

Béguin, P., Charpin, N., Koonin, E.V., Forterre, P.,

and Krupovic, M. (2016). Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Res. *44*, 10367–10376.

Bérard, S., Gallien, C., Boussau, B., Szöllősi, G.J., Daubin, V., and Tannier, E. (2012). Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics *28*, i382–i388.

Berthelot, C., Muffato, M., Abecassis, J., and Roest Crollius, H. (2015). The 3D organization of chromatin explains evolutionary fragile genomic regions. Cell Rep. *10*, 1913–1924.

Betrán, E., Santos, M., and Ruiz, A. (1998). ANTAGONISTIC PLEIOTROPIC EFFECT OF SECOND-CHROMOSOME INVERSIONS ON BODY SIZE AND EARLY LIFE-HISTORY TRAITS IN DROSOPHILA BUZZATII. Evol. Int. J. Org. Evol. *52*, 144–154.

Bhutkar, A., Schaeffer, S.W., Russo, S.M., Xu, M., Smith, T.F., and Gelbart, W.M. (2008). Chromosomal Rearrangement Inferred From Comparisons of 12 Drosophila Genomes. Genetics *179*, 1657–1680.

Biémont, C. (2010). A Brief History of the Status of Transposable Elements: From Junk DNA to Major Players in Evolution. Genetics *186*, 1085–1093.

Blanchette, null, Bourque, null, and Sankoff, null (1997). Breakpoint Phylogenies. Genome Inform. Workshop Genome Inform. *8*, 25–34.

Blanchette, M., Green, E.D., Miller, W., and Haussler, D. (2004). Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res. *14*, 2412–2423.

Bleykasten-Grosshans, C., and Neuvéglise, C. (2011). Transposable elements in yeasts. C. R. Biol. *334*, 679–686.

Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.-L.V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. Nature *494*, 234–237.

Blount, B.A., Gowers, G.-O.F., Ho, J.C.H., Ledesma-Amaro, R., Jovicevic, D., McKiernan, R.M., Xie, Z.X., Li, B.Z., Yuan, Y.J., and Ellis, T. (2018). Rapid host strain improvement by in vivo rearrangement of a synthetic yeast chromosome. Nat. Commun. *9*.

Bolton, E.C., and Boeke, J.D. (2003). Transcriptional Interactions Between Yeast tRNA Genes, Flanking Genes and Ty Elements: A Genomic Point of View. Genome Res. *13*, 254–263.

Bon, E., Casaregola, S., Blandin, G., Llorente, B., Neuvéglise, C., Munsterkotter, M., Guldener, U., Mewes, H.-W., Helden, J.V., Dujon, B., et al. (2003). Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. Nucleic Acids Res. *31*, 1121–1135.

Bonnet, A., Grosso, A.R., Elkaoutari, A., Coleno, E., Presle, A., Sridhara, S.C., Janbon, G., Géli, V., Almeida, S.F. de, and Palancade, B. (2017). Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability. Mol. Cell *67*, 608-621.e6.

Bordelet, H., and Dubrana, K. (2018). Keep moving and stay in a good shape to find your homologous recombination partner. Curr. Genet.

Boulé, J.-B., and Zakian, V.A. (2006). Roles of Pif1-like helicases in the maintenance of genomic stability. Nucleic Acids Res. *34*, 4147–4153.

Boulé, J.-B., and Zakian, V.A. (2007). The yeast Pif1p DNA helicase preferentially unwinds RNA–DNA substrates. Nucleic Acids Res. *35*, 5809–5818.

Bourque, G., and Pevzner, P.A. (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res. *12*, 26–36.

Bowen, N.J., Jordan, I.K., Epstein, J.A., Wood, V., and Levin, H.L. (2003). Retrotransposons and Their Recognition of pol II Promoters: A Comprehensive Survey of the Transposable Elements From the Complete Genome Sequence of Schizosaccharomyces pombe. Genome Res. *13*, 1984–1997.

Branzei, D., and Foiani, M. (2008). Regulation of DNA repair throughout the cell cycle. Nat. Rev. Mol. Cell Biol. *9*, 297–308.

Britten, R.J., and Davidson, E.H. (1969). Gene regulation for higher cells: a theory. Science *165*, 349–357.

Britten, R.J., and Davidson, E.H. (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q. Rev. Biol. *46*, 111–138.

Brown, K.S., and Benson, W.W. (1974). Adaptive Polymorphism Associated with Multiple Mullerian Mimicry in Heliconius numata (Lepid. Nymph.). Biotropica *6*, 205.

Brown, W.R.A., Liti, G., Rosa, C., James, S., Roberts, I., Robert, V., Jolly, N., Tang, W., Baumann, P., Green, C., et al. (2011). A Geographically Diverse Collection of Schizosaccharomyces pombe Isolates Shows Limited Phenotypic Variation but Extensive Karyotypic Diversity. G3 Bethesda Md *1*, 615–626.

Brunet, E., and Jasin, M. (2018). Induction of Chromosomal Translocations with CRISPR-Cas9 and Other Nucleases: Understanding the Repair Mechanisms That Give Rise to Translocations. In Chromosome Translocation, Y. Zhang, ed. (Singapore: Springer Singapore), pp. 15–25.

Brunet, E., Simsek, D., Tomishima, M., DeKelver, R., Choi, V.M., Gregory, P., Urnov, F., Weinstock, D.M., and Jasin, M. (2009). Chromosomal translocations induced at specified loci in human stem cells. Proc. Natl. Acad. Sci. *106*, 10620–10625.

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

Bushman, F.D. (2003). Targeting Survival: Integration Site Selection by Retroviruses and LTR-Retrotransposons. Cell *115*, 135–138.

Butler, G., Kenny, C., Fagan, A., Kurischko, C., Gaillardin, C., and Wolfe, K.H. (2004). Evolution of the MAT locus and its Ho endonuclease in yeast species. Proc. Natl. Acad. Sci. U. S. A. *101*, 1632–1637.

Butler, G., Rasmussen, M.D., Lin, M.F., Santos, M.A.S., Sakthikumar, S., Munro, C.A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J.L., et al. (2009). Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature *459*, 657–662.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012–2018.

Cannavo, E., and Cejka, P. (2014). Sae2 promotes dsDNA endonuclease activity within Mre11-Rad50-Xrs2 to resect DNA breaks. Nature *514*, 122–125.

Carmel, L., Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. (2007). Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. *17*, 1034–1044.

Carr, M., Bensasson, D., and Bergman, C.M. (2012). Evolutionary Genomics of Transposable Elements in Saccharomyces cerevisiae. PLoS ONE *7*.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol. Biol. Evol. *17*, 540–552.

Ceccaldi, R., Rondinelli, B., and D'Andrea, A.D. (2016). Repair Pathway Choices and Consequences at the Double-Strand Break. Trends Cell Biol. *26*, 52–64.

Chambers, S.R., Hunter, N., Louis, E.J., and Borts, R.H. (1996). The mismatch repair system reduces meiotic homeologous recombination and stimulates recombination-dependent chromosome loss. Mol. Cell. Biol. *16*, 6110–6120.

Chapman, J.R., Taylor, M.R.G., and Boulton, S.J. (2012). Playing the end game: DNA double-strand break repair pathway choice. Mol. Cell *47*, 497–510.

Chatterjee, S., Alampalli, S.V., Nageshan, R.K., Chettiar, S.T., Joshi, S., and Tatu, U.S. (2015). Draft genome of a commonly misdiagnosed multidrug resistant pathogen Candida auris. BMC Genomics *16*, 686.

Chauve, C., Gavranovic, H., Ouangraoua, A., and Tannier, E. (2010). Yeast ancestral genome reconstructions: the possibilities of computational methods II. J. Comput. Biol. J. Comput. Mol. Cell Biol. *17*, 1097–1112.

Chen, J., and Stubbe, J. (2005). Bleomycins: towards better therapeutics. Nat. Rev. Cancer *5*, 102–112.

Chiruvella, K.K., Liang, Z., and Wilson, T.E. (2013). Repair of Double-Strand Breaks by End Joining. Cold Spring Harb. Perspect. Biol. *5*.

Chylinski, K., Makarova, K.S., Charpentier, E., and Koonin, E.V. (2014). Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Res. *42*, 6091–6105.

Ciccia, A., Constantinou, A., and West, S.C. (2003). Identification and characterization of the human mus81-eme1 endonuclease. J. Biol. Chem. *278*, 25172–25178.

de Clare, M., Pir, P., and Oliver, S.G. (2011). Haploinsufficiency and the sex chromosomes from yeasts to humans. BMC Biol. *9*, 15.

Clarke, L., and Carbon, J. (1980). Isolation of a yeast centromere and construction of functional small circular chromosomes. Nature *287*, 504–509.

Coghlan, A., Eichler, E.E., Oliver, S.G., Paterson, A.H., and Stein, L. (2005). Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet. TIG *21*, 673–682.

Cole, F., Keeney, S., and Jasin, M. (2010). Evolutionary conservation of meiotic DSB proteins: more than just Spo11. Genes Dev. *24*, 1201–1207.

Colson, I., Delneri, D., and Oliver, S.G. (2004). Effects of reciprocal chromosomal translocations on the fitness of Saccharomyces cerevisiae. EMBO Rep. *5*, 392–398.

Costantino, L., and Koshland, D. (2015). The Yin and Yang of R-loop Biology. Curr. Opin. Cell Biol. *34*, 39–45.

Crick, F.H.C. (1968). The origin of the genetic code. J. Mol. Biol. *38*, 367–379.

Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A., and Micklem, G. (2015). Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biol. *16*.

Csűrös, M., and Miklós, I. (2009). Streamlining and Large Ancestral Genomes in Archaea Inferred with a Phylogenetic Birth-and-Death Model. Mol. Biol. Evol. *26*, 2087–2095.

Curtin, C.D., Borneman, A.R., Chambers, P.J., and Pretorius, I.S. (2012). De-Novo Assembly and Analysis of the Heterozygous Triploid Genome of the Wine Spoilage Yeast Dekkera bruxellensis AWRI1499. PLOS ONE *7*, e33840.

Dai, J., Xie, W., Brady, T.L., Gao, J., and Voytas, D.F. (2007). Phosphorylation Regulates Integration of the Yeast Ty5 Retrotransposon into Heterochromatin. Mol. Cell *27*, 289–299.

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

Darai-Ramqvist, E., Sandlund, A., Müller, S., Klein, G., Imreh, S., and Kost-Alimova, M. (2008). Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. Genome Res. *18*, 370–379.

Daulny, A., Mejía-Ramírez, E., Reina, O., Rosado-Lugo, J., Aguilar-Arnal, L., Auer, H.,

Zaratiegui, M., and Azorin, F. (2016). The fission yeast CENP-B protein Abp1 prevents pervasive transcription of repetitive DNA elements. Biochim. Biophys. Acta *1859*, 1314–1321.

Davis, M.M., Chien, Y.H., Gascoigne, N.R., and Hedrick, S.M. (1984). A murine T cell receptor gene complex: isolation, structure and rearrangement. Immunol. Rev. *81*, 235–258.

Deem, A., Keszthelyi, A., Blackgrove, T., Vayl, A., Coffey, B., Mathur, R., Chabes, A., and Malkova, A. (2011). Break-induced replication is highly inaccurate. PLoS Biol. *9*, e1000594.

DeFilippis, V., and Villarreal, L.P. (2001). Lateral Gene Transfer or Viral Colonization? Science *293*, 1048–1048.

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

Derr, L.K. (1998). The involvement of cellular recombination and repair genes in RNA-mediated recombination in Saccharomyces cerevisiae. Genetics *148*, 937–945.

DiCarlo, J.E., Norville, J.E., Mali, P., Rios, X., Aach, J., and Church, G.M. (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. *41*, 4336–4343.

Dietrich, F.S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pöhlmann, R., Luedi, P., Choi, S., et al. (2004). The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science *304*, 304–307.

Dietrich, F.S., Voegeli, S., Kuo, S., and Philippsen, P. (2013). Genomes of Ashbya Fungi Isolated from Insects Reveal Four Mating-Type Loci, Numerous Translocations, Lack of Transposons, and Distinct Gene Duplications. G3 Genes Genomes Genet. *3*, 1225–1239.

Dion, V., Kalck, V., Horigome, C., Towbin, B.D., and Gasser, S.M. (2012). Increased mobility of double-strand breaks requires Mec1, Rad9 and the homologous recombination machinery. Nat. Cell Biol. *14*, 502–509.

Dobzhansky, T., and Sturtevant, A.H. (1938). Inversions in the Chromosomes of Drosophila Pseudoobscura. Genetics *23*, 28–64.

Donnianni, R.A., and Symington, L.S. (2013). Break-induced replication occurs by conservative DNA synthesis. Proc. Natl. Acad. Sci. U. S. A. *110*, 13475–13480.

Doudna, J.A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096–1258096.

Downs, J.A., and Jackson, S.P. (2004). A means to a DNA end: the many roles of Ku. Nat. Rev. Mol. Cell Biol. *5*, 367–378.

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

Drillon, G., Carbone, A., and Fischer, G. (2014). SynChro: A Fast and Easy Tool to Reconstruct and Visualize Synteny Blocks along Eukaryotic Chromosomes. PLoS ONE *9*.

Drinnenberg, I.A., Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R., and Bartel, D.P. (2009). RNAi in budding yeast. Science *326*, 544–550.

Drinnenberg, I.A., Fink, G.R., and Bartel, D.P. (2011). Compatibility with Killer explains the Rise of RNAi-deficient fungi. Science *333*, 1592.

Dujon, B. (2006). Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet. TIG *22*, 375–387.

Dujon, B.A., and Louis, E.J. (2017). Genome Diversity and Evolution in the Budding Yeasts (Saccharomycotina). Genetics *206*, 717–750.

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

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

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.

Elias, I., and Tuller, T. (2007). Reconstruction of Ancestral Genomic Sequences Using Likelihood. J. Comput. Biol. *14*, 216–237.

Engels, W.R., and Preston, C.R. (1984). Formation of Chromosome Rearrangements by P Factors in Drosophila. Genetics *107*, 657–678.

Fabre, E., and Zimmer, C. (2018). From dynamic chromatin architecture to DNA damage repair and back. Nucleus *9*, 161–170.

Fairhead, C., Llorente, B., Denis, F., Soler, M., and Dujon, B. (1996). New vectors for combinatorial deletions in yeast chromosomes and for gap-repair cloning using “split-marker” recombination. Yeast Chichester Engl. *12*, 1439–1457.

Faraut, T. (2008). Addressing chromosome evolution in the whole-genome sequence era. Chromosome Res. *16*, 5–16.

Farlow, A., Meduri, E., and Schlötterer, C. (2011). DNA double-strand break repair and the evolution of intron density. Trends Genet. *27*, 1–6.

Feng, G., Leem, Y.-E., and Levin, H.L. (2013). Transposon integration enhances expression of stress response genes. Nucleic Acids Res. *41*,

775–789.

Ferguson, D.O., and Alt, F.W. (2001). DNA double strand break repair and chromosomal translocation: Lessons from animal models. Oncogene *20*, 5572–5579.

Feschotte, C. (2008). The contribution of transposable elements to the evolution of regulatory networks. Nat. Rev. Genet. *9*, 397–405.

Figueiró, H.V., Li, G., Trindade, F.J., Assis, J., Pais, F., Fernandes, G., Santos, S.H.D., Hughes, G.M., Komissarov, A., Antunes, A., et al. (2017). Genome-wide signatures of complex introgression and adaptive evolution in the big cats. Sci. Adv. *3*.

Fink, G.R. (1987). Pseudogenes in yeast? Cell *49*, 5–6.

Finnis, M., Dayan, S., Hobson, L., Chenevix-Trench, G., Friend, K., Ried, K., Venter, D., Woollatt, E., Baker, E., and Richards, R.I. (2005). Common chromosomal fragile site FRA16D mutation in cancer cells. Hum. Mol. Genet. *14*, 1341–1349.

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

Fischer, G., Rocha, E.P.C., Brunet, F., Vergassola, M., and Dujon, B. (2006). Highly Variable Rates of Genome Rearrangements between Hemiascomycetous Yeast Lineages. PLoS Genet. *2*.

Fitch, W.M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Syst. Zool. *20*, 406.

Ford, C.B., Funt, J.M., Abbey, D., Issi, L., Guiducci, C., Martinez, D.A., Delorey, T., Li, B. yu, White, T.C., Cuomo, C., et al. (2011). The evolution of drug resistance in clinical isolates of Candida albicans. ELife *4*.

Fostier, J., Proost, S., Dhoedt, B., Saeys, Y., Demeester, P., Van de Peer, Y., and Vandepoele, K. (2011). A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. Bioinforma. Oxf. Engl. *27*, 749–756.

Fraczek, M.G., Naseeb, S., and Delneri, D. (2018). History of genome editing in yeast. Yeast *35*, 361–368.

Frank, A.C., and Wolfe, K.H. (2009). Evolutionary Capture of Viral and Plasmid DNA by Yeast Nuclear Chromosomes. Eukaryot. Cell *8*, 1521–1531.

Freel, K.C., Sarilar, V., Neuvéglise, C., Devillers, H., Friedrich, A., and Schacherer, J. (2014). Genome Sequence of the Yeast Cyberlindnera fabianii (Hansenula fabianii). Genome Announc. *2*.

Freel, K.C., Friedrich, A., Sarilar, V., Devillers, H., Neuvéglise, C., and Schacherer, J. (2016). Whole-Genome Sequencing and Intraspecific Analysis of the Yeast Species Lachancea quebecensis. Genome Biol. Evol. *8*, 733–741.

Friedrich, A., Jung, P., Reisser, C., Fischer, G., and Schacherer, J. (2015). Population Genomics Reveals Chromosome-Scale Heterogeneous Evolution in a Protoploid Yeast. Mol. Biol. Evol. *32*, 184–192.

Fugmann, S.D. (2010). The origins of the RAG genes – from transposition to V(D)J recombination. Semin. Immunol. *22*, 10–16.

Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K.A., and Makova, K.D. (2012). A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome? Genome Res. *22*, 993–1005.

Gabaldón, T., Martin, T., Marcet-Houben, M., Durrens, P., Bolotin-Fukuhara, M., Lespinet, O., Arnaise, S., Boisnard, S., Aguileta, G., Atanasova, R., et al. (2013). Comparative genomics of emerging pathogens in the Candida glabrata clade. BMC Genomics *14*, 623.

Gabaldón, T., Naranjo-Ortíz, M.A., and Marcet-Houben, M. (2016). Evolutionary genomics of yeast pathogens in the Saccharomycotina. FEMS Yeast Res. *16*.

Gagnon, Y., Blanchette, M., and El-Mabrouk, N. (2012). A flexible ancestral genome reconstruction method based on gapped adjacencies. BMC Bioinformatics *13*, S4.

Galeote, V., Bigey, F., Devillers, H., Neuvéglise, C., and Dequin, S. (2013). Genome Sequence of the Food Spoilage Yeast Zygosaccharomyces bailii CLIB 213T. Genome Announc. *1*.

Garcia, V., Phelps, S.E.L., Gray, S., and Neale, M.J. (2011). Bidirectional resection of DNA double-strand breaks by Mre11 and Exo1. Nature *479*, 241–244.

Gasior, S.L., Wakeman, T.P., Xu, B., and Deininger, P.L. (2006). The human LINE-1 retrotransposon creates DNA double-strand breaks. J. Mol. Biol. *357*, 1383–1393.

Gellert, M. (2002). V(D)J recombination: RAG proteins, repair factors, and regulation. Annu. Rev. Biochem. *71*, 101–132.

Genereux, D.P., and Logsdon, J.M. (2003). Much ado about bacteria-to-vertebrate lateral gene transfer. Trends Genet. *19*, 191–195.

Gilbert, N., Lutz-Prigge, S., and Moran, J.V. (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell *110*, 315–325.

Ginno, P.A., Lim, Y.W., Lott, P.L., Korf, I., and Chédin, F. (2013). GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome Res. *23*, 1590–1600.

Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., Brandt, P., Chakraborty, T., et al. (2001). Comparative genomics of Listeria species. Science *294*, 849–852.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546, 563–567.

Gordon, J.L., Byrne, K.P., and Wolfe, K.H. (2009). Additions, Losses, and Rearrangements on the Evolutionary Route from a Reconstructed Ancestor to the Modern Saccharomyces cerevisiae Genome. PLOS Genet. *5*, e1000485.

Gordon, J.L., Byrne, K.P., and Wolfe, K.H. (2011a). Mechanisms of Chromosome Number Evolution in Yeast. PLoS Genet. *7*.

Gordon, J.L., Armisén, D., Proux-Wéra, E., ÓhÉigeartaigh, S.S., Byrne, K.P., and Wolfe, K.H. (2011b). Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. Proc. Natl. Acad. Sci. *108*, 20024–20029.

Gray, Y.H. (2000). It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet. TIG *16*, 461–468.

Greig, D. (2007). A Screen for Recessive Speciation Genes Expressed in the Gametes of F1 Hybrid Yeast. PLOS Genet. *3*, e21.

Greig, D. (2009). Reproductive isolation in *Saccharomyces*. Heredity *102*, 39–44.

Greig, D., Borts, R.H., Louis, E.J., and Travisano, M. (2002). Epistasis and hybrid sterility in Saccharomyces. Proc. R. Soc. Lond. B Biol. Sci. *269*, 1167–1171.

Gresham, D., Desai, M.M., Tucker, C.M., Jenq, H.T., Pai, D.A., Ward, A., DeSevo, C.G., Botstein, D., and Dunham, M.J. (2008). The Repertoire and Dynamics of Evolutionary Adaptations to Controlled Nutrient-Limited Environments in Yeast. PLOS Genet. *4*, e1000303.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Haber, J.E. (2012). Mating-Type Genes and MAT Switching in Saccharomyces cerevisiae. Genetics *191*, 33–64.

Hahn, P.J. (1993). Molecular biology of double-minute chromosomes. BioEssays News Rev. Mol. Cell. Dev. Biol. *15*, 477–484.

Hall, C., and Dietrich, F.S. (2007). The Reacquisition of Biotin Prototrophy in Saccharomyces cerevisiae Involved Horizontal Gene Transfer, Gene Duplication and Gene Clustering. Genetics *177*, 2293–2307.

Hall, C., Brachat, S., and Dietrich, F.S. (2005). Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae. Eukaryot. Cell *4*, 1102–1115.

Hane, J.K., Rouxel, T., Howlett, B.J., Kema, G.H., Goodwin, S.B., and Oliver, R.P. (2011). A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. Genome Biol. *12*, R45.

Hannenhalli, S., and Pevzner, P.A. (1999). Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. J. ACM *46*, 1–27.

Hanson, S.J., and Wolfe, K.H. (2017). An Evolutionary Perspective on Yeast Mating-Type Switching. Genetics *206*, 9–32.

Hanson, S.J., Byrne, K.P., and Wolfe, K.H. (2014). Mating-type switching by chromosomal inversion in methylotrophic yeasts suggests an origin for the three-locus Saccharomyces cerevisiae system. Proc. Natl. Acad. Sci. U. S. A. *111*, E4851–E4858.

Harewood, L., and Fraser, P. (2014). The impact of chromosomal rearrangements on regulation of gene expression. Hum. Mol. Genet. *23*, R76–R82.

Hastings, P.J., Ira, G., and Lupski, J.R. (2009). A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. PLoS Genet. *5*.

Hazen, K.C. (1995). New and emerging yeast pathogens. Clin. Microbiol. Rev. *8*, 462–478.

Hedges, S.B., Blair, J.E., Venturi, M.L., and Shoe, J.L. (2004). A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol. Biol. *4*, 2.

Helmrich, A., Ballarino, M., and Tora, L. (2011). Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. Mol. Cell *44*, 966–977.

Hickman, A.B., and Dyda, F. (2015). The casposon-encoded Cas1 protein from Aciduliprofundum boonei is a DNA integrase that generates target site duplications. Nucleic Acids Res. *43*, 10576–10587.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. *106*, 9362–9367.

Hirakawa, M.P., Martinez, D.A., Sakthikumar, S., Anderson, M.Z., Berlin, A., Gujja, S., Zeng, Q., Zisson, E., Wang, J.M., Greenberg, J.M., et al. (2015). Genetic and phenotypic intra-species variation in Candida albicans. Genome Res. *25*, 413–425.

Hittinger, C.T., Rokas, A., and Carroll, S.B. (2004). Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. Proc. Natl. Acad. Sci. U. S. A. *101*, 14144–14149.

Hittinger, C.T., Rokas, A., Bai, F.-Y., Boekhout, T., Gonçalves, P., Jeffries, T.W., Kominek, J., Lachance, M.-A., Libkind, D., Rosa, C.A., et al. (2015). For the "Genomes and Evolution" Special Issue of Current Opinion in Genetics and Development. Curr. Opin. Genet. Dev. *35*, 100–109.

Hochrein, L., Mitchell, L.A., Schulz, K., Messerschmidt, K., and Mueller-Roeber, B. (2018). L-SCRaMbLE as a tool for light-controlled Cre-mediated recombination in yeast. Nat. Commun. *9*.

Hoeijmakers, J.H.J. (2009). DNA Damage, Aging,

and Cancer. N. Engl. J. Med. *361*, 1475–1485.

Hooks, K.B., Delneri, D., and Griffiths-Jones, S. (2014). Intron Evolution in Saccharomycetaceae. Genome Biol. Evol. *6*, 2543–2556.

Hou, J., Friedrich, A., de Montigny, J., and Schacherer, J. (2014). Chromosomal Rearrangements as a Major Mechanism in the Onset of Reproductive Isolation in Saccharomyces cerevisiae. Curr. Biol. *24*, 1153–1159.

Hou, J., Sigwalt, A., Fournier, T., Pflieger, D., Peter, J., de Montigny, J., Dunham, M.J., and Schacherer, J. (2016). The Hidden Complexity of Mendelian Traits across Natural Yeast Populations. Cell Rep. *16*, 1106–1114.

Hu, K. (2006). Intron exclusion and the mystery of intron loss. FEBS Lett. *580*, 6361–6365.

Hu, F., Lin, Y., and Tang, J. (2014). MLGO: phylogeny reconstruction and ancestral inference from gene-order data. BMC Bioinformatics *15*.

Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H.A., Zhang, Y., Yu, W., Pontarotti, P., Escriva, H., et al. (2016). Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. Cell *166*, 102–114.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldón, T. (2007). The human phylome. Genome Biol. *8*, R109.

Hunter, N., Chambers, S.R., Louis, E.J., and Borts, R.H. (1996). The mismatch repair system contributes to meiotic sterility in an interspecific yeast hybrid. EMBO J. *15*, 1726–1733.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Ira, G., Pellicioli, A., Balijja, A., Wang, X., Fiorani, S., Carotenuto, W., Liberi, G., Bressan, D., Wan, L., Hollingsworth, N.M., et al. (2004). DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1. Nature *431*, 1011–1017.

Iwasaki, H., Takahagi, M., Shiba, T., Nakata, A., and Shinagawa, H. (1991). Escherichia coli RuvC protein is an endonuclease that resolves the Holliday structure. EMBO J. *10*, 4381–4389.

Jackson, A.P., Gamble, J.A., Yeomans, T., Moran, G.P., Saunders, D., Harris, D., Aslett, M., Barrell, J.F., Butler, G., Citiulo, F., et al. (2009). Comparative genomics of the fungal pathogens Candida dubliniensis and Candida albicans. Genome Res. *19*, 2231–2244.

Jakočiūnas, T., Bonde, I., Herrgård, M., Harrison, S.J., Kristensen, M., Pedersen, L.E., Jensen, M.K., and Keasling, J.D. (2015). Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. Metab. Eng. *28*, 213–222.

Jeffries, T.W., Grigoriev, I.V., Grimwood, J., Laplaza, J.M., Aerts, A., Salamov, A., Schmutz, J., Lindquist, E., Dehal, P., Shapiro, H., et al. (2007). Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis. Nat. Biotechnol. *25*, 319–326.

Jia, B., Wu, Y., Li, B.-Z., Mitchell, L.A., Liu, H., Pan, S., Wang, J., Zhang, H.-R., Jia, N., Li, B., et al. (2018). Precise control of SCRaMbLE in synthetic haploid and diploid yeast. Nat. Commun. *9*, 1933.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature *473*, 97–100.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science *337*, 816–821.

Johnstone, I.L., McCabe, P.C., Greaves, P., Gurr, S.J., Cole, G.E., Brow, M.A., Unkles, S.E., Clutterbuck, A.J., Kinghorn, J.R., and Innis, M.A. (1990). Isolation and characterisation of the crnA-niiA-niaD gene cluster for nitrate assimilation in Aspergillus nidulans. Gene *90*, 181–192.

Jones, B.R., Rajaraman, A., Tannier, E., and Chauve, C. (2012). ANGES: reconstructing ANcestral GEnomeS maps. Bioinforma. Oxf. Engl. *28*, 2388–2390.

Jones, T., Federspiel, N.A., Chibana, H., Dungan, J., Kalman, S., Magee, B.B., Newport, G., Thorstenson, Y.R., Agabian, N., Magee, P.T., et al. (2004). The diploid genome sequence of Candida albicans. Proc. Natl. Acad. Sci. U. S. A. *101*, 7329–7334.

Joron, M., Papa, R., Beltrán, M., Chamberlain, N., Mavárez, J., Baxter, S., Abanto, M., Bermingham, E., Humphray, S.J., Rogers, J., et al. (2006). A Conserved Supergene Locus Controls Colour Pattern Diversity in Heliconius Butterflies. PLOS Biol. *4*, e303.

Joron, M., Frezal, L., Jones, R.T., Chamberlain, N.L., Lee, S.F., Haag, C.R., Whibley, A., Becuwe, M., Baxter, S.W., Ferguson, L., et al. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature *477*, 203–206.

Kachroo, A.H., Laurent, J.M., Yellman, C.M., Meyer, A.G., Wilke, C.O., and Marcotte, E.M. (2015). Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science *348*, 921–925.

Kapitonov, V.V., and Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. Proc. Natl. Acad. Sci. U. S. A. *103*, 4540–4545.

Karathanasis, E., and Wilson, T.E. (2002). Enhancement of Saccharomyces cerevisiae end-joining efficiency by cell growth stage but not by impairment of recombination. Genetics *161*, 1015–1027.

Kawaguchi, Y., Honda, H., Taniguchi-Morimura, J., and Iwasaki, S. (1989). The codon CUG is read as serine in an asporogenic yeast Candida cylindracea. Nature *341*, 164–166.

Keeney, S. (2008). Spo11 and the Formation of DNA Double-Strand Breaks in Meiosis. Genome Dyn. Stab. *2*, 81–123.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241–254.

Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature *428*, 617–624.

Keskin, H., Shen, Y., Huang, F., Patel, M., Yang, T., Ashley, K., Mazin, A.V., and Storici, F. (2014). Transcript RNA-templated DNA recombination and repair. Nature *515*, 436–439.

Kloosterman, W.P., Francioli, L.C., Hormozdiari, F., Marschall, T., Hehir-Kwa, J.Y., Abdellaoui, A., Lameijer, E.-W., Moed, M.H., Koval, V., Renkens, I., et al. (2015). Characteristics of de novo structural changes in the human genome. Genome Res. *25*, 792–801.

Kobayashi, N., Suzuki, Y., Schoenfeld, L.W., Müller, C.A., Nieduszynski, C., Wolfe, K.H., and Tanaka, T.U. (2015). Discovery of an Unconventional Centromere in Budding Yeast Redefines Evolution of Point Centromeres. Curr. Biol. *25*, 2026–2033.

Koç, A., Wheeler, L.J., Mathews, C.K., and Merrill, G.F. (2004). Hydroxyurea Arrests DNA Replication by a Mechanism That Preserves Basal dNTP Pools. J. Biol. Chem. *279*, 223–230.

Kollmar, M., and Mühlhausen, S. (2017). Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. BioEssays *39*, 1600221.

Koster, D.A., Palle, K., Bot, E.S.M., Bjornsti, M.-A., and Dekker, N.H. (2007). Antitumour drugs impede DNA uncoiling by topoisomerase I. Nature *448*, 213–217.

Koszul, R., Caburet, S., Dujon, B., and Fischer, G. (2004). Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. EMBO J. *23*, 234–243.

Krassowski, T., Coughlan, A.Y., Shen, X.-X., Zhou, X., Kominek, J., Opulente, D.A., Riley, R., Grigoriev, I.V., Maheshwari, N., Shields, D.C., et al. (2018). Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. Nat. Commun. *9*, 1887.

Krogh, B.O., and Symington, L.S. (2004). Recombination Proteins in Yeast. Annu. Rev. Genet. *38*, 233–271.

Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D., and Koonin, E.V. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol. *12*, 36.

Küberl, A., Schneider, J., Thallinger, G.G., Anderl, I., Wibberg, D., Hajek, T., Jaenicke, S., Brinkrolf, K., Goesmann, A., Szczepanowski, R., et al. (2011). High-quality genome sequence of Pichia pastoris CBS7435. J. Biotechnol. *154*, 312–320.

Kuhner, M.K., and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. *11*, 459–468.

Kunze, G., Gaillardin, C., Czernicka, M., Durrens, P., Martin, T., Böer, E., Gabaldón, T., Cruz, J.A., Talla, E., Marck, C., et al. (2014). The complete genome of Blastobotrys (Arxula) adeninivorans LS3 - a yeast of biotechnological interest. Biotechnol. Biofuels *7*, 66.

Kurtzman, C.P., Fell, J.W., and Boekhout, T. (2011). The yeasts a taxonomic study. Volume 1 (London; Burlington, Mass.: Elsevier Science).

Lam, I., and Keeney, S. (2015a). Non-paradoxical evolutionary stability of the recombination initiation landscape in yeast. Science *350*, 932–937.

Lam, I., and Keeney, S. (2015b). Mechanism and Regulation of Meiotic Recombination Initiation. Cold Spring Harb. Perspect. Biol. *7*.

Latysheva, N.S., and Babu, M.M. (2016). Discovering and understanding oncogenic gene fusions through data intensive computational approaches. Nucleic Acids Res. *44*, 4487–4503.

Le Tallec, B., Millot, G.A., Blin, M.E., Brison, O., Dutrillaux, B., and Debatisse, M. (2013). Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. Cell Rep. *4*, 420–428.

Le Tallec, B., Koundrioukoff, S., Wilhelm, T., Letessier, A., Brison, O., and Debatisse, M. (2014). Updating the mechanisms of common fragile site instability: how to reconcile the different views? Cell. Mol. Life Sci. *71*, 4489–4494.

Lertwattanasakul, N., Kosaka, T., Hosoyama, A., Suzuki, Y., Rodrussamee, N., Matsutani, M., Murata, M., Fujimoto, N., Suprayogi, Tsuchikane, K., et al. (2015). Genetic basis of the highly efficient yeast Kluyveromyces marxianus: complete genome sequence and transcriptome analyses. Biotechnol. Biofuels *8*.

Lesage, P., and Todeschini, A.L. (2005). Happy together: the life and times of Ty retrotransposons and their hosts. Cytogenet. Genome Res. *110*, 70–90.

Letessier, A., Millot, G.A., Koundrioukoff, S., Lachagès, A.-M., Vogt, N., Hansen, R.S., Malfoy, B., Brison, O., and Debatisse, M. (2011). Cell-type-specific replication initiation programs set fragility of the *FRA3B* fragile site. Nature *470*, 120–123.

Lewis, K.M., and Ke, A. (2017). Building the Class 2 CRISPR-Cas Arsenal. Mol. Cell *65*, 377–379.

Li, F., Dong, J., Pan, X., Oum, J.H., Boeke, J.D., and Lee, S.E. (2008). Microarray-based genetic screen defines SAW1, a new gene required for Rad1/Rad10-dependent processing of recombination intermediates. Mol. Cell *30*, 325–335.

Lieber, M.R., and Karanjawala, Z.E. (2004). Ageing,

repetitive genomes and DNA damage. Nat. Rev. Mol. Cell Biol. *5*, 69–75.

Lieber, M.R., Ma, Y., Pannicke, U., and Schwarz, K. (2004). The mechanism of vertebrate nonhomologous DNA end joining and its role in V(D)J recombination. DNA Repair *3*, 817–826.

Lim, J.K., and Simmons, M.J. (1994). Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. BioEssays News Rev. Mol. Cell. Dev. Biol. *16*, 269–275.

Lin, Y., Dent, S.Y.R., Wilson, J.H., Wells, R.D., and Napierala, M. (2010). R loops stimulate genetic instability of CTG·CAG repeats. Proc. Natl. Acad. Sci. U. S. A. *107*, 692–697.

Lin, Y., Nurk, S., and Pevzner, P.A. (2014). What is the difference between the breakpoint graph and the de Bruijn graph? BMC Genomics *15*, S6.

Ling, J., O'Donoghue, P., and Söll, D. (2015). Genetic code flexibility in microorganisms: novel mechanisms and impact on physiology. Nat. Rev. Microbiol. *13*, 707–721.

Liti, G., Peruffo, A., James, S.A., Roberts, I.N., and Louis, E.J. (2005). Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the Saccharomyces sensu stricto complex. Yeast *22*, 177–192.

Liti, G., Barton, D.B.H., and Louis, E.J. (2006). Sequence Diversity, Reproductive Isolation and Species Concepts in Saccharomyces. Genetics *174*, 839–850.

Liti, G., Haricharan, S., Cubillos, F.A., Tierney, A.L., Sharp, S., Bertuch, A.A., Parts, L., Bailes, E., and Louis, E.J. (2009). Segregating YKU80 and TLC1 Alleles Underlying Natural Variation in Telomere Properties in Wild Yeast. PLoS Genet. *5*.

Liti, G., Ba, A.N.N., Blythe, M., Müller, C.A., Bergström, A., Cubillos, F.A., Dafhnis-Calas, F., Khoshraftar, S., Malla, S., Mehta, N., et al. (2013). High quality de novo sequencing and assembly of the Saccharomyces arboricolus genome. BMC Genomics *14*, 69.

Llorente, B., Malpertuy, A., Neuvéglise, C., de Montigny, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with Saccharomyces cerevisiae. FEBS Lett. *487*, 101–112.

Llorente, B., Smith, C.E., and Symington, L.S. (2008). Break-induced replication: What is it and what is it for? Cell Cycle *7*, 859–864.

Loewith, R., and Hall, M.N. (2011). Target of Rapamycin (TOR) in Nutrient Signaling and Growth Control. Genetics *189*, 1177–1201.

Loidl, J., Jin, Q.W., and Jantsch, M. (1998). Meiotic pairing and segregation of translocation quadrivalents in yeast. Chromosoma *107*, 247–254.

Louis, E.J. (2011). Population genomics and speciation in yeasts. Fungal Biol. Rev. *25*, 136–142.

Louis, V.L., Despons, L., Friedrich, A., Martin, T., Durrens, P., Casarégola, S., Neuvéglise, C., Fairhead, C., Marck, C., Cruz, J.A., et al. (2012). Pichia sorbitophila, an Interspecies Yeast Hybrid, Reveals Early Steps of Genome Resolution After Polyploidization. G3 GenesGenomesGenetics *2*, 299–311.

Luo, J., Sun, X., Cormack, B.P., and Boeke, J.D. (2018a). Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. Nature *560*, 392–396.

Luo, Z., Wang, L., Wang, Y., Zhang, W., Guo, Y., Shen, Y., Jiang, L., Wu, Q., Zhang, C., Cai, Y., et al. (2018b). Identifying and characterizing SCRaMbLEd synthetic yeast using ReSCuES. Nat. Commun. *9*.

Lynch, V.J., Nnamani, M.C., Kapusta, A., Brayer, K., Plaza, S.L., Mazur, E.C., Emera, D., Sheikh, S.Z., Grützner, F., Bauersachs, S., et al. (2015). Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. Cell Rep. *10*, 551–561.

Ma, H., Gutierrez, N.M., Morey, R., Van Dyken, C., Kang, E., Hayama, T., Lee, Y., Li, Y., Tippner-Hedges, R., Wolf, D.P., et al. (2016). Incompatibility Between Nuclear and Mitochondrial Genomes Contributes to Interspecies Reproductive Barrier. Cell Metab. *24*, 283–294.

Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, W.J., Blanchette, M., Haussler, D., and Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. Genome Res. *16*, 1557–1565.

Ma, J., Ratan, A., Raney, B.J., Suh, B.B., Zhang, L., Miller, W., and Haussler, D. (2008). DUPCAR: Reconstructing Contiguous Ancestral Regions with Duplications. J. Comput. Biol. *15*, 1007–1027.

Ma, W., Halweg, C.J., Menendez, D., and Resnick, M.A. (2012). Differential effects of poly(ADP-ribose) polymerase inhibition on DNA break repair in human cells are revealed with Epstein–Barr virus. Proc. Natl. Acad. Sci. U. S. A. *109*, 6590–6595.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986–D992.

Mackay, T.F.C., Stone, E.A., and Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. Nat. Rev. Genet. *10*, 565–577.

Maekawa, H., and Kaneko, Y. (2014). Inversion of the Chromosomal Region between Two Mating Type Loci Switches the Mating Type in Hansenula polymorpha. PLOS Genet. *10*, e1004796.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2013). The basic building blocks and evolution of CRISPR-CAS systems. Biochem. Soc. Trans. *41*, 1392–1400.

Malik, H.S., and Henikoff, S. (2009). Major evolutionary transitions in centromere complexity. Cell *138*, 1067–1082.

Malkova, A., Ivanov, E.L., and Haber, J.E. (1996). Double-strand break repair in the absence of RAD51 in yeast: a possible role for break-induced DNA replication. Proc. Natl. Acad. Sci. U. S. A. *93*, 7131–7136.

Malpertuy, A., Tekaia, F., Casarégola, S., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., de Montigny, J., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. FEBS Lett. *487*, 113–121.

Mani, R.-S., and Chinnaiyan, A.M. (2011). Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. Nat. Rev. Genet. *12*, 150.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Mans, R., van Rossum, H.M., Wijsman, M., Backx, A., Kuijpers, N.G.A., van den Broek, M., Daran-Lapujade, P., Pronk, J.T., van Maris, A.J.A., and Daran, J.-M.G. (2015). CRISPR/Cas9: a molecular Swiss army knife for simultaneous introduction of multiple genetic modifications in Saccharomyces cerevisiae. FEMS Yeast Res. *15*.

Mans, R., Wijsman, M., Daran-Lapujade, P., and Daran, J.-M. (2018). A protocol for introduction of multiple genetic modifications in Saccharomyces cerevisiae using CRISPR/Cas9. FEMS Yeast Res.

Marbouty, M., Cournac, A., Flot, J.-F., Marie-Nelly, H., Mozziconacci, J., and Koszul, R. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. ELife *3*.

Marcet-Houben, M., and Gabaldón, T. (2015). Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. PLOS Biol. *13*, e1002220.

Marie-Nelly, H., Marbouty, M., Cournac, A., Flot, J.-F., Liti, G., Parodi, D.P., Syan, S., Guillén, N., Margeot, A., Zimmer, C., et al. (2014). High-quality genome (re)assembly using chromosomal contact data. Nat. Commun. *5*, 5695.

Marnef, A., Cohen, S., and Legube, G. (2017). Transcription-Coupled DNA Double-Strand Break Repair: Active Genes Need Special Care. J. Mol. Biol. *429*, 1277–1288.

Marsit, S., Mena, A., Bigey, F., Sauvage, F.-X., Couloux, A., Guy, J., Legras, J.-L., Barrio, E., Dequin, S., and Galeote, V. (2015). Evolutionary Advantage Conferred by an Eukaryote-to-Eukaryote Gene Transfer Event in Wine Yeasts. Mol. Biol. Evol. *32*, 1695–1707.

McClintock, B. (1984). The significance of responses of the genome to challenge. Science *226*, 792–801.

Meaburn, K.J., Misteli, T., and Soutoglou, E. (2007). Spatial genome organization in the formation of chromosomal translocations. Semin. Cancer Biol. *17*, 80–90.

Mehta, A., and Haber, J.E. (2014). Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair. Cold Spring Harb. Perspect. Biol. *6*.

Mendoza, O., Bourdoncle, A., Boulé, J.-B., Brosh, R.M., and Mergny, J.-L. (2016). G-quadruplexes and helicases. Nucleic Acids Res. *44*, 1989–2006.

Meraldi, P., McAinsh, A.D., Rheinbay, E., and Sorger, P.K. (2006). Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. Genome Biol. *7*, R23.

Merrikh, H., Zhang, Y., Grossman, A.D., and Wang, J.D. (2012). Replication-transcription conflicts in bacteria. Nat. Rev. Microbiol. *10*, 449–458.

Milligan, J.R., Ng, J.Y., Wu, C.C., Aguilera, J.A., Fahey, R.C., and Ward, J.F. (1995). DNA repair by thiols in air shows two radicals make a double-strand break. Radiat. Res. *143*, 273–280.

Miné-Hattab, J., and Rothstein, R. (2012). Increased chromosome mobility facilitates homology search during recombination. Nat. Cell Biol. *14*, 510–517.

Miranda, I., Silva-Dias, A., Rocha, R., Teixeira-Santos, R., Coelho, C., Gonçalves, T., Santos, M.A.S., Pina-Vaz, C., Solis, N.V., Filler, S.G., et al. (2013). Candida albicans CUG Mistranslation Is a Mechanism To Create Cell Surface Variation. MBio *4*, e00285-13.

Mirkin, E.V., and Mirkin, S.M. (2007). Replication Fork Stalling at Natural Impediments. Microbiol. Mol. Biol. Rev. MMBR *71*, 13–35.

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*.

Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. Nat. Rev. Cancer *7*, 233–245.

Miyake, T., Mae, N., Shiba, T., and Kondo, S. (1987). Production of virus-like particles by the transposable genetic element, copia, of Drosophila melanogaster. Mol. Gen. Genet. MGG *207*, 29–37.

Modesti, M., and Kanaar, R. (2001). DNA repair: spot(light)s on chromatin. Curr. Biol. CB *11*, R229-232.

Modrzewska, B., and Kurnatowski, P. (2013). Selected pathogenic characteristics of fungi from the genus Candida. Ann. Parasitol. *59*, 57–66.

Morales, L., Noel, B., Porcel, B., Marcet-Houben, M., Hullo, M.-F., Sacerdot, C., Tekaia, F., Leh-Louis, V., Despons, L., Khanna, V., et al. (2013). Complete DNA Sequence of Kuraishia capsulata Illustrates Novel Genomic Features among Budding Yeasts

(Saccharomycotina). Genome Biol. Evol. *5*, 2524–2539.

Morel, G., Sterck, L., Swennen, D., Marcet-Houben, M., Onesime, D., Levasseur, A., Jacques, N., Mallet, S., Couloux, A., Labadie, K., et al. (2015). Differential gene retention as an evolutionary mechanism to generate biodiversity and adaptation in yeasts. Sci. Rep. *5*.

Moret, B.M., Wang, L.S., Warnow, T., and Wyman, S.K. (2001a). New approaches for reconstructing phylogenies from gene order data. Bioinforma. Oxf. Engl. *17 Suppl 1*, S165-173.

Moret, B.M., Wyman, S., Bader, D.A., Warnow, T., and Yan, M. (2001b). A new implementation and detailed study of breakpoint analysis. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 583–594.

Mossie, K.G., Young, M.W., and Varmus, H.E. (1985). Extrachromosomal DNA forms of copia-like transposable elements, F elements and middle repetitive DNA sequences in Drosophila melanogaster: Variation in cultured cells and embryos. J. Mol. Biol. *182*, 31–43.

Mourier, T., and Jeffares, D.C. (2003). Eukaryotic Intron Loss. Science *300*, 1393–1393.

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Mühlhausen, S., Findeisen, P., Plessmann, U., Urlaub, H., and Kollmar, M. (2016). A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. Genome Res.

Muller, L.A.H., and McCusker, J.H. (2009). A multi-species based taxonomic microarray reveals interspecies hybridization and introgression in Saccharomyces cerevisiae. FEMS Yeast Res. *9*, 143–152.

Muñoz, A., Zheng, C., Zhu, Q., Albert, V.A., Rounsley, S., and Sankoff, D. (2010). Scaffold filling, contig fusion and comparative gene order inference. BMC Bioinformatics *11*, 304.

Mural, R.J., Adams, M.D., Myers, E.W., Smith, H.O., Miklos, G.L.G., Wides, R., Halpern, A., Li, P.W., Sutton, G.G., Nadeau, J., et al. (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science *296*, 1661–1671.

Muramoto, N., Oda, A., Tanaka, H., Nakamura, T., Kugou, K., Suda, K., Kobayashi, A., Yoneda, S., Ikeuchi, A., Sugimoto, H., et al. (2018). Phenotypic diversification by enhanced genome restructuring after induction of multiple DNA double-strand breaks. Nat. Commun. *9*, 1995.

Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L., et al. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science *309*, 613–617.

Nagy, L.G., Ohm, R.A., Kovács, G.M., Floudas, D., Riley, R., Gácser, A., Sipiczki, M., Davis, J.M., Doty, S.L., Hoog, G.S. de, et al. (2014). Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. Nat. Commun. *5*, 4471.

Naim, V., Wilhelm, T., Debatisse, M., and Rosselli, F. (2013). ERCC1 and MUS81-EME1 promote sister chromatid separation by processing late replication intermediates at common fragile sites during mitosis. Nat. Cell Biol. *15*, 1008–1015.

Naseeb, S., and Delneri, D. (2012). Impact of Chromosomal Inversions on the Yeast DAL Cluster. PLoS ONE *7*, e42022.

Naseeb, S., Carter, Z., Minnis, D., Donaldson, I., Zeef, L., and Delneri, D. (2016). Widespread Impact of Chromosomal Inversions on Gene Expression Uncovers Robustness via Phenotypic Buffering. Mol. Biol. Evol. *33*, 1679–1696.

Neuvéglise, C., Marck, C., and Gaillardin, C. (2011). The intronome of budding yeasts. C. R. Biol. *334*, 662–670.

Nickoloff, J.A., Chen, E.Y., and Heffron, F. (1986). A 24-base-pair DNA sequence from the MAT locus stimulates intergenic recombination in yeast. Proc. Natl. Acad. Sci. U. S. A. *83*, 7831–7835.

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J.-L., Wincker, P., Casaregola, S., et al. (2009). Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. Proc. Natl. Acad. Sci. U. S. A. *106*, 16333–16338.

Ohno, S. (1970). Evolution by Gene Duplication (Berlin, Heidelberg: Springer Berlin Heidelberg).

Orgel, L.E., and Crick, F.H.C. (1980). Selfish DNA: the ultimate parasite. Nature *284*, 604–607.

Osawa, S., and Jukes, T.H. (1989). Codon reassignment (codon capture) in evolution. J. Mol. Evol. *28*, 271–278.

Pâques, F., and Haber, J.E. (1999). Multiple Pathways of Recombination Induced by Double-Strand Breaks in Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev. *63*, 349–404.

Parenteau, J., Durand, M., Véronneau, S., Lacombe, A.-A., Morin, G., Guérin, V., Cecez, B., Gervais-Bird, J., Koh, C.-S., Brunelle, D., et al. (2008). Deletion of Many Yeast Introns Reveals a Minority of Genes that Require Splicing for Function. Mol. Biol. Cell *19*, 1932–1941.

Park, H., Lopez, N.I., and Bakalinsky, A.T. (1999). Use of sulfite resistance in Saccharomyces cerevisiae as a dominant selectable marker. Curr. Genet. *36*, 339–344.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol.

Payen, C., Koszul, R., Dujon, B., and Fischer, G. (2008). Segmental Duplications Arise from Pol32-Dependent Repair of Broken Forks through Two Alternative Replication-Based Mechanisms. PLOS Genet. *4*, e1000175.

Payen, C., Fischer, G., Marck, C., Proux, C., Sherman, D.J., Coppée, J.-Y., Johnston, M., Dujon, B., and Neuvéglise, C. (2009). Unusual composition of a yeast chromosome arm is associated with its delayed replication. Genome Res. *19*, 1710–1721.

Payen, C., Di Rienzi, S.C., Ong, G.T., Pogachar, J.L., Sanchez, J.C., Sunshine, A.B., Raghuraman, M.K., Brewer, B.J., and Dunham, M.J. (2013). The Dynamics of Diverse Segmental Amplifications in Populations of Saccharomyces cerevisiae Adapting to Strong Selection. G3 GenesGenomesGenetics *4*, 399–409.

Pe'er, I., and Shamir, R. (1998). The median problems for breakpoints are NP-complete. Electron. Colloq. Comput. Complex. ECCC *5*.

Peng, Q., Pevzner, P.A., and Tesler, G. (2006). The Fragile Breakage versus Random Breakage Models of Chromosome Evolution. PLoS Comput. Biol. *2*.

Perez-Ortin, J.E. (2002). Molecular Characterization of a Chromosomal Rearrangement Involved in the Adaptive Evolution of Yeast Strains. Genome Res. *12*, 1533–1539.

Perrin, A., Varré, J.-S., Blanquart, S., and Ouangraoua, A. (2015). ProCARs: Progressive Reconstruction of Ancestral Gene Orders. BMC Genomics *16*, S6.

Pevzner, P., and Tesler, G. (2003). Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes. Genome Res. *13*, 37–45.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biol. *9*, e1000602.

Piazza, A., Wright, W.D., and Heyer, W.-D. (2017). Multi-invasions Are Recombination Byproducts that Induce Chromosomal Rearrangements. Cell *170*, 760-773.e15.

Piganeau, M., Ghezraoui, H., De Cian, A., Guittat, L., Tomishima, M., Perrouault, L., Rene, O., Katibah, G.E., Zhang, L., Holmes, M.C., et al. (2013). Cancer translocations in human cells induced by zinc finger and TALE nucleases. Genome Res. *23*, 1182–1193.

Ponting, C.P. (2001). Plagiarized bacterial genes in the human book of life. Trends Genet. *17*, 235–237.

Potter, S.S. (1982). DNA sequence of a foldback transposable element in Drosophila. Nature *297*, 201–204.

Poyatos, J.F., and Hurst, L.D. (2007). The determinants of gene order conservation in yeasts. Genome Biol. *8*, R233.

Prado, F., and Aguilera, A. (2005). Impairment of replication fork progression mediates RNA polII transcription-associated recombination. EMBO J. *24*, 1267–1276.

Priest, S.J., and Lorenz, M.C. (2015). Characterization of Virulence-Related Phenotypes in Candida Species of the CUG Clade. Eukaryot. Cell *14*, 931–940.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neandertal from the Altai Mountains. Nature.

Puchta, H. (2005). The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. J. Exp. Bot. *56*, 1–14.

Puigbò, P., Lobkovsky, A.E., Kristensen, D.M., Wolf, Y.I., and Koonin, E.V. (2014). Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol. *12*, 66.

Quintero-Rivera, F., Xi, Q.J., Keppler-Noreuil, K.M., Lee, J.H., Higgins, A.W., Anchan, R.M., Roberts, A.E., Seong, I.S., Fan, X., Lage, K., et al. (2015). MATR3 disruption in human and mouse associated with bicuspid aortic valve, aortic coarctation and patent ductus arteriosus. Hum. Mol. Genet. *24*, 2375–2389.

Ranz, J.M., Maurin, D., Chan, Y.S., Grotthuss, M. von, Hillier, L.W., Roote, J., Ashburner, M., and Bergman, C.M. (2007). Principles of Genome Evolution in the Drosophila melanogaster Species Group. PLOS Biol. *5*, e152.

Ratmeyer, L., Vinayak, R., Zhong, Y.Y., Zon, G., and Wilson, W.D. (1994). Sequence specific thermodynamic and structural properties for DNA.RNA duplexes. Biochemistry *33*, 5298–5304.

Ravin, N.V., Eldarov, M.A., Kadnikov, V.V., Beletsky, A.V., Schneider, J., Mardanova, E.S., Smekalova, E.M., Zvereva, M.I., Dontsova, O.A., Mardanov, A.V., et al. (2013). Genome sequence and analysis of methylotrophic yeast Hansenula polymorpha DL1. BMC Genomics *14*, 837.

Ray, S.M., Park, S.S., and Ray, A. (1997). Pollen tube guidance by the female gametophyte. Development *124*, 2489–2498.

Riccombeni, A., Vidanes, G., Proux-Wéra, E., Wolfe, K.H., and Butler, G. (2012). Sequence and Analysis of the Genome of the Pathogenic Yeast Candida orthopsilosis. PLOS ONE *7*, e35750.

Richard, G.-F., Viterbo, D., Khanna, V., Mosbach, V., Castelain, L., and Dujon, B. (2014). Highly Specific Contractions of a Single CAG/CTG Trinucleotide Repeat by TALEN in Yeast. PLoS ONE *9*, e95611.

Richardson, C., and Jasin, M. (2000). Frequent chromosomal translocations induced by DNA double-strand breaks. Nature *405*, 697–700.

Riley, R., Haridas, S., Wolfe, K.H., Lopes, M.R., Hittinger, C.T., Göker, M., Salamov, A.A., Wisecaver, J.H., Long, T.M., Calvey, C.H., et al. (2016). Comparative genomics of biotechnologically important yeasts. Proc. Natl. Acad. Sci. U. S. A. *113*, 9882–9887.

Rocha, E.P.C., Cornet, E., and Michel, B. (2005). Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems. PLoS Genet. *1*.

Rolland, T., and Dujon, B. (2011). Yeasty clocks: Dating genomic changes in yeasts. C. R. Biol. *334*, 620–628.

Rothkamm, K., Krüger, I., Thompson, L.H., and Löbrich, M. (2003). Pathways of DNA Double-Strand Break Repair during the Mammalian Cell Cycle. Mol. Cell. Biol. *23*, 5706–5715.

Rowley, J.D. (1998). The critical role of chromosome translocations in human leukemias. Annu. Rev. Genet. *32*, 495–519.

Roy, S.W. (2006). Intron-rich ancestors. Trends Genet. TIG *22*, 468–471.

Roy, K.R., Smith, J.D., Vonesch, S.C., Lin, G., Tu, C.S., Lederer, A.R., Chu, A., Suresh, S., Nguyen, M., Horecka, J., et al. (2018). Multiplexed precision genome editing with trackable genomic barcodes in yeast. Nat. Biotechnol.

Ruiz-Herrera, A., García, F., Giulotto, E., Attolini, C., Egozcue, J., Ponsà, M., and Garcia, M. (2005). Evolutionary breakpoints are co-localized with fragile sites and intrachromosomal telomeric sequences in primates. Cytogenet. Genome Res. *108*, 234–247.

Rupp, O., Brinkrolf, K., Buerth, C., Kunigo, M., Schneider, J., Jaenicke, S., Goesmann, A., Pühler, A., Jaeger, K.-E., and Ernst, J.F. (2015). The structure of the Cyberlindnera jadinii genome and its relation to Candida utilis analyzed by the occurrence of single nucleotide polymorphisms. J. Biotechnol. *211*, 20–30.

Sacerdot, C., Louis, A., Bon, C., Berthelot, C., and Roest Crollius, H. (2018). Chromosome evolution at the origin of the ancestral vertebrate genome. Genome Biol. *19*, 166.

Sadhu, M.J., Bloom, J.S., Day, L., Siegel, J.J., Kosuri, S., and Kruglyak, L. (2017). Highly parallel genome variant engineering with CRISPR/Cas9 in eukaryotic cells.

Sadhu, M.J., Bloom, J.S., Day, L., Siegel, J.J., Kosuri, S., and Kruglyak, L. (2018). Highly parallel genome variant engineering with CRISPR–Cas9. Nat. Genet. *50*, 510–514.

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.

Salzberg, S.L. (2017). Horizontal gene transfer is not a hallmark of the human genome. Genome Biol. *18*.

Salzberg, S.L., White, O., Peterson, J., and Eisen, J.A. (2001). Microbial genes in the human genome: lateral transfer or gene loss? Science *292*, 1903–1906.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular cloning: a laboratory manual. (Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press).

Sandmeyer, S. (2003). Integration by design. Proc. Natl. Acad. Sci. *100*, 5586–5588.

Sandmeyer, S.B., Aye, M., and Menees, T. (2002). Ty3, a Position-Specific, Gypsy-Like Element in Saccharomyces cerevisiae. Mob. DNA II 663–683.

Sankoff, D. (2009). Reconstructing the History of Yeast Genomes. PLOS Genet. *5*, e1000483.

Sankoff, D., and Trinh, P. (2005). Chromosomal breakpoint reuse in genome sequence rearrangement. J. Comput. Biol. J. Comput. Mol. Cell Biol. *12*, 812–821.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., et al. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science *274*, 765–768.

Santos, M.A.S., Gomes, A.C., Santos, M.C., Carreto, L.C., and Moura, G.R. (2011). The genetic code of the fungal CTG clade. C. R. Biol. *334*, 607–611.

Santos-Pereira, J.M., and Aguilera, A. (2015). R loops: new modulators of genome dynamics and function. Nat. Rev. Genet. *16*, 583.

Sanz, L.A., Hartono, S.R., Lim, Y.W., Steyaert, S., Rajpurkar, A., Ginno, P.A., Xu, X., and Chédin, F. (2016). Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. Mol. Cell *63*, 167–178.

Sargent, R.G., Brenneman, M.A., and Wilson, J.H. (1997). Repair of site-specific double-strand breaks in a mammalian chromosome by homologous and illegitimate recombination. Mol. Cell. Biol. *17*, 267–277.

Sarilar, V., Devillers, H., Freel, K.C., Schacherer, J., and Neuvéglise, C. (2015). Draft Genome Sequence of Lachancea lanzarotensis CBS 12615T, an Ascomycetous Yeast Isolated from Grapes. Genome Announc. *3*.

Sarni, D., and Kerem, B. (2016). The complex nature of fragile site plasticity and its importance in cancer. Curr. Opin. Cell Biol. *40*, 131–136.

Sasano, Y., Nagasawa, K., Kaboli, S., Sugiyama, M., and Harashima, S. (2016). CRISPR-PCS: a powerful new approach to inducing multiple chromosome splitting in Saccharomyces cerevisiae. Sci. Rep. *6*.

Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., and Kazazian, H.H. (1997). Many human L1 elements are capable of retrotransposition. Nat. Genet. *16*, 37–43.

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S., and Wolfe, K.H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature *440*, 341–345.

Scannell, D.R., Butler, G., and Wolfe, K.H. (2007a). Yeast genome evolution--the origin of the species.

Yeast Chichester Engl. *24*, 929–942.

Scannell, D.R., Frank, A.C., Conant, G.C., Byrne, K.P., Woolfit, M., and Wolfe, K.H. (2007b). Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc. Natl. Acad. Sci. *104*, 8397–8402.

Scannell, D.R., Zill, O.A., Rokas, A., Payen, C., Dunham, M.J., Eisen, M.B., Rine, J., Johnston, M., and Hittinger, C.T. (2011). The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 GenesGenomesGenetics *1*, 11–25.

Schacherer, J., Tourrette, Y., Souciet, J.-L., Potier, S., and de Montigny, J. (2004). Recovery of a Function Involving Gene Duplication by Retroposition in Saccharomyces cerevisiae. Genome Res. *14*, 1291–1297.

Schneider, J., Andrea, H., Blom, J., Jaenicke, S., Rückert, C., Schorsch, C., Szczepanowski, R., Farwick, M., Goesmann, A., Pühler, A., et al. (2012). Draft Genome Sequence of Wickerhamomyces ciferrii NRRL Y-1031 F-60-10. Eukaryot. Cell *11*, 1582–1583.

Schultz, D.W., and Yarus, M. (1994). Transfer RNA Mutation and the Malleability of the Genetic Code. J. Mol. Biol. *235*, 1377–1380.

Seoighe, C., Federspiel, N., Jones, T., Hansen, N., Bivolarovic, V., Surzycki, R., Tamse, R., Komp, C., Huizar, L., Davis, R.W., et al. (2000). Prevalence of small inversions in yeast gene order evolution. Proc. Natl. Acad. Sci. U. S. A. *97*, 14433–14437.

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.

Shao, Y., Lu, N., Wu, Z., Cai, C., Wang, S., Zhang, L.-L., Zhou, F., Xiao, S., Liu, L., Zeng, X., et al. (2018). Creating a functional single-chromosome yeast. Nature *560*, 331–335.

Sharon, E., Chen, S.-A.A., Khosla, N.M., Smith, J.D., Pritchard, J.K., and Fraser, H.B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. Cell *175*, 544-557.e16.

Shen, M.J., Wu, Y., Yang, K., Li, Y., Xu, H., Zhang, H., Li, B.-Z., Li, X., Xiao, W.-H., Zhou, X., et al. (2018a). Heterozygous diploid and interspecies SCRaMbLEing. Nat. Commun. *9*, 1934.

Shen, X.-X., Zhou, X., Kominek, J., Kurtzman, C.P., Hittinger, C.T., and Rokas, A. (2016a). Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. G3 GenesGenomesGenetics *6*, 3927–3939.

Shen, X.-X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B., Wisecaver, J.H., Wang, M., Doering, D.T., et al. (2018b). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell *0*.

Shen, Y., Stracquadanio, G., Wang, Y., Yang, K., Mitchell, L.A., Xue, Y., Cai, Y., Chen, T., Dymond, J.S., Kang, K., et al. (2016b). SCRaMbLE generates designed combinatorial stochastic diversity in synthetic chromosomes. Genome Res. *26*, 36–49.

Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A.M., and Fire, A.Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse-transcriptase-Cas1 fusion protein. Science *351*, aad4234.

Simillion, C., Janssens, K., Sterck, L., and Van de Peer, Y. (2008). i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. Bioinforma. Oxf. Engl. *24*, 127–128.

Slot, J.C., and Rokas, A. (2010). Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. Proc. Natl. Acad. Sci. U. S. A. *107*, 10136–10141.

Smith, C.E., Llorente, B., and Symington, L.S. (2007). Template switching during break-induced replication. Nature *447*, 102–105.

Snel, B., Bork, P., and Huynen, M.A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. *12*, 17–25.

Sniegowski, P.D., Dombrowski, P.G., and Fingerman, E. (2002). Saccharomyces cerevisiae and Saccharomyces paradoxus coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. FEMS Yeast Res. *1*, 299–306.

Sollier, J., and Cimprich, K.A. (2015). R-Loops Breaking Bad. Trends Cell Biol. *25*, 514–522.

Souciet, J.-L., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., de Montigny, J., Dujon, B., et al. (2000). Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies11Sequences and annotations are accessible at: Génoscope (http://www.genoscope.cns.fr), FEBS Letters Website (http://www.elsevier.nl/febs/show/), Bordeaux (http://cbi.genopole-bordeaux.fr/Genolevures) and were deposited into the EMBL database (accession number from AL392203 to AL441602). FEBS Lett. *487*, 3–12.

Souciet, J.-L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P.V., Cliften, P., Sherman, D.J., Weissenbach, J., Westhof, E., Wincker, P., et al. (2009). Comparative genomics of protoploid Saccharomycetaceae. Genome Res. *19*, 1696–1709.

Soulas-Sprauel, P., Rivera-Munoz, P., Malivert, L., Guyader, G.L., Abramowski, V., Revy, P., and Villartay, J.-P. de (2007). V(D)J and immunoglobulin class switch recombinations: a paradigm to study the regulation of DNA end-joining. Oncogene *26*, 7780–7791.

Spingola, M., Grate, L., Haussler, D., and Ares, M. (1999). Genome-wide bioinformatic and molecular

analysis of introns in Saccharomyces cerevisiae. RNA N. Y. N *5*, 221–234.

Stajich, J.E., Dietrich, F.S., and Roy, S.W. (2007). Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome Biol. *8*, R223.

Stanhope, M.J., Lupas, A., Italia, M.J., Koretke, K.K., Volker, C., and Brown, J.R. (2001). Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature *411*, 940–944.

Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. *61*, 437–455.

Storici, F., and Resnick, M.A. (2003). Delitto perfetto targeted mutagenesis in yeast with oligonucleotides. Genet. Eng. (N. Y.) *25*, 189–207.

Storici, F., and Resnick, M.A. (2006). The Delitto Perfetto Approach to In Vivo Site-Directed Mutagenesis and Chromosome Rearrangements with Synthetic Oligonucleotides in Yeast. In Methods in Enzymology, (Elsevier), pp. 329–345.

Sudbery, P.E. (2011). Growth of Candida albicans hyphae. Nat. Rev. Microbiol. *9*, 737–748.

Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., et al. (2015). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761.

Sugawara, N., Pâques, F., Colaiácovo, M., and Haber, J.E. (1997). Role of Saccharomyces cerevisiae Msh2 and Msh3 repair proteins in double-strand break-induced recombination. Proc. Natl. Acad. Sci. U. S. A. *94*, 9214–9219.

Syeda, A.H., Hawkins, M., and McGlynn, P. (2014). Recombination and Replication. Cold Spring Harb. Perspect. Biol. *6*.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327–338.

Symington, L.S. (2016). Mechanism and Regulation of DNA End Resection in Eukaryotes. Crit. Rev. Biochem. Mol. Biol. *51*, 195–212.

Thomas, J., and Pritham, E.J. (2015). Helitrons, the Eukaryotic Rolling-circle Transposable Elements. Microbiol. Spectr. *3*.

Thompson, L.H. (2012). Recognition, signaling, and repair of DNA double-strand breaks produced by ionizing radiation in mammalian cells: The molecular choreography. Mutat. Res. Mutat. Res. *751*, 158–246.

Thompson, D.S., Carlisle, P.L., and Kadosh, D. (2011). Coevolution of morphology and virulence in Candida species. Eukaryot. Cell *10*, 1173–1182.

Tobler, R., Franssen, S.U., Kofler, R., Orozco-terWengel, P., Nolte, V., Hermisson, J., and Schlötterer, C. (2014). Massive Habitat-Specific Genomic Response in D. melanogaster Populations during Experimental Evolution in Hot and Cold Environments. Mol. Biol. Evol. *31*, 364–375.

Toh, G.W.-L., Sugawara, N., Dong, J., Toth, R., Lee, S.E., Haber, J.E., and Rouse, J. (2010). Mec1/Tel1-dependent phosphorylation of Slx4 stimulates Rad1-Rad10-dependent cleavage of non-homologous DNA tails. DNA Repair *9*, 718–726.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Török, T., Rockhold, D., and King, A.D. (1993). Use of electrophoretic karyotyping and DNA-DNA hybridization in yeast identification. Int. J. Food Microbiol. *19*, 63–80.

Torres, J.Z., Bessler, J.B., and Zakian, V.A. (2004). Local chromatin structure at the ribosomal DNA causes replication fork pausing and genome instability in the absence of the *S.* cerevisiae DNA helicase Rrm3p. Genes Dev. *18*, 498–503.

Trinh, P., McLysaght, A., and Sankoff, D. (2004). Genomic features in the breakpoint regions between syntenic blocks. Bioinforma. Oxf. Engl. *20 Suppl 1*, i318-325.

Tuduri, S., Crabbé, L., Conti, C., Tourrière, H., Holtgreve-Grez, H., Jauch, A., Pantesco, V., De Vos, J., Thomas, A., Theillet, C., et al. (2009). Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. Nat. Cell Biol. *11*, 1315–1324.

Turner, S.A., and Butler, G. (2014). The Candida Pathogenic Species Complex. Cold Spring Harb. Perspect. Med. *4*.

Umezu, K., Hiraoka, M., Mori, M., and Maki, H. (2002). Structural analysis of aberrant chromosomes that occur spontaneously in diploid Saccharomyces cerevisiae: retrotransposon Ty1 plays a crucial role in chromosomal rearrangements. Genetics *160*, 97–110.

Vakirlis, N., Sarilar, V., Drillon, G., Fleiss, A., Agier, N., Meyniel, J.-P., Blanpain, L., Carbone, A., Devillers, H., Dubois, K., et al. (2016). Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. Genome Res. *26*, 918–932.

Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., and Lafontaine, I. (2018). A Molecular Portrait of De Novo Genes in Yeasts. Mol. Biol. Evol. *35*, 631–645.

VanHulle, K., Lemoine, F.J., Narayanan, V., Downing, B., Hull, K., McCullough, C., Bellinger, M., Lobachev, K., Petes, T.D., and Malkova, A. (2007). Inverted DNA Repeats Channel Repair of Distant Double-Strand Breaks into Chromatid Fusions and Chromosomal Rearrangements. Mol. Cell. Biol. *27*, 2601–2614.

Vanoli, F., Tomishima, M., Feng, W., Lamribet, K., Babin, L., Brunet, E., and Jasin, M. (2017). CRISPR-Cas9–guided oncogenic chromosomal translocations with conditional fusion protein expression in human mesenchymal cells. Proc. Natl.

Acad. Sci. *114*, 3696–3701.

Varmus, H. (1988). Retroviruses. Science *240*, 1427–1435.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates., EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. Genome Res. *19*, *19*, 327, 327–335.

Villarreal, L.P., and DeFilippis, V.R. (2000). A Hypothesis for DNA Viruses as the Origin of Eukaryotic Replication Proteins. J. Virol. *74*, 7079–7084.

Viollet, S., Monot, C., and Cristofari, G. (2014). L1 retrotransposition. Mob. Genet. Elem. *4*.

Wade, N. (2001). Link Between Human Genes and Bacteria Is Hotly Debated by Rival Scientific Camps. N. Y. Times.

Wahba, L., Amon, J.D., Koshland, D., and Vuica-Ross, M. (2011). RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. Mol. Cell *44*, 978–988.

Wang, F., and Qi, L.S. (2016). Applications of CRISPR Genome Engineering in Cell Biology. Trends Cell Biol. *26*, 875–888.

Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K., and Haussler, D. (2007). Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc. Natl. Acad. Sci. U. S. A. *104*, 18613–18618.

Ward, J.F. (1994). The complexity of DNA damage: relevance to biological consequences. Int. J. Radiat. Biol. *66*, 427–432.

Wei, X., and Zhang, J. (2017). The Genomic Architecture of Interactions Between Natural Genetic Polymorphisms and Environments in Yeast Growth. Genetics *205*, 925–937.

Weinfeld, M., and Soderlind, K.J. (1991). 32P-postlabeling detection of radiation-induced DNA damage: identification and estimation of thymine glycols and phosphoglycolate termini. Biochemistry *30*, 1091–1097.

Wellinger, R.E., Prado, F., and Aguilera, A. (2006). Replication Fork Progression Is Impaired by Transcription in Hyperrecombinant Yeast Cells Lacking a Functional THO Complex. Mol. Cell. Biol. *26*, 3327–3334.

Wendland, J., and Walther, A. (2011). Genome Evolution in the Eremothecium Clade of the Saccharomyces Complex Revealed by Comparative Genomics. G3 GenesGenomesGenetics *1*, 539–548.

Wenger, J.W., Schwartz, K., and Sherlock, G. (2010). Bulk Segregant Analysis by High-Throughput Sequencing Reveals a Novel Xylose Utilization Gene from Saccharomyces cerevisiae. PLoS Genet. *6*.

Westover, K.D., Bushnell, D.A., and Kornberg, R.D. (2004). Structural basis of transcription: separation of RNA from DNA by RNA polymerase II. Science *303*, 1014–1016.

Wohlbach, D.J., Kuo, A., Sato, T.K., Potts, K.M., Salamov, A.A., LaButti, K.M., Sun, H., Clum, A., Pangilinan, J.L., Lindquist, E.A., et al. (2011). Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. Proc. Natl. Acad. Sci. U. S. A. *108*, 13212–13217.

Wolfe, K.H. (2015). Origin of the Yeast Whole-Genome Duplication. PLOS Biol. *13*, e1002221.

Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature *387*, 708–713.

Wolfe, K.H., Armisén, D., Proux-Wera, E., ÓhÉigeartaigh, S.S., Azam, H., Gordon, J.L., and Byrne, K.P. (2015). Clade- and species-specific features of genome evolution in the Saccharomycetaceae. FEMS Yeast Res. *15*.

Wright, W.D., Shah, S.S., and Heyer, W.-D. (2018). Homologous recombination and the repair of DNA double-strand breaks. J. Biol. Chem. *293*, 10524–10535.

Wyatt, H.D.M., and West, S.C. (2014). Holliday Junction Resolvases. Cold Spring Harb. Perspect. Biol. *6*.

Wyrobek, A.J., Schmid, T.E., and Marchetti, F. (2005). Relative Susceptibilities of Male Germ Cells to Genetic Defects Induced by Cancer Chemotherapies. JNCI Monogr. *2005*, 31–35.

Xiao, A., Wang, Z., Hu, Y., Wu, Y., Luo, Z., Yang, Z., Zu, Y., Li, W., Huang, P., Tong, X., et al. (2013). Chromosomal deletions and inversions mediated by TALENs and CRISPR/Cas in zebrafish. Nucleic Acids Res. *41*, e141–e141.

Xiong, Y., and Eickbush, T.H. (1988). Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol. Biol. Evol. *5*, 675–690.

Xu, A.W., and Moret, B.M.E. (2011). GASTS: Parsimony Scoring under Rearrangements. In Algorithms in Bioinformatics, T.M. Przytycka, and M.-F. Sagot, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 351–363.

Yancopoulos, S., and Friedberg, R. (2009). DCJ Path Formulation for Genome Transformations which Include Insertions, Deletions, and Duplications. J. Comput. Biol. *16*, 1311–1338.

Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinforma. Oxf. Engl. *21*, 3340–3346.

Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat. Rev. Genet. *13*, 303–314.

Yang, N., Hu, F., Zhou, L., and Tang, J. (2014). Reconstruction of Ancestral Gene Orders Using

Probabilistic and Gene Encoding Approaches. PLOS ONE *9*, e108796.

Yu, X., Jacobs, S.A., West, S.C., Ogawa, T., and Egelman, E.H. (2001). Domain structure and dynamics in the helical filaments formed by RecA and Rad51 on DNA. Proc. Natl. Acad. Sci. U. S. A. *98*, 8419–8424.

Yue, J.-X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., Bergström, A., Coupland, P., Warringer, J., Lagomarsino, M.C., et al. (2017). Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat. Genet. *49*, 913–924.

Zeman, M.K., and Cimprich, K.A. (2014). Causes and consequences of replication stress. Nat. Cell Biol. *16*, 2–9.

Zhang, H., and Freudenreich, C.H. (2007). An AT-rich Sequence in Human Common Fragile Site FRA16D Causes Fork Stalling and Chromosome Breakage in S. cerevisiae. Mol. Cell *27*, 367–379.

Zhang, J., and Peterson, T. (2004). Transposition of reversed Ac element ends generates chromosome rearrangements in maize. Genetics *167*, 1929–1937.

Zheng, C., and Sankoff, D. (2011). On the PATHGROUPS approach to rapid small phylogeny. BMC Bioinformatics *12*, S4.

Zuckerkandl, E., and Pauling, L.B. (1962). Molecular disease, evolution, and genetic heterogeneity. In Horizons in Biochemistry, M. Kasha, and B. Pullman, eds. (New York: Academic Press), pp. 189–225.

Zufall, R.A., Robinson, T., and Katz, L.A. (2005). Evolution of developmentally regulated genome rearrangements in eukaryotes. J. Exp. Zoolog. B Mol. Dev. Evol. *304B*, 448–455.

(2002). Mobile DNA II (American Society of Microbiology).

ANNEXES

Article: Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus (Vakirlis et al., 2016)

Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus

Nikolaos Vakirlis,[1,6] Véronique Sarilar,[2,6] Guénola Drillon,[1,6] Aubin Fleiss,[1] Nicolas Agier,[1] Jean-Philippe Meyniel,[3] Lou Blanpain,[2] Alessandra Carbone,[1] Hugo Devillers,[2] Kenny Dubois,[4] Alexandre Gillet-Markowska,[1] Stéphane Graziani,[3] Nguyen Huu-Vang,[2] Marion Poirel,[3] Cyrielle Reisser,[5] Jonathan Schott,[4] Joseph Schacherer,[5] Ingrid Lafontaine,[1] Bertrand Llorente,[4] Cécile Neuvéglise,[2] and Gilles Fischer[1]

[1]*Sorbonne Universités, UPMC Univ. Paris 06, CNRS, Institut de Biologie Paris-Seine, Laboratory of Computational and Quantitative Biology, F-75005, Paris, France;* [2]*Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France;* [3]*ISoft, Route de l'Orme, Parc "Les Algorithmes" Bâtiment Euclide, 91190 Saint-Aubin, France;* [4]*CRCM, CNRS, UMR7258, Inserm, U1068; Institut Paoli-Calmettes, Aix-Marseille Université, UM 105, F-13009, Marseille, France;* [5]*Department of Genetics, Genomics and Microbiology, University of Strasbourg/CNRS, UMR 7156, 67083 Strasbourg, France*

Reconstructing genome history is complex but necessary to reveal quantitative principles governing genome evolution. Such reconstruction requires recapitulating into a single evolutionary framework the evolution of genome architecture and gene repertoire. Here, we reconstructed the genome history of the genus *Lachancea* that appeared to cover a continuous evolutionary range from closely related to more diverged yeast species. Our approach integrated the generation of a high-quality genome data set; the development of *AnChro*, a new algorithm for reconstructing ancestral genome architecture; and a comprehensive analysis of gene repertoire evolution. We found that the ancestral genome of the genus *Lachancea* contained eight chromosomes and about 5173 protein-coding genes. Moreover, we characterized 24 horizontal gene transfers and 159 putative gene creation events that punctuated species diversification. We retraced all chromosomal rearrangements, including gene losses, gene duplications, chromosomal inversions and translocations at single gene resolution. Gene duplications outnumbered losses and balanced rearrangements with 1503, 929, and 423 events, respectively. Gene content variations between extant species are mainly driven by differential gene losses, while gene duplications remained globally constant in all lineages. Remarkably, we discovered that balanced chromosomal rearrangements could be responsible for up to 14% of all gene losses by disrupting genes at their breakpoints. Finally, we found that nonsynonymous substitutions reached fixation at a coordinated pace with chromosomal inversions, translocations, and duplications, but not deletions. Overall, we provide a granular view of genome evolution within an entire eukaryotic genus, linking gene content, chromosome rearrangements, and protein divergence into a single evolutionary framework.

[Supplemental material is available for this article.]

Eukaryotic genomes evolve through the accumulation of point mutations and chromosomal rearrangements that ultimately contribute to the evolution of the gene repertoire. Point mutations can promote gene inactivation by pseudogenization of coding sequences (Mighell et al. 2000; Lafontaine and Dujon 2010) but also participate in gene gain by de novo gene creation from noncoding sequences (Khalturin et al. 2009; McLysaght and Guerzoni 2015). Balanced rearrangements—including translocations, inversions, and chromosome fusion/fission—modify gene order and orientation. Although these rearrangements are often thought to occur mostly in intergenic regions (Peng et al. 2006; Poyatos and Hurst 2007; Berthelot et al. 2015), they have the potential to modify gene expression, create new gene combinations, and disrupt genes at their breakpoints (Rowley 1998; Perez-Ortin et al. 2002; Avelar et al. 2013; Quintero-Rivera et al. 2015). Unbalanced chromosomal rearrangements include deletions and duplications of the chromosome segments, which promote reduction and expansion of the gene repertoire, respectively (Llorente et al. 2000; Dujon et al. 2004; Wapinski et al. 2007; Butler et al. 2009; Souciet et al. 2009; Scannell et al. 2011; Gabaldon et al. 2013). Whole-genome duplication (WGD) and hybridization events also affect gene repertoire, as well-documented in yeasts (Semon and Wolfe 2007; Louis et al.

[6]**These authors equally contributed to this work.**
Corresponding authors: bertrand.llorente@inserm.fr, ncecile@grignon.inra.fr, gilles.fischer@upmc.fr
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.204420.116.

© 2016 Vakirlis et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

26:918–932 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/16; www.genome.org

2012; Morales and Dujon 2012; Marcet-Houben and Gabaldon 2015). The impact of horizontal gene transfers (HGTs), although seemingly important in Pezizomycotina, is limited in Saccharomycotina, with only a few dozen reported events so far (Rolland et al. 2009; Galeote et al. 2010; Marcet-Houben and Gabaldon 2010; Wisecaver et al. 2014; Marsit et al. 2015).

Comparative genomics has been instrumental in identifying these mechanisms and deciphering their contribution to genome evolution. Notably, the study of synteny conservation across multiple species allowed critical conceptual advances in the understanding of genome dynamics. Comparative studies on synteny conservation revealed highly variable rates of chromosome rearrangements between individual lineages both in vertebrates and in yeasts (Bourque et al. 2005; Fischer et al. 2006). Interestingly, a comparative study between 12 *Drosophila* genomes reported that the disruption of synteny regions via chromosomal inversions approximated a linear process over time (Bhutkar et al. 2008). At a broader evolutionary scale, reconstructions of ancestral gene content in Proteobacteria and Archaea showed that gene losses and/or duplications correlated with amino acid substitution rates (Snel et al. 2002; Csuros and Miklos 2009). Similarly, linear correlations between the rates of genomic rearrangements such as gene duplications and losses, HGTs and gene creations, and the rate of nonsynonymous substitutions were recently reported in bacteria (Puigbo et al. 2014). By analogy to the traditional molecular clock (Zuckerkandl and Pauling 1962), the investigators coined the term "genomic clock" to describe the coordinated pace of fixation between point mutations and large-scale rearrangements. The first attempt to define a genomic clock in yeast was based on the correlation between synteny conservation and amino acid identity between orthologous genes (Rolland and Dujon 2011).

Reconstructing genome history is a rather difficult task requiring efficient reconstruction of ancestral genome organization and precise characterization of the chromosomal rearrangements that occurred along different lineages. Reconstruction of ancestral genome architecture has benefited from the development of several computational models (Ma et al. 2006; Faraut 2008; Muffato and Roest Crollius 2008; Alekseyev and Pevzner 2009; Ouangraoua et al. 2011; Gagnon et al. 2012). However, integrating the reconstruction of ancestral genome architecture into an evolutionary framework has only been achieved in a limited number of cases. In yeast, Wolfe and colleagues manually reconstructed the ancestral genome of *Saccharomyces cerevisiae* as it was before the WGD event and identified at least 144 structural rearrangements, as well as 124 genes that are present in the actual *S. cerevisiae* genome but absent from its ancestor (Gordon et al. 2009). These investigators also traced the complete rearrangement history of the *Lachancea kluyveri* genome since its common ancestor with *S. cerevisiae* (Gordon et al. 2011). In vertebrates, a recent study used ancestral genome reconstruction to explain the distribution pattern of rearrangement breakpoints in Boreoeutherian genomes (Berthelot et al. 2015). The investigators found a strong positive correlation between gene density and evolutionary rearrangement breakpoints and showed that this property could be extended to yeast genomes. Finally, Weng et al. (2014) recently reconstructed the ancestral genome organization of highly rearranged *Geraniaceae* plastid genomes and characterized the rearrangements unique to each genus. They found that the degree of plastid genome rearrangements was correlated with nonsynonymous substitution rates but not with synonymous substitution rates, compatible with the existence of a genomic clock in plastid genomes.

Based on genome comparison between three previously sequenced *Lachancea* species, we predicted that the number of rearrangements that reached fixation in this genus was sufficiently high, but not too high, to provide key information on the dynamics of chromosome evolution (Fischer et al. 2006; Payen et al. 2009; Souciet et al. 2009; Drillon and Fischer 2011; Gordon et al. 2011). Therefore, we undertook the reconstruction of genome history in this genus to seek for quantitative rules that govern the evolution of genomes. First, we sequenced, assembled, and annotated the genomes of seven additional *Lachancea* species. With 10 fully sequenced, assembled, and annotated genomes, the *Lachancea* clade is the most densely sampled yeast genus at the genomic level within the Saccharomycetaceae family. Second, we developed a new computational method called *AnChro*, to reconstruct ancestral genome organization and identify all balanced rearrangements that accumulated during evolution. We combined these reconstructions with an exhaustive survey of the gene repertoire and revealed general principles that govern genome evolution in this yeast genus.

Results

High-quality reference genomes of the *Lachancea* genus

We sequenced and assembled into one scaffold per chromosome the nuclear genomes of seven *Lachancea* species (see Methods). Haploid genome sizes range from 10.2 to 11.3 Mb (Fig. 1A). All *Lachancea* species have eight chromosomes, each containing one centromere with the three typical elements CDEI, CDEII, and CDEIII (Supplemental Fig. S1; Supplemental Table S1), except *Lachancea fantastica* that has only seven chromosomes because of a telomere-to-telomere fusion (Supplemental Fig. S2; Supplemental Table S2). The genomic GC content ranges from 41.2%–47.3% (Supplemental Table S2). In *L. kluyveri*, a region of 1 Mb containing the *MAT* locus and covering almost the whole left arm of chromosome C, has an average GC content of 52.9%, which is significantly higher than the 40.4% global GC content of the rest of the genome (Payen et al. 2009; Souciet et al. 2009). The orthologous counterpart of this chromosomal region is found in all other *Lachancea* species, but none of them presented the GC content heterogeneity characterized in *L. kluyveri*, reinforcing the hypothesis of an introgression event at the origin of this chromosomal arm (Friedrich et al. 2015).

We annotated the coding and noncoding elements of the seven newly sequenced genomes and reannotated the three previously sequenced genomes. Protein-coding genes range from 4997 in *Lachancea meyersii* to 5378 in *L. kluyveri*, and pseudogenes range from 52 in *Lachancea cidri* to 104 in *Lachancea nothofagi* (excluding *Lachancea waltii* where gaps in the original sequence [Kellis et al. 2004] artificially increase the number of pseudogenes to 295) (Supplemental Table S2). On average, coding sequences represent between 67% and 77% of the genome (Fig. 1A), similar to most Saccharomycotina sequenced genomes (Dujon et al. 2004). Finally, we found a small number of Class I retrotransposons (from one to 17) in all species except in *L. cidri* and *L. meyersii*, while Class II elements are more widespread, with at least one copy in *L. cidri* and *Lachancea fermentati* and up to 41 copies in *L. fantastica* (Supplemental Table S2; Sarilar et al. 2014).

We found no correlation between genome size and either the number of genes, the cumulative size of coding sequences, or the transposable element content. However, we found a clear positive correlation between genome sizes and the cumulative sizes of

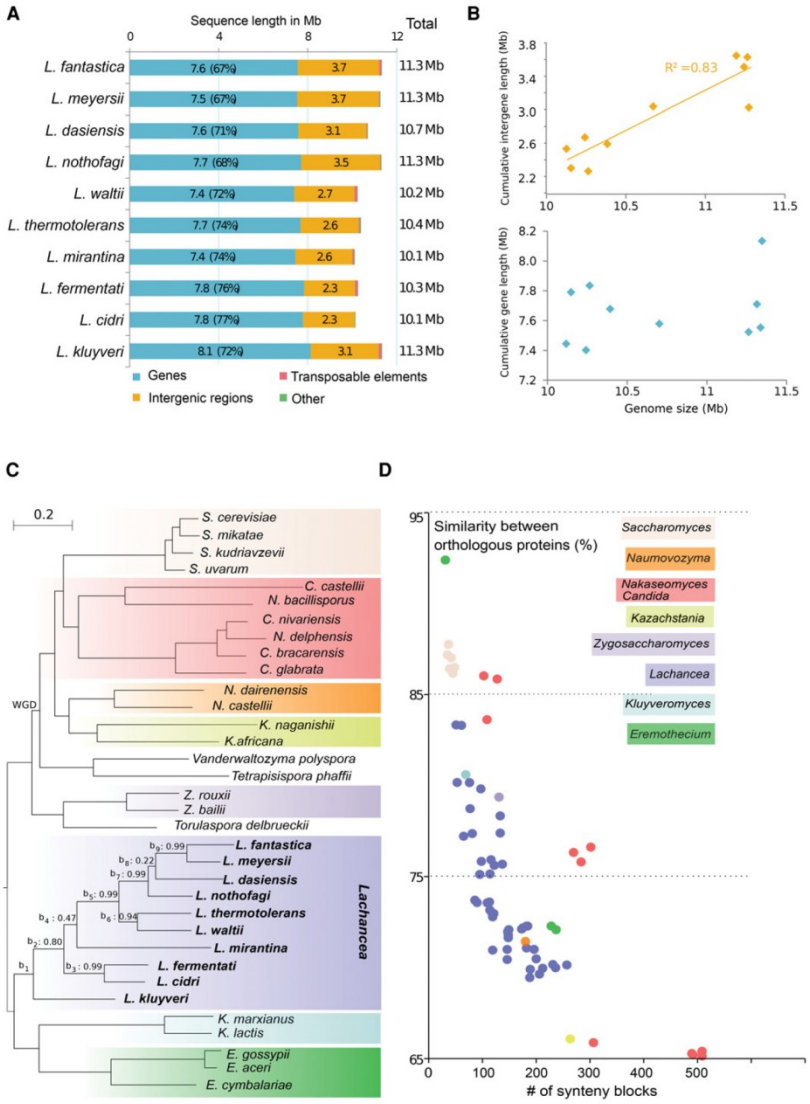


Figure 1. (*A*) Cumulative sequence length of the annotated genetic elements in the 10 *Lachancea* genomes. The percentages of protein-coding sequences are in parentheses. (*B*) Genome size in *Lachancea* positively correlates with intergene length (*top*) but not with cumulative gene length (*bottom*). (*C*) Phylogeny of 34 Saccharomycetaceae species inferred from a maximum likelihood analysis of a concatenated alignment of 756 families of syntenic homologs. The tree topology within the *Lachancea* clade remains identical for several reconstruction methods: concatenation tree, majority tree, and extended majority rule consensus (eMRC) tree (see Methods). Internal branches within the *Lachancea* clade are named b_1 to b_9. The corresponding internode certainty (IC) values, indicating the robustness of the eMRC tree topology, are given. (WGD) Whole-genome duplication. (*D*) Relationship between orthologous protein similarity and the number of conserved synteny blocks within different yeast genera. The *Lachancea* genus is the only clade showing a continuum of genome reorganization and pairwise protein similarities over a large evolutionary range.

intergenic regions and introns (Fig. 1B; Supplemental Fig. S3). The largest size variation in intergenic regions equals a total of 1.38 Mb between *L. fermentati* and *L. meyersii*, showing that the differences between genome sizes are mainly due to variations in noncoding sequence length and not to differential gene or transposable element content. Interestingly, other studies also reported genome

size changes targeted toward intergenic regions. A decrease in gene density with increasing genome size was observed in Ascomycota genomes, (Kelkar and Ochman 2012), and a similar correlation was observed for 81 Saccharomycotina mitochondrial genomes (Freel et al. 2015b).

A robust species tree to reliably reconstruct genome history

Establishing a robust species phylogenetic tree is a crucial prerequisite to any evolutionary reconstruction. The phylogenetic position of the *Lachancea* clade was first inferred from a maximum likelihood analysis of a concatenated alignment of 756 families of syntenic homologs (see Methods) shared by 36 species (Supplemental Table S3). The resulting tree shows that all *Lachancea* species share a monophyletic origin, supporting the existence of the genus (Fig. 1C).

We further reconstructed all gene trees for the 3598 sets of orthologs present in the 10 *Lachancea* species and in *S. cerevisiae* (see Methods). A total of 796 different topologies were observed among the 3598 individual gene trees. The most prevalent topology was shared by 472 gene trees (majority topology). The same topology was systematically retrieved from the concatenations of the corresponding 472 alignments, the 3598 alignments, or the 756 families of syntenic homologs (Supplemental Fig. S4). We also showed that the extended majority rule consensus (eMRC) tree (Felsentein 1995) topology was identical to both the concatenation and the majority topologies (see Methods). We applied the internode certainty (IC) measure (Salichos and Rokas 2013; Salichos et al. 2014) to estimate the robustness of the eMRC topology. Six of eight internal branches have good supporting values (IC higher than 0.8) (Fig. 1C). Branches b_4 and b_8 have the lowest IC values (0.47 and 0.22, respectively); however, their corresponding bipartitions are 7.2 and 3.2 times more frequent than their second most prevalent bipartitions, indicative of a weaker but still exploitable phylogenetic signal (Fig. 1C). Altogether these analyses show that the *Lachancea* phylogeny is robust and reliable for genome history reconstruction.

HGTs and gene creation events contributed to the gene repertoire evolution

We characterized 24 events of putative HGT that correspond to a total of 85 coding sequences (CDS) in *Lachancea* (0.2% of the CDS and pseudogenes; see Methods). Twenty-three events are novel compared to previously reported HGT cases (Rolland et al. 2009; Morel et al. 2015). The 24 HGT families are similar to proteins from *Pezizomycotina* species (10 cases), from other eukaryotes (three cases) and from bacteria (11 cases) (Supplemental Table S4). The phyletic patterns show that eight HGTs are common to several *Lachancea* species; therefore, the transfers would have happened along internal branches of the tree (Fig. 2). Nine HGT families are

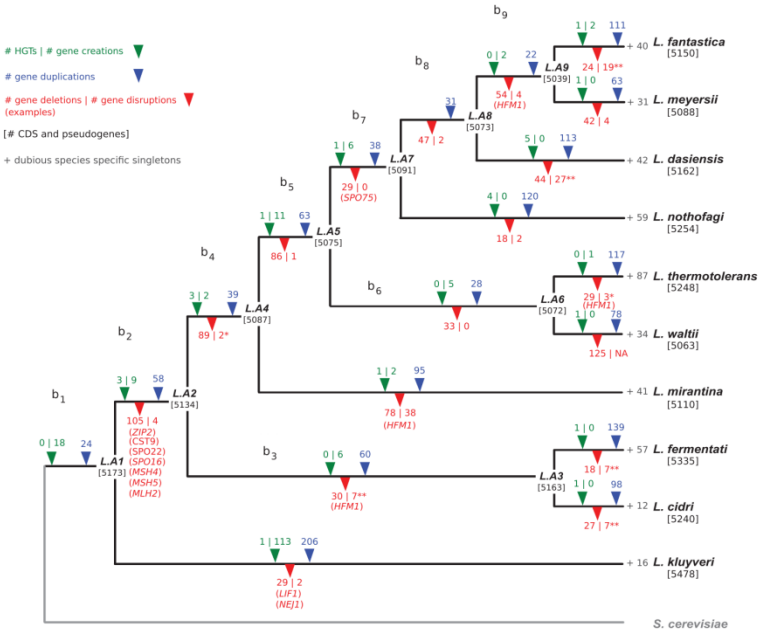


Figure 2. Evolution of the *Lachancea* gene repertoire. The number of gene duplications (in blue) and losses (in red) were estimated on each branch of the tree under a birth–death evolutionary model (Methods). The total number of gene losses indicated on the figure (1036) comprises the 107 cases of dubious loss (see text). The statistical significance of the enrichment of gene losses associated to rearrangements compared to the proportion of genes associated to rearrangements is indicated by an asterisk ($P < 0.05$) or double asterisk ($P < 10^{-4}$). Notable examples of gene losses are in parentheses *below* their corresponding branches. The phyletic patterns of the 51,110 CDS and 1018 pseudogenes were used to map HGTs and gene creation events (in green) in the different branches of the tree (Methods). The number of species-specific singletons is in gray at the tip of each terminal branch. The total number of genes in each ancestral genome and the number of genes and pseudogenes in extant species are indicated between squared brackets.

similar to proteins of unknown function; three are similar to proteins involved in DNA metabolism, i.e., a transcription regulator (pseudogene), an endonuclease, and a serine recombinase previously described (Rolland et al. 2009); and the 12 others are similar to proteins with catalytic activities mostly belonging to oxidation-reduction processes (Supplemental Table S4). Interestingly, homologs of the polysaccharide lyase family 3 of the phytopathogen fungus *Botrytis cinerea* (noble rot fungus) are present in *L. fantastica*, *L. meyersii*, *Lachancea thermotolerans*, and *L. waltii* (Family ID 4751) (Supplemental Table S4). Polysaccharide lyases are mostly found in phytopathogens because they catalyze the eliminative cleavage of pectin, which is a major component of the primary cell wall of many plants. *L. fantastica*, *L. waltii*, and *L. thermotolerans* were isolated from plant-associated habitats, while *L. meyersii* was isolated from seawater. Consistently with ecological distribution, the homolog in *L. meyersii* is highly diverged and only partially similar to the members in the three other species. Overall, our study suggests that the contribution of HGTs on the evolution of the gene repertoire in yeast is probably underestimated.

We also characterized 596 taxonomically restricted gene (TRG) families that are specific to the *Lachancea* clade without any detectable homolog in the nonredundant sequence database or conserved domain in the PFAM database (see Methods). Sixty-six TRG families (encompassing 316 CDS and four pseudogenes) comprise members in at least two different *Lachancea* species and/or several paralogs within a given *Lachancea* species (Supplemental Table S5). Therefore, they could result from de novo gene formation events that occurred in the *Lachancea* clade. Their phyletic patterns were used together with a birth–death–innovation evolutionary model (see Methods) to map these events on the *Lachancea* tree (Fig. 2). The evolutionary rates for these 66 TRG families are generally above the median evolutionary rate of the set of orthologous genes (see Methods) but remain within the distribution, suggesting that no remote homolog would have been missed because of unusually high divergence. Four families show a nonsaturated rate of synonymous substitutions ($d_S < 1$), and all of them have a mean pairwise ratio of nonsynonymous over synonymous substitution rates of $d_N/d_S < 1$, suggesting that they could be under purifying selection (Supplemental Table S5).

The remaining 530 singletons bear the usual characteristics of TRG (Khalturin et al. 2009); they globally have a lower GC content, a smaller size, and a lower codon adaptation index (CAI) value than orthologous gene sequences. With 131 predicted genes, the *L. kluyveri* branch encompasses the highest number of species-specific singleton genes. We used the available population genomic data (Friedrich et al. 2015) to check whether these CDS are conserved between the genomes of *L. kluyveri* isolates we recently sequenced. We found that 114 CDS have homologs conserved in several strains and, therefore, probably correspond to real genes. The remaining 17 CDS are absent or pseudogenized in all other sequenced genomes, suggesting that these genes should be considered dubious. For the other nine species for which no population data are available, all 403 species-specific singletons are also presently considered as dubious genes. Altogether, the nonvertically inherited genes in *Lachancea* would result from a minimum of 24 HGT and 159 gene creation events, which have enriched the genus' gene repertoire.

The genus *Lachancea* covers a unique continuous evolutionary range in Saccharomycotina

The number and size of conserved synteny blocks between *Lachancea* species reveal that they share a continuum of intermediate levels of genome reorganization, ranging from highly collinear genomes down to significantly reordered chromosome maps (Fig. 1D). This continuous range of relatedness is also recognizable through the pairwise protein similarities shared between *Lachancea* orthologs, ranging from 69%–83% (Fig. 1D). More importantly, divergence in the genus *Lachancea*, in both terms of protein sequences and chromosome reorganization, remains below the thresholds beyond which the accumulation of too many mutations and rearrangements leads to the progressive loss of detectable synteny blocks, which prevents any reliable reconstruction of genome history (Drillon and Fischer 2011). Such continuous evolutionary range is so far unique among sequenced Saccharomycotina species (Fig. 1D) and makes the genus *Lachancea* an ideal candidate for the evolutionary reconstruction of genome history.

AnChro, a new computational tool to reconstruct ancestral genome architecture

We developed a new computational method of ancestral genome reconstruction named *AnChro*. This tool is part of an integrated suite of software named CHROnicle (freely available at www.lcqb.upmc.fr/CHROnicle/). Briefly, in the first step of the reconstruction, *SynChro* identifies conserved synteny blocks between pairwise combinations of genomes (Drillon et al. 2014). In the second step, *ReChro* constructs cycles of linked breakpoints between adjacent synteny blocks, and in the last step, *AnChro* infers the ancestral gene order by comparison with external reference genomes (see Supplemental Information).

There are nine ancestral genomes in total in the *Lachancea* phylogenetic tree (named *L.A1* to *L.A9*) (Figs. 2, 3). Genome reconstructions for these nine ancestors resulted in eight ancestral genomes composed of eight chromosomes and in one composed of nine scaffolds, probably because one ancestral adjacency was not reconstructed (Fig. 3, *L.A4*). The number of genes per ancestral genome varies between 4446 for *L.A1* and 4799 for *L.A9* (Fig. 3). Each ancestral genome is provided as a list of ordered ancestral genes with their corresponding orthologous genes in all 10 extant *Lachancea* species (Supplemental Table S6). The robustness of *AnChro*'s reconstructions was comprehensively tested by (1) calculating the probability of reconstructing ancestral genome organization with a single centromere by chromosome, given that *AnChro* does not use any information of centromere position to reconstruct ancestral genome organization (see Supplemental Information); (2) comparing the reconstructions to previously published ancestral genomes (Fig. 4; Supplemental Information; Gordon et al. 2009, 2011; Jones et al. 2012); and (3) benchmarking *AnChro* against four existing reconstruction software tools on both real and simulated genome data sets (Fig. 4; Supplemental Information). All these tests showed that *AnChro* achieved the most reliable and complete reconstruction of ancestral chromosome architecture.

Unbalanced rearrangements outnumbered balanced rearrangements

We performed two independent analyses to identify both balanced chromosomal rearrangements, i.e., translocations, inversions, and chromosome fusion/fission, and unbalanced rearrangements, i.e., duplications and deletions, that occurred since divergence from the last common ancestor of the genus.

To identify balanced rearrangements, we used *SynChro* (Drillon et al. 2014) to compute the synteny blocks between

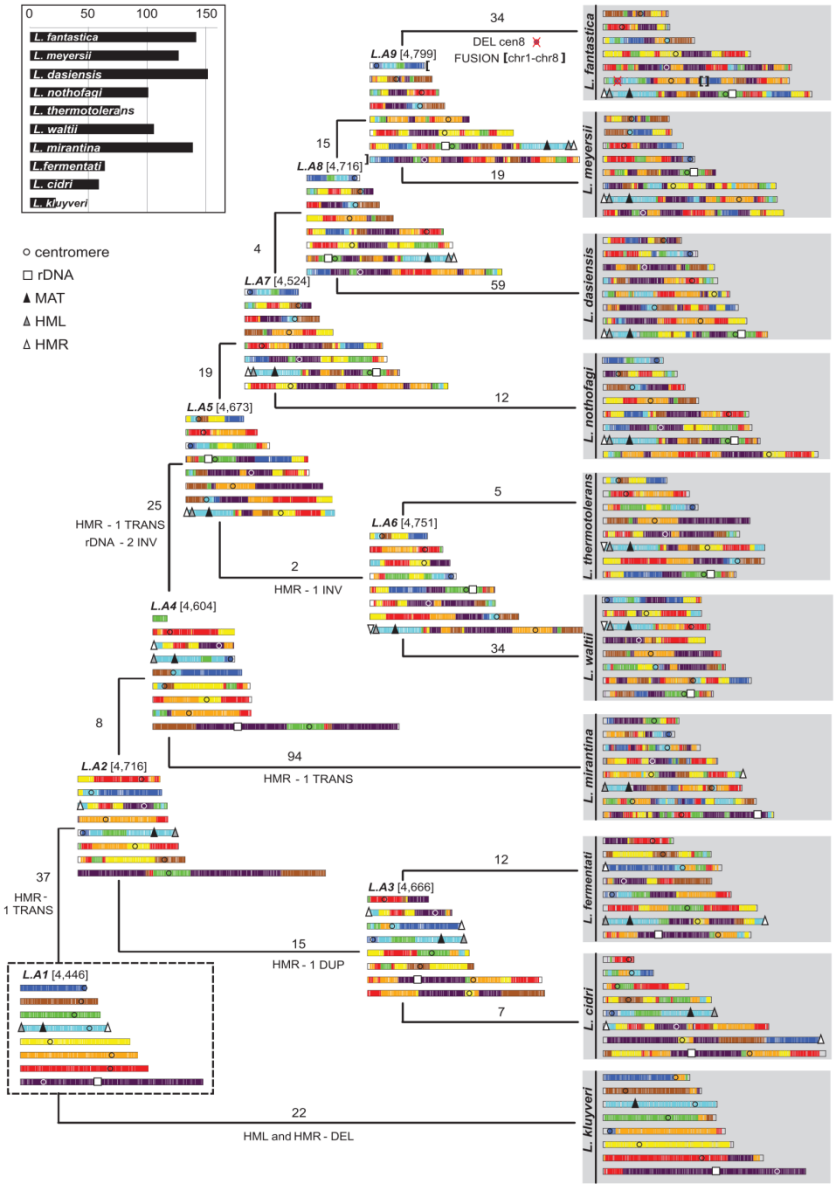


Figure 3. Chromosomal history of the *Lachancea* genomes. The chromosomal structures of the 10 extant species and the ancestral genomes *L.A2* to *L.A9* are represented as a function of the genome structure of *L.A1*, the last common ancestor of the clade. The number of genes of each ancestral genome is indicated with brackets. The total number of translocations and inversions accumulated between two genomes is indicated *above* each branch. Rearrangements involving *MAT*, *HML*, *HMR*, rDNA, or centromeres are indicated *below* each branch. The relocation of the rDNA array occurred in the branch between *L.A4* and *L.A5*. This transposition event occurred intra-chromosomally from an ancestral site, represented as a purple region in *L.A4*, to a new genomic location close to the green centromere in *L.A5*. The relative orientation of the rDNA and the centromere was inverted between *L.A4* and *L.A5*, suggesting that the rDNA relocation resulted from two intra-chromosomal inversions involving one breakpoint reuse (Supplemental Fig. S5; Supplemental Table S6). The interval between *MAT* and *HML* was never broken and was inherited intact from *L.A1* in all extant species except *L. kluyveri*, which lost both *HML* and *HMR*. The *HMR* cassette underwent many rearrangements (three translocations, one inversion, and one duplication) but always remained subtelomeric. The *HML*, *HMR*, and the *MAT* loci were located on the same chromosome in the last common ancestor of the genus, *L.A1*, with one silent cassette at each chromosome end. The only *Lachancea* species that also harbors the three sexual loci on a single chromosome is *L. fermentati*, but this organization is not inherited from *L.A1* as it results from additional translocations in the branch between *L.A3* and *L. fermentati*. Therefore, none of the present-day *Lachancea* species has retained the original chromosomal organization of the sexual loci. The *inset* plot recapitulates the total number of translocations and inversions that accumulated since each extant species diverged from the last common ancestor, *L.A1*.

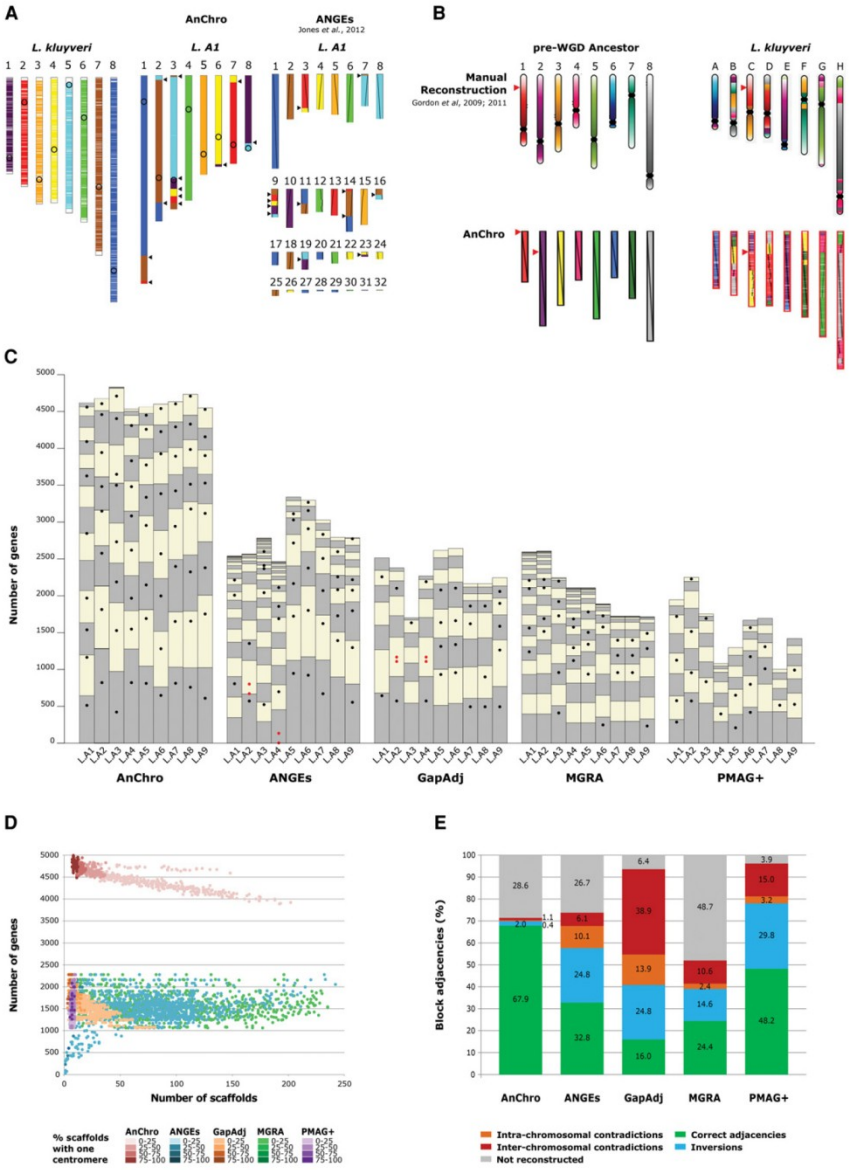


Figure 4. (*A*) Comparison between the two versions of the *L.A1* ancestral genome reconstructed by *AnChro* and by ANGES (Jones et al. 2012). Chromosome painting representations of ancestral genomes are colored relatively to the *L. kluyveri* chromosomes. The black triangles indicate the same 12 ancestral adjacencies that resulted from six translocations identically reconstructed by the two tools. (*B*) Comparison between the manually reconstructed (Gordon et al. 2009, 2011) and the *AnChro* version of the pre-WGD ancestral genome relative to the *L. kluyveri* genome. The only inter-chromosomal difference between the two reconstructions is indicated by the red triangles. (*C*) Comparison of the nine *Lachancea* ancestors (*L.A1* to *L.A9*) reconstructed by AnChro, ANGES, GapAdj, MGRA, and PMAG+. Synteny blocks were computed with I-ADHoRe for the five reconstruction tools. For *AnChro*, a single default reconstruction is presented. Each column represents the ancestral chromosomes of a given ancestor as an alternation of gray and beige boxes, with size being proportional to the number of reconstructed ancestral genes. The small black circles indicate the centromere position. The small red circles indicate the centromere positions when an ancestral chromosome was reconstructed with two centromeres. (*D*) Ancestral genome reconstructions on simulated genomes. The figure presents 900 reconstructed ancestral genomes corresponding to nine different ancestors per simulation and 100 simulations, performed with AnChro (default reconstructions), ANGES, GapAdj, MGRA, and PMAG+. Each genome reconstruction is represented by a dot. The quality of the reconstructions is assessed by three criteria: the number of ancestral scaffolds (ideally eight), the number of ancestral genes (ideally 5000), and the proportion of scaffolds per reconstruction with a single centromere (ideally 100%). (*E*) Average proportions of correctly and incorrectly reconstructed adjacencies for the 900 reconstructions obtained by the five tools. Incorrect adjacencies are decomposed in single block inversions and intra- and inter-chromosomal contradictions. The average proportion of adjacencies that were not reconstructed by the different software is also indicated.

consecutive ancestral genomes in internal branches of the tree and between an ancestor and an extant genome in the terminal branches. The *SynChro* stringency parameter was set to zero to allow building synteny blocks comprising a single inverted orthologous gene-pair (see Supplemental Information). These blocks subsequently served as inputs for *ReChro* to identify all the balanced rearrangements that occurred in each branch of the tree, including single gene inversions. We identified a total of 423 balanced rearrangements (Fig. 3). The number of rearrangements accumulated between *L.A1* and the different *Lachancea* species was highly variable, from 22 in *L. kluyveri* up to 152 in *Lachancea dasiensis* (Fig. 3, inset plot). These rearrangements correspond to 136 inversions, including nine with endpoints at telomeres; 147 translocations, including 140 reciprocal translocations; seven telomeric nonreciprocal translocations; and 140 rearrangements for which it was not possible to discriminate between inversions and translocations because of breakpoint reuse. We identified 102 cases of inversion corresponding to individual chromosomal events with no overlap or breakpoint reuse with other rearrangements. The size distribution of these 102 inversions fits a power law, clearly showing that small inversions are favored over longer ones (Fig. 5A). Only two very large inversions of 318 and 351 genes were found.

We used a birth–death evolutionary model on the gene family classification of the complete set of protein-coding genes from the 10 *Lachancea* genomes to identify unbalanced rearrangements (see Methods). We characterized 1503 gene duplications and 1036 gene losses. We checked all gene losses by looking for syntenic homologs that would have been missed during either genome annotation or gene family clustering because of a level of divergence that could have exceeded the threshold. We filtered out 107 cases of such dubious losses, leaving a total of 929 gene losses, clearly outnumbered by the 1503 gene duplications. We then determined their positions in the phylogenetic tree (Fig. 2). For 132 gene families where the phyletic patterns clearly indicated which members of the family corresponded to the duplicated copies, we found 94 inter- and 38 intra-chromosomally duplicated copies. The distribution of the distances between intra-chromosomally duplicated copies is bimodal, with 20 events separated by <10 kb, possibly resulting from tandem duplications (Supplemental Fig. S6).

At the level of the entire clade, unbalanced rearrangements are six times more abundant than inversions and translocations. Note that this ratio might be overestimated because the number of gene duplications and losses characterized in this work does not necessarily correspond to the number of events that occurred since some duplications and deletions could have involved several genes at the same time. Altogether, this detailed and exhaustive catalog of balanced and unbalanced chromosomal rearrangements positioned on the different branches of the phylogenetic tree provides the opportunity to identify quantitative principles governing genome evolution.

The number of genes in extant genomes is driven by the number of ancestral gene losses

The *Lachancea* gene repertoire underwent 1686 expansion events due to 1503 gene duplications, 159 putative gene creations, 24 HGTs, and 1947 reduction events, corresponding to 1018 pseudogenizations by point mutations or small indels and 929 gene losses by deletion or disruption of the coding sequence by a rearrangement breakpoint.

By integrating all these gene expansions and reductions, we estimated that the last common ancestor of the genus *Lachancea*, *L.A1*, contained about 5173 genes (Fig. 2). The number of genes in extant genomes ranges from 4768 in *L. waltii* to 5378 in *L. kluyveri*, not very different from the estimated ancestral number of genes. These comparable figures might give the impression that the total gene number in *Lachancea* genomes has reached equilibrium where gene expansion events roughly equal gene reduction events. However, we observed between 200 and 400 gene gains per lineage since the divergence from *L.A1*, while gene losses were highly variable, ranging from 31 in *L. kluyveri* to 466 in *L. fantastica*. As a result, we found a negative correlation between the number of genes in extant genomes and the number of gene losses that occurred since the divergence from the last common ancestor of the clade ($R^2 = 0.85$, $P = 1.4 \times 10^{-4}$) (Fig. 5B), but not with the number of gene duplications or HGTs/de novo creations. Therefore, despite more abundant gene gains, the variation of the number of genes between extant genomes mainly results from the number of gene losses that occurred along the different branches of the tree. Remarkably, the especially low level of gene losses in *L. kluyveri* raises the exciting possibility that most gene losses occur by nonhomologous end joining (NHEJ), since essential components of this DSB repair pathway have been specifically lost in *L. kluyveri* (Gordon et al. 2011, and see below).

Finally, we found a negative correlation between the number of genes and pseudogenes present in extant genomes and the number of balanced rearrangements (inversions and translocations) that accumulated since the divergence from the last ancestor *L.A1* ($R^2 = 0.70$, $P = 1.7 \times 10^{-5}$) (Fig. 5B). This relationship was rather unexpected and suggested that fixed balanced rearrangements could be responsible for a significant proportion of gene losses.

Balanced rearrangements frequently disrupt genes at their breakpoints

We tested whether some gene losses could result from gene disruption caused by inversions or translocations with endpoints within coding sequences. From our initial estimate of 929 gene losses, we excluded the 125 losses specific to *L. waltii* because the sequencing gaps in the current genome assembly artificially increase the number of gene losses in this branch (Kellis et al. 2004; Di Rienzi et al. 2011), yielding 804 gene losses, which were positioned in the different branches of the tree. For each lost gene, we considered its two flanking genes in the species that did not undergo the loss. We then determined the positions of the orthologs of these flanking genes in the genome that underwent the loss. For each position, we looked at whether at least one of these orthologs was at the extremity of a synteny block involved in a balanced rearrangement predicted by *ReChro* on the corresponding branch. We found that 109 losses colocalized with a rearrangement breakpoint in a given branch of the tree (Fig. 2). This result suggests that up to 14% of all gene losses (109/804) could result from the disruption of a coding sequence by an inversion or a translocation event. We calculated the statistical significance of this result by testing the global enrichment of gene losses associated to rearrangements compared with the proportion of genes associated to rearrangements (χ^2 test, $P < 1.9 \times 10^{-30}$). The same calculation performed on each branch of the tree showed significativity in only seven out of the 17 branches because of the small sample size in each branch (Fig. 2). Remarkably, we found three gene relics, i.e., highly degenerated remnants of genes (Lafontaine et al. 2004), within intergenic sequences corresponding to rearrangement breakpoints. Such relics

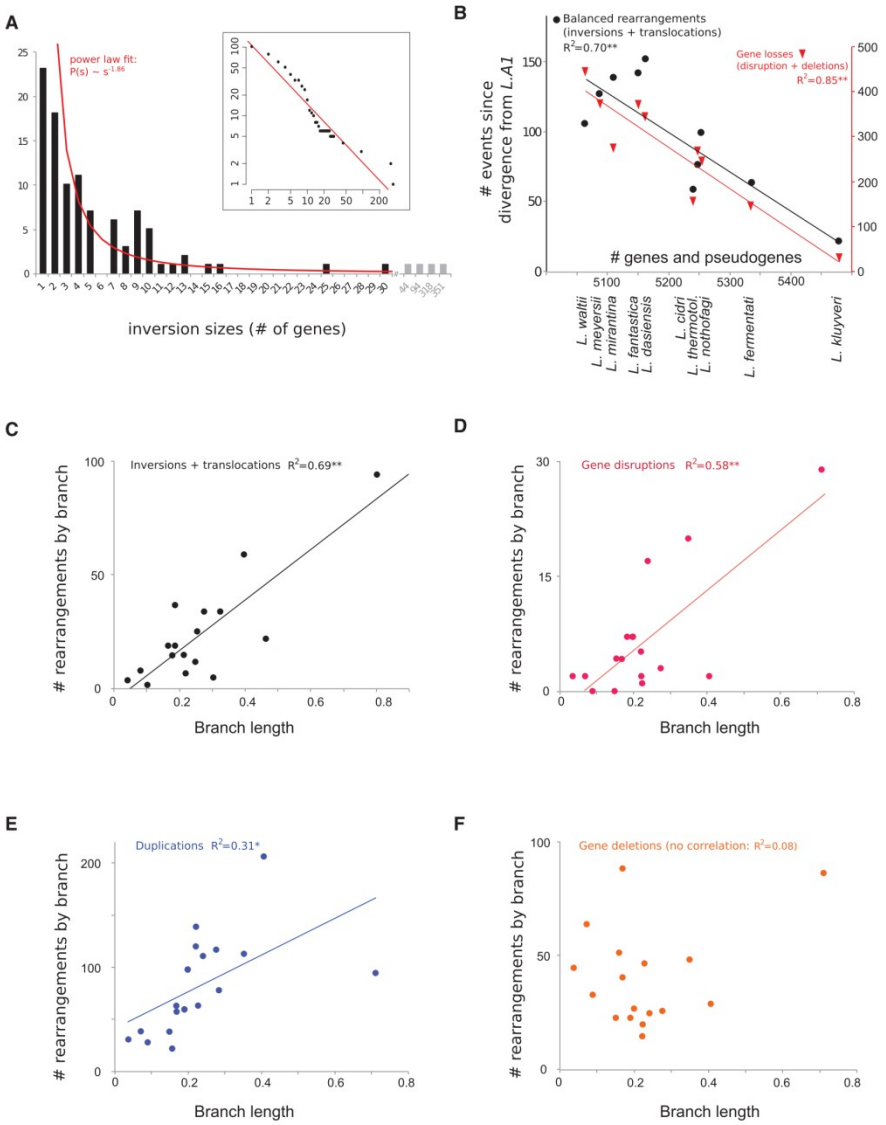


Figure 5. (*A*) Distribution of inversion sizes (in number of genes) accumulated since the divergence from the last common ancestor of the genus (*L.A1*). The red line symbolizes a power law fit to the data ($P(s) = C*s^{-\alpha}$, with C = 106.4 and $\alpha = 1.86$), which represents the probability of an inversion having its two end-points at *s* genes apart. The *inset* plot shows a cumulative histogram of the same data plotted with logarithmic scale. (*B*) Correlations between the number of genes and pseudogenes in extant genomes and number of balanced rearrangements, i.e., inversions and translocations (*left y*-axis) and with the numbers of gene losses, i.e., gene disruptions and deletions (*right y*-axis). (**) $P < 10^{-4}$. No correlation was found between the number of genes in extant genomes and the number of gene duplications (data not shown). (*C*) Correlation between the number of balanced rearrangements (inversions and translocations) and the corresponding individual branch lengths from the *Lachancea* species tree based on the concatenation of 3598 orthologous genes corresponding to 1,983,702 aligned positions (Supplemental Fig. S4). (**) $P < 10^{-4}$. (*D*) Correlation between the number of gene disruptions resulting from balanced rearrangements and branch lengths estimated as in *C*. (**) $P < 10^{-4}$. (*E*) Correlation between the number of gene duplications and branch lengths estimated as in *C*. (*) $P < 0.05$. (*F*) No significant correlation between the number of gene deletions and branch lengths estimated as in *C*.

could correspond to ancient genes disrupted by a rearrangement (Supplemental Fig. S7). These three cases correspond to gene disruptions that occurred in terminal branches of the tree and could therefore correspond to the most recent events that were not yet erased by the accumulation of subsequent point mutations and indels. Overall, these three gene relics support our finding that balanced rearrangements contributed to a significant number of gene losses.

Nonsynonymous substitution rates correlate with the number of inversions, translocations, and duplications, but not deletions

We tested whether the accumulation of large-scale chromosomal mutations and small-scale point mutations were coordinated during evolution, individually for each type of rearrangement. We correlated the number of rearrangements on each branch with branch lengths in the concatenation phylogenetic tree that represent the rates of fixed nonsynonymous substitutions. The number of balanced rearrangements per branch shows a significant positive correlation with the branch lengths ($R^2 = 0.69$, $P = 1.7 \times 10^{-5}$) (Fig. 5C). This correlation holds when inversions and translocations are treated separately ($R^2 = 0.57$ and $R^2 = 0.52$, respectively; data not shown). Remarkably, the 109 gene disruptions resulting from balanced rearrangements show a similar positive correlation with branch lengths ($R^2 = 0.61$, $P = 3.5 \times 10^{-4}$) (Fig. 5D). We also found that the number of gene duplications positively correlated with branch lengths ($R^2 = 0.31$, $P = 0.016$) (Fig. 5E). In contrast, the subset of 695 (=804 – 109) gene deletions that are not associated with breakpoints show no significant correlation with the branch lengths ($R^2 = 0.04$, $P = 0.39$) (Fig. 5F). We identified 45 gene relics among these 695 losses, resulting from the accumulation of point mutations and/or small deletions rather than from large deletions of entire ORFs. Removing these events from the analysis does not result in a significant correlation between DNA deletions and branch lengths ($R^2 = 0.05$, $P = 0.37$; data not shown). These 45 gene relics are equivalent to the 723 annotated pseudogenes (excluding the *L. waltii* genome). We tested the correlation between these losses and branch lengths, focusing on terminal branches of the tree encompassing 103 out of 723 species-specific pseudogenes and 35 out of the 45 detected relics. No correlation was observed between those 138 events and terminal branch lengths ($R^2 = 0.008$, $P = 0.82$), similarly to what was found for gene deletions.

Overall, these observations suggest the existence of a conserved genomic clock that applies to nonsynonymous substitutions, inversions, translocations, and duplications along each branch of the tree. However, deletions and pseudogenizations seem to accumulate independently from the other types of mutational events.

Functional consequences of gene repertoire evolution in *Lachancea* include the loss of the NHEJ and the crossover interfering pathways

Identifying all events that contributed to the evolution of the *Lachancea* gene repertoire allowed us to establish which main functional categories could be affected by gene losses and gains. We found that all essential genes in the S288c *S. cerevisiae* reference strain are conserved in all 10 *Lachancea* species, except 46 cases in *L. waltii* probably due to sequencing gaps (Kellis et al. 2004). No major change was observed in the gene repertoire involved in DNA replication, cell cycle checkpoints, or DNA repair, except the NHEJ pathway that is missing from *L. kluyveri* (Gordon et al. 2011). We confirmed that the orthologs of *LIF1* and *NEJ1* were missing from the *L. kluyveri* reference genome (Fig. 2) and from the 28 sequenced strains of *L. kluyveri* (Friedrich et al. 2015). A relic of *NEJ1* was found next to a rearrangement breakpoint in the *L. kluyveri* genome, suggesting that the loss of *NEJ1* resulted from a gene disruption event (Supplemental Fig. S7). *DNL4* was found as a pseudogene, while a truncated copy of *POL4* of 104 amino acids was annotated as a gene even if the average length in the other *Lachancea* species is 561 amino acids. All NHEJ factors *LIF1*, *NEJ1*, *DNL4*, and *POL4* were found in the other nine *Lachancea* species (except *LIF1* in *L. waltii* that is annotated as a pseudogene because of a sequencing gap) (Supplemental Table S7). Interestingly, the lower number of gene losses in *L. kluyveri* explains its larger gene complement of about 250 genes compared with the other *Lachancea* species (Fig. 2). This raises the exciting possibility that most gene losses occurred by NHEJ in the other species. Moreover, *L. kluyveri* underwent the smallest number of inversions and translocations among all *Lachancea* species since they diverged from their last common ancestor (Fig. 3, inset plot), suggesting that the NHEJ pathway could also participate in the formation of balanced rearrangements. On the contrary, the *L. kluyveri* lineage shows no clear deficit of duplication events compared with other *Lachancea* species such as *Lachancea mirantina*, *L. cidri*, or *L. fermentati* (Fig. 2), which is consistent with previous evidence that segmental duplications result from a replicative mechanism independent from the NHEJ pathway (Payen et al. 2008).

Remarkably, most genes from the ZMM pathway that generates interfering crossovers during meiosis are lost in *Lachancea* species except in *L. kluyveri* (Supplemental Table S8), suggesting a major change in the regulation of meiotic crossover within the genus *Lachancea*. While *ZIP1* is ubiquitous, *ZIP2*, *CST9*, *SPO22*, *SPO16*, and the *MutS* homologs *MSH4/5* are present in *L. kluyveri* only. In addition, *MLH2*, whose function seems to be related to meiotic recombination and to mismatch repair, is also absent from all the *Lachancea* species except *L. kluyveri*. These seven genes were probably lost after the divergence of *L. kluyveri* from the rest of the clade (along the b_2 branch in Fig. 2). The ZMM pathway also comprises *HFM1/MER3* that has homologs in *L. kluyveri*, *L. waltii*, *L. dasiensis*, and *L. nothofagi* (Supplemental Table S8). These genes are conserved in synteny in these species, suggesting that they were inherited vertically from their last common ancestor. Therefore, the phyletic pattern of *HFM1* involves four independent losses (Fig. 2). A gene relic corresponding to *HFM1* (also known as *MER3*) was only found in *L. meyersii*; none of the ZMM gene losses were found associated to a rearrangement breakpoint. Altogether, this suggests a major change in the regulation of meiotic crossover distribution between *L. kluyveri* and the other species of the clade. Interestingly, the ZMM pathway is found in many eukaryotes, including *S. cerevisiae*, plants, and mammals, but it has been lost independently several times during evolution, notably in yeasts, where it is absent from *Schizosaccharomyces pombe*, *Yarrowia lipolytica*, and *Debaryomyces hansenii* (Munz 1994; Richard et al. 2005).

Discussion

Our work combined a significant methodological contribution and a comprehensive comparative genomic analysis on a high-quality genome data set that we generated to achieve a detailed reconstruction of genome history in the model *Lachancea* yeast genus. We discovered relationships between genome size, gene content, chromosomal rearrangements, and rates of protein

divergence that suggest the existence of several evolutionary principles so far uncharacterized.

Our methodological contribution consists in the development of *AnChro* for the reconstruction of ancestral genome architecture. *AnChro* proposes a new conceptual framework based on two original principles. First, our algorithm uses synteny blocks resulting from pairwise comparisons between extant genomes. This preserves the information of synteny conservation shared between closely related genomes even if more distantly related species are present in the clade. In contrast, for algorithms that use universal blocks, the presence of more distant species in the analysis restricts the synteny information to the highest common denominator to all species. Second, *AnChro* combines the advantage of reconstructing reliable adjacencies as in synteny-based methods and of identifying the balanced rearrangements on the branches of the tree as in distance-based methods. The combination of these two approaches presents the additional advantage of allowing the reconstruction of more ancestral adjacencies than by each method alone (Supplemental Information). The association of *AnChro*'s reconstructions with an independent inference of gene duplications and losses under a birth–death evolutionary model using a third-party tool and the identification of candidates for HGT and de novo gene creation events allowed us to achieve a detailed reconstruction of genome history in the model yeast genus *Lachancea*.

Gene duplication is a major driving force in genome evolution as previously anticipated by Ohno (1970). Yeast has exemplified the evolutionary importance of gene duplications and losses since the demonstration of an ancestral WGD in the *S. cerevisiae* lineage (Wolfe and Shields 1997; Fischer et al. 2001; Dietrich et al. 2004; Kellis et al. 2004; Scannell et al. 2007). Interestingly, a study performed at a larger evolutionary scale in fungi reported an excess of gene losses over gene duplications in lineages that diverged before the WGD (Wapinski et al. 2007). However, this analysis relied on published genomes with highly heterogeneous annotations, which may have had a deep impact on the inference of the number of evolutionary events, as acknowledged by the investigators themselves. In our case, the high-quality genome data set coupled with accurate annotations across more closely related species allowed a more precise quantification of the different types of genomic events. We found that gene duplications outnumbered gene losses, suggesting that gene duplication would also be the dominant evolutionary process in a protoploid genus that diverged from the *S. cerevisiae* lineage before the WGD. A similar trend was observed in the CTG yeast clade that did not undergo WGD and comprises most of the *Candida* species, including *Candida albicans* (Butler et al. 2009). These findings confirm the previously anticipated quantitative importance of segmental duplications in yeast genomes (Llorente et al. 2000; Dujon et al. 2004; Souciet et al. 2009).

We characterized 102 chromosomal inversions at single gene resolution. Previous estimates of the distribution of inversion length were constrained by the size of the synteny blocks into which inverted segments were identified (Fischer et al. 2006; Bhutkar et al. 2008). Furthermore, we found that the distribution of inversion lengths fits a power law of coefficient $\alpha = 1.86$ (Fig. 5A). It is tempting to speculate that the power law relationship between the number and the length of fixed inversions indicates that inversions preferentially occur between regions coming into 3D contact in the nucleus since the 3D contact probability between two regions in the yeast nucleus decays with increasing genomic distance as a power law of coefficient 1.5 (Wong et al. 2012). Obviously, other parameters are likely to influence inversions such as chromatin accessibility as suggested by a recent study showing that the distribution of evolutionary breakpoints between five mammalian genomes depends on the 3D contact probability but also on the DNA accessibility in regions of open chromatin (Berthelot et al. 2015). Another parameter could be a higher selective cost associated with large heterozygous inversions compared with small inversions that could be better tolerated during the pairing of homologous chromosomes during meiotic prophase.

Our reconstruction of the *Lachancea* genome history sheds light on several genome evolution principles. We found that the gene number in extant genomes is negatively correlated to both the number of gene losses and the number of balanced rearrangements (inversions and translocations) that were fixed since divergence from the last common ancestor (Fig. 5B). On the contrary, while gene duplications were more abundant than gene losses, their number remained relatively homogeneous in all lineages, and therefore, they do not correlate with the gene complement in extant species. Remarkably, we found that gene losses are significantly enriched at balanced rearrangement breakpoints, representing 14% of the total gene losses. This strongly suggests that translocations and inversions contribute to the reduction of the gene repertoire by disrupting genes at their breakpoints. Further support comes from the identification of three truncated gene relics present at rearrangement breakpoints (Supplemental Fig. S7). In humans, numerous abnormal phenotypes, including intellectual disability and congenital anomalies, are caused by gene disruptions resulting from balanced rearrangements (Fruhmesser et al. 2013; Schluth-Bolard et al. 2013; Moyses-Oliveira et al. 2015; Schneider et al. 2015). This would occur in 6% of de novo reciprocal translocations and 9% of de novo inversions, but these events are detrimental and therefore remain rare in the population. By opposition, balanced rearrangements that reach fixation in populations are thought to be less detrimental because they are usually considered to occur in intergenic regions (Peng et al. 2006; Poyatos and Hurst 2007; Berthelot et al. 2015). Here, we show that numerous balanced rearrangements that occurred within coding sequences reached fixation in yeast populations. Our study provides the first genome-scale quantification of this phenomenon in a eukaryotic genus.

Finally, we showed that the number of balanced and unbalanced rearrangements varies greatly between lineages, leading to genomes in extant species that were differently rearranged compared with the ancestral genome of the genus (Figs. 2, 3). Such variable rates of genome rearrangements were already described in vertebrates and in yeasts (Bourque et al. 2005; Fischer et al. 2006). Furthermore, we found that nonsynonymous substitutions and inversions, translocations, and duplications reach fixation at a coordinated pace within each branch of the phylogenetic tree (Fig. 5). Previous works reported comparable correlations in *Drosophila*, bacteria, Archaea, and plastid genomes (Snel et al. 2002; Bhutkar et al. 2008; Csuros and Miklos 2009; Puigbo et al. 2014; Weng et al. 2014). Puigbo et al. (2014) coined the term genomic clock to describe the concept of coordinated pace of fixation between amino acid substitutions and large-scale rearrangements. This term might be misleading in the sense that a clock-like process is expected to follow a constant rate in time. This is clearly not the case here because the rates of substitution and number of rearrangements vary between branches. In bacteria, gene loss has been reported to be a more uniform, "clock-like" process than gene gain, suggesting that gene loss would be mostly neutral, whereas gene gain would be under positive selection or controlled

by genetic drift enabled by population bottlenecks (Puigbo et al. 2014). In contrast, we found that gene deletion and pseudogenization are the only types of events that show no apparent correlation with protein sequence divergence. Overall, our findings open new questions on the respective selective value of various mutational events in eukaryotes. Further work is now needed to determine whether a genomic clock can be observed in a wider number of taxa. So far, the complete genome of approximately 100 yeast species have been published, and this number is still increasing. There are about 1200 known Saccharomycotina yeast species (Hittinger et al. 2015), and the project to sequence and analyze their genomes was recently initiated (http://www.y1000plus.org/). This will allow testing in this entire yeast subphylum of the existence of the evolutionary principles that we uncovered in the genus *Lachancea*. Further work will be needed to determine whether these principles also apply in other organisms such as vertebrates.

Methods

Strain selection, ploidy, karyotypes, and culture conditions

We selected seven *Lachancea species: L. cidri*, *L. fermentati* (both formerly called *Zygosaccharomyces* species) (Kurtzman 2003), *L. meyersii* (Fell et al. 2004), *L. dasiensis* (Lee et al. 2009), *L. mirantina* (Pereira et al. 2011), *L. nothofagi* (Mestre et al. 2010), and *L. fantastica nomen nudum* (Fig. 1). We renamed the strain CBS6924 as *L. fantastica nomen nudum* because it was erroneously classified as *L. thermotolerans*. These species were isolated worldwide often in association with plants, plant products, or insects. Several isolates from all different species were collected except for *L. fantastica* and *L. mirantina*, which were represented only by one strain. Electrophoretic karyotyping was performed for all strains as previously described (Neuveglise et al. 2000) (Supplemental Fig. S8). The ploidy of each strain was assessed using flow cytometry as previously described (Agier et al. 2013). Natural isolates were mainly haploids in all 10 species, while diploids were found in five species only (Supplemental Table S9). One haploid strain per species was selected for sequencing: *L. meyersii* CBS 8951^T, *L. fantastica nomen nudum* CBS 6924, *L. nothofagi* CBS 11611^T, *L. dasiensis* CBS 10888, *L. fermentati* CBS 6772, *L. cidri* CBS 2950, and *L. mirantina* CBS 11717 (Supplemental Table S10). Note that two other species, *Lachancea lanzarotensis* and *Lachancea quebecensis*, were described and sequenced during the course of this work and are not taken into consideration in this study (Gonzalez et al. 2013; Freel et al. 2015a, 2016; Sarilar et al. 2015).

DNA extraction, sequencing, and assembly of *Lachancea* genomes

Nuclear DNA was separated from mitochondrial and plasmid DNA by CsCl gradient (Supplemental Methods). Sequencing was carried out with a combination of Roche 454 in single-read and paired-end 8 kb on a GS-Flex+ apparatus, and Illumina in single read of 50 bp on a HiSeq2000 apparatus. Illumina reads allowed the correction of sequencing errors in homopolymer blocks that generated erroneous frameshifted genes. Genome assemblies were achieved with Celera Assembler version 6.1 (Myers et al. 2000) and Newbler v2.7 (454 Life Sciences) (Supplemental Methods).

Annotation of *Lachancea* genomes

The genomes of *L. kluyveri* and *L. thermotolerans* were used as references for gene structure annotation in the seven newly sequenced genomes and for the reannotation of *L. waltii*. We first completed the two reference genome annotations by detecting genes in intergenic regions through BLASTX against the UniProt fungi database. Gene models were annotated for the seven newly sequenced genomes with an annotation transfer pipeline that we developed with the AMADEA Biopack platform (Isoft, http://www.isoft.fr/bio/biopack_en.htm) (Supplemental Methods). Manual curation of gene models consisted of resolving gene models with missing start or stop codons, with not properly defined introns or with frameshifts. Additional CDSs were identified in intergenic regions of the 10 species by BLASTX search against the nr database and manual curation. Moreover, ORFs longer than 150 amino acids without any homologs in the nr database were predicted with Orffinder (NCBI). tRNA genes were predicted with tRNAscan-SE (v.1.3.1) (Lowe and Eddy 1997) with default searching parameters of tRNAscan and EufindtRNA; covariance model: tRNA2-euk.cm. The snRNA genes were searched by BLASTN using *L. kluyveri* and *L. thermotolerans* known snRNA sequences as query. We identified complete and partial transposable elements as well as solo-LTRs using BLAST against known transposable elements of *Ty1/Copia*, *Ty3/Gypsy*, and class-II superfamilies. Elements of the *Rover* and *Roamer* families are described elsewhere (Sarilar et al. 2014). The position of centromeres in the seven newly sequenced *Lachancea* genomes and in *L. waltii* was inferred from synteny conservation with already annotated centromeres in *L. thermotolerans*, *L. kluyveri*, and *Zygosaccharomyces rouxii* (Souciet et al. 2009). CDEI, CDEII, and CDEIII motifs were identified with the MEME program (Bailey and Elkan 1994), using the oops mode on both strands (Supplemental Methods).

The functional annotation of protein-coding genes has been established on the basis of homology with *S. cerevisiae* genes (SGD S288C ORF translations, release February 3, 2011, available at http:// downloads.yeastgenome.org/sequence/S288C_reference/orf_protein/) or the NCBI Reference Sequence (RefSeq) database (release 58 of March 11, 2013, available at http://www.ncbi.nlm.nih.gov/refseq/) for putative genes without homologs in *S. cerevisiae* (Supplemental Methods).

Phylogenetic analyses

Orthologous genes were defined as syntenic homologs. Synteny block reconstructions were computed with the *SynChro* algorithm (Drillon et al. 2014) for all pairwise combinations between 36 yeast species (Supplemental Table S3). We inferred by transitivity 756 and 3598 groups of syntenic homologs composed of only one gene per species in the 36 yeast and 10 *Lachancea* species, respectively.

A multiple alignment of each group of orthologs was generated at the amino acid level with the MAFFT algorithm (linsi implementation, default parameters) (Katoh and Toh 2008). The best substitution model was determined by ProtTest (Abascal et al. 2005). PhyML reconstructions were performed from the concatenation of the multiple alignments for the Saccharomycotina set (756 orthologous groups: 486,399 aligned positions) and for the *Lachancea* set (either all 3598 orthologous genes—1,983,702 aligned positions—or the 472 orthologous groups whose individual trees have the eMRC topology—387,091 aligned positions), using the LG model and a gamma-law distribution with four categories of evolution rates (Guindon and Gascuel 2003). In all cases, 500 bootstrap replicates were performed.

We selected the 15,227 most strongly supported bipartitions (bootstrap value >0.95) out of the 32,391 bipartitions present in the 3598 individual gene trees to construct an unrooted eMRC tree that displays the most prevalent bipartitions in the data set. Each internal branch in the eMRC tree is associated with its gene support frequency (GSF), i.e., the number of gene trees supporting it (Gadagkar and Kumar 2005). To estimate the level of

incongruence in this set of gene trees, we calculated the IC as recently proposed by Salichos et al. (2014). Tree certainty (TC) values are the sum of the IC values for all internodes. TC values range from zero (maximum conflict among individual gene trees) to eight (total number of internal nodes in our 11 taxon eMRC tree, no conflict among the individual gene trees). The eMRC tree, IC, and TC values were calculated with RAxML V8 (Stamatakis 2014). All phylogenetic reconstructions were achieved by considering all aligned positions as homologous characters, i.e., no removal of gap positions because identical tree topologies with negligible variations of TC values were obtained with or without considering gapped positions.

Gene families

An all-against-all BLASTP (version 2.2.28+) comparison was performed between amino acid translations of all CDS from the 10 *Lachancea* species and *S. cerevisiae*, with default options and Smith-Waterman alignment (Altschul et al. 1997). Hits with an *E*-value lower than 10^{-3} were clustered with TribeMCL with an inflation value I = 6.5 (Supplemental Methods; Enright et al. 2003). The detailed composition of all gene families and singletons is provided in Supplemental Table S12; their repartition among the 10 species, in Supplemental Table S2.

Systematic search for homologs to the *Lachancea* protein families was performed in the nr database with PSI-BLAST using a position-specific scoring matrix (PSSM) built from the family multiple alignments (only one iteration is performed). A search for homologs to *Lachancea* singletons was performed in the nr database with BLASTP (Altschul et al. 1997). Hits with an *E*-value lower than 10^{-3} with at least 25% sequence identity and coverage of the longest sequence of at least 50% were considered as significant. Similarity search against the PFAM database (version 27.0) was performed with hmmsearch from the HMMER3 package (Mistry et al. 2013), and hits with an *E*-value lower than 10^{-5} were considered significant. Search for conserved protein domains was also performed with rpsblast from the BLAST 2.2.29+ distribution, against the CDD database, version 01/10/2014.

Gene content evolution

Gene acquisitions and losses were inferred with the BadiRate program (Librado et al. 2012). For nonvertically inherited gene families (HGTs and TRGs), we used the birth, death, and innovation (BDI) model with free rates (FRs) estimated by the Wagner parsimony method (CWP). For vertically inherited families, we used the birth and death model with FR and CWP, assuming that all families derived from ancestral genes present in the common ancestor of all *Lachancea* species.

CAI values were calculated for all TRGs using CAIJava, (Carbone et al. 2003). Rates of synonymous substitutions (d_S) and rates of nonsynonymous substitutions (d_N) were estimated with the yn00 program from the PAML package (Yang 1997).

For the 127 *L. kluyveri*–specific TRGs, the CDS were considered physically absent from a given *L. kluyveri* strain if their sequencing coverage (estimated by mpileup in samtools) in the BAM files from the 28 sequenced *L. kluyveri* strains (Friedrich et al. 2015) was lower than the mean coverage minus two standard deviations for the core genome of that strain (*L. kluyveri* syntenic homologs).

Ancestral genome reconstruction

Ancestral genome reconstruction was performed with the CHROnicle suite of programs freely available at www.lcqb.upmc.fr/CHROnicle/ that comprises *SynChro*, *ReChro*, and *AnChro*. All the details about the ancestral gene order reconstruction steps, the identification of chromosomal rearrangements in the different branches of the tree, and the validation of the reconstructions are in the Supplemental information file. *AnChro* source code can also be found in the Supplemental Material.

Data access

Accession numbers and/or website sources for all yeast species used in this work are listed in Supplemental Table S3. Genome sequences and (re)annotations of the 10 *Lachancea* species are available on the GRYC website: http://gryc.inra.fr. The sequencing reads and the seven new genome assemblies from this study have been submitted to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) under the following accession numbers: PRJEB12910, PRJEB12928, PRJEB12929, PRJEB12930, PRJEB12931, PRJEB12932, and PRJEB12933.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche (GB-3G, ANR-10-BLAN-1606). We thank Guillaume Achaz, Gilles Charvin, Frédéric Devaux, Cécile Fairhead, Romain Koszul, Gianni Liti, Marie-Claude Marsolier Kergoat, and Conrad Nieduszynski for fruitful discussions.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21:** 2104–2105.

Agier N, Romano OM, Touzain F, Lagomarsino MC, Fischer G. 2013. The spatio-temporal program of replication in the genome of *Lachancea kluyveri*. *Genome Biol Evol* **5:** 370–388.

Alekseyev MA, Pevzner PA. 2009. Breakpoint graphs and ancestral genome reconstructions. *Genome Res* **19:** 943–957.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Avelar AT, Perfeito L, Gordo I, Ferreira MG. 2013. Genome architecture is a selectable trait that can be maintained by antagonistic pleiotropy. *Nat Commun* **4:** 2235.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2:** 28–36.

Berthelot C, Muffato M, Abecassis J, Roest Crollius H. 2015. The 3D organization of chromatin explains evolutionary fragile genomic regions. *Cell Rep* **10:** 1913–1924.

Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM. 2008. Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. *Genetics* **179:** 1657–1680.

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

Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, et al. 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* **459:** 657–662.

Carbone A, Zinovyev A, Kepes F. 2003. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* **19:** 2005–2015.

Csuros M, Miklos I. 2009. Streamlining and large ancestral genomes in *Archaea* inferred with a phylogenetic birth-and-death model. *Mol Biol Evol* **26:** 2087–2095.

Di Rienzi SC, Lindstrom KC, Lancaster R, Rolczynski L, Raghuraman MK, Brewer BJ. 2011. Genetic, genomic, and molecular tools for studying the protoploid yeast, *L. waltii*. *Yeast* **28:** 137–151.

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, et al. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304:** 304–307.

Drillon G, Fischer G. 2011. Comparative study on synteny between yeasts and vertebrates. *C R Biol* **334:** 629–638.

Drillon G, Carbone A, Fischer G. 2014. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* **9:** e92621.
Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, et al. 2004. Genome evolution in yeasts. *Nature* **430:** 35–44.
Enright AJ, Kunin V, Ouzounis CA. 2003. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* **31:** 4632–4638.
Faraut T. 2008. Addressing chromosome evolution in the whole-genome sequence era. *Chromosome Res* **16:** 5–16.
Fell JW, Statzell-Tallman A, Kurtzman CP. 2004. *Lachancea meyersii* sp. nov., an ascosporogenous yeast from mangrove regions in the Bahama Islands. *Stud Mycol* **50:** 359–363.
Felsentein J. 1995. *Phylogenetic inference package (PHYLIP), version 3.5*. University of Washington, Seattle, WA.
Fischer G, Neuveglise C, Durrens P, Gaillardin C, Dujon B. 2001. Evolution of gene order in the genomes of two related yeast species. *Genome Res* **11:** 2009–2019.
Fischer G, Rocha EP, Brunet F, Vergassola M, Dujon B. 2006. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet* **2:** e32.
Freel KC, Charron G, Leducq JB, Landry CR, Schacherer J. 2015a. *Lachancea quebecensis* sp. nov., a yeast species consistently isolated from tree bark in the Canadian province of Quebec. *Int J Syst Evol Microbiol* **65:** 3392–3399.
Freel KC, Friedrich A, Schacherer J. 2015b. Mitochondrial genome evolution in yeasts: an all-encompassing view. *FEMS Yeast Res* **15:** fov023.
Freel KC, Friedrich A, Sarilar V, Devillers H, Neuveglise C, Schacherer J. 2016. Whole-genome sequencing and intraspecific analysis of the yeast species *Lachancea quebecensis*. *Genome Biol Evol* **8:** 733–741.
Friedrich A, Jung P, Reisser C, Fischer G, Schacherer J. 2015. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol Biol Evol* **32:** 184–192.
Fruhmesser A, Blake J, Haberlandt E, Baying B, Raeder B, Runz H, Spreiz A, Fauth C, Benes V, Utermann G, et al. 2013. Disruption of *EXOC6B* in a patient with developmental delay, epilepsy, and a *de novo* balanced t(2;8) translocation. *Eur J Hum Genet* **21:** 1177–1180.
Gabaldon T, Martin T, Marcet-Houben M, Durrens P, Bolotin-Fukuhara M, Lespinet O, Arnaise S, Boisnard S, Aguileta G, Atanasova R, et al. 2013. Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics* **14:** 623.
Gadagkar SR, Kumar S. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Mol Biol Evol* **22:** 2139–2141.
Gagnon Y, Blanchette M, El-Mabrouk N. 2012. A flexible ancestral genome reconstruction method based on gapped adjacencies. *BMC Bioinformatics* **13**(Suppl 19): S4.
Galeote V, Novo M, Salema-Oom M, Brion C, Valerio E, Goncalves P, Dequin S. 2010. *FSY1*, a horizontally transferred gene in the *Saccharomyces cerevisiae* EC1118 wine yeast strain, encodes a high-affinity fructose/H^+ symporter. *Microbiology* **156**(Pt 12): 3754–3761.
Gonzalez SS, Alcoba-Florez J, Laich F. 2013. *Lachancea lanzarotensis* sp. nov., an ascomycetous yeast isolated from grapes and wine fermentation in Lanzarote, Canary Islands. *Int J Syst Evol Microbiol* **63**(Pt 1): 358–363.
Gordon JL, Byrne KP, Wolfe KH. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet* **5:** e1000485.
Gordon JL, Byrne KP, Wolfe KH. 2011. Mechanisms of chromosome number evolution in yeast. *PLoS Genet* **7:** e1002190.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696–704.
Hittinger CT, Rokas A, Bai FY, Boekhout T, Goncalves P, Jeffries TW, Kominek J, Lachance MA, Libkind D, Rosa CA, et al. 2015. Genomics and the making of yeast biodiversity. *Curr Opin Genet Dev* **35:** 100–109.
Jones BR, Rajaraman A, Tannier E, Chauve C. 2012. ANGES: reconstructing ANcestral GEnomeS maps. *Bioinformatics* **28:** 2388–2390.
Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9:** 286–298.
Kelkar YD, Ochman H. 2012. Causes and consequences of genome expansion in fungi. *Genome Biol Evol* **4:** 13–23.
Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428:** 617–624.
Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. 2009. More than just orphans: Are taxonomically-restricted genes important in evolution? *Trends Genet* **25:** 404–413.
Kurtzman CP. 2003. Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Res* **4:** 233–245.
Lafontaine I, Dujon B. 2010. Origin and fate of pseudogenes in Hemiascomycetes: a comparative analysis. *BMC Genomics* **11:** 260.
Lafontaine I, Fischer G, Talla E, Dujon B. 2004. Gene relics in the genome of the yeast *Saccharomyces cerevisiae*. *Gene* **335:** 1–17.
Lee CF, Yao CH, Liu YR, Hsieh CW, Young SS. 2009. *Lachancea dasiensis* sp. nov., an ascosporogenous yeast isolated from soil and leaves in Taiwan. *Int J Syst Evol Microbiol* **59**(Pt 7): 1818–1822.
Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics* **28:** 279–281.
Llorente B, Durrens P, Malpertuy A, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, Brottier P, Casaregola S, et al. 2000. Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Lett* **487:** 122–133.
Louis VL, Despons L, Friedrich A, Martin T, Durrens P, Casaregola S, Neuveglise C, Fairhead C, Marck C, Cruz JA, et al. 2012. *Pichia sorbitophila*, an interspecies yeast hybrid, reveals early steps of genome resolution after polyploidization. *G3 (Bethesda)* **2:** 299–311.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25:** 955–964.
Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent WJ, Blanchette M, Haussler D, Miller W. 2006. Reconstructing contiguous regions of an ancestral genome. *Genome Res* **16:** 1557–1565.
Marcet-Houben M, Gabaldon T. 2010. Acquisition of prokaryotic genes by fungal genomes. *Trends Genet* **26:** 5–8.
Marcet-Houben M, Gabaldon T. 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the Baker's yeast lineage. *PLoS Biol* **13:** e1002220.
Marsit S, Mena A, Bigey F, Sauvage FX, Couloux A, Guy J, Legras JL, Barrio E, Dequin S, Galeote V. 2015. Evolutionary advantage conferred by an eukaryote-to-eukaryote gene transfer event in wine yeasts. *Mol Biol Evol* **32:** 1695–1707.
McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci* **370:** 20140332.
Mestre MC, Ulloa JR, Rosa CA, Lachance MA, Fontenla S. 2010. *Lachancea nothofagi* sp. nov., a yeast associated with *Nothofagus* species in Patagonia, Argentina. *Int J Syst Evol Microbiol* **60:** 2247–2250.
Mighell AJ, Smith NR, Robinson PA, Markham AF. 2000. Vertebrate pseudogenes. *FEBS Lett* **468:** 109–114.
Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41:** e121.
Morales L, Dujon B. 2012. Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. *Microbiol Mol Biol Rev* **76:** 721–739.
Morel G, Sterck L, Swennen D, Marcet-Houben M, Onesime D, Levasseur A, Jacques N, Mallet S, Couloux A, Labadie K, et al. 2015. Differential gene retention as an evolutionary mechanism to generate biodiversity and adaptation in yeasts. *Sci Rep* **5:** 11571.
Moyses-Oliveira M, Guilherme RS, Meloni VA, Di Battista A, de Mello CB, Bragagnolo S, Moretti-Ferreira D, Kosyakova N, Liehr T, Carvalheira GM, et al. 2015. X-linked intellectual disability related genes disrupted by balanced X-autosome translocations. *Am J Med Genet B Neuropsychiatr Genet* **168:** 669–677.
Muffato M, Roest Crollius H. 2008. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. *Bioessays* **30:** 122–134.
Munz P. 1994. An analysis of interference in the fission yeast *Schizosaccharomyces pombe*. *Genetics* **137:** 701–707.
Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al. 2000. A whole-genome assembly of *Drosophila*. *Science* **287:** 2196–2204.
Neuveglise C, Bon E, Lepingle A, Wincker P, Artiguenave F, Gaillardin C, Casaregola S. 2000. Genomic exploration of the hemiascomycetous yeasts: 9. *Saccharomyces kluyveri*. *FEBS Lett* **487:** 56–60.
Ohno S. 1970. *Evolution by gene duplication*. Springer Verlag, New York.
Ouangraoua A, Tannier E, Chauve C. 2011. Reconstructing the architecture of the ancestral amniote genome. *Bioinformatics* **27:** 2664–2671.
Payen C, Koszul R, Dujon B, Fischer G. 2008. Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. *PLoS Genet* **4:** e1000175.
Payen C, Fischer G, Marck C, Proux C, Sherman DJ, Coppee JY, Johnston M, Dujon B, Neuveglise C. 2009. Unusual composition of a yeast chromosome arm is associated with its delayed replication. *Genome Res* **19:** 1710–1721.
Peng Q, Pevzner PA, Tesler G. 2006. The fragile breakage versus random breakage models of chromosome evolution. *PLoS Comput Biol* **2:** e14.
Pereira LF, Costa CR Jr, Brasileiro BT, de Morais MA Jr. 2011. *Lachancea mirantina* sp. nov., an ascomycetous yeast isolated from the cachaca fermentation process. *Int J Syst Evol Microbiol* **61:** 989–992.

Perez-Ortin JE, Querol A, Puig S, Barrio E. 2002. Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome Res* **12:** 1533–1539.
Poyatos JF, Hurst LD. 2007. The determinants of gene order conservation in yeasts. *Genome Biol* **8:** R233.
Puigbo P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol* **12:** 66.
Quintero-Rivera F, Xi QJ, Keppler-Noreuil KM, Lee JH, Higgins AW, Anchan RM, Roberts AE, Seong IS, Fan X, Lage K, et al. 2015. *MATR3* disruption in human and mouse associated with bicuspid aortic valve, aortic coarctation and patent ductus arteriosus. *Hum Mol Genet* **24:** 2375–2389.
Richard GF, Kerrest A, Lafontaine I, Dujon B. 2005. Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination. *Mol Biol Evol* **22:** 1011–1023.
Rolland T, Dujon B. 2011. Yeasty clocks: dating genomic changes in yeasts. *C R Biol* **334:** 620–628.
Rolland T, Neuveglise C, Sacerdot C, Dujon B. 2009. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. *PLoS One* **4:** e6515.
Rowley JD. 1998. The critical role of chromosome translocations in human leukemias. *Annu Rev Genet* **32:** 495–519.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497:** 327–331.
Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol* **31:** 1261–1271.
Sarilar V, Bleykasten-Grosshans C, Neuveglise C. 2014. Evolutionary dynamics of *hAT* DNA transposon families in *Saccharomycetaceae*. *Genome Biol Evol* **7:** 172–190.
Sarilar V, Devillers H, Freel KC, Schacherer J, Neuveglise C. 2015. Draft genome sequence of *Lachancea lanzarotensis* CBS 12615^T, an ascomycetous yeast isolated from grapes. *Genome Announc* **3:** pii:e00292-15.
Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci* **104:** 8397–8402.
Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, Rine J, Johnston M, Hittinger CT. 2011. The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces sensu stricto* genus. *G3 (Bethesda)* **1:** 11–25.
Schluth-Bolard C, Labalme A, Cordier MP, Till M, Nadeau G, Tevissen H, Lesca G, Boutry-Kryza N, Rossignol S, Rocas D, et al. 2013. Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. *J Med Genet* **50:** 144–150.
Schneider A, Puechberty J, Ng BL, Coubes C, Gatinois V, Tournaire M, Girard M, Dumont B, Bouret P, Magnetto J, et al. 2015. Identification of disrupted *AUTS2* and *EPHA6* genes by array painting in a patient carrying a de novo balanced translocation t(3;7) with intellectual disability and neurodevelopment disorder. *Am J Med Genet A* **167:** 3031–3037.
Semon M, Wolfe KH. 2007. Reciprocal gene loss between *Tetraodon* and zebrafish after whole genome duplication in their ancestor. *Trends Genet* **23:** 108–112.
Snel B, Bork P, Huynen MA. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12:** 17–25.
Souciet JL, Dujon B, Gaillardin C, Johnston M, Baret PV, Cliften P, Sherman DJ, Weissenbach J, Westhof E, Wincker P, et al. 2009. Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res* **19:** 1696–1709.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313.
Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449:** 54–61.
Weng ML, Blazier JC, Govindu M, Jansen RK. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Mol Biol Evol* **31:** 645–659.
Wisecaver JH, Slot JC, Rokas A. 2014. The evolution of fungal metabolic pathways. *PLoS Genet* **10:** e1004816.
Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708–713.
Wong H, Marie-Nelly H, Herbert S, Carrivain P, Blanc H, Koszul R, Fabre E, Zimmer C. 2012. A predictive computational model of the dynamic 3D interphase yeast nucleus. *Curr Biol* **22:** 1881–1890.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555–556.
Zuckerkandl E, Pauling LB. 1962. Molecular disease, evolution, and genetic heterogeneity. In *Horizons in biochemistry* (ed. Kasha M, Pullman B), pp. 189–225. Academic Press, New York.

Received January 14, 2016; accepted in revised form April 28, 2016.