Génomique des populations : étude comparative au sein du sous-phylum des Saccharomycotina

Jean-Sébastien Gounot

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

UMR 7156

THÈSE présentée par :

JEAN-SEBASTIEN GOUNOT

soutenue le : 21 SEPTEMBRE 2018

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Bioinformatique

Génomique des populations : Étude comparative au sein du sous-phylum des Saccharomycotina

THÈSE dirigée par :

Pr. SCHACHERER Joseph	Professeur, Université de Strasbourg

RAPPORTEURS :

Dr. KOSZUL Romain	Directeur de recherches, Institut Pasteur
Dr. MARULLO Philippe	Chargé de recherches, Université de Bordeaux

---

AUTRES MEMBRES DU JURY :

Pr. FAIRHEAD Cécile	Professeur, Université Paris Sud
Pr. LECOMPTE Odile	Professeur, Université de Strasbourg
Dr. FRIEDRICH Anne	Maître de conférence, Université de Strasbourg

Remerciements

Le travail présenté dans ce mémoire a été réalisé au sein du laboratoire de Génétique Moléculaire, Génomique et Microbiologie, UMR7156/CNRS, Université de Strasbourg, sous la direction du Pr. Joseph Schacherer et l'encadrement du Dr. Anne Friedrich dans l'équipe Variation intra-specifique et évolution des génomes.

Tout d'abord j'aimerais remercier tous les membres de mon comité de thèse, Dr. Philippe Marullo, Dr. Romain Koszul, Pr. Cécile Fairhead et Pr. Odile Lecompte pour avoir accepté d'examiner mon travail. Je souhaite par ailleurs remercier plus particulièrement Mme Cécile Fairhead pour avoir aussi participé à mon comité de mi-thèse. Merci également à Mme Odile Lecompte pour l'intérêt à la génomique qu'elle a su me transmettre à travers ces cours.

Je tiens à te remercier particulièrement Anne, bien que malheureusement tu ne sois pas officiellement ma directrice de thèse, tu m'as encadré avec patience durant ces 4 dernières années et pendant les différents stages que j'ai réalisés auparavant. Merci de m'avoir fait confiance il y a maintenant 6 ans en me proposant d'intégrer l'équipe, je n'aurais jamais pu imaginer l'aventure qui allait s'en suivre à ce moment-là ! Ta disponibilité et ton implication dans mon projet de thèse m'ont été (*in fine*) inestimables. Bien évidemment, merci beaucoup à toi aussi Joseph pour avoir été mon directeur de thèse. Tes indications et tes remarques m'ont énormément aidé dans la réalisation de ce travail (et dans ma compréhension du microcosme de la madeleine Strasbourgeoise) et m'ont poussé à aller plus loin !

Merci également à l'équipe enseignante de l'ESBS que j'ai eu la chance d'intégrer dans le cadre de ma dernière année de thèse. Je tiens plus particulièrement à remercier Bruno Kieffer, Yves Nominé et Claude Ling pour le temps qu'ils m'ont consacré lors de la mise en place et la réalisation des différents cours, ainsi que leur soutien lorsque je n'étais pas forcément à mon aise. Cette expérience a été très enrichissante et c'est en grande partie grâce à vous.

Un grand merci à toute l'équipe qui gère le cluster de calcul de l'IBMP sur lequel j'ai réalisé l'ensemble de mes analyses et plus spécialement à Valérie, Magali et David sans lesquels je n'aurais jamais pu aller aussi vite. Merci d'avoir pris le temps de répondre à tous mes mails afin de mettre à disposition des logiciels toujours plus obscurs et d'avoir subi mon utilisation parfois un peu trop excessive des machines. David, que ce soit autour d'un verre à refaire le monde ou juste autour d'une manette, ça a toujours été sympa de passer ces moments avec toi, on se revoit bientôt j'espère !

Bien sûr j'aimerais remercier les différents membres de l'équipe et tout d'abord ceux qui ont su me supporter au bureau. Téo, au-delà de me sauver de mes crises de nerfs sur Illustrator, ça a été vraiment cool de partager des moments en dehors du labo, que ce soit en soirée, sur un terrain de Bad ou sur Sanhok. Bon courage pour ta fin de thèse même si je suis sûr que tu vas gérer ! J'en profite pour faire un big up à la fine équipe du Bad (Antoine, Thomas, Damien et Loïc) avec qui la rage s'est toujours mêlée à la bonne humeur ! Elodie, ça a été un réel plaisir

de partager ce bureau avec toi cette dernière année, mais aussi durant les différents moments qu'on a passé ensemble en dehors. Je n'oublierai pas notre retour nocturne et épique à vélo même s'il a été ponctué de quelques frayeurs. Bon courage pour ta thèse, ça semble long mais ça va plus vite qu'on ne le croit ! Merci Claudine pour avoir répondu avec enthousiasme à mes différentes questions, que ce soit sur les transposons ou non. Omar, bon courage pour ta thèse à toi aussi ! Ta maîtrise des arts mystiques du bash m'étonnera toujours. Merci Sabrina pour ta bonne humeur constante dans le labo et ton aide lorsque j'en avais besoin. Grâce à toi les secrets des traditions nordiques me sont moins méconnus. Finalement, merci également à toi aussi Claudia pour ta gentillesse et ton support dans les moments difficiles malgré mon absence à la paillasse que je regrette un peu du coup !

Je n'oublie pas non plus les « anciens » du labo. Jing, ton talent et ta passion pour la science ont rendu chacune de nos discussions sur le sujet enrichissantes. Je ne te fais plus confiance concernant le menu dans les restaurants chinois par contre ! Christian, ta bonne humeur et ton optimisme ont rendu l'ambiance au labo plus agréable, un jour je suis sûr que l'on arrivera à être d'accord sur la qualité d'un film. Kelle, thanks for everything you've done for my project and for making me discover what an American thanksgiving looks like. Our dance with the crazy Balkan music fan, among other moments, was really fun ! Jack, ça a été vraiment cool de partager le bureau avec toi, ainsi que tous les moments à côté bien sûr. Bon courage à vous tous pour la suite et on se retrouve autour d'une bière quand vous le voulez !

J'aimerais bien évidemment remercier l'ensemble des mes amis qui m'ont soutenu d'une manière ou d'une autre durant ces 4 années. D'abord bien évidemment le club des Jackies. Arnaud et Alex, que ce soit autour d'une bière ou sous les pluies torrentielles dans le Jura, ça a toujours été des bons moments malgré tout. Caro et Julien, merci infiniment pour tout votre soutien lorsque j'étais au bout, surtout à la fin ! Je saurai me racheter ! Merci à toi aussi Pierrot pour ta gentillesse et pour toutes les discussions que l'on a eu, pendant mais aussi avant la thèse, qui m'ont été particulièrement importantes. Merci bien évidemment à vous aussi, Marion, Coralie et Séverine sans qui les moments passés tous ensemble n'auraient pas été pareils. Je tiens aussi à te remercier Maud, ton oreille attentive autour d'une bouteille de vin (bio) m'est toujours précieuse ! A toi aussi Julie, merci d'être toujours là malgré le fait que je ne sois pas toujours dispo, notre virée parisienne de dernière minute a été mythique malgré le fait que l'on n'était pas sûr de s'en sortir vivant. En parlant de Paris, merci à toi aussi Fabien ! Ces dernières années sur les bancs de la fac ont été moins longues avec toi et c'est toujours un plaisir de passer te voir dans la capitale.

Un coucou aussi à tous mes amis d'Abribus et plus particulièrement à la CP et aux membres de l'équipe du samedi qui ont rendu cette expérience unique. Malgré les galères hebdomadaires qui font partie du job, ça a été un réel plaisir de passer ces deux dernières années avec vous et ça m'a permis de penser à autre chose ! Merci particulièrement à toi Charlotte pour m'avoir initié aux tournées, d'abord tranquilles, puis à 250 personnes ou encore sous la neige. J'espère que l'on se reverra sous le soleil mauricien pour changer ! Merci également à toi Sumitra pour ta bonne humeur et ta sympathie constantes qui ont rendu les tournées plus agréables. Et puis

bien sûr, merci à toi Pauline, qui bien loin d'être une ancienne collègue de taff ou une amie d'Abribus, tu es devenue ma sœur des galères. Merci pour ton soutien durant ces dernières années, je t'en suis vraiment reconnaissant !

Avant de finir, un mot pour ma famille. Merci Auré pour t'être rendu dispo quand j'étais au bout, ça m'a fait plaisir. Merci mamie pour tes encouragements et tes propositions d'aller bosser dans ton appartement ventilé lorsque le destin a voulu que la chaleur se rajoute au stress. Merci maman pour ton soutien indéfectible sur tous les sujets malgré le fait que mon travail de thèse a toujours semblé être un mystère pour toi. J'espère que la lecture de ce manuscrit rendra les choses plus claires ! Et puis merci à toi aussi papa pour ton aide, que ce soit durant mes années d'études ou durant ma thèse, bien que ça n'est pas toujours été facile il faut se l'avouer ! Tu as su me transmettre ton intérêt pour la science et cette thèse t'est en partie dédiée.

Finalement, merci à tous les autres, ceux que j'aurais oublié mais aussi toutes les petites rencontres qui m'ont fait progresser et m'ont permis de terminer cette thèse. Cette thèse, en dehors du travail, est aussi une formidable aventure humaine qui m'a particulièrement enrichi. Merci encore à vous tous !

ETAT DE L’ART

Introduction

La génomique, à travers l'exploration et la comparaison de séquences complètes de génomes d'un grand nombre d'espèces, a bouleversé la biologie. Dans les années 90, l'amélioration des technologies de séquençage a encouragé la mise en place de consortia ayant pour but de déchiffrer le génome complet de plusieurs organismes modèles. Les efforts réalisés ont permis la publication des séquences de référence relatives à *Saccharomyces cerevisiae*[1] (1996 – 12 Mb), *Escherichia coli*[2] (1997 – 4,6 Mb), *Caenorhabditis elegans*[3] (1998 – 97 Mb), *Drosophila melanogaster*[4] (2000 – 120 Mb), *Arabidopsis thaliana*[5] (2000 – 115 Mb) ou encore celle de *Mus musculus*[6] (2002 – 3 Gb). En parallèle, un effort considérable a été déployé pour décrypter le génome humain après une dizaine d'années de travail de la communauté internationale (2001 – 3,2 Gb)[7,8]. Ces séquences ont dans un premier temps permis d'avoir une vision plus claire de la structure, de l'organisation et du contenu des génomes dans différents embranchements du vivant. Le séquençage du génome complet d'un nombre toujours plus important d'espèces a rapidement permis la comparaison de leur séquence. Sur des échelles évolutives proches, ces comparaisons permettent une meilleure identification d'éléments génétiques et l'annotation d'éléments fonctionnels déjà caractérisés dans une des espèces alors que la comparaison d'organismes plus éloignés informe sur l'organisation et l'évolution des génomes au cours du temps. Par exemple, la comparaison du génome du yack avec celui du chien, de la vache et de l'Homme a permis l'identification, chez le yack, d'une expansion des gènes impliqués dans le stress hypoxique causé par l'absence d'oxygène et permettant l'adaptation du yack à de hautes altitudes[9].

Néanmoins, le génome d'un individu ne reflète pas toute la complexité de l'histoire évolutive de son espèce. Dans ce cadre, des études de génomique des populations ont été initiées à travers la comparaison des génomes d'un grand nombre d'individus. Originellement proposées en 1998 dans le cadre de l'analyse des maladies humaines[10], ces études sont maintenant devenues incontournables dans l'exploration de l'évolution des génomes. Premièrement, elles permettent de mieux comprendre les relations entre les individus et peuvent donner une vision globale de l'histoire évolutive d'une espèce. Deuxièmement, la génomique des populations permet d'étudier les processus évolutifs impliqués dans la génération et le maintien de la diversité génomique, tels que la sélection, la variation du taux de mutation et de recombinaison. Troisièmement, l'identification des variants génétiques responsables des différences phénotypiques permet d'apporter des éléments de réponse quant aux relations qui lient le génotype et le phénotype. Lors des 10 dernières années, la démocratisation des techniques de séquençage à haut-débit a conduit à un nombre important d'études de génomique des populations. Ces études ont permis d'explorer la diversité intraspécifique sur la base d'un très grand nombre d'individus au sein de l'espèce humaine[11–13], mais aussi au sein d'espèces modèles en génétique telles que *A. thaliana*[14,15], *C. elegans*[16] ou *D. melanogaster*[17]. Chez l'Homme, ces données ont aidé à mieux comprendre les changements adaptatifs ayant eu lieux au cours de notre Histoire[18], tels que notre tolérance à différentes altitudes[19] ou aux produits laitiers[20]. De plus, ces études ont permis d'identifier en partie les bases génétiques impliquées dans la manifestation de certaines maladies humaines. Par exemple, 241 loci susceptibles de

jouer un rôle dans les maladies inflammatoires chroniques de l'intestin incluant la maladie de Crohn ont pu être identifiés[21]. Au sein d'autres espèces, les bases génétiques de traits impliqués dans des procédés industriels ont été étudiés, telles que la croissance de la pomme de terre[22], la taille des grains de riz[23], ou la composition en huile des graines d'*A. thaliana*[24].

Dans le domaine de la génomique, les levures, et plus particulièrement celles appartenant au sous-phylum des Saccharomycotina, sont des organismes de choix. En effet, ces levures possèdent un génome de petite taille (généralement inférieur à 20 Mb) avec peu d'introns et des régions intergéniques courtes. Cette composition rend le séquençage de leur génome facile et relativement peu couteux en comparaison à celui d'autres eucaryotes tels que l'Homme. De plus, ce sont des organismes unicellulaires avec une reproduction clonale, rendant leur manipulation en laboratoire aisée. Enfin, les espèces qui composent ce sous-phylum peuvent être retrouvées dans différents environnements naturels, dans des processus de fermentation (vin, bière, boulangerie)[25] ou en tant que pathogènes opportunistes[26], pouvant amener à différentes évolutions de leur génome. Ces caractéristiques ont amené à l'établissement d'un grand nombre de projets et de consortia internationaux tels que Génolevures visant à l'obtention et l'analyse comparative des génomes de plusieurs espèces de levures. Ces études ont permis d'explorer les différences génomiques à la fois entre des espèces proches ou recouvrant l'ensemble du sous-phylum des Saccharomycotina et à ce jour, les génomes de 218 espèces sont maintenant disponibles (NCBI Genome, TaxID = 147537, au 30 juillet 2018).

Dans une première partie de cette introduction, je décrirai comment les études de génomique comparative établies au sein du sous-phylum des Saccharomycotina ont permis d'obtenir aujourd'hui une vue globale de la diversité interspécifique et des mécanismes moléculaires responsables des différences génomiques au sein de ce sous-phylum. Dans un second temps, je me focaliserai sur l'étude de la variabilité intraspécifique au sein du sous-phylum. Ces travaux se sont majoritairement focalisés sur l'exploration de la variabilité intraspécifique de *S. cerevisiae,* aboutissant récemment à l'étude de plus de 1000 isolats chez cette espèce[27]. Les résultats de ces études, résumés dans une seconde partie, ont permis une dissection précise de son histoire évolutive et des différences génomiques observées entre sous-populations. Ces analyses ont conduit à une meilleure compréhension de l'impact de l'environnement et des processus de fermentation sur l'évolution des génomes de cette espèce. De plus, l'importante quantité de données de séquençage générée a permis l'établissement de plusieurs études visant à identifier avec précision les variants génétiques responsables de certains phénotypes. En parallèle, l'étude de la variabilité intraspécifique au sein d'espèces non-modèles de levures suscite un intérêt croissant. Ces analyses, discutées dans une troisième et dernière partie de cette introduction, bien que limitées à une dizaine d'espèces et à l'étude de collections contenant un nombre réduit d'isolats (généralement inférieur à 50) en comparaison à *S. cerevisiae*, ont déjà donné une première vue des différences évolutives des espèces au sein des Saccharomycotina.

Exploration de la diversité interspécifique chez les levures

L'obtention de la séquence du génome de *S. cerevisiae* en 1996, après des années d'effort de la communauté internationale, a donné un point de départ dans l'exploration de la diversité génétique à l'échelle du génome complet chez les levures. Depuis, les séquences complètes des génomes de plus de 200 espèces de levures ont été générées. L'obtention de ces séquences a permis d'établir un nombre conséquent d'études comparatives permettant à la fois l'obtention d'une vue globale de l'évolution des génomes entre les différentes espèces, ainsi que l'examen précis des différences dans certaines fonctions cellulaires telle que le changement de signe sexuel chez la levure. Dans cette partie, je me focaliserai sur l'intérêt de ces études dans l'exploration de la variabilité interspécifique au sein d'espèces du sous-phylum des Saccharomycotina. Dans ce cadre, une description globale des principaux groupes retrouvés dans ce sous-phylum ainsi que des différences observées dans la constitution des génomes des différentes espèces seront décrites. Dans un dernier point, j'évoquerai les différents projets visant à l'obtention d'une vue plus large de la variabilité interspécifique.

Résolution moléculaire de la classification des espèces de levures

Le séquençage de plusieurs espèces permet d'établir une classification plus évidente des levures. La classification de ces organismes a longtemps reposé sur un ensemble de caractères morphologiques et métaboliques, basés essentiellement sur des propriétés fermentatives et l'assimilation de différentes sources de carbone[28]. Une telle approche présente cependant des biais, du fait par exemple de la présence de caractères phénotypiques similaires entre plusieurs espèces mais acquis indépendamment, faussant les relations de parenté inférées. Dans ce cadre, la résolution du génome complet de plusieurs espèces et la confrontation de leurs séquences permettent de retracer avec précision leurs liens de parenté. L'analyse successive de ces relations a permis l'identification de 4 groupes majeurs au sein du sous-phylum des Saccharomycotina (Figure 1) : (i) les *Saccharomycetaceae,* comportant notamment *S. cerevisiae,* est le groupe le plus étudié jusqu'à présent ; (ii) le clade « CTG », un groupe diversifié ayant en commun une modification de leur code génétique ; (iii) le groupe des méthylotrophes récemment mis en évidence et présentant une architecture génétique intermédiaire ; (iv) les espèces dites basales du sous-phylum, moins étudiées mais qui montrent une importante variabilité génétique.

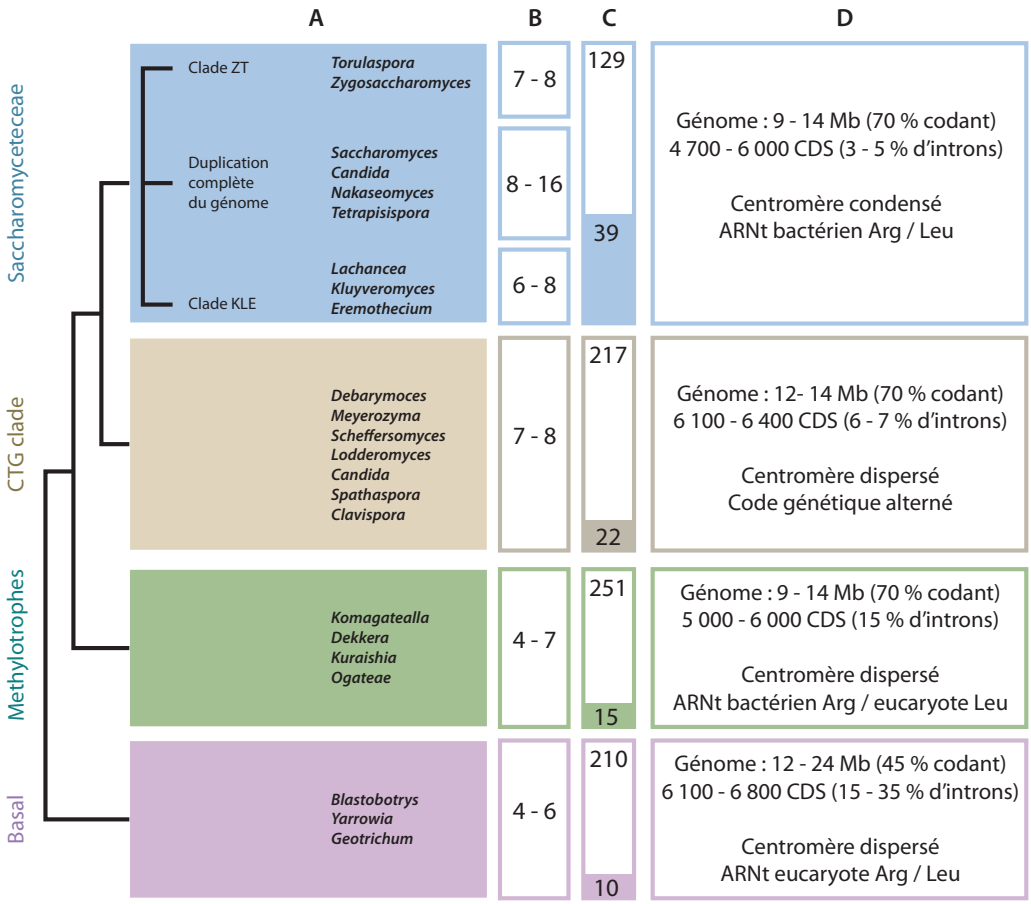


Figure 1. Caractéristiques principales des différents groupes au sein des Saccharomycotina. Pour chaque groupe est présenté : (A) une liste non exhaustive des genres (B) la variation du nombre de chromosomes, (C) le nombre d'espèces séquencées par rapport au nombre d'espèces identifiées (données de 2015[29]), (D) les caractéristiques principales. Adaptée de Dujon et Louis, 2017.

Ces différents groupes se distinguent par la présence de signatures génomiques spécifiques[30] (Figure 1.D). Alors, que les génomes des *Saccharomycetaceae* sont petits et compacts (9 à 14 Mb avec environ 70 % de séquences codantes), certaines espèces du groupe basal, telle que *Yarrowia lipolytica,* présentent un génome bien plus grand et moins condensé (20.5 Mb avec 46 % de séquences codantes)[31]. Ces groupes se différencient aussi par des spécificités dans la constitution de certaines régions génomiques ou dans des fonctions vitales à la cellule, notamment dans la machinerie traductionnelle et la constitution des centromères. Par exemple, les *Saccharomycetaceae* présentent un centromère condensé en un seul locus et la structure de l'anticodon des ARNt affectés à l'arginine et à la leucine est similaire à celui des bactéries[32]. Au contraire, les souches basales possèdent un centromère dispersé composé notamment d'éléments transposables, ainsi qu'une structure d'ARNt similaire à celle des eucaryotes. De manière intéressante, le groupe des méthylotrophes, composé d'espèces provenant de plusieurs

clades, présente une évolution intermédiaire entre les deux groupes précédemment cités. Les génomes de ce groupe sont de petite taille (entre 9 et 13 Mb) mais possèdent un plus grand nombre d'introns que ceux du groupe des *Saccharomycetaceae*. Par ailleurs, tandis que les ARNt affectés à l'arginine ont une structure similaire à celle des *Saccharomycetaceae*, ceux liés à la leucine se caractérisent par un mode de fonctionnement identique au groupe basal. Finalement, l'analyse des espèces au sein du clade CTG a montré que celles-ci présentent aussi des signatures génomiques intermédiaires : leurs génomes sont petits et compacts mais possèdent un centromère dispersé similaire aux souches basales et méthylotrophes. Ce groupe se distingue cependant par l'utilisation d'un code génétique alternatif dans lequel le codon CUG est décodé en sérine plutôt qu'en leucine, grâce à l'utilisation d'un ARNt additionnel au sein de ces espèces[33].

Au sein des Saccharomycotina, des modes de vie similaires sont retrouvés entre différents clades, suggérant l'homoplasie de plusieurs traits au cours de l'histoire évolutive des espèces. Par exemple, *Candida albicans* et *Candida glabrata* sont deux espèces commensales de l'Homme, retrouvées respectivement dans le clade CTG et celui des *Nakaseomyces*, présent au sien du groupe des *Saccharomycetaceae*, suggérant l'acquisition indépendante des caractères pathogènes chez ces espèces. L'analyse comparative des génomes des 6 espèces contenues dans le clade des *Nakaseomyces* a montré que les espèces pathogènes sont polyphylétiques et sont retrouvées dans des groupes contenant aussi des espèces isolées dans des environnements naturels[34]. De manière intéressante, une conclusion similaire a pu être obtenue dans l'étude comparative des génomes de 8 espèces au sein du clade des CTG[35] . Par ailleurs, l'isolement récent de 3 souches de *C. albicans* sur des chênes[36] ou de plusieurs souches de *C. glabrata* sur des grains de café[37] indiquent que ces espèces, bien que retrouvées essentiellement en tant que pathogènes, ont colonisé d'autres habitats au cours de leur histoire évolutive. *C. albicans* et *C. glabrata* se distinguent sur plusieurs plans au niveau génomique. Par exemple, alors que *C. albicans* est retrouvée essentiellement en tant que diploïde, les souches de *C. glabrata* privilégient une forme haploïde. De plus, leurs génomes sont très divergents : 1557 gènes de *C. glabrata* n'ont pas d'orthologue chez *C. albicans* et l'inverse est aussi retrouvé pour 2257 gènes de *C. albicans*[38]. Ces différences font de ces espèces des modèles d'intérêt pour mieux comprendre les mécanismes évolutifs responsables de l'adaptation des espèces à ce mode de vie.

De grandes variations dans l'architecture des génomes ont aussi été observées au sein des différents groupes. Par exemple, l'ancêtre commun d'un sous-ensemble de souches au sein du groupe des *Saccharomycetaceae* dont fait partie *S. cerevisiae,* a subi une duplication complète de son génome[39]. Cette modification majeure de la structure du génome, retrouvée aussi chez les Ciliés[40] ou les plantes[41], se caractérise par un dédoublement du jeu de chromosomes suivi rapidement par des délétions successives de larges portions du génome. Cet événement a amené à plusieurs spécificités au sein de ce clade, notamment à la présence d'un nombre plus variable de chromosomes dans ce groupe (6 à 16) ainsi qu'à la présence de gènes dits ohnologues, résultant de la duplication complète du génome mais dont les deux copies ont été conservées. L'exploration des génomes au sein de ce groupe a par ailleurs révélé des différences dans le

mode de fonctionnement de certains mécanismes moléculaires. Par exemple, l'étude du changement de signe sexuel chez les levures a permis de mettre en évidence différents procédés selon les espèces considérées[42]. Chez *S. cerevisiae*, l'expression d'un signe sexuel par rapport à un autre dépend de la version allélique présente au sein d'un locus spécifique : le locus *MAT*. La conversion d'un type sexuel repose sur la présence de deux versions alléliques silencieuses de ce locus (désignées *HML* et *HMR*). Le clivage du locus actif induit par une protéine (HO) suivi de la réparation de cette région par homologie avec comme modèle une des copies silencieuses permet le changement du signe sexuel. Récemment, la dissection de ce mécanisme chez *K. lactis*, une espèce appartenant au groupe des *Saccharomycetaceae* mais inclus dans le clade des *Kluyveromyces* n'ayant pas subie la duplication complète du génome (Figure 1), a révélé un fonctionnement très différent[43]. Dans cette espèce, le génome ne contient pas le gène codant la protéine HO et le changement de signe sexuel est assuré par deux éléments transposables qui permettent chacun d'activer l'une ou l'autre version allélique[44]. Dans d'autres espèces, cette capacité semble avoir été perdue. Par exemple, bien que des homologues des deux cassettes *MAT* et de la protéine HO de *S. cerevisiae* aient été retrouvés chez *C. glabrata*, une reproduction sexuée n'a jamais été mise en évidence en laboratoire chez cette espèce[45]. Une étude se basant sur l'induction de la version allélique de la protéine HO de *S. cerevisiae* chez *C. glabrata* a montré que celle-ci induit une forte létalité (taux de survie de 0,1 %) dans cette dernière[46]. Cette létalité résulterait notamment de l'absence du gène *SIR1* chez *C. glabrata* impliqué dans la régulation des cassettes HML et HMR chez *S. cerevisiae*, et dont l'absence amène à un grand nombre de coupures de ces cassettes. L'analyse d'une population de *C. glabrata* a cependant montré la présence de souches possédant les deux signes sexuels[47] et des traces de recombinaison ont été identifiées au sein de cette espèce[48], suggérant qu'une reproduction sexuelle, bien que rare, soit possible dans certaines conditions.

Bases moléculaires de l'évolution des génomes

Au cours du temps, plusieurs mécanismes évolutifs sont responsables de l'évolution des génomes d'une espèce. Ces mécanismes, tels que la génération de variants structuraux (insertions, délétions, inversions, translocations) ou l'accumulation de mutations ponctuelles (Single Nucleotide Variants - SNV), impactent les génomes à différentes échelles. Alors que les premiers modifient de manière importante la structure des génomes, le second amène à une divergence progressive de la séquence des génomes et ainsi des séquences protéiques entre les espèces. Dans ce cadre, l'étude de la conservation de l'ordre des gènes (synténie) et l'analyse de la conservation des séquences homologues permet une quantification de la divergence génétique entre les différentes espèces.

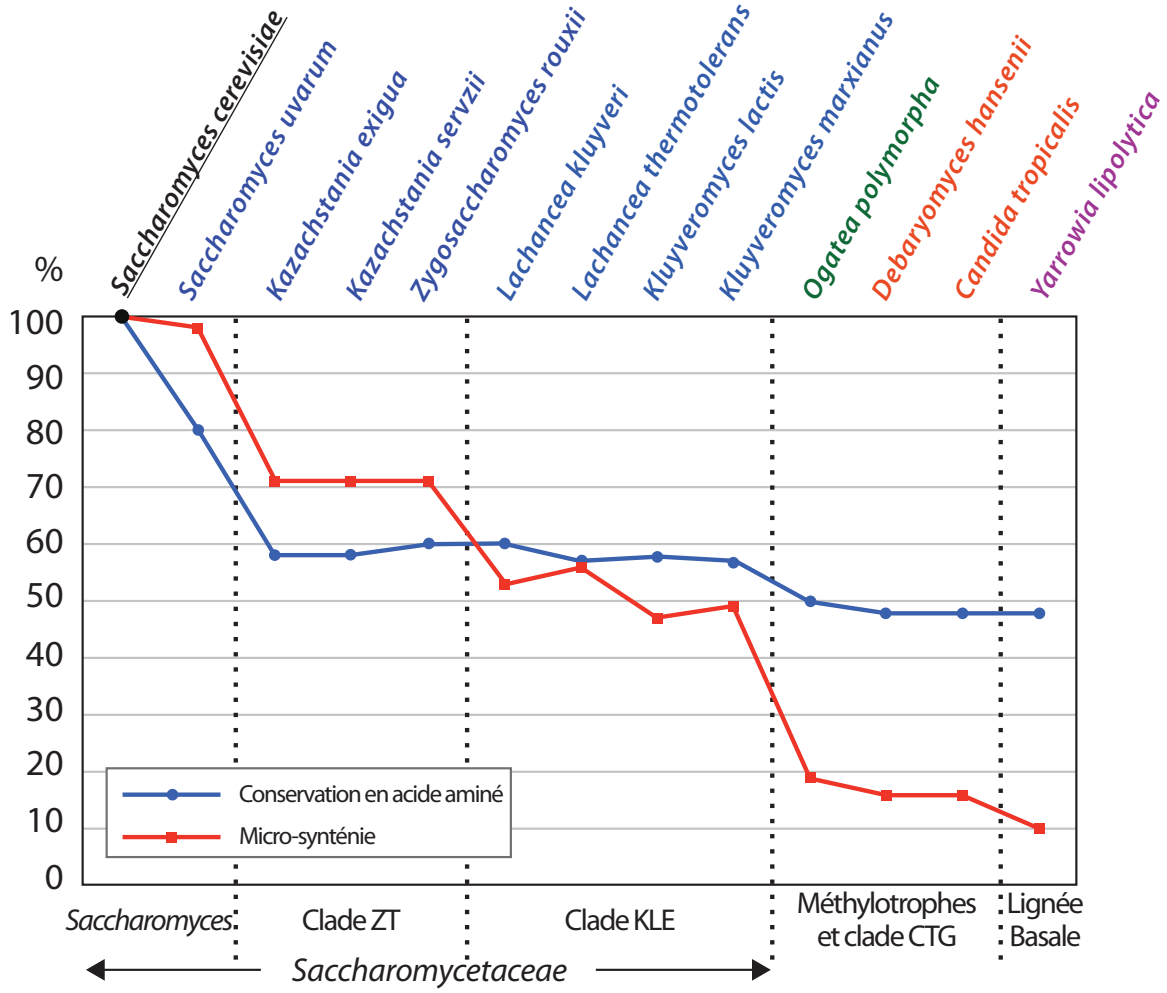


Figure 2 : Conservation des séquences protéiques et de la synténie de plusieurs espèces en comparaison à *S. cerevisiae*. Figure adaptée de Dujon et Louis, 2017.

La comparaison des premiers génomes séquencés a rapidement montré une diversité génétique particulièrement importante au sein du sous-phylum des Saccharomycotina. Par exemple, l'analyse comparative des séquences codantes entre *S. cerevisiae* et de l'espèce proche *S. uvarum* a mis en évidence une divergence nucléotidique d'environ 33 %, bien au-delà de celle observée entre l'Homme et le chimpanzé (1,8 %). Cette divergence augmente significativement au sein même des *Saccharomycetaceae* avec un pourcentage de conservation en acides aminés entre *S. cerevisiae* et *C. glabrata* de 65 %, similaire à celui observé entre l'Homme et certains animaux marins de la sous-famille des *Tetraodontinae*[31], et atteint 50 % entre les groupes les plus éloignés (*S. cerevisiae* et *Y. lipolytica*), une divergence représentative de celle observable au sein des Chordés[49] (Figure 2). Par ailleurs, la diversité génétique au sein de chaque genre n'est pas du même ordre. Récemment, la comparaison des génomes de 34 espèces au sein des *Saccharomycetaceae* a ainsi permis d'identifier et de comparer la divergence entre les génomes au sein des différents genres composant ce groupe[50] (Figure 3). Alors que les 4 espèces appartenant au genre des *Saccharomyces* (dont fait partie *S. cerevisiae*) présentent une divergence de l'ordre de 15 % entre elles, les espèces du genre des *Lachancea,*

dont aucun ancêtre n'a subi la duplication complète du génome, présentent des divergences allant de 17 à 31 %. Cette tendance est partiellement retrouvée dans l'analyse de la synténie entre les espèces. Alors qu'un grand nombre de gènes sont partagés et que l'ordre de ceux-ci est essentiellement conservé au sein des *Saccharomyces*[51], la taille des blocs de synténie diminue rapidement au sein des clades plus éloignés pour lesquels l'accumulation de points de cassure chromosomique a entrainé la présence d'un nombre plus important de blocs contenant moins de gènes[50] (Figure 2). La comparaison d'espèces éloignées ne permet pas de retrouver des traces de synténie, indiquant que le nombre total de réarrangements dans l'évolution des génomes depuis le dernier ancêtre commun approche ou dépasse le nombre de gènes dans ces génomes. Ces résultats indiquent une importante plasticité des génomes de levures. L'analyse comparative des blocs de synténie entre 18 espèces de levures et entre 13 espèces du sous-embranchement des vertébrés a par ailleurs montré que le taux de réarrangements chromosomiques est environ 50 fois plus élevé dans les génomes de levures que dans les génomes de vertébrés[52]. L'analyse approfondie des répertoires de gènes dans le genre des *Lachancea* a récemment permis l'étude de l'impact de ces réarrangements dans l'évolution des différentes espèces qui le composent[50]. Cette étude a mis en évidence 1686 évènements impliqués dans l'expansion de ces répertoires et dont la majorité (1503) correspond à la duplication de gènes. A l'opposé, 1018 pseudo-gènes, correspondant à des gènes dont la fonction a été perdue par l'apparition de mutations ponctuelles ou de petites délétions, ainsi que 929 pertes de gènes ont pu être mis en évidence. Par ailleurs, ces réarrangements non-balancés, impliquant un gain ou une perte d'une partie du génome, sont présents de manière bien plus fréquente que les réarrangements balancés comme les inversions ou les translocations (423 évènements). Ces derniers jouent cependant un rôle non négligeable dans l'évolution du répertoire des gènes. En effet, en modifiant les gènes au niveau de leur point de rupture, ces réarrangements seraient responsables de la perte de 14 % des gènes dans ces espèces. Ces résultats soulignent l'importante variabilité du répertoire des gènes, notamment par des évènements de duplication, au cours de l'évolution de ces espèces.

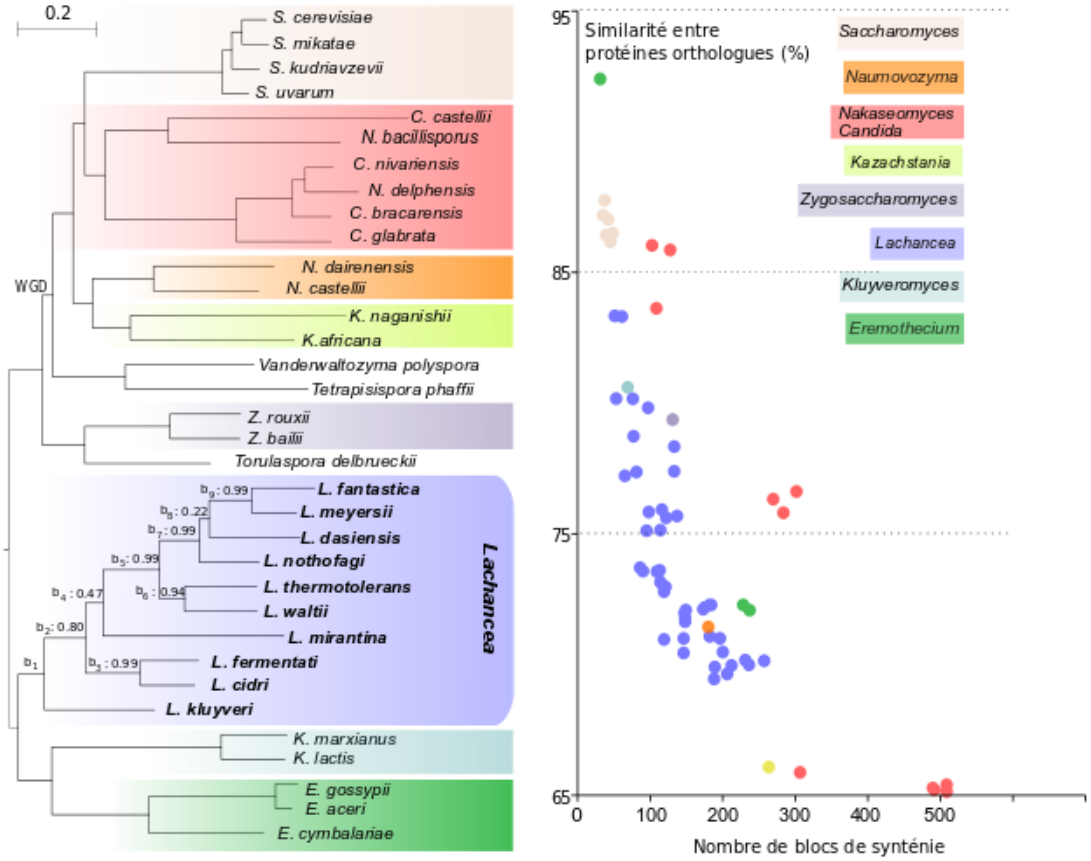


Figure 3. Distribution du taux de similarité entre protéines orthologues et du nombre de blocs de synténie entre les espèces pour chaque clade appartenant au groupe des *Saccharomycetaceae*[50]. Adaptée de Vakirlis *et al.* (2016).

L'adaptation à de nouvelles conditions repose notamment sur l'incorporation de nouveaux gènes ou la combinaison de plusieurs génomes en un hybride. Plusieurs mécanismes sont impliqués dans cette dynamique. D'une part, la formation de gène *de novo* au cours de l'évolution, longtemps considérée comme improbable, se révèle être finalement un mécanisme ayant eu un rôle dans l'évolution du répertoire du gène. Par exemple, le gène *BSC4* codant une enzyme impliquée dans la réparation de l'ADN ou le gène *MDF1* codant une protéine impliquée dans la régulation de l'assimilation du glucose ont été identifiés comme ayant émergé lors de l'évolution de *S. cerevisiae*[53,54]. Récemment, l'étude de 15 espèces comprises dans les genres *Saccharomyces* et *Lachancea* a permis l'identification de 703 candidats potentiels, suggérant l'apparition d'un nombre non négligeable de ces gènes au cours de l'évolution des espèces[55]. De manière générale, ces gènes sont cependant de plus petite taille et subissent une pression de sélection plus faible en comparaison au reste du génome. Un autre mécanisme d'adaptation repose sur l'hybridation de génomes de différentes espèces, pouvant résulter en l'émergence d'une nouvelle espèce en tant que telle. Par exemple, *Saccharomyces pastorianus*, une espèce d'intérêt dans les processus de fermentation de par sa résistance aux températures basses, résulte d'un croisement entre *S. cerevisiae* et de l'espèce cryo-tolérante *Saccharomyces eubayanus*[56,57]. D'autres génomes allopolyploïdes impliquant davantage d'espèces telles que

Saccharomyces carlsbergensis[58], résultant du croisement de *S. cerevisiae, S. uvarum* et *S. eubayanus*, ou au sein d'autres groupes, tels que *Millerozyma sorbitophila*[59] ou *Candida orthopsilosis*[60,61] présents dans le groupe CTG ont été identifiés. L'évolution des génomes de ces hybrides reste cependant encore globalement méconnue bien que certaines études aient donné des premières indications à ce sujet. Par exemple, l'analyse du génome de l'espèce osmotolérante *Pichia sorbitophila*, pour laquelle les génomes des deux parents diffèrent de 12 à 15 % dans leur séquence nucléotidique, a montré des traces importantes de perte d'hétérozygotie[59]. Ce mécanisme a modifié de manière importante le génome de chaque parent : 40,3 % du génome de cet hybride est affecté et certains chromosomes ont subi des pertes d'hétérozygotie sur l'ensemble de leur séquence. De manière intéressante, cette trajectoire évolutive est aussi retrouvée chez d'autres hybrides tels que *S. pastorianus*[62], suggérant que ce mécanisme jouerait un rôle important dans l'évolution de plusieurs espèces résultant d'un phénomène d'hybridation. Dans certains cas, la recombinaison entre les génomes parentaux suivi de la transmission de ces versions alléliques dans une des populations d'origine permet l'incorporation de ces gènes d'une espèce à une autre. Les études de génomique des populations, par l'identification de ces introgressions de manière systématique pour l'ensemble des individus, offrent l'occasion de mieux quantifier ce phénomène. L'impact de ces variants dans l'évolution des génomes de différentes espèces sera à cet égard discuté dans le chapitre 3 cette introduction. De manière surprenante, l'incorporation de séquences génomiques d'organismes très divergents par transferts horizontaux a aussi été observée. Par exemple, le gène *URA1* nécessaire à la propagation des levures en condition strictement anaérobique a ainsi été acquis par transfert horizontal d'une bactérie[63]. Ces phénomènes sont cependant rares, du fait de l'importante divergence entre les génomes parentaux.

Vers une exploration plus large de la diversité interspécifique

Le séquençage progressif et l'étude du génome d'un nombre important d'espèces a élargi de manière conséquente notre vision de la diversité génétique au sein du sous-phylum des Saccharomycotina[30]. L'étude de cette diversité s'est cependant longtemps focalisée sur les *Saccharomycetaceae* et la variabilité interspécifique au sein des autres groupes reste encore grandement inexplorée[29] (Figure 1). En effet, alors que les génomes de 39 des 129 espèces connues dans ce groupe ont été séquencés (30 %), cette proportion est réduite à 5 à 10 % pour les espèces des autres groupes (données de 2015). De par l'amélioration des technologies de séquençage et la diminution des coûts nécessaires à l'obtention d'un génome, l'étude d'autres espèces suscite un intérêt croissant. En effet, l'exploration d'espèces plus éloignées de *S. cerevisiae* permettrait d'obtenir une vue plus large de l'évolution des espèces au sein du sous-phylum. Dans ce cadre, plusieurs projets se sont focalisés sur l'obtention et la comparaison d'un nombre conséquent de génomes. Ainsi, le projet « 1000 fungal genomes » (1000.fungalgenomes.org) a déjà permis le séquençage de plus de 1000 génomes présents dans ce groupe. Plus récemment, le projet « Y1000+ » (y1000plus.wei.wisc.edu), toujours en cours de réalisation, vise à mieux appréhender la diversité génétique au sein du sous-phylum des Saccharomycotina, notamment par le séquençage de génomes au sein des groupes les moins étudiés.

Dans un premier temps, les ressources générées par le projet « Y1000+ » permettront de mieux caractériser les relations phylogénétiques entre les différentes espèces du sous-phylum. Dans ce cadre, une étude s'est basée sur l'exploration des relations d'orthologie entre 96 espèces afin de produire une classification recouvrant l'ensemble du sous-phylum des *Saccharomycotina*[64]. Cette analyse a d'abord permis l'identification de 1233 groupes d'orthologues, présents de manière variable entre les espèces. Alors que 88 espèces partagent plus de 1000 de ces groupes (81 %), certaines espèces contiennent moins de 750 (60,5 %) d'entre eux. Ces données ont par la suite été utilisées dans différentes approches phylogénétiques (coalescence et concaténation), basées à la fois sur les séquences protéiques et nucléiques des gènes, permettant la détermination des relations phylogénétiques pour l'ensemble des espèces. D'un point de vue technique, cette étude démontre la faisabilité d'obtenir une classification robuste des espèces présentes sur l'ensemble du sous-phylum. L'adaptation de cette méthode à un plus grand nombre d'espèces permettra à terme d'établir avec précision les relations phylogénétiques du sous-phylum des Saccharomycotina. Cette ressource sera utile pour l'étude de l'évolution des différentes espèces, permettant par exemple l'identification d'espèces phylogénétiquement éloignées mais présentant une adaptation aux mêmes environnements.

En parallèle, une exploration globale de la diversité phénotypique au sein du sous-phylum a été réalisée. Dans cette étude, une analyse globale des caractères phénotypiques a été réalisée par la mesure du taux de croissance de 240 espèces sur 48 conditions ainsi que la collecte de données phénotypiques, telles que la croissance en présence de différents composés carbonés ou à différentes températures, relatives à 784 espèces[65]. Dans un premier temps, ces données ont permis la comparaison globale des données dans chaque condition. Cette approche a permis d'identifier les associations positives ou négatives entre deux conditions, correspondant à la capacité des espèces à pouvoir dégrader ou non ces composants (Figure 4). De telles associations sont particulièrement retrouvées entre les sources de carbone et suggèrent la présence de gènes ayant un effet pléiotropique dans la dégradation de ces composés. A l'opposé, les composés impliqués dans les processus de fermentation sont négativement corrélés avec la dégradation des pentoses, suggérant une interaction négative dans la mise en place simultanée de ces deux métabolismes. Cet aspect est particulièrement important dans certains procédés industriels tels que la production de biocarburant pour laquelle les sucres et plus spécifiquement les pentoses sont particulièrement présents. Dans un second temps, la comparaison entre les modes de vie des différentes espèces et les conditions phénotypiques a été réalisée. De manière intéressante, les levures pathogènes présentent une tolérance importante à de hautes températures (37°) et sont capables de dégrader un nombre plus faible de sources de carbone en comparaison aux autres espèces. Ce résultat peut notamment être expliqué par l'absence récurrente de sources de carbone rencontrées lors de la colonisation d'organismes vivants et notamment des mammifères.

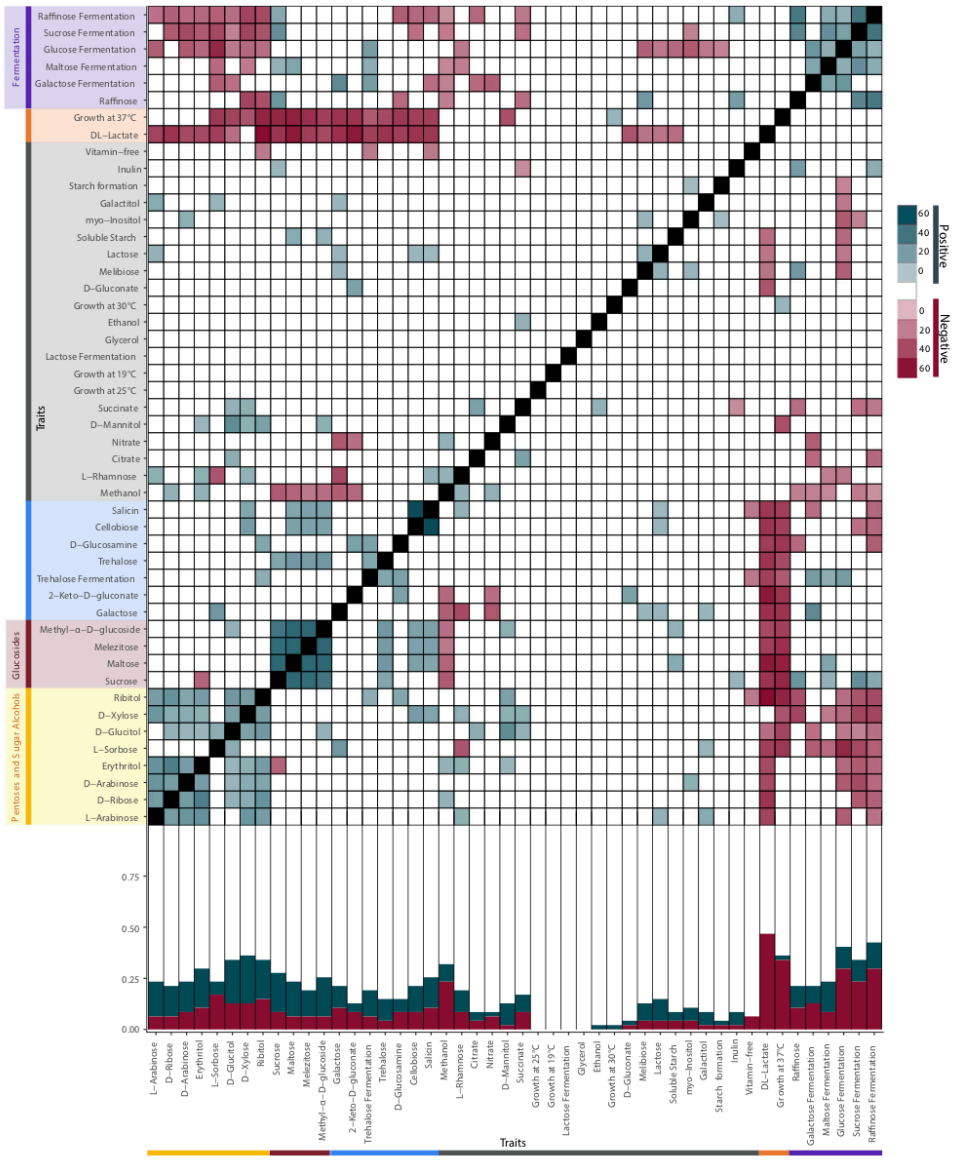


Figure 4. Distribution des associations négatives (rouge) et positives (bleu) ou non significatives (blanc) entre les différentes conditions dans l'analyse de données phénotypiques chez 240 espèces[65]. En bas du graphique est représentée la proportion de ces associations pour chacun des traits. En haut est représentée l'association deux à deux de chaque phénotype, la saturation des couleurs est liée à l'importance de ces associations. Adaptée de Opulente *et al.* (2018).

Finalement, l'exploration de la diversité génétique au sein de clades jusqu'alors peu explorés permet d'identifier de nouveaux évènements adaptatifs dans l'évolution de certaines espèces de levures. Dans ce cadre, l'analyse du métabolisme impliqué dans la dégradation des sources de carbone au sein des clades *Wickerhamiella* et *Starmerella* (W/S) s'est révélée particulièrement intéressante. Ces deux clades, proches du groupe des *Saccharomycetaceae*, sont composés d'espèces présentes majoritairement sur les fleurs et sont capables de dégrader rapidement le fructose, un composant majeur de ces environnements. L'adaptation à cette niche particulière

repose notamment sur l'intégration du gène *FFZ1* codant un transporteur permettant la dégradation rapide de cette source de carbone. Celui-ci, perdu dans l'ancêtre commun des Saccharomycotina, a été réintégré dans l'ancêtre commun des clades W/S à partir d'un transfert horizontal provenant d'une espèce d'un autre sous-phylum appartenant au sous-embranchement des *Pezizomycotina*[66]. Récemment, la comparaison globale des génomes de plusieurs espèces de ces deux clades a permis l'identification de 52 gènes résultant d'un transfert horizontal de bactéries[67] et un enrichissement pour des fonctions liées à la fermentation alcoolique a été retrouvé. De manière intéressante, plusieurs gènes ancestraux impliqués dans ce type de métabolisme ont été perdus au cours de l'évolution de l'espèce, indiquant la ré-acquisition de ces fonctions par transfert horizontal. Par exemple, la version originelle du gène *ADH1* codant une déshydrogénase impliquée dans la réduction de l'acétaldéhyde en éthanol est absente dans ces espèces mais a été réintégrée via une version bactérienne. Par ailleurs, ce gène est absent chez certaines espèces proches de ces clades, confirmant la perte puis la réintégration de ce gène au cours de l'évolution de ces espèces.

Étude de génomique des populations au sein de *S. cerevisiae*

L'analyse comparative inter-espèce a grandement enrichi nos connaissances relatives à la diversité génétique au sein du sous-phylum des Saccharomycotina. Néanmoins, le génome d'un seul individu d'une espèce ne reflète pas la variabilité génétique présente au sein de celle-ci. Dans ce cadre, les études de génomique des populations, à travers la comparaison des génomes de plusieurs individus, se révèlent être particulièrement intéressantes. En effet, l'exploration de la diversité intraspécifique permet l'étude de l'histoire évolutive et notamment des mouvements de migration au sein d'une espèce. Par ailleurs, l'analyse de la diversité génétique le long du génome et entre sous-populations permet de mieux comprendre comment la diversité génétique est générée et maintenue. Enfin, l'obtention d'un catalogue exhaustif de variants génétiques, en comparaison avec des données phénotypiques, permet de mieux comprendre la relation qui lie le génotype et le phénotype. Dans l'histoire de la génomique des levures, une attention particulière a été portée à l'espèce modèle *S. cerevisiae*. La caractérisation précise de cette espèce, notamment associée à des processus de fermentation (vin, bière, boulangerie), ainsi que l'isolement d'un grand nombre de souches a rapidement poussé à l'exploration de la variabilité génétique présente au sein de *S. cerevisiae*. Dans ce cadre, plusieurs études de génomique des populations ont été initiées lors de la dernière décennie[27,68–79], aboutissant récemment à l'analyse comparative du génome de plus de 1000 isolats[27]. Ces études offrent une vue sans précédent des processus évolutifs et de leurs impacts sur la diversité génétique au cours de l'évolution de *S. cerevisiae*.

Reconstruction de l'histoire évolutive de *S. cerevisiae*

Les relations entre les isolats peuvent être élucidées à travers la comparaison des variants génétiques d'un grand nombre d'individus. D'une part, les liens de parenté entre les individus et leurs ancêtres communs peuvent être inférés, représentés par un arbre phylogénétique où la distance entre les nœuds est relative à la divergence génétique entre les individus. Dans un deuxième temps, la population peut être divisée en un nombre fixe de groupes représentatifs de génomes ancestraux et la proportion de ces génomes est déterminée individuellement pour l'ensemble des isolats. La confrontation de ces résultats avec les origines écologiques et géographiques permet de tester l'influence de ces facteurs sur la diversité génétique au sein de la population. De telles analyses ont révélé une fragmentation importante des isolats de *S. cerevisiae* en différentes sous-populations[27,70,74–78] présentant des compositions variables : certains groupes présentent un enrichissement important pour les souches impliquées dans un même processus de fermentation alors que d'autres groupes contiennent des isolats aux origines variées[27]. Ces derniers présentent un nombre plus important de souches dites mosaïques, résultant du croisement et de la recombinaison de génomes provenant de lignées pures (Figure 5). De manière intéressante, les souches cliniques présentent un enrichissement pour ces génomes mosaïques[27,68,70,74]. Ce résultat indique que ces isolats, retrouvés chez des patients immunodéprimés, ne dérivent pas d'un ancêtre unique mais résultent plutôt de la colonisation successive de différentes souches naturelles d'origines écologiques variées. À l'opposé, les sous-populations impliquées dans des processus de fermentation, forment des groupes

génétiquement distincts avec peu ou pas de souches mosaïques. Par ailleurs, les souches de bières sont polyphylétiques et sont ainsi retrouvées dans 3 sous-populations distinctes, indiquant que l'utilisation de *S. cerevisiae* dans le brassage de bière résulte de plusieurs évènements de domestication indépendants.

Figure 5. Structure retrouvée dans l'analyse des 1011 génomes obtenue en utilisant un nombre de populations K allant de 5 à 12[27]. Adaptée de Peter *et al.* (2018).

De par l'absence de fossile et le peu de relevés historiques dans l'utilisation des levures, l'origine de *S. cerevisiae* a longtemps été méconnue. Récemment, la découverte de souches chinoises présentant une forte divergence génétique (1,1 %) par rapport aux souches déjà connues a suggéré une origine asiatique de cette espèce[80]. Cette hypothèse est soutenue par l'isolement de plusieurs espèces proches de *S. cerevisiae* en Asie de l'Est, comme *S. mikatae*[81] et *S. arboricola*[82], ainsi que par l'importante diversité génétique retrouvée chez les souches japonaises de *S. kudriavzevii*[83]. L'exploration des génomes de plus de 1000 souches au sein de l'espèce a conforté cette théorie[27]. Tout d'abord, la comparaison d'assemblages provenant de souches représentatives des différentes sous-populations ainsi que d'espèces proches de *S. cerevisiae* montre une divergence plus faible entre les souches isolées en Asie de l'Est comparées aux autres souches (Figure 6). Par ailleurs, l'analyse en composante principale de l'ensemble des souches permet la distinction entre les souches asiatiques et les autres isolats. Plus récemment, l'exploration de 266 isolats provenant majoritairement d'Asie a confirmé cette analyse[71]. L'étude de ces génomes suggèrent par ailleurs que la domestication de *S. cerevisiae* aurait aussi débutée dans cette région. En effet, plusieurs souches chinoises peuvent être retrouvées dans la sous-population liée à la production de vin et des régions génomiques résultant de transferts horizontaux peuvent être retrouvées communément dans les souches de vin et dans les souches naturelles chinoises. L'ensemble de ces résultats conforte l'idée que *S. cerevisiae* et les espèces qui lui sont proches, ont pour origine l'Asie de l'Est, suivi d'une migration importante de *S. cerevisiae* à travers le monde permettant la diversification de l'espèce.

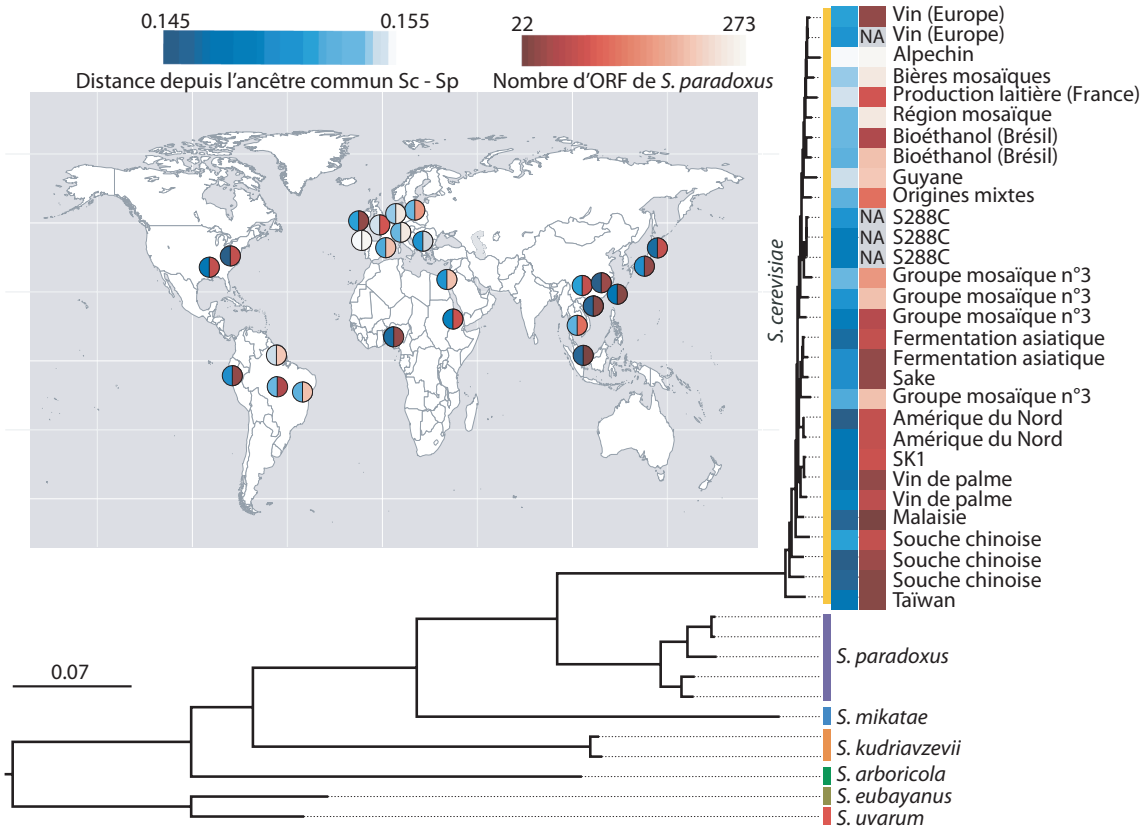


Figure 6. Origine chinoise de *S. cerevisiae*. Arbre phylogénétique basé sur l'alignement de 2018 gènes conservés entre les espèces. Les couleurs reflètent la distance depuis le dernier ancêtre commun entre *S. cerevisiae* (Sc) et *S. paradoxus* (Sp) (blanc - bleu) et le nombre d'ORF introgressées de *S. paradoxus* (blanc – rouge). La carte représente les origines géographiques des souches. Adaptée de Peter *et al.* (2018).

Variabilité génétique et génomique au sein de l'espèce

Pour la première fois, l'exploration de la diversité génétique de 1011 isolats de *S. cerevisiae* a permis d'obtenir une vue exhaustive de la distribution de la variabilité nucléotidique au sein d'une espèce de levure[27]. Au total, 1625809 variants nucléotidiques (SNP) et 125701 insertions ou délétions de petite taille ont ainsi pu être caractérisés dans cette population. De manière intéressante, les variants nucléotidiques sont très peu partagés entre les individus. En effet, 31,3 % de ces SNP sont des singletons, c'est-à-dire présents chez un seul individu et presque 93 % des sites polymorphiques présentent une fréquence allélique inférieure à 5 %. Cette distribution est fortement impactée par le nombre important de souches provenant de niche viticole dans la population globale et pour lesquelles cette tendance est particulièrement forte. La présence d'un grand nombre de variants génétiques rares au sein de cette sous-population pourrait notamment résulter d'un « bottleneck » suivi d'une expansion récente de la population. Cette tendance est cependant observée pour une majorité des souches. Ces résultats, couplés au

continuum observé dans l'analyse phylogénétique suggèrent ainsi une expansion globale de la population amenant à l'acquisition de nouveaux variants génétiques.

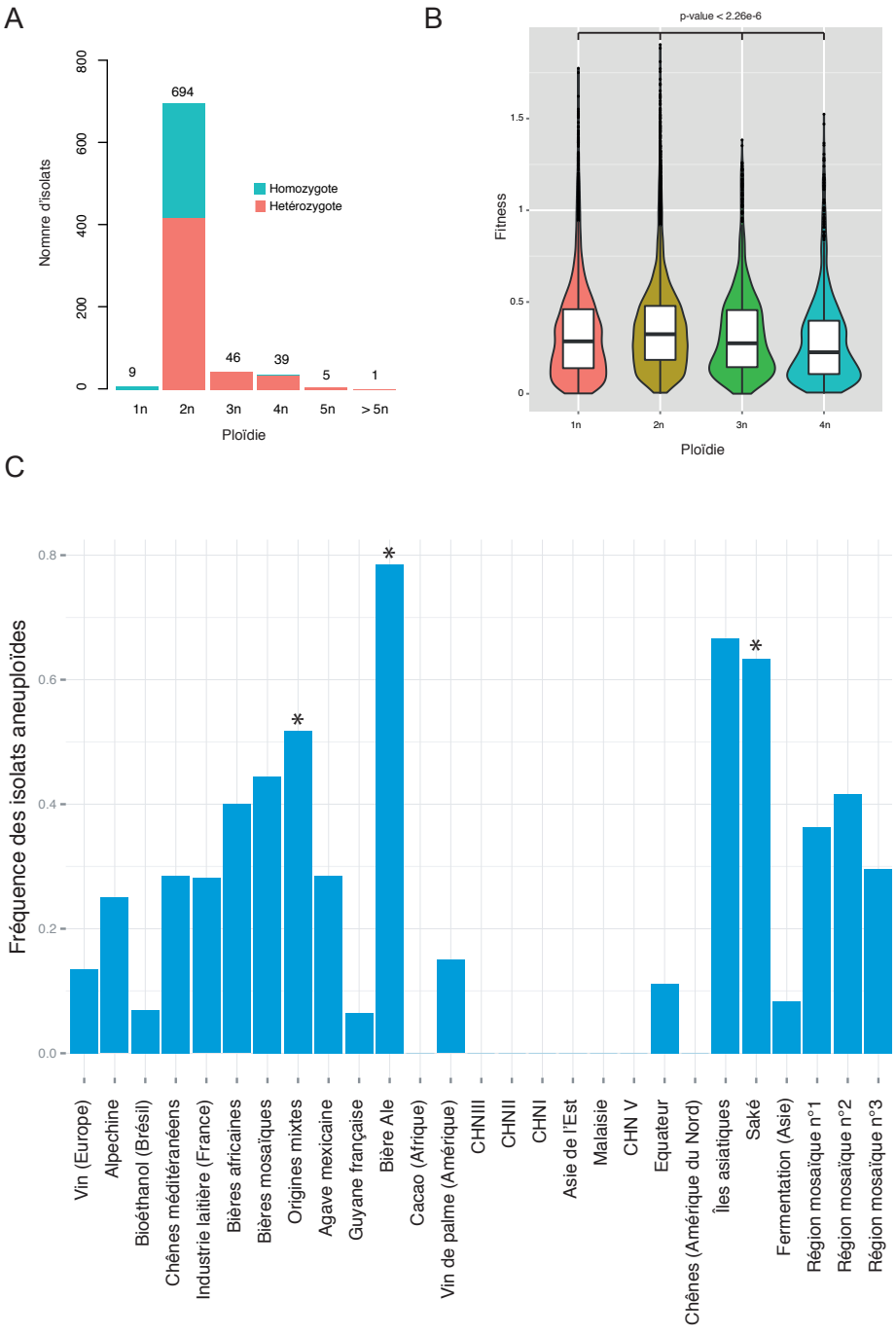


Figure 7. Distribution de la ploïdie et des aneuploïdies au sein des 1011 souches : (A) Distribution des différentes ploïdies au sein des isolats. (B) Mesures de fitness obtenues en fonction des ploïdies des souches. (C) Fréquences des souches possédant une aneuploïdie en fonction de l'origine écologique ou géographique des isolats. Adaptée de Peter *et al.* (2018).

Le cycle haplodiplobiontique des levures permet la propagation des isolats sous forme haploïde ou diploïde. L'examen systématique de la ploïdie de 794 isolats naturels a cependant montré que les souches diploïdes sont majoritaires dans la population (N = 694) en comparaison aux autres ploïdies (haploïdes = 9, polyploïdes = 91)[27] (Figure 7.A). Les souches diploïdes présentent par ailleurs un « fitness » plus élevé en comparaison aux autres ploïdies (Figure 7.B), indiquant que le maintien d'une ploïdie plus importante a un effet négatif sur la croissance des souches. De plus, une plus forte proportion de souches polyploïdes est retrouvée dans les sous-populations impliquées dans des processus de fermentation, telles que les souches de bières. Ces résultats suggèrent que la présence de ploïdies anormales est essentiellement liée aux processus de fermentation pour lesquels des pressions de sélection jouent un rôle important dans la structure des génomes des souches impliquées. Au sein des souches diploïdes, 60 % des isolats sont hétérozygotes mais cette proportion varie au sein de la population. De manière intéressante, alors que les souches retrouvées dans des processus de fermentation présentent une proportion plus importante de SNP hétérozygotes, ce taux varie entre les sous-populations concernées[27,84]. Par exemple, cette proportion est faible pour les souches de vin et de saké (moyenne de 2515 SNP hétérozygotes par souche) alors que les souches de bières sont fortement hétérozygotes (moyenne de 31374 SNP hétérozygotes par souche), pouvant résulter du nombre important des souches polyploïdes chez ces dernières. Par ailleurs, les variants hétérozygotes ne sont pas répartis de manière homogène le long du génome et plusieurs régions ayant subi une perte d'hétérozygotie peuvent être identifiées. Ces régions recouvrent en moyenne 50 % des génomes au sein de l'ensemble des individus hétérozygotes, allant jusqu'à 80 % chez les souches de saké, suggérant que ces évènements jouent un rôle clé dans l'évolution des génomes. En effet, une perte d'hétérozygotie permet l'expression d'allèles récessifs ainsi que la production de nouvelles combinaisons alléliques pouvant avoir un impact important sur les phénotypes.

De par la plasticité de son génome, la présence d'aneuploïdie ou d'un nombre variable de copie de certains gènes peut être fréquemment observée chez *S. cerevisiae*. Les aneuploïdies, résultant d'un dysfonctionnement des fuseaux mitotiques durant la ségrégation des chromosomes[85], modifient le dosage génique d'un grand nombre de gènes et sont ainsi souvent associées à une réponse rapide et transitoire face à un stress environnemental[86,87]. L'examen des 1011 isolats a montré des enrichissements significatifs de ces variations génétiques dans les sous-populations domestiquées, notamment de saké, de certaines bières (ale), ainsi que dans un groupe contenant des souches utilisées en boulangerie[27]. En accord avec d'autres études[86,88–90], les génomes des souches cliniques sont aussi significativement plus affectées par ces variations, suggérant que la présence d'aneuploïdies est probablement une des conséquences d'une réponse adaptative à l'utilisation de traitements antifongiques (Figure 7.C). Les aneuploïdies ne sont cependant pas restreintes à ces groupes et environ un quart des souches naturelles présentent elles aussi ce type de variation, suggérant que les souches de *S. cerevisiae* tolèrent relativement bien ces variants génomiques. Une tendance similaire peut être observée lors de l'analyse systématique du nombre de copies de chaque gène (CNV) dans la population et pour lequel un enrichissement est aussi observé pour les souches impliquées dans les processus fermentaires. Ces variations affectent de manière globale le génome et presque tous les gènes présentent un

nombre de copie variable dans au moins une des 1011 souches. La distribution de ces éléments le long du génome n'est cependant pas homogène et les gènes localisés à proximité des télomères sont bien plus affectés. Par ailleurs, seule une faible proportion de ces gènes présente un nombre très important de copies (> 20) et ceux-ci sont essentiellement retrouvés dans le génome mitochondrial ou dans le plasmide 2 micron. Ces gènes sont associés à des fonctions ribosomales ou à des éléments répétés tels que les transposons, résultant respectivement de l'utilisation intensive de la machinerie de traduction et de la nature égoïste des éléments transposables.

Au sein de *S. cerevisiae*, la variabilité du contenu en gène est aussi caractérisée par la présence d'un nombre important d'introgressions, découlant du transfert d'un ou de plusieurs gènes d'une autre espèce lors d'évènements d'hybridation. L'identification systématique de ces gènes au sein des 1011 souches a permis de mettre en évidence 913 gènes introgressés au sein de la population et indique ainsi l'importance de ces évènements dans la structure actuelle des génomes de *S. cerevisiae*[27]. Près de 97 % de ces introgressions proviennent de *S. paradoxus* et confirment ainsi l'important flux de gènes existant entre ces deux espèces phylogénétiquement proches. Ces introgressions ont tendance à remplacer la copie orthologue de *S. cerevisiae*, suggérant que leur intégration dépend essentiellement d'une recombinaison homologue entre les deux génomes. De plus, 183 gènes résultant du transfert horizontal d'espèces phylogénétiquement plus éloignées ont été identifiés. Ces évènements affectent majoritairement les souches impliquées dans des processus de fermentation et plus de 30 % de ces gènes proviennent d'espèces appartenant aux genre *Zygosaccharomyces* et *Torulaspora*, coexistant avec *S. cerevisiae* durant ces processus. L'analyse conjointe de ces résultats avec ceux obtenus par l'étude du nombre de copie des gènes permet l'obtention d'une vue globale du répertoire des séquences codantes (ORF) au sein de l'espèce. Chez *S. cerevisiae*, celui-ci est constitué de 7797 ORF (pangénome) dont 4940 sont présentes au sein des tous les individus (core génome)[27]. Ce résultat a amené à l'identification de 2857 gènes accessoires présents de manière variable au sein des différentes souches. Ces gènes accessoires présentent des signatures génétiques particulières par rapport au reste du génome. Premièrement, un taux plus important de mutations non-synonymes et de mutations délétères y est associé, indiquant une sélection moins forte sur ces gènes en comparaison à ceux du core génome. Deuxièmement, un enrichissement des gènes impliqués dans les interactions cellulaires, le métabolisme secondaire et la réponse au stress est retrouvé au sein des gènes accessoires[27]. Troisièmement, les gènes accessoires sont localisés préférentiellement dans les régions télomériques[27,91], renforçant l'idée que les régions télomériques jouent un rôle important dans la dynamique des génomes et l'adaptation à de nouvelles conditions de par la plasticité qu'elles apportent.

Adaptations locales dans les différentes sous-populations

L'exploration du génome de plusieurs centaines d'isolats offre l'occasion d'identifier avec précision les différences évolutives entre sous-populations. Ces analyses chez *S. cerevisiae* ont ainsi mis en évidence des signatures évolutives distinctes au sein des sous-populations domestiquées[27]. Par exemple, les génomes des isolats de bière se caractérisent par une ploïdie

plus importante (≥ 3n) ainsi que la présence d'un nombre élevé d'aneuploïdies. Par ailleurs, la diversité génétique nucléotidique estimée pour l'ensemble des isolats est particulièrement élevée ($\pi = 2{,}8 \times 10^{-3}$) et les génomes sont fortement hétérozygotes. De manière intéressante, alors que les souches de bière sont polyphylétiques, les signatures génomiques associées à ce processus de fermentation sont retrouvées dans les différentes sous-populations associées. La génération et le maintien de génomes polyploïdes, favorisant l'acquisition et l'accumulation de nouvelles mutations, sont probablement liés à l'usage intensif durant une longue période des mêmes isolats lors du brassage des bières.

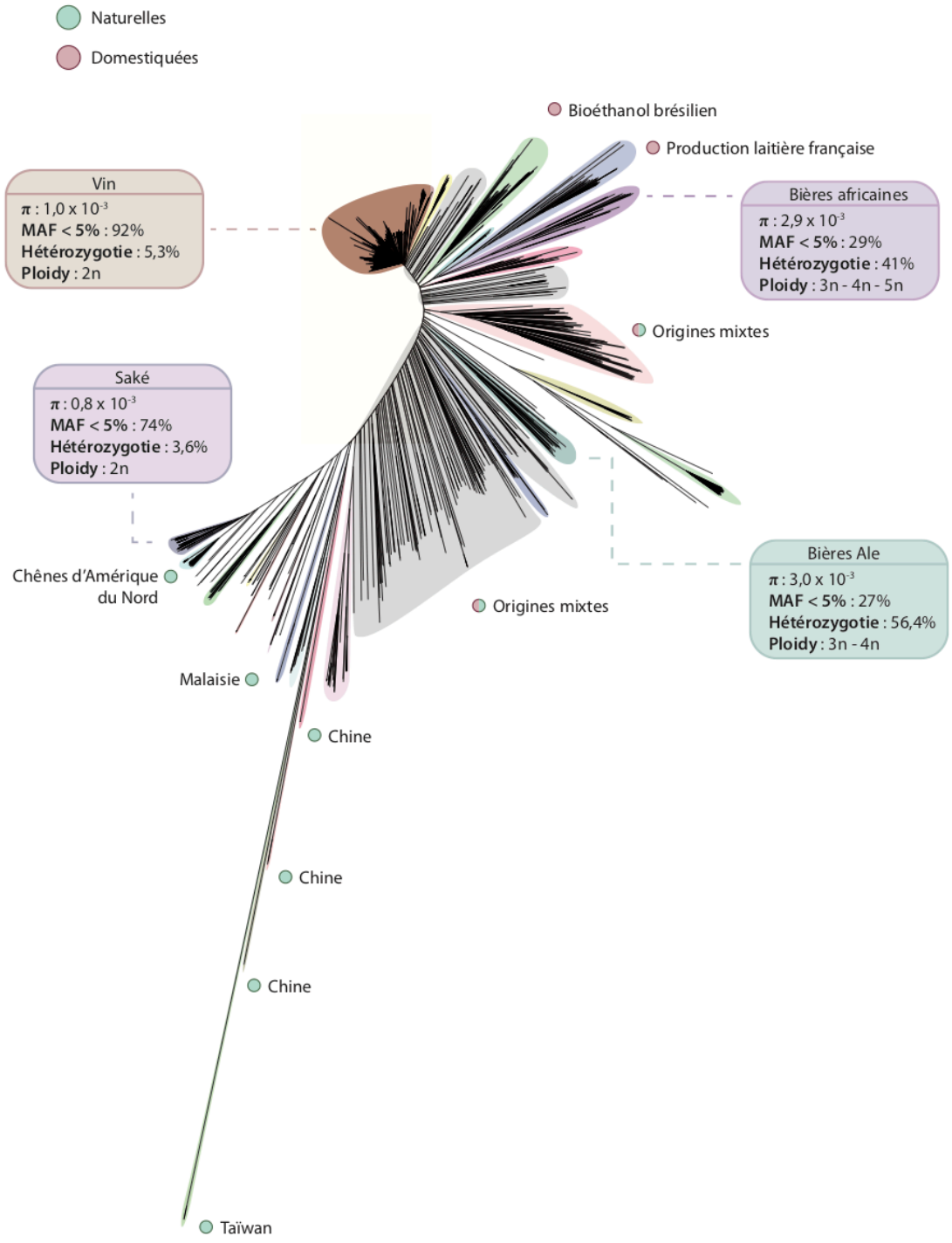


Figure 8. Relations phylogénétiques entre les 1011 souches de *S. cerevisiae*. Des métriques de la diversité nucléotidique (π), de la fréquence allélique mineure (MAF), du taux d'hétérozygotie et de la ploïdie majoritairement retrouvée sont données pour plusieurs sous-populations. Adaptée de Peter *et al.* (2018).

Par opposition, les sous-populations de saké et de vin sont monophylétiques et sont constituées majoritairement d'isolats diploïdes. La diversité génétique en leur sein est très faible (vin : $\pi = 1 \times 10^{-3}$ – saké : $\pi = 0{,}8 \times 10^{-3}$) et on comptabilise un nombre important de variants ayant une faible fréquence allélique, en particulier pour les souches de vin pour lesquelles 95 % des sites polymorphiques sont associés à une fréquence allélique mineure inférieure à 1 % (Figure 8). Ces souches sont par ailleurs très faiblement hétérozygotes (vin : 5,3 % - saké = 3,6 %) et un grand nombre de régions ayant subi une perte d'hétérozygotie peut être retrouvé dans leur génome. Ces observations suggèrent une expansion récente de ces populations pour ces deux groupes, favorisant l'accumulation de nouvelles mutations chez les différents individus.

La caractérisation des gènes introgressés ou présentant un nombre variable de copies entre les sous-populations permet l'identification de gènes jouant un rôle important dans l'adaptation à certains environnements. Par exemple, l'introgression de plusieurs gènes impliqués dans le métabolisme du galactose provenant de *S. eubayanus* et de *S. uvarum* au sein de souches fromagères de *S. cerevisiae* permet une meilleure assimilation de ce substrat[70]. De manière comparable, il a été démontré qu'une introgression du gène *PDR5* codant un transporteur pléiotropique de diverses drogues offre une résistance accrue à plusieurs antibiotiques[74]. Les évènements de duplication ont aussi un impact considérable dans l'adaptation des souches à certains environnements, de par la modification du dosage génique engendré. Ainsi, plusieurs gènes impliqués dans le métabolisme du maltose, un composant présent dans la production de saké et de bière mais absent dans le raisin, sont amplifiés dans les souches liées à ces deux premiers processus, mais absents dans les souches de vin[78]. Similairement, la duplication de groupes de gènes jouant un rôle dans la résistance au cuivre, dans la floculation ou dans le métabolisme du glucose a aussi été retrouvée dans plusieurs populations domestiquées[27].

L'étude de la variabilité au sein des génomes de plusieurs dizaines d'individus a permis l'établissement d'études d'association en confrontant les variants génétiques identifiés à des données phénotypiques (GWAS, Genome Wide Association Studies)[74,92–95]. Chez *S. cerevisiae*, la première étude d'association, basée sur la comparaison de 44 souches cliniques contre 44 souches non cliniques a été réalisée à partir de données de puce à ADN[92]. Cette analyse a mis en évidence une association entre plusieurs polymorphismes retrouvés dans les souches cliniques, notamment dans des gènes impliqués dans la formation de pseudo-hyphes ainsi que la détoxification de la cellule. Plus récemment, les génomes complets de 100 souches et de données de phénotypage sur 49 conditions ont permis d'identifier plusieurs variants du nombre de copies significativement associés aux phénotypes dans cette population[74]. Parmi les gènes impliqués, une association entre les gènes *ENA* codant une pompe Li^+ Na^+ et la résistance au lithium, ainsi qu'une association entre le gène *CUP1*, un gène impliqué dans la fixation et la médiation du cuivre dans la cellule, et la résistance au cuivre ont pu être identifiées. Ces travaux

ont posé les fondements de ces études d'association chez *S. cerevisiae* mais sont restés limités notamment à cause du faible nombre de génomes étudiés (< 100). Dans ce contexte, le séquençage du génome complet et le phénotypage de 971 isolats sur 35 conditions ayant un impact sur divers processus physiologiques et cellulaires (sources de carbone, transcription, traduction, stress osmotique ou oxydant) a donné la possibilité de dépasser cette restriction[27]. La comparaison systématique des variants génétiques pour l'ensemble des conditions a ainsi mis en évidence 35 variants génétiques associés à 14 conditions, avec une plus forte proportion de CNV (N = 22) en comparaison aux SNP (N = 13). Ce résultat confirme l'impact important des CNV dans la variabilité phénotypique, bien que celui-ci soit aussi une conséquence de la forte héritabilité de ces variants. Par exemple, une association significative entre la variation du nombre de copies du gène *CUP1* et la présence de sulfate de cuivre a une nouvelle fois été retrouvée. Ce variant explique à lui seul 44,5 % de la variance phénotypique mesurée. De manière similaire, un nombre supérieur de copies du cluster *ARR* (*AAR1*, *AAR2* et *ARR3*) permettant une résistance accrue au méta-arsénite de sodium a été retrouvé, comme précédemment observé[88]. Par ailleurs, quatre des 35 variants retrouvés sont associés à des gènes absents dans la souche de référence, suggérant un impact fonctionnel non négligeable de la plasticité génomique dans l'adaptation des souches.

Génomique des populations au sein d'espèces de levures non-modèles

L'analyse comparative des génomes de plusieurs centaines d'isolats chez *S. cerevisiae* a permis une exploration globale de la diversité génétique au sein de cette espèce. Aujourd'hui, il est cependant clair que la connaissance des processus impliqués dans l'évolution des génomes peut être améliorée par l'exploration d'un plus grand nombre d'espèces. En effet, une espèce n'est pas représentative de l'évolution d'un phylum ou d'un clade. Dans ce contexte, l'étude d'espèces non-modèles au sein du sous-phylum des Saccharomycotina est particulièrement intéressante. En effet, les études de génomique comparative ont montré que la divergence génétique entre les espèces qui le constituent est importante. Par ailleurs, les espèces de levures présentent une diversité importante relative à leurs milieux de colonisation et plusieurs cas d'homoplasie ont pu être mis en évidence, suggérant la présence de plusieurs évènements d'adaptations indépendants dans ce phylum. La mise en place de telles études est cependant encore grandement limitée par différents facteurs. Tout d'abord, ces études dépendent de la disponibilité d'une séquence de référence et d'annotations de bonne qualité, ce qui n'est pas le cas pour la majorité des espèces du sous-phylum. De plus, le nombre de souches isolées n'est parfois pas suffisant pour produire une analyse représentative de la diversité génétique de l'espèce.

Les efforts importants réalisés depuis le début des années 2000 pour obtenir des séquences de références de qualité et bien annotées pour des dizaines d'espèces de levures, en particulier dans le cadre des études menées par le consortium Génolevures, ont permis la mise en place de plusieurs études de génomique des populations au sein d'espèces non-modèles au cours des dernières années[38,62,72,83,96–110]. Bien que ces études soient limitées par des collections souvent réduites en comparaison à celles utilisées dans l'étude de *S. cerevisiae*, celles-ci ont déjà permis d'obtenir un meilleur aperçu de plusieurs aspects relatifs à l'évolution globale des génomes : (i) l'histoire évolutive des espèces ; (ii) les forces impliquées dans l'évolution des génomes et (iii) la relation entre le génotype et le phénotype. De par l'indépendance de ces études, notamment dans les méthodes utilisées, des variants étudiés et de l'axe de recherche abordé, il est cependant difficile de comparer de manière exhaustive les résultats obtenus dans ces différents travaux. Dans ce contexte, je me focaliserai uniquement sur les résultats obtenus chez quelques espèces, représentatifs de chacun des aspects abordés.

Exploration de l'histoire évolutive des espèces

L'analyse de la distribution des variants génétiques au sein d'une population permet d'inférer les relations entre les souches et d'identifier des sous-populations. La comparaison de sous-populations peut ensuite permettre de mieux appréhender et comprendre l'histoire évolutive des différentes espèces.

Histoire évolutive chez S. paradoxus

La levure bourgeonnante *Saccharomyces paradoxus* est l'espèce la plus proche de *S. cerevisiae* connue à ce jour. Cette espèce, isolée majoritairement dans la sève ou dans l'écorce d'arbres à travers le monde (Amérique, Europe de l'Est et Asie), n'est cependant pas retrouvée dans des processus de fermentation dirigés par l'Homme, contrairement à *S. cerevisiae*. Ces caractéristiques ont fait de *S. paradoxus* la première espèce non-modèle utilisée dans des études de génomique des populations, à travers l'exploration parallèle de la diversité intraspécifique de 36 isolats de *S. cerevisiae* et de 35 isolats de *S. paradoxus*[72]. Cette étude a donné un premier aperçu des différences dans l'histoire évolutive entre ces espèces. En effet, alors que les souches de *S. paradoxus* se regroupent en trois clusters présentant une structure extrêmement conservée et une divergence allant jusqu'à 3,5 %, l'analyse des relations entre les isolats de *S. cerevisiae* montre un continuum composé de plusieurs souches mosaïques.

Plus récemment, l'exploration de la diversité génétique et phénotypique au sein d'une population d'isolats de *S. paradoxus* d'Amérique du Nord a donné un exemple intéressant de l'adaptation d'une espèce à son environnement et de l'impact que peut avoir cet environnement dans son histoire évolutive. L'analyse phénotypique de 27 souches au sein de cette population a montré un gradient Nord-Sud dans la résistance des souches à différentes températures. Les souches isolées au Sud de la région présentent une meilleure résistance à de hautes températures et à des variations importantes de celles-ci, en accord avec les relevés météorologiques retrouvés dans ces régions[107]. De manière intéressante, ces données phénotypiques permettent la distinction des souches en 3 sous-populations ayant une conformation similaire à celle observée lors de l'analyse de la structure obtenue dans l'étude de 35 loci chez 62 souches[111]. L'importante divergence génétique ainsi que la conservation forte au sein des différentes sous-populations suggéraient alors un isolement reproductif entre les individus qui les composent. Les origines d'un tel phénomène peuvent être multiples, allant de l'incompatibilité génétique entre deux génomes parentaux à un dysfonctionnement global du mécanisme de réparation des mésappariements aboutissant à une mauvaise ségrégation des chromosomes[112]. L'analyse de la viabilité de la descendance de 25 souches représentatives des différentes sous-populations a confirmé cette hypothèse et une corrélation négative entre la viabilité des spores et la divergence génétique entre les parents a pu être établie[113], suggérant que la spéciation observée entre ces sous-populations résulterait en partie de l'accumulation de mutations indépendantes dans chacune d'entre elles. Plus récemment, une analyse de génomique des populations portant sur l'analyse du génome complet de 161 souches a permis de caractériser les origines de cet évènement. L'assemblage *de novo* des génomes des souches a permis d'identifier plusieurs variants structuraux, notamment une inversion de 42 kb et une translocation entre le chromosome 9 et 13, retrouvés de manière spécifique au sein de certaines populations. L'ensemble des résultats trouvés dans cette étude a permis de dresser un scénario expliquant la distribution actuelle de la population (Figure 9) : la population ancestrale d'Amérique du Nord a été séparée par un épisode majeur de glaciation et les deux sous-populations obtenues (nommées SpC et SpB) ont accumulé un grand nombre de mutations. Au cours de cette période, la translocation et l'inversion se sont fixées au sein de la sous-population SpB. Après la fonte

des glaces, les deux populations ancestrales ont pu rentrer de nouveau en contact et une zone d'hybridation a donné place à une troisième sous-population (SpC*). Finalement, la perte progressive des variants structuraux dans le fonds génétique des populations parentales a conduit à l'isolement reproductif entre les différentes populations.

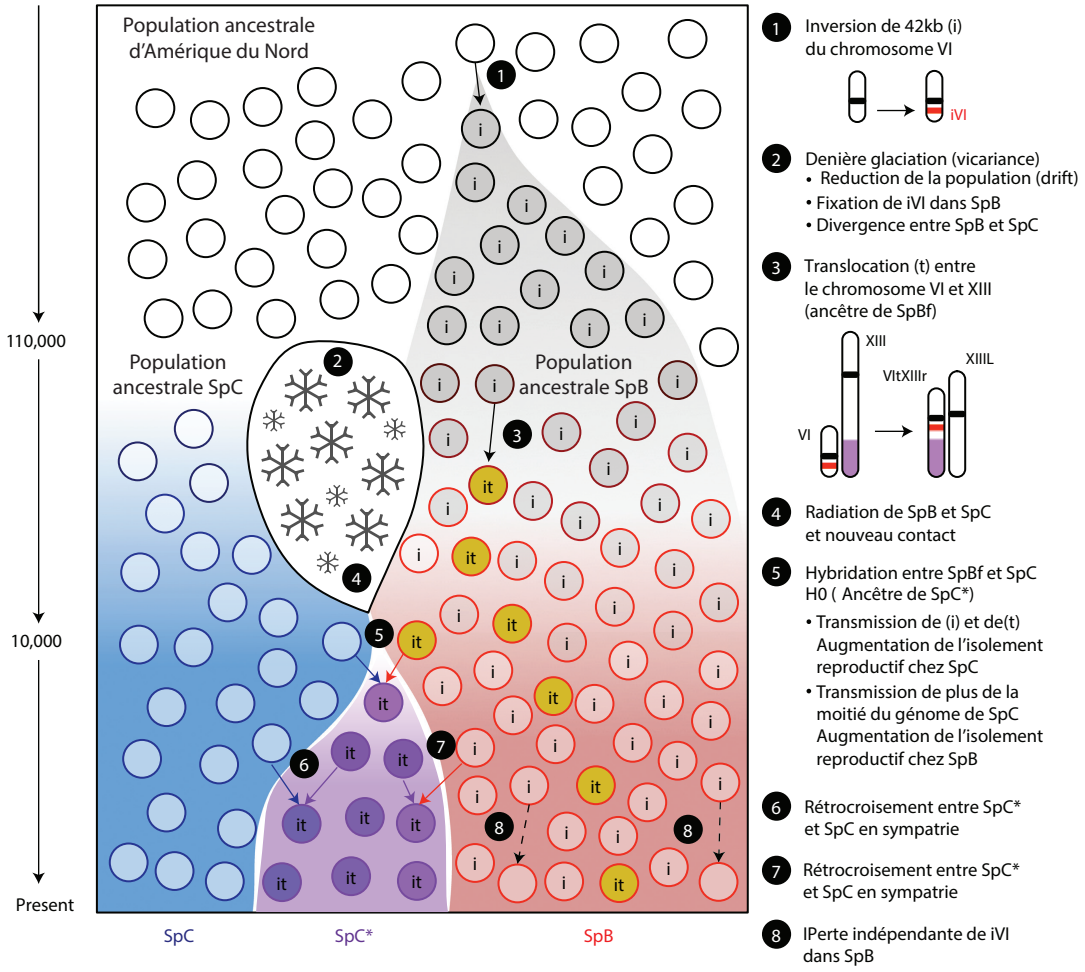


Figure 9. Scénario envisagé de l'origine des différentes sous-populations d'Amérique du Nord de *S. paradoxus*[109]. Adaptée de Leducq *et al.* (2016).

Une généralisation de l'histoire évolutive de ces sous-populations reste cependant impossible tant les évènements de spéciation se révèlent spécifiques aux fonds génétiques impliqués. Par exemple, l'exploration de plusieurs génomes de *Saccharomyces kudriavzevii* a révélé la présence de deux sous-populations présentant d'importantes variations génétiques, notamment de l'ensemble du réseau impliqué dans la dégradation du galactose[83]. Bien que la structure de la population démontre une absence de croisement entre les deux populations dans la nature, une descendance viable peut être obtenue en laboratoire[83].

Origine de l'hybridation de S. pastorianus

S. pastorianus est une espèce allopolyploïde résultant du croisement de *S. cerevisiae* et de l'espèce cryo-tolérante *S. eubayanus*[58]. La combinaison de ces génomes fait de *S. pastorianus* une espèce utilisée de manière intensive dans le brassage de bière de type lager de par sa capacité à fermenter à basse température (2-4°C). L'origine de cet hybride est pourtant restée pendant longtemps en grande partie méconnue et ce ne sera qu'en 2011 que l'identification du deuxième parent de cet hybride[58] sera obtenue à partir de la découverte d'une souche de *S. eubayanus* en Patagonie. L'origine de cet isolat était cependant en contradiction avec l'utilisation de *S. pastorianus* en Europe avant la découverte de l'Amérique, suggérant la présence d'autres sous-populations de *S. eubayanus*. Récemment, l'isolement et le séquençage de nouvelles souches de *S. eubayanus* au Tibet[114] a mis en évidence la forte proximité génétique entre ces isolats et les génomes actuels de *S. pastorianus*, suggérant une origine asiatique à cette espèce, similaire à celle de *S. cerevisiae*[27,71]. Les premières études de génomique des populations de *S. pastorianus*, basées sur le génotypage par puce à ADN d'une vingtaine de souches, ont permis d'explorer une partie de la diversité génétique de cette espèce[57,115]. Les isolats se regroupent en deux sous-populations : le groupe 1 correspondant aux bières de type Saaz se caractérise par une perte importante de régions génomiques relatives à *S. cerevisiae*. Le groupe 2 contient des souches provenant de brasseries hollandaises ou danoises et présente quant à lui une conservation presque intégrale du génome des deux espèces. Cette distribution a suggéré l'implication d'au moins deux évènements d'hybridation dans la formation de cet hybride. Ces études ont cependant été limitées par l'analyse d'une portion du génome des souches ainsi que par l'absence du génome de *S. eubayanus*, amenant à l'utilisation à la place de celui de *S. uvarum* présentant une constitution similaire.

Récemment, le séquençage du génome complet de 10 souches (5 isolats pour chaque groupe) de *S. pastorianus* a offert la possibilité d'étudier plus précisément l'histoire évolutive de cette espèce[62]. Cette étude a d'abord permis de mieux caractériser les sous-populations. Alors que les souches du groupe 1 possèdent une version haploïde du génome de *S. cerevisiae* et diploïde de *S. eubayanus*, les isolats du groupe 2 sont diploïdes pour les deux parents. Pour les deux groupes, des aneuploïdies et des duplications ou délétions segmentales sont fréquemment observées, notamment pour le génome de *S. cerevisiae* chez les espèces du groupe 1. De manière intéressante, 7 translocations ont aussi été retrouvées dans les génomes et trois d'entre elles sont partagées dans les deux groupes, suggérant au moins un évènement d'hybridation commun pour l'ensemble de la population. Cette hypothèse est appuyée par l'analyse phylogénétique du génome mitochondrial et du génome nucléaire de *S. eubayanus* pour lesquels une séparation entre les deux groupes n'est pas observable. Ce résultat contraste cependant avec celui observé dans l'analyse du génome nucléaire de *S. cerevisiae* pour lequel deux clusters correspondant aux deux groupes sont retrouvés. Par ailleurs un grand nombre de régions ayant subi une perte d'hétérozygotie a pu être retrouvé dans le groupe 2. Ces résultats suggèrent un évènement commun d'hybridation entre les deux groupes, suivi rapidement d'une évolution différente et indépendante des génomes des deux sous-populations. Bien que ces résultats ne

permettent pas de déterminer avec précision les origines génétiques responsables de la distinction entre les deux groupes, plusieurs scénarii peuvent être imaginés (Figure 10). Dans un premier cas (Figure 10. A), l'hybride ancestral aurait été formé par la combinaison de deux souches diploïdes, suivi par une perte d'une des deux copies du génome de *S. cerevisiae* dans le groupe 1. Une seconde hypothèse (Figure 10. B) serait que *S. pastorianus* résulterait du croisement entre un isolat diploïde de *S. eubayanus* et d'une souche haploïde de *S. cerevisiae*. Au cours de l'évolution, un second évènement d'hybridation spécifique au groupe 2 aurait eu lieu, apportant une copie supplémentaire du génome de *S. cerevisiae*.

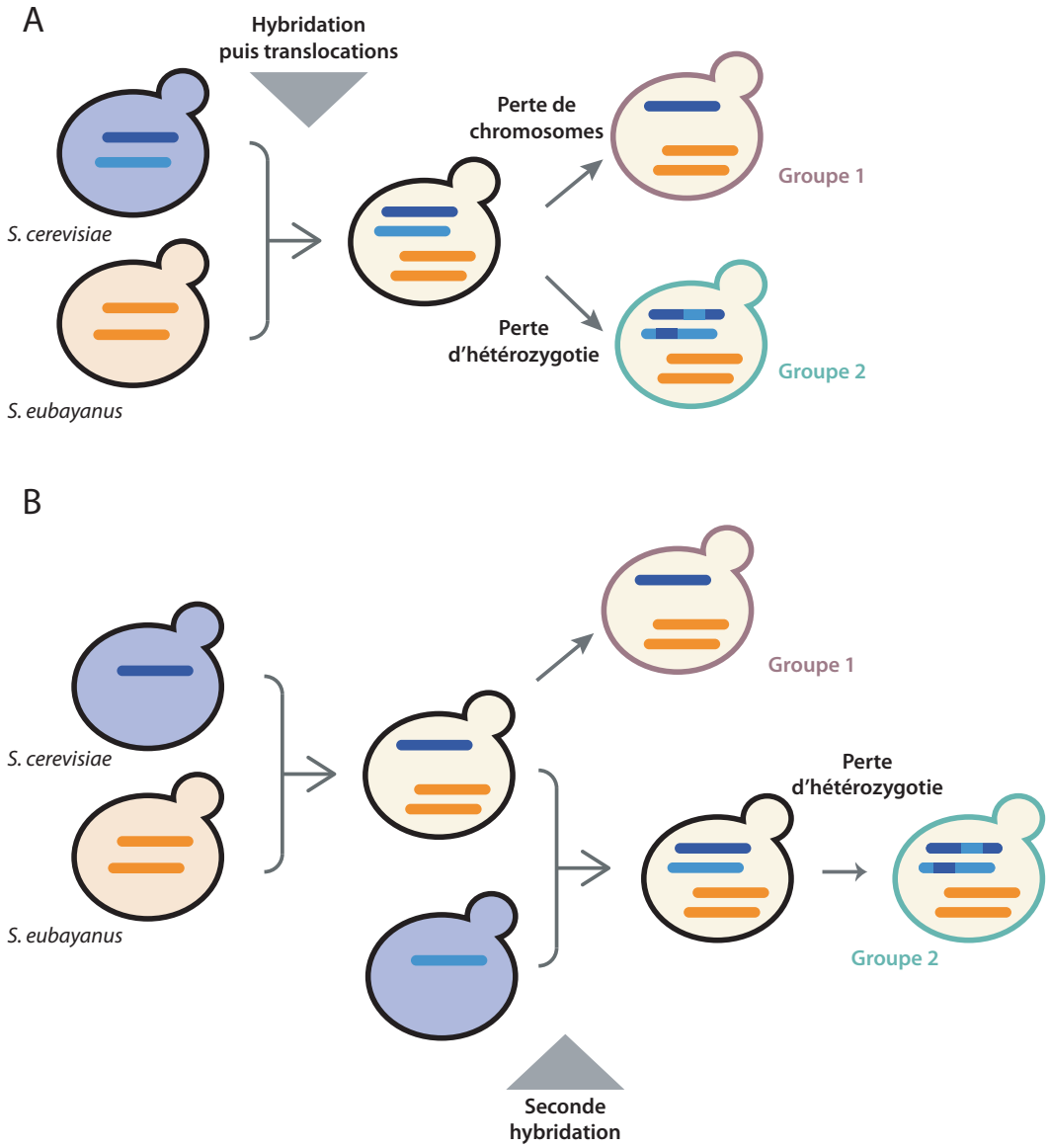


Figure 10. Hypothèses de l'origine de la formation de l'hybride *S. pastorianus* : (A) formation basée après un événement d'hybridation entre deux souches diploïdes et une perte partielle du génome de *S. cerevisiae* dans l'une des sous-populations (Groupe 1). (B) Scénario impliquant

le croisement d'une souche haploïde de *S. cerevisiae* et d'une souche diploïde de *S. eubayanus* suivi d'un second événement d'hybridation de *S. cerevisiae* chez l'ancêtre commun de la seconde sous-population (Groupe 2). Adaptée de Okuno *et al.* (2016).

Variabilité dans les forces impliquées dans l'évolution des génomes

L'évolution des génomes résultent de différentes forces, telles que la recombinaison ou l'accumulation de mutations, permettant notamment l'adaptation des individus à leurs environnements. L'exploration de populations au sein de différentes espèces a permis d'obtenir des premiers indices sur l'étendue des différents processus évolutifs au sein du sous-phylum des Saccharomycotina.

Adaptation des souches impliquées dans les processus de fermentation chez S. uvarum

Le reséquençage de 54 souches de *S. uvarum*, une espèce mise à profit dans le cadre de différents processus de fermentation (vin et cidre), a permis l'exploration de la phylogéographie et des traces de domestication dans cette espèce[96]. Dans cette étude, des régions introgressées, dont l'origine a été attribuée à plusieurs espèces incluant *Saccharomyces kudriavzevii* (14 souches), *S. cerevisiae* (3 souches) et *S. eubayanus* (20 souches), ont été identifiées au sein de 20 souches. L'impact de ces introgressions varient en terme de taille et de nombre entre les souches mais les localisations de celles-ci ont tendance à être partagées entre les différents isolats. Alors que l'ensemble du génome de *S. uvarum* est impacté par de tels évènements, certaines introgressions sont particulièrement grandes et peuvent atteindre plusieurs centaines de kb. De manière intéressante, presque toutes les souches affectées sont impliquées dans des processus de fermentation. Ces résultats offrent un parallèle intéressant avec ceux obtenus chez *S. cerevisiae* pour laquelle 913 introgressions ont pu être déterminées dans l'analyse de 1011 souches et provenant essentiellement de son espèce proche *S. paradoxus*[27]. Par ailleurs, certaines de ces introgressions peuvent apporter un avantage sélectif, notamment dans certains processus de fermentation telle que la production de vin[116]. L'analyse des zones introgressées dans la population de *S. uvarum* a permis de retrouver 34 gènes impliqués dans la fermentation du moût de raisin, notamment dans le métabolisme et le transport des sucres et du nitrogène. Ces résultats suggèrent que les introgressions jouent un rôle important dans l'adaptation de ces deux espèces dans les processus de fermentation. La signature adaptative de la sous-population domestiquée de *S. uvarum* diffère cependant de celle observée pour les souches de vigne de *S. cerevisiae*. Alors que ces dernières présentent une diversité nucléique plus faible en comparaison au reste de la population, une telle différence n'est pas observée pour les souches de *S. uvarum*.

Introgression dans une population naturelle de levure

L'étude d'espèces naturelles est nécessaire afin d'avoir une vue d'ensemble des processus évolutifs impliqués dans l'évolution des génomes. Dans ce cadre, l'espèce *L. kluyveri* présente des propriétés intéressantes : les isolats peuvent être retrouvés à travers le monde et sont

principalement issus d'exsudats d'arbre ou d'insecte. À l'opposé de *S. cerevisiae*, cette espèce ne fermente pas les sucres comme le glucose en présence d'oxygène[117] et se différencie du groupe des *Saccharomycetaceae* par sa capacité à utiliser les pyrimidines et ses dérivés comme unique source d'azote[118–120]. C'est une espèce présente au sein des *Saccharomycetaceae* mais n'ayant pas d'ancêtre ayant subi la duplication totale du génome contrairement à *S. cerevisiae*. De manière intéressante, le séquençage de son génome a révélé une large région de 1 Mb correspondant au bras gauche du chromosome C et ayant un taux de GC supérieur comparé au reste du génome[121] (désignée C-left par la suite). L'analyse de la phylogénie, de la conservation de la synténie et du contenu en gène de cette région a suggéré que celle-ci résulterait probablement d'une large introgression[122].

Ces caractéristiques ont poussé à l'établissement d'une étude de génomique des populations basée sur la comparaison du génome complet d'une collection de 31 isolats provenant d'origines géographiques (Europe, Asie, États-Unis) et écologiques (arbre, sol, insectes) variées[123]. Cette population présente une diversité nucléotidique bien plus importante que celle observée au sein de *S. cerevisiae* ($\pi_{\textit{L. kluyveri}} = 0{,}017$, $\pi_{\textit{S. cerevisiae}} = 0{,}003$). Cette diversité est accompagnée d'une divergence nucléotidique de 2,8 % entre les deux souches les plus éloignées alors que celle-ci atteint 1,2 % chez *S. cerevisiae*[27]. De manière intéressante, le contenu en GC est plus élevé pour l'ensemble de la collection au niveau du C-left, suggérant que l'événement évolutif qui est à son origine a précédé la diversification de l'espèce (Figure 11. C). Cette hypothèse est renforcée par l'observation d'une topologie similaire entre l'arbre phylogénétique obtenu par l'analyse du génome complet et celui généré avec le C-left pris individuellement. De manière additionnelle, cette région se caractérise par une diversité nucléotidique et une recombinaison ancestrale plus élevées en comparaison au reste du génome (Figure 11. A-B). Par ailleurs, la fréquence de substitution GC vers AT est supérieure à celle d'AT vers GC au sein de cette région, alors que les fréquences sont similaires pour le reste du génome. Ces données suggèrent que la composition en base est à l'équilibre pour l'ensemble des chromosomes sauf pour le C-left, qui présente une tendance à la diminution du pourcentage en GC.

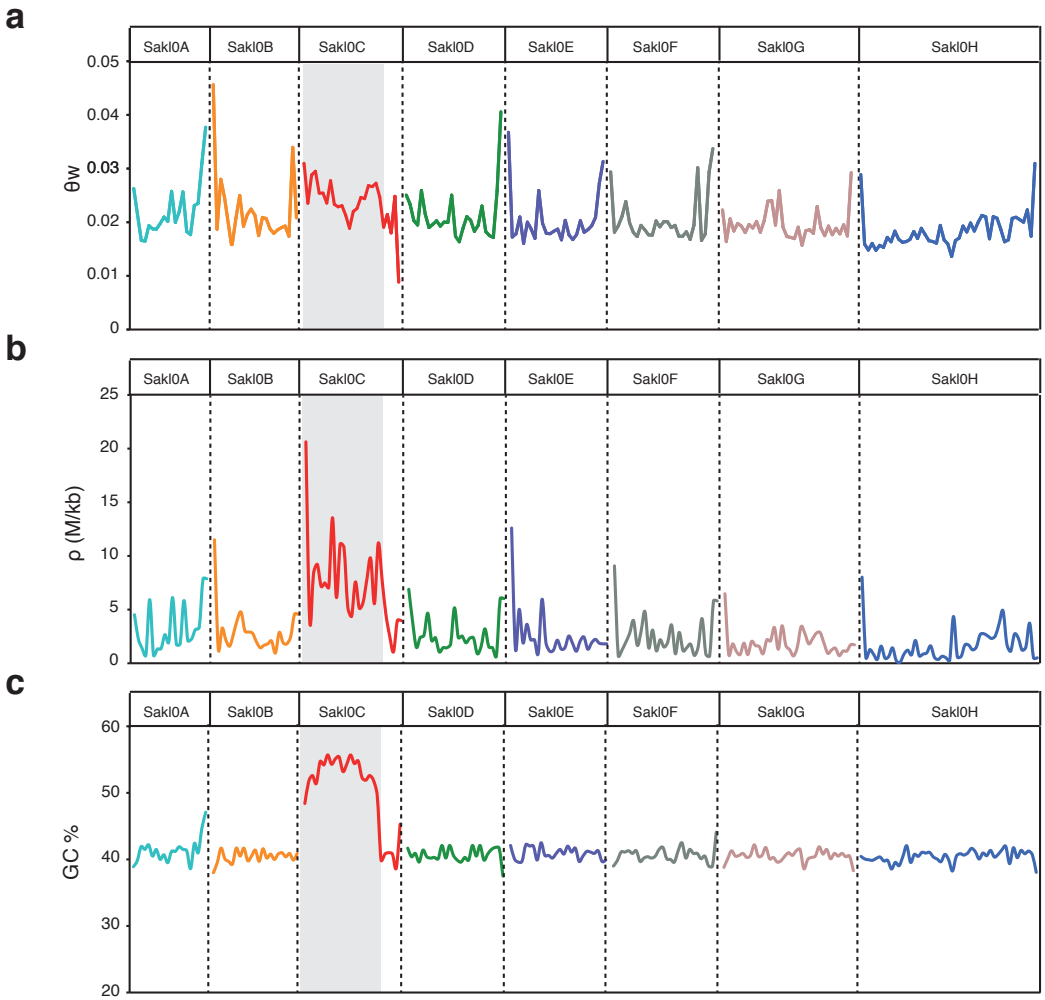


Figure 11. Variabilité des métriques le long du génome de *L. kluyveri* à partir de fenêtre de 50 kb. La partie grisée représente le C-left. (A) Proportion de sites polymorphiques. (B) Taux de recombinaison ancestrale. (C) Pourcentage en GC. Adaptée de Friedrich *et al*. (2015).

Afin de mieux comprendre l'influence du fonds génétique sur le transcriptome de cette espèce, l'ensemble des transcrits de 20 isolats naturels ont été séquencés après croissance des souches en milieu complet par RNA-seq[124]. Les gènes retrouvés dans la région introgressée présentent une variation intraspécifique de leur expression plus élevée (plasticité) et sont moins impliqués dans les modules de co-expression ainsi que dans la régulation d'autres gènes (connectivité). L'ensemble de ces observations suggère que ces gènes ne sont pas encore complètement intégrés dans les réseaux métaboliques du génome et qu'ils présentent une plus grande variabilité dans leur expression et leur séquences comparés au reste du génome. Dans une autre étude, la recombinaison méiotique au sein de l'espèce a été explorée à travers le séquençage du génome complet de 49 spores résultant du croisement de deux souches naturelles[125]. Cette analyse a montré dans un premier temps que le taux de recombinaison est plus faible chez *L. kluyveri* en comparaison à *S. cerevisiae* (~ 1,6 contre ~ 6,0 cross-over par Mb respectivement). Cette différence pourrait être liée au fait que *L. kluyveri* présente moins de chromosome (N = 8) comparé à *S. cerevisiae* (N = 16) pour une taille de génome similaire (~

12 Mb). En effet, les espèces possédant un plus grand nombre de chromosomes ont tendance à posséder un taux de recombinaison plus élevé afin d'assurer la présence d'un cross-over pour au moins chacun d'entre eux lors de la méiose et assurer ainsi une ségrégation correcte de ceux-ci. De manière surprenante, cette étude a aussi révélé une absence de crossing-over au sein du C-left alors que le taux de recombinaison ancestrale avait été déterminé comme plus élevé, suggérant une variation du taux de recombinaison au cours de l'évolution de l'espèce. De plus, cette absence amènerait sur le long terme à une conservation de l'ensemble des allèles dans la région et limiterait ainsi la diversité génomique présente en celle-ci. Cette observation contraste avec la présence d'une diversité génétique plus élevée au sein de cette région en comparaison au reste du génome. L'origine de cette absence de recombinaison reste cependant encore méconnue.

Étude de la relation génotype-phénotype au sein d'espèces non-modèles

L'exploration d'espèces non-modèles et notamment d'espèces présentant des cas d'homoplasie est particulièrement intéressante afin d'appréhender la relation qui lie le génotype et le phénotype. En effet, il est maintenant admis que la diversité phénotypique des espèces est liée de manière importante à leur histoire évolutive et ne corrèle pas simplement avec la variabilité génétique intraspécifique[126]. Au sein des espèces non-modèles des Saccharomycotina, l'exploration de la relation génotype-phénotype s'est notamment portée sur les espèces ayant un rôle important dans les infections humaines. Sur le plan médical, ces analyses visent à mieux comprendre comment ces organismes pathogènes évoluent et s'adaptent lors des infections et de la prise de traitements, nous donnant des renseignements importants dans la mise en place de futures thérapies. Dans le domaine de la biologie évolutive, l'étude de ces espèces offre l'occasion d'identifier les processus évolutifs responsables de l'adaptation face à ces pressions de sélection.

Dans ce cadre, l'exploration des génomes de l'espèce pathogène *C. albicans* est particulièrement intéressante. En effet, cette espèce fait partie du groupe des CTG, un clade diversifié responsable d'un grand nombre d'infections[127] . Parmi les espèces pathogènes, *C. albicans* joue un rôle prédominant et est le pathogène fongique le plus retrouvé lors d'infections chez l'Homme. Cette caractéristique a amené au séquençage rapide de son génome complet en 2004[128] et à l'établissement de plusieurs études de génomique des populations. Une première étude, basée sur la comparaison des génomes complets de 21 souches échantillonnées chez des patients infectées et ayant pris un traitement médical a mis en évidence un nombre important d'aneuploïdies au sein de cette espèce, permettant une adaptation rapide des souches à la présence d'antifongique[98]. Le reséquençage de souches cliniques en absence d'un tel traitement ne permet cependant pas de retrouver de telles variations génomiques, indiquant que la présence d'aneuploïdies résulte principalement d'une réponse rapide à un stress important chez cette espèce[99,129]. De manière intéressante, ces différentes études ont aussi mis en évidence un nombre important d'évènements de perte d'hétérozygotie, en particulier sur une partie des chromosomes 3 et 5. Ces régions comprennent des gènes clés dans la résistance aux antifongiques, notamment plusieurs gènes impliqués dans la mise en place des pompes à efflux

(*CDR1*, *CDR2, TAC1* et *MRR1*), ainsi que des gènes spécifiques à la réponse à certaines molécules retrouvées dans les traitements médicaux, tels que *ERG11* dans la résistance aux composants azolés[99]. Plus récemment, le reséquençage de 182 isolats de cette espèce a permis une exploration plus détaillée de ces évènements au sein de la population[129]. L'exploration comparée des différents clusters a montré que bien qu'une majorité des régions présentant une perte d'hétérozygotie soit unique à chaque population, certaines d'entre elles sont partagées entre différents clusters, suggérant une sélection de ces signatures génétiques au cours de l'évolution. Dans la même étude, la comparaison des génomes de souches présentant une virulence plus faible avec le reste de la population a permis d'identifier des gènes responsables de ce trait phénotypique. Cette analyse a amené à l'identification des facteurs de transcription SFL1 et ZCF29, impliqués dans la réponse globale lors d'une infection et à la régulation de la morphogénèse.

Parallèlement, des études portant sur l'exploration des processus évolutif d'autres espèces présentant un mode de vie similaire ont été initiées. Dans ce cadre, plusieurs études se sont focalisées sur l'espèce *Candida glabrata* qui appartient au clade des *Nakaseomyces*. Malgré sa dénomination proche de celle de *C. albicans*, l'étude de son génome a montré que cette espèce est phylogénétiquement plus proche de *S. cerevisiae* que de *C. albicans*[31], suggérant que les caractères infectieux de ces deux espèces ont été acquis indépendamment. La comparaison des 6 espèces contenues dans le clade des *Nakaseomyces* a donné des premiers indices sur les bases génétiques impliquées dans le mode de vie infectieux de *C. glabrata*[34]. Dans cette étude, un plus grand nombre de gènes de la famille *EPA,* impliqués dans l'adhésion aux parois cellulaires, a été retrouvé chez les espèces pathogènes et 18 membres de cette famille ont pu être identifiés chez *C. glabrata*. En comparaison, l'espèce non-pathogène *Nakaseomyces delphensis* appartenant au même clade possède une seule copie, suggérant un avantage sélectif de ce caractère dans la pathogénicité chez les espèces de ce clade. Une étude de génomique des populations de *C. glabrata*, basée sur la comparaison entre la variabilité intraspécifique et la résistance d'une douzaine de souches cliniques a permis d'identifier plusieurs variants génétiques potentiellement impliqués dans la pathogénicité de l'espèce[130]. Ces variants impactent notamment des gènes impliqués dans la résistance aux composant azolés et aux antifongiques du groupe des échinocandines. Plus récemment, le reséquençage de 33 isolats de cette espèce a permis l'obtention d'une vue plus exhaustive des processus évolutifs impliqués chez *C. glabrata*[103]. De manière intéressante, tous les isolats retenus sont haploïdes, alors que les isolats de *C. albicans* sont majoritairement retrouvés sous forme diploïde. Ces génomes présentent cependant une plasticité importante et plusieurs aneuploïdies ainsi que des variants structuraux ont pu être identifiés. Comme mentionné précédemment, ce type de variants est aussi communément retrouvé chez *C. albicans* et pour les souches cliniques de *S. cerevisiae*, suggérant un recours similaire à ce mécanisme adaptatif pour l'ensemble des espèces pathogènes au sein du sous-phylum. Parallèlement, l'analyse comparative des variants génétiques entre les souches a permis l'identification de gènes impliqués dans le caractère infectieux de l'espèce. Par exemple, des variants du nombre de copies ont été détectés pour les gènes *PWP4* et *AWP13*, impliqués dans l'adhésion des parois cellulaires, nécessaires lors des processus d'infection. La majorité de ces gènes est retrouvée dans les régions télomériques,

confirmant une nouvelle fois l'intérêt de ces régions et de ce type de variants structuraux dans l'adaptation globale des levures.

Références

1. Goffeau, A. *et al.* Life with 6000 genes. *Science (80-.).* **274,** 546–567 (1996).
2. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–62 (1997).
3. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–8 (1998).
4. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science (80-.).* **287,** 2185–2195 (2000).
5. Initiative, T. A. G. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
6. Chinwalla, A. T. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
7. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
8. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–51 (2001).
9. Qiu, Q. *et al.* The yak genome and adaptation to life at high altitude. *Nat. Genet.* **44,** 946–949 (2012).
10. Gulcher, J. & Stefansson, K. Population genomics: laying the groundwork for genetic disease modeling and targeting. *Clin. Chem. Lab. Med.* **36,** 523–527 (1998).
11. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
12. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
13. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).
14. Alonso-Blanco, C. *et al.* 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **0,** 249–258 (2016).
15. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43,** 956–63 (2011).
16. Andersen, E. C. *et al.* Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nat. Genet.* **44,** 285–90 (2012).
17. Mackay, T. F. C. *et al.* The *Drosophila melanogaster* genetic reference panel. *Nature* **482,** 173–8 (2012).
18. Lachance, J. & Tishkoff, S. A. Population genomics of human adaptation. *Annu. Rev. Ecol. Evol. Syst.* **44,** 123–143 (2013).
19. Beall, C. M. Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. *Integr. Comp. Biol.* **46,** 18–24 (2006).
20. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39,** 31–40 (2007).
21. Verstockt, B., Smith, K. G. & Lee, J. C. Genome-wide association studies in Crohn's disease: Past, present and future. *Clin. Transl. Immunol.* **7,** e1001 (2018).
22. Mosquera, T. *et al.* Targeted and untargeted approaches unravel novel candidate genes and diagnostic snps for quantitative resistance of the potato (*Solanum tuberosum* L.) to phytophthora infestans causing the late blight disease. *PLoS One* **11,** e0156254 (2016).
23. Pantalião, G. F. *et al.* Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. *Genetica* **144,** 651–664 (2016).
24. Branham, S. E., Wright, S. J., Reba, A. & Linder, C. R. Genome-Wide Association Study of *Arabidopsis thaliana* identifies determinants of natural variation in seed oil composition. *J. Hered.* **107,** 248–256 (2016).
25. Hittinger, C. T., Steele, J. L. & Ryder, D. S. Diverse yeasts for diverse fermented beverages and foods. *Curr. Opin. Biotechnol.* **49,** 199–206 (2018).
26. Taylor, L. H., Latham, S. M. & Woolhouse, M. E. Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **356,** 983–9 (2001).
27. Peter, J. *et al.* Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556,** 339–344 (2018).

28. Vaughan-Martini, A. & Martini, A. Facts, myths and legends on the prime industrial microorganism. *J. Ind. Microbiol.* **14,** 514–522 (1995).
29. Hittinger, C. T. *et al.* Genomics and the making of yeast biodiversity. *Curr. Opin. Genet. Dev.* **35,** 100–109 (2015).
30. Dujon, B. A. & Louis, E. J. Genome diversity and evolution in the budding yeasts (Saccharomycotina). *Genetics* **206,** (2017).
31. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430,** 35–44 (2004).
32. Marck, C. *et al.* The RNA polymerase III-dependent family of genes in hemiascomycetes: comparative RNomics, decoding strategies, transcription and evolutionary implications. *Nucleic Acids Res.* **34,** 1816–35 (2006).
33. Santos, M. A. S., Gomes, A. C., Santos, M. C., Carreto, L. C. & Moura, G. R. The genetic code of the fungal CTG clade. *Comptes Rendus - Biologies* **334,** 607–611 (2011).
34. Gabaldón, T. *et al.* Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics* **14,** 623 (2013).
35. Butler, G. *et al.* Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* **459,** 657–662 (2009).
36. Bensasson, D. *et al.* Diverse lineages of *Candida albicans* live on old oaks. *bioRxiv* 341032 (2018). doi:10.1101/341032
37. de Melo Pereira, G. V. *et al.* Isolation, selection and evaluation of yeasts for use in fermentation of coffee beans by the wet process. *Int. J. Food Microbiol.* **188,** 60–66 (2014).
38. Gabaldón, T. & Carreté, L. The birth of a deadly yeast: tracing the evolutionary emergence of virulence traits in *Candida glabrata*. *FEMS Yeast Res.* **16,** fov110 (2016).
39. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. **387,** 708–713 (1997).
40. Aury, J.-M. *et al.* Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444,** 171–8 (2006).
41. Vision, T. J., Brown, D. G. & Tanksley, S. D. The origins of genomic duplications in *Arabidopsis*. *Science* **290,** 2114–7 (2000).
42. Hanson, S. J. & Wolfe, K. H. An evolutionary perspective on yeast mating-type switching. *Genetics* **206,** 9–32 (2017).
43. Barsoum, E., Martinez, P. & Aström, S. U. Alpha3, a transposable element that promotes host sexual reproduction. *Genes Dev.* **24,** 33–44 (2010).
44. Rusche, L. N. & Rine, J. Switching the mechanism of mating type switching: a domesticated transposase supplants a domesticated homing endonuclease. *Genes Dev.* **24,** 10–4 (2010).
45. Fabre, E. *et al.* Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. *Mol. Biol. Evol.* **22,** 856–873 (2005).
46. Boisnard, S. *et al.* Efficient mating-type switching in *Candida glabrata* induces cell death. *PLoS One* **10,** e0140990 (2015).
47. Brisse, S. *et al.* Uneven distribution of mating types among genotypes of *Candida glabrata* isolates from clinical samples. *Eukaryot. Cell* **8,** 287–95 (2009).
48. Dodgson, A. R., Pujol, C., Pfaller, M. A., Denning, D. W. & Soll, D. R. Evidence for recombination in *Candida glabrata*. *Fungal Genet. Biol.* **42,** 233–243 (2005).
49. Dujon, B. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet.* **22,** 375–387 (2006).
50. Vakirlis, N. *et al.* Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res.* **26,** 918–32 (2016).
51. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440,** 341–5 (2006).
52. Drillon, G. & Fischer, G. Comparative study on synteny between yeasts and vertebrates. *C. R. Biol.* **334,** 629–638 (2011).
53. Cai, J., Zhao, R., Jiang, H. & Wang, W. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179,** 487–96 (2008).
54. Li, D. *et al.* A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* **20,** 408–420 (2010).

55. Vakirlis, N. *et al.* A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.* **35,** 631–645 (2018).
56. Gibson, B. R., Storgårds, E., Krogerus, K. & Vidgren, V. Comparative physiology and fermentation performance of Saaz and Frohberg lager yeast strains and the parental species *Saccharomyces eubayanus*. *Yeast* **30,** 255–266 (2013).
57. Dunn, B. & Sherlock, G. Reconstruction of the genome origins and evolution of the hybrid lager yeast *Saccharomyces pastorianus*. *Genome Res.* **18,** 1610–23 (2008).
58. Libkind, D. *et al.* Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 14539–14544 (2011).
59. Louis, V. L. *et al. Pichia sorbitophila*, an interspecies yeast hybrid, reveals early steps of genome resolution after polyploidization. *G3 (Bethesda).* **2,** 299–311 (2012).
60. Pryszcz, L. P., Németh, T., Gácser, A. & Gabaldón, T. Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome Biol. Evol.* **6,** 1069–78 (2014).
61. Schröder, M. S. *et al.* Multiple origins of the pathogenic yeast *Candida orthopsilosis* by separate hybridizations between two parental species. *PLOS Genet.* **12,** e1006404 (2016).
62. Okuno, M. *et al.* Next-generation sequencing analysis of lager brewing yeast strains reveals the evolutionary history of interspecies hybridization. *DNA Res.* **23,** 67–80 (2016).
63. Gojković, Z. *et al.* Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. *Mol. Genet. Genomics* **271,** 387–393 (2004).
64. Shen, X.-X. *et al.* Reconstructing the backbone of the saccharomycotina yeast phylogeny using genome-scale data. *G3;#58; Genes|Genomes|Genetics* 070235 (2016). doi:10.1534/g3.116.034744
65. Opulente, D. A. *et al.* Factors driving metabolic diversity in the budding yeast subphylum. *BMC Biol.* **16,** 26 (2018).
66. Gonçalves, C., Coelho, M. A., Salema-Oom, M. & Gonçalves, P. Stepwise functional evolution in a fungal sugar transporter family. *Mol. Biol. Evol.* **33,** 352–366 (2016).
67. Gonçalves, C. *et al.* Evidence for loss and adaptive reacquisition of alcoholic fermentation in an early-derived fructophilic yeast lineage. *Elife* **7,** e33034 (2018).
68. Schacherer, J., Shapiro, J. a, Ruderfer, D. M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458,** 342–5 (2009).
69. Bergström, A. *et al.* A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31,** 872–888 (2014).
70. Legras, J.-L. *et al.* Adaptation of *S . cerevisiae* to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. *Mol. Biol. Evol.* (2018). doi:10.1093/molbev/msy066
71. Duan, S.-F. *et al.* The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nat. Commun.* **9,** 2690 (2018).
72. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458,** 337–41 (2009).
73. Zhu, Y. O., Sherlock, G. & Petrov, D. A. Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3 (Bethesda).* (2016). doi:10.1534/g3.116.029397
74. Strope, P. K. *et al.* The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* gr.185538.114- (2015). doi:10.1101/gr.185538.114
75. Almeida, P. *et al.* A population genomics insight into the mediterranean origins of wine yeast domestication. *Mol. Ecol.* (2015). doi:10.1111/mec.13341
76. Ludlow, C. L. *et al.* Independent origins of yeast associated with coffee and cacao fermentation. *Curr. Biol.* **26,** 965–71 (2016).
77. Gonçalves, M. *et al.* Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Curr. Biol.* **26,** 2750–2761 (2016).
78. Gallone, B. *et al.* Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* **166,** 1397–1410.e16 (2016).

79. Barbosa, R. *et al.* Evidence of natural hybridization in Brazilian wild lineages of *Saccharomyces cerevisiae*. *Genome Biol. Evol.* evv263 (2016). doi:10.1093/gbe/evv263
80. Wang, Q. M., Liu, W. Q., Liti, G., Wang, S. A. & Bai, F. Y. Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. *Mol. Ecol.* **21,** 5404–5417 (2012).
81. Boynton, P. J. & Greig, D. The ecology and evolution of non-domesticated *Saccharomyces* species. *Yeast* **31,** n/a-n/a (2014).
82. Liti, G. *et al.* High quality de novo sequencing and assembly of the *Saccharomyces arboricolus* genome. *BMC Genomics* **14,** 69 (2013).
83. Hittinger, C. T. *et al.* Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* **464,** 54–58 (2010).
84. Magwene, P. M. *et al.* Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1987–1992 (2011).
85. Mulla, W., Zhu, J. & Li, R. Yeast: a simple model system to study complex phenomena of aneuploidy. *FEMS Microbiol. Rev.* **38,** 201–212 (2014).
86. Yona, A. H. *et al.* Chromosomal duplication is a transient evolutionary solution to stress. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 21010–5 (2012).
87. Rancati, G. *et al.* Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. *Cell* **135,** 879–93 (2008).
88. Hughes, T. R. *et al.* Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat. Genet.* **25,** 333–337 (2000).
89. Dunham, M. J. *et al.* Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 16144–9 (2002).
90. Pavelka, N. *et al.* Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature* **468,** 321–5 (2010).
91. Brown, C. A., Murray, A. W. & Verstrepen, K. J. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr. Biol.* **20,** 895–903 (2010).
92. Muller, L. A. H., Lucas, J. E., Georgianna, D. R. & McCusker, J. H. Genome-wide association analysis of clinical vs. nonclinical origin provides insights into *Saccharomyces cerevisiae* pathogenesis. *Mol. Ecol.* **20,** 4085–97 (2011).
93. Connelly, C. F. & Akey, J. M. On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*. *Genetics* **191,** 1345–53 (2012).
94. Diao, L. & Chen, K. C. Local ancestry corrects for population structure in *Saccharomyces cerevisiae* genome-wide association studies. *Genetics* **192,** 1503–1511 (2012).
95. Pérez-Ortín, J. E., Querol, A., Puig, S. & Barrio, E. Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome Res.* **12,** 1533–9 (2002).
96. Almeida, P. *et al.* A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. *Nat. Commun.* **5,** 4044 (2014).
97. Peris, D. *et al.* Population structure and reticulate evolution of *Saccharomyces eubayanus* and its lager-brewing hybrids. *Mol. Ecol.* **23,** 2031–2045 (2014).
98. Hirakawa, M. P. *et al.* Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Res.* gr.174623.114- (2015). doi:10.1101/gr.174623.114
99. Ford, C. B. *et al.* The evolution of drug resistance in clinical isolates of *Candida albicans*. *Elife* **4,** e00662 (2015).
100. Fawcett, J. A. *et al.* Population genomics of the fission yeast *Schizosaccharomyces pombe*. *PLoS One* **9,** e104241 (2014).
101. Jung, P. P., Friedrich, A., Reisser, C., Hou, J. & Schacherer, J. Mitochondrial genome evolution in a single protoploid yeast species. *G3 (Bethesda)*. **2,** 1103–11 (2012).
102. Jeffares, D. C. *et al.* The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. *Nat. Genet.* **advance on,** (2015).
103. Carreté, L. *et al.* Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. *Curr. Biol.* **28,** 15–27.e7 (2018).
104. Ortiz-Merino, R. A. *et al.* Ploidy variation in *Kluyveromyces marxianus* separates dairy and non-dairy isolates. *Front. Genet.* **9,** 94 (2018).

105. Tsai, I. J., Bensasson, D., Burt, A. & Koufopanou, V. Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 4957–62 (2008).
106. Vishnoi, A., Sethupathy, P., Simola, D., Plotkin, J. B. & Hannenhalli, S. Genome-wide survey of natural selection on functional, structural, and network properties of polymorphic sites in *Saccharomyces paradoxus*. *Mol. Biol. Evol.* **28,** 2615–27 (2011).
107. Leducq, J.-B. *et al.* Local climatic adaptation in a widespread microorganism. *Proc. Biol. Sci.* **281,** 20132472 (2014).
108. Bergström, A. *et al.* A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31,** 872–888 (2014).
109. Leducq, J.-B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat. Microbiol.* **1,** 15003 (2016).
110. Xia, W. *et al.* Population genomics reveals structure at the individual, host-tree scale and persistence of genotypic variants of the undomesticated yeast *Saccharomyces paradoxus* in a natural woodland. *Mol. Ecol.* **26,** 995–1007 (2017).
111. Kuehne, H. A., Murphy, H. A., Francis, C. A. & Sniegowski, P. D. Allopatric divergence, secondary contact, and genetic isolation in wild yeast populations. *Curr. Biol.* **17,** 407–411 (2007).
112. Hou, J., Fournier, T. & Schacherer, J. Species-wide survey reveals the various flavors of intraspecific reproductive isolation in yeast. *FEMS Yeast Res.* **16,** (2016).
113. Charron, G., Leducq, J. B. & Landry, C. R. Chromosomal variation segregates within incipient species and correlates with reproductive isolation. *Mol. Ecol.* 4362–4372 (2014). doi:10.1111/mec.12864
114. Bing, J., Han, P.-J., Liu, W.-Q., Wang, Q.-M. & Bai, F.-Y. Evidence for a Far East Asian origin of lager beer yeast. *Curr. Biol.* **24,** R380-1 (2014).
115. Tadami, H., Shikata-Miyoshi, M. & Ogata, T. Aneuploidy, copy number variation and unique chromosomal structures in bottom-fermenting yeast revealed by array-CGH. *J. Inst. Brew.* **120,** 27–37 (2014).
116. Novo, M. *et al.* 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,** 16333–8 (2009).
117. Møller, K., Sharif, M. Z. & Olsson, L. Production of fungal α-amylase by *Saccharomyces kluyveri* in glucose-limited cultivations. *J. Biotechnol.* **111,** 311–318 (2004).
118. Gojkovic, Z., Paracchini, S. & Piskur, J. in 475–479 (Springer, Boston, MA, 1998). doi:10.1007/978-1-4615-5381-6_94
119. Gojkovic, Z., Jahnke, K., Schnackerz, K. D. & Piškur, J. *PYD2* encodes 5,6-dihydropyrimidine amidohydrolase, which participates in a novel fungal catabolic pathway. *J. Mol. Biol.* **295,** 1073–1087 (2000).
120. Beck, H., Dobritzsch, D. & PiÅ¡kur, J. *Saccharomyces kluyveri* as a model organism to study pyrimidine degradation. *FEMS Yeast Res.* **8,** 1209–1213 (2008).
121. Souciet, J.-L. *et al.* Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res.* **19,** 1696–709 (2009).
122. Payen, C. *et al.* Unusual composition of a yeast chromosome arm is associated with its delayed replication. *Genome Res.* **19,** 1710–21 (2009).
123. Friedrich, A., Jung, P., Reisser, C., Fischer, G. & Schacherer, J. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol. Biol. Evol.* **32,** 184–92 (2015).
124. Brion, C., Pflieger, D., Friedrich, A. & Schacherer, J. Evolution of intraspecific transcriptomic landscapes in yeasts. *Nucleic Acids Res.* **43,** 4558–68 (2015).
125. Brion, C. *et al.* Variation of the meiotic recombination landscape and properties over a broad evolutionary distance in yeasts. *PLoS Genet.* **13,** e1006917 (2017).
126. Warringer, J. *et al.* Trait variation in yeast is defined by population history. *PLoS Genet.* **7,** e1002111 (2011).
127. Mendes Giannini, M. J. S., Bernardi, T., Scorzoni, L., Fusco-Almeida, A. M. & Sardi, J. C. O. *Candida* species: current epidemiology, pathogenicity, biofilm formation, natural antifungal products and new therapeutic options. *J. Med. Microbiol.* **62,** 10–24 (2013).
128. Jones, T. *et al.* The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci.*

101, 7329–7334 (2004).
129. Ropars, J. *et al.* Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. *Nat. Commun.* **9,** 2253 (2018).
130. Biswas, C. *et al.* Identification of genetic markers of resistance to echinocandins, azoles and 5-fluorocytosine in *Candida glabrata* by next-generation sequencing: a feasibility study. *Clin. Microbiol. Infect.* **23,** 676.e7-676.e10 (2017).

VUE D’ENSEMBLE DU PROJET

La génomique des populations offre la possibilité d'obtenir une vue globale de l'évolution au sein d'une même espèce. Au niveau des levures du sous-phylum des Saccharomycotina, l'exploration de plusieurs centaines de génomes de *S. cerevisiae* a d'ores et déjà permis une dissection précise de son histoire évolutive. Des différences évolutives entre les sous-populations ont ainsi été mises en évidence, révélant l'impact des environnements associés à chacune d'entre elles. En parallèle, plusieurs études de génomique des populations menées sur des espèces non-modèles ont révélé des différences évolutives entre les espèces tant dans leur histoire évolutive que dans les modifications génomiques impliquées dans l'adaptation des espèces. L'exploration et la comparaison de l'évolution d'un plus grand nombre d'espèces se révèlent ainsi critiques afin de mieux comprendre les processus responsables de l'évolution des espèces. Cet axe de recherche est cependant toujours limité par différentes contraintes. Tout d'abord, d'importantes différences dans les outils et les analyses employés ainsi que des variants étudiés rendent difficile une comparaison précise de l'évolution entre les espèces. Par ailleurs, l'absence de séquence de référence et d'annotations pour un grand nombre d'espèces limitent encore l'établissement d'études de génomique des populations et seule une dizaine d'espèces ont été étudiées jusqu'à présent.

C'est dans ce cadre que se positionne mon projet de thèse. **Dans un premier chapitre**, je me suis concentré sur l'étude comparative de l'évolution au sein de 6 espèces non-modèles du sous-phylum des Saccharomycotina pour lesquelles une séquence de référence et des annotations de bonnes qualités avaient déjà été préalablement établies : *Candida albicans*, *Kluyveromyces lactis*, *Lachancea kluyveri*, *Lachancea thermotolerans*, *Saccharomyces paradoxus* et *Saccharomyces uvarum*. Ces espèces sont retrouvées dans différents habitats et certaines d'entre elles sont partiellement domestiquées. De plus, ces espèces couvrent différentes échelles évolutives, permettant la comparaison de la variabilité intraspécifique entre des espèces phylogénétiquement proches ou éloignées au sein du sous-phylum. Afin d'effectuer une comparaison globale de l'évolution de ces espèces, plusieurs types de variants allant de la duplication complète d'un chromosome à la diversité nucléotidique ont été quantifiés de manière systématique. Dans le cadre de cette étude, je me suis concentré sur l'analyse bio-informatique des données de séquençage (acquises au sein du laboratoire ou disponibles publiquement) et j'ai notamment développé une librairie informatique dédiée à l'exploration et l'exploitation de ces données. Cette étude a permis de mettre en relief des différences dans l'histoires évolutive des espèces. À l'opposé, certaines trajectoires évolutives semblent partagées entre les espèces. Par exemple, l'analyse comparative du pangénome de chaque espèce a ainsi révélé l'importance des évènements d'introgressions dans la dynamique des génomes au cours du temps pour l'ensemble des espèces étudiées.

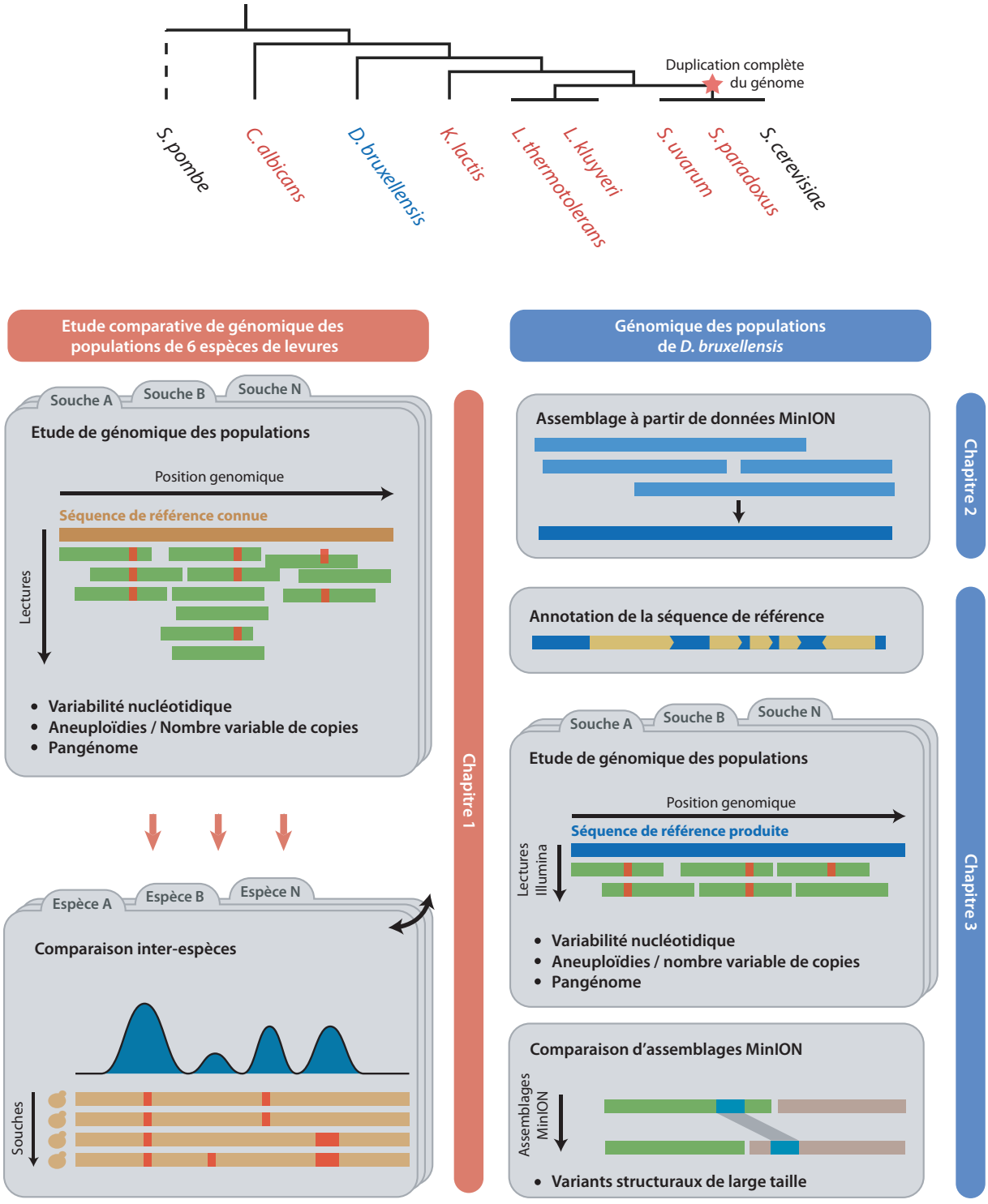


Figure 1. Représentation schématique de mon projet de thèse.

Dans un second temps, je me suis concentré sur l'exploration de la variabilité génétique au sein de l'espèce *Dekkera bruxellensis*, qui présente des propriétés industrielles et des particularités évolutives intéressantes. En effet, *D. bruxellensis* contribue à la fermentation de certaines bières belges, mais est aussi connue comme un contaminant majeur du vin. De plus, cette espèce présente un intérêt au niveau évolutif : les isolats naturels arborent différents niveaux de ploïdie et des réarrangements chromosomiques étendus ont pu être observés. Dans ce cadre, une

collection de 56 souches a été constituée afin de réaliser une analyse de génomique des populations. Cependant, alors que les séquences des génomes de plusieurs souches ont déjà été déterminées et mises à disposition de la communauté scientifique, aucune n'associait une résolution au niveau chromosomique et des annotations de qualité, nécessaires à l'établissement d'une étude de génomique des populations. Le développement des technologies dites de 3ème génération, qui génèrent des lectures longues (plusieurs kb) permettent aujourd'hui d'envisager l'obtention rapide et aisée d'assemblages de bonnes qualités. **Dans un second chapitre**, je me suis ainsi attelé à la création d'un génome de référence de qualité pour l'espèce *D. bruxellensis* à l'aide de données de séquençage produites par la technologie Oxford Nanopore. La comparaison de différents logiciels d'assemblage développés récemment et dédiés à l'exploitation de ces lectures a permis l'obtention d'une séquence de référence d'une qualité supérieure en comparaison à celles obtenues précédemment. Par cette étude, nous avons montré que la mise à disposition de ces nouvelles technologies de séquençage offre la possibilité d'explorer la variabilité génomique au sein d'espèces jusqu'alors inaccessibles de par l'absence de séquence de référence de bonne qualité.

Enfin, dans **un troisième chapitre**, les données de séquençage des 56 souches ont été exploitées afin de réaliser une étude de génomique des populations au sein de *D. bruxellensis*. L'exploration de la diversité génétique a dans un premier temps permis la mise en évidence de plusieurs sous-populations et l'identification et la caractérisation de plusieurs évènements d'hybridations au cours de l'histoire évolutive de l'espèce. Les données de séquençage ont par la suite été exploitées afin de caractériser le pangénome ainsi que d'étudier la variabilité du contenu en gènes au sein de l'espèce. Cette analyse a été complétée par le séquençage de 3 souches à partir d'une technologie de type Oxford Nanopore permettant l'identification des réarrangements chromosomiques indétectables avec des données de type Illumina. L'ensemble de ces données ont permis de caractériser des variants structuraux, affectant notamment des transporteurs de sucres ou de drogues et jouant potentiellement un rôle dans l'adaptation des souches dans les différents processus de fermentation dans lesquels intervient *D. bruxellensis*.

CHAPTER 1

Pangenomes of Saccharomycotina yeast species are mainly shaped by introgression events

Résumé

Au sein de chaque espèce, la diversité phénotypique est liée à la variabilité génétique observée entre individus. Cette dernière repose sur des facteurs propres à l'espèce, comme ses taux de mutation et de recombinaison, son cycle sexuel, sa démographie et les éventuelles pressions de sélection auxquelles elle peut être soumise. Le poids de ces facteurs varie d'une espèce à l'autre rendant leur évolution unique. Afin d'appréhender leur impact relatif au sein de chaque espèce, des études dites de génomique des populations peuvent être réalisées en se basant sur l'analyse comparative des génomes d'un grand nombre d'individus. Depuis une dizaine d'années, plusieurs études de génomique des populations ont ainsi été produites au sein d'espèces non-modèles, permettant d'obtenir une première vue des différences dans l'histoire évolutive entre les espèces. Cependant, d'importantes disparités dans les méthodologies utilisées ainsi que le choix des variants génétiques analysés rendent difficile une comparaison globale de la variabilité intraspécifique entre les espèces.

Dans ce contexte, nous nous sommes intéressés à l'établissement d'une étude de génomique des populations comparative à partir de données de séquençage de génomes complets provenant à la fois du domaine public ou produites au sein du laboratoire. Plus précisément, six espèces ont été analysées : *K. lactis*, *L. thermotolerans* pour lesquelles des collections ont été constituées spécifiquement dans le cadre de cette étude, ainsi que *C. albicans*, *L. kluyveri*, *S. paradoxus* et *S. uvarum* dont des génomes ont été préalablement séquencés pour de précédentes études. Pour chacune d'entre elles, plusieurs types de variants allant de la duplication complète d'un chromosome à la divergence nucléotidique ont été quantifiés. De manière intéressante, l'impact des différents facteurs est variable entre les espèces, démontrant d'importantes différences dans l'évolution des espèces. Dans un second temps, nous nous sommes focalisés sur l'analyse de la variabilité du contenu en gènes au sein des différentes espèces dont la diversité est encore largement inexplorée au sein des espèces non-modèles de levure. Dans ce cadre, le nombre de copies a été identifié pour l'ensemble des gènes de la référence et les gènes supplémentaires ont été identifiés à travers la construction et la comparaison d'assemblages *de novo* de chaque souche contre les génomes de référence. Cette approche a permis d'obtenir le core génome (gènes présents systématiquement dans l'ensemble des souches) et le pan génome (ensemble de gènes retrouvés dans la population) pour chacune des espèces. Les résultats ont montré d'importantes variations dans la constitution des core et pangénomes entre les espèces, indiquant que la variabilité du contenu en gènes est une composante importante de la diversité génétique au sein des levures. De manière intéressante, un grand nombre d'introgressions provenant d'espèces proches a pu être retrouvé, suggérant la présence de multiples évènements d'hybridation entre les différentes espèces de levure. Dans l'ensemble, ces résultats démontrent l'importance de l'étude de variabilité du répertoire de gènes dans l'étude de l'évolution et de l'adaptation des espèces.

Introduction

Genetic variation results from a wide assortment of evolutionary forces, among which mutation, structural variation and selection play major roles in shaping genomes. These factors impact differently genome evolution of various species, making their evolutionary history unique. Using large-scale polymorphic data, population genomic surveys aim to dissect the evolutionary history by analyzing genetic variants from a high number of individuals. Due to their small and compact genome, yeasts are well-suited organisms for population genomic studies. Moreover, the considerable levels of genetic and phenotypic diversity displayed in yeast natural populations allows them to adapt to changing environments and to new ecological niches. During the last decade, yeast population genomics has been mainly focused on the model organism *S. cerevisiae* and recently more than a thousand of isolates have been sequenced and analyzed[1]. This recent study allowed a deep investigation of its evolutionary history as well as a better understanding of the driving forces of genome evolution, the genomic variations between subpopulations, and the genotype-phenotype relationship[1].

While *S. cerevisiae* population genomic studies have provided a comprehensive view of multiple aspects of this species[1–8], less is known about other non-conventional yeast species. However, it is now increasingly clear that our knowledge in biology will be greatly improved by exploring a broader diversity of species. In this context, yeast species from the Saccharomycotina subphylum show interesting characteristics. Indeed, thanks to their diverse metabolisms, yeasts exhibit a remarkable range of ecologies including natural niches (plants, soil, insects), domesticated processes (fermented beverage, bread, dairies products) and some species are mostly found as opportunistic pathogenic organisms[9]. Moreover, extensive comparative genomic analyses during the last decades have been produced, revealing significant variations in the genomic architectures between species of this subphylum[10]. During the past years, several populations genomic studies have been established, aiming to have a better view of the evolutionary patterns of non-model yeast species[11–17]. While these studies have been mainly restricted to reduced datasets compared to *S. cerevisiae* analyses (usually under a hundred of isolates), results have already enhanced insights into the differences and similarities which can be found in the evolutionary histories and forces involved in genome evolution[10,14–16,18,19]. These studies are valuable assets in our understanding of the genetic variability within yeasts, however the task of comprehensively exploring yeast genetic diversity remains uncomplete. First only a limited number of species have been investigated and the global view of the genomic variation occurring in yeasts is still unknown. Secondly, these analyses have been completed using independent methods, making a systematic comparison between species tedious.

Here, we therefore decided to undertake the population genomic analysis of two unexplored non-conventional yeast species: *K. lactis* and *L. thermotolerans*. The *K. lactis* yeast is a protoploid heterothallic species with the ability to ferment several sugars including lactose, which make it widely used in dairy processes[20,21]. However, *K. lactis* isolates can also be found in natural habitats[22] and most of the wild isolates have lost the ability to ferment lactose. This

characteristic leads to the formation of two distinct groups: the var. *drosophila* group containing wild isolates and the var. *lactis* representing dairy strains. Similarly, *L. thermotolerans* is a protoploid species found commonly in natural habitats. However, its ability to produce lactic acid during alcoholic fermentation lead to an increased use of *L. thermotolerans* during wine-making[23]. *K. lactis* and *L. thermotolerans* genome are both compact and distributed along 6 and 8 chromosomes respectively, with few duplicated blocks and a low sequence identity compared to *S. cerevisiae*[24,25]. To get a global view of the genomic variability within the entire Saccharomycotina subphylum, we additionally included sequencing data from two pre-duplicated species: *S. paradoxus*[26] and *S. uvarum*[5] and two others protoploid species: *L. kluyveri*[16] and *C. albicans*[13] for which publicly whole genome sequencing data are available, providing a multispecies comparison englobing a large part of the subphylum.

Using these datasets, we systematically explored the intraspecific variability within each species at the genome-scale level, including single nucleotide polymorphisms (SNPs), aneuploidies and copy number variants (CNVs), providing a complete multispecies comparison covering a broad evolutionary distance. Moreover, we generated for each species *de novo* genome assemblies, allowing us to construct the pangenome and explore the gene content within these different yeast species. Our results highlight that pangenomes are variable across species. Interestingly, within all species, accessory genomes are composed of introgressed genes coming from closely related species. This result suggests the extent of interspecific hybridization events across the entire subphylum and its strong impact on the gene content. Finally, we found a functional enrichment in accessory genomes of genes involved in secondary metabolism and flocculation, providing a source for adaptive evolution. Overall, the constructed pangenomes highlight an essential source of genetic variability poorly explored so far and show that reference genomes significantly underestimate the gene space of any yeast species.

Results

For this study, two collections of 41 and 57 isolates were assembled for *K. lactis* and *L. thermotolerans*, respectively. For both species, we deeply sequenced the genomes of each isolate using an Illumina HiSeq 100-bp sequencing strategy with a median 113 and 78-folds coverage (Figure S1). To produce a cross-species comparison along the Saccharomycotina subphylum, we have compared these generated data with whole genome sequences of *C. albicans*, *L. kluyveri, S. paradoxus* and *S. uvarum* leading to the analysis of 6 collections ranging from 22 to 57 genomes (Figure 1, Table S1).

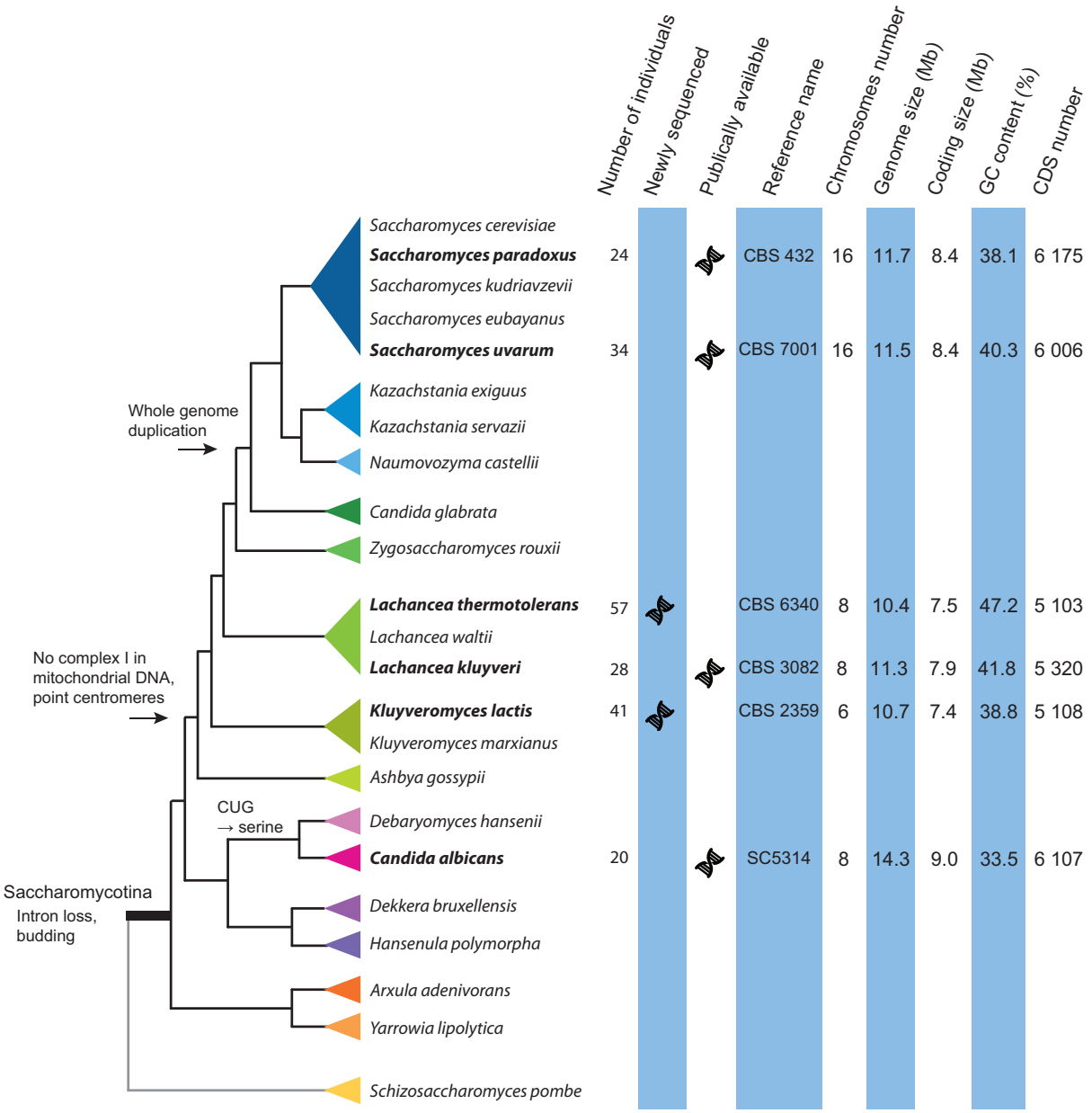


Figure 1. Species used in this study. On the left, the location of each species within the Saccharomycotina subphylum. On the right, main characteristics of the populations and of the reference genome sequences.

Samples were mostly isolated worldwide from natural habitats (*e.g.* plants, soil, insect). A subset of 15 *K. lactis* strains have been isolated in dairy-related processes while a small number of *L. thermotolerans* strains were collected in wine fermentation environments. Since *C. albicans* is an opportunistic pathogenic species, every isolate has been collected from healthy or immuno-compromised patients. Except for *C. albicans* for which samples were diploid, only haploid clones were sequenced producing medium to high coverage ranging from 39X to 170X for *S. uvarum* and *C. albicans* respectively (Figure S1).

Variation of the global pattern of polymorphisms

To obtain a first view of the nucleotide diversity at the genome-scale level, the global genetic diversity was estimated by determining the average pairwise divergence (π) and the proportion of polymorphic sites per base (θ_w) in each species. Additionally, insertions and deletions (indels) smaller than 50 bp have also been identified. To assess the impact of variants on protein sequences, annotation has been performed using SnpEff[27] and the impact on protein function for non-synonymous SNPs has been predicted using SIFT[28].

Both *K. lactis* and *L. thermotolerans* show a very high genetic diversity ($\pi = 2.8 \times 10^{-2}$ and $\pi = 1.5 \times 10^{-2}$, respectively), which is almost 10 times higher compared to the one observed in the *S. cerevisiae* species[1] ($\pi = 3 \times 10^{-3}$) (Figure 2.A). In these two species, the pairwise divergence (π) is slightly lower compared to the polymorphic site variability (θ_w), resulting to negative values of Tajima's D, sign of an excess of low-frequency polymorphisms (Table S2)

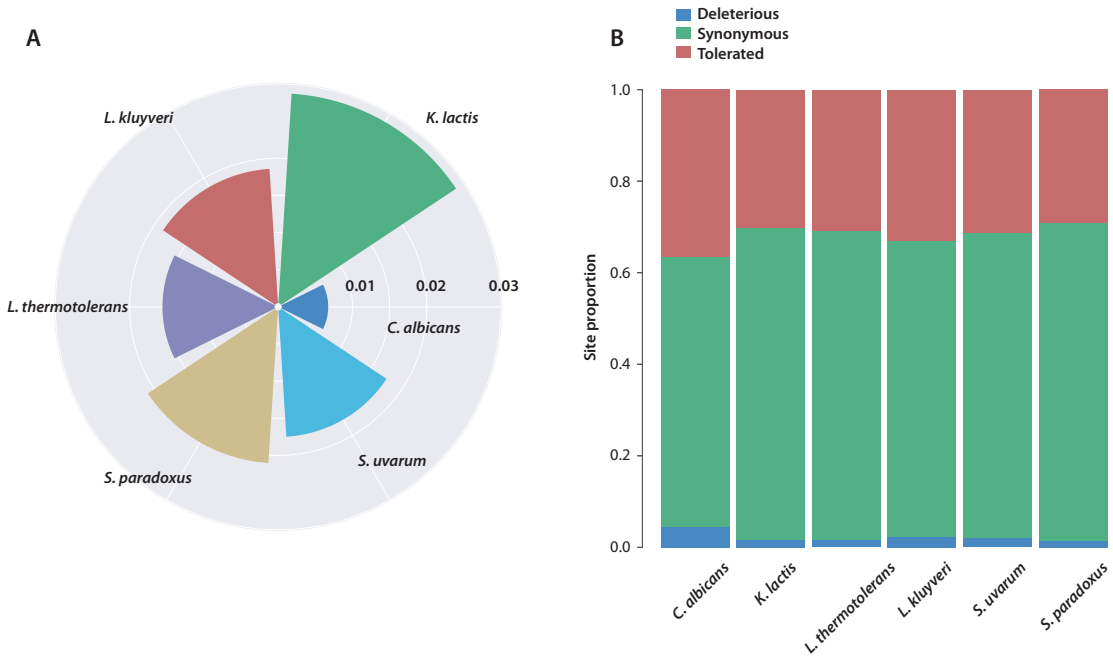


Figure 2. Nucleic metrics within the different species. (A) Estimation of the nucleic diversity π for the whole genome. (B) Proportion of synonymous (green), tolerated (ref) and deleterious (blue) SNPs.

Comparison of the nucleotide diversity between coding and non-coding region showed an increase variability in non-coding region in all species, illustrating the impact of purifying selection (Figure S3). However, the pairwise diversity remains extremely high in *K. lactis* (π_{coding} = 2.4 x 10^{-2}) and both *K. lactis* and *L. thermotolerans* display a higher number of low frequency variants in coding region, resulting to the low Tajima's D in these regions. For all species, more than 95% of the SNPs are either synonymous or do not impact protein function (Figure 2.B). Additionally, sites containing deleterious SNPs and more globally non-synonymous SNPs are enriched in low frequency, reflecting again the strong impact of purifying selection (Figure S3). However, *C. albicans* seems more permissive to non-synonymous (~ 41%) and deleterious SNPs (4.4%) compared to other species (~ 34.7% and ~ 1.8%, respectively).

In addition, *C. albicans* shows a very high number of indels with an average of 22,973 indels per isolate, which is way above of what can be observed in other species (t-test, p-value = 2.3 x 10^{-10}) (Figure 3) and might be linked to the almost obligatory diploid state of this species. Moreover, a significant higher number of insertions compared to deletions can be observed in *C. albicans* and *L. kluyveri*, while other species show a similar rate of both kind of indels (t-test, p-value = 5.6 x 10^{-4} and p-value = 8.5 x 10^{-4} respectively).

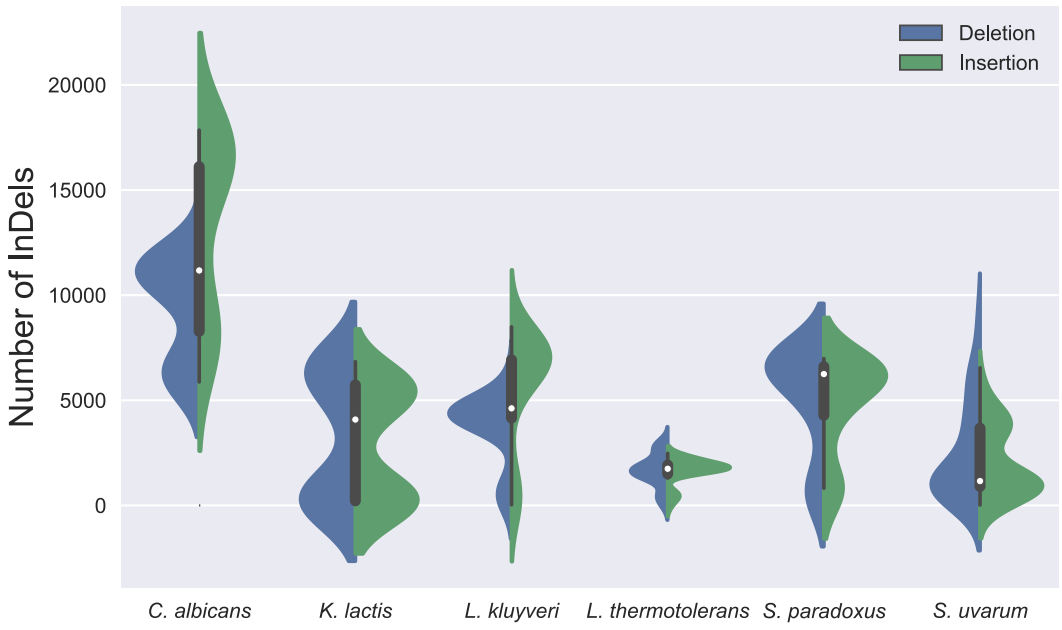


Figure 3. Distribution of the number of deletion (blue) and insertion (green) per strains across species.

Additionally, indels are strongly impacted by purifying selection in all species, non-coding regions showed a higher number of indels compared to coding region (Figure S4) and a clear enrichment of indels of multiple-of-three length can be observed within each species (chi2 – p-value = 0 for all species). Moreover, indels found in coding regions correspond mostly to low

frequency variants (Figure S5) and genes containing frameshift indels display a higher dN/dS values (t-value = 119, p-value = 0) (Figure S5). Overall, these results gave a first view of the similarities and disparities of the polymorphism landscape between yeast species. While a common pattern of selection can be observed in all species, a clear variation of the distribution of variants can be observed, especially for *C. albicans* compared to other species. Interestingly, this distribution is balanced differently between species and is likely linked either to the species lifestyle and their genome properties.

Strains relationship and population structure

To examine strains relationships within each species, we built a majority-rule consensus trees based on the bio neighbour-joining method (Figure S6). Alongside, the structure of each species has been examined based on the clustering method implemented in the STRUCTURE software[29] (Figure 4).

These analyses revealed complex patterns of differentiation including distinct subpopulations. However, these subpopulations correlate with geography for most the studied species with an exception, namely *K. lactis*. As already mentioned, this species is partially domesticated, with isolates involved in dairy processes.

The *K. lactis* species displays 5 clusters which are highly divergent between each other with a nucleotide divergence going up to 8.3% between the Asian cluster and a cluster composed of two samples of unknown origins. Within this species, clusters are highly conserved with a very low SNPs diversity, around 0.3%, between strains and almost no admixture. As observed for domesticated strains in *S. cerevisiae*, strains linked to dairy processes group into a single cluster, with an extremely low genetic divergence of 0.07, suggesting the usage of closely related isolates in this industrial field.

Similarly, *L. thermotolerans* displays 6 distinct clusters with an average of 2.1% divergence (with a maximum of 4% of genetic divergence) with a low intra-clade diversity around 0.25%. Nonetheless, as previously observed with mitochondrial genomes[30], 4 clusters out of 6 have mixed geographical origins without clear relationship with the substrates where sample were isolated. While *L. thermotolerans* is sporadically used for wine making, our dataset contains only a few number of samples directly isolated from industrial process, suggesting that the observed phylogeny could result of strains which were human associated and reintroduced in a natural environment. This hypothesis is correlated with previous clustering obtained using microsatellite markers[31], however a bigger sample size is required to fully resolve the evolutionary history of this species. Such strains distribution leads to very conserved structure for this species, with almost no admixture. As previously reported, similar clean lineages can also be observed in *L. kluyveri*[16], *S. paradoxus*[32] and *S. uvarum*[19] with a genetic diversity up to 2.9%, 3.6% and 4.9% respectively. However, strains with mixed origins can also be found within the *L. kluyveri* Asian cluster and the *S. uvarum* Argentina cluster, suggesting recent cross-breeding inside these populations.

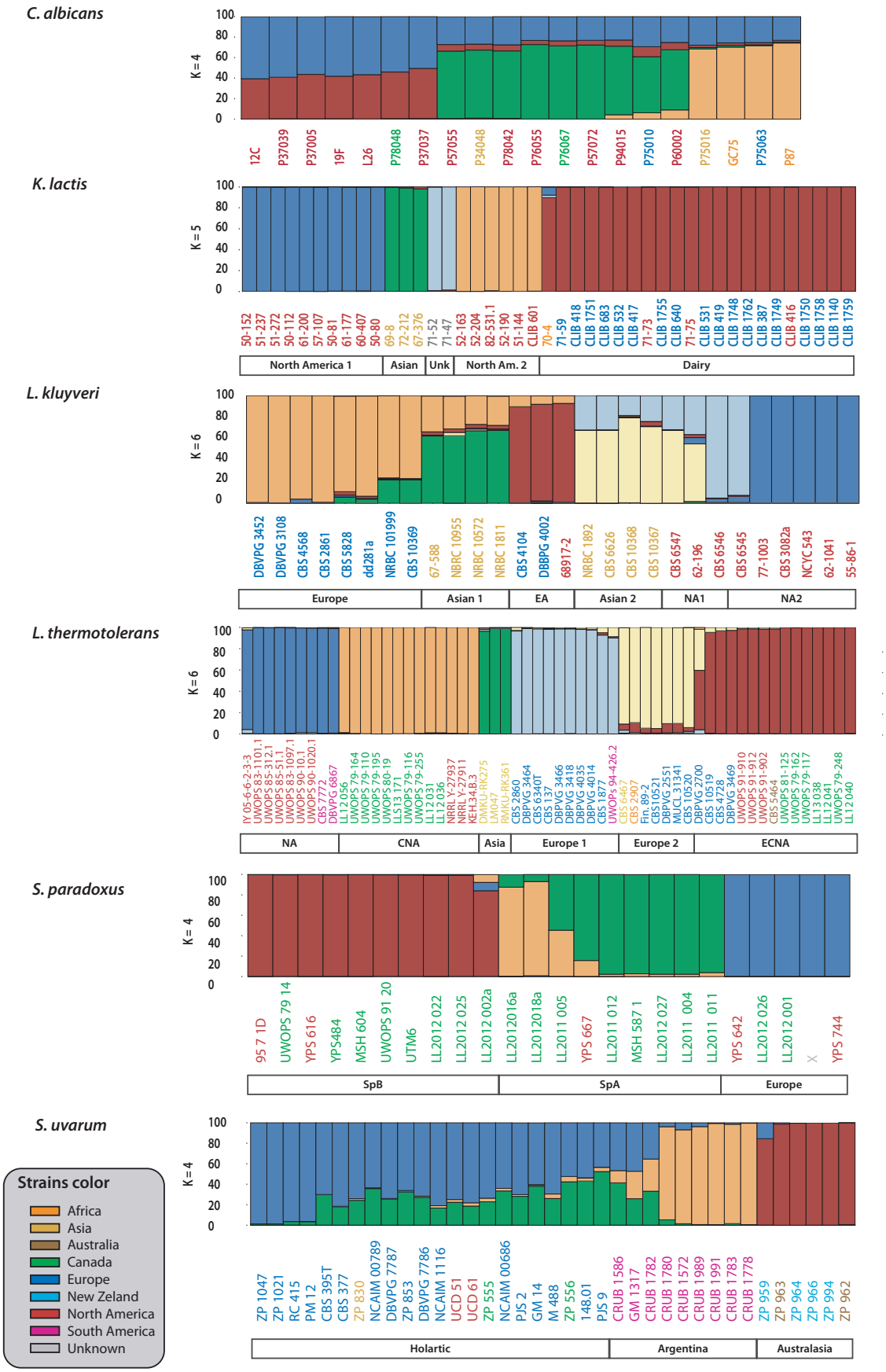


Figure 4. Population structure of all species. For each structure, the number of subpopulations was adjusted (K value, from 4 to 6). Geographical origin for each sample is represented by the color code.

Natural ploidy and large-scale duplication events

Among yeast species, different life cycle can be observed, yielding to potential different natural ploidy state. Most of the species in our studies have an haplo-diploid life cycle similar to *S cerevisiae*, allowing strain to reproduce asexually either as haploid or diploid or sexually with the mating of two haploids. At the other end, *C. albicans* undergoes a parasexual cycle wherein two diploid cells mate, and is almost only found in diploid state with some exceptions. Recent analysis of 1,011 genomes of *S. cerevisiae* demonstrates the predominance of the diploid state in natural isolates of this species and similar results have been obtained for *S. paradoxus* and *S. uvarum* suggesting a predominance of the diploid state in the *Saccharomyces* subphylum. However, little is known about the natural ploidy level in other yeast species. We therefore determined the ploidy of sequenced samples of *K. lactis, L. kluyveri* and *L. thermotolerans* using FACS analysis (Figure 5).

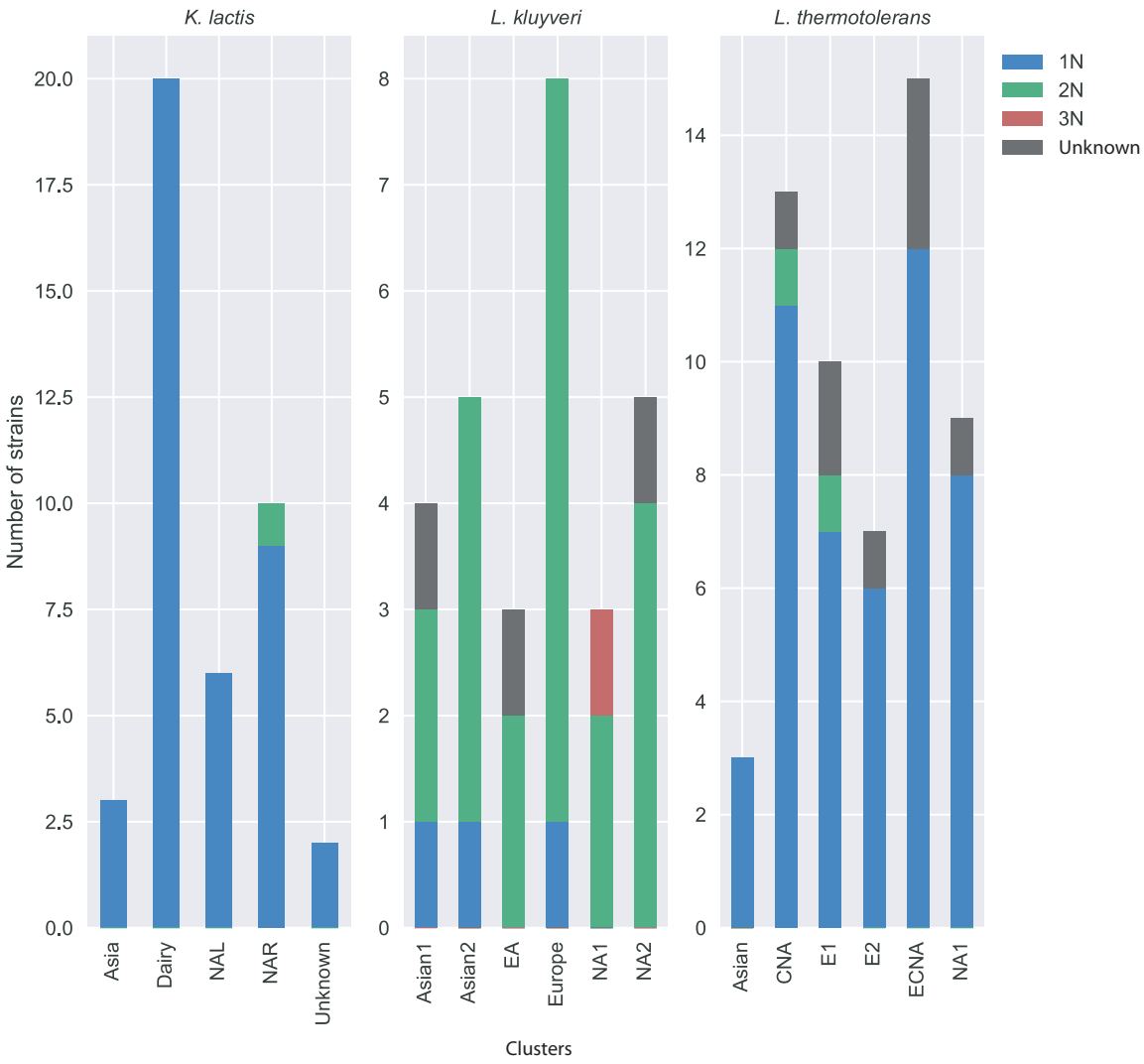


Figure 5. Natural ploidy distribution within *K. lactis*, *L. kluyveri* and *L. thermotolerans* populations based on FACS analysis and grouped by cluster.

While almost all the samples in *K. lactis* and *L. thermotolerans* have been determined as haploid, our results reveal that *L. kluyveri* samples are mostly encounter in a diploid state, showing a variation of the natural ploidy level among the Saccharomycotina subphylum. Moreover, no significant variation in the ploidy level can be observed in the dairy samples of *K. lactis*. This result is similar to previous observations for other domesticated yeast isolates like wine and sake related strains in *S. cerevisiae*, while ale beer isolates display a higher ploidy level, and can be linked to the open environment in which dairy strains are living, where *K. lactis* dairy strains are not directly monitored during cheese fermentation process, but are living freely in the cheese caves.

Aneuploidies and segmental duplications, by quickly modifying gene dosage of large chromosomal segments, are known to be a fast, common and effective response against new environmental stress in the *Saccharomyces cerevisiae* yeast. To examine the frequency of aneuploidies in our populations, we assigned the presence of aneuploidy and segmental duplications in each strain based on the coverage distribution along the genome using non-overlapping 20 kb windows. As previously observed[13], 40% of *C. albicans* strains are affected by either one or multiples aneuploidies, suggesting a predisposition of this species to generate such event (Table 1). However, this result seems to be linked to a bias in the sampling due to isolates exposed to antifungal treatments. In a recent survey, it has been shown that aneuploidies are rather rare, with only 2,2% impacted by such events[15]. In fact, our data suggests that yeast species usually display none or a very low level of aneuploidies. In *K. lactis*, 7 strains (16% of the population) are affected by segmental duplications and no aneuploidies have been found in this population. While an increase number of aneuploidies has been observed in *S. cerevisiae* human-associated clades such as for ale beer, sake or baker isolates, only two dairy strains out of 17 carry segmental duplications in *K. lactis*, suggesting that aneuploidies are not a driving evolutionary process for dairy strains in this species. Overall, aneuploidies are quite rare events in studied population, and only segmental duplications can be observed in WGD species. These results support the hypothesis that aneuploidies arise and are maintained only in very specific and harsh environments and, furthermore, that segmental duplications are preferred in post-WGD species, for which strains have longer chromosomes and therefore can lead to more inappropriate gene dosage due to the number of genes impacted by a complete aneuploidy.

Species	# Individuals	Aneuploidies	Segmental duplications	Strains affected	Prc strains
C. albicans	20	7	2	8	40%
K. lactis	43	0	14	7	16%
L. kluyveri	28	0	1	1	4%
L. thermotolerans	57	0	4	4	7%
S. paradoxus	24	0	1	1	4%
S. uvarum	37	5	1	4	10%

Table 1. Number of aneuploidies and segmental duplication across each species based on manual coverage analysis.

Pangenome construction highlights variability across species

Whole genome sequencing data provided an opportunity to build the pangenome[33] for each studied species using two complementary approaches. First, copy number variants were determined for all reference genes based on a coverage analysis powered by Control-Freec[34]. Secondly, *de novo* assemblies were produced and compared to the corresponding reference genome to determine non-reference materials. Since the genetic divergence is very high in some species, especially in *K.* lactis and *L. thermotolerans,* a given threshold based on the SNPs number was used for each strain (See Method). Putative ORFs were then predicted for each strain and a non-redundant set of supplemental ORFs was determined for each species. Combining these two analyses, it was possible to determine the pangenome of each species, including the accessory genome (*i.e.* supplemental and reference variable ORFs) as well as the core genome, representing the genes found in all strains of the surveyed population (Table 2).

Species	Core	Ref variable	Supplemental	Accessory	PanGenome	Core (%)
C. albicans	6,144	50	202	252	6,396	96.1
K. lactis	4,889	187	566	753	5,642	86.7
L. kluyveri	5,197	124	598	722	5,919	87.8
L. thermotolerans	4,965	127	476	603	5,568	89.2
S. paradoxus	5,928	241	318	559	6,487	91.4
S. uvarum	5,516	474	1,034	1,508	7,024	78.5
Mean	5,439	200	532	733	6,172	88.3
Mean without *S. uvarum*	5,424	146	432	578	6002	90.2

Table 2. Main characteristics of the pangenome composition in each studied species.

The pangenome size is variable across yeast species, ranging from 5,568 to 7,024 genes for *L. thermotolerans* and *S. uvarum*, respectively. This variation is mainly due to the high number of supplemental genes in *S. uvarum* (N = 1,034 ORFs), which is more than two-fold higher compared to the other species (N = 432 ORFs on average, Table 2). Combined with the reference variable genes, 1,508 accessory genes were found in the *S. uvarum* pangenome. By contrast, *C. albicans* genome is way more conserved and only 50 genes of the reference genome are part of the accessory genome (*i.e.* only 0.8 % of the reference genome). However, *C. albicans* is the only species with diploid isolates in our analysis, which led to a better conservation of at least one copy of each gene during evolution and it is worth noting that 648 genes (10.4 %) are in fact impacted by at least one deletion event.

The size of the core genome is also variable across species and ranges from 4,889 to 6,144 genes for *K. lactis* and *C. albicans*, respectively. While the high number of supplemental genes in *S. uvarum* results in a high proportion of accessory genes, its core gene size is in the average (N = 5,516 ORFs). Moreover, the lower core genome size of *K. Lactis* and *L. thermotolerans* is mostly linked to the low reference genome size (~ 5,000 genes) compared to *Saccharomyces*

species or *C. albicans* (~ 6,000 genes). Therefore, while the core genome of *S. uvarum* is slightly bigger compared to the two-former species, a higher number of variable genes can be found (474, 187 and 127 ORFs, respectively). Overall, we found that accessory genes are mostly composed of supplemental ORFs within all species (Table S4), ranging from 57 % to 83 % for *S. paradoxus* and *L. kluyveri*, respectively (74 % on average), clearly pointing out that genome content variation in the reference genome is an underestimation of the gene space.

Genomic characteristics of reference accessory genes

To determine if accessory and core genes present different evolutionary patterns, we looked at the genetic variability between them in each species by comparing the distribution of nucleotide diversity estimates. Within all species, π values, corresponding to the average pairwise divergence, are higher in accessory genes (Mann-Whitney test, p-value = 2×10^{-75}), suggesting that the core genome is more conserved compared to other genes. Interestingly, the Tajima's D values are roughly equal for each species and no significant differences can be observed between these two groups. The absence of variability for this estimate, corresponding to the difference between π and the proportion of polymorphic sites (θ_w), indicates that while a higher nucleotide diversity is found in accessory genes, there is no bias in allele frequency, which could be caused by an overrepresentation of low frequency variants.

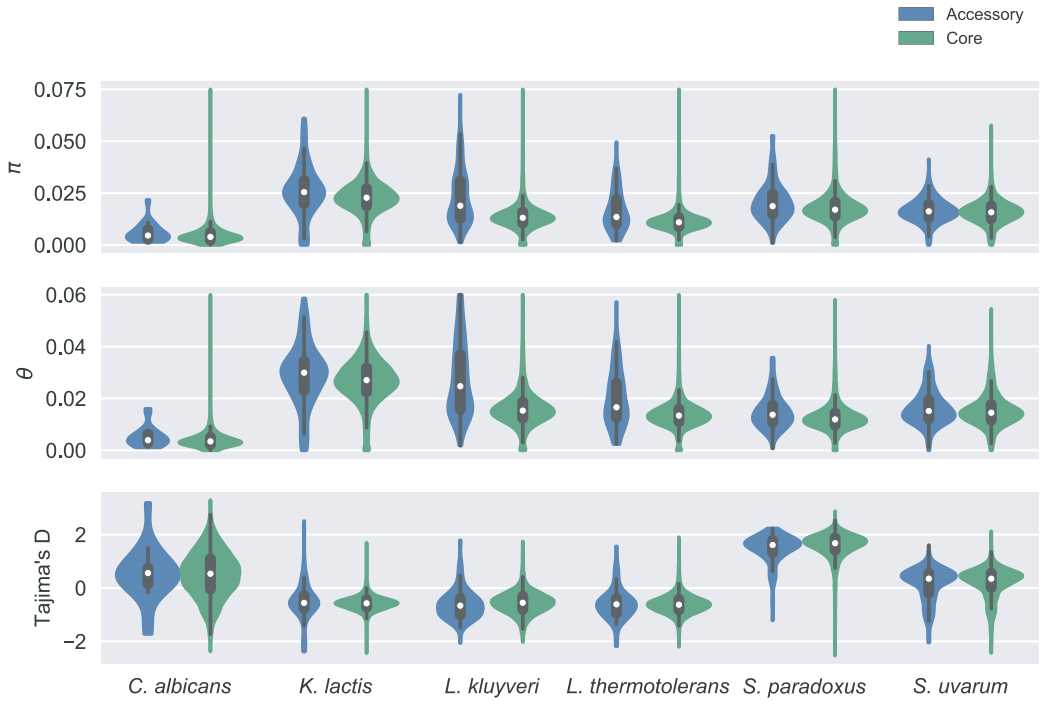


Figure 6. Distribution of nucleic diversity estimators between accessories and core genes in all species. π corresponds to the pairwise genetic diversity. θ_w is relative to the number of polymorphic sites. Tajima's D is the difference between π and θ_w.

Subtelomeric regions are known to show a higher genetic plasticity allowing the fast integration or loss of genes into the genome. To determine if subtelomeric regions are enriched in accessory genes, we looked at the proportion between the core and accessory genes along the genome based on the distance of each gene to their closest telomere (Figure 7). Our result shows that accessory genes are significantly enriched in these regions, as previously observed in *S. cerevisiae*[1]. This trend is especially true from *K. lactis* for which all the genes with a distance lower to 10 kb from the nearest telomere are deleted in at least one sample. Moreover, accessory genes are more subject to duplications (p-value = 1.9×10^{-11}), especially for *K. lactis, L. kluyveri* and *L. thermotolerans* for which a 5-fold increase can be observed (Table S5). Combined with our analysis of the nucleotide diversity, these results confirm the idea that subtelomeric regions are hotspots of structural variation in most of the yeast species, in which genes evolved and expanded much faster compared to genes from the core genome[35].

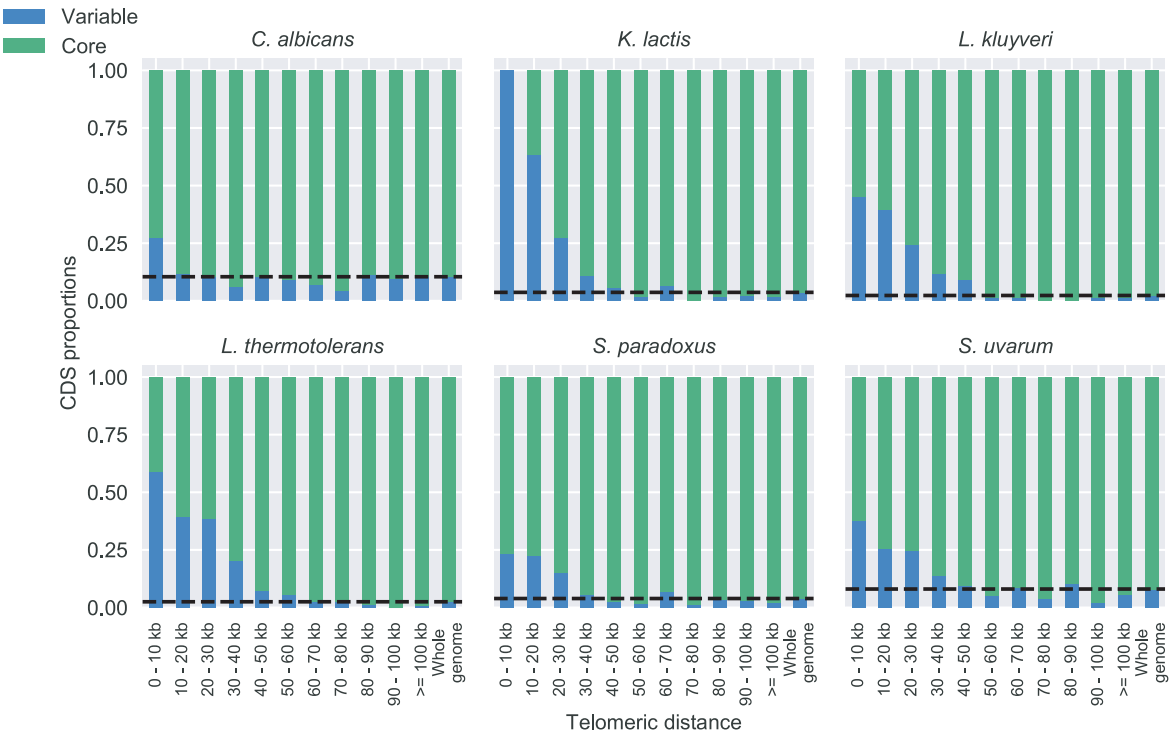


Figure 7. Proportion of variable and core genes based on the distance of each gene to the nearest telomere using windows of 10 kb. The dashed lines correspond to the mean proportion of variable genes at the whole genome scale for each species.

Modeling pangenomes

To determine a model describing the pangenome increase and the core genome decrease for each species, we fitted different non-linear regression model to our data. To that end, we first produced a distribution of the core and pan-genome size for all possible population size available by randomly sampling each sample sets 5,000 times, allowing to reduce the modelling

bias induced by populations stratification. We then tried to fit several power and exponential laws in our dataset (See Method) to estimate the pan and core genome trajectories (Figure 8).

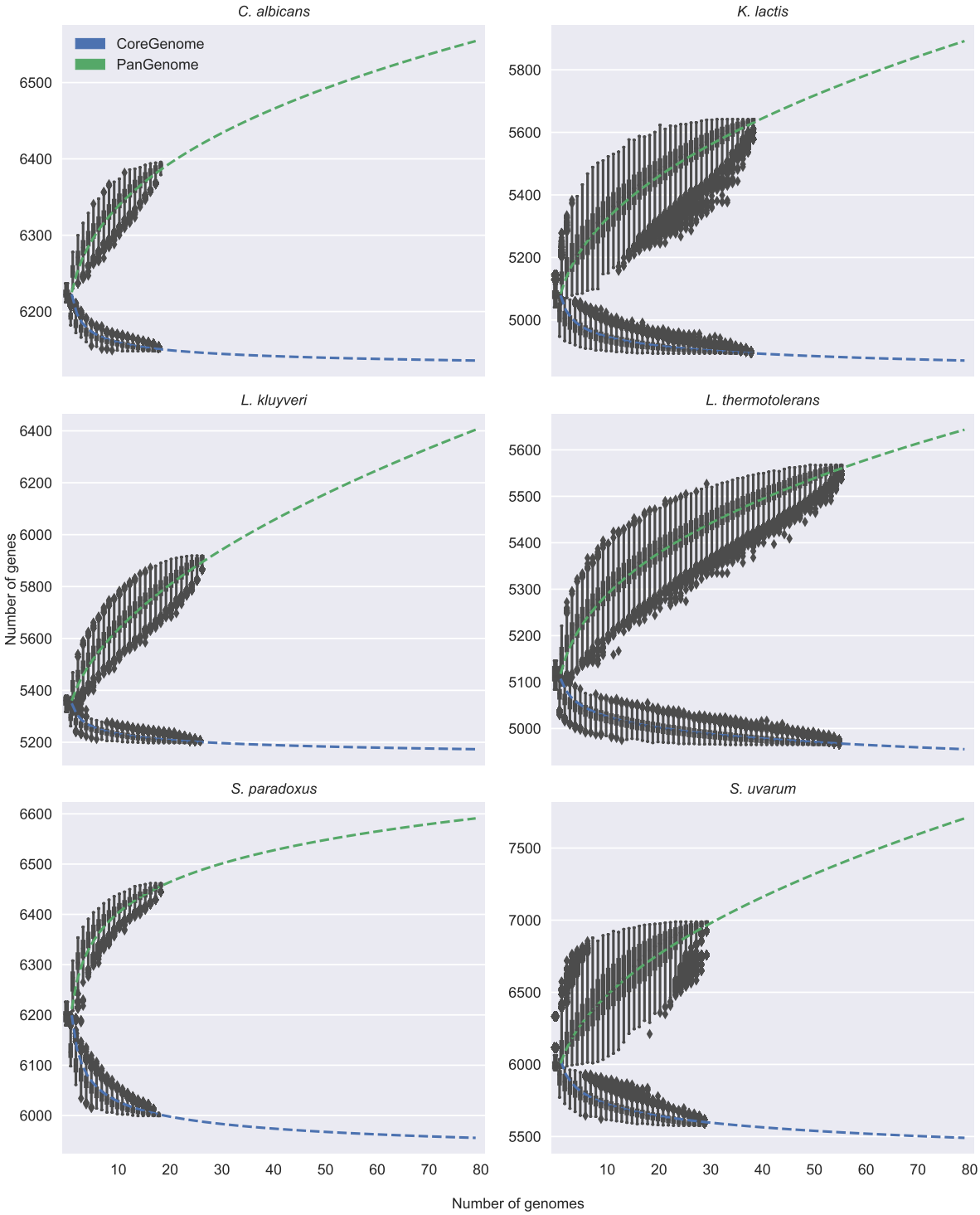


Figure 8. Estimation of the size of the core genomes and pangenomes within in each species. For both the core and pan genome, the power law $y = Ax^{-b} + c$ gave the best fit and was used to estimate the core and pangenome size, represented in the graph by dashed lines.

For all the different dataset, the power law $y = Ax^{-b} + c$ gave the best fit while we were not able to correctly fit the exponential laws with our data. Based on this model, we looked at the

population size required to stabilize the genome, corresponding to the sample size where the addition of a new sample does not change the pangenome size. The number of genomes is highly variable between species, ranging from 95 genomes to almost 10,000 genomes for *S. paradoxus* and *S. uvarum* respectively (Table S6). The high variation observed between the species suggests that the sample size is not sufficient enough to make an accurate prediction of the genome size, especially for *S. uvarum* for which the high number of accessory genes are more likely leading to a wrong prediction. On the other side, the core genome is stabilizing really quickly in our populations, ranging from 14 genomes to 103 genomes for *C. albicans* and *S. uvarum* respectively (N = 41 genomes on average), suggesting that the major part of the genome is conserved within yeast species during evolution.

Content of accessory genes point out introgression from very closely species

Comparison of the protein sequence of each supplemental ORF against several proteomes database allowed the precise identification of donor species and the classification of them in 4 distinct groups. The introgression events correspond to genes resulting from a hybridization event with a closely related species from the same clade and ancestral ORFs correspond to genes more closely related to the reference sequence (See Method). The horizontal gene transfers (HGT) refer to genes originating from divergent species (not the same clade). Lastly, genes for which no similarity was found compared to other species or the reference sequence were classified as unknown (Figure 9).

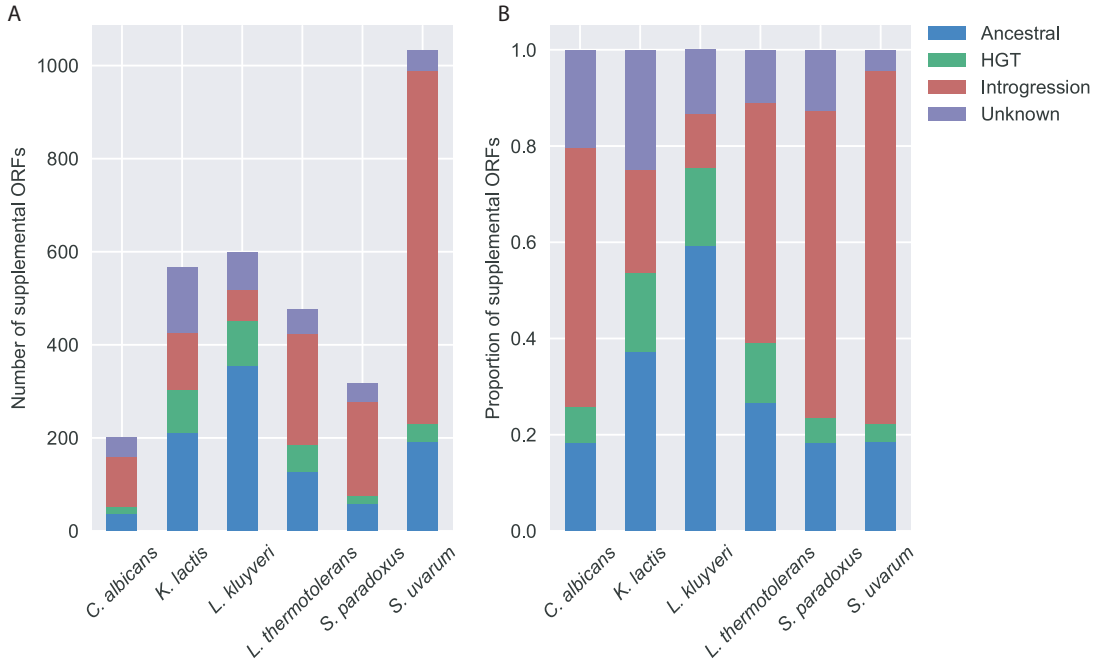


Figure 9. Predicted origins for supplemental ORFs. (A) Number and (B) proportion of each supplementals ORFs categories (Blue : Ancestral; Green : HGT; Red : Introgression; Purple : Unknown) across species.

We found that supplemental ORFs are mainly composed of introgressions with a mean value of 45 % among species, ranging from 11 % to 74 % for *L. kluyveri* and *S. uvarum* respectively (Table S7), suggesting that introgressions play a major role in gene content variation within yeasts. While the majority of supplemental genes are ancestral ORFs within *K. lactis* and *L. kluyveri*, introgressions represent the major categories within other species. For *K. lactis*, almost half of the ancestral ORFs (N = 126 ORFs) have been found in the Asian subpopulation. This cluster is the most divergent compared to the reference sequence with 6% of nucleotide divergence, which could result in the presence of highly diverged genes in this population. On the other side, most of the ancestral genes found in *L. kluyveri* (N = 140 ORFs, Table S8) are located on the left side of the chromosome C. This region is known to correspond to a 1 Mb introgressed region for which the donor species is still unknown[16], displaying higher genetic variability compared to the rest of the genome.

We investigated the supplemental ORFs distribution among each population by identifying whether each ORF is related to a single subpopulation or shared between them (Figure 10). Most of the genes can be found in one single cluster, suggesting that these genes result from independent events during the evolution of each subpopulation. *S. paradoxus* is the only species for which the biggest group has been found to be shared between subpopulation (SpA and SpB) and can be linked to the close proximity of the isolation locations for the samples in this species compared to other collections.

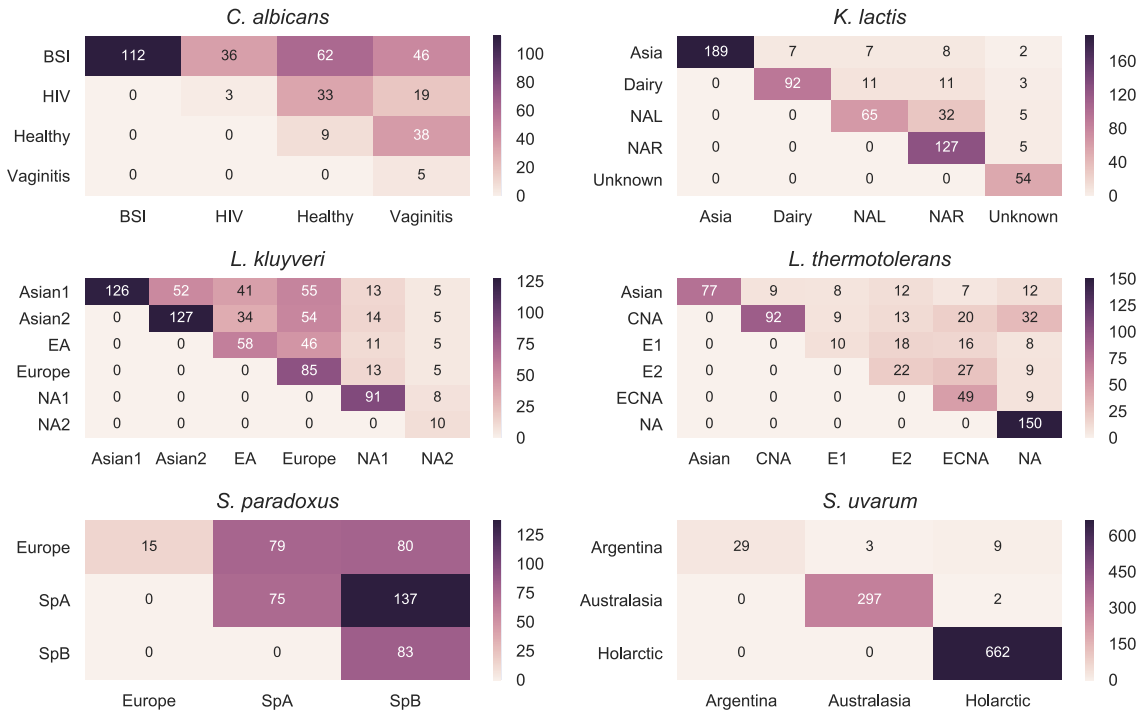


Figure 10. Supplemental ORFs distribution between cluster within each species. Colors are relative to the number of ORFs found within each group, independently for each species.

We then further examined the origin of introgressions and horizontal gene transfers by analysing the donor species (Figure 11, Table S9). For all species, several introgressions originating from closely related species have been found, suggesting a high genes flow between species from the same clade. Interestingly, introgressions from a wide range of closed species can also be found in wild species not impacted by domesticated processes. For example, *L. thermotolerans* display introgressions coming from 9 other species within the *Lachancea* clade, indicating that introgression events are not limited to domestication process and that gene flow between species is not uncommon in nature.

Within *S. paradoxus*, more than 150 introgressions from its sister species *S. cerevisiae* have been found (61 %), which is congruent with analysis of *S. cerevisiae* pangenome for which most of the introgressions originated from *S. paradoxus*[1]. This result reinforces the idea that an important gene flow exists between these two species. For *S. uvarum*, most of the introgressions comes from *S. eubayanus* (N = 647 ORFs) and are mostly located in the Holarctic cluster, representing 67 % of the supplementary ORFs within the species. This result is in accordance with observations found in the previous study using the same dataset and is linked to several domesticated processes associated to this subpopulation[19]. While introgressions are found in natural population, *S. uvarum* displays by far the highest number of events, accounting for almost half of total number of introgressed genes within all species. This result confirms that domestication process, by producing environment where multiple species coexist, favours the apparition of hybrids and introgression events.

Interestingly, horizontal gene transfers can also be found in our results, corresponding to around 10 % of the supplemental genes across species, and are mainly found in *K. lactis* (N = 93 ORFs) and *L. thermotolerans* (N = 97 ORFs). Many HGT events in *K. lactis* are originated from *Lachancea* species, suggesting that inter-species exchanges are not limited to species from the same clade and that gene flow exist between the *Lachancea* and *Kluyveromyces* species.

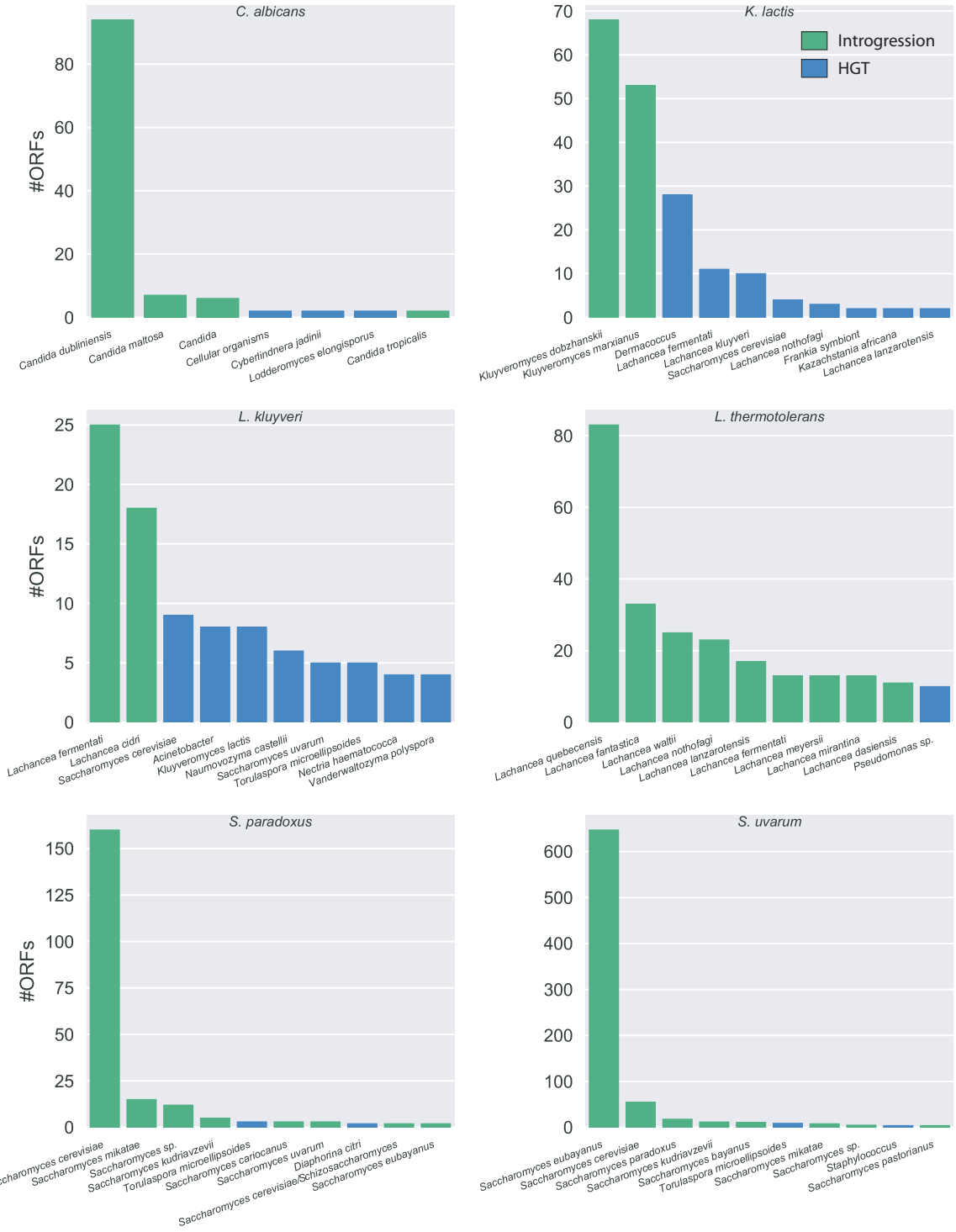


Figure 11. Distribution of the number of introgressed or HGT genes by donor species across species for both introgression (green) and HGT (blue) events.

Functional analysis of accessory genes

Finally, we analysed both CNVs and supplementary genes to identify genomic signatures of phenotypic adaptations. Indeed, while CNVs are known to be a quick response to specific environments by modifying gene dosage, supplementary genes allow the incorporation of new functions favouring the adaptation to new conditions. In *S. cerevisiae*, such associations have been found, including CNVs of the Li^+ and Na^+ pump-encoding *ENA* genes linked to lithium resistance[4], or an introgression of *PDR5* conferring resistance to translation inhibitor cycloheximide[36].

Examination of the gene content across populations allowed us to identify a cluster of genes found in dairy and lost in wild isolates of the *K. lactis* species (Figure 12, Table S10). This cluster encompasses genes coding for maltose and lactose permeases, as well as for drug resistance such as arsenic response metabolism. More precisely, both the *LAC12* and *LAC4* genes coding for a lactose permease and a beta-galactosidase are present in the dairy but absent in the wild *K. lactis* isolates. The same observation holds true for the two *ARR1* and *ARR2* genes involved in arsenic resistance. Concerning the *MAL* genes (*MAL11*, *MAL21* and *MAL22*) coding for maltose permeases, a higher CN can be found in dairy compared to wild isolates. Interestingly, an analysis of *S. cerevisiae* cheese isolates also highlighted amplification of specific genes compared to other subpopulations, including the *MAL11, MAL12* and *MAL13* genes[37]. This result confirms that the plasticity of gene content plays a major role in the adaptation to domestication processes.

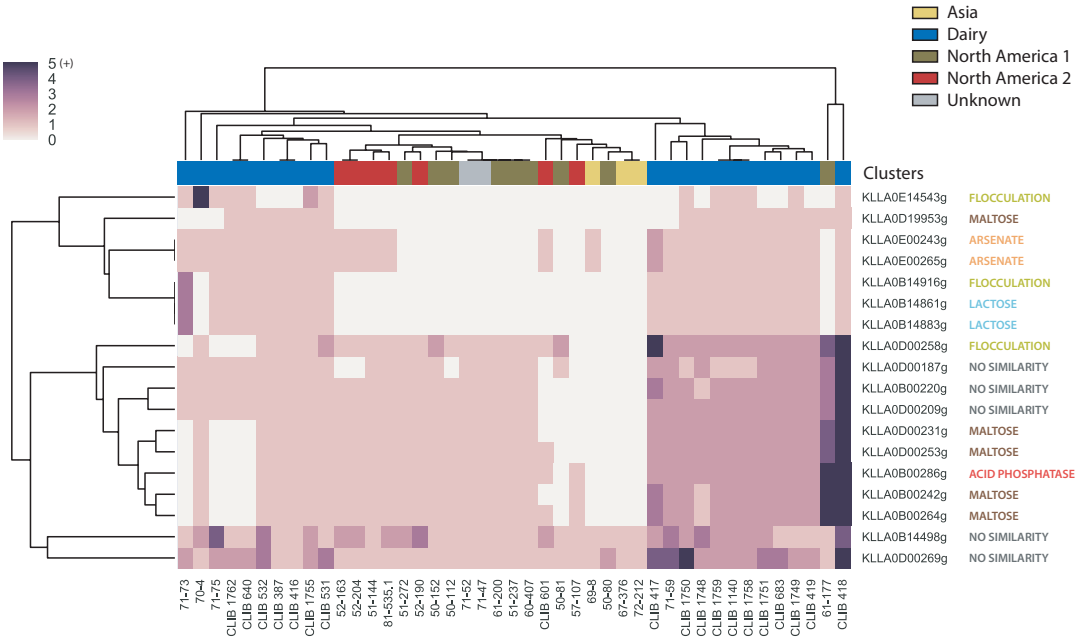


Figure 12. CNVs heatmap of genes linked to dairy samples in *K. lactis*. Clusters in which strain can be found are represented by the color code on top of the heatmap. For graphical purpose, genes which have a number of copy higher or equal to 5 were put in the same category (5+).

To identify potential functional enrichment for additional genes, GO-Terms analyses were performed using *S. cerevisiae* orthologs. Our result shows several enrichments for genes family, including *FLO* genes involved in flocculation or the *HXT* superfamily coding for sugar transporters. However, GO analysis of supplemental genes failed to yield overrepresented genes GO categories that might be specifically linked to adaptation. It is more likely that the high hitchhiking resulting from the transfer of large regions during introgression and HGT events considerably diminished the power of such analysis.

Conclusion

By combining population genomic analysis of six yeast species of the Saccharomycotina phylum, we have acquired a global view of the genetic variability within and between species covering a broad evolutionary distance. Comparative analysis of several types of variant showed that evolutionary process took multiple forms for which the impact can be specific to each yeast species. To get a global picture of the genetic diversity within each population, we then focused on gene content variability across species by constructing to pangenome, *i.e* the complete set of ORFs within each population. Analysis of this genetic variability component, still poorly studied in yeast, highlights that the pangenome and core genome size is variable across species, indicating that genome plasticity in yeast genome is high and that yeast evolution is also mainly shaped by genes content variation. Furthermore, introgression events can be found in all species, including natural populations, suggesting that interspecific hybrids leading to gene flow between closed species is not uncommon even in natural environment. It is therefore more likely that this process is an important factor driving yeast evolution, however determining functional impact of introgression remains difficult, partially due to hitchhiking genes obtained when introgression and HGT events took place. On the other side, analysis of CNVs is more straightforward and our results show a striking example of such genomic adaptation into the dairy strains within *K. lactis*, similar to what was observed in *S. cerevisiae* cheese subpopulation. This result indicates that CNV mechanism is a common adaptive process within yeast, relying on the genome plasticity of these organisms. Overall, these results support the idea that gene content variation is a major component of yeast genome evolution and play a significant role in yeast adaptation. This observation is especially true for highly variable species such as *K. lactis* and *L. thermotolerans* for which the reference genomes significantly underestimate the gene space of the species.

References

1. Peter, J. *et al.* Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556,** 339–344 (2018).
2. Skelly, D. A. *et al.* Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* **23,** 1496–504 (2013).
3. Bergström, A. *et al.* A High-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31,** 872–888 (2014).
4. Strope, P. K. *et al.* The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* gr.185538.114- (2015). doi:10.1101/gr.185538.114
5. Almeida, P. *et al.* A population genomics insight into the mediterranean origins of wine yeast domestication. *Mol. Ecol.* (2015). doi:10.1111/mec.13341
6. Zhu, Y. O., Sherlock, G. & Petrov, D. A. Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3 (Bethesda)*. (2016). doi:10.1534/g3.116.029397
7. Gonçalves, M. *et al.* Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Curr. Biol.* **26,** 2750–2761 (2016).
8. Gallone, B. *et al.* Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* **166,** 1397–1410.e16 (2016).
9. Riley, R. *et al.* Comparative genomics of biotechnologically important yeasts. *Proc. Natl. Acad. Sci.* **113,** 9882–9887 (2016).
10. Dujon, B. A. & Louis, E. J. Genome diversity and evolution in the budding yeasts (Saccharomycotina). *Genetics* **206,** (2017).
11. Leducq, J.-B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat. Microbiol.* **1,** 15003 (2016).
12. Carreté, L. *et al.* Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. *Curr. Biol.* **28,** 15–27.e7 (2018).
13. Hirakawa, M. P. *et al.* Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Res.* gr.174623.114- (2015). doi:10.1101/gr.174623.114
14. Ford, C. B. *et al.* The evolution of drug resistance in clinical isolates of *Candida albicans*. *Elife* **4,** e00662 (2015).
15. Ropars, J. *et al.* Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. *Nat. Commun.* **9,** 2253 (2018).
16. Friedrich, A., Jung, P., Reisser, C., Fischer, G. & Schacherer, J. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol. Biol. Evol.* **32,** 184–92 (2015).
17. Ortiz-Merino, R. A. *et al.* Ploidy variation in *Kluyveromyces marxianus* separates dairy and non-dairy isolates. *Front. Genet.* **9,** 94 (2018).
18. Peter, J. & Schacherer, J. Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast* **33,** 73–81 (2016).
19. Almeida, P. *et al.* A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. *Nat. Commun.* **5,** 4044 (2014).
20. Nonklang, S. *et al.* High-temperature ethanol fermentation and transformation with linear DNA in the thermotolerant yeast *Kluyveromyces marxianus* DMKU3-1042. *Appl. Environ. Microbiol.* **74,** 7514–7521 (2008).
21. Spohner, S. C., Schaum, V., Quitmann, H. & Czermak, P. *Kluyveromyces lactis*: An emerging tool in biotechnology. *J. Biotechnol.* **222,** 104–116 (2016).
22. Naumov, G. I., Naumova, E. S., Glushakova, A. M., Kachalkin, A. V. & Chernov, I. Y. Finding of dairy yeasts *Kluyveromyces lactis var. lactis* in natural habitats. *Microbiology* **83,** 782–786 (2014).
23. Jolly, N. P., Varela, C. & Pretorius, I. S. Not your ordinary yeast: non- *Saccharomyces* yeasts in wine production uncovered. *FEMS Yeast Res.* **14,** 215–237 (2014).
24. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430,** 35–44 (2004).
25. Souciet, J.-L. *et al.* Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res.* **19,** 1696–709 (2009).

26. Leducq, J.-B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *bioRxiv* (Cold Spring Harbor Labs Journals, 2015). doi:10.1101/027383
27. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* **6,** 80–92
28. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–81 (2009).
29. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–59 (2000).
30. Freel, K. C., Friedrich, A., Hou, J. & Schacherer, J. Population genomic analysis reveals highly conserved mitochondrial genomes in the yeast species *Lachancea thermotolerans*. *Genome Biol. Evol.* **6,** 2586–2594 (2014).
31. Hranilovic, A., Bely, M., Masneuf-Pomarede, I., Jiranek, V. & Albertin, W. The evolution of *Lachancea thermotolerans* is driven by geographical determination, anthropisation and flux between different ecosystems. *PLoS One* **12,** e0184652 (2017).
32. Leducq, J.-B. *et al.* Local climatic adaptation in a widespread microorganism. *Proc. Biol. Sci.* **281,** 20132472 (2014).
33. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome";. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 13950–5 (2005).
34. Boeva, V. *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27,** 268–269 (2011).
35. Brown, C. A., Murray, A. W. & Verstrepen, K. J. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr. Biol.* **20,** 895–903 (2010).
36. Pérez-Ortín, J. E., Querol, A., Puig, S. & Barrio, E. Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome Res.* **12,** 1533–9 (2002).
37. Legras, J.-L. *et al.* Adaptation of *S . cerevisiae* to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. *Mol. Biol. Evol.* (2018). doi:10.1093/molbev/msy066

Supplementary materials

Supplementary tables

Supplementary tables are available at https://bit.ly/2L5wc0D and https://bit.ly/2MXRajr

Supplementary figures

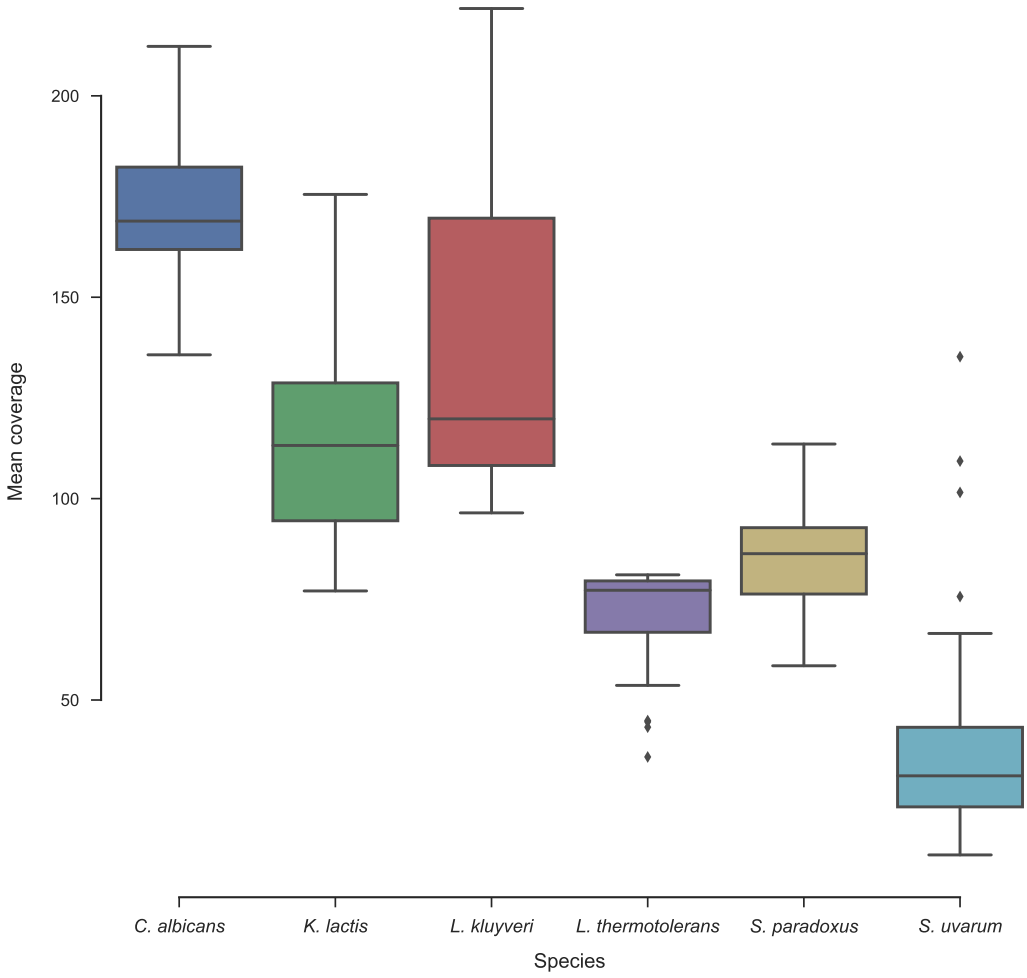


Figure S1. Mean coverage distribution within all species.

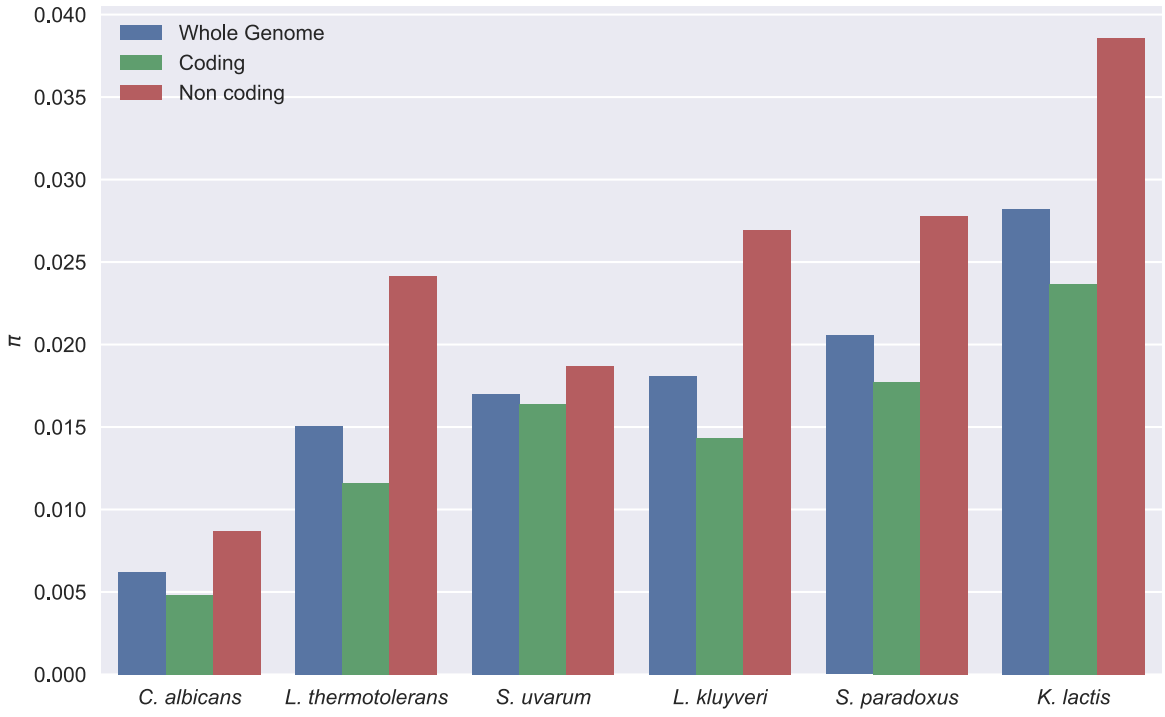


Figure S2. Pairwise nucleic diversity within the whole genome (blue), coding regions (green) for non-coding regions (red) across species.

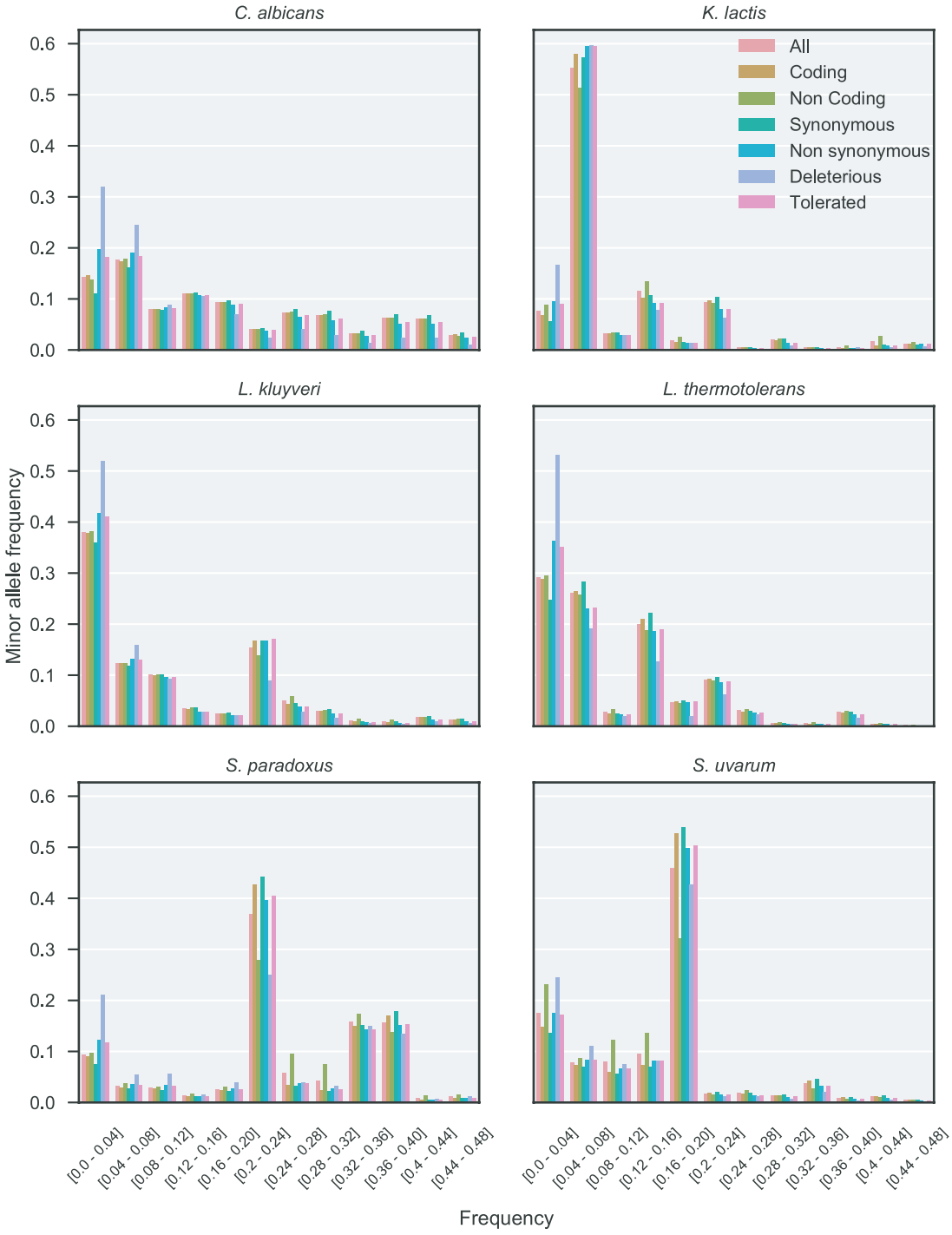


Figure S3. Frequency spectrum of SNPs across species. Minor allele frequency (MAF) of polymorphisms was determined from SNPs with difference functional annotations.

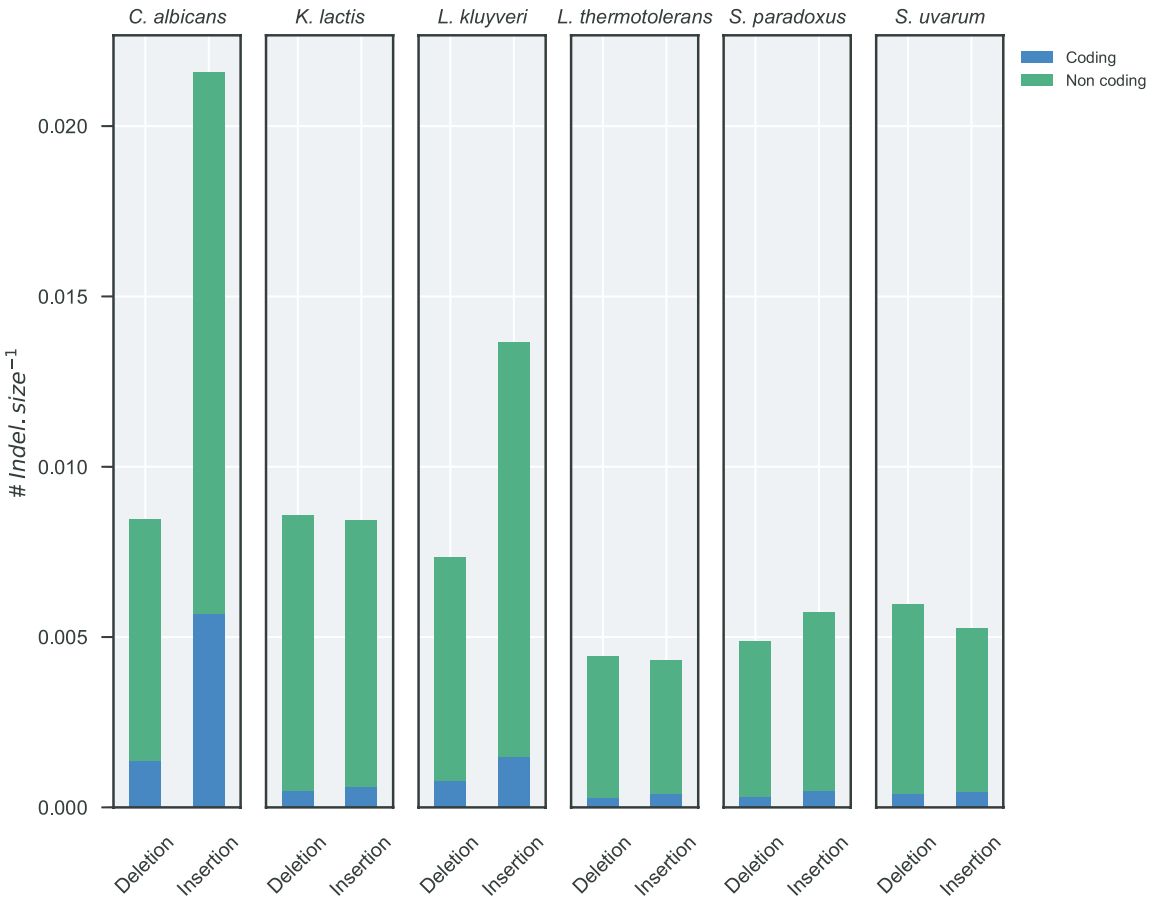


Figure S4. Distribution of short deletions and insertions (InDels) within coding and non-coding sequences across species.

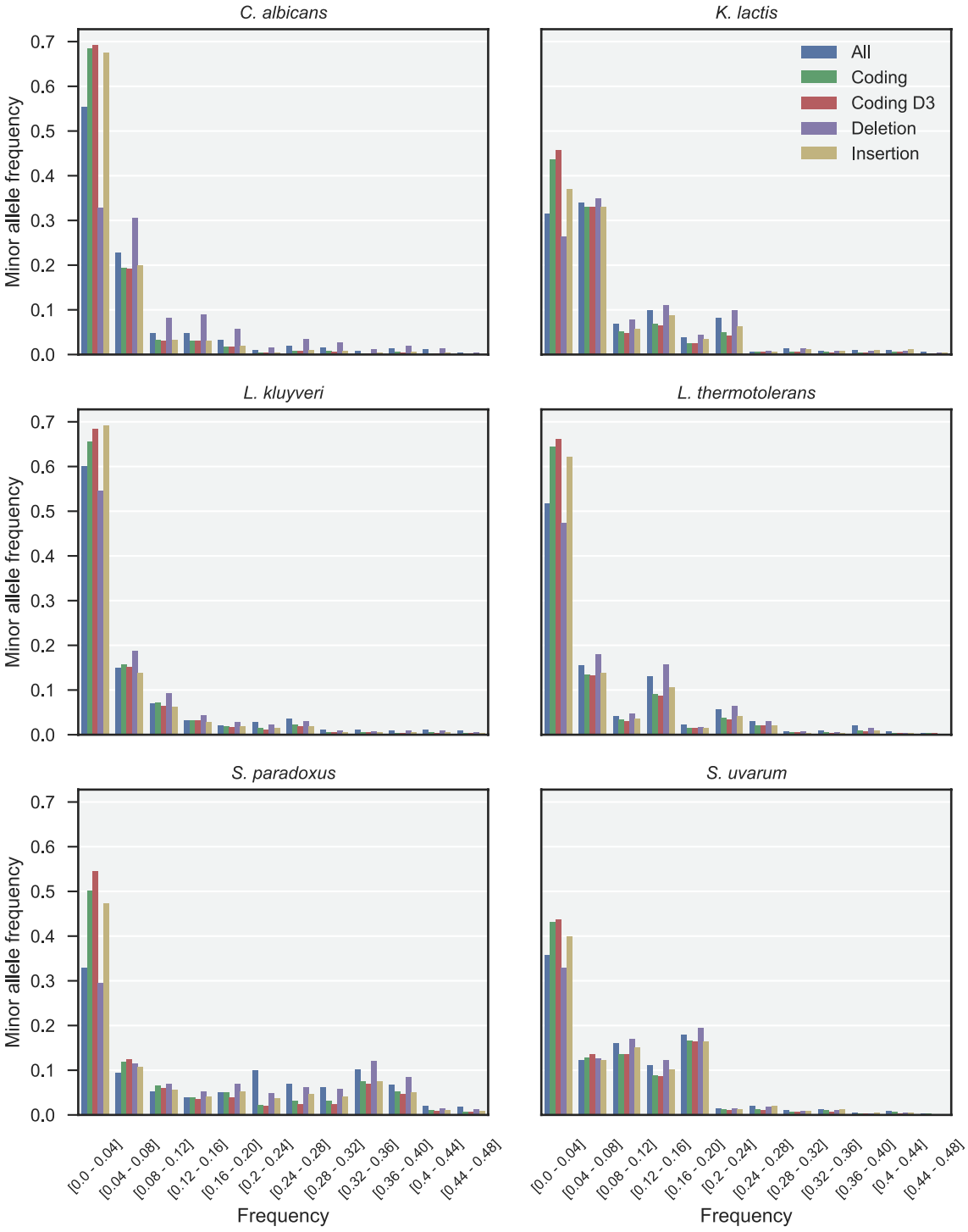


Figure S5. Frequency spectrum of short insertions and deletions (InDels) across each species grouped by kind (insertion, deletions) or by functional impact (coding and coding with a size multiple of 3 (D3))

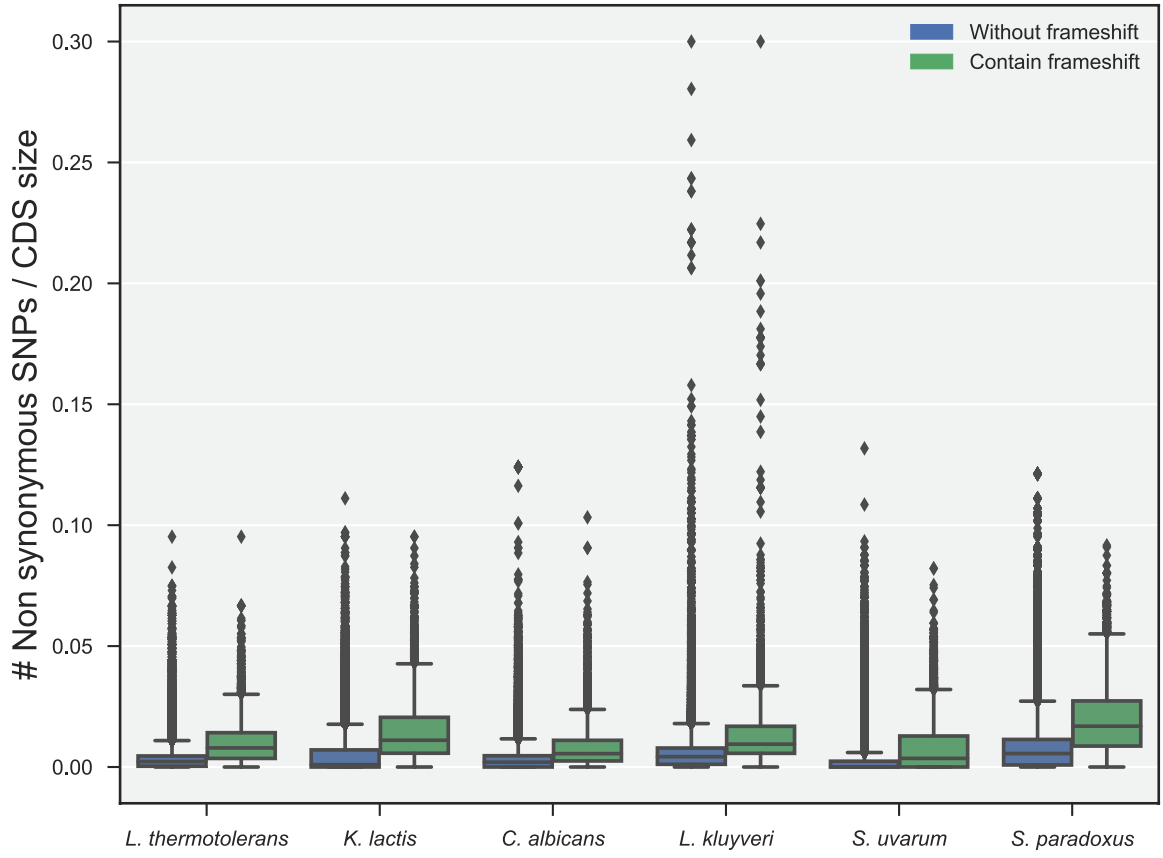
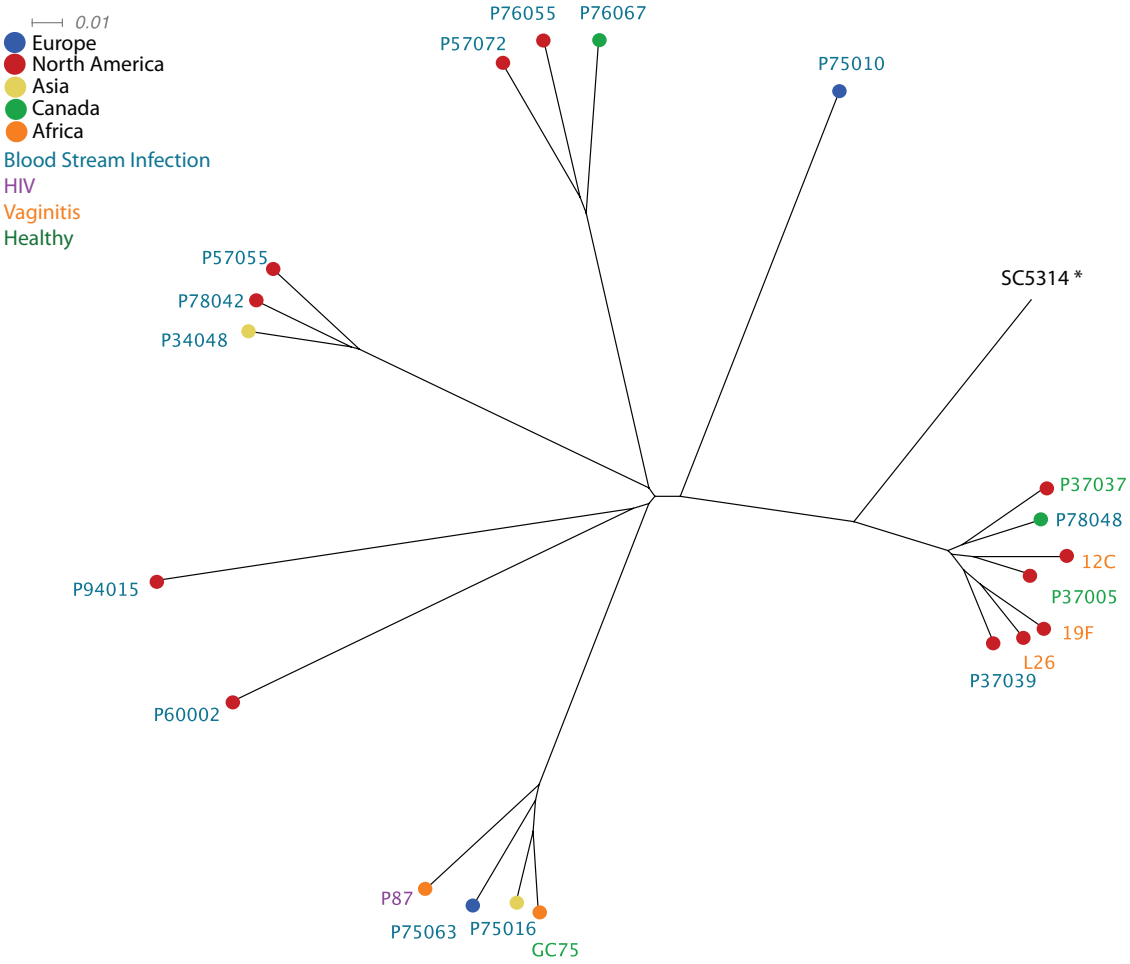
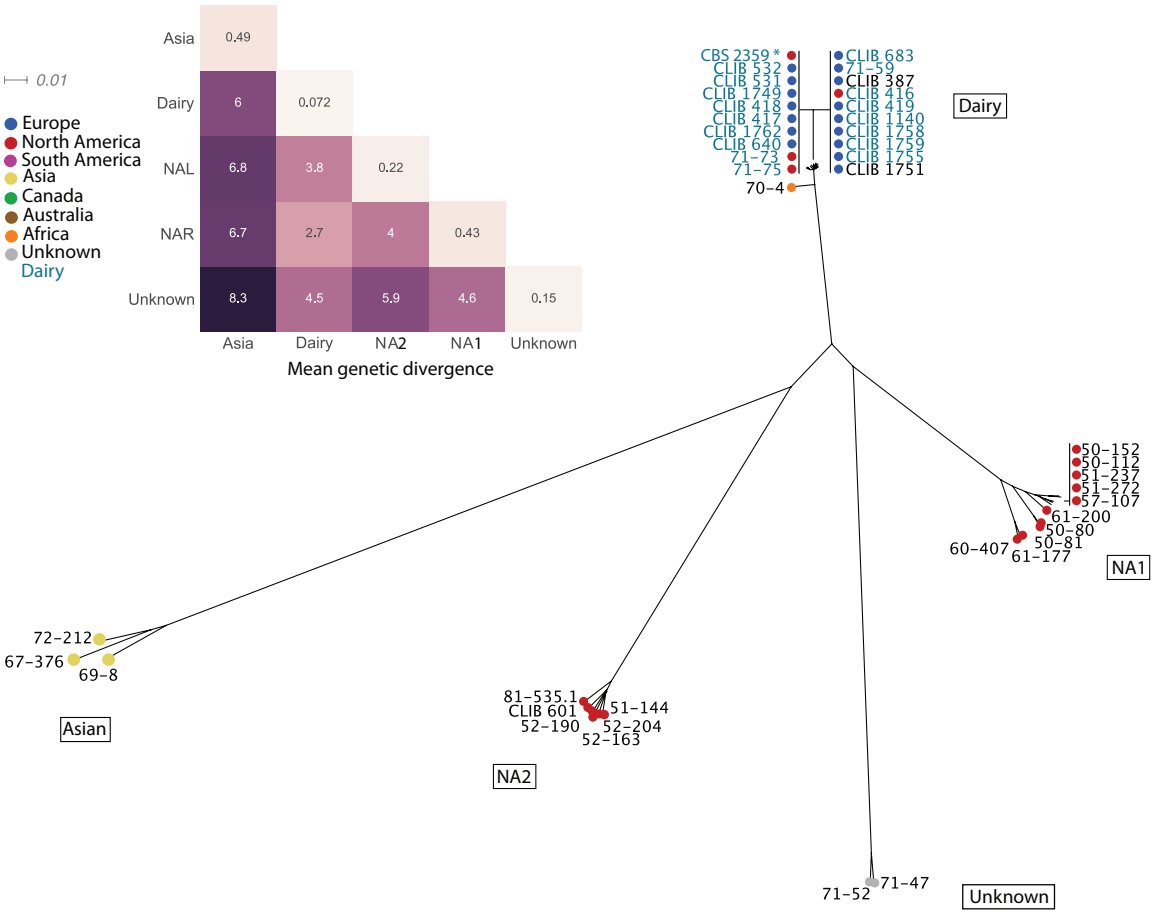


Figure S6. Distribution of the ratio of non-synonymous (dN) to synonymous (dS) substitution rate (dN/dS) for each gene grouped by the presence or not of frameshift InDels across populations.

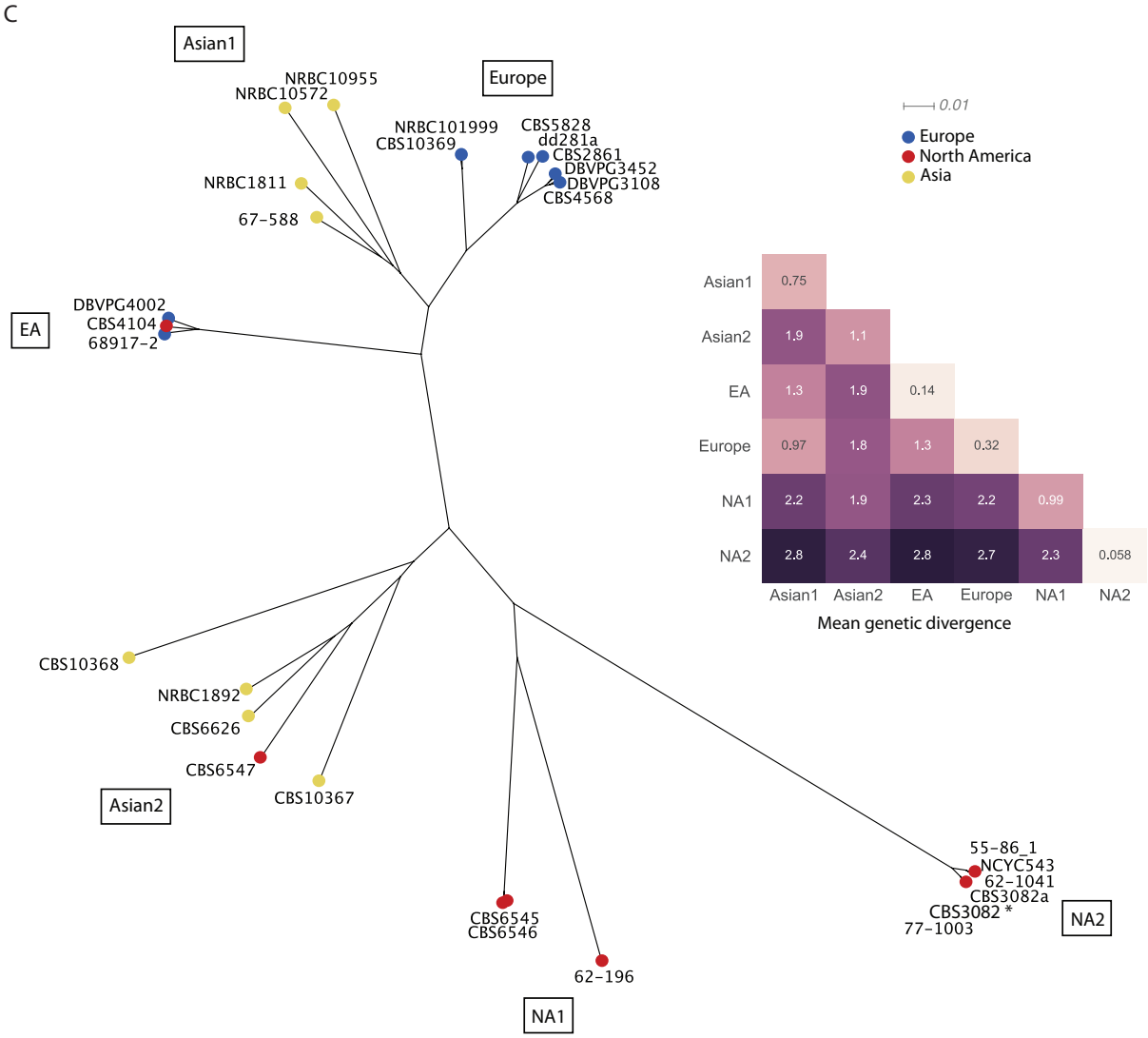
A


0.01
Europe
North America
Asia
Canada
Africa
Blood Stream Infection
HIV
Vaginitis
Healthy
P76055
P76067
P57072
P75010
P57055
P78042
P34048
SC5314 *
P37037
P78048
12C
P37005
19F
L26
P37039
P94015
P60002
P87
P75063
P75016
GC75

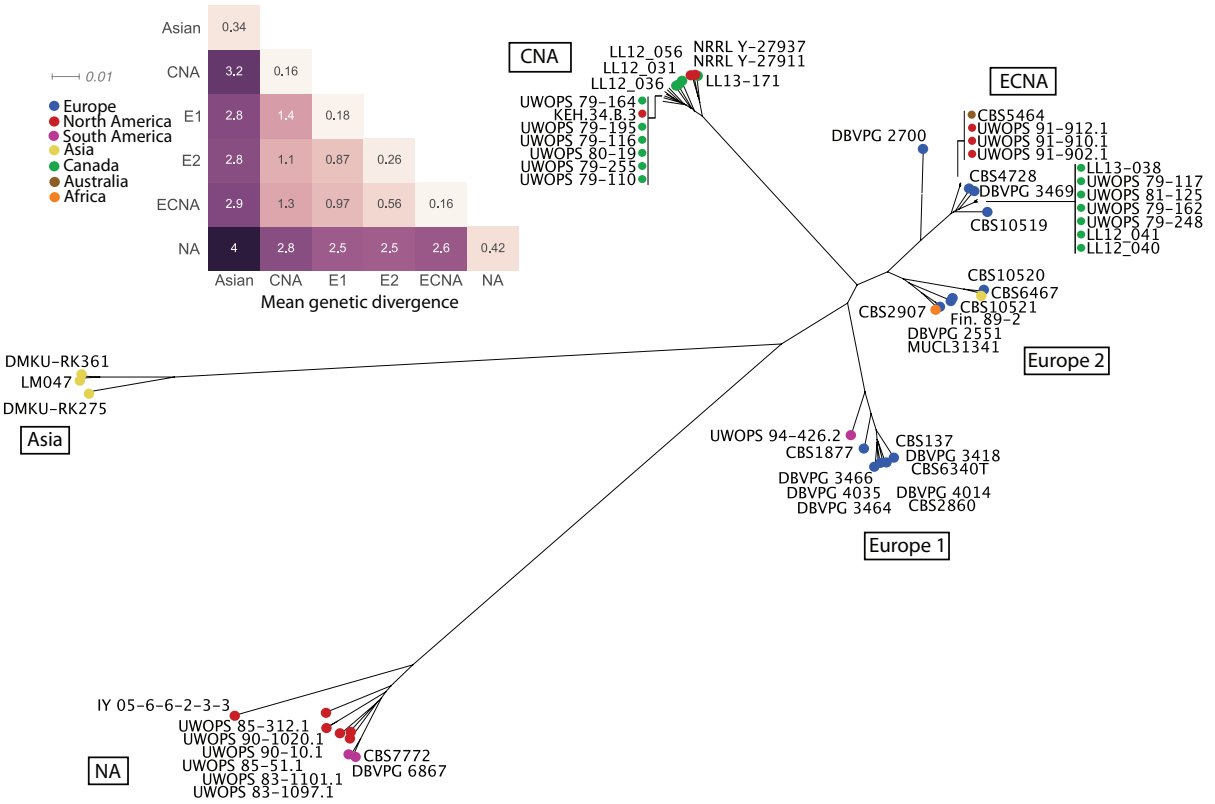
B


0.01
Europe
North America
South America
Asia
Canada
Australia
Africa
Unknown
Dairy
Asia
Dairy
NAL
NAR
Unknown
0.49
6
0.072
6.8
3.8
0.22
6.7
2.7
4
0.43
8.3
4.5
5.9
4.6
0.15
Asia
Dairy
NA2
NA1
Unknown
Mean genetic divergence
CBS 2359*
CLIB 532
CLIB 531
CLIB 1749
CLIB 418
CLIB 417
CLIB 1762
CLIB 640
71-73
71-75
70-4
CLIB 683
71-59
CLIB 387
CLIB 416
CLIB 419
CLIB 1140
CLIB 1758
CLIB 1759
CLIB 1755
CLIB 1751
Dairy
50-152
50-112
51-237
51-272
57-107
61-200
50-80
50-81
61-177
60-407
NA1
72-212
67-376
69-8
Asian
81-535.1
CLIB 601
52-190
51-144
52-204
52-163
NA2
71-52
71-47
Unknown

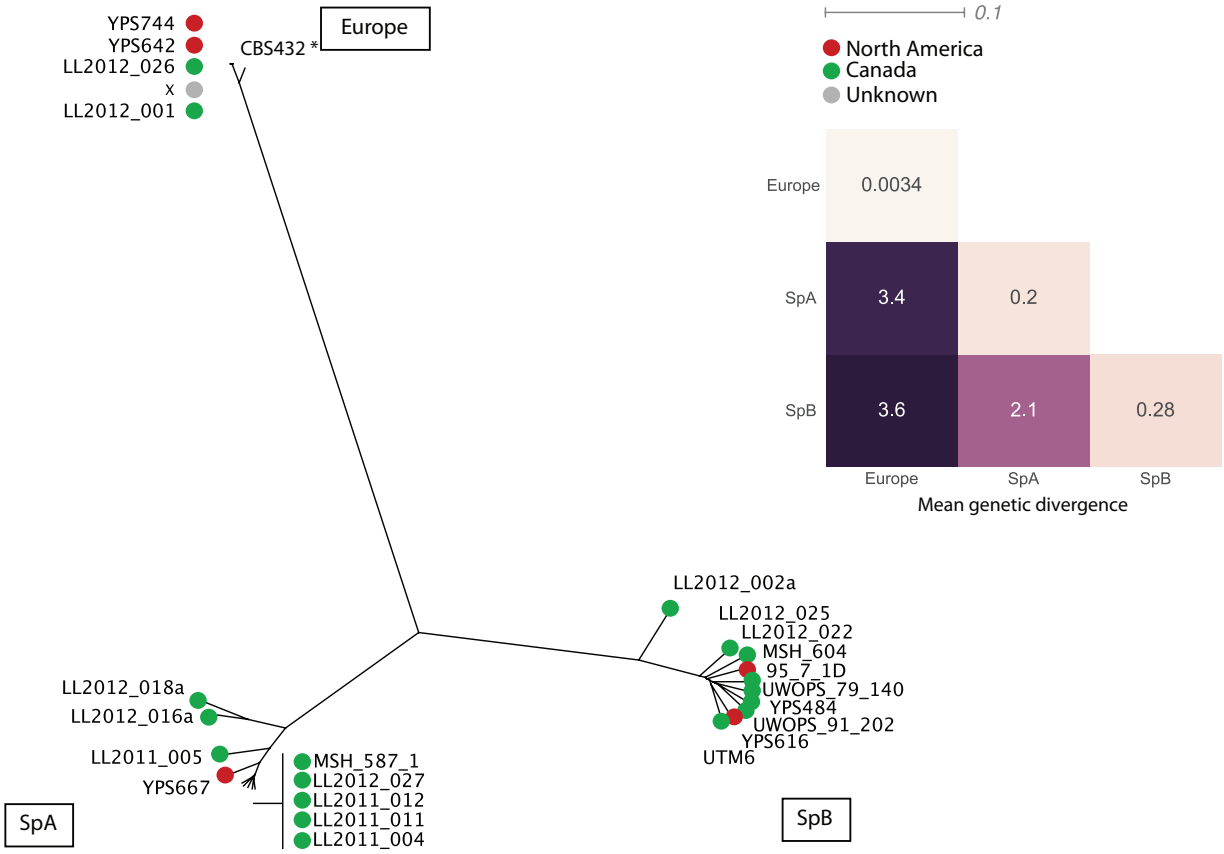
C


Asian1
NRBC10955
NRBC10572
Europe
NRBC101999
CBS10369
CBS5828
dd281a
CBS2861
DBVPG3452
DBVPG3108
CBS4568
NRBC1811
67-588
EA
DBVPG4002
CBS4104
68917-2
CBS10368
NRBC1892
CBS6626
CBS6547
CBS10367
Asian2
CBS6545
CBS6546
62-196
NA1
55-86_1
NCYC543
62-1041
CBS3082a
CBS3082 *
77-1003
NA2
0.01
Europe
North America
Asia
Asian1 0.75
Asian2 1.9 1.1
EA 1.3 1.9 0.14
Europe 0.97 1.8 1.3 0.32
NA1 2.2 1.9 2.3 2.2 0.99
NA2 2.8 2.4 2.8 2.7 2.3 0.058
Asian1 Asian2 EA Europe NA1 NA2
Mean genetic divergence

D


0.01
Europe
North America
South America
Asia
Canada
Australia
Africa
Asian 0.34
CNA 3.2 0.16
E1 2.8 1.4 0.18
E2 2.8 1.1 0.87 0.26
ECNA 2.9 1.3 0.97 0.56 0.16
NA 4 2.8 2.5 2.5 2.6 0.42
Asian CNA E1 E2 ECNA NA
Mean genetic divergence
CNA
LL12_056
NRRL Y-27937
LL12_031
NRRL Y-27911
LL12_036
LL13-171
UWOPS 79-164
KEH 34.B.3
UWOPS 79-195
UWOPS 79-116
UWOPS 80-19
UWOPS 79-255
UWOPS 79-110
ECNA
DBVPG 2700
CBS5464
UWOPS 91-912.1
UWOPS 91-910.1
UWOPS 91-902.1
CBS4728
DBVPG 3469
CBS10519
LL13-038
UWOPS 79-117
UWOPS 81-125
UWOPS 79-162
UWOPS 79-248
LL12_041
LL12_040
CBS10520
CBS6467
CBS10521
Fin. 89-2
CBS2907
DBVPG 2551
MUCL31341
Europe 2
DMKU-RK361
LM047
DMKU-RK275
Asia
UWOPS 94-426.2
CBS1877
CBS137
DBVPG 3418
CBS6340T
DBVPG 3466
DBVPG 4035
DBVPG 4014
DBVPG 3464
CBS2860
Europe 1
IY 05-6-6-2-3-3
UWOPS 85-312.1
UWOPS 90-1020.1
UWOPS 90-10.1
UWOPS 85-51.1
UWOPS 83-1101.1
UWOPS 83-1097.1
CBS7772
DBVPG 6867
NA

E


YPS744
YPS642
LL2012_026
X
LL2012_001
CBS432 *
Europe
0.1
North America
Canada
Unknown
Europe 0.0034
SpA 3.4 0.2
SpB 3.6 2.1 0.28
Europe SpA SpB
Mean genetic divergence
LL2012_002a
LL2012_025
LL2012_022
MSH_604
95_7_1D
UWOPS_79_140
YPS484
UWOPS_91_202
YPS616
UTM6
LL2012_018a
LL2012_016a
LL2011_005
YPS667
MSH_587_1
LL2012_027
LL2011_012
LL2011_011
LL2011_004
SpA
SpB

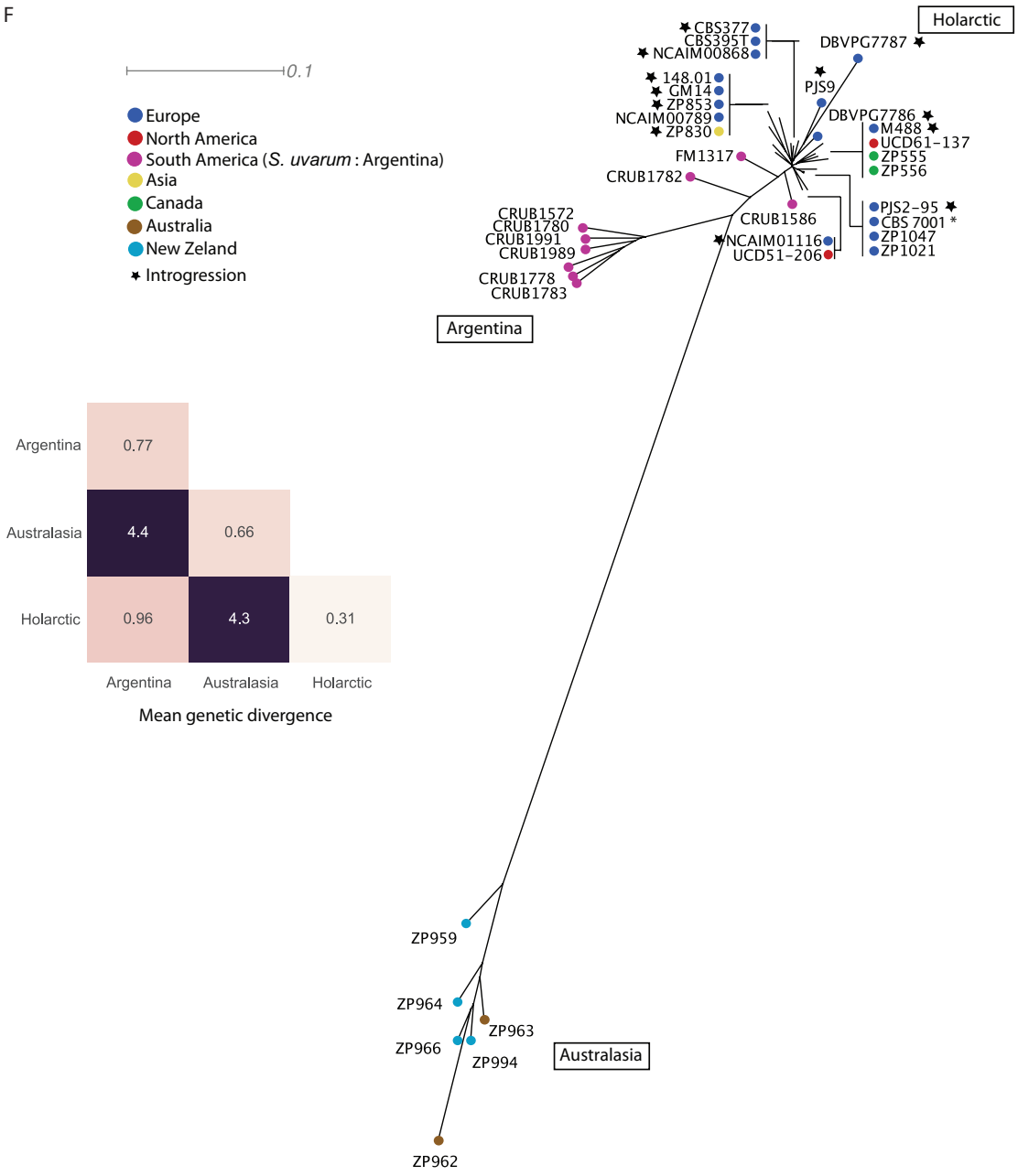


Figure S7. Phylogenetic relationships for each species based on a bio Neighbor-joining method (not at the same scale between species) : (A) *C. albicans* – (B) *K. lactis* – (C) *L. kluyveri* – (D) *L. thermotolerans* – (E) *S. paradoxus* and (F) *S. uvarum*. For each species the mean genetic divergence between clusters is represented by a heatmap.

CHAPTER 2

High-quality *de novo* genome assembly of *Dekkera bruxellensis* using Nanopore MinION Sequencing

Résumé

L'exploration de la diversité génétique au sein d'espèces non-modèles se révèle cruciale afin de mieux comprendre l'évolution des génomes et la relation qui lie génotype et phénotype. Dans ce cadre, l'espèce *D. bruxellensis* présente des propriétés intéressantes. En effet, des isolats de cette espèce peuvent être retrouvés dans plusieurs processus de fermentation (vin et bière) mais aussi en tant que contaminant majeur dans la production de certains vins de par l'odeur forte des composés phénols dégagés par les souches. Par ailleurs, des études préalables de plusieurs génomes ont révélé d'importantes variations dans le caryotype de certaines souches ainsi que des évènements d'hybridation, motivant l'exploration de la variabilité intraspécifique au sein de cette espèce. L'établissement d'une étude de génomique des populations est cependant limité de par l'absence d'une séquence de référence et d'annotations de bonne qualité. L'émergence récente des techniques de séquençage dites de 3ème génération offre aujourd'hui la possibilité d'obtenir des lectures de grandes tailles facilitant grandement la production d'assemblages. Pour *D. bruxellensis*, des assemblages ont déjà été produits pour 4 isolats de cette espèce avec des technologies de séquençage variées, mais ces derniers sont incomplets et présentent une faible contiguïté ainsi qu'une annotation très partielle de leur génome. En l'état, ils ne permettent pas de poser les fondations pour une étude de génomique des populations.

Dans ce contexte, nous nous sommes intéressés à la production d'un assemblage complet du génome d'un isolat de *D. bruxellensis*. Le génome de la souche de vin UMY321 a ainsi été séquencé à travers une combinaison des technologies Oxford Nanopore et Illumina. Ces données ont permis la construction d'un assemblage contenant 8 chromosomes sur environ 13 Mb et présentant ainsi une meilleure contiguïté par rapport aux assemblages précédemment produits. La qualité de notre assemblage dans le cadre d'études de génomiques des populations a par la suite été testée à travers l'alignement de données Illumina provenant de plusieurs isolats de *D. bruxellensis*. De manière générale, un plus faible pourcentage de lecture non-mappées a été retrouvé pour l'assemblage produit, indiquant que celui-ci est adapté à ce type d'études. À travers ce travail, nous démontrons la possibilité de produire une séquence de référence de qualité à partir de l'utilisation de nouvelles technologies de séquençage. L'application de ces méthodes à un plus grand nombre d'espèces permettra ainsi à terme de faciliter l'étude de la diversité intraspécifique au sein d'espèces non-modèles.

Introduction

Knowledge in biology has been greatly improved by exploring a large diversity of species as well as evolutionary contexts. No single species is representative of the evolution of either an entire phylum or a whole genus. Exploration of the genetic diversity of nonmodel species is essential to have a better insight into the variation of the population history, recombination, selection, mutation, and the genotype-phenotype relationship. In this context, the Saccharomycotina subphylum (budding yeasts), which includes the baker's yeast *Saccharomyces cerevisiae*, represents an ideal group of nonmodel organisms for population genomic studies[1].

Recent years have seen a burst of population genomic surveys focusing on various nonconventional yeasts associated with different objectives. This has a bearing on several aspects of evolutionary biology. Analysis of resequencing data of a large sample of isolates from the same species has been focused on yeast model organisms such as *S. cerevisiae*[2–10] and the fission yeast *Schizosaccharomyces pombe*[11,12], as well as on the nonmodel yeast species *Saccharomyces paradoxus*[13], *Saccharomyces uvarum*[14], *Candida albicans*[15,16] , and *Lachancea kluyveri*[17–19]. Altogether, these data and analysis enhanced our knowledge about the evolutionary history of species[14], the forces involved in genome evolution[17] and the genetic basis of the phenotypic diversity[15].

Among the Saccharomycotina, *Dekkera bruxellensis* is a yeast species associated with human fermentation processes that is well known as a major cause of wine spoilage, and also as an essential contributor to Belgium lambic and gueuze beer fermentation[20,21]. In addition to its industrial properties, this species is of interest at the evolutionary level. Natural isolates show different ploidy levels[22,23] and extensive chromosomal rearrangements, which were observed through electrophoretic karyotypes[24]. These observations indicate a rapid evolution at the intraspecific level. Recent findings suggest that the ploidy level could be linked to the substrate of origin of the strain and related to adaptive processes linked to specific environments[25]. Consequently, a genome-wide polymorphism survey based on a representative set of *D. bruxellensis* individuals would be of interest. The exploration of single nucleotide polymorphisms (SNPs), small indels, as well as structural variants such as large indels, and inversions and translocations at the species level would help provide insight into the forces that shape genomic architecture and evolution. However, to conduct a population genomic survey, the availability of a high-quality reference sequence for the species at a completeness level to cover the majority of the genomic variation and a contiguity level to efficiently detect structural variants, is a prerequisite.

To date, population genomic studies have mostly been performed on species for which chromosomal-scale genome assemblies were available; however, this necessary high-quality assembly was unfortunately not yet available for the *D. bruxellensis* species. Here, we present the *de novo* sequence and high-quality genome assembly of the UMY321 *D. bruxellensis* isolate with a combination of long Oxford Nanopore and short Illumina reads. By aligning the short-

read sequencing data from a total of eight sequenced natural isolates on the generated assembly, as well as other previously available assemblies[22,26–29], we tested the capacity of our assembly to be used as a reference assembly for future population genomic studies of this nonmodel species. The results showed that we generated the most complete and contiguous de novo assembly of *D. bruxellensis* necessary to explore the intraspecific genetic diversity of this unique and economically relevant species.

Results and discussion

Three *D. bruxellensis* isolates (UMY321, UMY315, and 133) were sequenced in this study (Table 1). These strains were determined to be diploid based on flow cytometry analysis and were all isolated from wine or grape must in Italy or South Africa. The genome of the UMY321 isolate was sequenced using a combination of Nanopore long-read and Illumina short-read sequencing data to obtain a high-quality assembly. By contrast, the UM315 and 133 isolates were only sequenced using a short-read strategy. In addition, these genomes were compared to previously sequenced genomes of six other *D. bruxellensis* isolates[22,26–29] (Table 1).

Strain	Ploidy	Ecological origin	Geographical origin	Reference
AWRI1499	3n	wine	Australia	Curtin *et al.* (2012)
AWRI1608	3n	wine	Australia	Borneman *et al.* (2014)
AWRI1613	2n	wine	Australia	Borneman *et al.* (2014)
CBS11270	2n	industrial ethanol	Sweden	Olsen *et al*. (2015)
CBS2499	2n	wine	France	Piskur *et al.* (2012)
ST05_12_22	2n	lambic beer	Belgium	Crauwels *et al.* (2014)
UMY315	2n	must	Italy	this study
UMY321	2n	red wine	Italy	this study
133	2n	merlot wine	South Africa	this study

Table 1. Description of the *D. bruxellensis* isolates used in this study

de novo genome assembly construction and comparison

For the UMY321 isolate, a total of three MinION Mk1 runs were performed with the R7.3 chemistry using 2D library types with an 8 kb mean fragmentation size. A total of 115,559 reads representing a cumulative size of 1.15 Gb were generated, among which 41,686 2D reads showed an average quality greater than nine (2D pass reads). We focused on these 2D pass reads representing a total of 376.8Mb, with the longest read being 70,058 bp (mean = 9033 bp and median = 8676 bp) (Figure S1). Four subsets of our 2D pass reads (10X, 15X, and 20X of the longest 2D pass reads, and all of them, *i.e.* ~25X·) (Table S1) were submitted to four assemblers: ABruijn[30], Canu[31], miniasm[32], and SMARTdenovo (https://github.com/ruanjue/smartdenovo). As the MinION sequencing technology is known to be associated with high error rates[33] (~10% for 2D pass reads), we polished the assemblies with Pilon[34] using 100X of Illumina paired-end reads. The lengths of the constructed assemblies were all in the same order of magnitude and ranged from 11.7 to 13.7 Mb (Table S2).

Using these various datasets and assemblies, the objective was to define the best assembler and the minimal coverage needed. Hence, we computed the standard contiguity metrics for all assemblies to evaluate their quality, which is related to both the assembler and the dataset (Figure 1 and Table S2). First, we observed that, considering the results by assembler, the

number of scaffolds obtained with the 10X dataset is much higher compared to the other datasets, which suggests that a 10X coverage of MinION reads is too low to obtain a good quality assembly. By assembler, the results obtained for the higher coverages are comparable. Using Canu, the number of scaffolds is much higher and the N90 as well as N50 are much lower, producing less connected assemblies (Figure 1 and Table S2). The contiguity metrics associated with the assemblies constructed with SMARTdenovo, ABruijn, and miniasm were closely related, and it seemed difficult to select a single best assembly on the sole basis of these measurements, especially since good contiguity metrics are not necessarily associated with assembly completeness.

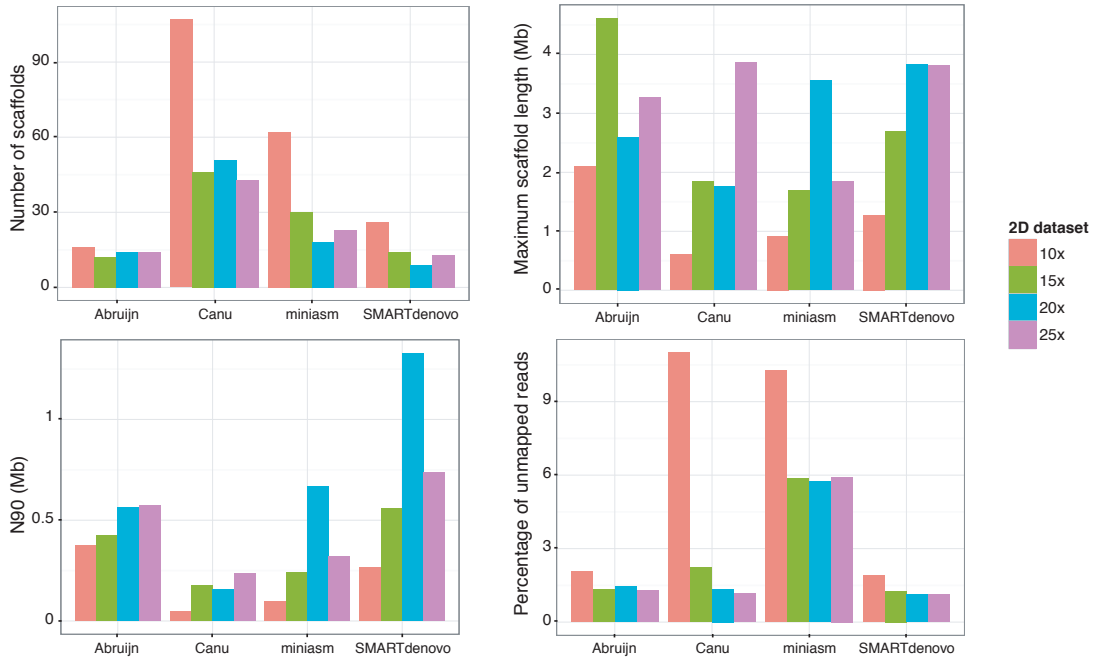


Figure 1. Metrics related to the constructed assemblies, per assembler and dataset.

Considering these results, we decided to map the Illumina paired-end reads back to the generated assemblies with BWA[35]. Among all the assemblies, the proportion of unmapped reads ranged from 1.12 to 11% (Figure 1 and Table S2). Surprisingly, the assemblies constructed with miniasm were less complete, as .5% of the reads did not map back, compared to 1.5% for the ABruijn and SMARTdenovo assemblies.

By comparing standard metrics and the proportion of unmapped reads, the most accurate assembly was obtained with the 20X 2D reads dataset combined with the SMARTdenovo assembler. This assembly is composed of nine scaffolds, *i.e.*, very close to the estimated number of chromosomes, which appears to vary between four and nine among different strains of this species[24], for a complete assembly size of 12.97 Mb. This was then submitted to SSPACE-long reads, which reduced the number of scaffolds to eight after grouping the two smallest ones, based on our long-read information, and a further Pilon run. The final assembly contains eight scaffolds and shows a cumulative size of 12,965,163 bp (Table 2). We also evaluated the completeness of our assembly at the gene content level by running CEGMA[36]: 245 out of the

248 most extremely conserved genes in eukaryotes were detected in our assembly, through 242 complete and three partial alignments. Altogether, these results reveal a high level of completeness of our assembly.

Strain	# Scaffolds	Assembly size (Mb)	Maximum scaffold size	N50	N90	# N
AWRI1499[26]	324	12.7	170,307	65,420	22,583	57
CBS11270[29]	15	17.3	4,993,495	3,706,654	944,992	2,497,785
CBS2499[27]	84	13.4	2,877,306	1,792,735	190,560	586,105
ST05_12_22[28]	85	13.1	1,439,423	732,210	177,142	218,317
UMY321	8	13	3,829,289	1,917,156	1,329,398	2,708

Table 2. Metrics associated to the *D. bruxellensis* publicly available assemblies

Comparison with available assemblies of *D. bruxellensis*

To date, several assemblies of the *D. bruxellensis* species have already been released[22,26–29]. These assemblies are related to isolates from different ecological and geographical origins (Table 1). They were mostly constructed by combining several sequencing methods, such as 454, PacBio, and Illumina, as well as optical mapping in the most recently published assembly[29].

The assemblies have very variable metrics associated with each of them (Table 2). In terms of contiguity, our assembly and the assembly generated for the CBS11270 isolate are close, and reach a chromosome- scale resolution. However, the CBS11270 assembly is much larger than the others (17.3 Mb vs. 12.7–13.4 Mb), although it does also contain ~2.5 Mb of undetermined (N) residues. By comparing the assembly metrics, we determined that our assembly is closer to that of CBS11270, which was generated by combining PacBio and Illumina sequencing methods as well as optical mapping. In addition it was also much better than the other three available for comparison, which were much more fragmented and comprised at least 84 scaffolds.

A MUMmer comparison of our UMY321 assembly to that of CBS11270 indicates that 91% and 99.6% of the assemblies aligned, respectively, with one another and revealed that the scaffolds are mostly colinear (Figure 2). However, some large repetitive regions can be observed in the CBS11270 assembly, e.g., on chromosome 1, between chromosomes 1 and 6, and between chromosomes 4 and 5 (Figure 2 and Figure S2) that are absent in our assembly, and could explain the size differences between the assemblies (17.3 Mb vs. 12.97 Mb). Moreover, some synteny breaks can be observed, at the level of scaffolds, specifically between three and four. All the inconsistencies between the assemblies could be related either to

structural rearrangements between the isolates or to assembly errors, and would require further investigations to reach a conclusion as to their most likely source.

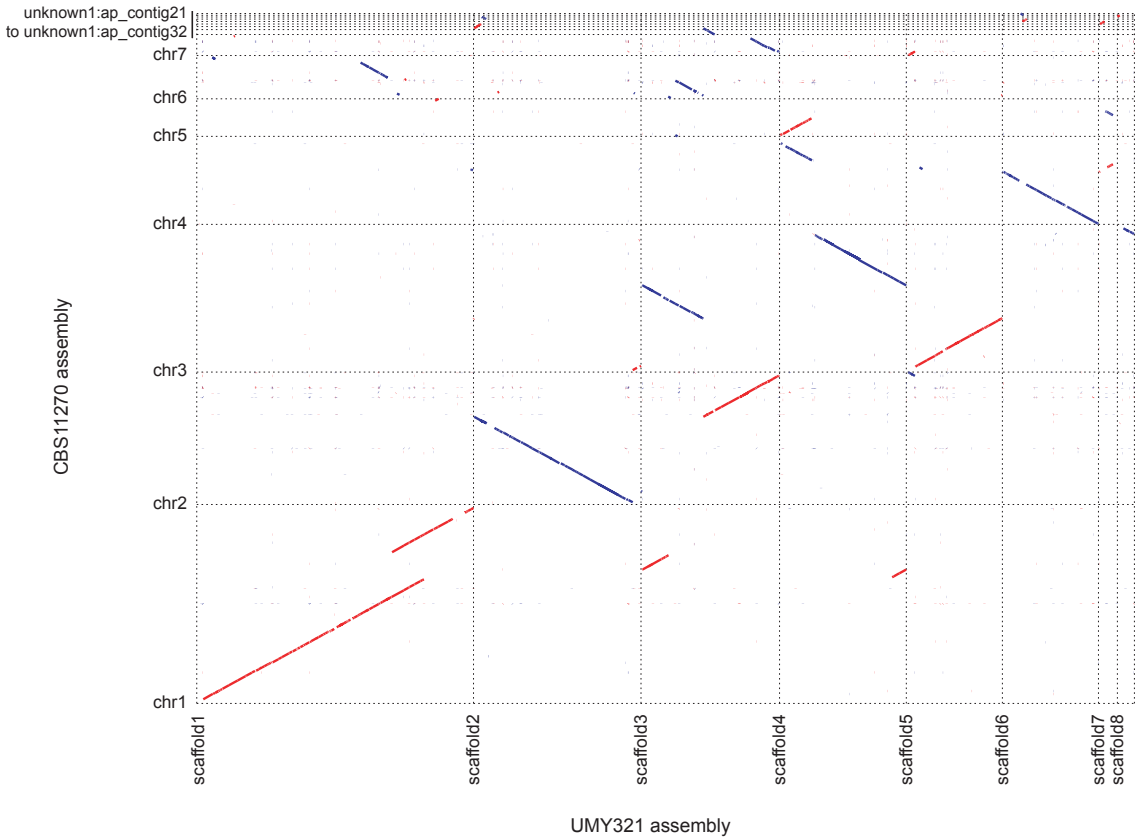


Figure 2. Comparison of the CBS11270 and UMY321 assemblies. The alignments and the pilot were generated with the MUMmer software suite. Red lines: sequences aligning in the same direction. Blue lines: sequences aligning in the opposite.

Suitability of our assembly for population genomics studies

As previously mentioned, to function as a valuable resource for conducting population genomics studies, a reference genome should combine high contiguity (for the detection of structural variants) and completeness (for the efficient detection of SNPs and small indels). At the contiguity level, our assembly is close from a chromosomal-scale resolution, which suggests that it would be highly suitable for gross structural rearrangement detection (translocations, inversions, and long insertions/deletions).

To test our assembly for the detection of polymorphism along the genome, we further investigated the mapping of the Illumina reads. As previously mentioned, 98.89% of the UMY321 Illumina reads mapped to our assembly. The read coverage was homogeneous along the scaffolds (Figure 3A), which suggests that the strain is devoid of aneuploidy and segmental duplication, and confirms the lack of large repetitive regions within our assembly.

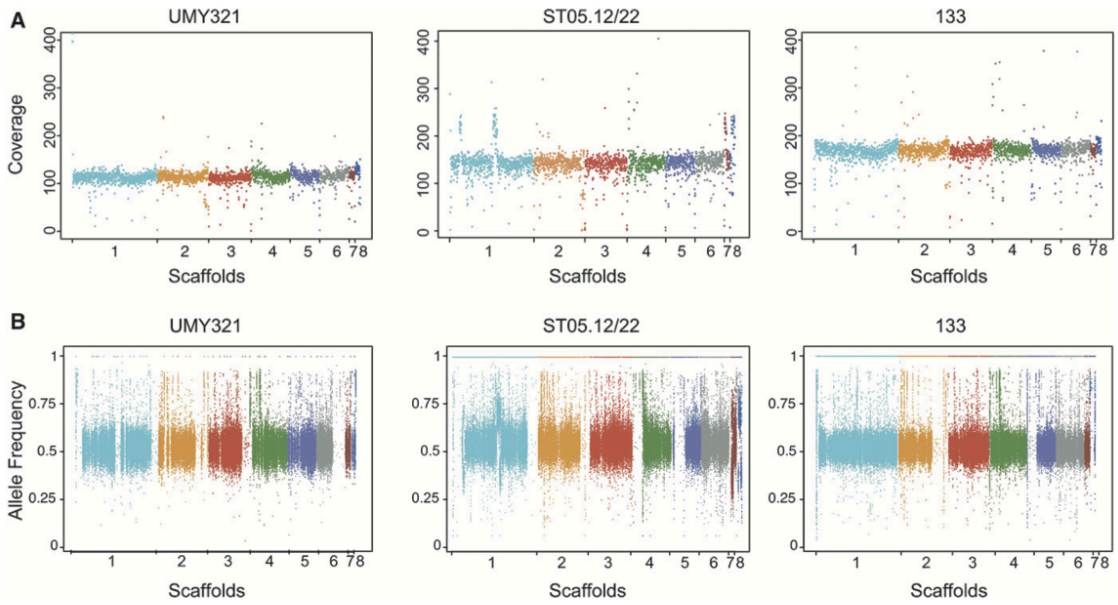


Figure 3. Mapping of the Illumina reads *vs.* the UMY321 reference assembly. (A) Illumina reads coverage along the reference genome (B) Frequency of the reference allele at heterozygous sites along the genome.

A total of 83,006 SNPs were detected with GATK[37], among which 374 were homozygous and 82,632 were heterozygous (Table S3). The 374 homozygous SNPs could be considered as false positives. Although not completely negligible, this number is very low and could be related to the high error rate of the MinION technology, which is not completely compensated by using Illumina short reads[38].

The UMY321 isolate that we sequenced is diploid, and the detection of these 82,632 heterozygous SNPs revealed that the two genomic copies are not identical and have a high heterozygosity level. These heterozygous positions are mostly evenly distributed all along the genome, with several regions showing loss of heterozygosity (LOH) on scaffolds 1, 2, 3, and 6 (Figure 3B).

		Assemblies				
		UMY321	**CBS11270**	**CBS2499**	**ST05.12/22**	**AWRI1499**
Illumina PE reads	**UMY321**	1.11	9.95	4.13	2.12	5.29
	CBS11270	4.74	12.43	5.88	3.48	9.49
	CBS2499	1.68	9.4	4.92	2.45	5.68
	ST05.12/22	1.9	10.83	7	3.78	11.97
	UMY315	0.66	10.00	4.04	2.02	5.35
	133	0.82	8.82	3.11	1.57	4.44
	AWRI1608	14.87	22.65	17.42	15.39	19.91
	AWRI1613	9.69	16.89	11.04	8.98	13.38

Table 3. Proportion of *D. bruxellensis* unmapped Illumina reads on the available assemblies

Altogether, these results confirmed that our assembly performs well when mapping the reads that were used for its construction. However, to determine if an assembly is relevant in the context of population genomic studies, we also analyzed its performance when mapping reads from other isolates. To survey polymorphisms within a species, resequencing projects rely mainly on Illumina sequencing technology, therefore we mapped the short reads related to this species that were publically available as well as from two isolates we sequenced in the context of this project (Table S4) against our assembly and reported the proportion of unmapped reads. We also aligned these reads against the publicly available assemblies to perform a comparative analysis (Table 3). As expected, the UMY321 Illumina paired-end reads mapped better on our assembly with only 1.11% of unmapped reads. More surprisingly, short reads generated in the context of the other projects also mapped better on our assembly compared to their related assemblies, and more generally compared to all other assemblies (Figure 4). It is also worth noting that all the reads, including those related to the CBS11270 isolate, mapped less efficiently to the CBS11270 assembly compared to all other assemblies, which suggests that although this assembly is highly contiguous and much larger than the others available, it is less complete.

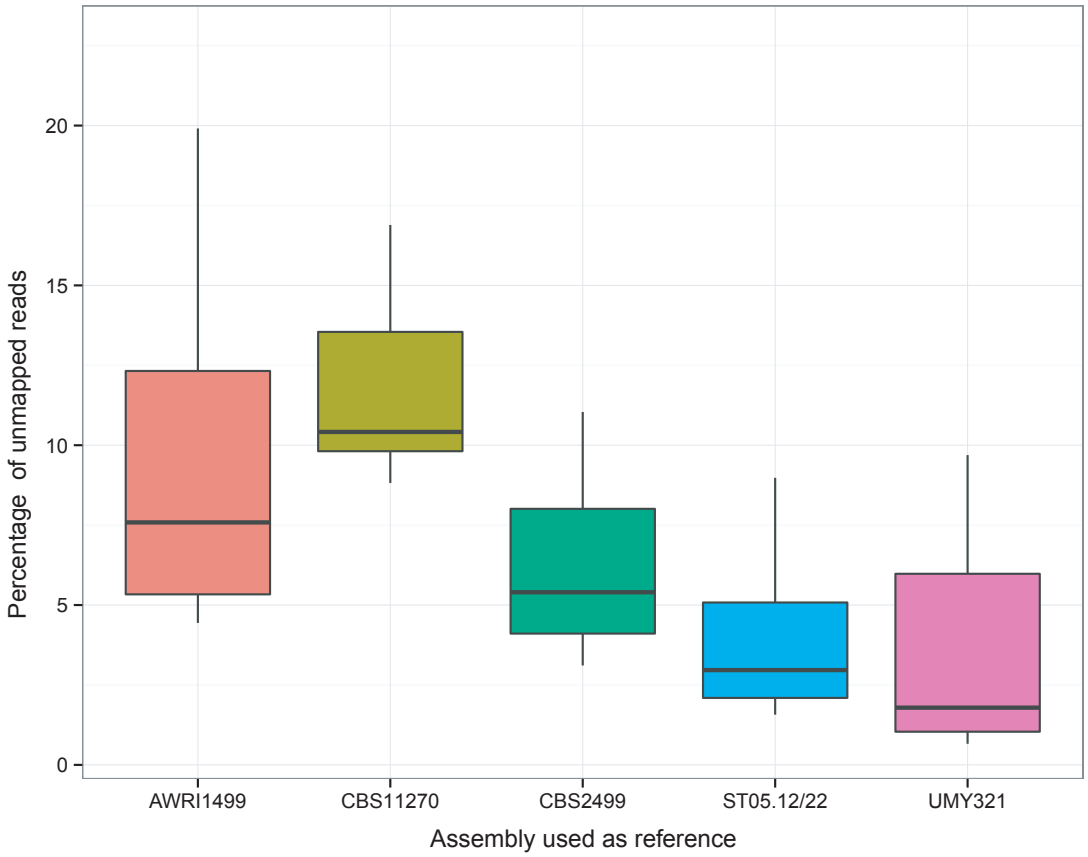


Figure 4. Illumina unmapped reads per assembly. Boxplot of the percentage of unmapped Illumina reads, according to the assembly used for the mapping.

Insight into the intraspecific genetic variability

Finally, we took advantage of the availability of Illumina reads related to different isolates in order to obtain a first glimpse into the genomic variability within this species, using our UMY321 assembly as a reference. The read coverage along the reference sequence was mostly homogeneous for all isolates, and only few deviations were observed, limited to small genomic regions, which are characteristic of segmental duplications, in the ST05.12/22 isolate (Figure 3A). This suggests that the structural variants within this species are mostly balanced. It can also be noticed that the coverage plot obtained from the CBS11270 Illumina reads did not show twofold deviations on scaffolds 1, 3, or 4 (data not shown), as expected from the comparison of the CBS11270 and UMY321 assemblies (Figure 2 and Figure S2), suggesting that the repetitive regions highlighted in the CBS11270 assembly are most probably related to assembly errors.

Among the eight studied isolates, one is triploid (AWRI1608) and all the others are diploid (Table 1). A total of 1,268,172 SNPs were detected across these eight isolates, among which 82% are heterozygous (Table S3). These SNPs are distributed over 500,707 polymorphic positions, with a majority present as singletons (68.8% of the polymorphic sites). However, a significant proportion of this variability is related to the triploid strain AWRI1608. Indeed, when this strain was not included in the analysis, 829,313 SNPs were detected over 188,717 polymorphic positions with only 50,702 singletons (27%). This is in agreement with the proposition that AWRI1608 consists of a slightly heterozygous diploid set of chromosomes with an additional full set of more distantly related chromosomes[22]. The phylogenetic relationships between this small sample of isolates based on the whole set of polymorphic positions also reflect the high divergence of this triploid isolate (Figure S3A). Ploidy levels across the genomes were also confirmed by taking advantage of allele frequency at heterozygous positions, which was ~ 0.5 for diploid isolates and 0.33/0.66 for the AWRI1608 genome (Figure S3B). These heterozygous positions are evenly distributed along the genome; however, LOH regions were detected in all the diploid isolates (Figure 3B).

Conclusion

D. bruxellensis is a yeast species of great importance in fermented beverage industries, largely thought of as a contaminant organism[20,21]. This species is also an interesting model to study genome evolution and dynamics as it is characterized by a large genomic plasticity. For these reasons, we sought to generate a high-quality genome assembly and ultimately obtain a suitable reference genome for population genomics. Our analyses show that the *D. bruxellensis* assembly that we generated with a combination of moderate coverage (20X) MinION long-reads in addition to a higher coverage (100X) of Illumina reads utilized for sequence polishing purposes, is highly valuable for population genomic studies and outperforms previously available sequences. Preliminary comparison among a small set of nine isolates already highlights the presence of large regions of LOH, which appears to be key factors in the genome evolution and adaptation of a large number of yeast species[15,39,40]. To obtain a species-wide view of the genetic variability of *D. bruxellensis*, many more isolates should be surveyed using both short-read as well as long-read sequencing techniques, which will allow for the exploration of the structural variant landscape.

References

1. Peter, J. & Schacherer, J. Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast* **33,** 73–81 (2016).
2. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458,** 337–41 (2009).
3. Schacherer, J., Shapiro, J. a, Ruderfer, D. M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458,** 342–5 (2009).
4. Skelly, D. A. *et al.* Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* **23,** 1496–504 (2013).
5. Bergström, A. *et al.* A High-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31,** 872–888 (2014).
6. Almeida, P. *et al.* A Population genomics insight into the mediterranean origins of wine yeast domestication. *Mol. Ecol.* (2015). doi:10.1111/mec.13341
7. Strope, P. K. *et al.* The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* gr.185538.114- (2015). doi:10.1101/gr.185538.114
8. Gallone, B. *et al.* Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* **166,** 1397–1410.e16 (2016).
9. Gonçalves, M. *et al.* Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Curr. Biol.* **26,** 2750–2761 (2016).
10. Zhu, Y. O., Sherlock, G. & Petrov, D. A. Whole Genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3 (Bethesda)*. (2016). doi:10.1534/g3.116.029397
11. Fawcett, J. A. *et al.* Population genomics of the fission yeast *Schizosaccharomyces pomb*e. *PLoS One* **9,** e104241 (2014).
12. Jeffares, D. C. *et al.* The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. *Nat. Genet.* **advance on,** (2015).
13. Leducq, J.-B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat. Microbiol.* **1,** 15003 (2016).
14. Almeida, P. *et al.* A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. *Nat. Commun.* **5,** 4044 (2014).
15. Ford, C. B. *et al.* The evolution of drug resistance in clinical isolates of *Candida albicans*. *Elife* **4,** e00662 (2015).
16. Hirakawa, M. P. *et al.* Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Res.* gr.174623.114- (2015). doi:10.1101/gr.174623.114
17. Friedrich, A., Jung, P., Reisser, C., Fischer, G. & Schacherer, J. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol. Biol. Evol.* **32,** 184–92 (2015).
18. Brion, C., Pflieger, D., Friedrich, A. & Schacherer, J. Evolution of intraspecific transcriptomic landscapes in yeasts. *Nucleic Acids Res.* **43,** 4558–68 (2015).
19. Brion, C., Pflieger, D., Souali-Crespo, S., Friedrich, A. & Schacherer, J. Differences in environmental stress response between yeasts is consistent with species-specific lifestyles. *Mol. Biol. Cell* (2016). doi:10.1091/mbc.E15-12-0816
20. Schifferdecker, A. J., Dashko, S., Ishchuk, O. P. & Piškur, J. The wine and beer yeast *Dekkera bruxellensis*. *Yeast* **31,** 323–332 (2014).
21. Masneuf-Pomarede, I., Bely, M., Marullo, P. & Albertin, W. The genetics of non-conventional wine yeasts: Current Knowledge and Future Challenges. *Front. Microbiol.* **6,** 1563 (2016).
22. Borneman, A. R., Zeppel, R., Chambers, P. J. & Curtin, C. D. Insights into the *Dekkera bruxellensis* genomic landscape: comparative genomics reveals variations in ploidy and nutrient utilisation potential amongst wine isolates. *PLoS Genet.* **10,** e1004161 (2014).
23. Curtin, C. D. & Pretorius, I. S. Genomic insights into the evolution of industrial yeast species *Brettanomyces bruxellensis*. *FEMS Yeast Res.* **14,** 997–1005 (2014).
24. Hellborg, L. & Piskur, J. Complex nature of the genome in a wine spoilage yeast, *Dekkera bruxellensis*. *Eukaryot. Cell* **8,** 1739–49 (2009).

25. Albertin, W. *et al.* Development of microsatellite markers for the rapid and reliable genotyping of *Brettanomyces bruxellensis* at strain level. *Food Microbiol.* **42,** 188–195 (2014).
26. Curtin, C. D., Borneman, A. R., Chambers, P. J. & Pretorius, I. S. De-novo assembly and analysis of the heterozygous triploid genome of the wine spoilage yeast *Dekkera bruxellensis* AWRI1499. *PLoS One* **7,** e33840 (2012).
27. Piškur, J. *et al.* The genome of wine yeast *Dekkera bruxellensis* provides a tool to explore its food-related properties. *Int. J. Food Microbiol.* **157,** 202–9 (2012).
28. Crauwels, S. *et al.* Assessing genetic diversity among *Brettanomyces* yeasts by DNA fingerprinting and whole-genome sequencing. *Appl. Environ. Microbiol.* **80,** 4398–413 (2014).
29. Olsen, R.-A. *et al.* De novo assembly of *Dekkera bruxellensis*: a multi technology approach using short and long-read sequencing and optical mapping. *Gigascience* **4,** 56 (2015).
30. Lin, Y. *et al.* Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E8396–E8405 (2016).
31. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33,** 623–630 (2015).
32. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32,** 2103–10 (2016).
33. Jain, M. *et al.* The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17,** 239 (2016).
34. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9,** e112963 (2014).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).
36. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067 (2007).
37. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).
38. Istace, B. *et al.* de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* **6,** 1–13 (2017).
39. Magwene, P. M. *et al.* Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1987–1992 (2011).
40. Smukowski Heil, C. S. *et al.* Loss of heterozygosity drives adaptation in hybrid yeast. *Mol. Biol. Evol.* **34,** 1596–1612 (2017).

Supplementary material

Supplementary tables

2D pass reads dataset	Number of reads	Sum of reads size (Mb)	Reads mean size (bp)
10X	10 379	150	14 452
15X	17 719	225	12 698
20X	26 637	300	11 262
all	41 686	376.8	9 038

Supplementary table 1: Description of the 2D pass reads datasets

Assemblers	2D reads dataset	# scaffolds	ace length	max scaff size	min scaff size	N50	N90	mean size	% Unmapped Illumina PE reads
miniasm	10X	62	11 995 694	920 223	22 080	280 336	98 676	193 478	10.27
	15X	30	12 688 787	1 691 626	21 543	697 670	239 673	422 959	5.85
	20X	18	12 780 442	3 560 421	19 183	1 231 147	665 673	710 024	5.73
	all	23	12 900 249	1 842 309	19 189	1 000 815	319 373	560 880	5.93
SMARTdenovo	10X	26	12 763 255	1 280 032	37 561	698 086	267 638	490 894	1.90
	15X	14	12 934 829	2 690 205	171 269	1 484 593	558 866	923 916	1.27
	20X	**9**	**12 968 841**	**3 829 828**	**63 051**	**1 917 610**	**1 329 299**	**1 440 982**	**1.12**
	all	13	13 092 154	3 812 985	19 560	1 917 659	738 108	1 007 088	1.14
Abruijn	10X	16	12 843 903	2 096 071	187 086	1 181 299	376 514	802 743	2.06
	15X	12	12 858 037	4 613 209	230 802	1 162 267	423 352	1 071 503	1.34
	20X	14	12 891 057	2 599 228	194 230	1 030 314	564 767	920 789	1.45
	all	14	12 891 513	3 270 218	137 679	1 926 745	574 037	920 822	1.28
Canu	10X	107	11 665 570	608 446	6 322	162 168	47 877	109 024	11.00
	15X	46	13 064 238	1 846 744	9 058	624 297	175 675	284 005	2.22
	20X	51	13 621 047	1 762 944	13 433	815 578	156 999	267 079	1.36
	all	43	13 730 558	3 866 340	16 769	1 635 393	237 482	319 315	1.16
Final assembly (SMARTdenovo + SSPACE-longreads)	20X	8	12 965 163	3 829 289	236 484	1 917 156	1 329 398	1 620 645	1.11

Supplementary table 2: Metrics of the constructed assemblies, after a polishing step using Pilon

	# SNPs	# heterozygous SNPs	# homozygous SNPs
UMY321	83 006	82 632	374
CBS11270	112 289	75 565	36 724
CBS2499	126 944	94 797	32 147
ST05.12/22	124 598	86 467	38 131
UMY315	125 900	99 985	25 915
133	128 345	96 175	32 170
AWRI1608	438 859	413 600	25 259
AWRI1613	128 231	96 433	31 798
Total	1 268 172	1 045 654	222 518

Table S3. Number of SNPs detected for the studied isolates compared to the UMY321 reference sequence

		SRA ID	
Strain	**Reference**	**for proportion of unmapped reads**	**for population genomics study**
AWRI1608	Borneman et al., 2014	SRR945538	SRR945538
AWRI1613	Borneman et al., 2014	SRR945561	SRR945561
CBS11270	Olsen et al., 2015	ERR1102659, ERR1102660, ERR1102661, ERR1102662,ERR1102663, ERR1102664	ERR1102661
CBS2499	Piskur et al., 2012	SRR427169, SRR427170, SRR3927080, SRR3927081	SRR3927081
ST05_12_22	Crauwels et al., 2014	SRR1222151, SRR1222155, SRR1222162	SRR1222151
UMY315	this study	ERR2004535	ERR2004535
UMY321	this study	ERR2004534	ERR2004534
133	this study	ERR2004536	ERR2004536

Table S4. SRA accession of the Illumina reads used to determine the proportion of unmapped reads and for population genomic purposes

Supplementary figures

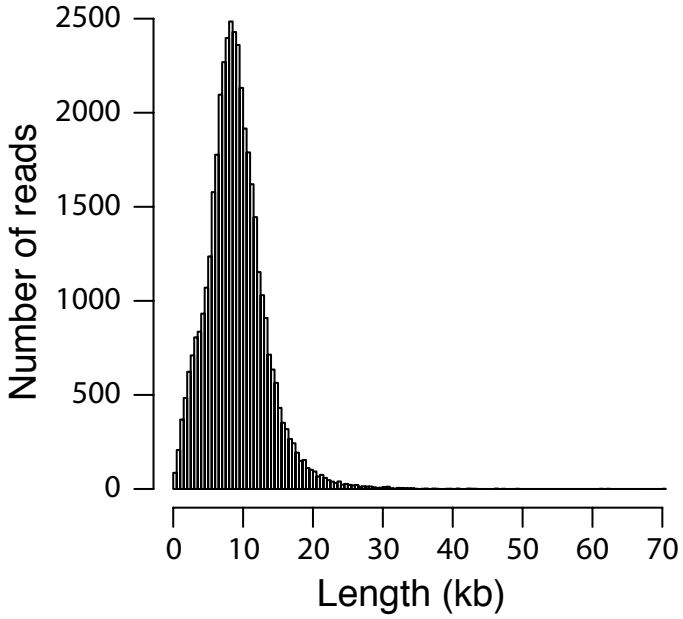


Figure S1. Distribution of the lengths of the 41,686 2D pass reads.

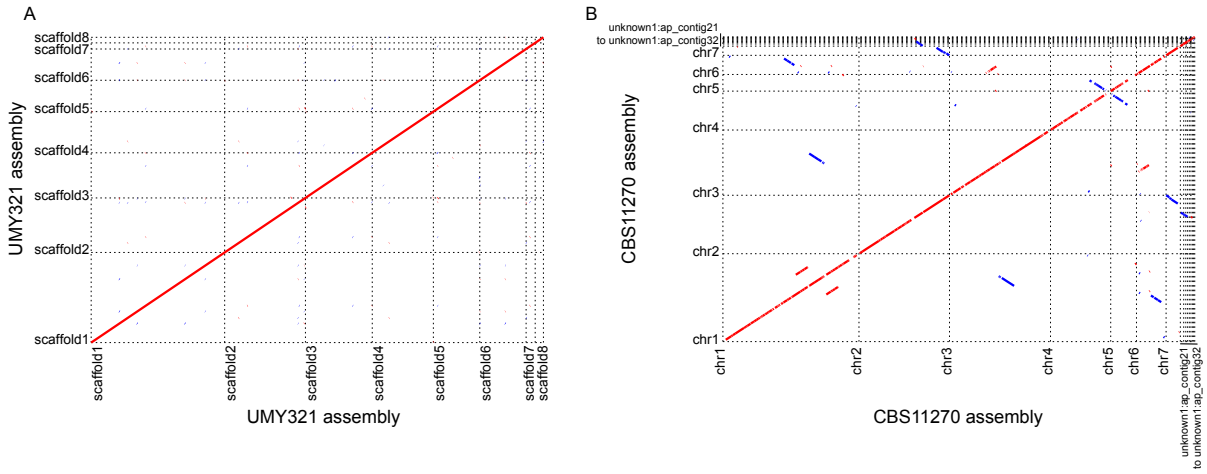


Figure S2. MUMmer-based genomic comparison of assemblies for large repeated regions detection. A. The UMY321 assembly is compared to itself. B. The CBS11270 assembly is compared to itself.

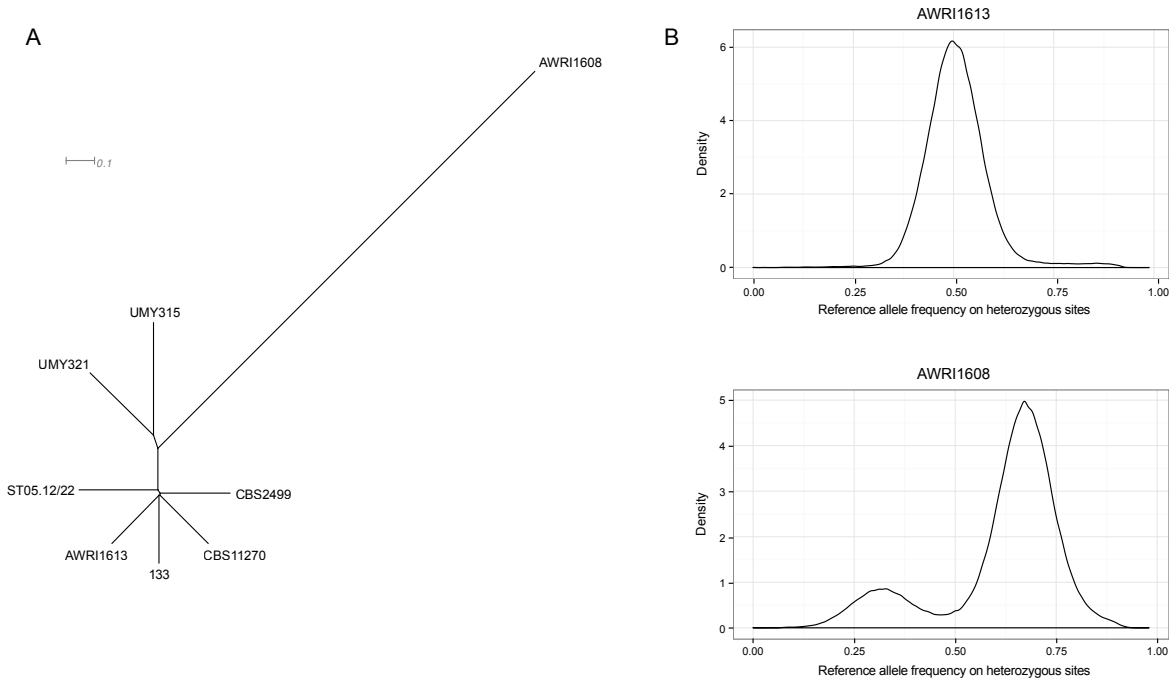


Figure S3. Intraspecific genetic variability. (A) Neighbor-joining tree of *D. bruxellensis* isolates constructed based on 500,707 polymorphic sites. (B) Density of the reference allele frequency for a diploid isolate (AWRI1613) and for a triploid isolate (AWRI1608).

Publication related to this chapter:

Fournier, T.*, Gounot, J.-S.*, Freel, K., Cruaud, C., Lemainque, A., Aury, J.-M., Wincker, P., Schacherer, J., and Friedrich, A. (2017). High-quality *de novo* genome assembly of the *Dekkera bruxellensis* yeast using Nanopore MinION Sequencing. G3 (Bethesda). *7*, 3243–3250.

* These authors contributed equally to this work

CHAPTER 3

High complexity and degree of genetic variation in *Dekkera bruxellensis* population

Résumé

Le génome de référence que nous avons généré pour l'espèce *D. bruxellensis* présente une contiguïté et une séquence relativement complète. Elle pose désormais les fondations pour l'exploration de la diversité génétique présente entre individus appartenant à cette espèce. De manière intéressante, cette dernière possède un ensemble de caractéristiques motivant la mise en place d'une telle étude. Premièrement, les isolats présentent une importante variabilité phénotypique et sont associés à de nombreux processus de fermentations tels que la production de vin ou de bière. Deuxièmement, plusieurs études ont révélé la présence d'évènements d'hybridation et de perte d'hétérozygotie au sein de plusieurs souches. De manière additionnelle, l'étude du caryotype de plusieurs isolats a montré de nombreux réarrangements chromosomiques au sein de l'espèce, suggérant une dynamique chromosomique importante dans la structure des génomes. L'étude de la variabilité génétique au sein de cette espèce est cependant restée limitée à la comparaison d'assemblages d'un faible nombre d'isolats ($N < 5$) ou à l'analyse d'un nombre limité de loci chez plusieurs centaines de souches. L'exploration de la variabilité du génome complet au sein d'une population se révèle ainsi nécessaire afin de mieux comprendre l'évolution de cette espèce. Dans ce cadre, nous avons séquencé le génome complet de 53 souches de *D. bruxellensis* isolées à travers le monde à partir d'une stratégie de type Illumina, ainsi que le génome de trois souches grâce à la technologie Oxford Nanopore. Le génome de référence généré (chapitre 2) a dans un premier temps été annoté afin de pouvoir réaliser une analyse fonctionnelle. Dans un second temps, cette étude a permis l'identification de plusieurs sous-populations caractérisées notamment par la ploïdie des souches (diploïdes et triploïdes). L'étude de la variabilité génétique le long du génome a permis de retracer l'origine d'un de ces évènements, partagée entre deux sous-populations. Dans un troisième temps, le pangénome de l'espèce a été reconstruit et un ensemble de variants structuraux ont été identifiés (aneuploïdies, CNV, translocations de large taille), permettant d'étudier l'étendu des variations structurales au sein de la population. Ces variations sont importantes au sein de *D. bruxellensis* et affectent principalement les souches triploïdes. Finalement, la comparaison de certains de ces variants entre souches a permis d'identifier des gènes ayant un rôle dans la variabilité phénotypique de l'espèce. Par exemple, plusieurs gènes impliqués dans le transport de sucres ou de drogues ont été retrouvés avec un nombre de copies élevé, notamment chez une sous-population triploïde impliquée dans la production de vin. L'ensemble de ces résultats démontre l'intérêt de l'établissement d'études de génomique des populations au sein de *D. bruxellensis*, et plus globalement d'espèces non-modèles.

Introduction

The yeast species *Dekkera bruxellensis* (anamorph *Brettanomyces bruxellensis*) is a distant relative of *Saccharomyces cerevisiae* since lineages diverged more than 200 million years ago. Interestingly, these species share characteristics that have been independently acquired during evolution, such as the metabolic ability to produce ethanol in the presence of oxygen and excess of glucose[1]. Both species also share the capacity to efficiently catabolize the produced ethanol, their corresponding life style being described as 'make-accumulate-consume (ethanol)'[2]. This strategy allows *D. bruxellensis*, which is associated to several human fermentation processes, to survive and develop in harsh and limiting environmental conditions and for now, this yeast species has exclusively been isolated from anthropized niches.

Its impact on the fermentative processes is however contrasted. As an example, it has a positive contribution to brewing of Belgium Lambic and Gueuze beers, but it is also well known as a major wine-spoilage by producing odorant molecules (volatile phenol) described as barnyard or horse sweat[3]. It has also been found associated with other food or industrial processes, such as kombucha or bioethanol, for which its contribution is still unclear[4,5]. The opposite contributions of *D. bruxellensis* in fermentation processes raised a growing interest in this species, and high phenotypic variability regarding for example sugar metabolism or nitrogen source utilization has been highlighted between the isolates in various studies[6–8].

The observed phenotypic variability is undoubtedly, at least partly, related to its genomic plasticity, another peculiarity of this species that defines it as a good model at the evolutionary level. Indeed, the first proof of this high genomic variability was provided through the comparison of electrophoretic karyotypes from different strains that showed extensive chromosomal rearrangements[9], indicating a rapid evolution at the intraspecific level. Whole genome sequencing of a small number of isolates revealed a variation of ploidy between isolates with some triploid individuals deriving from an allotriploidization event involving a moderately heterozygous diploid and a more distantly related haploid[6]. The level of ploidy is supposed to be linked to the substrate of isolation and geographical distribution[10]. For now, genomic variability has essentially been restricted to the exploration of small regions of the genomes[11–16]. More recently, the complete genomes of a limited number of strains have been sequenced and assembled[6,11,17–19]. Comparison of these genomes have suggested that polyploid and hybridization events might play a significant role in this species evolution[6]. The most recent population genomic study was performed on a collection of almost 1,500 isolates based on 12 microsatellite regions and showed that the population is structured according to ploidy level, substrate of isolation and geographical origin of the strain[12].

With the currently available sequencing technologies, it is now possible to explore the intraspecific variability of a species at the genome-wide level. Such population genomic studies have been performed on multiple yeast species, including *S. cerevisiae*[20–27] and *Schizosaccharomyces pombe*[28,29] but also non-conventional yeast species[30–36], granting better insights into their respective evolutionary histories as well as their genotype-phenotype

relationships. Here, we conducted a population genomic survey of *D. bruxellensis* based on whole genome sequencing data. In total, 53 isolates collected worldwide were sequenced with Illumima HiSeq technology. The data shows a high degree and complexity of genetic variation among isolates. This species displays a variable level of nucleotide diversity, heterozygosity and copy number variants, depending on the ploidy of the isolates. In addition, structural variant content has been further examined by sequencing three strains using the Oxford Nanopore sequencing technology, producing long reads which allow a comparative assembly analysis. This dataset offers a first view of the genomic variants at the genome-scale within *D. bruxellensis*. Overall, this study provides insights into the evolutionary history of the species and the identification of genetics contents linked to subpopulation adaptations.

Results and discussion

In order to survey genome-wide variability within this species, we gathered a collection of 53 *D. bruxellensis* isolates from diverse origins (Table S1), with a large part of our collection being wine-related. This collection is representative of all five major clusters previously defined using microsatellites and based on almost 1,500 isolates[12]. These strains were mostly isolated in Europe (*e.g.* Belgium, Italy, Spain) but also South Africa, Australia and Chile. We sequenced the genome of all strains with a short-read sequencing strategy to at least 40-fold coverage, with a mean of 98-fold coverage. With the intent to capture large-scale genomic rearrangements, a subset of 3 isolates was also submitted to Oxford Nanopores long-read sequencing strategy. The genetic diversity was explored through the comparison of the isolates sequences with a recently published reference sequence of the species[37] *via* the mapping of the short-reads, but also through the construction of genome assemblies in order to define the gene repertoire of the species (assembly of Illumina short-reads) and to detect gross structural variants (assembly of Oxford Nanopore long-reads).

Gene content of the *Dekkera bruxellensis* genome

The recent release of a highly contiguous reference assembly for *D. bruxellensis*[37] opens the way to a genome-wide exploration of intraspecific variability. However, functional analyses of the genomic variability require genetic elements to be located on the genome sequence, which led us to carry out a complete annotation of this assembly. A total of 5,226 protein-coding genes were predicted, among which 1,427 were interrupted by frameshifts or stop codons probably due to sequencing or assembly errors linked to the heterozygous diploid state of the original isolate. This high proportion of out-of-frame genes encouraged us to refine the reference sequence. This was performed by retrieving the sequences of the concerned genes in an independent assembly constructed with Illumina reads and scaffolded with Redundans[38]. This procedure allows inferring 872 in-frame gene sequences in the initial assembly leading to a total of 4,671 in-frame protein coding genes (90%) within the genome. The remaining 555 out-of-frame genes were considered as pseudogenes.

Publically available RNA-seq data[18] allowed the identification of 509 introns in 472 genes, *i.e.* 9% of the genes, with 24 introns located in UTR. In total, 99 tRNA genes with 39 different anticodons were listed. The 26 transposable elements found in UMY321 belong to the LTR retrotransposons. Most of them are closely related to *D. hansenii* Tdh5. In addition to the intact and degenerate copies, at least 96 solo LTR were detected, sometimes grouped into large regions of up to 40 kb that may correspond to centromeres as reported for *D. hansenii*[39].

Genetic diversity variation is related to ploidy level within *D. bruxellensis*

The mapping of the Illumina paired-end reads on the reference sequence allows for single nucleotide variants detection. From 84,890 to 502,399 SNPs were highlighted per isolate,

distributed among 811,159 polymorphic positions. A minimum of 84,500 heterozygous variants was called per sample, revealing that our dataset was devoid of haploid or homozygous isolates. The allele frequency at heterozygous sites was used to infer the ploidy of each isolate: 39 isolates showed values centered on 0.5 and were therefore considered as diploid. The remaining 14 isolates had an enrichment of values centered on 0.33 and 0.66 and were consequently defined as triploid (Figure 1).

Figure 1. Allele frequency distribution for each sample. Colors correspond to each defined cluster (see Figure 2).

Interestingly, the number of homozygous variants was in the same range for the whole strain collection except for a group of 10 diploid strains, which includes the reference, and for which fewer than 350 homozygous variants were called, suggesting that they are all closely related

(Table S2). By contrast, the number of heterozygous sites was much larger for triploid isolates compared to the diploid strains (Figure S1), which is in accordance with previous studies that showed that two triploid Australian isolates were constituted of a moderately heterozygous diploid genome in combination with a divergent haploid genome[6,17]. Interestingly, the isolate YJS5382 also displays a triploid genome while being more closely related to diploid strains, suggesting a more recent and independent triploidisation event for this strain.

Overall, the genetic diversity estimated by the average pairwise difference between strains is relatively high ($\pi = 1.2 \times 10^{-2}$) compared to *S. cerevisiae* ($\pi = 3 \times 10^{-3}$). As expected, the genetic diversity is lower in coding regions ($\pi = 1.0 \times 10^{-2}$) compared to intergenic regions ($\pi = 1.5 \times 10^{-2}$). In addition, SNPs found in CDS (CoDing Sequence) are mostly synonymous (64%) while nonsense mutations are rare (1%). The latter also show a lower allele frequency within the population and a more contrasted but similar pattern can be observed for missense variants (Figure S2).

Phylogeny and strain relatedness in *D. bruxellensis*

The 811,159 polymorphic positions were used to infer phylogenetic relationships between the isolates. To handle heterozygosity within our samples, heterozygous SNPs were encoded using the IUPAC code and the average state method was used for the distance calculation. Five distinct clusters were highlighted in the neighbor-joining tree, three of them being composed of diploid isolates while the remaining two contained exclusively triploid isolates (Figure 2).

These clusters are mostly in accordance with the ones recently described[12]. For example, the two triploid clusters G3N1 and G3N2 correspond respectively to the triploid Beer AWRI1608-like cluster and the triploid Wine AWRI1499-like cluster. Moreover, 3 distinct diploid clusters can be observed allowing a better discrimination of this population. The G2N3 corresponds to the wine CBS2499-like cluster, and G2N1, which contains the wine reference strain YJS5341 isolated in Italy is separated from the G2N3 cluster, both of them being grouped in a single Kombucha-like cluster in the former analysis.

The YJS5382 isolate could not be assigned to any cluster in our analysis, which is also in accordance with a previous study[12], as it is the only representative of the Wine L0308-like cluster in our dataset. Interestingly, this isolate is the only triploid that does not group with our G3 clusters, on the contrary it's in fact closer to the diploid clusters, suggesting an independent triploidisation event for this strain. Additionally, the YJS5416 strain is not closely related to any cluster in our tree and is located between the G2N1 and G2N2 clusters and may therefore be a representative of another subpopulation.

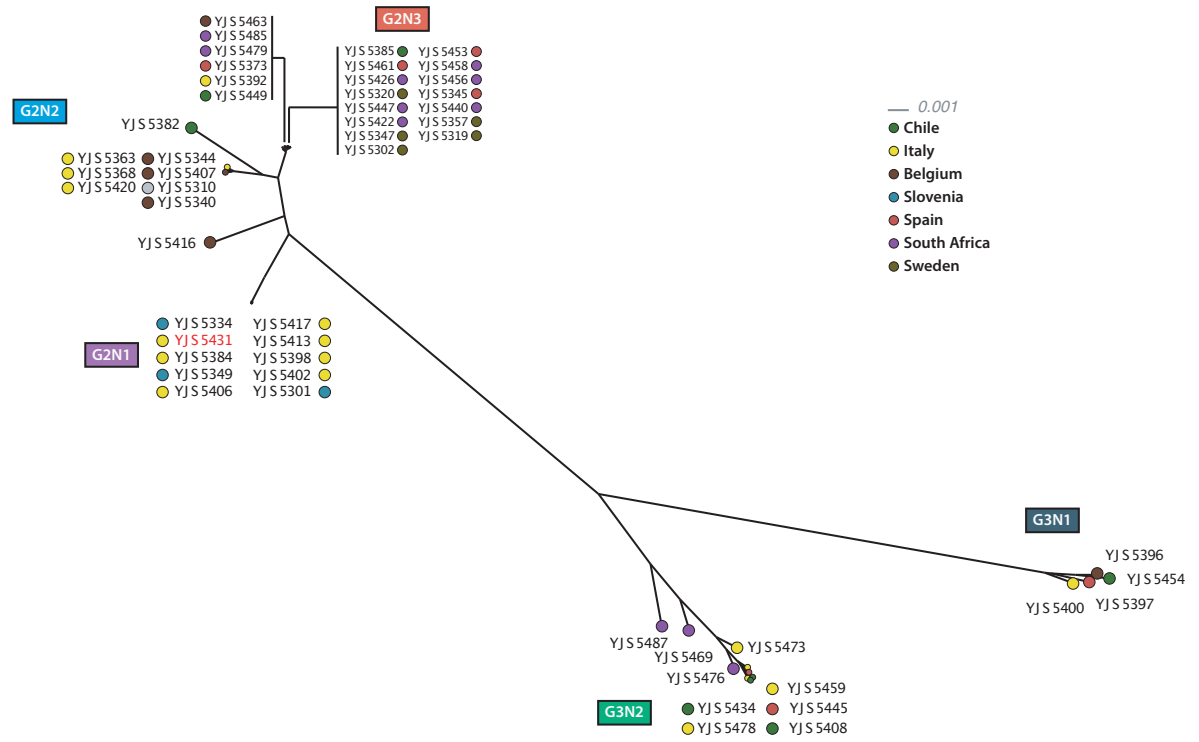


Figure 2. Neighbor-joining tree based on the SNPs of the 53 isolates. We identified 5 clusters composed of diploid (G2) or triploid (G3) strains. YJS5382 is the only triploid strain found among the diploid clusters (left side).

Genomes are punctuated by a few loss-of-heterozygosity regions

To have a better insight into the nucleotide variation along the genome, we examined the genetic diversity across non-overlapping windows of 10 kb for each cluster (Figure 3). Overall, the triploid subpopulations show a higher genetic diversity (π_{G3N1} & π_{G3N2} = 2 x 10^{-2}) compared to diploid clusters (π_{G2N1} = 3.5 x 10^{-3}, π_{G2N2} = 6 x 10^{-3}, π_{G2N3} = 4 x 10^{-3}). However, the nucleotide diversity is not homogeneous along the genome (Figure 3). In fact, the two triploid clusters display a lower genetic diversity (π < 1 x 10^{-2}) within a large region (~ 1.2 Mb) on the left side of the scaffold 1. Interestingly, two similar regions can be observed in the triploid wine subpopulation (G3N1) on the right extremity of the scaffolds 3 and 5. In the diploid subpopulations, similar regions exhibiting the same pattern can also be observed on almost all chromosomes, especially in the G2N2 and G2N3 clusters for which a high decrease of the Tajima's D values can also be observed, indicating a high proportion of rare alleles in these regions compared to the rest of the genome. While the Tajima's D value never drops below 0 with the exception of the right side of the chromosome 6 in the G2N1 cluster, the impacted region showed a lower Tajima's D value. This difference might be imputed to the high proximity between the strains belonging to this cluster and the reference sequence.

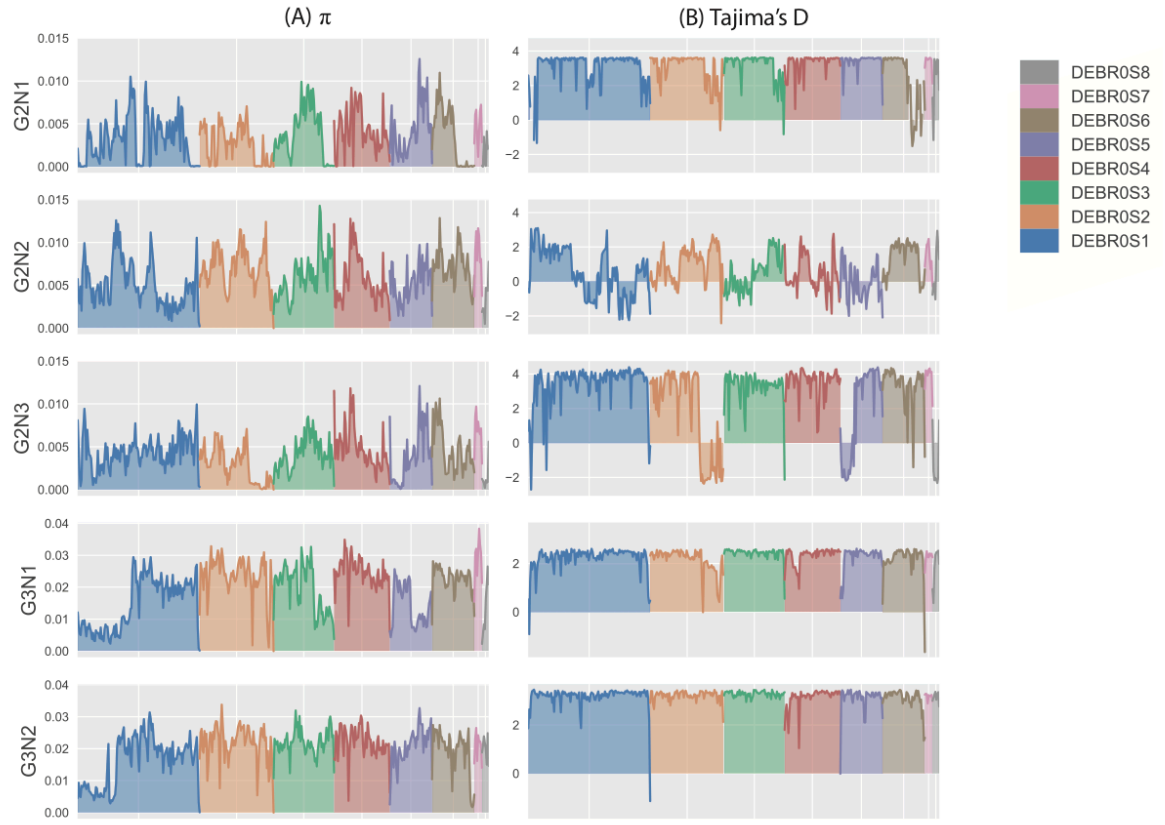


Figure 3. Variation of nucleotide diversity metrics for each cluster using non-overlapping window (30 kb). (A) Pairwise nucleotide diversity (π), (B) Tajima's D values.

To get insights into the origins of this genetic diversity decrease, we investigated these regions for the presence of loss of heterozygosity events (LOH). This mechanism leads to a decrease of the genetic variability allowing the expression of recessive alleles and therefore can result in a beneficial adaptation in some environments. The prevalence of LOH events has recently been observed in several yeast species such as *S. cerevisiae*[27], *C. albicans*[34] and *K. marxianus*[36]. Moreover, previous studies of *D. bruxellensis* isolates revealed the presence of several regions which underwent LOH, especially for the wine strains AWRI1613 and CBS2499 for which 17.9% and 16.3% of the genome are impacted[6]. To detect LOH regions in our collection, the allele frequency of the polymorphic sites was plotted along their chromosomal location in the reference assembly and the number of heterozygous SNPs along the genome was furthermore examined using sliding windows of 10 kb (Figure S3).

In all diploid isolates, several regions of LOH were observed representing a total of almost 0.7 to 2.7 Mb (mean value of 1.8 Mb, 13% of the genome) distributed among 6 to 18 regions (mean value of 11.7) (Table S3). This value is low compared to *S. cerevisiae* for which LOH cover approximately 50% of the genome on average[27]. Patterns of LOH are very well conserved within clusters and only very few regions are strain-specific. Some LOH events seem to have arisen before cluster expansion as they are shared between several clusters. Moreover, most of the identified regions also showed a low nucleotide diversity, suggesting that LOH is the main reason for the nucleotide variability patterns mentioned above. In triploid isolates, regions

exhibiting a complete loss of heterozygosity are uncommon and only 1 to 3 LOH regions were detected covering 29 to 279 kb. These regions are once more very well conserved between the strains. Interestingly, several regions in the triploid strains, corresponding to the low nucleotide diversity locations, showed a reduced heterozygous rate but were not considered as LOH regions with our criteria. These regions could result from an ancient LOH event that occurred before the separation of the two triploid clusters and the accumulation of new mutations, especially for the left part of the chromosome 1.

Masking LOH regions allowed us to precisely determine the level of heterozygosity, which ranges from 6.2 to 36.3 heterozygous sites per kb (Table S4). As expected, all diploid isolates share the same level of heterozygosity, with a mean value of 7.1. Much higher heterozygosity rates were associated with triploid strains: YJS5382 has 11.4 heterozygous SNPs/kb, and this level is much higher within the triploid clusters. Indeed, a mean of 34 heterozygous SNPs/kb can be observed, this level being slightly higher in the G3N2 cluster compared to the other triploid subpopulations with 31 and 35 heterozygous SNPs/kb respectively. This difference is most likely due to the regions which have undergone LOH in the G3N1 cluster and which are not found in the G3N2 strains.

Aneuploidies and segmental copy variants are not common in *D. bruxellensis*

To determine the frequency of aneuploidy as well as of segmental copy variants in our collection, we examined the coverage distribution along the reference genome using non-overlapping 20 kb windows. Coverage deviations were confirmed with the allele frequency variation in these regions. We observed such deviations at whole scaffold level for only 3 isolates (5,6% of the 53 strains), these being related to the smallest ones (scaffolds 7 and 8). This result suggests that aneuploidies are rare in *D. bruxellensis* compared to other species such as *S. cerevisiae* for which analysis of 1,011 strains has revealed that around 20% of the population is affected by aneuploidies[27]. Segmental variations were more prevalent and detected in 17 isolates (Figure 4), 9 of them showing several regions affected. However, the relatively weak prevalence of these events and the size of the affected regions (mean size around 350 kb) suggest that they could not explain the highly variable karyotypes observed within this species on their own and that balanced rearrangements may also occur. Moreover, triploid clusters display a higher proportion of samples carrying segmental variants with more than half of the samples affected by such variants (Figure S4). This result supports the idea that hybrids and generally, polyploid strains, are affected by structural changes[40].

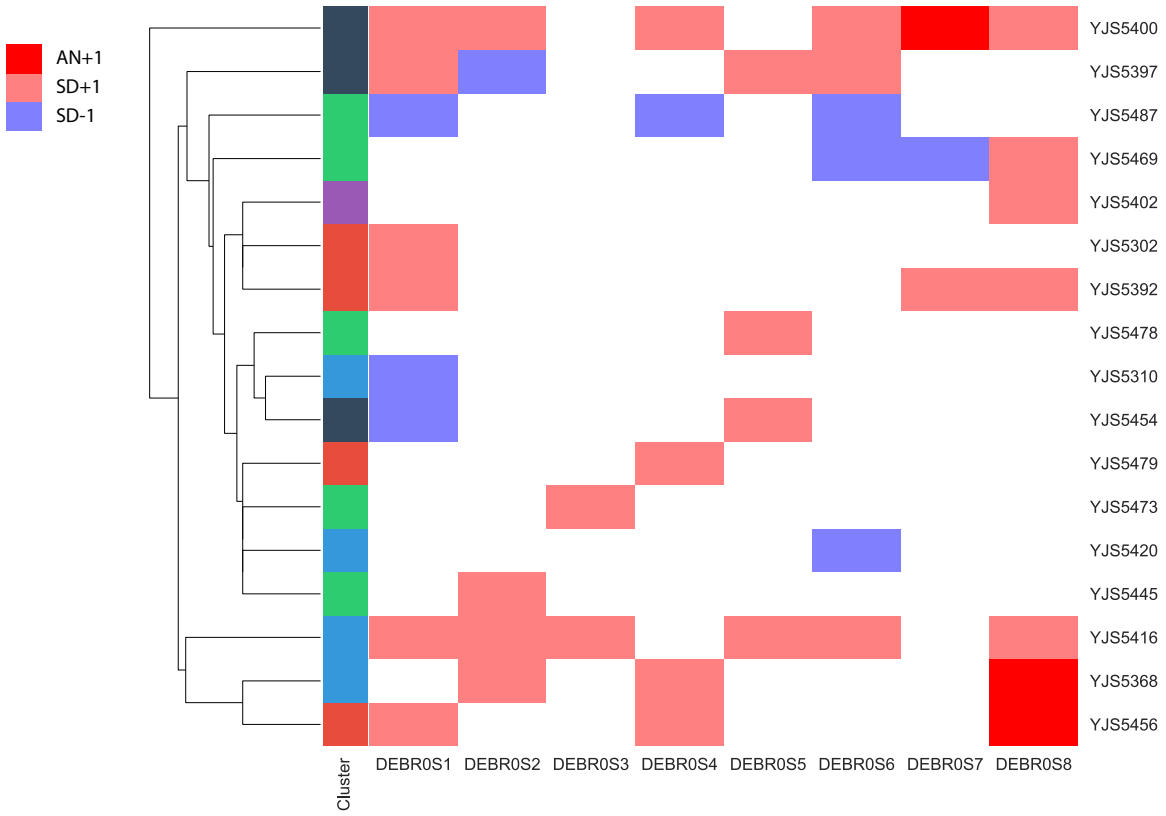


Figure 4. Distribution of aneuploidies (AN) or segmental duplications (SD) for each sample.

We then focused on triploid isolates to determine whether these copy variants affected the haploid or diploid genomic version by taking advantage of the allele frequency within these regions. Indeed, while a duplication of the haploid version will result in an equal genome version ratio (2:2) and therefore in a frequency shift to 0.5, a supplemental copy of the diploid genome will lead to a frequency of 0.25 and 0.75 (3:1). On the other hand, a deletion of the haploid and diploid versions will result in an allele frequency of 1 (2:0) and 0.5 (1:1) in these regions, respectively. Nine triploid strains carrying segmental duplications were investigated (Figure S5, Table S5). A total of seven duplicated regions showed a 3:1 ratio *vs* four a 2:2 ratio, suggesting that duplication of the diploid genome is more common. Moreover, six out of the seven deleted regions showed a ratio close to 0.5, corresponding to the loss of one copy of the diploid genome. This result suggests that the haploid version of hybrids is mostly conserved compared to the diploid regarding of deleterious structural variants.

Triploid genomes are more subject to gene copy number variation than diploids

Our sequencing data also provided the opportunity to get a more precise view of the variation in terms of copy number of every gene present in the reference genome. By scanning the coverage for each strain using a GC content normalized approach (see Methods), we detected the genes impacted by a variation of the CN. Among the 5,226 annotated genes in the reference

genome, 4,088 genes (78.2%) showed a variation in the number of copies in at least one strain, most of them being duplicated (3,587). Deleted copies were found for 1,734 genes, among which only 100 are totally absent from the genomes. However, gene deletions are more shared across isolates compared to duplications (mean of 4.70 *vs* 2.72, respectively). In addition, CNVs are enriched in subtelomeric regions as previously observed in other yeast species such as *S. cerevisiae*[27] and LTR, mobile elements and tRNA are proportionally more impacted by these variants compared to CDS (Figure S6).

Interestingly, CDS in triploid isolates are more subjected to CN variation (Figure 5.B, p-value = 2.17×10^{-128}), which is consistent with the segmental duplication profiles. While the number of CNVs in diploid strains is 5.6 times lower compared to triploid strains (mean of 169 *vs* 881, respectively), this value is variable among strains and some diploid isolates, such as YJS5456, display a number of CNVs similar to the one observed in triploid strains (Figure 5A, Table S7). Moreover, the ratio between deleted and duplicated CDS is variable among clusters (Table S7) and the G3N1 cluster shows a significant higher number of duplicated CDS compared to the deleted ones (ratio = 2.6, Table S7). This could obviously be linked to the high number of segmental duplications found in the G3N1 cluster.

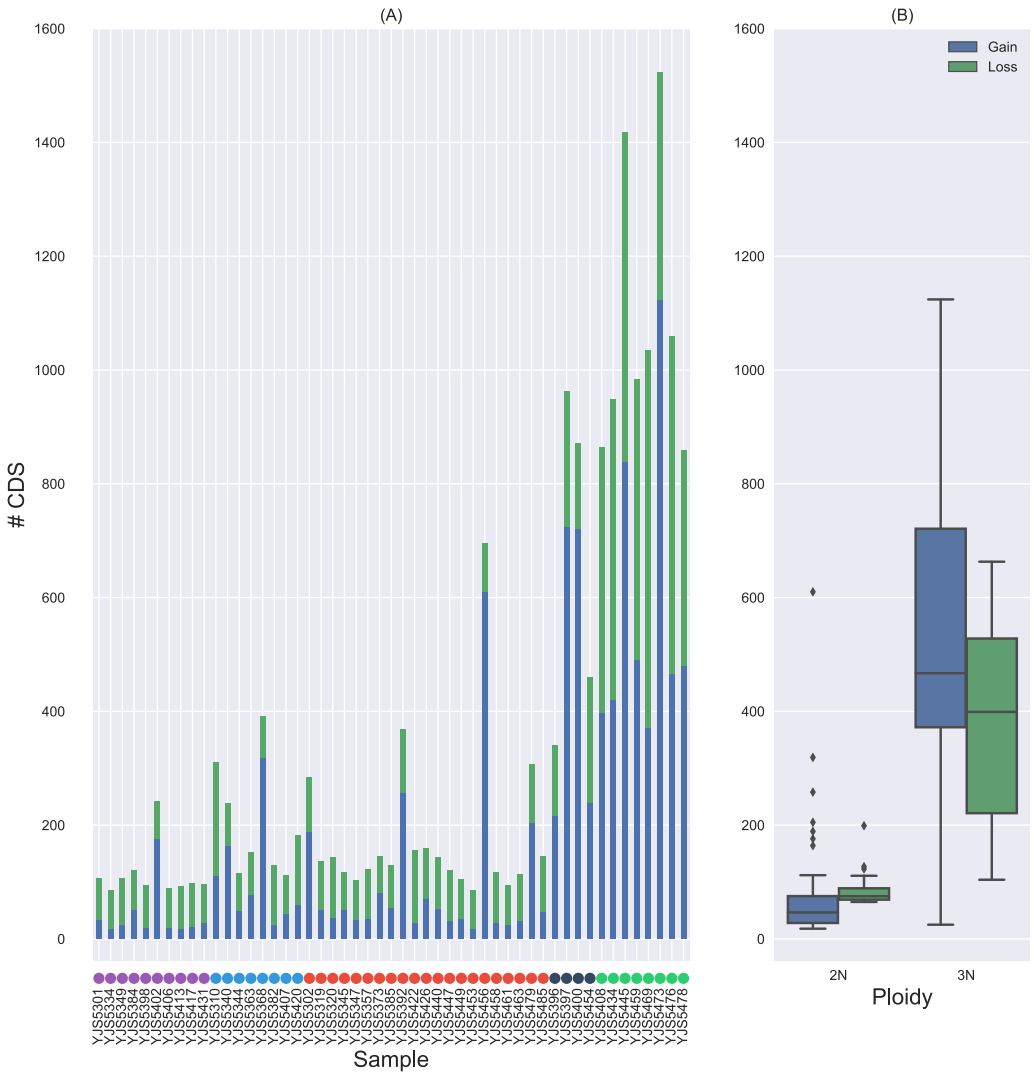


Figure 5. CNVs distribution among the population. (A) Number of CNVs for each sample. (B) Distribution of CNVs by ploidy.

Functional insight into the duplicated and deleted genes

Gene copy variants are known to be a driving mechanism of genomic adaptation to changing environments in yeast[41] and are frequently found to be associated with domesticated processes. For example, in *S. cerevisiae*, duplications of the *CUP* genes have been repeatedly associated with resistance to copper[22,27] and a *MAL* cluster duplication which facilitates the utilization of maltose can be found in beer ale isolates for which this sugar is the main carbon source during the fermentation process[26]. Within *D. bruxellensis,* the investigation of the genome assemblies related to different isolates already highlighted cases of gene content variations between strains. For example, an expansion in the alcohol dehydrogenase family in the wine isolate AWRI1499 has been found to similarly to *S. cerevisiae* enabling greater control over ethanol formation and

consumption[17]. In another study, comparisons of two wine and one beer strains have highlighted 20 genes encoding sugar metabolic processes and nitrogen consumption which were found to be present in the wine strains AWRI1499 and CBS2499 but missing in the beer strain ST05.12/2 2[11].

To determine whether these variations are shared between samples or are independent events, we first looked at the copy number in the whole population using the 20 genes identified in this previous study (Figure 6). Interestingly, all the genes are missing for the strain YJS5392 (G2N3) which has been isolated in Belgium and is closely related to the STO5.12/22 beer isolate. However, this strain is an exception within the whole population and most of the genes are present in the different subpopulations. Interestingly, two genes coding for MFS drug transporters and one gene coding for a high-affinity glucose transporter have more than 4 copies in all triploid strains. Moreover, some strains from the triploid cluster G3N2 related to the wine strain AWRI1499 have multiple copies of genes coding for galactose, glucose and hexose transporter and metabolism while diploid strains have 2 copies and even 1 copy for the G2N1 cluster. Finally, genes involved in nitrogen metabolism are completely lost in 7 diploid strains independently in the 3 different clusters.

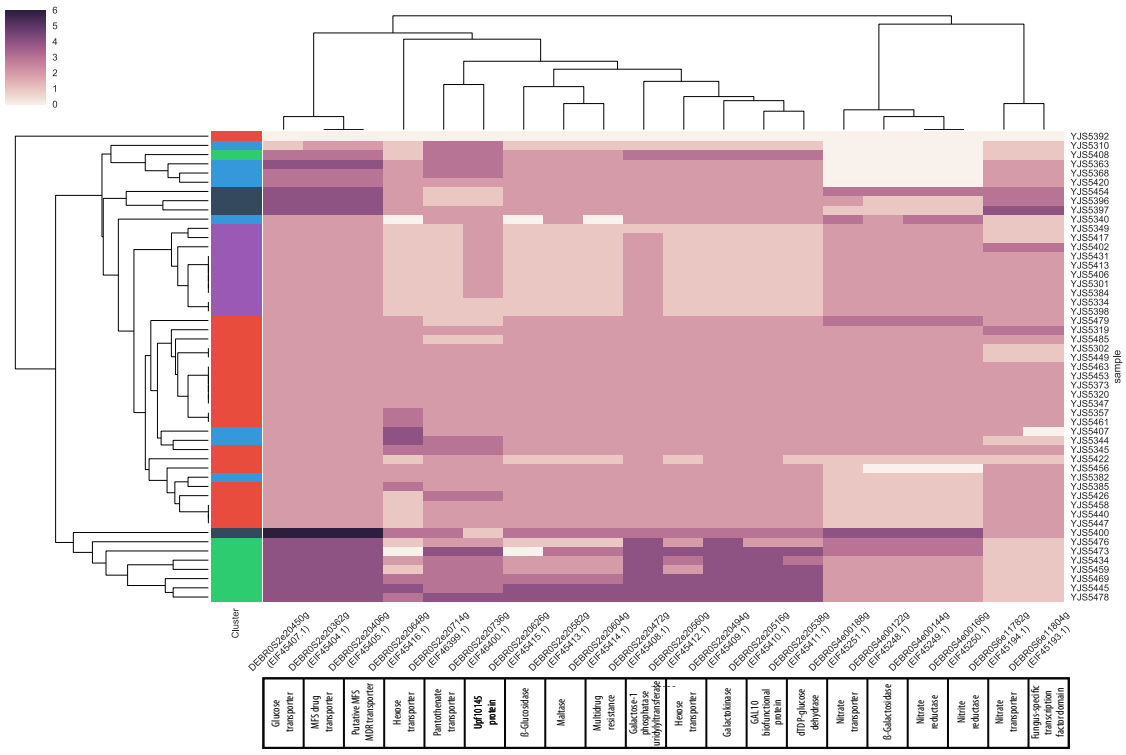


Figure 6. CNVs within the whole population for 20 genes as previously found to be present in the two strains AWRI1499 and CBS2499 but missing in the ST05.12/22a isolate[11].

To determine if other genes are under selective pressure within a whole subpopulation, we examined all duplicated or deleted genes within each cluster for which at least 70% of the strains display the same structural changes (Table S8). Most of them were found in the triploid clusters,

especially for the G3N2 cluster encompassing strains related to the wine strain AWRI1499. Several genes involved in core pathway such as histones H3 and H4 or ribosomal proteins can be found in the G2N3 and G3N2 clusters. Moreover, Go-Term analysis of genes found in the G3N2 cluster revealed enrichment for several biological processes including galactose metabolism (*GAL7*, *GAL3*, p-value = 0.003), nitrogen utilization (*UGA1*, *ATO2*, p-value= 0.004) and transmembrane transport (9 genes, p-value = 0.005).

Chromosomal rearrangement detection using long-read sequencing

With the goal to detect chromosomal rearrangements based on the comparison of highly contiguous assemblies, we applied Oxford Nanopore long-reads sequencing to two diploid (YJS5301 and YJS5345) and one triploid (YJS5476) isolates, representative of the G2N1, G2N3 and G3N2 clusters, respectively. The sequencing yield was very variable across the runs (from 320 Mb to 7 Gb) (Table S6A), and subsets of the longest reads were tested for assemblies, with several assemblers and options. The best assemblies were obtained with SMARTdenovo for the diploid strains, while Flye performed better for the triploid one (Table S6B).

The sizes of the constructed assemblies were in the same range (from 12.89 to 13.39 Mb) but only assemblies related to the diploid strains were almost resolved at the chromosomal-scale level and composed of 7 and 8 scaffolds. Indeed, the assembly of the triploid strain was constituted of 24 scaffolds, which could be linked to the highly heterozygous state of this strain. The pairwise comparison of these assemblies with the reference sequence allowed detecting potential structural variants (Figure S6). *de novo* assemblies revealed 6 and 11 gross rearrangements respectively in the diploid strains YJS5301 and YJS5345. Among them, 3 are related to the same 30 kb genomic regions. The survey of structural rearrangements in the fragmented assembly related to the triploid isolate cannot give an exhaustive view of these events. However, some of the observed rearrangements are shared with the diploid isolates and 5 out of the 7 rearrangements observed are related to genomic regions highlighted in the diploid strain, that could reveal breakpoint hotspots.

***D. bruxellensis* pangenome is small with a few accessory genes**

Finally, to complete our analysis of the gene content within *D. bruxellensis*, we determine the species pangenome, *i.e*. the global set of ORFs (Open Reading Frame) present within the species. To that end, *de novo* assemblies for all isolates have been constructed and scanned to detect non-reference materials (see Methods). A total of 203 (3.7%) supplemental protein-coding genes were highlighted, which lead to 303 variable genes (5,5%) with the previously mentioned fully deleted genes (N = 100). This result clearly shows that the pangenome is much smaller compared to the one defined in *S. cerevisiae*, showing 1,712 accessory genes[27]. In *D. bruxellensis*, 4,923 genes were found in all isolates at least once and consequently were assigned to the core-genome (94,5%). Supplemental ORFs are mostly found in triploid strains, with 173 and 49 of ORFs being specific to the triploid and diploid respectively. A significant

part of these supplemental ORFs can be found for only one isolate (33%), however clustering analysis still revealed groups of genes associated to the different subpopulations (Figure S7). To determine if these genes provide adaptive advantages, we searched for putative functions based on a Blast analysis of the protein sequences. Several transporters have been found in the whole population or within clusters, for example 2 accessory genes having similarities with MFS drug transporters are shared into the triploid wine related cluster (G3N2). However, significant enrichments were not found and multiple genes are linked to transposon or flocculation proteins, which are known to easily degenerate during evolution and could therefore be false positives in our analysis. While current data does not provide the opportunity to know the precise origin of supplemental genes, the high number of them in triploid isolates suggest that a significant part of them may result from the hybridization event. Indeed, it is possible that the formation and selection of hybrids is linked to the contribution of the two genomes, sharing both new alleles and genes into one single individuals, conferring beneficial advantages to specific environments. In this regard, further studies based on haplotype phasing in hybrid genomes can be produced, and will provide valuable insights into the genomic adaptations driven by hybridization events in this species.

Conclusion

With the advent of affordable sequencing technologies, it is now possible to explore and analyze the genome-wide variability within non-model but industrially relevant species. In this study, we provide for the first time a comprehensive description of the genetic variability at the genome-scale in a population of *D. bruxellensis*, giving us a better view of its evolutionary history and the genetic variations underlying the phenotypic diversity within this species. Our results show the presence of at least two hybridization events, which is one of the main factors involved in the division classification of this species into subpopulations. Whether these hybridization events occur spontaneously or not is still unknown, however it is more likely that they are a driving mechanism of *D. bruxellensis* evolution as an adaptive response to the harsh environments found in the domestication processes. Interestingly, similar patterns of genetic variability are observed in both wine and beer triploid subpopulations, suggesting that these strains derived from a single hybridization event that occurred in a common ancestor of these strains. Nevertheless, significant variations between the genomes of the two subpopulations can be found, suggesting the presence of potential genomic adaptations specific to each of them. In addition, our results indicate that loss of heterozygosity is also present in *D. bruxellensis* evolution, impacting the genetic diversity within the species. Moreover, several aneuploidies, segmental duplications and CNVs were also found in the whole population, especially in the triploid strains, indicating that the genome dynamic is important within the species and that triploid hybrids favor these kinds of structural variations. These observations are similar to what has been shown in the hybrid species *S. pastorianus* resulting from the combination of the genomes of *S. cerevisiae* and *S. eubayanus*, for which subpopulations show extensive chromosome loss and LOH events[42]. However, whether the hybridization events in *D. bruxellensis* derived from isolates of the same or two different species remains unknown and both a deeper analysis of triploid isolates genomes and the sequencing of closely related species genome would be needed to investigate this aspect. Although our subset is small, analysis of the allele frequency of aneuploidies in triploid strains suggest that the haploid copy is mainly conserved in the triploid genomes, and therefore confers a significant adaptive advantage in some conditions. At the whole population scale, our data provides the opportunity to get a deeper view of the genetic variants involved in the phenotypic diversity of *D. bruxellensis*. Analysis of CNVs and accessory ORFs in the populations highlights several genes involved in drug and sugar transports as previously found in other analyses. However, these genes are mostly found in the wine related triploid clusters. Interestingly, the nitrogen pathway is independently lost in several diploid isolates within different subpopulations, suggesting that nitrate assimilation is not a common requirement for *D. bruxellensis* isolates. This result is in accordance with phenotypic analysis for which up to a third of the isolates failed to growth on nitrate[14], which could result from the reduction of ethanol and the production of acetic acid during anaerobic fermentation nitrate assimilation[43].

References

1. Schifferdecker, A. J., Dashko, S., Ishchuk, O. P. & Piškur, J. The wine and beer yeast *Dekkera bruxellensis*. *Yeast* **31,** 323–332 (2014).
2. Thomson, J. M. *et al.* Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat. Genet.* **37,** 630–635 (2005).
3. Heresztyn, T. Metabolism of volatile phenolic compounds from hydroxycinnamic acids byBrettanomyces yeast. *Arch. Microbiol.* **146,** 96–98 (1986).
4. Beckner, M., Ivey, M. L. & Phister, T. G. Microbial contamination of fuel ethanol fermentations. *Lett. Appl. Microbiol.* **53,** 387–394 (2011).
5. Teoh, A. L., Heard, G. & Cox, J. Yeast ecology of Kombucha fermentation. *Int. J. Food Microbiol.* **95,** 119–126 (2004).
6. Borneman, A. R., Zeppel, R., Chambers, P. J. & Curtin, C. D. Insights into the *Dekkera bruxellensis* genomic landscape: comparative genomics reveals variations in ploidy and nutrient utilisation potential amongst wine isolates. *PLoS Genet.* **10,** e1004161 (2014).
7. Crauwels, S. *et al.* Comparative phenomics and targeted use of genomics reveals variation in carbon and nitrogen assimilation among different *Brettanomyces bruxellensis* strains. *Appl. Microbiol. Biotechnol.* **99,** 9123–9134 (2015).
8. Crauwels, S. *et al.* Fermentation assays reveal differences in sugar and (off-) flavor metabolism across different *Brettanomyces bruxellensis* strains. *FEMS Yeast Res.* **17,** fow105 (2017).
9. Hellborg, L. & Piskur, J. Complex nature of the genome in a wine spoilage yeast, *Dekkera bruxellensis*. *Eukaryot. Cell* **8,** 1739–49 (2009).
10. Albertin, W. *et al.* Development of microsatellite markers for the rapid and reliable genotyping of *Brettanomyces bruxellensis* at strain level. *Food Microbiol.* **42,** 188–195 (2014).
11. Crauwels, S. *et al.* Assessing genetic diversity among *Brettanomyces* yeasts by DNA fingerprinting and whole-genome sequencing. *Appl. Environ. Microbiol.* **80,** 4398–413 (2014).
12. Avramova, M. *et al. Brettanomyces bruxellensis* population survey reveals a diploid-triploid complex structured according to substrate of isolation and geographical distribution. *Sci. Rep.* **8,** 4136 (2018).
13. Agnolucci, M. *et al.* Genetic diversity and physiological traits of *Brettanomyces bruxellensis* strains isolated from Tuscan Sangiovese wines. *Int. J. Food Microbiol.* **130,** 238–244 (2009).
14. Conterno, L., Joseph, C. M. L., Arvik, T. J., Henick-kling, T. & Bisson, L. F. Genetic and physiological characterization of *Brettanomyces bruxellensis* strains isolated from wines. *Am. J. Enol. Vitic.* **57,** 139–147 (2006).
15. Curtin, C. D., Bellon, J. R., Henschke, P. A., Godden, P. W. & de Barros Lopes, M. A. Genetic diversity of *Dekkera bruxellensis* yeasts isolated from australian wineries. *FEMS Yeast Res.* **7,** 471–481 (2007).
16. Vigentini, I. *et al.* Intraspecific variations of *Dekkera/Brettanomyces bruxellensis* genome studied by capillary electrophoresis separation of the intron splice site profiles. *Int. J. Food Microbiol.* **157,** 6–15 (2012).
17. Curtin, C. D., Borneman, A. R., Chambers, P. J. & Pretorius, I. S. De-novo assembly and analysis of the heterozygous triploid genome of the wine spoilage yeast *Dekkera bruxellensis* AWRI1499. *PLoS One* **7,** e33840 (2012).
18. Piškur, J. *et al.* The genome of wine yeast *Dekkera bruxellensis* provides a tool to explore its food-related properties. *Int. J. Food Microbiol.* **157,** 202–9 (2012).
19. Olsen, R.-A. *et al.* De novo assembly of *Dekkera bruxellensis*: a multi technology approach using short and long-read sequencing and optical mapping. *Gigascience* **4,** 56 (2015).
20. Skelly, D. A. *et al.* Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* **23,** 1496–504 (2013).
21. Bergström, A. *et al.* A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31,** 872–888 (2014).
22. Strope, P. K. *et al.* The 100-genomes strains, an *S. cerevisiae* resource that illuminates

its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* gr.185538.114- (2015). doi:10.1101/gr.185538.114
23. Almeida, P. *et al.* A population genomics insight into the mediterranean origins of wine yeast domestication. *Mol. Ecol.* (2015). doi:10.1111/mec.13341
24. Zhu, Y. O., Sherlock, G. & Petrov, D. A. Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3 (Bethesda)*. (2016). doi:10.1534/g3.116.029397
25. Gonçalves, M. *et al.* Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Curr. Biol.* **26,** 2750–2761 (2016).
26. Gallone, B. *et al.* Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* **166,** 1397–1410.e16 (2016).
27. Peter, J. *et al.* Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556,** 339–344 (2018).
28. Fawcett, J. A. *et al.* Population genomics of the fission yeast *Schizosaccharomyces pombe*. *PLoS One* **9,** e104241 (2014).
29. Jeffares, D. C. *et al.* The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. *Nat. Genet.* **advance on,** (2015).
30. Leducq, J.-B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat. Microbiol.* **1,** 15003 (2016).
31. Carreté, L. *et al.* Patterns of Genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. *Curr. Biol.* **28,** 15–27.e7 (2018).
32. Hirakawa, M. P. *et al.* Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Res.* gr.174623.114- (2015). doi:10.1101/gr.174623.114
33. Ford, C. B. *et al.* The evolution of drug resistance in clinical isolates of *Candida albicans*. *Elife* **4,** e00662 (2015).
34. Ropars, J. *et al.* Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. *Nat. Commun.* **9,** 2253 (2018).
35. Friedrich, A., Jung, P., Reisser, C., Fischer, G. & Schacherer, J. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol. Biol. Evol.* **32,** 184–92 (2015).
36. Ortiz-Merino, R. A. *et al.* Ploidy variation in *Kluyveromyces marxianus* separates dairy and non-dairy isolates. *Front. Genet.* **9,** 94 (2018).
37. Fournier, T. *et al.* High-Quality de NovoGenome Assembly of the *Dekkera bruxellensis* Yeast Using Nanopore MinION Sequencing. *G3 (Bethesda)*. **7,** 3243–3250 (2017).
38. Pryszcz, L. P. & Gabaldón, T. Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44,** e113 (2016).
39. Lynch, D. B., Logue, M. E., Butler, G. & Wolfe, K. H. Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. *Genome Biol. Evol.* **2,** 572–83 (2010).
40. Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131,** 452–462 (2007).
41. Kondrashov, F. A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings. Biol. Sci.* **279,** 5048–57 (2012).
42. Okuno, M. *et al.* Next-generation sequencing analysis of lager brewing yeast strains reveals the evolutionary history of interspecies hybridization. *DNA Res.* **23,** 67–80 (2016).
43. Galafassi, S., Capusoni, C., Moktaduzzaman, M. & Compagno, C. Utilization of nitrate abolishes the "Custers effect" in *Dekkera bruxellensis* and determines a different pattern of fermentation products. *J. Ind. Microbiol. Biotechnol.* **40,** 297–303 (2013).

Supplementary materials

Supplementary tables

Supplementary tables can be found at https://bit.ly/2MxaVSh

Supplementary figures

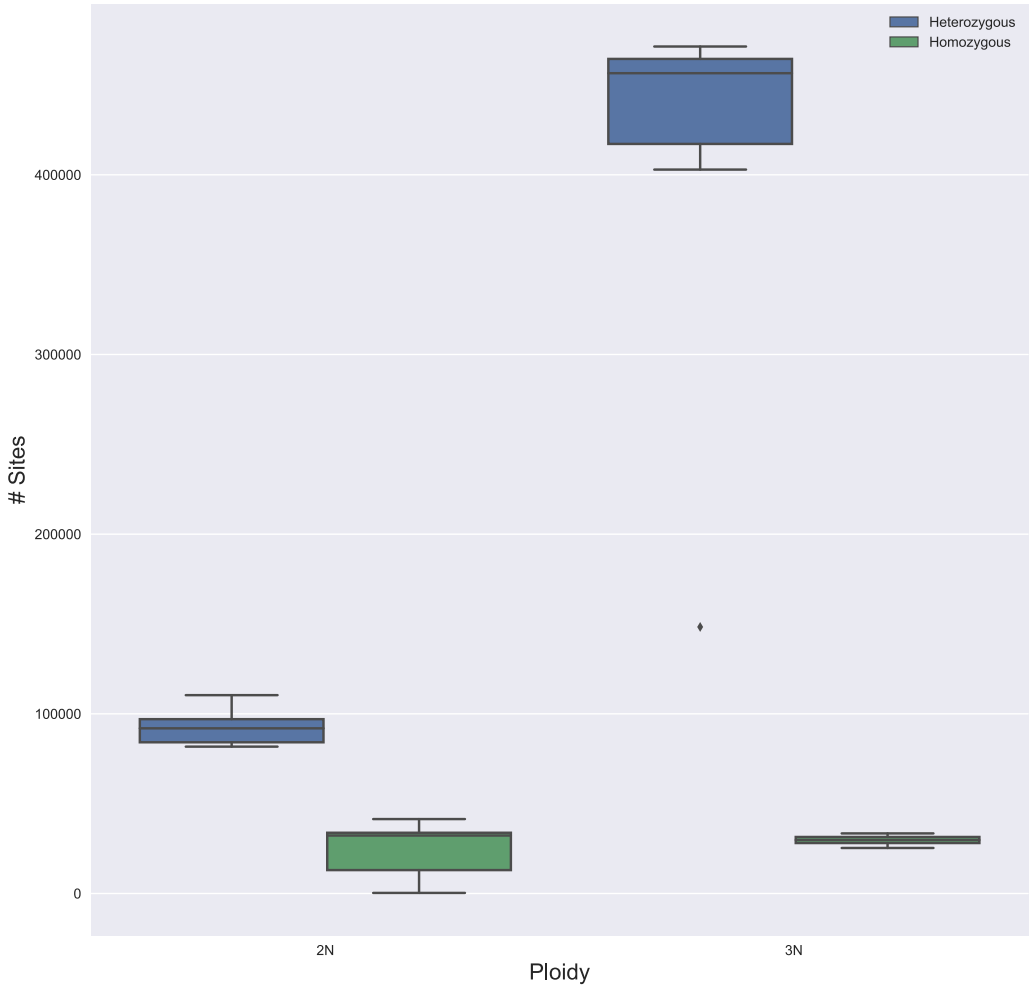


Figure S1. Distribution of the number of heterozygous sites for each strain by ploidy.

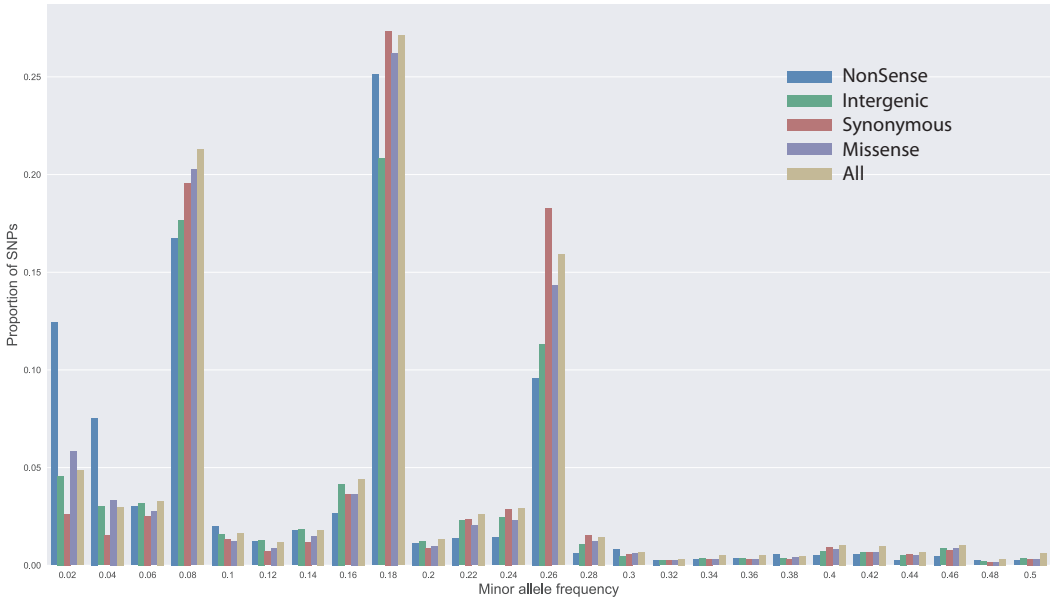


Figure S2. Minor allele frequency of single nucleic polymorphism (SNP) with different genomic locations and functional annotation.

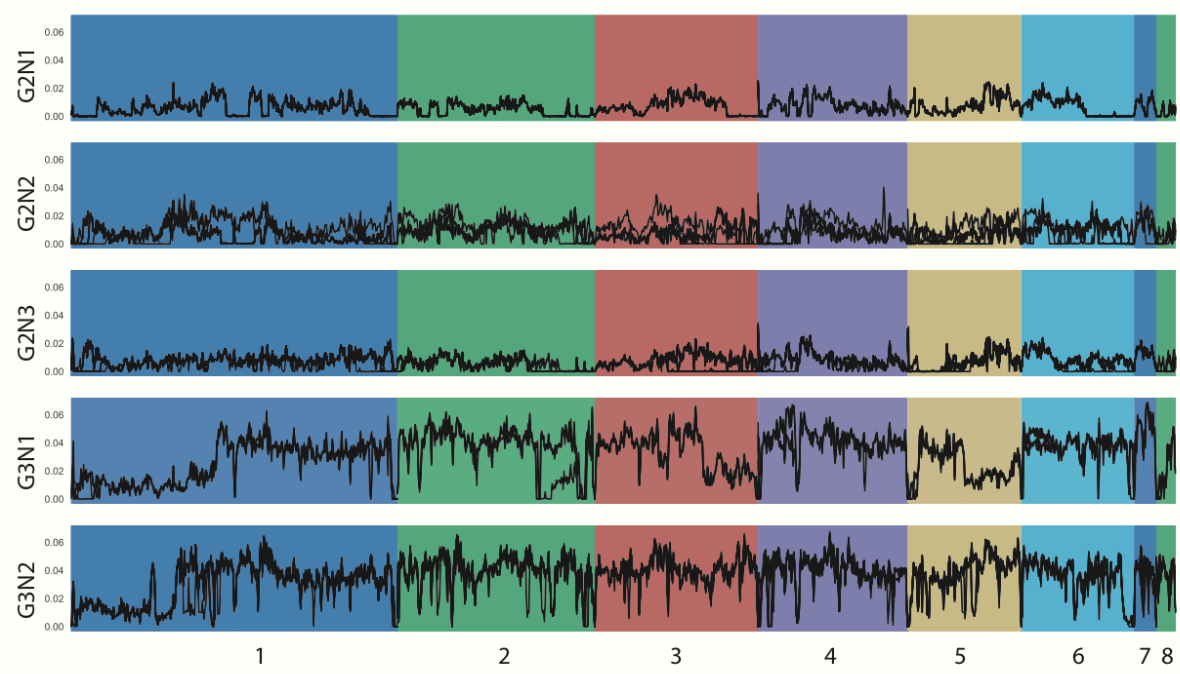


Figure S3. Heterozygosity level along the genome for each cluster using sliding window of 10kb. For each window, the number of heterozygous SNPs has been divided by the window size.

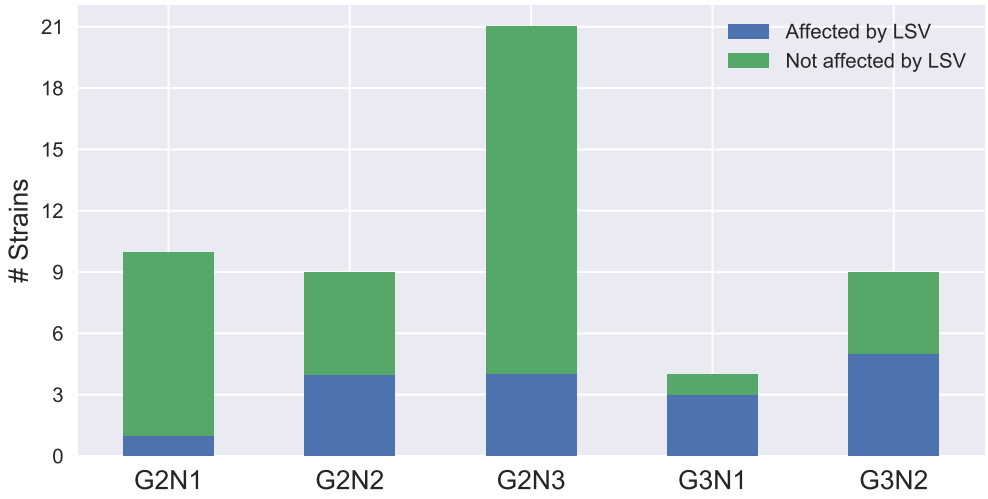


Figure S4. Number of strains affected by aneuploidies or segmental duplications by cluster.

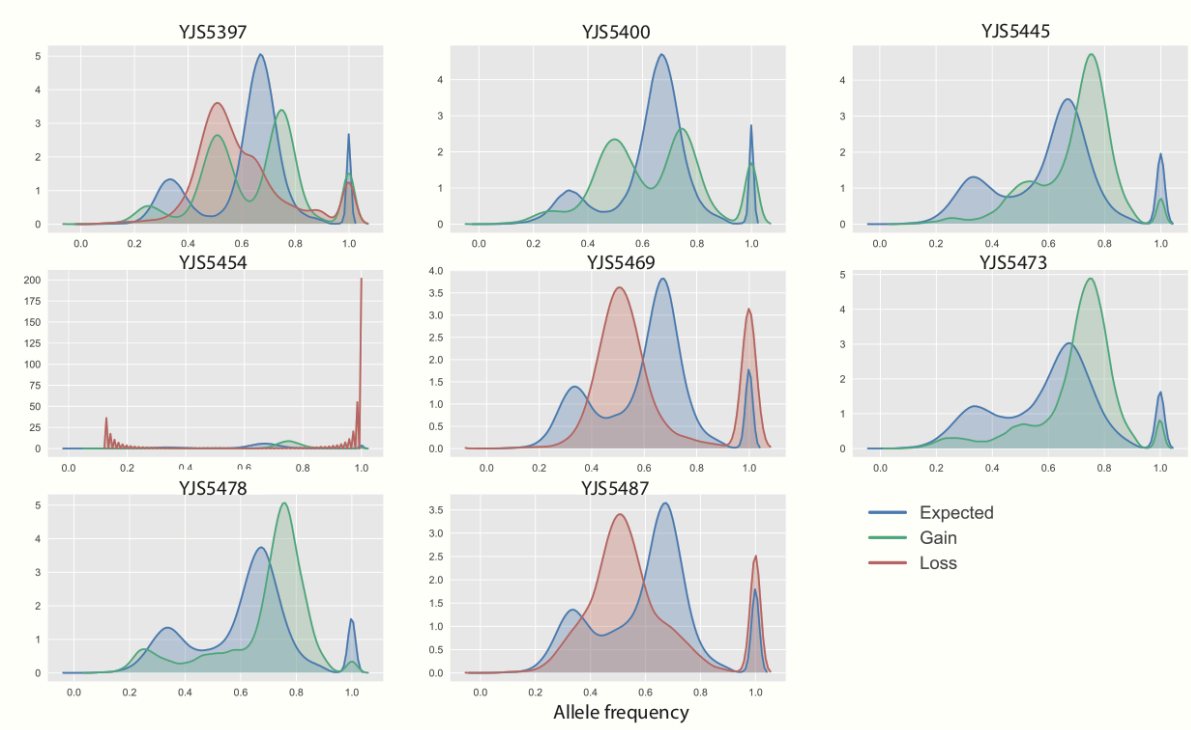


Figure S5. Allele frequency distributions of heterozygous SNPs in aneuploidy regions of triploid strains

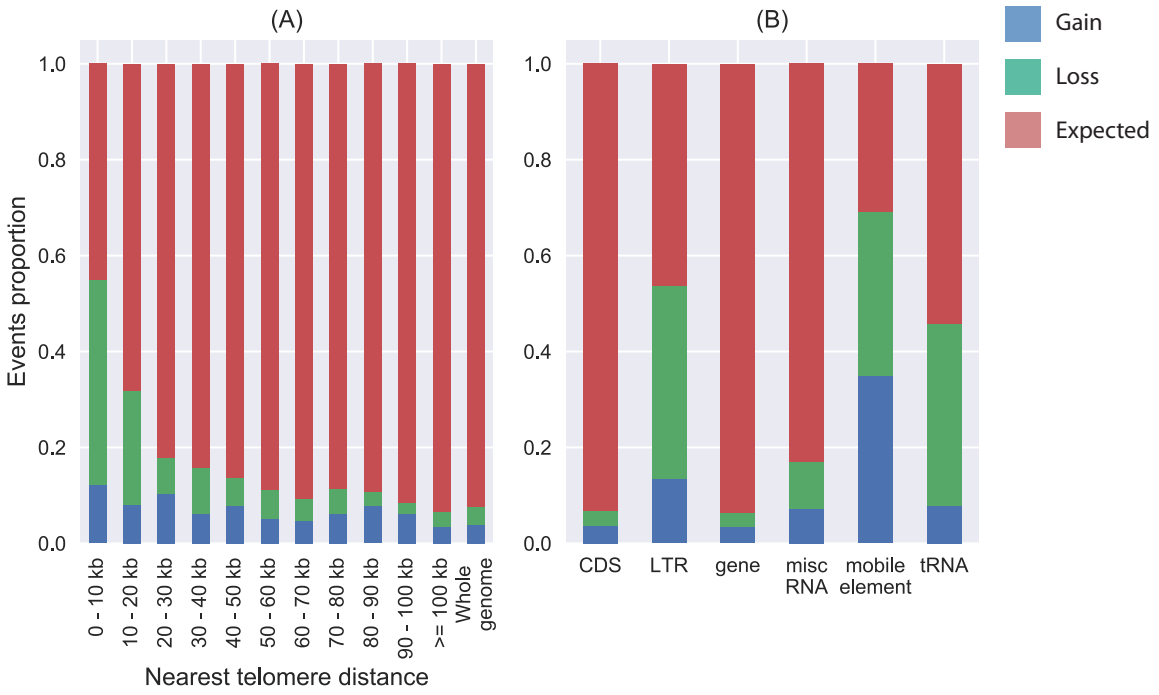


Figure S6. CNV distribution in the genome. (A) Fraction of CNVs affecting all genetic elements within the whole population based on their distance from the nearest telomere. (B) Proportion of features affected within the whole population for each feature type.

MATERIELS ET METHODES

Données de séquençage

Collections pour les études de génomique de population

Dans le cadre du chapitre 1, des collections d'isolats pour *K. lactis* et *L. thermotolerans* ont été spécifiquement constituées au laboratoire. A l'inverse, les isolats de *C. albicans, L. kluyveri, S. paradoxus et S. uvarum* proviennent de différentes collections constituées dans le cadre de précédentes études de génomique de population[1–4]. Le nombre de souches par collection varie, allant de 24 isolats pour *S. paradoxus* à 57 pour *L. thermotolerans* (Tableau 1).

Espèce	Nombre d'isolats	Nom de la souche de référence	Nombre de chromosomes	Taille du génome (Mb)
C. albicans	20	SC 5314	8	14,3
D. bruxellensis	53	UMY321	8	13,0
K. lactis	41	CBS 2359	6	10,7
L. kluyveri	28	CBS 3082	8	11,3
L. thermotolerans	57	CBS 6340	8	10,4
S. paradoxus	24	CBS 432	16	11,7
S. uvarum	34	CBS 7001	16	11,5

Tableau 1. Caractéristiques principales des collections et des espèces étudiées

Pour les collections de *K. lactis* et *L. thermotolerans*, un panel de souches représentatif des origines écologiques et géographiques chez ces espèces a été constitué. Dans ce cadre, les isolats de *K. lactis* proviennent d'environnements naturels (exsudats de chêne ou d'équipements de vin) ou dans la production laitière (fromage, crème de lait, petit-lait) récupérés à travers le monde. Les souches de *L. thermotolerans* ont été isolées dans un ensemble d'environnements naturels incluant des exsudats d'arbre, des grappes de raisin, d'insectes ou de feuilles de riz. A l'exception de *S. paradoxus* pour laquelle les souches proviennent d'Amérique du Nord, les souches des autres espèces ont été isolées à travers le monde (Figure 1). Pour le chapitre 3, une collection de *D. bruxellensis* a été composée à partir de 56 isolats. Les souches proviennent d'origines géographiques variées (Europe, Afrique du Sud, Australie, Chile) et une large partie des isolats est associée aux processus de vinification.

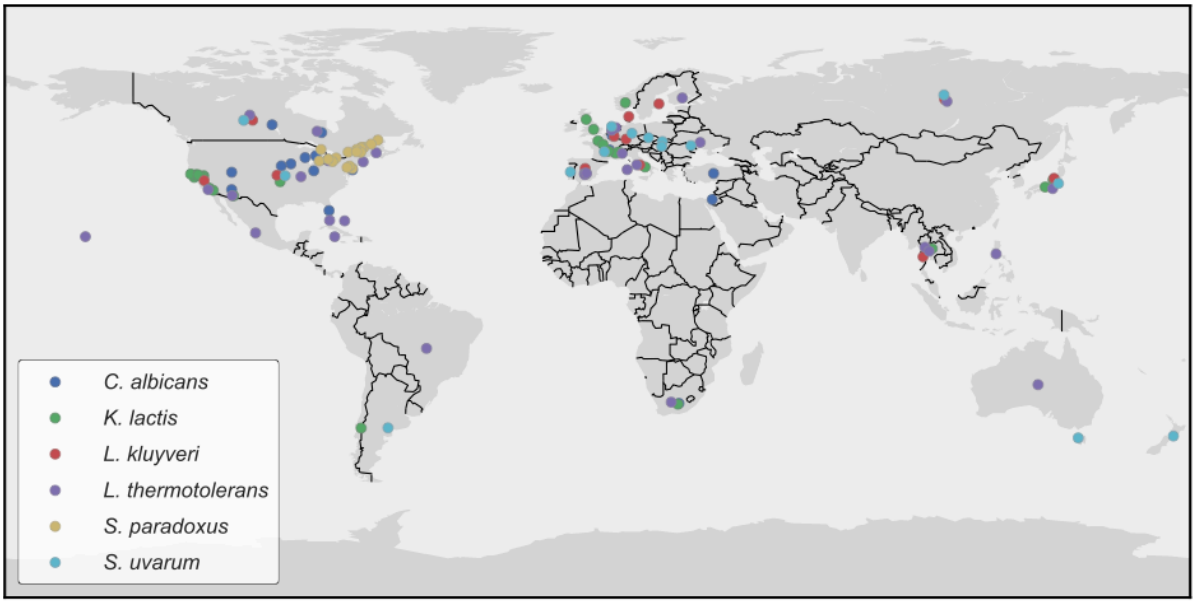


Figure 1. Origines géographiques des souches utilisées dans le cadre du chapitre 1.

Analyse de la ploïdie naturelle par cytométrie

La détermination de la ploïdie a été réalisée par une mesure cellule par cellule de l'intensité de fluorescence émise suite à un marquage de l'ADN à l'iodure de propidium, un agent intercalant de l'ADN fluorescent. Plus précisément, les cellules ont été mises en culture sur milieu YPG à 30°C afin d'obtenir une phase exponentielle de croissance. Les cellules ont par la suite été nettoyées avec 1 ml d'eau et le culot a été de nouveau suspendu dans 1 ml d'éthanol à 70 %. Après centrifugation, le surnageant a été enlevé et les cellules ont été encore suspendues dans 1 ml de tampon de citrate de sodium (trisodium citrate 50 mM ; pH 7,5). Après rajout de 10 µl de RNase A (100 mg/ml), les souches ont été incubées à 37°C pendant 2 heures. Les souches ont par la suite été soniquées pendant 20 secondes avec une amplitude de 20 %. Après sonication, les cellules ont été plongées dans 1 ml de citrate de sodium supplémenté de 10 µl d'iodure de propidium (1,6 mg/ml) et laissées dans le noir à 4°C pendant 12 heure. Une fois terminé, le contenu de l'ADN a été déterminé en mesurant l'intensité de la fluorescence grâce à une analyse de cytométrie en flux (CyFlow Space ; Partec).

Séquençage Illumina

Le séquençage des souches de *K. lactis, L. thermotolerans* (chapitre 1) et *D. bruxellensis* (chapitre 2 et 3) a été réalisé avec la technologie Illumina HiSeq 2000 au BGI (Beijing Genomics Institute), basée sur le séquençage par synthèse. Dans cette approche, le génome d'intérêt est fragmenté de manière aléatoire. Chaque fragment, auquel des adaptateurs ont été fusionnés avant amplification, est ensuite mis en présence de 4 types de nucléotide associé à un fluorophore et à un terminateur réversible. Le nucléotide à incorporer pour l'élongation du brin complémentaire est alors insèré. Un signal lumineux spécifique du nucléotide présent est

ensuite émis et détecté. En associant une séquence d'index aux échantillons, il est possible de séquencer simultanément plusieurs échantillons. Cette pratique, appelée multiplexage, permet de réduire les coûts en divisant le nombre d'expériences de séquençage.

Les souches ont préalablement été séparément mises en culture dans 20 ml de milieu YPG liquide et incubées à 30°C sous agitation pendant une nuit. Les cellules ont par la suite été centrifugées et l'ADN génomique a été extrait grâce au kit de purification MasterPure Yeast DNA (Cat No MPY80200). Pour l'ensemble des génomes, des librairies d'insertion de 280-bp ont été produites et les deux extrémités de chaque fragment ont systématiquement été séquencées, produisant ainsi des lectures dites pairées, de 96 à 102 pb.

Séquençage MinION de la souche UMY321

Pour le chapitre 2,2 µg d'ADN génomique a été coupé en fragment de 8000 bp dans des g-TUBE. Les fragments ont par la suite été nettoyés en utilisant des billes 1X AMPure XP et une librairie Nanopore à deux dimension (2D) de 8 kb a été préparée en suivant le protocole SQK-MAP005- MinION gDNA Sequencing Kit.

Le mix de séquençage a été préparé en utilisant 8 µl de la librairie d'ADN génomique, de l'eau, un Fuel mix et un running buffer, en accord avec le protocole SQK-MAP005. Ce mix a été ajouté dans une flowcell R7.3 pendant 48 heures. La flowcell a été rechargée une fois toutes les 24 heures par l'addition de 8 µl de la librairie d'ADN.

Séquençage MinION dans l'analyse de variants structuraux

Pour le chapitre 3, 5 µg d'ADN génomique de trois souches a été coupé en fragments de 20 kb dans des g-UBE. Les fragments ont par la suite été nettoyés en utilisant des billes 1X AMPure XP et une librairie Nanopore à une dimension (1D) a été préparée en suivant le protocole SQK-LSK108.

Le mix de séquençage a été préparé en utilisant 12 µl de la librairie d'ADN génomique, de l'eau, un Fuel mix et un running buffer, en accord avec le protocole SQK-LSK108. Pour les souches YJS5301 et YJS5416, une flowcell R9 a été utilisée, alors qu'une flowcell R9.4 a été utilisée pour la souche YJ5345. Dans les deux cas, les séquençages ont duré 48 heures.

Récupération des données externes

Pour le chapitre 1, les données de séquençage relatives aux génomes des isolats de *C. albicans, L. kluyveri, S. paradoxus* et *S. uvarum* proviennent du domaine public et ont été téléchargées à partir de la banque de données SRA qui stocke les données brutes issues de séquençage de nouvelle génération à partir d'un identifiant spécifique à chaque projet :

- *C. albicans* : PRJNA193498
- *L. kluyveri* : PRJEB5130
- *S. paradoxus* : PRJNA277692
- *S. uvarum* : PRJNA230139 et PRJEB5133.

La séquence et les annotations des génomes de référence pour les différentes espèces du chapitre 1 ont été produites dans le cadre de précédentes études et sont disponibles publiquement sur les sites web suivants :

- IGénolevure (http://gryc.inra.fr) pour *K. lactis*[5], *L. kluyveri*[6] et *L. thermotolerans*[6]
- SSS (http://sss.genetics.wisc.edu/cgi-bin/s3.cgi) pour *S. paradoxus*[7] et *S. uvarum*[7]
- Candida Genome Database (http://www.candidagenome.org) pour *C. albicans*[8]

Traitement des données de séquençage

Nettoyage des lectures

Pour le chapitre 1, un prétraitement de nos lectures brutes a été réalisé. Ce dernier consiste à supprimer les nucléotides de mauvaise qualité à la fin des lectures puis d'éliminer les lectures devenues trop courtes, après avoir préalablement retiré les lectures contenant des amorces de PCR séquencées par erreur. Les lectures ont tout d'abord été analysées grâce au logiciel FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), nous permettant de récupérer les séquences des amorces PCR et de vérifier la qualité des résidus. Le nettoyage des lectures a par la suite été réalisé avec le logiciel cutAdapt[9] (v. 1.9) : les séquences présentant un taux de similarité supérieur à 90% avec les amorces ont été supprimées et les résidus de mauvaise qualité en fin de lectures ont été éliminés. Seules les lectures d'une taille supérieure ou égale à 50 pb ont été conservées pour la suite des analyses

Alignement contre le génome de référence

Pour le chapitre 1 et 3, le même pipeline a été utilisé afin d'aligner les lectures contre les génomes de référence. Pour chaque souche, les jeux de lectures ont tout d'abord été alignés contre le génome de référence grâce à BWA mem (v. 0.7.12) avec les paramètres par défaut. Les alignements ont par la suite été nettoyés en utilisation successivement les fonctions fixmate de samtools[10] (v. 1.1), realignement de GATK[11] (v. 3.3.0) et MarkDuplicates de Picard (http://broadinstitute.github.io/picard) (v. 1.140). Pour chaque souche, les alignements produits ont par la suite été concaténés grâce à la fonction merge de samtools et les lectures ont été de nouveau réalignées en utilisant la fonction realign du package GATK.

Étude de la couverture : Aneuploïdies et duplications segmentales

Afin d'examiner la présence d'aneuploïdies ou de duplications segmentales pour chaque souche, un scan global de la couverture a été réalisé. La couverture de chaque position du génome de référence a été obtenue grâce à la fonction depth de samtools. Les aneuploïdies ou de duplications segmentale ont par la suite été systématiquement déterminées par l'analyse manuelle de la distribution de la couverture à travers des fenêtres non chevauchantes de 10kb le long du génome (Figure 2).

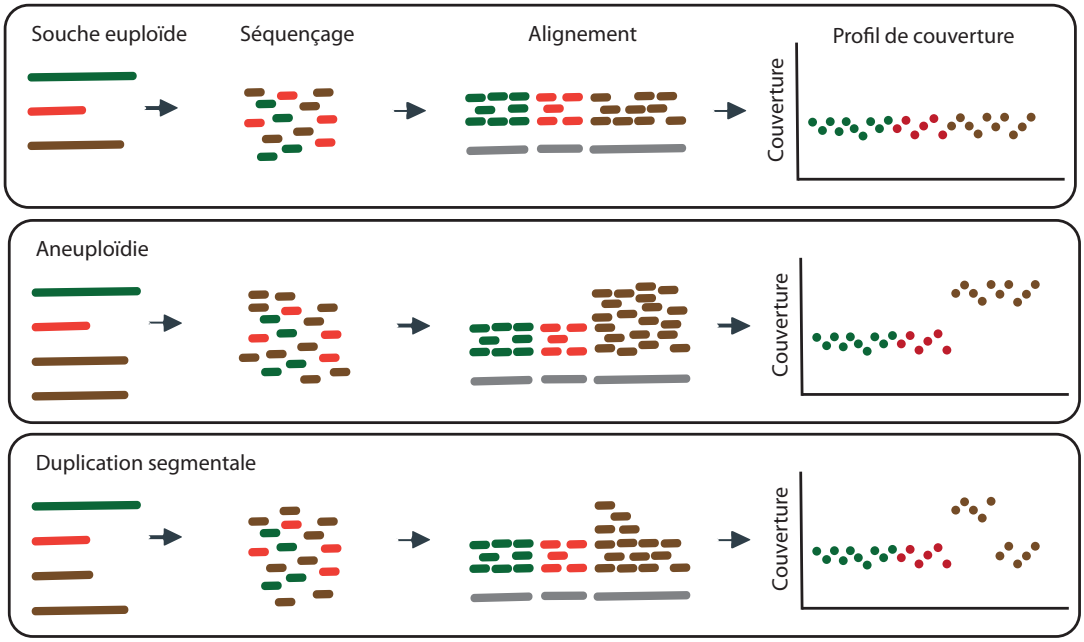


Figure 2. Profils théoriques de couverture obtenus dans l'identification d'aneuploïdies et de duplications segmentales.

Étude de la couverture : Nombre variable de copies

La détermination du nombre de copie de chaque gène repose tout d'abord sur l'analyse des fichiers bam par le logiciel Control-FREEC[12] (v. 10.6). Cette approche, basée sur une normalisation du contenu en GC et sur la distribution de la couverture le long de chaque contig permet l'identification précise des coordonnées de régions génomiques présentant significativement une couverture différente en comparaison du reste du génome. Dans ce cadre, les paramètres suivants ont été utilisés : « `breakPointThreshold = 0.6, window = 1000, telocentromeric = 6000` et `step = 200` ». Le pourcentage en GC indiqué pour chaque espèce a été déterminé en fonction de la composition en GC des génomes de référence de chaque espèce (Tableau 2).

Espèces	Ploïdie	minExpectedGC	maxExpectedGC
C. albicans	2N	0.25	0.40
D. bruxellensis	2N / 3N	0.30	0.50
K. lactis	1N	0.35	0.45
L. kluyveri	1N	0.35	0.65
L. thermotolerans	1N	0.40	0.60
S. paradoxus	1N	0.35	0.45
S. uvarum	1N	0.35	0.50

Tableau 2. Paramètres spécifiques à chaque espèce, utilisés dans l'utilisation de Control-FREEC.

Pour *K. lactis*, *L. kluyveri*, *L. thermotolerans*, *S. paradoxus* et *S. uvarum* une ploïdie de 1N a été renseignée. Pour *C. albicans* une ploïdie de 2N a été donnée. Pour *D. bruxellensis*, la ploïdie spécifique à chaque souche (2N ou 3N) a été renseignée à partir de l'estimation *in silico* de la ploïdie des souches (voir section *Détermination in silico de la ploïdie des souches*). Pour cette espèce, les résultats des souches YJS5416 et YJS5487 ont été ignorées dans l'analyse dû au nombre important de duplications segmentales dans ces souches, induisant de nombreux faux-positifs.

Pour chaque espèce, les régions obtenues ont par la suite été comparées aux fichiers d'annotations associées. Un nombre de copies variable a été assigné à chaque élément fonctionnel si au moins 50% de sa séquence est recouverte par une des régions déterminées précédemment. Si deux ou d'avantage de régions ont été retrouvées pour le même élément fonctionnel, la région présentant le plus grand recouvrement a été sélectionnée.

Assemblage des lectures longues et détection des réarrangements chromosomiques

Les séquences d'adaptateurs ont été supprimées des reads MinION avec PoreChop (v0.2.3) (https://github.com/rrwick/Porechop). Les 25X plus longs reads de YJS5301 et les 30X plus longs reads de YJS5345 ont été assemblés par SMARTdenovo (https://github.com/ruanjue/smartdenovo) avec les paramètres « -c 1 -e zmo -k 14 -J 2500 ». Les 40X plus longs reads de YJS5476 ont été assemblés par Flye (v2.3.5b) (https://github.com/fenderglass/Flye) avec les paramètres « `-m 1000 --genome-size 13m` ».

Ces assemblages ont ensuite été alignés contre le génome de référence de *D. bruxellensis* par NUCmer[13] avec les paramètres « `--maxmatch -c 5000` » et les dotplots ont été générés par mummerplot. Les variants structuraux ont ensuite été identifiés visuellement et leurs points de rupture ont été déterminés par show-coords avec les paramètres « `-r -c -l` ». Toutes les informations génomiques dans les régions à ± 5 kb autour de ces points de rupture ont été extraites de l'annotation de la référence.

Détection des polymorphismes nucléotidiques

Les fichiers d'alignement bam produits lors de l'alignement des lectures contre les génomes de références constituent les fichiers d'entrée utilisés afin d'extraire le polymorphisme nucléotidique (SNP, insertions et délétions de petite taille). Pour ce faire, la fonction HaplotypeCaller de GATK a été utilisée en précisant une ploïdie de 1N pour la majorité des espèces à l'exception de *C. albicans* et *D. bruxellensis* pour lesquelles une ploïdie de 2N a été donnée. Pour *D. bruxellensis*, les variants ont par la suite été annotés grâce à la fonction VariantAnnotator de GATK, permettant notamment l'obtention de la fréquence allélique de chacun d'entre eux. La fonction filter de bcftools[10] (v. 1.5) a par la suite été utilisée afin que seuls les variants présentant une couverture supérieure ou égale à 10X et un score de qualité supérieur à 25 soient retenus. Finalement, les fichiers obtenus pour chaque souche ont été

concaténés par la fonction merge de vcftools[14] (v. 0.1.13) dans un seul fichier VCF (Variant Call Format) où l'ensemble du polymorphisme nucléotidique de l'espèce peut être retrouvé.

Annotation du polymorphisme nucléotidique

L'impact fonctionnel des variants nucléotidiques sur les génomes a été déterminé par SNPEff variant annotator[15] (v 4.3i). Dans ce cadre, les fichiers d'annotations de chaque espèce ont été soumis. Pour le chapitre 3, les pseudo-gènes de *D. bruxellensis* ont été ignorés dans l'analyse de l'annotation des SNP (N = 555). Afin de prédire l'effet fonctionnel des mutations non-synonymes, les différents changements d'acide aminé présents dans la référence ont été analysés par SIFT[16] (v 6.2.1). Ce programme se base sur les propriétés physicochimiques des acides aminés et de la conservation de ceux-ci dans la séquence protéique. Pour chaque gène de chaque espèce, les différentes substitutions d'acide aminé ont été données avec un fichier d'alignement protéique produit spécifiquement dans le cadre de cette étude à partir de la comparaison des séquences codantes contre une banque de données protéique réalisée au sein du laboratoire et contenant les protéomes de plusieurs espèces de levures et de champignons.

Librairie informatique

Les analyses de génomique des populations dépendent notamment de la manipulation d'un sous-set ou de l'ensemble des variants nucléotidiques au sein des espèces. D'un point de vue technique, cette manipulation pose un certain nombre de contraintes. Premièrement, l'importante quantité de données générée nécessite un traitement rapide et efficace des informations. Deuxièmement, les variants nucléotidiques sont utilisés dans différentes approches informatiques, demandant ainsi un traitement malléable des données brutes.

Afin de faciliter la comparaison et l'analyse des données de l'ensemble des génomes séquencés pour chaque espèce, j'ai développé une librairie informatique dédiée à l'extraction et l'analyse du polymorphisme nucléotidique. Celle-ci possède une interface avec le langage de programmation python, utilisé fréquemment dans le milieu de la bio-informatique et présentant plusieurs avantages : c'est un langage multiplateforme qui supporte à la fois la programmation orientée objet et la programmation fonctionnelle. Ceci permet le traitement rapide de problèmes complexes et la conception de scripts courts ou plus complexes en fonction des tâches à réaliser.

Le cœur de la librairie repose sur le module pysam (http://pysam.readthedocs.io) et les différentes fonctions associées ont été codées à partir du langage Cython, un langage de programmation compilé à plus bas niveau permettant un traitement rapide et efficace des données. Cette couche est entièrement transparente pour l'utilisateur et permet ainsi un gain important de performance sans affecter les utilisations sous-jacentes. Par ailleurs, ce module repose lui-même sur l'API C de la librairie htslib proposée par le package samtools[10] offrant à la fois l'accès et l'analyse d'une région précise du génome mais aussi de l'ensemble des variants, rendant ainsi la libraire modulable et extensible pour différentes approches informatiques.

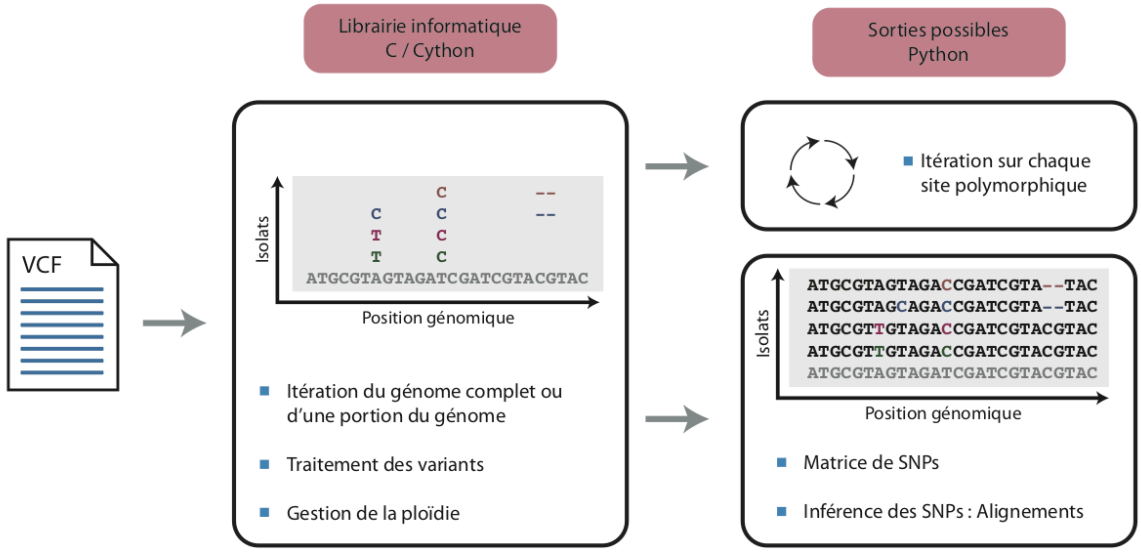


Figure 3. Représentation schématique du traitement des données par la librairie

Afin de répondre aux différents besoins, plusieurs sorties sont disponibles sous la forme d'objet Python permettant le traitement rapide des informations (Figure 3). Dans un premier temps, la librairie offre la possibilité d'itérer sur chaque site polymorphique. Des fonctions associées à ces sites, telle que le décompte de la fréquence allélique, facilitent l'exploration des données. Dans un second temps, l'ensemble des SNP peuvent être inférés au sein d'un alignement contenant ou non l'ensemble des positions génomiques du génome. Cette fonction permet l'obtention d'un objet MultiSeqAlignment de la librairie biopython (https://biopython.org) possédant un ensemble de fonctions permettant l'encodage des données sous différents formats. Cette possibilité offre l'opportunité de produire rapidement des fichiers d'entrée pour différents logiciels tiers dédiés à la génomique des populations. De manière additionnelle, l'itération interne des variants peut être modifiée, par exemple l'ensemble des sites polymorphiques présentant une fréquence allélique inférieure à 5% peuvent être ignorés dans la production des alignements, permettant la conception rapide de fichiers répondant au différents besoins.

Analyses de la variabilité nucléotidique

Phylogénie et structure des population.

Afin d'étudier la relations entre les différents individus de chaque espèce, nous avons construit un arbre en nous basant sur une méthode de calcul des distances de type neighbor-joining (BioNJ) proposée par le programme Splitstree[17]. Cette distribution phylogénétique est inférée à partir de l'ensemble des sites polymorphiques retrouvés pour chaque isolat et permet l'obtention d'un arbre, où la longueur des branches est proportionnelle au nombre de sites polymorphiques qui différencie chaque nœud. Pour chaque espèce, un alignement a été créé où pour chaque isolat les SNP ont été inférés dans la séquence de référence. Pour *C. albicans* et *D. bruxellensis*, les SNP hétérozygotes ont été encodés grâce au code IUPAC et le mode « MatchState » a été sélectionné dans le calcul des distances.

Dans un second temps, l'analyse de la division des espèces étudiées en sous-populations a été réalisée grâce à un algorithme de clustering implémenté dans le programme Structure[18]. Cette méthode se base sur la définition d'un nombre précis de sous-populations. Pour chaque individu analysé, le programme donne alors un pourcentage d'appartenance de cet individu à chaque sous-population définie sur la base des fréquences alléliques observées. Plusieurs profils caractéristiques peuvent être mis en évidence : les individus peuvent présenter une appartenance spécifique et entière à une sous-population, révélant une homogénéité génétique. *A contrario*, un individu peut être attribué à plusieurs sous-populations avec un pourcentage d'appartenance variable, on parle alors d'individus présentant des génomes mosaïques. Pour chaque espèce, nous avons demandé à obtenir une estimation de la structure sur la base de 2 à 5 populations.

Divergence nucléotidique

Dans le chapitre 1, la divergence nucléotidique entre chaque pair d'isolats d'une même espèce a été déterminée au sein des différentes populations. Pour les souches haploïdes, l'ensemble des positions polymorphiques divergentes entre chaque pair a été décompté et la valeur a été divisée par la taille totale de la séquence de référence. Pour les souches diploïdes, une somme des SNP divergents divisées par l'ensemble des possibilités a été calculée. Par exemple, dans le cas de deux souches diploïdes : AT/AA amène à 0,5 ((0 + 0 + 1 + 1) / 4). Cette valeur a par la suite été divisée par la taille totale du génome.

Estimation de la variabilité nucléotidique

Des estimateurs de la diversité nucléotidique pour chaque espèce ont été calculés grâce au logiciel VariScan[19]. Ce programme permet de déterminer deux valeurs représentatives de la diversité nucléotidique au sein d'un échantillon : θ et π. La valeur θ est déterminée sur la base du nombre de sites polymorphiques présents dans l'ensemble de l'échantillon étudié. Par contre, la valeur π est calculée sur la base du nombre mais aussi de la fréquence des SNP pour chaque

site polymorphique. π correspond ainsi au nombre moyen de différences nucléotidiques par site entre paires de séquences dans la même espèce.

La comparaison des valeurs de π et θ permet de déterminer la valeur *D* de Tajima[20]. Cette valeur permet de voir si une région précise ou encore une population a évolué de manière neutre, c'est-à-dire que le polymorphisme génétique observé s'explique par un équilibre entre mutation et dérive. Dans ce cas, les paramètres π et θ sont égaux et la valeur *D* de Tajima est nulle. Par contre, un excès de variants peu divergents va conduire à une valeur négative. Cet excès peut être dû à une sélection négative au niveau d'un locus ou à une expansion démographique récente de la population. Finalement, un excès de variants intermédiaires entraîne une valeur *D* de Tajima positive, signe d'une sélection positive au niveau d'un locus ou à une diminution démographique de la population.

Afin d'explorer la variabilité génétique du génome complet, les SNP ont été inférés dans la séquence de référence pour chaque isolat et les séquences relatives à chaque isolat ont été concaténé dans un même alignement pour chaque contig. Ces alignements ont par la suite été utilisés par VariScan et les options « runmode = 12, usemuts =1, widthSW = 10000 et JumpSW = 10000 » ont été spécifiées afin de calculer les différents estimateurs sur des fenêtres glissantes non chevauchantes de 10kb le long du génome. Afin d'obtenir ces estimateurs pour le génome complet, les alignements relatifs à chaque contig ont été concaténés et le logiciel a été relancé avec l'option « `SlidingWindow = 0` ». Une approche similaire a été utilisée afin de déterminer la variabilité nucléotidique de chaque séquence codante.

dN/dS

Afin d'obtenir une estimation de la pression de sélection pour chaque séquence codante, le ratio entre le polymorphisme non-synonyme (dN) et synonyme (dS). Pour chaque CDS, un alignement des séquences nucléiques correspondant à la version allélique de chaque isolat a été produit. Cet alignement a par la suite été soumis au programme yn00 du package PAML[21]. Les arguments par défaut ont été utilisés à l'exception de *C. albicans* pour lequel le code génétique alternatif a été indiqué. La médiane des valeurs obtenues pour chaque gène a par la suite été calculée. Une valeur de dN/dS supérieure à 1 reflète un excès de mutation non-synonyme et est ainsi indicateur d'une sélection positive de la protéine. A l'inverse, un ratio inférieur à 1 indique une conservation de la séquence protéique résultant d'une sélection négative sur celle-ci.

Détermination *in silico* de la ploïdie des souches

Dans le chapitre 3, la ploïdie des souches a été déterminée dans un premier temps par l'analyse de la fréquence allélique des SNP hétérozygotes. Dans ce cadre, la distribution de la fréquence allélique a été analysée manuellement pour chacune des souches de la population. Alors qu'une distribution centrée sur une valeur de 0,5 indique la présence d'une souche diploïde, une distribution bimodale centrée autour d'une valeur de 0,33 et 0,66 indique la présence d'une

souche triploïde. Cette analyse a été confirmée par l'étude manuelle des aneuploïdies des souches.

Absence d'hétérozygotie

Afin d'identifier la présence de région ayant subi une perte d'hétérozygotie, le génome de chaque souche a été examiné à travers des fenêtre glissante de 50 kb avec un pas de 25 kb. Au sein de chaque fenêtre, le nombre de SNP hétérozygote a été déterminé et une région a été annotées comme ayant subi un LOH si moins de 10 SNP hétérozygotes ont été retrouvés dans la fenêtre. Par la suite, l'ensemble des fenêtres chevauchantes ont été fusionnées et les fenêtres ayant une taille inférieure à 25 kb, présentes en fin de contigs, ont été ignorées.

Construction des pangénomes

Pour les chapitres 1 et 3, un pipeline a été développé afin d'identifier et annoter les séquences supplémentaires de la référence. Ce pipeline repose sur (i) l'assemblage du génome de chaque souche, (ii) l'identification et la conservation des régions absentes dans le génome de référence, (iii) la prédiction des séquences codantes dans ces régions et finalement (iv) l'identification de l'origine et de la fonction des séquences codantes à travers la comparaison à des banques de données protéiques (Figure 4).

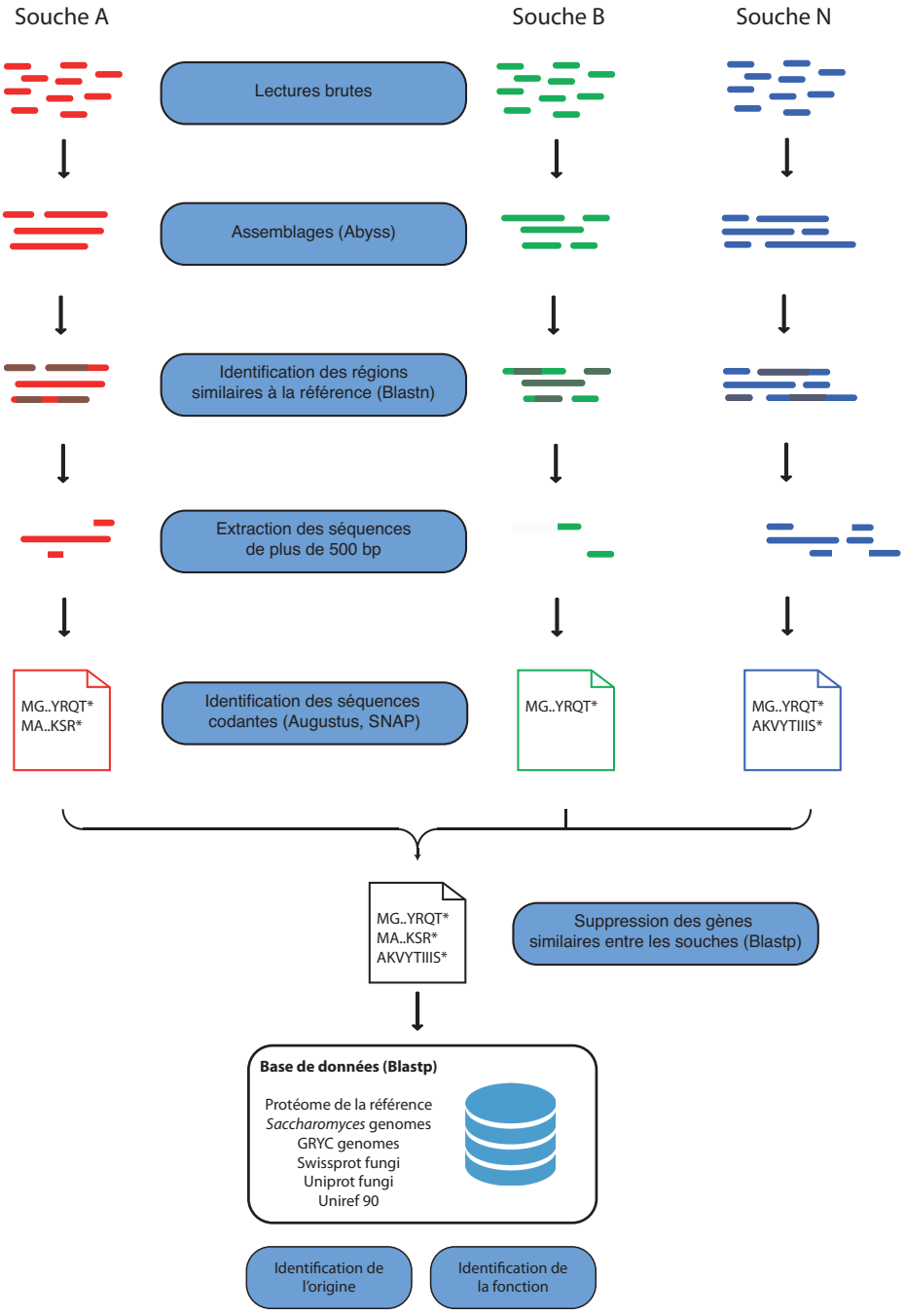


Figure 4. Représentation schématique des principales étapes du pipeline permettant l'identification du pangénome au sein de chaque espèce.

Assemblages et identification des séquences absentes dans la référence

Pour chaque souche, un assemblage a été réalisé grâce au logiciel ABySS[22] (v. 2.0.2). Pour le chapitre 1, une taille de kmer de 64 bp a été privilégié à l'exception de quelques souches de *C. albicans* et *S. uvarum* pour lesquelles une taille de 51 bp a été utilisées. Pour *D. bruxellensis*, les souches ont été assemblées à partir d'une taille de kmer de 67 bp.

Chaque assemblage a par la suite été comparé au génome de référence grâce à blastn (version 2.2.31) avec les options « `gapopen 5, gapextend 5, reward 1, penalty -5, evalue 2` et `word size 11` ». Les assemblages ont par la suite été comparés à la séquence de référence afin d'identifier les régions génomiques présentant une similarité avec celles-ci. Puisque la divergence nucléotidique peut être importante au sein de certaines populations entre certains individus et la souche de référence, un seuil a été spécifié pour chaque souche, basé sur la divergence nucléotidique déterminée dans l'analyse des SNP suivant cette formule : PID = 100 – 5 – divergence (souche, référence). Les régions couvertes par une HSPs (High Scoring Pairs) ayant une identité supérieure au PID et une taille supérieure ou égale à 50 bp ont par la suite été sélectionnées et regroupées en deux catégories : Si l'HSP obtenu est plus petit que 1 kb, l'HSP a été ignorée, sinon la séquence a été fragmentée en segments de 250 bp qui ont été réalignés contre la séquence de référence en utilisant les mêmes arguments. Si l'identité d'une HSP était supérieure au PID et que l'HSP recouvrait au moins 75% de la taille du segment, celui-ci a été annoté comme appartenant à la référence. A la fin du processus, les assemblages ont été coupés en fonction des positions génomiques des segments similaires à la référence et ayant une taille supérieure ou égale à 100 bp. Afin de supprimer l'ensemble des segments similaires, ce processus a été renouvelé 3 fois. Pour chaque souche, les différentes séquences obtenues ont été comparées contre elles-même en utilisant les mêmes options blast afin de supprimer les séquences dupliquées. Si deux séquences possédaient une HSP ayant une identité supérieure à 98 %, la séquence la plus courte a été supprimée.

Prédiction des gènes et nettoyage

Pour chaque segment, les gènes potentiels ont été prédits en utilisant à la fois les logiciels Augustus[23] (v. 3.2) et SNAP[24] (v. 2006.07.28). Pour chaque programme, les logiciels ont été entraînés à partir d'un jeu de données spécifique à chaque espèce. Pour *S. uvarum* et *S. paradoxus*, le fichier d'entraînement déjà disponible de *S. cerevisiae* a été utilisé. Pour Augustus, les fichiers d'entraînement inclus dans le logiciel pour *S. cerevisiae* et *C. albicans* ont été utilisés. Pour les autres espèces, un sous-ensemble aléatoire de 1000 CDS a été utilisé afin de produire les fichiers d'entraînement en suivant la méthodologie spécifique à chaque programme.

Afin de supprimer les gènes présentant des régions de faible complexité, le logiciel SEG disponible dans la suite Blast a été utilisé. Les protéines ayant moins de 50 acides aminés présents en dehors d'une région de faible complexité ont ainsi été supprimées.

L'ensemble des protéines ont par la suite été regroupées dans un seul fichier et comparées entre elles afin de déterminer les protéines similaires entre les souches d'une même espèce. Dans ce cadre, l'ensemble des séquences ont été comparées grâce à Blastp. Les séquences présentant une similarité supérieure à 98% ont par la suite été regroupées dans un même groupe et un réseau a été construit pour chaque groupe. Pour chaque réseau, le nœud central correspondant à la protéine la plus proche de l'ensemble des protéines de chaque groupe a été sélectionné comme représentative, tandis que les autres séquences ont été éliminées.

Identification de l'origine et la fonction des gènes prédits

La fonction et l'origine des protéines supplémentaires ont par la suite été déterminées grâce à la comparaison de leur séquence protéique contre plusieurs banques de données protéiques :

- Le protéome de chaque séquence de référence (A)
- Le protéome de *S. cerevisiae* (B)
- L'ensemble des protéomes retrouvés sur la plateforme GRYC (C) (http://gryc.inra.fr)
- Six espèces appartenant au groupe des *Saccharomyces sensus stricto* (D) (http://sss.genetics.wisc.edu/cgi-bin/s3.cgi)
- La banque de données SwissProt Fungi et Uniprot Fungi (E)
- Uniref 90 (F)

Afin de déterminer la fonction des CDS supplémentaires, une analyse itérative a été réalisée sur les différentes banques de données suivant cet ordre : (B – C – D – E – F – A). Pour chaque protéine, si une similarité supérieure à 70% a été retrouvée à partir d'une analyse Blastp avec une protéine de la banque de données, celle-ci a été annotée par rapport à la séquence retrouvée et ignorée pour les prochaines itérations.

Afin d'identifier l'origine des protéines, l'ensemble des résultats Blastp contre les banques de données ont été regroupés et seuls les résultats présentant une similarité supérieure à 30 % ont été gardés. Pour chaque protéine, le meilleur résultat a été sélectionné en fonction de la similarité par rapport à la protéine de la banque de données. Chaque protéine a par la suite été catégorisée en quatres groupes : Ancestrale si le meilleur résultat correspond à une protéine de la séquence de référence avec une similarité minimum de 70 % ; HGT (Horizontal Gene Transfert) si le meilleur résultat ne provient pas d'une espèce appartenant au même genre que l'espèce étudiées ; Introgression si dans le cas contraire, le genre est le même. Les gènes n'appartenant à aucune des précédentes catégories ont été classifiés comme Unknown.

Courbes de saturation

Pour le chapitre 1, la taille du pan et du core génome de chaque espèce pour un plus grand nombre d'individus a été modélisée à travers l'utilisation d'une régression non-linéaire. Dans ce cadre une distribution pour chaque valeur N d'individus compris entre 1 et la taille de la

population de chaque espèce a été calculée. Pour chaque nombre N, 5000 itérations ont été réalisées et le nombre total de CDS non redondantes a été calculé pour un nombre N d'individus choisis aléatoirement au sein de la population. Ces distributions ont par la suite été utilisées afin d'ajuster différents modèles de régression non-linéaires :

- $y = a \times e^{\left(-\frac{x}{b}\right)} + c$
- $y = a \times x^{-b} + c$
- $y = a \times e^{(b \times x)} + c$
- $y = a \times x^{b} + c$

Établissement d'une séquence de référence

Assemblage de génome de novo à partir des données MinION

Les différents jeux de données de lectures MinION 2D, correspondant aux couvertures théoriques du génome (10, 15, 20 et l'ensemble des lectures 2D) ont été sujets à 4 assembleurs : ABruijin[25] (v 0.3b), Canu[26] (v 1.1), miniasm[27] (v 0.2-r137-dirty) et SMARTdenovo (https://github.com/ruanjue/smartdenovo).

ABruijin et miniasm ont été lancés avec les paramètres par défaut. Pour Canu et SMARTdenovo les options « `genomeSize = 13m, minReadLength = 2500, mhapSensitivity = high, corMhapSensitivity = high corOutCoverage = 500` » et « `-c 1 -k 14 -J 2500 -e zmo` » ont été utilisées respectivement. Après assemblage, les contigs ont été corrigés par Pilon[28] (v 1.18) à partir de lectures Illumina pairées de 100 bp. Finalement, SSPACE-LongRead[29] (v 1.1) a été utilisé afin de regrouper les assemblages grâce aux informations provenant des lectures longues.

Évaluation de la qualité des assemblages

La qualité des différents assemblages a tout d'abord été évaluée par l'analyse des lectures courtes non mappées déterminées grâce à Samtools[30] (v 0.1.19) avec les options « `view -f 4 -c` ». Cette analyse a été complétée par l'examination des gènes eucaryotes très conservés retrouvés par CEGMA[31] (v 2.5) en utilisant les paramètres par défaut.

Comparaison des génomes complets

La comparaison des génomes obtenus a été réalisée grâce à MUMmer[13] (v 3.0). Nucmer avec l'option « `-maxmatch` » a permis l'alignement des séquences. Les coordonnées des alignements ont été extraits afin de déterminer la proportion des résidus non-ambigües (N) pour chaque assemblage. Les fichiers delta ont par la suite été filtrés afin d'ignorer les alignements d'une taille inférieure à 5 kb. Les données obtenues ont par la suite été soumises à mummerplot afin d'obtenir les représentations graphiques de la colinéarité des génomes.

Alignement des lectures courtes

Les lectures ont été alignées contre le génome grâce à BWA[30] (v 0.7.4) et le nombre de lectures non-alignées a été estimé grâce à Samtools[10] (v 0.1.19). GATK[11] (v 3.3) a par la suite été utilisé afin de réaligner les lectures dans les régions présentant des insertions et délétions de petites taille, de déterminer les variants nucléotidiques et d'obtenir la fréquence allélique de ces derniers.

Accessibilité des données

L'ensemble des données de séquençage générées dans cette étude, ainsi que l'assemblage de référence obtenue pour la souche UMY321 ont été déposés sur la base de données "European Nucleotide Archive" (ENA) sous l'identifiant PRJEB21262.

Annotation du génome

Afin de produire une annotation des éléments fonctionnels au sein de la séquence de référence de *D. bruxellensis*, le logiciel Amadea Annnotation transfer tool (Isoft, France) a été utilisé. Les génomes de la souche CNS3082^T (*Lachancea kluyveri*) et celui de la souche CBS767^T (*Debaryomyces hansenii*) ont été utilisés comme génome de référence (versions corrigées, disponible sur GRYC (http://gryc.inra.fr)). Cette étape a été suivie d'un nettoyage manuel à partir de données de RNA-Seq provenant de la souche de *D. bruxellensis* CBS2499[32] (SRA : SRR427169 – Projet : PRJNA76499). Le logiciel d'alignement Tophat2[33] (v. 2.1.0) a par la suite été utilisé afin d'aligner les lectures contre le génome assemblé de YJS5431. Afin de modifier les fichiers bam, le logiciel Artemis[34] (v. 16.0.0) a été utilisé, permettant de corriger les coordonnées des exons et des introns, ainsi que d'identifier les IncRNA. Les ARNt ont par la suite été identifiés grâce à tRNA-scan-SE[35] (v. 1.3.1) et les éléments transposables ont été déterminés grâce à une comparaison Blast contre des éléments transposables de levures provenant de différentes familles, tels que *Ty1-copia*, *Ty3-gypsy*, et *hAT*.

Correction de la séquence de référence

Après la première étape d'annotation, plusieurs pseudo-gènes présentant une taille de CDS non divisible par trois ont été déterminés (1427 / 5226), résultant essentiellement d'erreurs d'assemblage de par la forte hétérozygotie de la souche de référence séquencée. Afin de corriger ces annotations, nous avons réassemblé le génome grâce à SOAPdenovo à partir des données de type Illumina. Cet assemblage a par la suite été soumis au logiciel Redundans[36] (v. 0.13c) en utilisant la séquence de référence produite initialement comme modèle. Les différentes séquences des pseudo-gènes ont par la suite été confrontées à l'assemblage produit par Redundans grâce au logiciel Blastn, et seuls les HSP couvrant de manière complète les séquences des pseudo-gènes ont été conservés. Un alignement global à partir du logiciel needle de la suite EMBOSS[37] a été réalisé à partir de la séquence protéique des CDS sélectionnés et seuls ceux présentant une similarité supérieure à 80 % ont été conservés. Dans ce cas, la séquence nucléotidique a été remplacée dans la séquence de référence et l'ensemble des coordonnées des différents éléments fonctionnels ont été mis à jour. Cette procédure a permis de corriger la séquence de 872 gènes dans l'assemblage initial, et l'annotation comme pseudo-gène a été conservée pour les 555 gènes restants.

Références

1. Hirakawa, M. P. *et al.* Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Res.* gr.174623.114- (2015). doi:10.1101/gr.174623.114
2. Almeida, P. *et al.* A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. *Nat. Commun.* **5,** 4044 (2014).
3. Leducq, J.-B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat. Microbiol.* **1,** 15003 (2016).
4. Friedrich, A., Jung, P., Reisser, C., Fischer, G. & Schacherer, J. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol. Biol. Evol.* **32,** 184–92 (2015).
5. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430,** 35–44 (2004).
6. Souciet, J.-L. *et al.* Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res.* **19,** 1696–709 (2009).
7. Scannell, D. R. *et al.* The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces sensu stricto* genus. *G3 (Bethesda).* **1,** 11–25 (2011).
8. Jones, T. *et al.* The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci.* **101,** 7329–7334 (2004).
9. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
10. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).
11. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).
12. Boeva, V. *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27,** 268–269 (2011).
13. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
14. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
15. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* **6,** 80–92
16. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–81 (2009).
17. Huson, D. H. Drawing rooted phylogenetic networks. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* (2009). doi:10.1109/TCBB.2008.58
18. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–59 (2000).
19. Vilella, A. J., Blanco-Garcia, A., Hutter, S. & Rozas, J. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics* **21,** 2791–3 (2005).
20. Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123,** 585–595 (1989).
21. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–91 (2007).
22. Jackman, S. D. *et al.* ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res.* **27,** 768–777 (2017).
23. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2,** ii215-25 (2003).
24. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5,** 59 (2004).
25. Lin, Y. *et al.* Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E8396–E8405 (2016).
26. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27,** 722–736 (2017).
27. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long

sequences. *Bioinformatics* **32,** 2103–10 (2016).
28. Walker, B. J. *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* **9,** e112963 (2014).
29. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15,** 211 (2014).
30. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).
31. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067 (2007).
32. Piškur, J. *et al.* The genome of wine yeast *Dekkera bruxellensis* provides a tool to explore its food-related properties. *Int. J. Food Microbiol.* **157,** 202–9 (2012).
33. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
34. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28,** 464–9 (2012).
35. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44,** W54-7 (2016).
36. Pryszcz, L. P. & Gabaldón, T. Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44,** e113 (2016).
37. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).

CONCLUSION ET PERSPECTIVES

Au sein de toute espèce, les génomes des individus résultent de l'accumulation de mutations génomiques constituant le matériel brut sur lequel la sélection agit, permettant l'adaptation à leur environnement. L'importante diversité à la fois des pressions de sélection, ainsi que facteurs impliqués dans la génération, la maintenance et la sélection de nouvelles mutations rend l'évolution de chaque espèce unique. L'étude précise de ces facteurs au sein de chaque espèce et de leur incidence sur leur évolution est maintenant rendue possible par le séquençage du génome complet d'un grand nombre d'individus. À travers l'établissement d'études de génomique des populations au sein de plusieurs espèces de levures du sous-phylum des Saccharomycotina, nous avons cherché à déterminer l'importance de ces différences évolutives entre les espèces.

Analyse comparative de génomique des populations

Dans ce contexte, nous avons réalisé dans un premier temps une étude comparative de génomique des populations à travers l'analyse systématique de la variabilité intraspécifique chez 6 espèces de levures présentes au sein du sous-phylum des Saccharomycotina et pour lesquelles une séquence de référence et des annotations de bonnes qualités étaient disponibles. Parmi les espèces étudiées, des collections pour *K. lactis* et *L. thermotolerans* ont été spécifiquement constituées tandis que les données de séquençage de *C. albicans*, *L. kluyveri*, *S. paradoxus* et *S. uvarum* proviennent du domaine public. Pour chaque espèce, plusieurs variants génétiques tels que les mutations ponctuelles ou les aneuploïdies ont été systématiquement identifiés. La comparaison de ces données a révélé d'importantes différences dans l'impact de ces mécanismes évolutifs sur le génome des espèces amenant à des histoires évolutives divergentes entre elles. Afin d'obtenir une vue exhaustive de la variabilité intraspécifique responsable de la diversité phénotypique, nous nous sommes intéressés plus précisément à la variation du répertoire de gènes au sein de chaque espèce. Pour se faire, le nombre variable de copies des gènes a été déterminé et les gènes absents dans la référence ont été identifiés à partir de l'analyse d'assemblages *de novo*, permettant la construction des différents pangénomes. Les résultats ont montré d'importantes variations dans la constitution des core et pangénomes entre les espèces, indiquant que la variabilité du contenu en gènes est une composante importante de la diversité génétique au sein des levures. De manière intéressante, un grand nombre d'introgressions provenant d'espèces proches a pu être retrouvé dans les génomes accessoires, suggérant la présence de multiples évènements d'hybridation entre les différentes espèces de levure. Dans l'ensemble, ces résultats démontrent l'importance de l'étude de variabilité du répertoire de gènes chez les espèces, et il serait désormais intéressant de connaître plus précisément l'impact que cette diversité a sur la variabilité phénotypique.

Exploration de la diversité intraspécifique chez *D. bruxellensis*

Une exploration globale de la variabilité intraspécifique au sein des différentes espèces composant le sous-phylum des Saccharomycotina est cependant encore grandement limitée, notamment par l'absence de séquence de référence et d'annotation de qualité pour une majorité d'espèces non-modèles. Les nouvelles technologies de séquençage offrent aujourd'hui la

possibilité de surmonter cette limitation à travers la production de lectures plus longues facilitant les processus d'assemblage. Dans ce cadre, nous avons essayé de prouver la faisabilité d'une telle méthode à travers l'obtention d'un assemblage de qualité pour l'espèce *D. bruxellensis* à travers la production et l'analyse de données de type Oxford Nanopore avec comme objectif d'initier une étude de génomique des populations pour cette espèce. En effet, cette espèce est impliquée dans plusieurs processus de fermentation et présente des propriétés évolutives intéressantes, notamment la présence d'évènements d'hybridation, motivant l'étude de son évolution. L'assemblage généré présente une plus forte contiguïté et est davantage complet en comparaison aux assemblages produits dans de précédentes études. Par ailleurs, l'alignement de données de séquençage de plusieurs souches a montré que cet assemblage est approprié dans la réalisation d'une étude de génomique des populations. Ce résultat ouvre des perspectives intéressantes dans l'exploration de la variabilité intraspécifique au sein d'espèces non-modèles, confirmant l'intérêt de ce type de séquençage dans l'obtention de séquence de référence.

Finalement, la séquence de référence obtenue a été utilisée afin de produire une étude de génomique des populations au sein de *D. bruxellensis*. Dans ce cadre, 53 souches ont été séquencées à travers un séquençage de type Illumina, permettant pour la première fois l'étude de la variabilité génomique à l'échelle du génome chez cette espèce. De manière additionnelle, un séquençage de type Oxford Nanopore a été réalisé pour 3 souches afin de pouvoir réaliser une analyse comparative de leur assemblage. Cette étude a permis l'identification d'au moins deux évènements d'hybridation indépendants au sein de l'espèce, dont un partagé entre deux sous-populations triploïdes. De manière intéressante, de nombreux évènements de perte d'hétérozygotie peuvent être observés dans les différentes sous-populations, suggérant l'importance de ce mécanisme dans l'évolution et l'adaptation de plusieurs espèces de levures[1–4]. Par ailleurs, la détermination du pangénome et du nombre variable de copies des gènes a permis d'identifier plusieurs gènes impliqués notamment dans le transport d'antibiotiques ou dans le métabolisme de sucres chez les souches triploïdes isolées lors du processus fermentaire du vin. Pris ensemble, ces résultats permettent de mieux caractériser l'histoire évolutive ainsi que les variants impliqués dans la relation génotype-phénotype chez *D. bruxellensis*.

Vers une exploration plus large de la diversité intraspécifique

L'amélioration et la réduction des coûts de séquençage ces dernières années ont donné l'occasion d'explorer plus en détail l'évolution de différentes espèces. Chez *S. cerevisiae*, l'analyse de plus de 1000 génomes a permis une dissection précise de son histoire évolutive et des variants impliqués dans son adaptation[1]. En parallèle, l'étude d'espèces non-modèles est en plein essor, amenant à l'exploration d'un nombre toujours plus important d'espèces au sein du vivant et chez la levure[5,6], nous renseignant sur les différences évolutives présentes entre les espèces. Cet axe de recherche qui se révèlera crucial dans notre compréhension de l'évolution du vivant est cependant encore en plein développement, ouvrant la voie à un grand nombre de perspectives intéressantes.

Dans un premier temps, il se révèlera essentiel d'explorer la variabilité génétique d'un plus grand nombre d'individus au sein des différentes espèces. En effet, la majorité des études de génomique des populations d'espèces non-modèles est pour l'instant limitée par un jeu de données restreint, habituellement bien inférieur à 100 souches (N = 41 en moyenne), ne permettant pas une analyse complète de l'ensemble des processus évolutifs impliqués dans l'évolution de ces espèces. Cet aspect reste cependant inhérent au nombre de souches isolées, et une collection plus importante sera nécessaire afin de produire de telles analyses pour une majorité des espèces. Néanmoins, certaines espèces possèdent déjà un nombre suffisant d'isolats, par exemple près de 1500 individus sont disponibles pour l'espèce *D. bruxellensis*. Chez *S. cerevisiae*, l'exploration de plus de 1000 souches a montré l'intérêt d'un tel jeu de données dans l'exploration de l'histoire évolutive et des processus adaptatifs entre les différentes sous-populations. Par ailleurs, ces données se révèleront particulièrement intéressantes dans l'étude de la relation génotype-phénotype. En effet, la comparaison entre les variants génomiques de plusieurs centaines individus et des données de phénotypage permettra l'établissement d'études d'association pangénomique. Chez les levures, ces études ont pour l'instant été principalement réalisées au sein de *S. cerevisiae* et ont permis d'identifier avec précision des associations entre des variants génétiques et l'adaptation des souches à certaines conditions[1,7]. Par ailleurs, ces études ont permis l'identification de différentes contraintes techniques, telle que la structure de la population, ce qui facilitera la mise en place d'études d'association pangénomique chez d'autres espèces.

Dans ce cadre, l'identification d'un plus grand nombre de variants génomiques et notamment des variants structuraux de large taille telles que les phénomènes d'inversions ou de translocations sera particulièrement intéressante. En effet, plusieurs études ont montré l'implication de ces mutations dans la variabilité phénotypique, par exemple plusieurs versions de translocations impliquant le gène *SSU1* chez certains isolats de vin de *S. cerevisiae* sont associées à une résistance plus importante au sulfite[8]. Ces variants génétiques, difficilement indétectables avec un séquençage de type Illumina, sont maintenant identifiables à moindre coût grâce aux nouvelles technologies de séquençage telle que celle proposée par Oxford Nanopore. La mise en place d'une telle approche a ainsi été possible dans le cadre de notre étude de génomique des populations de *D. bruxellensis* ou dans l'analyse de 22 isolats naturels de *S. cerevisiae*[9], confirmant la possibilité d'identifier de tels variants. L'amélioration de ces technologies offriront à l'avenir la possibilité de déterminer avec précision l'ensemble de ces variants génomiques à l'échelle d'une population qui pourront notamment être incorporés dans les études d'association, permettant ainsi de déterminer avec plus de précision leur rôle dans la variation phénotypique à l'échelle d'une espèce.

De manière additionnelle, l'obtention d'un assemblage de qualité pour plusieurs souches d'une même espèce facilitera de manière globale les analyses de génomique des populations au sein d'espèces présentant une forte variabilité génomique. Le passage d'une séquence de référence unique à l'alignement sur plusieurs génomes de référence permettra une meilleure identification des variants génétiques au sein de la population. Une telle approche, dénommée « genome graph » a ainsi été proposée dans l'analyse des variants génétiques chez l'Homme et a déjà

montré l'intérêt de cette méthode, permettant d'identifier avec une plus grande précision l'ensemble des variants génomiques[10,11]. Chez les levures, cette approche est particulièrement intéressante de par l'importante diversité génomique observable au sein de certaines espèces, telle que *K. lactis* pour laquelle une divergence de plus de 8% a été déterminée entre deux sous-populations. À terme, l'application de ce type d'approche permettra une meilleure identification des variants, mais aussi de mieux caractériser le pangénome de ces espèces.

Finalement, l'amélioration de ces techniques permettra d'étendre l'exploration de la variabilité intraspécifique à un plus grand nombre d'espèces. Au sein du groupe des *Saccharomycetaceae*, des génomes de référence de bonne qualité ainsi que des isolats sont déjà disponibles pour un grand nombre d'espèces[12] (Figure 1). Par exemple, une large collection d'individus est disponible pour l'espèce *Torulaspora delbrueckii*, impliquée dans plusieurs processus de fermentation, notamment comme levure de boulangerie ou la production de boissons fermentées (cidre, bière, kéfir). À l'échelle du sous-phylum des Saccharomycotina, l'exploration d'espèces plus éloignées se révélera aussi intéressante, notamment au sein des espèces du groupe basal pour lesquelles la variabilité intraspécifique est encore largement inexplorée. L'analyse comparative au sein de ce groupe a révélé d'importantes différences dans la constitution de leur génome, par exemple le génome de *Yarrowia lipolytica* possède un nombre de CDS supérieur à *S. cerevisiae* (6618 *vs* 5769 respectivement), riches en intron (35% *vs* 3%), ainsi qu'un génome bien plus large (20,6 *vs* 12,1 Mb), amenant potentiellement à une évolution différente de leur génome. Dans ce cadre, l'obtention plus aisée d'une séquence de référence à partir de lectures longues facilitera la mise en place de telles études. L'établissement de tels jeux de données demande cependant encore un travail important, notamment de par les processus d'annotations encore complexes des génomes mais nécessaires à la mise en place d'une analyse fonctionnelle. Ces méthodes reposent notamment sur l'identification de protéines déjà annotées et ayant une similarité de séquence proche et l'obtention d'un nombre croissant de génomes annotés, pour faciliter la réalisation de ces annotations.

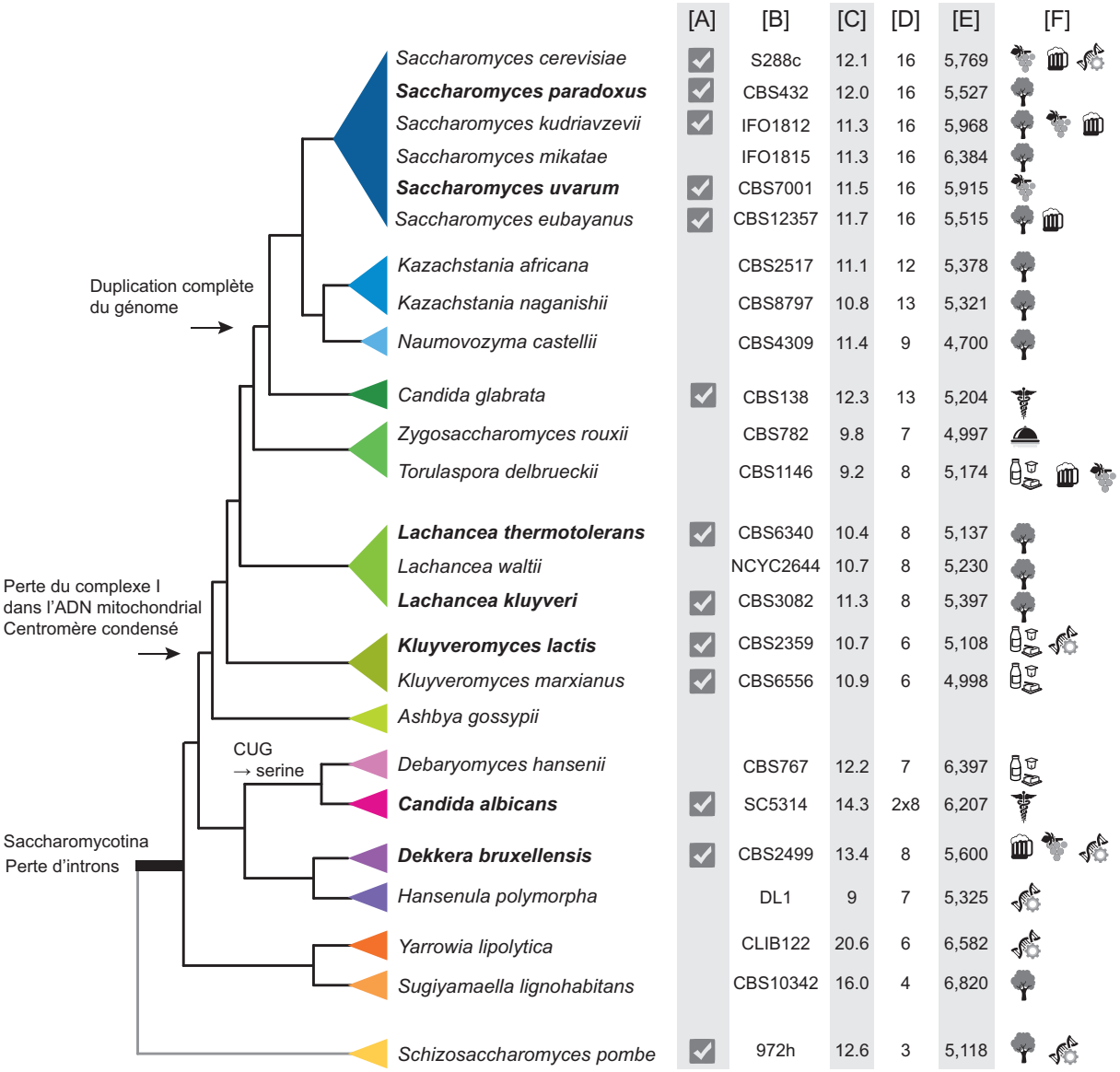


Figure 1. Arbre phylogénétique non exhaustif des espèces dont le génome a été séquencé au sein du sous-phylum des Saccharomycotina. Les espèces étudiées dans mes travaux de thèse sont en gras. (A) Espèces pour lesquelles une ou plusieurs études de génomique des populations ont déjà été réalisées. (B) Souche de référence. (C) Taille en Mb du génome de référence. (D) Nombre de chromosomes. (E) Nombre de CDS identifiées dans les génomes de référence. (F) Aperçu des environnements retrouvés pour chaque espèce.

Ainsi, l'exploration de la variabilité intraspécifique au sein des différentes espèces n'en est qu'à son début et ouvre la voie à un ensemble de perspectives intéressantes. D'un point de vue technique, l'amélioration des techniques de séquençage, des ressources informatiques et des algorithmes utilisés rendront possible l'analyse fine de la variabilité génomique chez un nombre toujours plus important d'individus. Sur le long terme, la mise en place de différentes études enrichiront grandement notre vision sur les différences évolutives présentent au sein des levures, nous renseignant de manière sans précédent sur les mécanismes impliqués et leur impact dans l'évolution des génomes.

Références

1. Peter, J. *et al.* Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556,** 339–344 (2018).
2. Ford, C. B. *et al.* The evolution of drug resistance in clinical isolates of *Candida albicans*. *Elife* **4,** e00662 (2015).
3. Okuno, M. *et al.* Next-generation sequencing analysis of lager brewing yeast strains reveals the evolutionary history of interspecies hybridization. *DNA Res.* **23,** 67–80 (2016).
4. Smukowski Heil, C. S. *et al.* Loss of heterozygosity drives adaptation in hybrid yeast. *Mol. Biol. Evol.* **34,** 1596–1612 (2017).
5. Ellegren, H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* **29,** 51–63 (2014).
6. Peter, J. & Schacherer, J. Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast* **33,** 73–81 (2016).
7. Strope, P. K. *et al.* The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* gr.185538.114- (2015). doi:10.1101/gr.185538.114
8. Pérez-Ortín, J. E., Querol, A., Puig, S. & Barrio, E. Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome Res.* **12,** 1533–9 (2002).
9. Istace, B. *et al. de novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* **6,** 1–13 (2017).
10. Novak, A. M. *et al.* Genome Graphs. *bioRxiv* 101378 (2017). doi:10.1101/101378
11. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27,** 665–676 (2017).
12. Dujon, B. A. & Louis, E. J. Genome diversity and evolution in the budding yeasts (Saccharomycotina). *Genetics* **206,** (2017).

ANNEXES

Liste des publications

Publiées :

- Hou J, Friedrich A, **Gounot J-S**, Schacherer J (2015). Comprehensive survey of condition-specific reproductive isolation reveals genetic incompatibility in yeast. ***Nature Communications***. 6, 7214.

- Fournier T*, **Gounot J-S***, Freel K, Cruaud C, Lemainque A, Aury J.-M, Wincker P, Schacherer J and Friedrich A (2017). High-quality *de novo* genome assembly of the *Dekkera bruxellensis* yeast using Nanopore MinION Sequencing. ***G3*** (Bethesda). *7*, 3243–3250.

En cours de rédaction :

- **Gounot J-S**, Freel K, Friedrich A, Schacherer J. Yeast pangenomes are mainly shaped by introgression events

- **Gounot J-S**, Neuvéglise C, Abou-Saada O, Fournier T, Friedrich A, Schacherer J. High complexity and degree of genetic variation in Dekkera bruxellensis population

* Les auteurs ont contribué de manière égale à ce travail.

Liste des communications

1. **Gounot J-S**, Freel K, Friedrich A, Schacherer J. (2016). Yeast Population Genomics Reveals Distinct Evolutionary Histories across Multiple Species. EMBO conferences on Experimental Approaches to Evolution and Ecology (Heidelberg, Germany). **Poster**

2. **Gounot J-S**, Freel K, Friedrich A, Schacherer J. (2016). Yeast Population Genomics Reveals Distinct Evolutionary Histories across Multiple Species. Levures, modèles, outils (Bruxelles, Belgique). **Oral**

3. **Gounot J-S**, Freel K, Friedrich A, Schacherer J. (2016). Yeast Population Genomics Reveals Distinct Evolutionary Histories across Multiple Species. Séminaire de microbiologie de Strasbourg (Strasbourg, France). **Oral**

Enseignements

Les différents enseignements ont été réalisés au sein de l'École Supérieure de Biotechnologie de Strasbourg (ESBS) dans le cadre d'un contrat ATER.

- Data processing using Python (TP)
- Structural biology (TD)
- Protein structure (TD)
- Travaux pratiques d'instrumentation et de biophysique (TP)

Jean-Sébastien GOUNOT

Génomique des populations : Étude comparative au sein du sous-phylum des Saccharomycotina

Résumé

Les améliorations des technologies de séquençage offrent aujourd'hui la possibilité d'explorer la variabilité intraspécifique au sein d'une espèce à travers le séquençage complet du génome d'un grand nombre d'individus. Dans ce contexte, mes travaux de thèse se sont basés sur l'étude et la comparaison de la variabilité génomique à travers des études de génomique des populations au sein de plusieurs espèces de levures. Dans un premier temps, j'ai réalisé une étude systématique de la variabilité intraspécifique au sein de 6 espèces de levures, me donnant notamment la possibilité d'étudier la variabilité du contenu en gènes entre les espèces. Dans un second temps, je me suis focalisé sur l'utilisation des dernières technologies de séquençage dans l'objectif de produire une séquence de référence de Dekkera bruxellensis, dont l'absence pour un grand nombre d'espèces limite l'établissement d'étude de génomique des populations. Cette séquence a été utilisée dans un dernier temps afin d'étudier l'évolution de l'espèce. Dans l'ensemble, ces travaux apportent de solides fondations dans l'exploration de la diversité génétique au sein d'espèces non-modèles.

Mots clés : génomique des populations, bioinformatique, levure

Résumé en anglais

Advent of high throughput technologies as well as the reduction of their price open the way to the exploration of the intraspecific genetic variation at the species level by sequencing the complete genome of a wide range of individuals. Doing so, I first produced populations genomics studies of 6 yeast species based on the same framework, allowing the exploration and comparison of the genes repository of each species. I then used new sequencing technologies to produce a reference sequence for the yeast species *Dekkera bruxellensis*. Using this sequence, I was then able to produce for the first time a population genomic study at the genome wide scale for this species.

Keywords : Population genomic, bioinformatic, yeast