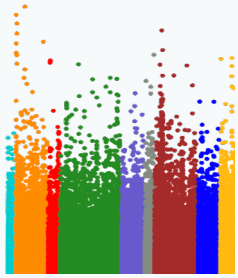
Vers une meilleure dissection des traits complexes au sein de populations naturelles de levures

Elodie CAUDAL

HaploTeam | UMR 7156 | CNRS | Université de Strasbourg | 2021



UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE ED414

UMR7156 Génétique Moléculaire Génomique Microbiologie

THÈSE présentée par :

Elodie CAUDAL

soutenue le : **22 septembre 2021**

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

Spécialité : Aspects moléculaires et cellulaires de la biologie

Caractérisation de la variation d'expressivité au sein d'une population

THÈSE dirigée par :

Pr. SCHACHERER Joseph	Professeur, Université de Strasbourg

RAPPORTEURS :

Pr. DUJON Bernard	Professeur, Institut Pasteur, Paris
Dr. MARULLO Philippe	Chargé de Recherche, Université de Bordeaux

AUTRES MEMBRES DU JURY :

Pr. MEIGNIN Carine	Professeur, Université de Strasbourg
Dr. WINCKER Patrick	Directeur de Recherche, Génoscope, Evry

A toi mon Juju,
mon petit frère extra-ordinaire,
merci d'avoir toujours guidé mes choix,
notamment celui-là !

Résumé

L'élucidation des origines génétiques des traits complexes reste une problématique majeure en biologie. Ces dix dernières années, les projets de reséquençage de milliers d'individus d'une même espèce ont permis d'explorer la diversité génétique au sein des génomes. Des études d'association pangénomique ont alors été initiées pour identifier les variants génétiques impliqués dans la variance phénotypique observée dans des populations étudiées. Cependant, les variants détectés n'expliquent qu'une portion plus ou moins importante de la variance phénotypique. Mieux caractériser la relation entre génotypes et phénotypes a ainsi été la base de mes travaux de thèse. Pour cela, une collection de plus de 1000 isolats naturels de *Saccharomyces cerevisiae* divers et séquencés a été une ressource clé pour disséquer dans un premier temps les origines génétiques de la variation du niveau d'expression de plus de 6000 gènes au sein de cette espèce. Dans un second temps, une stratégie de mutagénèse par insertion de transposons réalisée dans une centaine d'isolats de levure a permis d'estimer dans quelles proportions le fonds génétique impacte la variation des phénotypes.

Mots-clés : relations génotype-phénotype, génomique des populations, levure, GWAS, transcriptome, mutagénèse par insertion de transposons

Summary

Unravelling the genetic basis under complex traits is a major issue in biology. This last decade, resequencing projects including thousands of individuals from the same species allowed the survey of the genetic diversity along the genomes. Genome-wide association studies (GWAS) were then initiated to identify genetic variants involved in the phenotypic variance observed across natural populations. However, these genetic variants only explained a more or less extensive fraction of the observed phenotypic variances. The goal of my thesis was to obtain a better overview of the genotype-phenotype relationship. To this end, a collection of more than 1,000 *Saccharomyces cerevisiae* natural isolates previously sequenced was at first powerful enough to explore the genetic basis of gene expression variation in more than 6000 genes across this species. Secondly, a transposon mutagenesis strategy was used in one hundred of yeast natural isolates to provide an estimation of the genetic background effect on phenotypic variation after gene loss-of-functions.

Keywords: genotype-phenotype relationship, population genomics, yeast, GWAS, transcriptome, transposon mutagenesis

Remerciements

L'ensemble de ce travail a été réalisé au sein du laboratoire de Génétique Moléculaire, Génomique et Microbiologie, UMR7156 / Université de Strasbourg / CNRS dans l'équipe Variation intra-spécifique et évolution des génomes, sous la direction de Joseph Schacherer.

Je tiens tout d'abord à remercier les membres de mon comité de thèse, Pr. Bernard Dujon, Dr. Philippe Marullo, Pr. Carine Meignin et Dr. Patrick Wincker, pour avoir accepté d'évaluer mon travail. Je souhaiterais remercier plus particulièrement Pr. Bernard Dujon d'avoir fait partie de mes comités de suivi de thèse et d'avoir su me conseiller et me remotiver pour terminer cette thèse. Je remercie également Dr. Bertrand Séraphin d'avoir participé à ces comités de suivi de thèse.

Joseph, merci de m'avoir fait confiance il y a presque 6 ans, d'abord pour un stage, puis un autre, puis la thèse avec tous ces gros projets entrepris. Merci aussi pour ta patience face à toutes mes interrogations et tout le temps que tu accordes à ton équipe, à la recherche, à faire avancer les projets ! Et merci de dynamiser l'équipe comme tu le fais… Jamais je n'aurais couru 5km avant, depuis je ne pourrais plus me passer de sport, comme quoi, l'Ekiden a du bon (le 17/10/21, si jamais) !

Au gang de filles du bureau 228, merci d'avoir été au top pour créer une très bonne ambiance de travail (et pas que), pourtant la barre était haute après le bureau de l'enfer en Bota ! Anne, je te dois bien plus qu'un merci autant professionnellement que personnellement. Je crois que dans les semaines / mois / années à venir je vais continuer de me retourner pour poser une question en disant « Anne, j'ai une question sur les 1011… ». Merci d'avoir toujours répondu à mes questions même quand il fallait soulever des tapis bien planqués. Merci aussi pour toutes les corrections, relectures mais aussi les mapping, re-mapping, re-re-re-mapping. Et plus personnellement, merci pour ta bonne humeur, ta motivation, tes encouragements et toutes les histoires sur les étudiants et leur passion pour les verres à pied, sur tes enfants ou les salades de pâtes (maintenant c'est écrit, tu t'en souviendras !). Emna, merci aussi pour ta bonne humeur, ta gentillesse et de m'avoir souvent faire rire pour tes soucis de papiers à rendre à l'école doc ou de son pendant les visio. J'espère que tu continueras de porter ce projet Kombucha avec autant d'enthousiasme qu'actuellement et je te souhaite plein de courage pour terminer ta thèse. Marion, alors d'abord merci pour les dilutions, PCR, dissections et

compagnie… Vraiment, tu m'as été d'une grande aide sur le projet TranSat, souvent alors que j'étais démotivée. Bien entendu, merci d'avoir contribué à la bonne ambiance du bureau, merci aussi de compatir à tous mes moments de stress. J'espère que tu nous feras à nouveau part de tes prouesses de gymnaste lors d'une sortie au trampoline park. Et qui sait, peut-être à bientôt en Bretagne !
En parlant de toi Marion, ça me fait penser aux membres de « Joga Bonita Do Brasil » et des jeux des départements, des capitales et du fameux cookie game. J'espère que ces traditions perdureront encore longtemps. Elie, je compte sur toi ! Merci pour tous ces délires souvent basés sur de la beauferie ultime mais là-dessus on se complète bien, zéro jugement. D'ailleurs je rends ce manuscrit un vendredi, la chanson de cette semaine se doit d'être pépite. Merci aussi de m'avoir tant aidé dans les analyses pour le RNAseq, d'avoir souvent débloqué mes scripts, pour les relectures et ton soutien ces derniers temps. Fabien (ou Pabien je sais pas trop), merci aussi pour les relectures et d'avoir répondu à toutes mes questions que tu aurais sûrement aussi préféré laisser sous le tapis. Bon courage avec Elie, les complots et matchs de foot. Tu m'as bien impressionné à te passionner autant pour le foot et Lens en moins d'un an. Emilien, merci d'avoir été nul au cookie game et d'avoir vécu avec moi le suspens de fin de Ligue 1 pour savoir si Bordeaux ou Lorient allaient descendre (et pour la victoire 4-1 au Moustoir). Je te souhaite plein de belles choses pour les mois et années à venir. Andreas, le descendant, ça y est tu vas être le plus vieux des jeunes de l'équipe, je te laisse ma mémoire des souches et d'où sont rangés les tubes, fais-en bon usage. Bon courage pour t'en sortir avec toutes ces manips *1000 et les analyses (et aussi pour négocier le robot) !
Omar, ça y est, on y est, on a survécu à nos 4 années de thèse ! Merci d'avoir été un voisin de galère et j'ai hâte de te voir performer au karaoké, peut-être que tu rentreras chez toi dans le mauvais sens et sans ton sac, la boucle sera bouclée. Abhishek, I won't thank you for the journal club but thanks for your incredible memory on articles and for being as desperate as me because of bench troubles, it made me feel less lonely. Claudine, merci pour ces moments partagés dans le bureau en Bota et après dans le labo, je dois t'avouer que j'avais un peu peur de tout ce que tu avais à déménager. Merci aussi de nous avoir laissé gérer les coups de fil de Christine, à notre manière parfois disons-le ! Claudia, je te laisse le Gwenn Ha Du pour représenter la Bretagne fièrement. Merci d'avoir été là, de m'avoir reboosté quand c'était dur, d'avoir compris mon manque de la mer, des galettes et de tout ce qui nous ramène toujours à la maison. On se donne rdv à « Brest même » très vite. J'ai aussi une pensée pour Serge, merci pour ta bienveillance et tes encouragements que ce soit lors de repas d'équipe ou pour OpenLAB. Jing, t'es une parfaite transition

pour boucler la boucle de ces dernières années ! Je suis arrivée, tu partais, tu reviens, je m'en vais… Il doit y avoir un truc là-dessous. Et pourtant c'est bien avec toi et Loulou que je ne connaissais pas du tout que je suis allée boire ma première Haplo bière et j'ai su que cette équipe déchirait ! Merci aussi pour toute l'aide ces 3 derniers mois. Un conseil à tous les autres doctorants, profitez de toute l'aide possible, n'attendez pas vos derniers mois de thèse.

Et là, je vais pleurer en pensant à ceux qui ont été bien plus que des collègues ces dernières années, Téo, Jean-Seb et Sabrina. Merci d'avoir été là pour tous ces micmacs, je ne vous remercie pas d'être tous partis en même temps par contre. Téo, malgré les critères douteux, merci de m'avoir fait confiance pour être ta première stagiaire (enfin je crois… Ça m'a quand même bien cassé les c****) et surtout pour tous les moments passés depuis. Bibi-truc, merci d'être aussi spontanée et optimiste, ne change jamais. J'arrive bientôt te faire un coucou dans ton palais à la campagne où il y a plein de bêtes et de verdure. Jean-Seb, malgré la distance, merci d'avoir été parmi les plus présents ces derniers mois. Et maintenant que je suis une pilote, je t'attends pour prendre ma revanche à vélo, de nuit, le long de la Bruche. Merci aussi aux plus anciens jeunes, Jack, David et Fifou, c'était toujours un plaisir d'aller boire une bière avec vous. Et big up à Alain !
Merci aussi à tous les stagiaires passés dans l'équipe pour leur bonne humeur et particulièrement à Inès, Robin, Tristan et Dung d'avoir participé à mes projets ! Je remercie aussi toutes les personnes de l'unité que j'ai croisé durant ces 5 ans et demi et qui ont d'une manière ou d'une autre contribué à cette thèse. Bon courage aussi à tous les nouveaux qui vont arriver très vite dans l'équipe, Victor, Arthur, Gauthier, Sam…

Parce que cette thèse, ce n'est pas que la vie au labo (ou presque) ! Merci à toute l'équipe d'OpenLAB, les organisateurs, les doctorants et les élèves pour ces 2 années d'interventions tellement enrichissantes qui donnaient une grosse bouffée d'air frais. J'espère que ces interventions reprendront pour les futures générations. Merci particulièrement à toi Lucile, pour ta bienveillance, ton optimiste et tes conseils précieux. Merci aussi au reste de l'équipe du Jardin des Sciences, même si l'année 2020 a été un peu perturbée, c'était un plaisir de vous accompagner dans vos projets.

Merci aussi Nathalie pour ta gentillesse, tes conseils et tous ces repas de fou. Marion, ma Jeanine, que ce soit à Strasbourg, en Irlande ou en Bretagne, je ne garde en tête que de bons souvenirs avec toi, merci d'avoir été là depuis le début. Amélie et

Mickaële, les compatriotes représentantes de l'Ouest dans le master, merci pour les moments passés ensemble même s'ils étaient rares ces derniers temps.

Un gros merci aux copines de Brest pour leur folie et nos retrouvailles au détour d'un coin de la rue Jean Jaurès ou d'un mariage. J'ai hâte au prochain pour vous revoir, ou avant j'espère ?!

Je pense très fort à tous les copains de Locminé aussi qui m'ont encouragé et écouté avec attention quand j'essayais d'expliquer mon sujet de thèse. Bien entendu, les filles, je ne peux pas vous oublier, Maëlle, Anaïs, Marion, Morgane et Manon. Je vous remercie d'être dans ma vie depuis presque 25 ans pour certaines et j'ai si hâte de vous retrouver le plus vite et le plus souvent possible !

Je crois que c'est votre tour les Potos Spaghetto… On dit les meilleurs pour la presque fin non ? Merci d'être là pour le meilleur et pour le rire mais aussi pour tout le reste. Parce que vos petites pépites trouvées sur 9gag font toujours beaucoup de bien. Merci aussi pour tous nos moments ensemble, sur Houseparty ou en vrai, nouvel an et vacances. Hâte de reprendre tout ça le 22/02/22 ou avant !

Bien entendu, je remercie ma famille de m'avoir accompagnée et soutenue jusqu'ici. Tous ces moments à Larmor, Gagny ou à la maison me remettaient toujours sur pieds. Merci aussi à la belle-famille de Bubry pour le soutien et les bons moments passés ensemble. Maman et papa, merci de m'avoir laissé faire mes choix d'orientation pour en être là aujourd'hui. Merci aussi à vous deux pour tous ces bons plats quand je rentrais et pour m'avoir fait aimer cuisiner. Ça m'a bien fait décompresser ces derniers temps. Arnaud, je suis tellement reconnaissante d'avoir pu me retrouver dans la même ville que toi pour traverser ces 6 années ici, heureusement que tu étais là bien souvent (et Mimi aussi). Julien, mon Juju, juste merci d'être toi, maintenant j'espère te voir plus souvent, tu me manques.

Yann, un merci ne serait pas suffisant pour t'exprimer toute ma reconnaissance pour le soutien sans faille que tu m'as apporté ces dernières années et particulièrement ces derniers mois qui ont été si difficiles. Les plus belles choses restent à venir et j'ai hâte de les réaliser avec toi.

Sommaire

ÉTAT DE L'ART

Une *pas si* simple histoire de génétique

Dès l'Antiquité, avec Pythagore, Hippocrate ou encore Aristote, la question de l'hérédité des traits au fil des générations engendre de nombreuses théories. Alors que Pythagore suggère une hérédité paternelle complète quand la mère permet de nourrir l'embryon, Hippocrate propose une hérédité partagée par le père et la mère. Cette hypothèse est suivie par les traités d'Aristote sur l'Histoire des animaux. Aristote va même encore plus loin dans le traité de la Génération des animaux et évoque des phénomènes d'interactions entre les différents acteurs de l'hérédité dans la formation de l'individu. Ce n'est cependant qu'au milieu du XIX^{ème} siècle que Gregor Mendel théorise les lois de l'hérédité (Mendel, 1866), grâce à une démarche scientifique de l'étude de la descendance de croisements de petits pois d'aspects différents (couleur et texture, par exemple). À la même époque, Charles Darwin imagine les « gemmules » logées dans les organes comme particules porteuses de l'hérédité et des caractères acquis (Darwin, 1868). Les lois de l'hérédité de Mendel ne sont considérées par la communauté scientifique que 30 ans après leur découverte et sont adoptées à partir de 1900, notamment par William Bateson. Ce dernier introduit le terme de génétique et décrit des concepts fondamentaux comme celui d'allèles homozygote ou hétérozygote, et également d'épistasie, qui se réfère à l'interaction entre les gènes (Bateson, 1909). De nombreux chercheurs s'impliquent pour définir les premières bases de cette discipline. Aux côtés de Bateson, nous pouvons par exemple citer Sturtevant, Muller, Bridges et Morgan qui démontrent en laboratoire l'existence des mutations et confirment la théorie des chromosomes grâce à l'organisme modèle *Drosophila melanogaster*. En étudiant la variation de la couleur des yeux dans la descendance de mouches du vinaigre, Morgan met notamment en évidence la présence de chromosomes sexuels ainsi que la notion de dominance et de récessivité des allèles (Morgan et al., 1915).

Au sein des individus d'une même espèce, une grande diversité des caractères est observée. Ces caractères, encore appelés traits ou phénotypes, correspondent à toutes les caractéristiques observables ou mesurables d'un individu à différentes échelles : macroscopique, cellulaire ou moléculaire. Certains phénotypes sont dits qualitatifs, comme la couleur des petits pois étudiés par Mendel alors que d'autres sont dits quantitatifs, comme la taille humaine (Figure 1A). La variance phénotypique observée au sein d'une population repose ainsi sur une part génétique, mais aussi sur une part environnementale et sur l'interaction entre ces deux parts (Figure 1B). Si on se focalise sur la part génétique, la variance de cette dernière se décompose en effets

d'additivité, de dominance et d'épistasie (Figure 1B). Dans le cadre des traits quantitatifs, la complexité du phénotype peut être déduite de la distribution des valeurs phénotypiques au sein de la population. En effet, des traits monogéniques, ou mendéliens, c'est-à-dire gouvernés par un seul gène, peuvent être prédits sur la base de la distribution bimodale du phénotype dans la population. À l'inverse, les traits sous l'influence de multiples gènes / allèles, ou traits complexes, présentent une distribution normale des valeurs phénotypiques (Figure 1A).

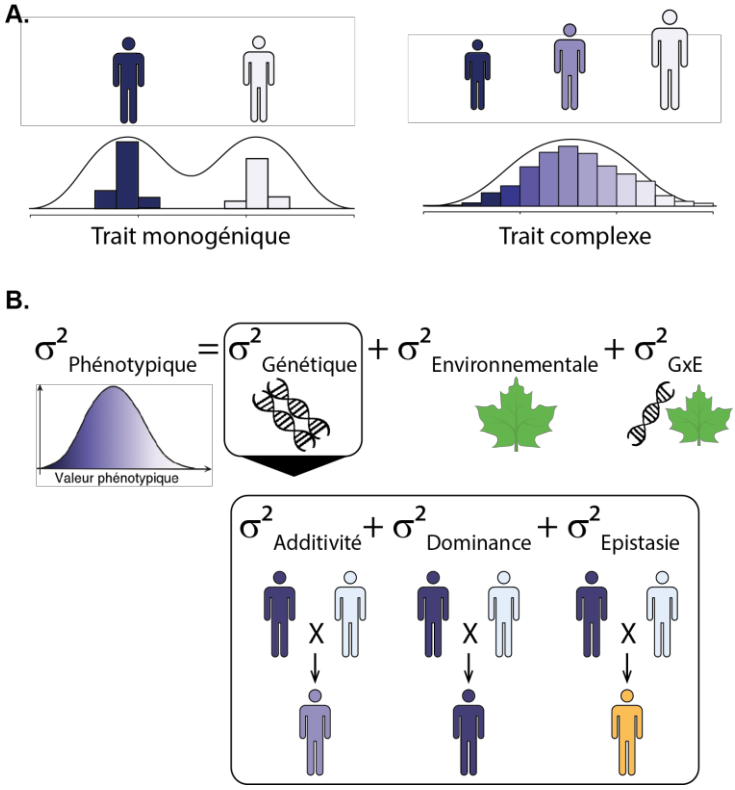


Figure 1. Origine de la variance phénotypique (σ^2_P).
(A.) Distribution de différents types de phénotypes dans une population. Un trait monogénique, ou mendélien, gouverné par un seul gène possède une distribution bimodale. Un trait complexe, gouverné par plusieurs gènes, possède, quant à lui, une distribution normale du phénotype. On peut par exemple citer la taille humaine. (B.) Décomposition de l'origine de la variance phénotypique (σ^2_P) observée dans une population pour un trait donné.

Chez l'Homme, le système ABO des groupes sanguins est un exemple de trait monogénique suivant une hérédité mendélienne (Yamamoto et al., 1990). En effet, la glycosyltransférase produisant l'antigène à la surface des globules rouges peut être codée par 3 allèles d'un même gène situé sur le chromosome 9 (Ferguson-Smith et al., 1976). L'allèle A code l'enzyme produisant l'antigène A ; l'allèle B, l'antigène B. Ces deux allèles étant dominants, les 2 antigènes seront produits chez les individus hétérozygotes, pour lesquels chaque parent aura transmis un des deux allèles. L'allèle O est quant à lui récessif et silencieux, c'est-à-dire qu'aucun antigène n'est produit dans un contexte homozygote où seul cet allèle est présent. La mucoviscidose ou encore la neurofibromatose de type 1 sont des maladies génétiques humaines qui suivent ce type d'hérédité, où la présence de mutations délétères dans un seul gène, *CFTR* ou *NF1* respectivement, engendre la maladie (Pasmant et al., 2012; Ratjen et al., 2015). À l'heure actuelle, plus de 5000 maladies monogéniques sont associées à leur gène causal (https://www.omim.org/).

De nombreuses maladies humaines telles que l'autisme, la schizophrénie ou encore la maladie d'Alzheimer ont une origine complexe ou multigénique (Lichtenstein et al., 2009; Nikolac Perkovic and Pivac, 2019; Ramaswami and Geschwind, 2018). Élucider les origines génétiques des traits complexes reste encore aujourd'hui une problématique fondamentale, les gènes causaux étant bien plus difficiles à cartographier. Au cours des 10 dernières années, le nombre de variants génétiques associés à la variance de la taille humaine est par exemple passé de 40 à plusieurs milliers (Akiyama et al., 2019; Manolio et al., 2009; Wood et al., 2014; Yengo et al., 2018). Cependant, ces milliers de variants génétiques n'expliquent qu'une partie relativement faible (à savoir 25%) de la variance phénotypique de ce trait dans la population étudiée (Yengo et al., 2018).

Le terme d'héritabilité dite au sens large (H^2) est utilisé pour définir la proportion de la variance phénotypique dans la population étudiée qui repose sur des facteurs génétiques. L'héritabilité au sens strict (h^2) correspond à la contribution des effets génétiques uniquement additifs à la variance phénotypique. En plus de l'additivité, des effets de dominance et d'épistasie sont également responsables de la variance génétique à l'origine de la variabilité des traits. Des travaux récents chez la levure *Saccharomyces cerevisiae* ont permis d'estimer que la variance phénotypique est en moyenne expliquée à 55% par des effets additifs et à 29% par des effets non additifs (Bloom et al., 2013; Fournier et al., 2019).

Études des variations du génome au sein d'une population

Afin de déterminer les bases génétiques à l'origine de la variance phénotypique observée entre les individus d'une même espèce, une comparaison des traits d'une part et des variants génétiques d'autre part est indispensable. Afin d'obtenir la vue la plus exhaustive possible des variants génétiques qui pourraient expliquer la variance phénotypique, il est nécessaire de disposer des séquences nucléotidiques des génomes de grandes cohortes d'individus.
Dans cette partie, nous décrirons les avancées technologiques ayant permis d'établir les catalogues des variants génétiques au sein des populations. Nous discuterons ensuite de l'histoire évolutive des espèces à partir de la diversité et de la structure des génomes au sein d'une population. Enfin, nous verrons comment des sous-populations ont génétiquement évolué au fil des générations.

Le séquençage et ses avancées au cours des dernières décennies

Les nombreuses avancées dans le domaine de la génétique au XX^{ème} siècle valident la théorie des chromosomes comme porteurs de l'hérédité. La molécule support de ces chromosomes est identifiée comme de l'ADN (acide désoxyribonucléique) et sa structure est décrite en 1953 par Watson, Crick, Franklin et Wilkins (Watson and Crick, 1953). L'ADN est un polymère constitué d'une succession de nucléotides composés d'une base nucléique sous 4 formes possibles : l'adénine, la thymine, la cytosine ou la guanine. Cette découverte révolutionne la discipline et très vite, le passage de l'ADN en protéine *via* l'ARNm est élucidé (Jacob and Monod, 1961) ainsi que le code génétique (Martin et al., 1961). Ce dernier permet de décrypter la correspondance entre la succession de nucléotides et d'acides aminés d'une protéine. Il devient ainsi essentiel de pouvoir déterminer l'enchaînement des nucléotides dans les génomes pour comprendre la diversité des phénotypes.

Le séquençage de 1ère génération

Une vingtaine d'années après la caractérisation de la structure de l'ADN, les premières stratégies de séquençage sont développées en parallèle par Frederick Sanger (Sanger et al., 1977) et par Allan Maxam et Walter Gilbert (Maxam and Gilbert, 1977) (Figure 2A). Cette dernière consiste à cliver une molécule d'ADN par différents réactifs chimiques selon la base nucléotidique (ou combinaison de bases), puis à détecter ces bases grâce à un marquage radioactif en 5' permettant de

reconstituer la séquence (Maxam and Gilbert, 1977). La stratégie de Sanger repose, quant à elle, sur l'incorporation de dNTP (désoxyribonucléotide) ou de ddNTP (didésoxyribonucléotide) lors de la polymérisation de l'ADN. Les ddNTP bloquant l'élongation, des fragments de toutes les tailles sont alors produits. Des marqueurs, radioactifs à l'époque, fluorescents aujourd'hui, permettent la lecture du ddNTP terminateur de la chaine d'ADN et donc de déduire la succession des nucléotides le long du fragment d'intérêt (Sanger et al., 1977). Cette stratégie a permis de séquencer dès 1977 le premier génome complet : celui du virus bactériophage φX174, constitué de 5386 nucléotides (Sanger et al., 1977, 1978). Vingt ans plus tard, en 1996, la levure *Saccharomyces cerevisiae,* dont le génome est composé de 12 Mb, est le premier organisme eucaryote entièrement séquencé grâce à une collaboration internationale d'environ 600 chercheurs (Dujon, 2019; Goffeau et al., 1996). Les génomes de nombreux autres organismes modèles tels que *Escherichia coli* (Blattner et al., 1997), *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998), *Drosophila melanogaster* (Adams et al., 2000), *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000) ou *Mus musculus* (Waterston et al., 2002) sont ensuite séquencés et publiés. Malgré l'ampleur de la tâche, le projet de séquençage du génome humain, constitué d'environ 3 Gb, est initié en 1985 et donne lieu à une première publication en 2001 (Lander et al., 2001; Venter et al., 2001). Cette période verra également l'émergence de la génomique comparative, qui repose sur la comparaison des génomes de diverses espèces évolutivement éloignées (Rubin et al., 2000) ou évolutivement plus proches, au sein d'un même phylum, comme celui des levures (Dujon et al., 2004; Souciet et al., 2000). Au vu de l'ampleur et des coûts induits par les projets de séquençage à cette époque, seul un génome, dit alors « de référence », est assemblé par espèce, prévenant à ce niveau les études comparatives entre individus à l'échelle du génome complet.

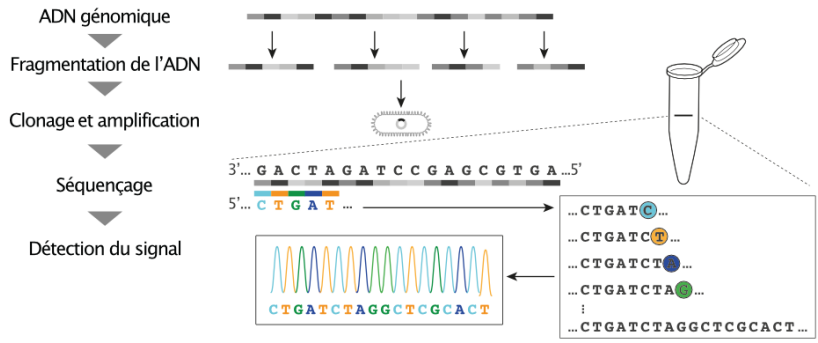
Le séquençage de 2ème génération

L'émergence, au début des années 2000, des techniques de séquençage dites de 2ème génération (Figure 2B) a permis d'augmenter le débit tout en réduisant les coûts de séquençage, induisant un changement d'échelle dans ce domaine (Shendure and Ji, 2008; Wheeler et al., 2008). Parmi ces techniques à haut-débit, le séquençage de type Illumina / Solexa repose sur une amplification préalable de courts fragments d'ADN puis d'une synthèse des brins dénaturés à partir de nucléotides marqués émettant un signal fluorescent. Plusieurs centaines de millions de lectures peuvent être séquencées en parallèle. Des projets basés sur le séquençage de plusieurs individus

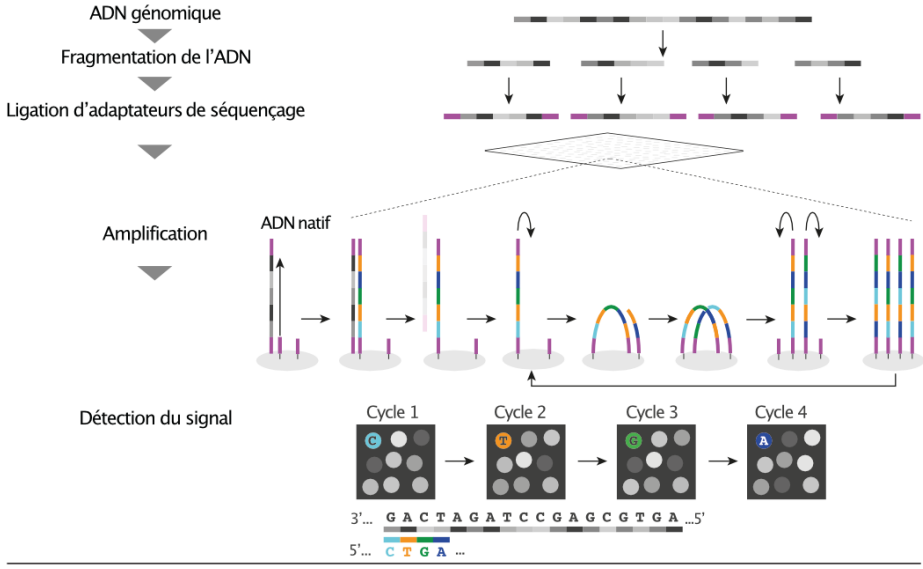
d'une même espèce sont alors initiés, permettant de poser les bases de la génomique des populations décrites quelques années auparavant (Gulcher and Stefansson, 1998). Rapidement, les échelles augmentent et le millier d'individus dont le génome est séquencé est atteint pour plusieurs espèces dont l'Homme (Auton et al., 2015), la plante *A. thaliana* (Alonso-Blanco et al., 2016) ou encore la levure *S. cerevisiae* (Peter et al., 2018). Le séquençage de 2ème génération présente des avantages majeurs tels que le haut-débit et un coût réduit par génome séquencé. Cependant, la petite taille des lectures générées, au maximum 300 paires de bases (pb), rend difficile les assemblages de génome *de novo*. L'alignement de ces lectures sur un génome de référence représente une stratégie de choix pour la détection efficace de variants génétiques courts comme les SNP (pour Single-Nucleotide Polymorphisms) et les indels (pour insertion / délétions). La détection des variants plus longs comme par exemple les grands réarrangements chromosomiques est néanmoins limitée par ces stratégies.

Figure 2. Évolution et principe des stratégies de séquençage.
(A.) Le séquençage de 1ère génération de type Sanger repose la synthèse d'un brin d'ADN à partir d'une matrice en incorporant des dNTP (désoxyribonucléotides) ou des ddNTP (didésoxyribonucléotides) pour soit continuer la synthèse soit la bloquer. Une lecture des ddNTP terminaux permet de reconstituer la séquence. (B.) La technologie de séquençage Illumina repose sur une amplification préalable de courts fragments d'ADN puis d'une synthèse des brins dénaturés à partir de nucléotides marqués émettant un signal fluorescent. (C.) Le séquençage de type Oxford Nanopore repose sur le passage de la molécule d'ADN natif à travers un nanopore biologique permettant l'émission d'un courant électrique variable selon les nucléotides le traversant. Figure adaptée de Shendure et al. (Shendure et al., 2017).

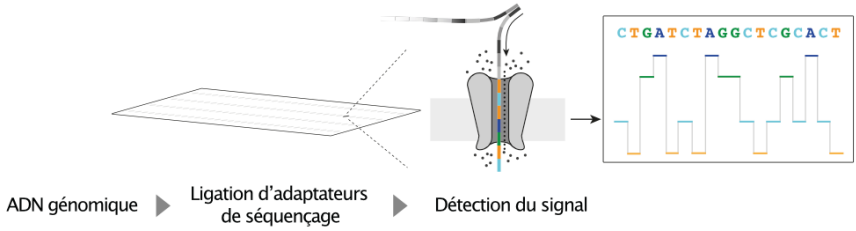
A. Le séquençage de 1ère génération (Sanger)



B. Le séquençage de 2ème génération (Illumina)



C. Le séquençage de 3ème génération (Oxford Nanopore)



Le séquençage de 3ème génération

En 2010, l'arrivée du séquençage de 3ème génération (Figure 2C) permet de générer des lectures longues, de plusieurs kilobases (kb) (Wenger et al., 2019) à plusieurs mégabases (Mb) (Payne et al., 2019) avec principalement deux technologies : Pacific Biosciences (PacBio) et Oxford Nanopore (Schadt et al., 2010). Aucune amplification préalable de l'ADN n'est nécessaire pour ces 2 méthodes. L'ADN génomique peut directement être fragmenté à la longueur voulue et préparé pour le séquençage. La technologie PacBio est basée sur un support de séquençage permettant la détection de signal fluorescent en temps réel lors de l'incorporation d'un nucléotide. Ce support de séquençage permet de séparer chaque molécule d'ADN dans un puit dans lequel une polymérase est immobilisée pour synthétiser l'ADN à partir d'un simple brin (Eid et al., 2009; Rhoads and Au, 2015). La technologie Oxford Nanopore repose quant à elle sur le passage de la molécule d'ADN à travers un nanopore biologique permettant l'émission d'un courant électrique variable selon les nucléotides qui traversent le pore (Branton et al., 2009; Jain et al., 2016). Malgré les améliorations constantes de ces technologies, le taux d'erreur de séquençage reste encore important, de l'ordre de 5%. Quoiqu'il en soit, la longueur des lectures fournies par ces nouvelles stratégies permettent de générer des assemblages de génome de bonne qualité, pouvant atteindre une résolution au niveau chromosomique. À partir de ces assemblages, la détection de longs réarrangements chromosomiques dans les génomes est alors facilitée. Ces technologies représentent une perspective intéressante dans l'étude des génomes à l'échelle de populations. Dans ce but, différents protocoles couplant les stratégies de 2ème et 3ème génération sont couramment mis en place et permettent de corriger les erreurs de séquençage retrouvées dans les lectures longues (De Coster et al., 2021). Par exemple, 20 ans après le premier génome humain séquencé, le couplage des technologies Illumina, Oxford Nanopore et PacBio a permis d'atteindre un nouveau niveau de résolution en séquençant l'intégralité du génome de 3 Gb incluant les régions d'hétérochromatine ainsi que les régions structurellement complexes (Nurk et al., 2021).

Variabilité des génomes au sein des populations

En 1998, Gulcher introduit la notion de « génomique des populations » en recherchant les causes génétiques de maladies complexes au sein d'une population islandaise. À partir de populations vivante et ancestrale de 600 000 islandais dont les relations généalogiques sont connues, l'objectif consiste à associer 2000 à 3000 variants nucléotidiques et 600 à 1000 marqueurs microsatellites avec les informations médicales connues dans la population (Gulcher and Stefansson, 1998). Vingt ans plus tard, les génomes d'un grand nombre d'individus de diverses espèces ont été séquencés grâce aux avancées technologiques considérables. Ces projets de grande ampleur ont rendu la génomique des populations incontournable dans l'exploration de la construction des génomes et de la diversité génétique intra-spécifique. En effet, ces projets de séquençage massif de génomes ont permis d'établir des catalogues de variants génétiques au sein de plusieurs espèces comme l'Homme (Auton et al., 2015) ou d'autres espèces modèles (Alonso-Blanco et al., 2016; Cook et al., 2017; Peter et al., 2018). Chez l'Homme, le séquençage du génome de 2504 individus issus de 26 populations a permis d'identifier plus de 88 millions de variants génétiques de différents types : 84,7 millions de sites de polymorphisme nucléotidique (SNP), 3,6 millions de courtes indels et 60 000 variants structurels (Auton et al., 2015) (Figure 3). La constitution de ces catalogues de variants génétiques permet d'envisager d'associer ces variants à leurs conséquences phénotypiques et de mieux appréhender les mécanismes évolutifs.

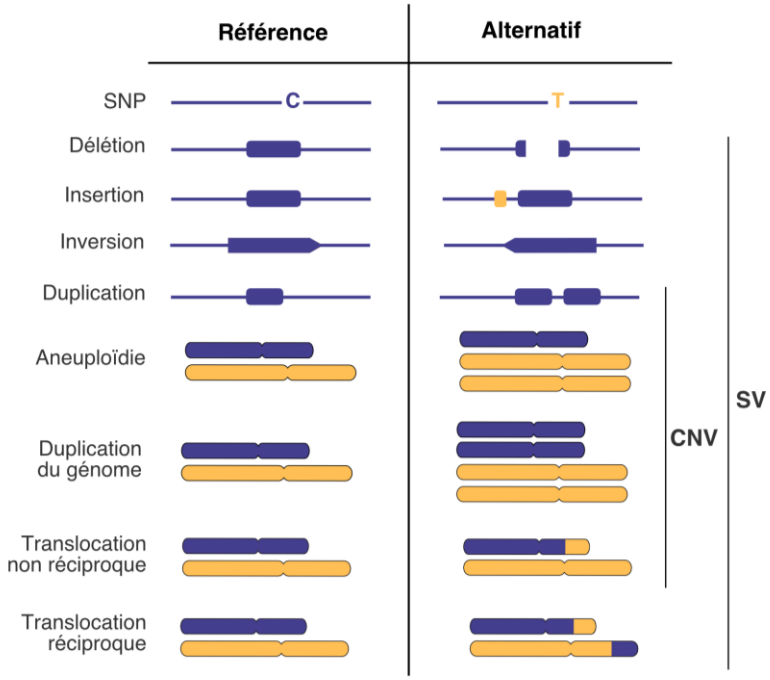


Figure 3. Description des types de variants génétiques.
Vue d'ensemble schématique des variants génétiques retrouvés dans les génomes comprenant les sites de polymorphismes nucléotidiques (SNP) ainsi que les variants structurels (SV) avec ou sans changement du nombre de copies (CNV pour Variant du Nombre de Copies).

Le polymorphisme nucléotidique

Parmi les variants génétiques, les SNP, qui correspondent au changement d'un seul nucléotide (Figure 3), sont les plus fréquents. Ces mutations peuvent être induites par des agents mutagènes extérieurs, comme les UV, ou par des erreurs de réplication ou de réparation de l'ADN. Chez l'Homme, un génome contient de 4,1 à 5 millions de SNP par rapport au génome de référence (Auton et al., 2015), ce qui correspond environ à une mutation toutes les 700 pb. La diversité nucléotidique est très variable entre les différentes espèces. À titre d'exemple, chez la levure *S. cerevisiae*, 1,6 millions de SNP sont retrouvés dans une population de 1011 isolats naturels et en moyenne, nous retrouvons une mutation toutes les 200 pb (Peter et al., 2018). Les conséquences de ce type de mutations dépendent d'une part de la localisation du résidu affecté et notamment s'il se situe dans une région codante (CDS pour Coding DNA Sequence), dans le promoteur d'un gène ou encore dans les régions

intergéniques. Une mutation est dite silencieuse si elle ne modifie pas la séquence protéique associée au gène, en se situant soit en dehors de la CDS ou en ne modifiant pas l'acide aminé grâce à la redondance du code génétique, on parle dans ce cas de mutation synonyme. À l'inverse, les mutations non-synonymes présentes dans la CDS peuvent être de deux types. Les mutations non-sens conduisent à l'apparition d'un codon stop et sont à l'origine de la production de protéines tronquées. Un changement d'acide aminé par un autre dans la séquence résulte d'une mutation faux-sens et peut avoir des conséquences diverses, selon la localisation du résidu affecté et la nature du changement. Les variations neutres n'affecteront pas la fonction de la protéine alors que les variations délétères induiront une altération de cette dernière. Dans le cas de la drépanocytose par exemple, une maladie génétique humaine, une mutation faux-sens dans le gène codant la β-globine entraîne le remplacement d'un acide glutamique par une valine (Herrick, 1910; Rees et al., 2010). Cette mutation entraîne une malformation des globules rouges et est associé à un grand panel de symptômes : de l'anémie à la défaillance d'organes. Des outils bio-informatiques, tels que SIFT (Sorting Tolerant From Intolerant) un algorithme basé sur les similarités de séquences et les propriétés chimiques des acides aminés (Kumar et al., 2009), permettent aujourd'hui de prédire les conséquences des différentes mutations sur la protéine, sa structure, sa fonction. Chez la levure *S. cerevisiae*, il a par exemple été décrit que parmi les 1,6 millions de SNP, les variants les plus rares dans la population étaient ceux prédits comme les plus délétères (Peter et al., 2018). Les mutations dans les régions intergéniques peuvent également avoir un impact phénotypique, notamment sur l'expression des gènes. En effet, l'expression des gènes peut être influencée par des variants génétiques présents dans la CDS mais aussi dans des régions proches (Figure 4), on parle alors de variants régulateurs locaux ou *cis*- (Albert and Kruglyak, 2015; Hill et al., 2021). Ces variants locaux peuvent affecter le promoteur direct (ou core promoter) (Lubliner et al., 2015; Tirosh et al., 2009) ainsi que les régions régulatrices adjacentes comme les sites de liaison des facteurs de transcription, les amplificateurs (ou enhancers) ou les régions terminatrices (Andersson and Sandelin, 2020; Hill et al., 2021; Wittkopp and Kalay, 2011). L'accessibilité de la chromatine autour du site d'initiation de la transcription peut également être modifiée et ainsi faire varier l'expression des gènes (Field et al., 2009; Hill et al., 2021; Lickwar et al., 2012). Des variants régulateurs distants ou *trans*- influent également sur le niveau d'expression des gènes en impactant directement des facteurs de transcription ou des protéines impliquées dans la signalisation de la régulation par exemple (Lutz et al., 2019) (Figure 4). De manière générale, les variants régulateurs distants sont majoritaires et ont des effets plus

pléiotropiques que les variants locaux (Albert et al., 2018; Hill et al., 2021; Signor and Nuzhdin, 2018; Wittkopp, 2005). Cependant, les variants locaux contribuent de façon plus importante à la variance phénotypique (Albert et al., 2018; Hill et al., 2021) en raison d'effets plus ciblés sur la transcription (Coolon et al., 2015; Emerson et al., 2010). L'analyse de l'expression du gène *TDH3* dans une population de plus de 50 isolats de la levure *S. cerevisiae* a récemment mis en évidence l'origine polygénique de l'expression du gène avec jusqu'à 100 variants régulateurs identifiés (Metzger and Wittkopp, 2019).

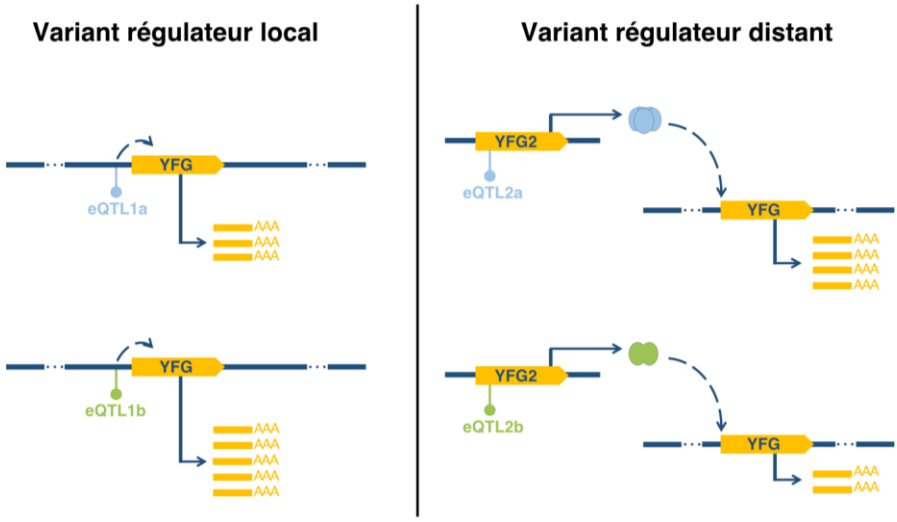


Figure 4. Classification des variants génétiques régulateurs.
Les variants génétiques régulant localement l'expression des gènes (gauche), ou *cis*-eQTL (pour expression Quantitative Trait Loci), ont un impact direct sur le niveau d'expression et peuvent par exemple être localisés dans le promoteur du gène. Les variants régulateurs distants (droite), ou *trans*-eQTL, sont situés dans une région éloignée par rapport au gène influencé, voire sur un autre chromosome. Ces *trans*-eQTL peuvent par exemple engendrer une modification d'un facteur de transcription régulant par la suite le niveau d'expression du gène considéré.

Les variants structurels

Différents types de variants structurels (SV pour Structural Variant) s'ajoutent aux SNP et participent activement à la variabilité des génomes (Figure 3). Parmi ces SV, on retrouve des remaniements réciproques, comme les inversions et les translocations réciproques qui ne modifient pas le nombre de copies dans le génome, et les remaniements non-réciproques, comme les translocations non réciproques et les variants du nombre de copies (CNV pour Copy Number Variant) : duplications,

insertions et délétions de tailles variables. Ces remaniements du génome sont issus d'erreurs de réparation de l'ADN ou de ségrégation des chromosomes mais sont aussi une réponse efficace d'adaptation à un stress ou à un processus de domestication industrielle (Gorkovskiy and Verstrepen, 2021). Par exemple, des isolats de *S. cerevisiae* impliqués dans les processus de vinification présentent des SV réciproques (translocations ou inversion) particulièrement conservés pour leurs intérêts industriels. En effet, certains de ces SV induisent la surexpression du gène *SSU1* impliqué dans la tolérance au sulfite, un composé utilisé pour son rôle antioxydant (Marullo et al., 2020; Pérez-Ortín et al., 2002). L'émergence du séquençage de lectures longues a largement contribué à la détection de ces variants structurels (SV) qui était jusqu'alors compliquée, en particulier pour les SV réciproques (De Coster et al., 2019). Des études plus larges des SV à l'échelle d'une population ont pu voir le jour chez l'Homme (Beyter et al., 2021) ou encore dans plus de 3000 génomes de riz (Fuentes et al., 2019).

Parmi les différents types de SV, les origines et les impacts des CNV ont été particulièrement caractérisés dans les génomes. Différentes expériences d'évolution chez la levure *S. cerevisiae* ont notamment révélé une propagation des CNV sans pression de sélection dans des contextes de compétition (Lauer et al., 2018; Payen et al., 2014; Thierry et al., 2015, 2016). La taille des régions affectées par ces CNV peut concerner le génome entier, un ou plusieurs chromosomes ou gènes ou seulement quelques paires de bases. Chez les plantes ou les levures par exemple, les duplications complètes du génome (WGD pour Whole-Genome Duplication) sont considérées comme un mécanisme évolutif clé (Albertin and Marullo, 2012; Ohno, 1970; Otto, 2007). De plus, au sein d'une même espèce, il est possible d'observer d'importantes variations de la ploïdie. Par exemple, dans la population des 1011 isolats de *S. cerevisiae*, 87% de la population naturelle est diploïde. Des isolats polyploïdes (3n, 4n et 5n) ont également été identifiés et représentent 11,5% de la population (Peter et al., 2018). Malgré un impact négatif de la polyploïdie sur le fitness des individus dans des conditions de laboratoire (Figure 5A), ces individus polyploïdes ont été spécifiquement isolés dans des environnements liés à l'activité humaine comme la production de bière, de vin de palme ou en boulangerie (Peter et al., 2018). Les aneuploïdies – modification du nombre d'un ou plusieurs chromosomes – provoquent également une diminution du fitness chez la levure *S. cerevisiae* (Figure 5B). De manière intéressante, le nombre d'aneuploïdies retrouvé dans la population par chromosome est inversement corrélé à la taille du chromosome concerné (Figure 5C) et suggère un effet plus délétère en présence d'un nombre plus grand de gènes dupliqués (Peter et al., 2018).

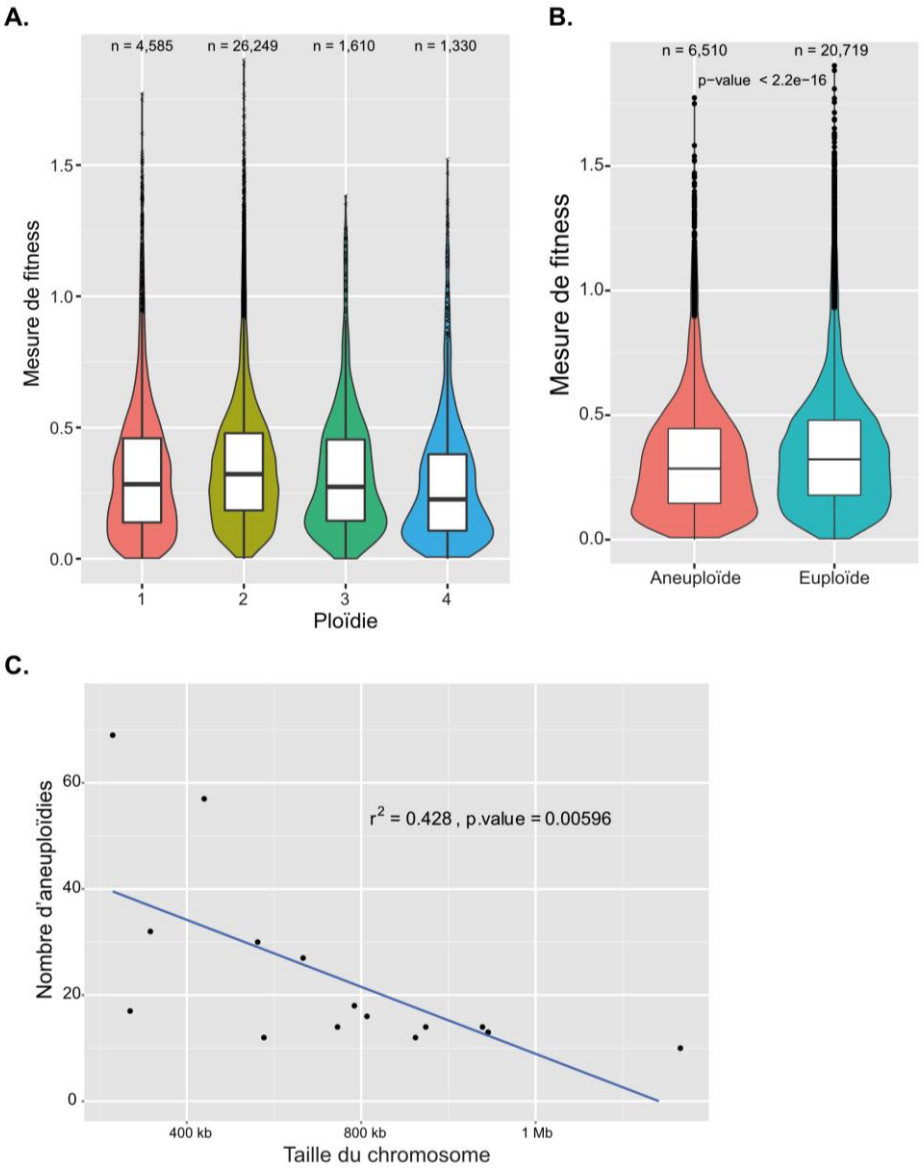


Figure 5. Variation de la ploïdie et des aneuploïdies dans une population de levures. (A.) Distribution des valeurs de croissance des isolats selon leur ploïdie dans 36 conditions de culture. Les isolats diploïdes possèdent une mesure de fitness globalement supérieure aux autres ploïdies. (B.) Distribution des valeurs de croissance des isolats dans 36 conditions de culture selon la présence ou non d'aneuploïdies dans leur génome. Un impact négatif sur le fitness des isolats est observé en cas d'aneuploïdie. (C.) Corrélation négative entre le nombre d'aneuploïdies détectées dans la population et la taille du chromosome concerné. Figures adaptées de Peter et al., 2018.

Une des seules aneuploïdies viables chez l'Homme affecte le chromosome 21, un des plus courts du génome humain. Dans le cas de la trisomie 21 ou syndrome de Down, le surdosage des gènes du chromosome 21 entraîne de nombreuses variations phénotypiques (Antonarakis, 2017; Patterson, 2009). De nombreux mécanismes régulant l'expression des gènes à différents niveaux sont cependant mis en place pour compenser le surdosage (Hose et al., 2015; Liu et al., 2017). Les mécanismes de compensation / tolérance à l'échelle chromosomique sont encore flous et des questions se posent également sur l'impact des CNV à l'échelle du gène. En effet, les CNV sont très répandus dans les génomes et concernent un grand nombre de gènes. Chez l'Homme, 9,5% du génome serait impliqué dans un gain ou une perte de copies (Zarrei et al., 2015). Dans la population de 1011 isolats de *S. cerevisiae*, presque chacun des ~ 6000 gènes est dupliqué ou supprimé dans au moins une souche avec une fréquence plus grande pour les gènes localisés dans les régions subtélomériques (Peter et al., 2018). Les conséquences de ces variants du nombre de copies peuvent être plus ou moins délétères. Chez les bactéries et les levures, les CNV sont fréquemment associés à des mécanismes d'adaptation à l'environnement, comme la résistance aux antibiotiques, aux antifongiques ou à d'autres composés chimiques (Fogel and Welch, 1982; Sandegren and Andersson, 2009; Soo et al., 2011; Todd and Selmecki, 2020). Chez l'Homme, les CNV sont largement impliqués dans les cancers, que ce soit comme déclencheurs de tumeur ou en conséquence du processus de tumorigenèse (Beroukhim et al., 2010). En plus des évènements de duplication ou de délétion de gènes, des insertions peuvent également avoir lieu. Elles peuvent concerner des éléments transposables ou être associées à des introgressions ou des transferts horizontaux de gènes (HGT pour Horizontal Gene Transfer).

Le pangénome

L'étude des variations génétiques au sein de populations naturelles a permis de mettre en évidence la variabilité du contenu en gènes entre individus appartenant à une même espèce. Ces observations introduisent pour la première fois la notion de pangénome chez la bactérie *Streptococcus agalactiae* (Tettelin et al., 2005). Le pangénome est constitué du génome de base (ou core genome), comprenant les gènes présents dans tous les individus d'une espèce, et du génome accessoire, c'est-à-dire des gènes dont la présence varie entre individus d'une population. Certains gènes peuvent être présents dans un seul individu de la population, faisant extrêmement varier les génomes. Ces gènes uniques peuvent représenter de 20 à 40% du

pangénome de certaines espèces bactériennes (Zou et al., 2019). Dans les génomes plus complexes majoritairement composés de régions non-codantes, la définition du pangénome est plus délicate et concerne aussi la variabilité du contenu en exons ou encore des régions intergéniques non-codantes. Ainsi, de nombreux consortiums incluant le séquençage de différentes populations participent à l'effort pour définir le pangénome humain (Sherman and Salzberg, 2020). L'étude des pangénomes peut avoir des intérêts biotechnologiques pour certaines espèces, comme celui des plantes utilisées en agriculture notamment. En effet, des génomes contenant certains gènes accessoires sont sélectionnés pour des phénotypes spécifiques à une culture (Bayer et al., 2020; Golicz et al., 2016; Tao et al., 2019). Dans différentes espèces de levures, une diversité importante des gènes accessoires est retrouvée bien que le génome de base constitue entre 75 et 90% des génomes (Mccarthy and Fitzpatrick, 2019; Peter et al., 2018). De manière générale, à l'échelle du génome de *S. cerevisiae*, les gènes accessoires ont une variabilité nucléotidique plus importante et subissent une pression de sélection moins importante que le génome de base (Peter et al., 2018). Parmi les gènes accessoires, on retrouve différents types d'insertions comme les HGT ou les introgressions. Ces introgressions proviennent de transfert d'un ou plusieurs gènes suite à un évènement d'hybridation entre espèces proches. Chez l'Homme, des introgressions provenant d'espèces archaïques Néandertaliennes ou Dénisoviennes sont retrouvées dans certaines sous-populations et peuvent impacter les phénotypes (Bergström et al., 2021; McCoy et al., 2017; Skov et al., 2020). La régulation de ces gènes est également spécifique par rapport au reste du génome. Une régulation locale de l'expression des gènes est prédominante indiquant la conservation des SNP régulateurs au cours des générations (McCoy et al., 2017).

Histoire évolutive des génomes

L'ensemble des variants génétiques identifiés grâce aux nombreux projets de séquençage est un excellent marqueur de l'histoire évolutive des espèces. Sur la base de ces variants, la divergence génétique entre chaque individu peut être estimée et regroupée dans une matrice de distances génétiques. Cette matrice permet alors la construction d'arbres de type neighbour joining. Des sous-groupes d'individus – ou sous-populations – peuvent y être identifiés et rapportés aux origines géographiques, écologiques ou des évènements de domestication / sélection. L'étude de la diversité génétique au sein de ces sous-populations et entre les différents sous-groupes permet d'obtenir des informations sur les variants génétiques uniques ou partagés entre les individus et de dater ces évènements. La recherche des ancêtres communs est aussi fondamentale pour retracer l'origine des génomes, soit par le séquençage de génomes ancestraux fossiles quand cela est possible, soit par l'analyse d'espèces proches révélant des évènements d'introgression par exemple.

De nombreux scénarios ont été proposés ces 20 dernières années sur l'origine et la construction du génome de l'Homme moderne avec le séquençage de plusieurs milliers d'individus. L'analyse des 1000 premiers génomes humains séquencés a mis en évidence une diversité génétique plus importante dans les sous-populations africaines (Figure 6A) (Auton et al., 2015). Ces observations corroborent les différentes hypothèses suggérant l'origine de l'Homme moderne en Afrique il y a 0,3 à 1 million d'années (Bergström et al., 2021). Des phases de dispersion, d'abord en Afrique, puis sur les autres continents ont suivi et créé des goulots d'étranglement génétiques dans les autres sous-populations expliquant la plus faible diversité génétique (Auton et al., 2015; Bergström et al., 2021). De plus, l'accès aux génomes d'individus archaïques fossiles (Hommes de Néandertal et Dénisova) datant de plusieurs dizaines, voire centaines de milliers d'années a permis de retracer encore plus efficacement l'histoire évolutive humaine. En effet, le génome de certaines sous-populations contient des fragments génomiques de ces espèces archaïques. D'un côté, les sous-populations eurasiennes contiennent environ 2% d'ADN néandertalien (Green et al., 2010; Sankararaman et al., 2012; Yang et al., 2012) alors que les populations d'Asie du Sud / Océanie ou d'Asie de l'Est comportent respectivement en moyenne 3 et 0,1% d'ADN dénisovien du Sud ou de Sibérie (Bergström et al., 2020; Browning et al., 2018; Massilani et al., 2020; Prüfer et al., 2017; Reich et al., 2010). Ces évènements d'introgression suggèrent ainsi une cohabitation et des croisements entre l'Homme moderne et des espèces archaïques

jusqu'à il y a -40/-60 000 ans (Bergström et al., 2021). Bien que ces fragments de génome aient été conservés au cours des générations, ils ont subi une forte sélection négative avec de nombreuses mutations provoquant des pertes de fonction des allèles introgressés (Harris and Nielsen, 2016; Juric et al., 2016; Petr et al., 2019; Sankararaman et al., 2016).

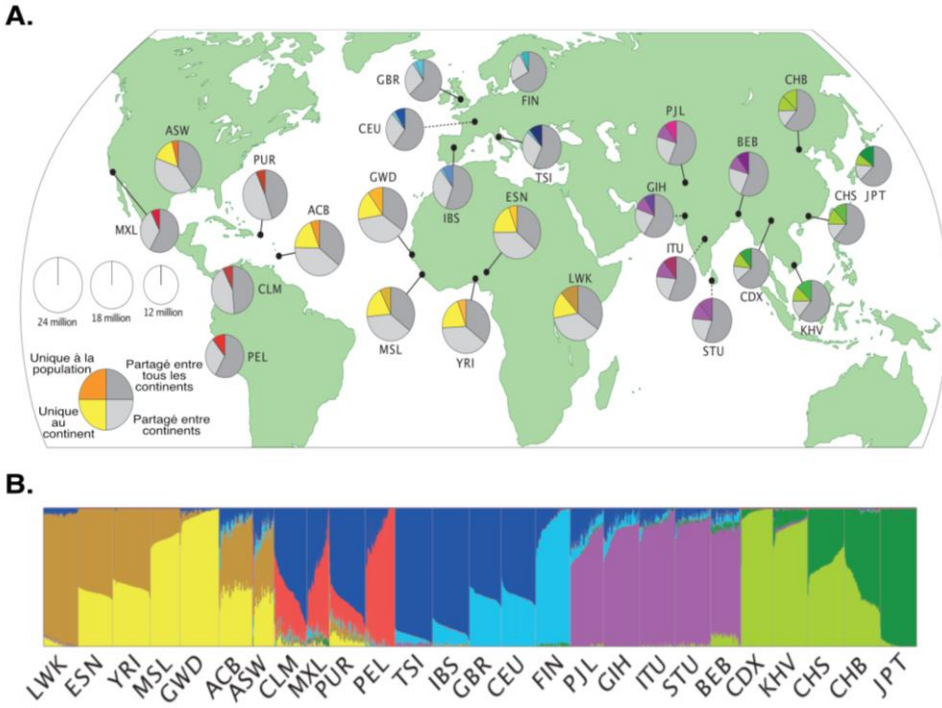


Figure 6. Distribution et structure de la population humaine.
(A.) Distribution des sites polymorphiques au sein de sous-groupes de la population humaine mondiale. Les diagrammes circulaires représentent les proportions de variants partagés entre les sous-populations. (B.) Structure de la population humaine basée sur une stratégie de calcul de maximum de similarité en inférant un nombre de sous-populations égal à 8. Figures adaptées de Auton et al., 2015.

Parmi les différentes sous-populations décrites, une structure importante se dégage avec des génomes ancestraux bien spécifiques pour chaque continent, à l'exception des sous-populations américaines qui présentent des mélanges plus importants et des génomes dits mosaïques (Figure 6B) (Auton et al., 2015).

De manière équivalente, une structure importante s'est dégagée de l'étude de 1011 génomes de la levure *S. cerevisiae* (Peter et al., 2018). En effet, 26 sous-populations – ou clades – ont été définies sur la base de la diversité nucléotidique entre isolats. Une corrélation importante est retrouvée entre ces clades et les origines géographiques, écologiques et les évènements de domestication. Cependant, 3 sous-groupes d'isolats ne présentent pas de structure bien définie. On y retrouve des souches d'origines écologiques ou géographiques variées constituées d'une mosaïque de génomes ancestraux (Figure 7A) (Peter et al., 2018). La diversité génétique intraspécifique importante (avec un maximum de 1,83% de divergence) et l'impact de l'Homme sur la sélection des isolats dans les processus industriels complexifient l'étude de l'histoire évolutive des génomes au sein de cette espèce. De plus, jusqu'à récemment, aucun ancêtre commun n'avait été identifié pour cette espèce. La découverte d'isolats provenant de Chine avec une importante divergence génétique par rapport au reste de la population suggère une origine asiatique de l'espèce. En cohérence avec cette hypothèse, des espèces proches de *S. cerevisiae*, telles que *Saccharomyces mikatae* ou *Saccharomyces arboricola*, sont elles-mêmes originaires d'Asie de l'Est (Figure 7B) (Duan et al., 2018; Peter et al., 2018). Comme dans le génome humain, certains clades contiennent un nombre important d'introgressions provenant notamment de l'espèce proche *Saccharomyces paradoxus*, avec un maximum de 300 ORF (pour Open Reading Frames) par génome. Un hybride entre les espèces *S. cerevisiae* et *S. paradoxus*, a récemment été étudié et permet de mieux comprendre l'origine des introgressions (D'Angiolo et al., 2020). Le génome de cet hybride a en effet révélé de larges régions de perte d'hétérozygotie (LOH pour Loss of Heterozygosity) en faveur de la version *S. paradoxus* pour environ 9,7% du génome (D'Angiolo et al., 2020). L'instabilité du génome de l'hybride conduisant à la formation de ces évènements de LOH favorise la recombinaison méiotique et mitotique, restaurant ainsi la fertilité de l'hybride et sa capacité à se croiser avec d'autres individus (D'Angiolo et al., 2020).

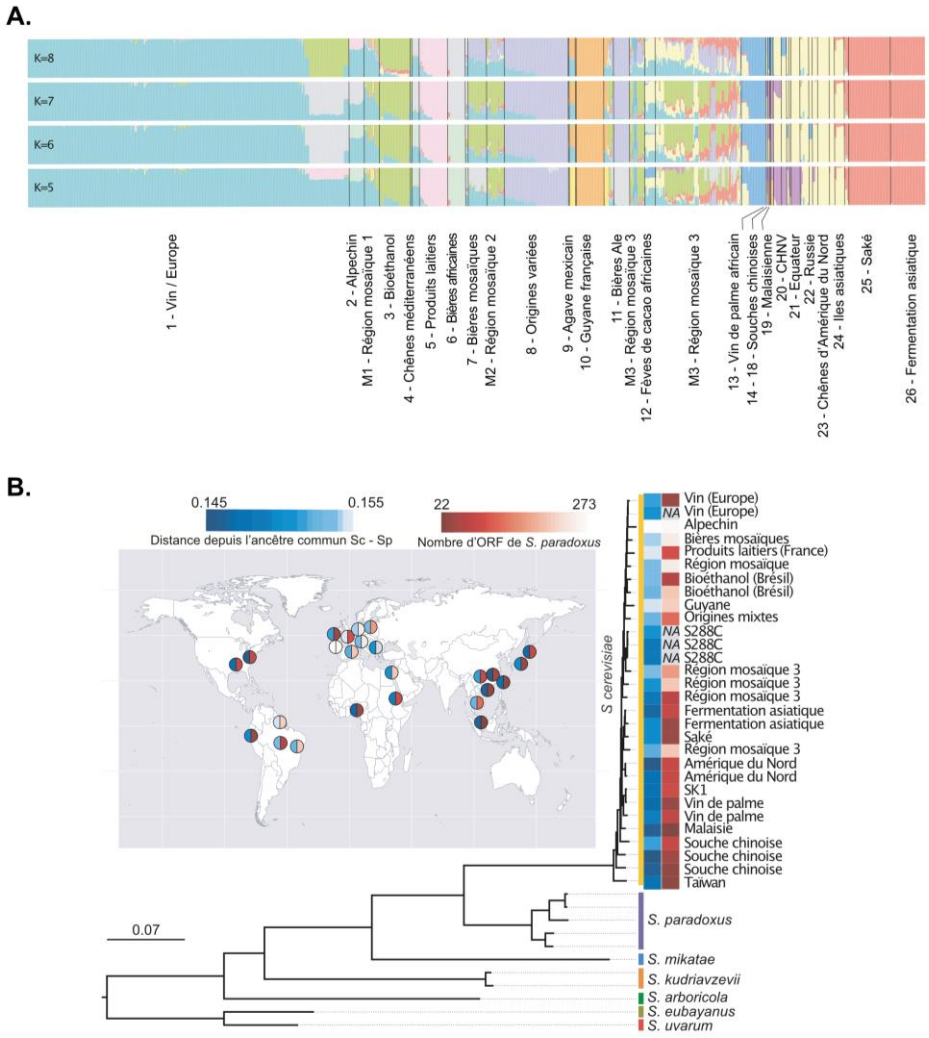


Figure 7. Structure et origine d'une population de 1011 isolats naturels de *Saccharomyces cerevisiae*.
(A.) Structure de la population de *S. cerevisiae* basée sur une stratégie de calcul de maximum de similarité en inférant un nombre de sous-populations entre 5 et 8 (valeur K). (B.) Arbre phylogénétique construit à partir de la comparaison des gènes communs entre les différentes sous-populations de *S. cerevisiae* et des espèces proches, du genre *Saccharomyces.* Origines géographiques des isolats de *S. cerevisiae* et calcul de la distance par rapport à l'ancêtre commun *S. cerevisiae* (Sc) / *Saccharomyces paradoxus* (Sp), en bleu, et le nombre d'ORF introgressées de *S. paradoxus*, en rouge. Figures adaptées de Peter et al., 2018.

Adaptations génétiques intra-spécifiques au sein de sous-populations

L'identification de sous-populations au sein des espèces révèle des pressions de sélection liées notamment à l'environnement de vie des individus. Différentes adaptations des génomes aux conditions environnementales peuvent ainsi être détectées dans les sous-populations.

Chez la levure *S. cerevisiae*, il est intéressant de noter que certaines sous-populations sont d'origine naturelle alors que d'autres ont été domestiquées par l'Homme. Ces isolats domestiqués sont impliqués dans des processus industriels divers comme par exemple l'élaboration de vin, de bière, de pain, de produits laitiers, de saké. Par conséquent, les génomes des isolats non-domestiqués ont évolué de façon distincte par rapport aux génomes des isolats domestiqués. Tandis que l'évolution par accumulation de SNP a prédominé dans les sous-populations naturelles, la domestication a participé à une variation importante du contenu en gènes (Peter et al., 2018). Ces différences illustrent la sélection subie par les populations naturelles par rapport à la sélection humaine qui favorise de meilleurs phénotypes pour des intérêts industriels précis. Les conditions de culture ont un impact considérable sur la sélection des isolats. Lors de la fabrication de produits laitiers, par exemple, les souches doivent être performantes dans des milieux qui contiennent notamment du galactose comme source de carbone. Différentes adaptations à ce milieu ont été identifiées dans les génomes de cette sous-population telles que des HGT de gènes impliqués dans le métabolisme du galactose (*GAL7*, *GAL10* et *GAL1*) ou encore des duplications de la D-lactate déshydrogénase, *DLD3* (Legras et al., 2018; Peter et al., 2018). Il est intéressant de noter que la divergence génétique au sein des sous-populations domestiquées peut être très variable. La sous-population des souches de vin et celle des souches de saké présentent par exemple une diversité génétique intra-clade très faible, $\pi = 1x10^{-3}$ et $0,8x10^{-3}$ respectivement. Cette valeur de diversité nucléotidique, π, représente la valeur moyenne de la divergence nucléotidique calculée pour chaque paire d'individus. À l'inverse, les isolats impliqués dans la fermentation de la bière se répartissent dans 3 sous-populations distinctes qui présentent une diversité nucléotidique intra-clade beaucoup plus importante, de l'ordre de $\pi = 2,8x10^{-3}$. Les souches de bière ont cependant des caractéristiques partagées entre ces clades, à savoir des génomes polyploïdes ($n > 2$), des aneuploïdies fréquentes et un degré d'hétérozygotie important. Cela illustre la coexistence de plusieurs évènements indépendants de domestication pour les souches de bière, dans le cadre desquels les mêmes caractéristiques ont été

sélectionnées, alors qu'un ancêtre commun unique serait à l'origine des sous-populations de vin ou de saké (Peter et al., 2018).

L'évolution des génomes est influencée par divers facteurs, parfois associés à l'Homme, mais aussi relatifs aux conditions de vie ou au climat. Chez l'Homme, par exemple, la couleur de la peau est liée à l'exposition variable aux UV d'une région géographique à l'autre (Jablonski and Chaplin, 2000). Il a ainsi été montré que l'évolution de certaines régions génomiques est liée au degré d'exposition aux UV des sous-populations et que l'importante variance phénotypique de ce trait est associée à une origine génétique complexe. De nombreux mécanismes de régulation du niveau de production de mélanine ont été mis en évidence et impliquent un grand nombre de variants génétiques (Beleza et al., 2013; Crawford et al., 2017; Liu et al., 2015; Lloyd-Jones et al., 2017). De manière intéressante, l'Homme moderne, originaire d'Afrique où l'exposition aux UV est importante, possède les variants ancestraux produisant plus de mélanine. Ces variants ancestraux ont notamment été remplacés par des variants introgressés à partir des espèces archaïques, Néandertal et Dénisova, moins exposées aux UV (Crawford et al., 2017; Simonti et al., 2016). D'autres variants introgressés maintenus au cours des générations en raison de la faible exposition aux UV se retrouvent dans les sous-populations eurasiennes. Une étude récente a par exemple permis d'associer des variants introgressés avec la prédisposition à la kératose actinique, maladie causée par l'exposition chronique aux UV (Simonti et al., 2016). Des questions se posent alors sur la conservation de variants génétiques, pourtant délétères, dans certaines sous-populations. Reprenons l'exemple de la drépanocytose, la mutation entraînant une altération de la β-globine doit être présente sous sa forme homozygote, HbS, pour provoquer la maladie. Cependant, la forme hétérozygote, un allèle sain HbA et un allèle muté HbS, a montré un avantage sélectif associé avec la résistance au paludisme (Flint et al., 1998; Kariuki and Williams, 2020). Cela explique ainsi la conservation et la prévalence plus importantes de l'allèle HbS dans les sous-populations africaines subsahariennes où le paludisme se propage toujours intensivement.

Ces observations illustrent ainsi la complexité derrière la constitution et le maintien de la diversité génétique au sein des populations. De plus, sur la grande diversité génétique présente au sein des espèces, peu de variants sont associés à un ou plusieurs phénotypes. Inversement, la variance phénotypique de nombreux traits complexes n'est expliquée qu'en partie par un nombre limité de variants génétiques.

Étude de la relation génotype-phénotype

Après l'établissement des premières bases de la génétique il y a 150 ans, la mise au point de différentes stratégies de séquençage et la démocratisation de la génomique des populations ont bouleversé les manières d'étudier les relations entre le génotype et le phénotype. Des catalogues de variants génétiques détectés au sein de populations naturelles ont permis de reconstruire une partie de l'histoire évolutive de certaines espèces et de mieux caractériser les régions génomiques qui ont subi diverses pressions de sélection. L'ensemble de cette diversité génétique est le fondement de la variance phénotypique observée entre les individus d'une même espèce. Un objectif majeur reste donc aujourd'hui de disséquer, sur la base de ces observations, la relation liant le génotype au phénotype et ce pour une multitude de traits, allant des plus simples, monogéniques, aux plus complexes (Mackay, 2001).
Après un rappel des méthodes de génétiques classiques directes ou indirectes pour identifier l'implication de variants génétiques dans un phénotype, nous évoquerons les deux stratégies majeures de mise en relation du génotype à la variance phénotypique. Tout d'abord, les analyses de liaison qui reposent sur les bases de l'hérédité et de la ségrégation des allèles dans la descendance d'un croisement. Ensuite, les analyses d'association qui reposent quant à elles sur la recombinaison ancestrale au sein d'une large population d'individus génétiquement différents. Chaque stratégie possède ses propres limites pour décomposer l'architecture génétique complète des phénotypes et les bases de l'héritabilité manquante seront alors exposées.

Stratégies de génétique classique

Dès 1941, et avant même la découverte de la structure de l'ADN, du code génétique puis des mécanismes de l'expression génétique dans les années 50, Beadle et Tatum émettent l'hypothèse d'un gène pour une enzyme chez le champignon *Neurospora crassa* (Beadle and Tatum, 1941). L'identification des étapes de transcription et de traduction pour le passage d'ADN en ARNm puis en protéine (Jacob and Monod, 1961) couplée à la mise au point du séquençage Sanger (Sanger et al., 1977) sont à l'origine de la biologie et la génétique moléculaires. Par conséquent, différentes stratégies de mutagénèse aléatoire ou dirigée ont été établies pour lier les variations génétiques et phénotypiques.

Les stratégies de mutagénèse aléatoire consistent à introduire des mutations dans le génome à des positions variables non contrôlées puis à associer un phénotype à ces mutations. L'effet mutagène d'agents physiques (rayons X) ou chimiques (sulfure de 2,2'-dichlorodiéthyle ou gaz moutarde) est décrit dès la première moitié du XX^{ème} siècle (Auerbach and Robson, 1946; Muller, 1928). Différents agents mutagènes rentrent alors dans les laboratoires pour réaliser des expériences de mutagénèse aléatoire sur les génomes. De grands criblages mettent en évidence l'association de plusieurs gènes et mutations à un phénotype dans les années 70. Environ 90 mutations dans 24 gènes différents, générées par un traitement au nitrosoguanidine, ont par exemple été identifiées pour leur implication dans la division cellulaire chez la levure *Schizosaccharomyces pombe* (Nurse and Thuriaux, 1980; Nurse et al., 1976). En parallèle, des criblages ont aussi été réalisés dans des organismes modèles multicellulaires plus complexes comme le nématode *C. elegans* ou encore la souris *M. musculus*. Brenner a par exemple identifié près de 250 mutations dans 77 gènes altérant les mouvements de *C. elegans* à partir d'un traitement à l'EMS (méthanesulfonate d'éthyle). Chez la souris, des criblages à grande échelle réalisés par traitement ENU (N-nitroso-N-éthylurée) ont permis d'identifier des gènes candidats associés à des maladies humaines (Hitotsumachi et al., 1985; Hrabé de Angelis et al., 2000; Nolan et al., 2000). La caractérisation moléculaire des mutants identifiés par mutagénèse aléatoire s'est développée grâce au séquençage de l'ADN, d'abord par les méthodes de Sanger ou de Maxam et Gilbert puis les méthodes à haut-débit. La démocratisation de l'utilisation des techniques de séquençage à haut-débit a permis la mise en place de nouvelles stratégies de mutagénèse aléatoire, permettant de limiter l'utilisation d'agents mutagènes et de caractériser directement le variant génétique responsable. Nous pouvons par exemple citer les stratégies de séquençage d'insertions de transposons (TIS) mises au point il y a 10 ans (Cain et al., 2020; van Opijnen and Levin, 2020; van Opijnen et al., 2009). Ces stratégies reposent sur l'introduction d'un transposon exogène dans les cellules ciblant aléatoirement des régions génomiques. L'impact phénotypique de l'insertion du transposon varie ainsi selon sa localisation et seules les cellules dont le transposon s'est inséré dans une région non-essentielle dans les conditions testées peuvent alors survivre. La sélection suivie du séquençage ciblé des positions d'insertion permettent ainsi d'identifier les régions génomiques sans conséquence sur la survie des cellules dans les conditions testées (Figure 8).

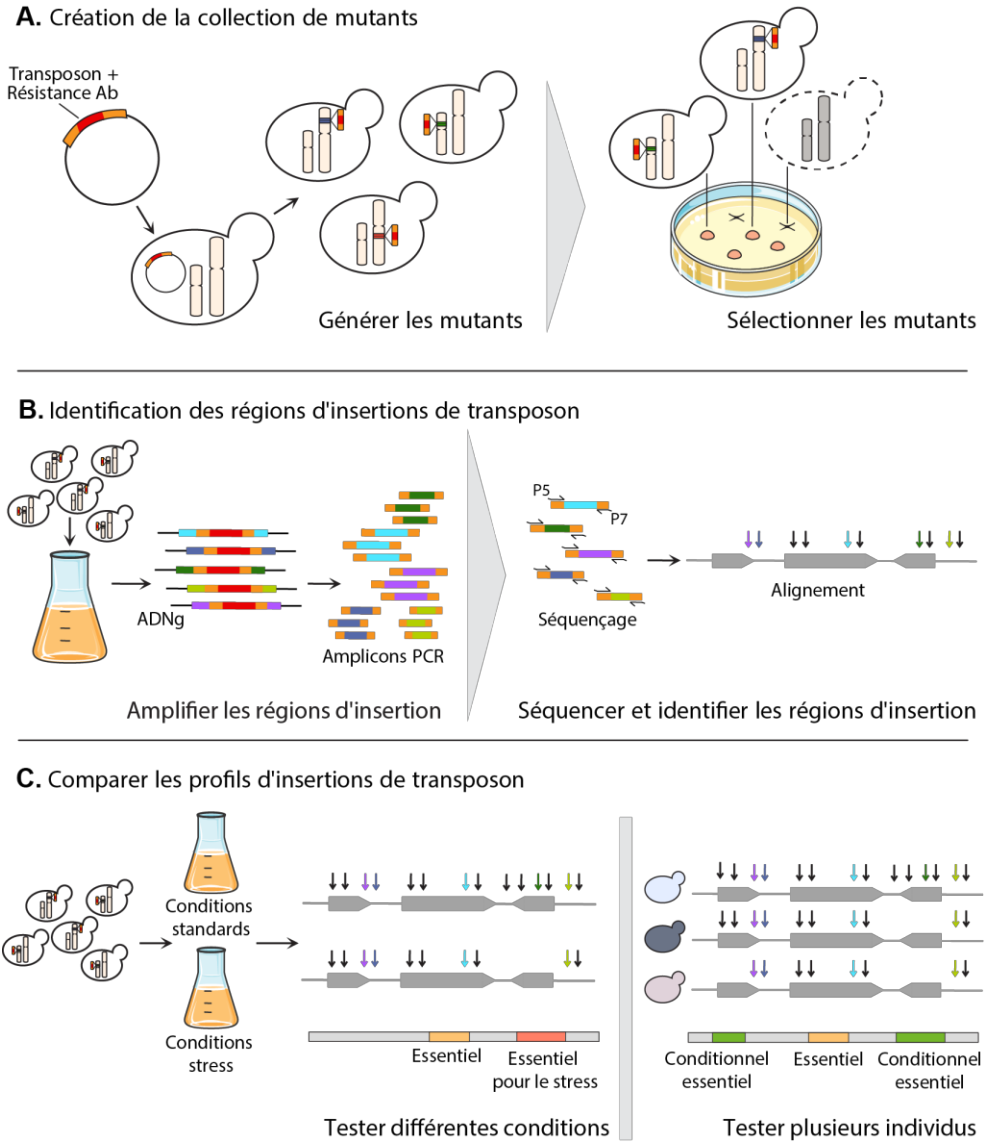


Figure 8. Stratégie de séquençage d'insertions de transposons.
(A.) Création d'une librairie de millions de mutants d'insertion de transposon à partir d'une construction composée d'un transposon pouvant s'insérer aléatoirement dans les génomes et d'un marqueur de sélection de l'insertion, une cassette de résistance à un antibiotique (Ab) par exemple. Sélection des mutants sur un milieu sélectif, supplémenté en antibiotique par exemple. Ces mutants possèdent d'une part l'insertion de transposon et d'autre part peuvent survivre à l'interruption du génome engendrée. (B.) Identification des régions d'insertion par regroupement des mutants survivants et extraction de leur ADN génomique pour en amplifier

spécifiquement les régions d'insertion du transposon. Ces régions génomiques sont alors séquencées et alignées le long du génome de l'individu considéré pour en déterminer les sites précis d'insertion du transposon. (C.) Différentes problématiques peuvent être étudiées avec cette stratégie. La survie des mutants d'insertion d'un même individu peut être testée dans plusieurs conditions, la comparaison des profils d'insertion du transposon permettra alors d'identifier des gènes essentiels uniques aux conditions testées, comme par exemple des gènes impliqués dans la résistance à un stress. Plusieurs individus d'une même espèce peuvent également subir la saturation en transposons et la comparaison des profils d'insertion révèlera des gènes essentiels conditionnels à certains fonds génétiques.

Cette approche à haut-débit permet ainsi de passer en revue l'ensemble du génome et de tester en parallèle différents individus et / ou dans différentes conditions expérimentales. À l'heure actuelle, les organismes unicellulaires, bactéries ou levures, sont les cibles principales pour la mutagénèse par insertion de transposons et de nombreux systèmes ont été élaborés (Cain et al., 2020; Christen et al., 2011; Evertts et al., 2007; Gangadharan et al., 2010; Guo et al., 2013; Michel et al., 2017; van Opijnen and Levin, 2020; van Opijnen et al., 2009) (Figure 8). Le génome de nombreuses bactéries pathogènes de l'Homme a ainsi été criblé afin d'identifier leurs gènes essentiels et ainsi fournir des cibles thérapeutiques (Cain et al., 2020; van Opijnen and Levin, 2020) comme par exemple chez *Staphylococcus aureus* (Coe et al., 2019; Santiago et al., 2015).

En parallèle des stratégies de mutagénèse aléatoire, des outils de biologie moléculaire basés sur des principes de recombinaison homologue ont permis la mise en place de stratégies de mutagénèse ciblée. Ces stratégies ont premièrement consisté à valider le rôle d'un gène dans un phénotype d'intérêt. Des mutants *knock-out* (KO) sont générés et permettent l'inactivation d'un gène, souvent par la délétion de ce dernier, à partir de système adapté selon l'espèce. Chez la levure *S. cerevisiae*, la méthode est simple et repose sur l'amplification PCR d'un marqueur d'auxotrophie ou d'une cassette de résistance à un antibiotique avec 50 pb d'homologie de chaque côté avec la région génomique à déléter. Ainsi, la transformation des cellules avec ce fragment d'ADN conduit à un remplacement de l'ADN génomique situé entre les 2 régions d'homologie par le fragment PCR grâce à la recombinaison homologue (Shortle et al., 1982; Wach et al., 1994). D'autres méthodes sont également disponibles pour d'autres espèces, comme le système de recombinaison Cre/Lox notamment utilisé pour générer des mutants chez la souris (Sauer and Henderson, 1988). Le séquençage des génomes de référence de nombreuses espèces à la fin des années 90 a permis de caractériser une grande partie des gènes et d'en définir leur localisation génomique précise. Ces données ont conduit à l'établissement de

collections de délétions, pour lesquelles chaque gène a été supprimé individuellement afin d'en étudier la conséquence phénotypique. Ces premières collections ont été rapidement générées chez les organismes unicellulaires, bactéries (Baba et al., 2006) et levures (Giaever et al., 2002; Kim et al., 2010; Winzeler et al., 1999) principalement. Les organismes modèles multicellulaires ont suivi avec la création de larges consortiums, le *International Knockout Mouse Consortium* (http://www.mousephenotype.org) pour *M. musculus* (Skarnes et al., 2011), la collection de délétion DrosDel pour *D. melanogaster* (Ryder et al., 2007) ou le *C. elegans Gene Knockout Consortium* pour le nématode *C. elegans* (Frøkjaer-Jensen et al., 2010). Des collections de doubles et triples mutants ont également été réalisées chez la levure *S. cerevisiae* afin d'étudier les réseaux d'interaction des gènes au sein de l'espèce (Costanzo et al., 2016; Kuzmin et al., 2018). D'autres stratégies ont été développées dans le but d'altérer l'action d'un gène, comme la technique de l'interférence à ARN largement utilisée et décrite chez *C. elegans* (Fire et al., 1998). Ce système repose sur l'introduction dans les cellules d'ARN double brin (dsRNA) qui vont interférer avec les ARNm complémentaires au dsRNA et inhiber l'expression du gène ciblé. Différents criblages des génomes ont ainsi été réalisés avec cette méthode dans plusieurs organismes afin d'étudier les phénotypes résultants (Dietzl et al., 2007; Kamath et al., 2003; Kiger et al., 2003; Timmons et al., 2001). Ce mécanisme d'interférence a été décrit chez de nombreux eucaryotes à l'exception de *S. cerevisiae*, où il n'existe pas. Enfin, l'outil biomoléculaire d'édition du génome CRISPR-Cas (pour Clustered Regularly Interspaced Short Palindromic Repeats) développé il y a moins de 10 ans est aujourd'hui incontournable et représente la méthode de choix pour induire des modifications génétiques chez les organismes eucaryotes (Doudna and Charpentier, 2014; Jinek et al., 2012). Cette stratégie dérive d'un mécanisme d'immunité adaptative des bactéries. Elle permet *via* un ARN guide de cibler une région spécifique du génome qui sera clivée par l'endonucléase Cas9. Les conséquences de la cassure double brin de l'ADN peuvent être multiples : une délétion de la région, le remplacement d'un seul ou de plusieurs nucléotides, l'inhibition de l'expression d'un gène ou encore l'introduction de variants structurels (CNV ou inversion par exemple). Grâce à l'efficacité de ce système, de nombreuses variations génétiques peuvent être induites à haut-débit dans les génomes dans un grand nombre d'individus en parallèle à l'aide d'optimisations de l'outil de base (Fleiss et al., 2019; Sharon et al., 2018).

Les analyses de liaison

L'identification de variants génétiques responsables de la variance phénotypique observée au sein d'une population se base sur l'étude de la ségrégation des marqueurs génétiques au sein de cette population. On définit alors comme QTL (pour Quantitative Trait Loci) la région génomique associée à la variation d'un trait quantitatif (ou complexe). La précision de la région génomique détectée varie selon la stratégie employée, le nombre de marqueurs génétiques et le déséquilibre de liaison, qui peut être défini comme l'association non aléatoire de 2 marqueurs au sein d'une population. La résolution du QTL peut aller d'une large région du génome contenant plusieurs gènes et éléments régulateurs, à un seul gène, QTG (pour Quantitative Trait Gene), voire au nucléotide près, QTN (pour Quantitative Trait Nucleotide).

Le principe des analyses de liaison

L'analyse de liaison permet l'identification de QTL en étudiant la descendance issue d'un croisement entre deux individus génétiquement différents. Les évènements de recombinaison se produisant lors de la méiose vont induire un mélange des variants génétiques parentaux au sein de la descendance, permettant ainsi d'obtenir une combinaison unique pour chaque descendant. Le phénotypage et le génotypage des descendants pourront alors permettre d'associer un phénotype particulier à des QTL. En effet, tous les descendants présentant un phénotype proche partageront les variants génétiques impliqués dans le trait. À l'inverse, les variants parentaux seront aléatoirement répartis sur le reste du génome (Figure 9A). L'intérêt de croiser deux parents génétiquement divergents permet d'augmenter le nombre de variants génétiques utilisés comme marqueurs dans la descendance. Les premiers marqueurs génétiques utilisés dans les stratégies d'analyses de liaison étaient des sites de restriction enzymatique, RFLP (pour Restriction Fragment Length Polymorphisms) (Botstein et al., 1980). Aujourd'hui, grâce aux stratégies de séquençage haut-débit, ce sont les SNP entre les génomes parentaux qui servent de marqueurs moléculaires. Tous les organismes ne sont pas adaptés à l'analyse de liaison. En effet, afin de définir une localisation précise des QTL, le taux de recombinaison doit être suffisamment élevé pour permettre un mélange efficace des marqueurs génétiques parentaux dans la descendance et éviter ainsi un déséquilibre de liaison trop important. Il est aussi possible d'augmenter le nombre de générations et donc de méioses et / ou de descendants pris en compte dans l'analyse pour améliorer la résolution. De manière intéressante, il a été mis en évidence que chez les eucaryotes

le taux de recombinaison est anti-corrélé avec la taille des génomes (Figure 9B) rendant ainsi les levures les modèles les plus appropriés pour ce type d'analyse (Fay, 2013; Lynch, 2006; Swinnen et al., 2012).

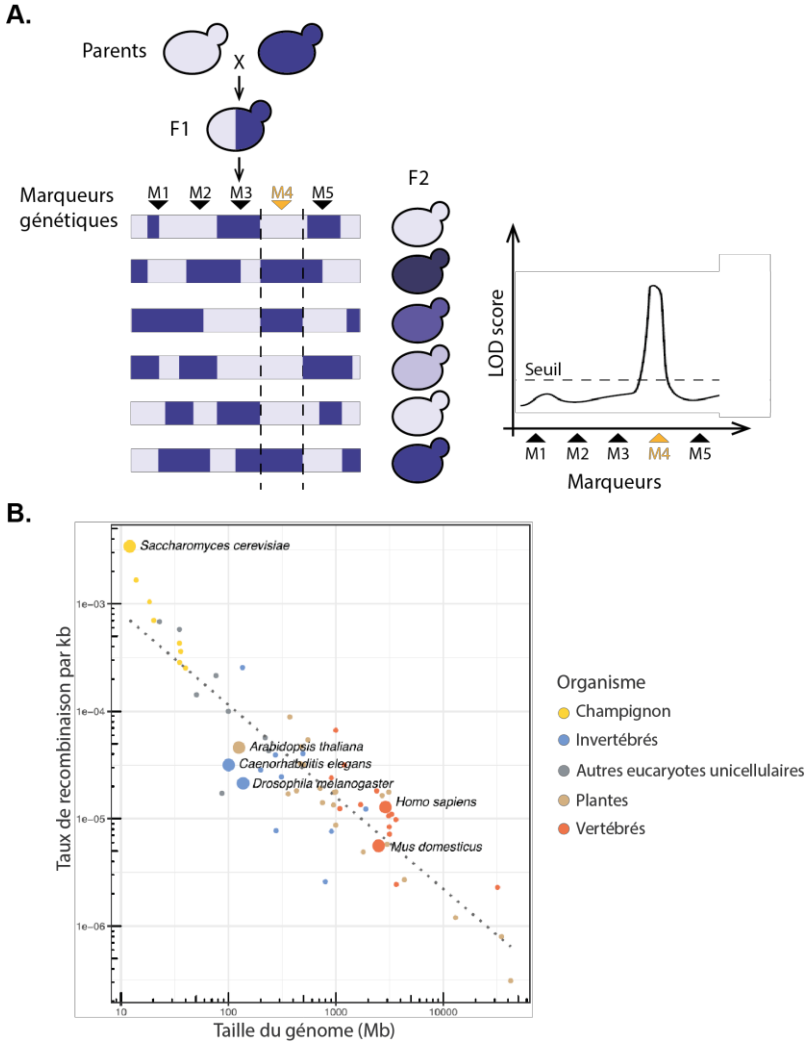


Figure 9. Principe des analyses de liaison.
(A.) A partir de la descendance d'un croisement entre 2 individus au génotype et au phénotype d'intérêt différents, un score de liaison (LOD score pour Logarithm Of Odds) est calculé entre les marqueurs génétiques de chaque descendant et sa valeur phénotypique. Un seuil de significativité est défini pour détecter le QTL (pour Quantitative Trait Locus) impliqué dans la variance phénotypique observée dans la population. (B.) Corrélation négative entre le taux de recombinaison par kilobase et la taille des génomes (échelle logarithmique) chez des organismes eucaryotes (coefficient de corrélation = 0,90, p-value = $6{,}3e^{-25}$). Figure adaptée de Lynch, 2006.

Les applications des analyses de liaison

Les analyses de liaison ont, depuis la fin des années 70, permis d'identifier de nombreux QTL associés à des traits complexes dans différents organismes modèles. On peut par exemple citer la mise en évidence de l'implication de plusieurs gènes homéotiques dans le développement des pattes et des antennes chez la drosophile (Lewis, 1978; Struhl, 1981). Cependant, pour avoir une résolution suffisante, un nombre important de descendants est nécessaire et le génotypage de chaque individu peut être laborieux et onéreux. Pour pallier à cette problématique, des analyses par pool de ségrégants, ou BSA (pour Bulk Segregant Analysis), ont été développées (Brauer et al., 2006; Magwene et al., 2011; Segrè et al., 2006). À partir d'une génération F2 de descendants, les individus présentant un phénotype « extrême » – c'est-à-dire aux extrémités de la distribution normale du trait – sont regroupés et séquencés en un seul groupe. La fréquence allélique des marqueurs génétiques impliqués dans le phénotype sera alors déviée vers le parent possédant la version allélique du marqueur associé au trait. Dans ce type d'analyse, le nombre de descendants et le nombre d'évènements de recombinaisons doivent être relativement importants pour limiter le déséquilibre de liaison. Dans ce contexte, un organisme avec un taux de recombinaison élevé, tel que la levure *S. cerevisiae*, est privilégié. Une alternative repose sur la constitution de plusieurs générations successives de descendants, afin d'augmenter le nombre d'évènements de recombinaison. Des stratégies plus poussées, appelées x-QTL, ont ainsi été développées afin de faciliter les étapes de phénotypage et génotypage en groupe. Des millions d'individus peuvent être inclus dans les analyses et cette méthode a montré son efficacité pour l'organisme modèle *S. cerevisiae* (Ehrenreich et al., 2010) mais aussi pour le nématode *C. elegans* (Burga et al., 2019).

Le niveau d'expression des gènes figure également parmi les phénotypes pour lesquels les analyses de liaison ont donné de nombreux résultats. En effet, des eQTL (pour expression Quantitative Trait Loci) impliqués dans la régulation de l'expression génique ont été identifiés grâce à des analyses de liaison. Dans les années 2000, le niveau d'expression de chaque gène dans une population de descendants était quantifié par les technologies de puces à ADN (Brem et al., 2002; Jansen and Nap, 2001; Steinmetz et al., 2002). La levure *S. cerevisiae* a été le modèle privilégié pour ce type d'analyse et a permis le premier criblage à l'échelle du génome des mécanismes complexes de régulation des gènes avec des eQTL agissant localement et d'autres à distance (Brem et al., 2002). D'autres organismes modèles ont également été soumis à ces analyses (Jansen and Nap, 2001; Rockman et al.,

2010; Schadt et al., 2003). La caractérisation de la régulation de l'expression des gènes peut révéler des intérêts biotechnologiques chez les plantes ou servir de modèle pour étudier des maladies humaines comme l'obésité dans le modèle murin par exemple (Schadt et al., 2003). Les stratégies de séquençage des ARN (RNA-seq) actuelles ont permis d'augmenter considérablement le débit et le nombre d'individus étudiés passant ainsi d'une centaine à plus de 1000 descendants (Albert et al., 2018). Les méthodes de quantification des niveaux de protéines telles que la spectrométrie de masse ont également fourni des valeurs phénotypiques (abondance protéique) permettant de réaliser des analyses de liaison. L'identification de pQTL (pour protein Quantitative Trait Loci) révèle ainsi de nouveaux réseaux de régulation au niveau protéique (Albert et al., 2018; Foss et al., 2007). Cependant ces techniques restent laborieuses à mettre en œuvre à haut-débit et restent donc rares.

Les limites des analyses de liaison

Les résultats des analyses de liaison réalisées ces 20 dernières années ont permis en partie d'élucider les bases génétiques à l'origine des traits complexes. En effet, des catalogues conséquents de QTL et même de QTN ont été associés à des phénotypes complexes. Les validations moléculaires de ces QTN donnent également l'occasion d'aller plus loin dans l'étude de l'impact de tels variants génétiques sur la variance phénotypique dans une espèce. Plusieurs études notamment chez l'Homme et la drosophile ont d'ailleurs montré que les variants impliqués dans les traits quantitatifs étaient des variants génétiques rares ou à faible fréquence, c'est-à-dire présents dans moins de 1 à 5% des individus de l'espèce (MacKay et al., 2009). De manière intéressante, parmi les 284 QTN validés fonctionnellement chez *S. cerevisiae*, plus de la moitié (environ 150 QTN) sont effectivement des variants génétiques rares (Peltier et al., 2019; Peter et al., 2018). L'élucidation complète de l'origine des traits complexes est ainsi rendue difficile par les nombreux variants rares impliqués avec une faible contribution à la variance phénotypique. De plus, ces systèmes restent limités à deux parents (et donc deux fonds génétiques) et ne permettent donc pas de représenter toute la diversité génétique d'une espèce. Ces dernières années, des designs expérimentaux innovants ont permis d'aller plus loin grâce au modèle *S. cerevisiae*, en particulier. D'un côté, une étude portant sur environ 20 000 descendants d'une génération F6 a considérablement amélioré la résolution de détection des QTL et illustre la pléiotropie et les caractéristiques de certains variants génétiques (Jakobson and Jarosz, 2019). D'un autre côté, pour mieux représenter la diversité au sein de l'espèce, une stratégie d'analyses de liaison a été élaborée à partir

de 16 parents génétiquement divers, chacun étant croisé à tour de rôle avec un autre parent. Deux populations différentes d'environ 1000 descendants ont été générées pour chaque croisement et révèlent ainsi une contribution importante des variants rares dans la population (Bloom et al., 2019). Malgré les innovations régulières, le principe des analyses de liaison reste limitant à la fois pour le choix de l'organisme modèle, en raison de la descendance conséquente à générer, mais aussi du fait du manque de représentativité de la diversité génétique de l'espèce.

Les études d'association pangénomique

Les nombreux projets de séquençage de plusieurs centaines d'individus d'une même espèce ont fourni une vue globale de la diversité génétique intra-spécifique. Les études d'association pangénomique (GWAS pour Genome-Wide Association Studies) se sont révélées une stratégie efficace afin d'explorer l'impact de cette variabilité sur certains phénotypes. Ces études permettent également de considérer un panel d'espèces pour lesquelles les analyses de liaison ne sont pas adaptées, comme l'Homme par exemple.

Le principe des études d'association pangénomique

Pour la première fois en 2005, les études d'association pangénomique sont décrites comme une stratégie adaptée pour disséquer l'origine de maladies génétiques humaines et identifier des cibles thérapeutiques (Hirschhorn and Daly, 2005). À partir d'un panel d'individus, les études d'association permettent d'associer statistiquement des variants génétiques causaux à la variation d'un phénotype, tel que la taille ou les symptômes d'une maladie. Des groupes d'individus sont créés selon la valeur phénotypique testée et sont mis en parallèle avec les variants génétiques présents dans la population. L'association est détectée comme statistiquement significative selon le nombre d'individus du sous-groupe phénotypique possédant le même variant (Figure 10). Les grands projets de séquençage qui concernent des centaines de génomes permettent ainsi de disposer des génotypes de suffisamment d'individus pour valider statistiquement l'association. Au sein d'une espèce, les évènements de recombinaison ancestrale accumulés au cours des générations participent activement à la diversité des génomes. Inclure un grand nombre de génomes dans les études d'association réduit ainsi le déséquilibre de liaison, augmente la puissance statistique et permet une identification plus précise du locus causal.

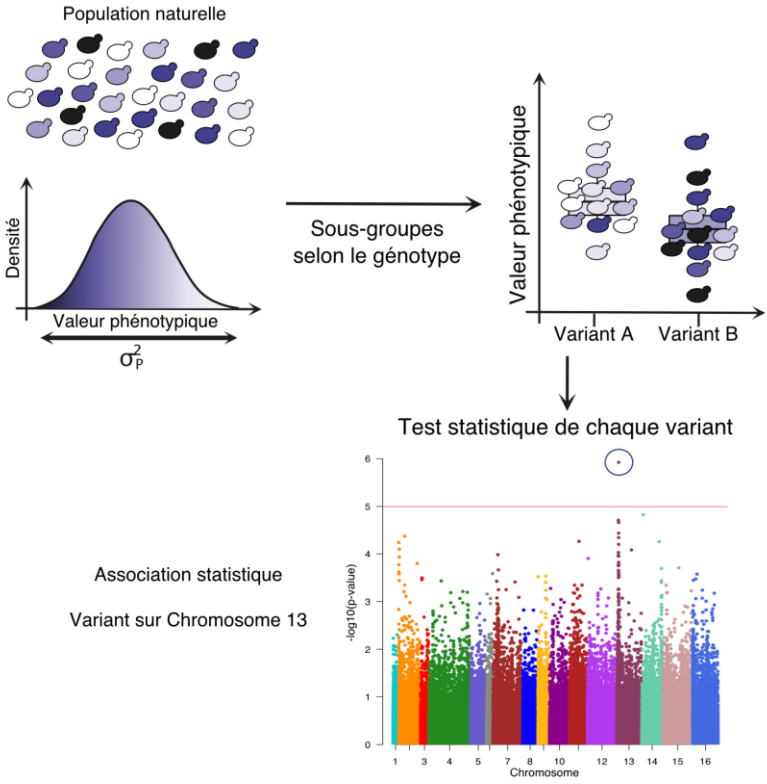


Figure 10. Principe des études d'association pangénomique (ou GWAS).
À partir de la variance observée pour un phénotype d'intérêt au sein d'une population naturelle, chaque variant génétique est associé aux valeurs du phénotype pour retrouver une différence significative entre les sous-groupes associés au génotype. Chaque variant est statistiquement testé et un seuil est établi pour déterminer le ou les variants significativement associés avec la variation du phénotype dans la population.

Les applications des études d'association pangénomique

Dès la mise en application de ces stratégies, de nouveaux gènes responsables de maladies génétiques humaines ont pu être mis en évidence. Par exemple, les premières études d'association ont permis d'identifier en partie les origines génétiques de la dégénérescence maculaire liée à l'âge (DMLA), jusqu'alors inconnues, associant ainsi à la maladie un variant présent dans un intron du gène *CFH* codant le facteur H (Klein et al., 2005). Depuis plus de 15 ans, les études d'association pangénomique ont bouleversé la caractérisation des bases génétiques des traits complexes en identifiant par exemple plus de 25 0000 variants associés à une multitude de traits chez l'Homme (https://www.ebi.ac.uk/gwas/, consulté le

07/07/2021) (Buniello et al., 2019). Au-delà de l'Homme, les organismes modèles comme la plante *A. thaliana* (Alonso-Blanco et al., 2016), la levure *S. cerevisiae* (Peter et al., 2018) ou encore la souris (Flint and Eskin, 2012; Gonzales et al., 2018) ont également été soumis à des études d'association pangénomique sur la base des grands projets de séquençage. L'explosion des études d'association a, de la même manière que le séquençage haut-débit, conduit à des listes de variants génétiques associés à une multitude de phénotypes et, inversement, de traits complexes associés à des centaines voire des milliers de variants causaux. Différents catalogues ont ainsi été établis afin de regrouper et de visualiser plus simplement les relations génotype-phénotype. Un atlas des résultats de ce type d'études chez l'Homme (https://atlas.ctglab.nl/) a par exemple permis de déceler le rôle pléiotropique de la majorité des variants génétiques causaux alors impliqués dans plusieurs phénotypes. De plus, ces travaux ont montré l'impact prépondérant des gènes et des éléments régulateurs dans l'établissement des phénotypes (Watanabe et al., 2019). Ces dernières années, les études d'association ont pris une nouvelle dimension. En effet, les données de transcriptomique, métabolomique, protéomique ou encore d'épigénomique ont permis d'accroître le nombre de phénotypes étudiés et d'individus inclus dans les analyses, que ce soit chez l'Homme ou encore dans différents modèles végétaux (Jansen et al., 2019; Kawakatsu et al., 2016; Scossa et al., 2021; The GTEx Consortium, 2017, 2020). Parmi les méta-analyses réalisées, l'étude portant sur les origines génétiques des troubles du sommeil considère les génomes, les transcriptomes et les épigénomes et est à ce jour, l'étude d'association la plus conséquente incluant plus de 1,3 millions d'individus (Jansen et al., 2019). Dans le cadre du GTEx consortium, l'analyse du transcriptome de 838 donneurs dans 49 tissus différents (de 73 à 670 individus/tissu) révèle une régulation en majorité locale et complexe de l'expression des gènes dans les tissus. Ces résultats illustrent également les limites de la détection des eQTL distants impactant faiblement la variance phénotypique en raison du pouvoir statistique restreint dans des tissus avec encore peu d'individus (The GTEx Consortium, 2017, 2020).

L'héritabilité manquante

La taille humaine figure parmi les phénotypes dont les origines génétiques ont été les plus disséquées, notamment par l'intermédiaire d'études d'association pangénomique. Plusieurs milliers de variants génétiques sont aujourd'hui associés à ce phénotype (Akiyama et al., 2019; Guo et al., 2021; Manolio et al., 2009; Wood et al., 2014; Yengo et al., 2018). Cependant, l'ensemble de ces variants associés ne permet d'expliquer que 25% de l'héritabilité (H^2) du trait dans la population (Yengo et al., 2018), ce qui illustre les limites des études d'association. Différentes variantes des études d'association classiques, comme les méta-analyses ou encore les enrichissements vers des sous-populations, permettent de détecter des variants plus rares dans l'espèce. Ces variants rares (trouvés dans moins de 5% de la population) ne peuvent en effet pas être détectés avec un pouvoir statistique suffisant par les études d'association classiques. Il en va de même pour les variants avec un impact faible sur la variance phénotypique. Par conséquent, une part d'héritabilité manquante reste ainsi à être explorée (MacKay et al., 2009; Manolio et al., 2009; Tam et al., 2019).

Origines de l'héritabilité manquante

Bien que de nombreuses stratégies aient été développées afin de disséquer l'origine génétique de la variance phénotypique, les variants génétiques identifiés ne permettent d'expliquer qu'une partie de l'héritabilité (H^2). L'héritabilité manquante correspond ainsi à la fraction de la variance génétique non expliquée. Diverses sources génétiques sont responsables de cette héritabilité manquante (Maher, 2008; Manolio et al., 2009). Parmi ces sources, nous pouvons par exemple citer les variants génétiques rares, les interactions génétiques, les variants structurels, l'effet des génomes mitochondriaux ou encore l'épigénétique. Les études d'association pangénomique classiques ne sont notamment pas adaptées à la détection de ces sources de variance génétique. En effet, ces dernières détectent effectivement en majorité des variants génétiques avec des effets additifs sur la variance phénotypique (Visscher et al., 2017). Des effets non-additifs, de dominance ou d'épistasie, participent cependant aussi à la variance génétique. En effet, chez la levure *S. cerevisiae*, un tiers de la variance phénotypique est en moyenne expliqué par des effets non-additifs (Bloom et al., 2013, 2015; Fournier et al., 2019).

L'effet des variants rares

Les grands projets de séquençage dans différentes espèces ont mis en évidence que la grande majorité des variants de type SNP avait une faible fréquence dans les populations étudiées. En effet, environ 90% des SNP ont une fréquence allélique mineure (MAF pour Minor Allele Frequency) inférieure à 5% dans les populations étudiées de *S. cerevisiae* ou chez l'Homme (Figure 11) (Auton et al., 2015; Peter et al., 2018). De plus, une importante contribution de ces variants rares dans les phénotypes complexes a été décrite chez l'Homme et la drosophile (MacKay et al., 2009). Récemment, l'analyse de 284 QTN identifiés par analyses de liaison et validés fonctionnellement a révélé une fréquence allélique inférieure à 5% dans la population des 1011 isolats naturels de *S. cerevisiae* pour plus de 50% d'entre eux (Peltier et al., 2019).

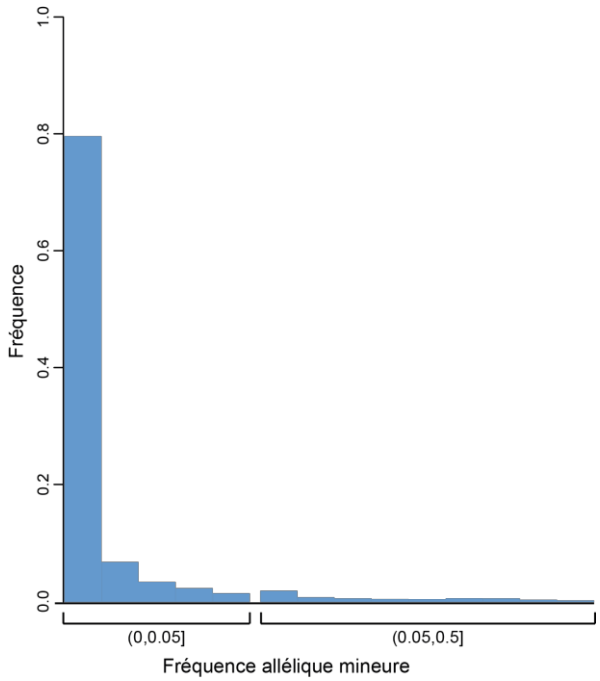


Figure 11. Distribution de la fréquence allélique mineure des variants dans 1011 isolats naturels de *Saccharomyces cerevisiae*.
Au total, 92% des SNP présents dans la population ont une fréquence allélique rare, < 0,05. Figure adaptée de Peter et al., 2018.

À l'échelle d'une population, la détermination de l'impact fonctionnel de ces variants rares sur la variance phénotypique est ainsi primordiale pour obtenir une vue globale des bases génétiques de la diversité phénotypique. Une limite majeure des analyses d'association repose cependant sur la fréquence allélique des variants génétiques dans la population nécessaire pour significativement associer le variant à un phénotype. En conséquence, seuls les variants dont la fréquence allélique mineure dans la population est supérieure à 0,05 (5%) sont inclus dans les stratégies classiques d'association de type GWAS. Le design expérimental du croisement diallèle a permis de contourner cette limite et d'inclure ces variants rares dans une analyse d'association (Fournier et al., 2019). En effet, cette stratégie consiste à réaliser des croisements dirigés entre une sélection d'individus les uns avec les autres et d'estimer l'impact des composantes génétiques pour chaque phénotype testé (Griffing, 1956). À partir des croisements entre 34 isolats de *S. cerevisiae* précédemment séquencés, le génotype de 595 hybrides a été recréé *in silico*. Chaque variant génétique présent dans un des parents est ainsi représenté dans 34 hybrides au minimum, et aura donc une fréquence allélique supérieure à 5% au sein de la population considérée. Les 31 000 SNP présents dans les 595 génotypes et incluant 3,8% de variants rares (< 5%) dans la population de 1011 isolats peuvent ainsi être inclus dans la matrice pour l'étude d'association pangénomique. De manière intéressante, plus de 16% des variants rares dans la population sont associés à la variance phénotypique observée dans 49 conditions de culture, confirmant ainsi le rôle notable de cet aspect de la diversité génétique souvent oublié (Fournier et al., 2019). D'autres stratégies d'inclusion des variants rares dans les analyses ont aussi été développées chez la levure (Bloom et al., 2019) ou encore chez l'Homme pour déterminer l'impact des variants rares sur l'expression des gènes (Ferraro et al., 2020; Li et al., 2021, 2017).

L'effet du fonds génétique

De manière évidente, l'élucidation de l'origine des traits complexes demande une résolution fine pour déterminer l'intégralité des variants génétiques impliqués. Concernant les traits monogéniques, bien que le variant causal associé au phénotype puisse plus simplement être identifié, une complexité sous-jacente peut apparaître. En effet, la présence d'une même mutation peut entraîner un phénotype avec un degré de gravité variable selon l'individu, on parle alors d'expressivité. Dans le cas d'une maladie génétique, la neurofibromatose de type I par exemple, une mutation dans le gène *NF1* induit la maladie. Cependant la sévérité des symptômes sera

variable entre individus porteurs d'une mutation, allant de l'apparition de tâches pigmentées au développement de tumeurs cutanées (Pasmant et al., 2012). Dans des cas plus extrêmes, malgré la présence de la mutation causale, le phénotype peut ne pas se manifester. La mutation du gène *BRCA1*, par exemple, prédispose au cancer du sein et des ovaires chez l'Homme. Cependant, 20% des individus possédant la mutation ne développeront pas ce cancer au cours de leur vie (Mavaddat et al., 2013). On parle alors de pénétrance incomplète du phénotype. Ces phénomènes suggèrent ainsi un effet du fonds génétique des individus sur le phénotype et donc la présence de variants modificateurs interagissant avec le variant causal. Différentes études ont permis d'identifier des variants modificateurs impliqués dans des maladies humaines ou encore dans les conséquences phénotypiques de mutations de type perte de fonction chez la levure et d'autres organismes (Chow et al., 2016; Cutting, 2010; Hamilton and Yu, 2012; Hou et al., 2019). L'induction de mutations perte de fonction – par délétion, inhibition ou interruption de gènes – dans différents fonds génétiques permet de mettre en évidence de l'expressivité phénotypique entre les individus testés (Chandler et al., 2014; Dowell et al., 2010; Galardini et al., 2019; Johnson et al., 2019; Mullis et al., 2018; Paaby et al., 2015; Parts et al., 2021; Vu et al., 2015). À l'échelle du génome, environ 20% de variation phénotypique entre plusieurs fonds génétiques a été observé après induction de mutations perte de fonction au sein des espèces modèles *C. elegans* (Vu et al., 2015) et *S. cerevisiae* (Galardini et al., 2019). De manière plus stricte, la délétion systématique de chacun des 5100 gènes dans 2 isolats de *S. cerevisiae* a révélé 5% de gènes conditionnels essentiels, c'est-à-dire dont la délétion ne conduit à la létalité que dans un seul des 2 fonds génétiques (Dowell et al., 2010). Pour l'instant, ces études ne considèrent qu'un faible nombre de fonds génétiques et ne sont par conséquent pas représentatifs de la diversité génétique des espèces.

L'effet des variants structurels

Du fait de la difficulté à caractériser les variants structurels dans les génomes, les études d'association pangénomique ne prennent couramment en compte que les SNP. Cependant, les SV ont un impact important sur les traits complexes (Jeffares et al., 2017; Peter et al., 2018). Chez la levure *S. cerevisiae*, les CNV ont été inclus dans une analyse d'association incluant 971 isolats naturels cultivés dans 36 conditions différentes (ou phénotypes). Les résultats soulignent un impact des CNV presque 10 fois supérieur aux SNP sur la variance phénotypique, avec une médiane de 36,8% contre 4,49% de la variance expliquée, respectivement (Peter et al., 2018).

Un enrichissement des eQTL concernés par des SV a récemment été identifié dans la régulation locale de l'expression des gènes, d'autant plus lorsque le SV impacte une région codante (Scott et al., 2021). L'arrivée des stratégies de séquençage de lectures longues permet aujourd'hui d'inclure avec plus de précision les SV dans les études d'association pangénomique (De Coster et al., 2021). Bien que ce type d'études soit encore rare, plusieurs traits ont été associés avec différents SV, délétions ou duplications, dont des effets sur la taille humaine (Beyter et al., 2021).

Saccharomyces cerevisiae comme modèle d'étude de la relation génotype-phénotype

L'étude des relations génotype-phénotype se révèle toujours extrêmement complexe malgré les avancées scientifiques et technologiques permettant d'approfondir les connaissances. La mise en place de stratégies novatrices est donc nécessaire dans le but de disséquer les bases génétiques des phénotypes. Dans ce contexte, la levure *S. cerevisiae* se révèle comme un modèle de choix. Nous exposerons ainsi dans un premier temps les sources de diversité génétique au sein de l'espèce et dans un deuxième temps la large palette d'outils disponibles pour cet organisme modèle.

Génomique des populations dans l'espèce *Saccharomyces cerevisiae*

Depuis des millénaires, la levure *S. cerevisiae* est couramment utilisée dans les processus de fermentation pour la fabrication de vin, de bière ou de pain (Giannakou et al., 2020). Au milieu du XIX^{ème} siècle, Louis Pasteur décrit pour la première fois le rôle de *S. cerevisiae* dans la fermentation alcoolique (Pasteur, 1858). Très rapidement au début du XX^{ème} siècle, ce champignon unicellulaire est devenu un modèle en laboratoire, d'abord pour des intérêts biotechnologiques, puis comme organisme modèle pour faire avancer les connaissances en génétique mais aussi dans différents domaines de la biologie. De par son génome petit (12 Mb) et compact (70% du génome est codant), *S. cerevisiae* a été le premier génome eucaryote entièrement séquencé en 1996 (Dujon, 2019; Goffeau et al., 1996). Ce séquençage a révélé la présence d'environ 6000 gènes répartis sur 16 chromosomes issus d'un processus de duplication complète du génome (Wolfe, 2015; Wolfe and Shields, 1997). Au total, les génomes de plus de 2000 isolats de *S. cerevisiae* ont été séquencés dans le cadre de différents projets (Duan et al., 2018; Gallone et al., 2016; Marsit et al., 2015; Peter et al., 2018). L'étude la plus complète réalisée à ce jour a été conduite sur 1011 isolats naturels représentatifs de la diversité écologique, géographique et génétique de l'espèce (Peter et al., 2018). Cette population contient à la fois des isolats issus de la nature, de patients humains présentant des pathologies cliniques ou utilisés dans le cadre de processus industriels / fermentaires. Une diversité génétique importante a été mise en évidence au sein de l'espèce avec plus de 1,6 millions de SNP et une diversité nucléotidique maximale de 1,83% entre les isolats, environ 125 000 insertions/délétions courtes et de nombreuses duplications de gènes – chaque gène est dupliqué dans au moins un isolat (Peter et al., 2018). Alors que la fréquence allélique de 92% des SNP présents dans la population est

faible (inférieure à 5%, Figure 11), cette diversité nucléotidique est représentative des évènements évolutifs qui ont façonné les génomes au sein de l'espèce, séparant l'espèce en au moins 26 sous-populations distinctes (Figure 12). Ces sous-populations opposent les isolats domestiqués par l'Homme aux isolats naturels, et des souches mosaïques, notamment d'origine clinique, s'intercalent entre ces 2 catégories (Peter et al., 2018).

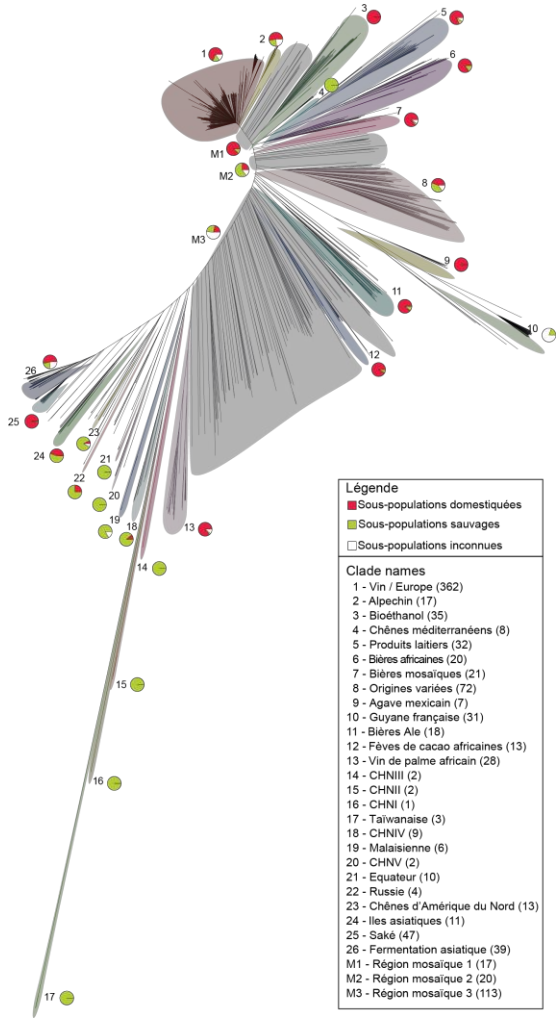


Figure 12. Arbre phylogénétique basé sur la diversité nucléotidique entre 1011 isolats naturels de *Saccharomyces cerevisiae*.
Cet arbre de type Neighbour Joining est construit à partir d'une matrice des distances établie entre les 1 544 489 sites biallélliques des 1011 souches. Plusieurs sous-populations domestiquées ou sauvages sont retrouvées et réparties au sein de 26 clades spécifiques et 3 régions mosaïques. Figure adaptée de Peter et al., 2018.

Grâce à l'étude de cette population, une large diversité de variants génétiques a été décrite et exploitée pour définir leurs origines et conséquences. Le pangénome de l'espèce est constitué d'environ 4900 gènes communs à tous les individus (génome de base) et de 2800 gènes accessoires regroupant à la fois des introgressions et des HGT (Li et al.; Peter et al., 2018). De manière intéressante, les gènes accessoires, moins conservés, impactent moins la survie des cellules et accumulent plus de mutations non-synonymes que le génome de base. La diversité des génomes de *S. cerevisiae* dévoile une multitude d'évènements d'adaptation à l'environnement, notamment par l'intermédiaire de variants structurels. À l'échelle du génome entier, des variations importantes de la ploïdie sont observées au sein de la population. Alors que les génomes diploïdes sont majoritaires, des individus polyploïdes (3n, 4n et 5n) sont fréquemment retrouvés dans des processus de domestication spécifiques. Dans des conditions expérimentales de laboratoire, le fitness de ces isolats polyploïdes est plus impacté (Figure 5A) (Fay et al., 2019; Gallone et al., 2016; Otto, 2007; Peter et al., 2018; Selmecki et al., 2015; Todd et al., 2017). Alors que la reproduction asexuée est privilégiée au sein de cette espèce, 63% des isolats, majoritairement domestiqués, présentent de l'hétérozygotie combinée à de plus ou moins larges régions de LOH (LOH pour Loss of Heterozygosity) selon les sous-populations. Par exemple, les LOH recouvrent jusqu'à 80% des génomes de souches utilisées dans la production de saké (Peter et al., 2018). Des aneuploïdies ainsi que des duplications de gènes sont également fréquemment observées dans les génomes. Les aneuploïdies ont un impact global négatif sur le fitness des individus (Figure 5B) (Peter et al., 2018; Scopel et al., 2021; Torres et al., 2007) alors que les CNV sont des réponses adaptatives à différents stress environnementaux et favorisent un intérêt biotechnologique pour les isolats domestiqués (Steenwyk and Rokas, 2018; Tan et al., 2013; Todd and Selmecki, 2020; Yona et al., 2012). Les souches de vin ont par exemple acquis de nombreux CNV pour s'adapter aux conditions environnementales, comme par exemple le gène *CUP1*, associé à la résistance au sulfate de cuivre (Fogel and Welch, 1982; Steenwyk and Rokas, 2018). En association avec ces CNV, un effet de compensation du nombre de copies a été décrit pour l'expression des gènes dupliqués, à la fois au niveau transcriptomique et protéomique (Ascencio et al., 2021; Dephoure et al., 2014; Hose et al., 2015).

En dehors des variants du nombre de copies, les réarrangements chromosomiques ont également d'importantes conséquences phénotypiques et influencent notamment l'expression des gènes adjacents (Gorkovskiy and Verstrepen, 2021; Hou et al.,

2014; Marullo et al., 2020; Naseeb and Delneri, 2012). Cependant, les génomes sont pour l'instant principalement construits sur la base de lectures courtes, limitant ainsi la détection de points d'inversion ou de translocation, entre autres évènements. De manière évidente, les perspectives à moyen terme impliquent le séquençage d'isolats avec les stratégies de type Oxford Nanopore ou PacBio afin de mieux caractériser les réarrangements chromosomiques.

Saccharomyces cerevisiae, comme outil génétique et moléculaire efficace

En plus de la diversité génétique importante présente au sein de l'espèce, la levure *S. cerevisiae* est particulièrement facile à manipuler en laboratoire. En effet, les outils de biologie moléculaire sont simples d'usage et optimisés dans cette espèce. Les remplacements alléliques par recombinaison homologue avec 50 pb d'homologie de chaque côté de la région ciblée ont longtemps été la stratégie privilégiée de mutagénèse des génomes (Shortle et al., 1982; Wach et al., 1994). Grâce à cette méthode, diverses collections de délétion des 6000 gènes ont été générées et ont notamment permis d'identifier les gènes essentiels dans la souche de référence S288C (Giaever et al., 2002; Winzeler et al., 1999). L'étude de ces gènes essentiels est également possible par l'intermédiaire d'autres techniques comme la mutagénèse par insertion de transposons. Divers outils de transposition tels que les systèmes *Hermes* (Gangadharan et al., 2010), *AcDs* (Michel et al., 2017), *PiggyBac* (Weiss et al., 2019) et *Tn7* (Sanchez et al., 2019) sont disponibles chez *S. cerevisiae*. Rapidement après son développement, le système CRISPR-Cas a aussi été adapté à la levure *S. cerevisiae* et permet de produire de nombreuses modifications génétiques différentes (DiCarlo et al., 2013). Un dérivé du système CRISPR, le système CRISPEY, a par exemple été développé pour réaliser des milliers de variants génétiques à haut-débit en une expérience unique (Sharon et al., 2018). Une approche innovante pour étudier l'impact des translocations sur les phénotypes a également été mise au point grâce à CRISPR-Cas9. De manière intéressante, la réorganisation 3D d'un même génome de *S. cerevisiae* par des translocations réciproques ciblant des régions répétées, les éléments transposables endogènes, induit une variance phénotypique considérable dans différentes conditions (Fleiss et al., 2019).

Un second avantage de cette levure repose sur son cycle cellulaire haplodiplobiontique, c'est-à-dire qu'elle peut se maintenir dans un état haploïde ou diploïde (Herskowitz, 1988). Le mode de reproduction asexuée par mitose est privilégié que ce soit à l'état haploïde ou diploïde avec un temps de génération de 1

à 2h. D'un côté, le passage de l'état diploïde à haploïde se produit par méiose lorsque les nutriments manquent dans le milieu. De l'autre côté, deux individus haploïdes de signe sexuel opposé, *MATa* et *MATα*, peuvent se croiser par reproduction sexuée pour former un diploïde hétérozygote stable. Il est ainsi simple de croiser des individus haploïdes entre eux en laboratoire, induire la méiose et disséquer les tétrades résultantes du croisement.

L'obtention rapide d'une large descendance d'un croisement (jusqu'à des milliers de ségrégants) associée à un taux de recombinaison élevé (Lynch, 2006) permet de réaliser des analyses de liaison à grande échelle pour étudier les relations génotype-phénotype (Fay, 2013; Liti and Louis, 2012; Swinnen et al., 2012). Par conséquent, de nombreux phénotypes ont été disséqués *via* les analyses de liaison. Les phénotypes ciblés sont fréquemment associés aux applications biotechnologiques de *S. cerevisiae* afin d'en améliorer différents aspects tels que la tolérance à la chaleur (Parts et al., 2011) ou encore différentes caractéristiques de la fermentation alcoolique (Bartle et al., 2021; Eder et al., 2018; Hu et al., 2007). Ces analyses de liaison permettent ainsi la détection de QTL impliqués dans ces phénotypes, qui peuvent être fonctionnellement vérifiés et précisés à l'échelle du QTN (Peltier et al., 2019). En dehors de l'étude de phénotypes d'intérêt biotechnologique, les premières bases de la régulation de l'expression des gènes ont pu être établies grâce à des analyses de liaison chez *S. cerevisiae*. Le niveau d'expression de chaque gène sert en effet de phénotype pour lequel des eQTL impliqués dans la régulation de la transcription ont été identifiés (Albert et al., 2018; Brem et al., 2002; Wittkopp, 2005; Yvert et al., 2003). À ce jour, l'étude incluant le plus d'individus – c'est-à-dire une descendance de plus de 1000 ségrégants – a révélé l'impact prédominant des variants génétiques locaux sur la variation de l'expression des gènes bien que les eQTL distants soient plus nombreux et pléiotropiques (Albert et al., 2018). Cependant, ces eQTL ne représentent pas la diversité génétique de l'espèce et la pression de sélection subie par les génomes au cours des générations. L'étude des origines de la variation des transcriptomes dans un peu moins d'une centaine d'isolats naturels (n = 85) *via* une étude d'association pangénomique a ainsi permis d'identifier avec une plus grande résolution des eQTL locaux au sein de l'espèce. Ces eQTL régulant la variation d'expression ont majoritairement été localisés dans les promoteurs et dans les régions 3'UTR des gènes impactés (Kita et al., 2017).

Le niveau d'expression des gènes fait partie de la diversité des phénotypes pouvant être étudiée et incluse dans des analyses d'association chez *S. cerevisiae* (Botstein

and Fink, 2011). En effet, des techniques de phénotypage adaptées à cette levure ont été décrites pour différents critères cellulaires telles que la croissance dans différentes conditions, le taux de sporulation ou encore la morphologie des cellules (Gerke et al., 2009; Ohya et al., 2005; Peter et al., 2018; Yvert et al., 2013), mais aussi des critères moléculaires parmi lesquels le transcriptome, le protéome ou le métabolome (Skelly et al., 2013). Les stratégies de phénotypage massives dans la population déjà génotypée de *S. cerevisiae* peuvent alors fournir de solides bases pour réaliser des études d'association pangénomique. La plus grande analyse implique actuellement 971 isolats naturels phénotypés dans 36 conditions de croissance affectant différents processus cellulaires tels que le métabolisme ou la réponse à des stress (Peter et al., 2018). L'exploration des 36 conditions testées a permis de mettre en évidence 35 variants génétiques – 22 CNV et 13 SNP – impliqués dans la variance phénotypique de 14 traits. De manière intéressante, les CNV ont un impact phénotypique environ 10 fois plus élevé que les SNP (Peter et al., 2018).

Références

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

Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat. Commun. *10*.

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

Albert, F.W., Bloom, J.S., Siegel, J., Day, L., and Kruglyak, L. (2018). Genetics of trans-regulatory variation in gene expression. Elife *7*, 1–44.

Albertin, W., and Marullo, P. (2012). Polyploidy in fungi: Evolution after whole-genome duplication. Proc. R. Soc. B Biol. Sci. *279*, 2497–2509.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K.M.M., Cao, J., Chae, E., Dezwaan, T.M.M., Ding, W., et al. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell *166*, 481–491.

Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. Nat. Rev. Genet. *21*, 71–87.

Antonarakis, S.E. (2017). Down syndrome and the complexity of genome dosage imbalance. Nat. Rev. Genet. *18*, 147–163.

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

Ascencio, D., Diss, G., Gagnon-Arsenault, I., Dubé, A.K., DeLuna, A., and Landry, C.R. (2021). Expression attenuation as a mechanism of robustness against gene duplication. Proc. Natl. Acad. Sci. U. S. A. *118*.

Auerbach, C., and Robson, J.M. (1946). Chemical Production of Mutations. Nature *157*, 302.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

Bartle, L., Peltier, E., Sundstrom, J.F., Sumby, K., Mitchell, J.G., Jiranek, V., and Marullo, P. (2021). QTL mapping: an innovative method for investigating the genetic determinism of yeast-bacteria interactions in wine. Appl. Microbiol. Biotechnol. *105*, 5053–5066.

Bateson, W. (1909). Mendel's principles of heredity, by W. Bateson. Cambridge [Eng.]University Press.

Bayer, P.E., Golicz, A.A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. Nat. Plants *6*, 914–920.

Beadle, G.W., and Tatum, E.L. (1941). Genetic Control of Biochemical Reactions in Neurospora. Proc. Natl. Acad. Sci. *27*, 499 LP – 506.

Beleza, S., Johnson, N.A., Candille, S.I., Absher, D.M., Coram, M.A., Lopes, J., Campos, J., Araújo, I.I., Anderson, T.M., Vilhjálmsson, B.J., et al. (2013). Genetic architecture of skin and eye color in an African-European admixed population. PLoS Genet. *9*, e1003372.

Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*.

Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E.M.L., and Skoglund, P. (2021). Origins of modern human ancestry. Nature *590*, 229–237.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat. Genet. *53*, 779–786.

Blattner, F.R., Plunkett, G. 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The complete genome sequence of Escherichia coli K-12. Science *277*, 1453–1462.

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

Bloom, J.S., Kotenko, I., Sadhu, M.J., Treusch, S., Albert, F.W., and Kruglyak, L. (2015). Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat. Commun. *6*, 8712.

Bloom, J.S., Boocock, J., Treusch, S., Sadhu, M.J., Day, L., Oates-Barker, H., and Kruglyak, L. (2019). Rare variants contribute disproportionately to quantitative trait variation in yeast. Elife *8*, 1–19.

Botstein, D., and Fink, G.R. (2011). Yeast: An experimental organism for 21st century biology. Genetics *189*, 695–704.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. Am J Hum Gen *32*, 314–331.

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Ventra, M. Di, Garaj, S., Hibbs, A., Jovanovich, S.B., et al. (2009). The potential and challenges of nanopore sequencing. Nat. Biotechnol. *26*, 1146–1153.

Brauer, M.J., Christianson, C.M., Pai, D.A., and Dunham, M.J. (2006). Mapping novel traits

by array-assisted bulk segregant analysis in Saccharomyces cerevisiae. Genetics *173*, 1813–1816.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science (80-.). *296*, 752–755.

Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S., and Akey, J.M. (2018). Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell *173*, 53-61.e9.

Buniello, A., Macarthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.

Burga, A., Ben-David, E., Lemus Vergara, T., Boocock, J., and Kruglyak, L. (2019). Fast genetic mapping of complex traits in C. elegans using millions of individuals in bulk. Nat. Commun. *10*, 2680.

Cain, A.K., Barquist, L., Goodman, A.L., Paulsen, I.T., Parkhill, J., and van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. Nat. Rev. Genet. *21*, 526–540.

Chandler, C.H., Chari, S., Tack, D., and Dworkin, I. (2014). Causes and consequences of genetic background effects illuminated by integrative genomic analysis. Genetics *196*, 1321–1336.

Chow, C.Y., Kelsey, K.J.P., Wolfner, M.F., and Clark, A.G. (2016). Candidate genetic modifiers of retinitis pigmentosa identified by exploiting natural variation in Drosophila. Hum. Mol. Genet. *25*, 651–659.

Christen, B., Abeliuk, E., Collier, J.M., Kalogeraki, V.S., Passarelli, B., Coller, J.A., Fero, M.J., McAdams, H.H., and Shapiro, L. (2011). The essential genome of a bacterium. Mol. Syst. Biol. *7*, 528.

Coe, K.A., Lee, W., Stone, M.C., Komazin-Meredith, G., Meredith, T.C., Grad, Y.H., and Walker, S. (2019). Multi-strain Tn-Seq reveals common daptomycin resistance determinants in Staphylococcus aureus. PLoS Pathog. *15*, e1007862.

Cook, D.E., Zdraljevic, S., Roberts, J.P., and Andersen, E.C. (2017). CeNDR, the Caenorhabditis elegans natural diversity resource. Nucleic Acids Res. *45*, D650–D657.

Coolon, J.D., Stevenson, K.R., McManus, C.J., Yang, B., Graveley, B.R., and Wittkopp, P.J. (2015). Molecular Mechanisms and Evolutionary Processes Contributing to Accelerated Divergence of Gene Expression on the Drosophila X Chromosome. Mol. Biol. Evol. *32*, 2605–2615.

Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*.

De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Sleegers, K., and Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore

PromethION sequencing of the human genome. Genome Res. *29*, 1178–1187.

De Coster, W., Weissensteiner, M.H., and Sedlazeck, F.J. (2021). Towards population-scale long-read sequencing. Nat. Rev. Genet. *30*.

Crawford, N.G., Kelly, D.E., Hansen, M.E.B., Beltrame, M.H., Fan, S., Bowman, S.L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., et al. (2017). Loci associated with skin pigmentation identified in African populations. Science *358*.

Cutting, G.R. (2010). Modifier genes in Mendelian disorders: the example of cystic fibrosis. Ann. N. Y. Acad. Sci. *1214*, 57–69.

D'Angiolo, M., Chiara, M. De, Yue, J., Irizar, A., Stenberg, S., Llored, A., Barré, B., Schacherer, J., Marangoni, R., and Gilson, E. (2020). A yeast living fossil reveals the origin of genomic introgressions. Nature *587*.

Darwin, C. (1868). The Variation of Animals and Plants Under Domestication. London John Murray. 1st Ed *2*.

Dephoure, N., Hwang, S., O'Sullivan, C., Dodgson, S.E., Gygi, S.P., Amon, A., and Torres, E.M. (2014). Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. Elife *3*, e03023.

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

Dietzl, G., Chen, D., Schnorrer, F., Su, K.-C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oppel, S., Scheiblauer, S., et al. (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature *448*, 151–156.

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

Dowell, R.D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D.A., Alexander Rolfe, P., Heisler, L.E., Chin, B., et al. (2010). Genotype to phenotype: A Complex problem. Science (80-.). *328*, 469.

Duan, S.F., Han, P.J., Wang, Q.M., Liu, W.Q., Shi, J.Y., Li, K., Zhang, X.L., and Bai, F.Y. (2018). The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. Nat. Commun. *9*.

Dujon, B. (2019). My route to the intimacy of genomes. FEMS Yeast Res. *19*.

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

Eder, M., Sanchez, I., Brice, C., Camarasa, C., Legras, J.-L., and Dequin, S. (2018). QTL mapping of volatile compound production in Saccharomyces cerevisiae during alcoholic fermentation. BMC Genomics *19*, 166.

Ehrenreich, I.M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J.A., Gresham, D., Caudy,

A.A., and Kruglyak, L. (2010). Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature *464*, 1039–1042.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science (80-.). *323*, 133–138.

Emerson, J.J., Hsieh, L.-C., Sung, H.-M., Wang, T.-Y., Huang, C.-J., Lu, H.H.-S., Lu, M.-Y.J., Wu, S.-H., and Li, W.-H. (2010). Natural selection on cis and trans regulation in yeasts. Genome Res. *20*, 826–836.

Evertts, A.G., Plymire, C., Craig, N.L., and Levin, H.L. (2007). The hermes transposon of Musca domestica is an efficient tool for the mutagenesis of Schizosaccharomyces pombe. Genetics *177*, 2519–2523.

Fay, J.C. (2013). The molecular basis of phenotypic variation in yeast. Curr. Opin. Genet. Dev. *23*, 672–677.

Fay, J.C., Liu, P., Ong, G.T., Dunham, M.J., Cromie, G.A., Jeffery, E.W., Ludlow, C.L., and Dudley, A.M. (2019). A polyploid admixed origin of beer yeasts derived from European and Asian wine populations. PLoS Biol. *17*, e3000147.

Ferguson-Smith, M.A., Aitken, D.A., Turleau, C., and de Grouchy, J. (1976). Localisation of the human ABO: Np-1: AK-1 linkage group by regional assignment of AK-1 to 9q34. Hum. Genet. *34*, 35–43.

Ferraro, N.M., Strober, B.J., Einson, J., Abell, N.S., Aguet, F., Barbeira, A.N., Brandt, M., Bucan, M., Castel, S.E., Davis, J.R., et al. (2020). Transcriptomic signatures across human tissues identify functional rare genetic variation. Science *369*.

Field, Y., Fondufe-Mittendorf, Y., Moore, I.K., Mieczkowski, P., Kaplan, N., Lubling, Y., Lieb, J.D., Widom, J., and Segal, E. (2009). Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. Nat. Genet. *41*, 438–445.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811.

Fleiss, A., O'Donnell, S., Fournier, T., Lu, W., Agier, N., Delmas, S., Schacherer, J., and Fischer, G. (2019). Reshuffling yeast chromosomes with CRISPR/Cas9. PLoS Genet. *15*, e1008332.

Flint, J., and Eskin, E. (2012). Genome-wide association studies in mice. Nat. Rev. Genet. *13*, 807–817.

Flint, J., Harding, R.M., Boyce, A.J., and Clegg, J.B. (1998). 1 The population genetics of the haemoglobinopathies. Baillieres. Clin. Haematol. *11*, 1–51.

Fogel, S., and Welch, J.W. (1982). Tandem gene amplification mediates copper resistance in yeast. Proc. Natl. Acad. Sci. U. S. A. *79*, 5342–5346.

Foss, E.J., Radulovic, D., Shaffer, S.A., Ruderfer, D.M., Bedalov, A., Goodlett, D.R., and Kruglyak, L. (2007). Genetic basis of proteome variation in yeast. Nat. Genet. *39*, 1369–

1375.

Fournier, T., Saada, O.A., Hou, J., Peter, J., Caudal, E., and Schacherer, J. (2019). Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. Elife *8*, 1–18.

Frøkjaer-Jensen, C., Davis, M.W., Hollopeter, G., Taylor, J., Harris, T.W., Nix, P., Lofgren, R., Prestgard-Duke, M., Bastiani, M., Moerman, D.G., et al. (2010). Targeted gene deletions in C. elegans using transposon excision. Nat. Methods *7*, 451–453.

Fuentes, R.R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J.F., Mohiyuddin, M., Wing, R.A., McNally, K.L., Tatarinova, T., Grigoriev, A., et al. (2019). Structural variants in 3000 rice genomes. Genome Res. *29*, 870–880.

Galardini, M., Busby, B.P., Vieitez, C., Dunham, A.S., Typas, A., and Beltrao, P. (2019). The impact of the genetic background on gene deletion phenotypes in Saccharomyces cerevisiae . Mol. Syst. Biol. *15*, 1–13.

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., et al. (2016). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. Cell *166*, 1397-1410.e16.

Gangadharan, S., Mularoni, L., Fain-Thornton, J., Wheelan, S.J., and Craig, N.L. (2010). DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. Proc. Natl. Acad. Sci. *107*, 21966–21972.

Gerke, J., Lorenz, K., and Cohen, B. (2009). Genetic interactions between transcription factors cause natural variation in yeast. Science *323*, 498–501.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387–391.

Giannakou, K., Cotterrell, M., and Delneri, D. (2020). Genomic Adaptation of Saccharomyces Species to Industrial Environments. Front. Genet. *11*, 1–10.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 Genes conveniently among the different interna- Old Questions and New Answers The genome . At the beginning of the se- of its more complex relatives in the eukary- cerevisiae has been completely sequenced Schizosaccharomyces pombe indicate. Science (80-.). *274*, 546–567.

Golicz, A.A., Batley, J., and Edwards, D. (2016). Towards plant pangenomics. Plant Biotechnol. J. *14*, 1099–1105.

Gonzales, N.M., Seo, J., Hernandez Cordero, A.I., St. Pierre, C.L., Gregory, J.S., Distler, M.G., Abney, M., Canzar, S., Lionikas, A., and Palmer, A.A. (2018). Genome wide association analysis in a mouse advanced intercross line. Nat. Commun. *9*.

Gorkovskiy, A., and Verstrepen, K.J. (2021). The Role of Structural Variation in Adaptation and Evolution of Yeast and Other Fungi. Genes (Basel). *12*.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N.,

Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.

Griffing, B. (1956). Concept of General and Specific Combining Ability in Relation to Diallel Crossing Systems. Aust. J. Biol. Sci. *9*, 463–493.

Gulcher, J., and Stefansson, K. (1998). Population genomics: Laying the groundwork for genetic disease modeling and targeting. Clin. Chem. Lab. Med. *36*, 523–527.

Guo, J., Bakshi, A., Wang, Y., Jiang, L., Yengo, L., Goddard, M.E., Visscher, P.M., and Yang, J. (2021). Quantifying genetic heterogeneity between continental populations for human height and body mass index. Sci. Rep. *11*, 1–9.

Guo, Y., Park, J.M., Cui, B., Humes, E., Gangadharan, S., Hung, S., FitzGerald, P.C., Hoe, K.L., Grewal, S.I.S., Craig, N.L., et al. (2013). Integration profiling of gene function with dense maps of transposon integration. Genetics *195*, 599–609.

Hamilton, B.A., and Yu, B.D. (2012). Modifier genes and the plasticity of genetic networks in mice. PLoS Genet. *8*, e1002644.

Harris, K., and Nielsen, R. (2016). The Genetic Cost of Neanderthal Introgression. Genetics *203*, 881–891.

Herrick, J.B. (1910). Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. 1910. Yale J. Biol. Med. *74*, 179–184.

Herskowitz, I. (1988). Life cycle of the budding yeast Saccharomyces cerevisiae. Microbiol. Rev. *52*, 536–553.

Hill, M.S., Vande Zande, P., and Wittkopp, P.J. (2021). Molecular and evolutionary processes generating variation in gene expression. Nat. Rev. Genet. *22*, 203–215.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.

Hitotsumachi, S., Carpenter, D.A., and Russell, W.L. (1985). Dose-repetition increases the mutagenic effectiveness of N-ethyl-N-nitrosourea in mouse spermatogonia. Proc. Natl. Acad. Sci. U. S. A. *82*, 6619–6621.

Hose, J., Yong, C.M., Sardi, M., Wang, Z., Newton, M.A., and Gasch, A.P. (2015). Dosage compensation can buffer copynumber variation in wild yeast. Elife *4*, 1–27.

Hou, J., Friedrich, A., de Montigny, J., and Schacherer, J. (2014). Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in Saccharomyces cerevisiae. Curr. Biol. *24*, 1153–1159.

Hou, J., Tan, G., Fink, G.R., Andrews, B.J., and Boone, C. (2019). Complex modifier landscape underlying genetic background effects. Proc. Natl. Acad. Sci. U. S. A. *116*, 5045–5054.

Hrabé de Angelis, M.H., Flaswinkel, H., Fuchs, H., Rathkolb, B., Soewarto, D., Marschall, S., Heffner, S., Pargent, W., Wuensch, K., Jung, M., et al. (2000). Genome-wide, large-scale production of mutant mice by ENU mutagenesis. Nat. Genet. *25*, 444–447.

Hu, X.H., Wang, M.H., Tan, T., Li, J.R., Yang, H., Leach, L., Zhang, R.M., and Luo, Z.W. (2007). Genetic dissection of ethanol tolerance in the budding yeast Saccharomyces cerevisiae. Genetics *175*, 1479–1487.

Jablonski, N.G., and Chaplin, G. (2000). The evolution of human skin coloration. J. Hum. Evol. *39*, 57–106.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.

Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. *17*, 1–11.

Jakobson, C.M., and Jarosz, D.F. (2019). Molecular Origins of Complex Heritability in Natural Genotype-to-Phenotype Relationships. Cell Syst. *8*, 363-379.e3.

Jansen, R.C., and Nap, J.P. (2001). Genetical genomics: the added value from segregation. Trends Genet. *17*, 388–391.

Jansen, P.R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A.R., de Leeuw, C.A., Benjamins, J.S., Muñoz-Manchado, A.B., Nagel, M., et al. (2019). Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. Nat. Genet. *51*, 394–403.

Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., and Sedlazeck, F.J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. *8*, 1–11.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Johnson, M.S., Martsul, A., Kryazhimskiy, S., and Desai, M.M. (2019). Higher-fitness yeast genotypes are less robust to deleterious mutations. Science *366*, 490–493.

Juric, I., Aeschbacher, S., and Coop, G. (2016). The Strength of Selection against Neanderthal Introgression. PLoS Genet. *12*, e1006340.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature *421*, 231–237.

Kariuki, S.N., and Williams, T.N. (2020). Human genetics and malaria resistance. Hum. Genet. *139*, 801–811.

Kawakatsu, T., Huang, S. shan C., Jupe, F., Sasaki, E., Schmitz, R.J.J., Urich, M.A.A., Castanon, R., Nery, J.R.R., Barragan, C., He, Y., et al. (2016). Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell *166*, 492–505.

Kiger, A.A., Baum, B., Jones, S., Jones, M.R., Coulson, A., Echeverri, C., and Perrimon, N. (2003). A functional genomic analysis of cell morphology using RNA interference. J. Biol. *2*, 27.

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., et al. (2010). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat. Biotechnol. *28*, 617–623.

Kita, R., Venkataram, S., Zhou, Y., and Fraser, H.B. (2017). High-resolution mapping of cis-regulatory variation in budding yeast. Proc. Natl. Acad. Sci. U. S. A. *114*, E10736–E10744.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385–389.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1082.

Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., Usaj, M., Balint, A., Usaj, M.M., Van Leeuwen, J., et al. (2018). Systematic analysis of complex genetic interactions. Science (80-.). *360*.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lauer, S., Avecilla, G., Spealman, P., Sethia, G., Brandt, N., Levy, S., and Gresham, D. (2018). Single-cell copy number variant detection reveals the dynamics and diversity of adaptation.

Legras, J.L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., et al. (2018). Adaptation of s. Cerevisiae to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. Mol. Biol. Evol. *35*, 1712–1727.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565–570.

Li, G., Ji, B., and Nielsen, J. The pan-genome of Saccharomyces cerevisiae.

Li, J., Kong, N., Han, B., and Sul, J.H. (2021). Rare variants regulate expression of nearby individual genes in multiple tissues. PLoS Genet. *17*, 1–26.

Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al. (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.

Lichtenstein, P., Yip, B.H., Björk, C., Pawitan, Y., Cannon, T.D., Sullivan, P.F., and Hultman, C.M. (2009). Common genetic influences for schizophrenia and bipolar disorder: A population-based study of 2 million nuclear families. Lancet *373*, 1–14.

Lickwar, C.R., Mueller, F., Hanlon, S.E., McNally, J.G., and Lieb, J.D. (2012). Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature *484*, 251–255.

Liti, G., and Louis, E.J. (2012). Advances in quantitative trait analysis in yeast. PLoS Genet.

8, e1002912.

Liu, F., Visser, M., Duffy, D.L., Hysi, P.G., Jacobs, L.C., Lao, O., Zhong, K., Walsh, S., Chaitanya, L., Wollstein, A., et al. (2015). Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. Hum. Genet. *134*, 823–835.

Liu, Y., Borel, C., Li, L., Müller, T., Williams, E.G., Germain, P.-L., Buljan, M., Sajic, T., Boersema, P.J., Shao, W., et al. (2017). Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. Nat. Commun. *8*, 1212.

Lloyd-Jones, L.R., Robinson, M.R., Moser, G., Zeng, J., Beleza, S., Barsh, G.S., Tang, H., and Visscher, P.M. (2017). Inference on the Genetic Basis of Eye and Skin Color in an Admixed Population via Bayesian Linear Mixed Models. Genetics *206*, 1113–1126.

Lubliner, S., Regev, I., Lotan-Pompan, M., Edelheit, S., Weinberger, A., and Segal, E. (2015). Core promoter sequence in yeast is a major determinant of expression level. Genome Res. *25*, 1008–1017.

Lutz, S., Brion, C., Kliebhan, M., and Albert, F.W. (2019). DNA variants affecting the expression of numerous genes in trans have diverse mechanisms of action and evolutionary histories. PLoS Genet. *15*, e1008375.

Lynch, M. (2006). The origins of eukaryotic gene structure. Mol. Biol. Evol. *23*, 450–468.

Mackay, T.F. (2001). The genetic architecture of quantitative traits. Annu. Rev. Genet. *35*, 303–339.

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

Magwene, P.M., Willis, J.H., and Kelly, J.K. (2011). The statistics of bulk segregant analysis using next generation sequencing. PLoS Comput. Biol. *7*, e1002255.

Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature *456*, 18–21.

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

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

Martin, R.G., Matthaei, J.H., Jones, O.W., and Nirenberg, M.W. (1961). Ribonucleotide composition of the genetic code. Biochem. Biophys. Res. Commun. *6*, 410–414.

Marullo, P., Claisse, O., Raymond Eder, M.L., Börlin, M., Feghali, N., Bernard, M., Legras, J.L., Albertin, W., Rosa, A.L., and Masneuf-Pomarede, I. (2020). SSU1 Checkup, a Rapid Tool for Detecting Chromosomal Rearrangements Related to the SSU1 Promoter in Saccharomyces cerevisiae: An Ecological and Technological Study on Wine Yeast. Front. Microbiol. *11*, 1–14.

Massilani, D., Skov, L., Hajdinjak, M., Gunchinsuren, B., Tseveendorj, D., Yi, S., Lee, J., Nagel, S., Nickel, B., Devièse, T., et al. (2020). Denisovan ancestry and population history of early East Asians. Science (80-.). *370*, 579 LP – 583.

Mavaddat, N., Peock, S., Frost, D., Ellis, S., Platte, R., Fineberg, E., Evans, D.G., Izatt, L., Eeles, R.A., Adlard, J., et al. (2013). Cancer risks for BRCA1 and BRCA2 mutation carriers: results from prospective analysis of EMBRACE. J. Natl. Cancer Inst. *105*, 812–822.

Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. U. S. A. *74*, 560–564.

Mccarthy, C.G.P., and Fitzpatrick, D.A. (2019). Pan-genome analyses of model fungal species. 1–23.

McCoy, R.C., Wakefield, J., and Akey, J.M. (2017). Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. Cell *168*, 916-927.e12.

Mendel, G. (1866). Versuche über Plflanzenhybriden. Verhandlungen Des Naturforschenden Vereines Brünn *Bd. IV*.

Metzger, B.P.H., and Wittkopp, P.J. (2019). Compensatory trans -regulatory alleles minimizing variation in TDH3 expression are common within Saccharomyces cerevisiae . Evol. Lett. *3*, 448–461.

Michel, A.H., Hatakeyama, R., Kimmig, P., Arter, M., Peter, M., Matos, J., De Virgilio, C., and Kornmann, B.T. (2017). Functional mapping of yeast genomes by saturated transposition. Elife *6*.

Morgan, T.H., Sturtevant, A.H., Muller, H.J., and Bridges, C.B. (1915). The Mechanism of Mendelian Heredity. New York,Holt.

Muller, H.J. (1928). The Production of Mutations by X-Rays. Proc. Natl. Acad. Sci. U. S. A. *14*, 714–726.

Mullis, M.N., Matsui, T., Schell, R., Foree, R., and Ehrenreich, I.M. (2018). The complex underpinnings of genetic background effects. Nat. Commun. *9*, 1–10.

Naseeb, S., and Delneri, D. (2012). Impact of chromosomal inversions on the yeast DAL cluster. PLoS One *7*, e42022.

Nikolac Perkovic, M., and Pivac, N. (2019). Genetic Markers of Alzheimer's Disease. Adv. Exp. Med. Biol. *1192*, 27–52.

Nolan, P.M., Peters, J., Strivens, M., Rogers, D., Hagan, J., Spurr, N., Gray, I.C., Vizor, L., Brooker, D., Whitehill, E., et al. (2000). A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. Nat. Genet. *25*, 440–443.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V, Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2021). The complete sequence of a human genome. BioRxiv 2021.05.26.445798.

Nurse, P., and Thuriaux, P. (1980). Regulatory genes controlling mitosis in the fission yeast schizosaccharomyces pombe. Genetics *96*, 627 LP – 637.

Nurse, P., Thuriaux, P., and Nasmyth, K. (1976). Genetic control of the cell division cycle in the fission yeast Schizosaccharomyces pombe. Mol. Gen. Genet. *146*, 167–178.

Ohno, S. (1970). Evolution by Gene Duplication. Heidelb. Springer 98–106.

Ohya, Y., Sese, J., Yukawa, M., Sano, F., Nakatani, Y., Saito, T.L., Saka, A., Fukuda, T., Ishihara, S., Oka, S., et al. (2005). High-dimensional and large-scale phenotyping of yeast mutants. Proc. Natl. Acad. Sci. U. S. A. *102*, 19015 LP – 19020.

van Opijnen, T., and Levin, H.L. (2020). Transposon Insertion Sequencing, a Global Measure of Gene Function. Annu. Rev. Genet. *54*, 337–365.

van Opijnen, T., Bodi, K.L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods *6*, 767–772.

Otto, S.P. (2007). The Evolutionary Consequences of Polyploidy. Cell *131*, 452–462.

Paaby, A.B., White, A.G., Riccardi, D.D., Gunsalus, K.C., Piano, F., and Rockman, M. V. (2015). Wild worm embryogenesis harbors ubiquitous polygenic modifier variation. Elife *4*, 1–17.

Parts, L., Cubillos, F.A., Warringer, J., Jain, K., Salinas, F., Bumpstead, S.J., Molin, M., Zia, A., Simpson, J.T., Quail, M.A., et al. (2011). Revealing the genetic structure of a trait by sequencing a population under selection. Genome Res. *21*, 1131–1138.

Parts, L., Batté, A., Lopes, M., Yuen, M.W., Laver, M., San Luis, B.-J., Yue, J.-X., Pons, C., Eray, E., Aloy, P., et al. (2021). Natural variants suppress mutations in hundreds of essential genes. Mol. Syst. Biol. *17*, e10138.

Pasmant, E., Vidaud, M., Vidaud, D., and Wolkenstein, P. (2012). Neurofibromatosis type 1: From genotype to phenotype. J. Med. Genet. *49*, 483–489.

Pasteur, L. (1858). Nouveaux faits concernant l'histoire de la fermentation alcoolique. Comptes Rendus Chim. *47*, 1011–1013.

Patterson, D. (2009). Molecular genetic analysis of Down syndrome. Hum. Genet. *126*, 195–214.

Payen, C., Di Rienzi, S.C., Ong, G.T., Pogachar, J.L., Sanchez, J.C., Sunshine, A.B., Raghuraman, M.K., Brewer, B.J., and Dunham, M.J. (2014). The Dynamics of Diverse Segmental Amplifications in Populations of Saccharomyces cerevisiae Adapting to Strong Selection. G3 Genes, Genomes, Genet. *4*, 399–409.

Payne, A., Holmes, N., Rakyan, V., and Loose, M. (2019). BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics *35*, 2193–2198.

Peltier, E., Friedrich, A., Schacherer, J., and Marullo, P. (2019). Quantitative trait nucleotides impacting the technological performances of industrial saccharomyces cerevisiaestrains. Front. Genet. *10*.

Pérez-Ortín, J.E., Querol, A., Puig, S., and Barrio, E. (2002). Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. Genome

Res. *12*, 1533–1539.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature *556*, 339–344.

Petr, M., Pääbo, S., Kelso, J., and Vernot, B. (2019). Limits of long-term selection against Neandertal introgression. Proc. Natl. Acad. Sci. U. S. A. *116*, 1639–1644.

Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. Science *358*, 655–658.

Ramaswami, G., and Geschwind, D.H. (2018). Genetics of autism spectrum disorder. Handb. Clin. Neurol. *147*, 321–329.

Ratjen, F., Bell, S.C., Rowe, S.M., Goss, C.H., Quittner, A.L., and Bush, A. (2015). Cystic fibrosis. Nat. Rev. Dis. Prim. *1*, 15010.

Rees, D.C., Williams, T.N., and Gladwin, M.T. (2010). Sickle-cell disease. Lancet *376*, 2018–2031.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature *468*, 1053–1060.

Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics, Proteomics Bioinforma. *13*, 278–289.

Rockman, M. V, Skrovanek, S.S., and Kruglyak, L. (2010). Selection at linked sites shapes heritable phenotypic variation in C. elegans. Science *330*, 372–376.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. (2000). Comparative genomics of the eukaryotes. Science *287*, 2204–2215.

Ryder, E., Ashburner, M., Bautista-Llacer, R., Drummond, J., Webster, J., Johnson, G., Morley, T., Chan, Y.S., Blows, F., Coulson, D., et al. (2007). The DrosDel deletion collection: a Drosophila genomewide chromosomal deficiency resource. Genetics *177*, 615–629.

Sanchez, M.R., Payen, C., Cheong, F., Hovde, B.T., Bissonnette, S., Arkin, A.P., Skerker, J.M., Brem, R.B., Caudy, A.A., and Dunham, M.J. (2019). Transposon insertional mutagenesis in Saccharomyces uvarum reveals trans-acting effects influencing species-dependent essential genes. Genome Res. *29*, 396–406.

Sandegren, L., and Andersson, D.I. (2009). Bacterial gene amplification: Implications for the evolution of antibiotic resistance. Nat. Rev. Microbiol. *7*, 578–588.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. *74*, 5463–5467.

Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes,

J.C., Hutchison, C.A. 3rd, Slocombe, P.M., and Smith, M. (1978). The nucleotide sequence of bacteriophage phiX174. J. Mol. Biol. *125*, 225–246.

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. PLoS Genet. *8*, e1002947.

Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Curr. Biol. *26*, 1241–1247.

Santiago, M., Matano, L.M., Moussa, S.H., Gilmore, M.S., Walker, S., and Meredith, T.C. (2015). A new platform for ultra-high density Staphylococcus aureus transposon libraries. BMC Genomics *16*, 252.

Sauer, B., and Henderson, N. (1988). Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. Proc. Natl. Acad. Sci. U. S. A. *85*, 5166–5170.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. Nature *422*, 297–302.

Schadt, E.E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. Hum. Mol. Genet. *19*, 227–240.

Scopel, E.F.C., Hose, J., Bensasson, D., and Gasch, A.P. (2021). Genetic variation in aneuploidy prevalence and tolerance across Saccharomyces cerevisiae lineages. Genetics *217*.

Scossa, F., Alseekh, S., and Fernie, A.R. (2021). Integrating multi-omics data for crop improvement. J. Plant Physiol. *257*, 153352.

Scott, A.J., Chiang, C., and Hall, I.M. (2021). Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. BioRxiv 2021.03.06.434233.

Segrè, A. V, Murray, A.W., and Leu, J.-Y. (2006). High-resolution mutation mapping reveals parallel experimental evolution in yeast. PLoS Biol. *4*, e256.

Selmecki, A.M., Maruvka, Y.E., Richmond, P.A., Guillet, M., Shoresh, N., Sorenson, A.L., De, S., Kishony, R., Michor, F., Dowell, R., et al. (2015). Polyploidy can drive rapid adaptation in yeast. Nature *519*, 349–352.

Sharon, E., Chen, S.-A.A., Khosla, N.M., Smith, J.D., Pritchard, J.K., and Fraser, H.B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. Cell 1–14.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. Nature *550*, 345–353.

Sherman, R.M., and Salzberg, S.L. (2020). Pan-genomics in the human genome era. Nat.

Rev. Genet. *21*, 243–254.

Shortle, D., Haber, J.E., and Botstein, D. (1982). Lethal disruption of the yeast actin gene by integrative DNA transformation. Science *217*, 371–373.

Signor, S.A., and Nuzhdin, S. V (2018). The Evolution of Gene Expression in cis and trans. Trends Genet. *34*, 532–544.

Simonti, C.N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D.S., Chisholm, R.L., Crosslin, D.R., Hebbring, S.J., Jarvik, G.P., Kullo, I.J., et al. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. Science (80-.). *351*, 737–741.

Skarnes, W.C., Rosen, B., West, A.P., Koutsourakis, M., Bushell, W., Iyer, V., Mujica, A.O., Thomas, M., Harrow, J., Cox, T., et al. (2011). A conditional knockout resource for the genome-wide study of mouse gene function. Nature *474*, 337–342.

Skelly, D.A., Merrihew, G.E., Riffle, M., Connelly, C.F., Kerr, E.O., Johansson, M., Jaschob, D., Graczyk, B., Shulman, N.J., Wakefield, J., et al. (2013). Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Res. *23*, 1496–1504.

Skov, L., Coll Macià, M., Sveinbjörnsson, G., Mafessoni, F., Lucotte, E.A., Einarsdóttir, M.S., Jonsson, H., Halldorsson, B., Gudbjartsson, D.F., Helgason, A., et al. (2020). The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. Nature *582*, 78–83.

Soo, V.W.C., Hanson-Manful, P., and Patrick, W.M. (2011). Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. *108*, 1484–1489.

Souciet, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., de Montigny, J., Dujon, B., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. FEBS Lett. *487*, 3–12.

Steenwyk, J.L., and Rokas, A. (2018). Copy Number Variation in Fungi and Its Implications for Wine Yeast Genetic Diversity and Adaptation. Front. Microbiol. *9*, 288.

Steinmetz, L.M., Sinha, H., Richards, D.R., Spiegelman, J.I., Oefner, P.J., McCusker, J.H., and Davis, R.W. (2002). Dissecting the architecture of a quantitative trait locus in yeast. Nature *416*, 326–330.

Struhl, G. (1981). A gene product required for correct initiation of segmental determination in Drosophila. Nature *293*, 36–41.

Swinnen, S., Schaerlaekens, K., Pais, T., Claesen, J., Hubmann, G., Yang, Y., Demeke, M., Foulquié-Moreno, M.R., Goovaerts, A., Souvereyns, K., et al. (2012). Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis. Genome Res. *22*, 975–984.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484.

Tan, Z., Hays, M., Cromie, G.A., Jeffery, E.W., Scott, A.C., Ahyong, V., Sirr, A., Skupin, A., and Dudley, A.M. (2013). Aneuploidy underlies a multicellular phenotypic switch. Proc. Natl. Acad. Sci. U. S. A. *110*, 12367–12372.

Tao, Y., Zhao, X., Mace, E., Henry, R., and Jordan, D. (2019). Exploring and Exploiting Pan-genomics for Crop Improvement. Mol. Plant *12*, 156–169.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S. V, Crabtree, J., Jones, A.L., Durkin, A.S., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U. S. A. *102*, 13950–13955.

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

The GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

Thierry, A., Khanna, V., Créno, S., Lafontaine, I., Ma, L., Bouchier, C., and Dujon, B. (2015). Macrotene chromosomes provide insights to a new mechanism of high-order gene amplification in eukaryotes. Nat. Commun. *6*, 6154.

Thierry, A., Khanna, V., and Dujon, B. (2016). Massive Amplification at an Unselected Locus Accompanies Complex Chromosomal Rearrangements in Yeast. G3 (Bethesda). *6*, 1201–1215.

Timmons, L., Court, D.L., and Fire, A. (2001). Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. Gene *263*, 103–112.

Tirosh, I., Barkai, N., and Verstrepen, K.J. (2009). Promoter architecture and the evolvability of gene expression. J. Biol. *8*, 95.

Todd, R.T., and Selmecki, A. (2020). Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs. Elife *9*, 1–33.

Todd, R.T., Forche, A., and Selmecki, A. (2017). Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution. Microbiol. Spectr. *5*.

Torres, E.M., Sokolsky, T., Tucker, C.M., Chan, L.Y., Boselli, M., Dunham, M.J., and Amon, A. (2007). Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science *317*, 916–924.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

Vu, V., Verster, A.J., Schertzberg, M., Chuluunbaatar, T., Spensley, M., Pajkic, D., Hart, G.T., Moffat, J., and Fraser, A.G. (2015). Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. Cell *162*, 391–402.

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

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. *51*, 1339–1348.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Watson, J.D., and Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature *171*, 737–738.

Weiss, C. V., Chuong, J.N., and Brem, R.B. (2019). Genetic Mapping of Thermotolerance Differences Between Species of Saccharomyces Yeast via Genome-Wide Reciprocal Hemizygosity Analysis. J. Vis. Exp. 1–8.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. *37*, 1155–1162.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science *285*, 901–906.

Wittkopp, P.J. (2005). Genomic sources of regulatory variation in cis and in trans. Cell. Mol. Life Sci. *62*, 1779–1783.

Wittkopp, P.J., and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. *13*, 59–69.

Wolfe, K.H. (2015). Origin of the yeast whole-genome duplication. PLoS Biol. *13*, 1–7.

Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature *387*, 708–713.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

Yamamoto, F.I., Clausen, H., White, T., Marken, J., and Hakomori, S.I. (1990). Molecular genetic basis of the histo-blood group ABO system. Nature *345*, 229–233.

Yang, M.A., Malaspinas, A.-S., Durand, E.Y., and Slatkin, M. (2012). Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity. Mol. Biol. Evol. *29*, 2987–2995.

Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., and Visscher, P.M. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. Hum. Mol. Genet. *27*, 3641–3649.

Yona, A.H., Manor, Y.S., Herbst, R.H., Romano, G.H., Mitchell, A., Kupiec, M., Pilpel, Y., and Dahan, O. (2012). Chromosomal duplication is a transient evolutionary solution to stress. Proc. Natl. Acad. Sci. U. S. A. *109*, 21010–21015.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R., and Kruglyak, L. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat. Genet. *35*, 57–64.

Yvert, G., Ohnuki, S., Nogami, S., Imanaga, Y., Fehrmann, S., Schacherer, J., and Ohya, Y. (2013). Single-cell phenomics reveals intra-species variation of phenotypic noise in yeast. BMC Syst. Biol. *7*, 54.

Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. *16*, 172–183.

Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., et al. (2019). 1,520 Reference Genomes From Cultivated Human Gut Bacteria Enable Functional Microbiome Analyses. Nat. Biotechnol. *37*, 179–185.

VUE D’ENSEMBLE DU PROJET

La compréhension de l'origine génétique de la variance phénotypique observée au sein de populations naturelles est une problématique majeure en biologie. En effet, cette variance repose en majorité sur des origines génétiques complexes, c'est-à-dire que plusieurs variants sont responsables du trait étudié par des effets d'additivité, de dominance ou même d'interaction. Ces dernières décennies, les nombreuses avancées technologiques, notamment dans le cadre du séquençage des génomes, ont permis d'avoir une première vue d'ensemble des variants génétiques présents au sein de populations naturelles. Le nombre et le type de variants génétiques, allant de variations nucléotidiques à de larges remaniements génomiques, révèlent ainsi une diversité importante des génomes au sein des espèces. À l'heure actuelle, diverses stratégies – de génétique classique, d'analyses de liaison ou encore d'études d'association pangénomique – ont permis de poser les premières bases permettant d'explorer la relation entre génotype-phénotype au sein d'organismes modèles ou de populations humaines. Cependant, l'ensemble de la complexité des traits reste difficile à capturer et les variants détectés n'expliquent qu'en partie la variance phénotypique observée. De nombreux facteurs sont limitants dans ces études tels que le nombre d'individus ou de phénotypes inclus par exemple.

L'objectif de mon projet de thèse a consisté à tirer profit des avancées technologiques et des ressources à disposition afin de mieux caractériser l'origine génétique de la complexité des traits. Pour cela, la levure *Saccharomyces cerevisiae* est un modèle de choix et permet la mise en place de stratégies innovantes pour explorer différents aspects de la variance phénotypique. De plus, au laboratoire, une collection de plus de 1000 isolats est disponible et a récemment été entièrement séquencée. Caractéristique de la diversité géographique et écologique de l'espèce, cette collection permet d'avoir une bonne représentativité de la diversité génétique, et représente les fondations de cette exploration.

Malgré la disponibilité de nombreux outils génétiques et moléculaires et d'une collection de plus de 1000 individus génétiquement divers pour cette espèce de levure, la caractérisation de l'architecture génétique des traits reste limitée au nombre de phénotypes étudiés. Une étude d'association pangénomique réalisée pour 36 conditions de croissance a, par exemple, permis de mettre en évidence des variants génétiques, SNP ou CNV, n'expliquant qu'une faible partie de la variance phénotypique observée, à savoir, en moyenne et respectivement, 4,49% ou 36,8%. Afin d'englober un nouvel aspect de l'établissement des phénotypes, nous avons alors considéré une des étapes intermédiaires entre le génotype et le phénotype final,

à savoir le niveau de transcription des ~6000 gènes constituant le génome de *S. cerevisiae*. L'analyse de la variation de l'expression des gènes et de leur régulation génétique fait l'objet de mon **premier chapitre**. Deux objectifs majeurs se dégagent de cet axe. Le premier repose sur des aspects techniques et le défi que représente la « génération » des transcriptomes d'un millier d'individus. Le second consiste en l'analyse des origines génétiques de la régulation de l'expression des gènes chez *S. cerevisiae* sur la base de l'ensemble des données générées. Une stratégie à haut-débit d'extraction et de séquençage (RNA-seq) des ARNm a été mise en place pour générer les transcriptomes de 1010 isolats naturels. Sur la base de ces données, nous nous sommes attelés à mettre en évidence les caractéristiques des variations de l'expression des gènes au sein de l'espèce et d'en déterminer les origines génétiques. Ainsi, une exploration des modules de co-expression dans l'espèce a révélé les processus les plus exprimés, à savoir la synthèse protéique et la glycolyse notamment, et les réseaux de gènes les moins exprimés. Ces gènes peu exprimés sont accessoires, c'est-à-dire absents de certains génomes, ou impliqués dans les processus méiotiques et la reproduction sexuée. Ces résultats ont pu être mis en lien et confirmés respectivement par l'analyse de l'expression des gènes à l'échelle du pangénome et par des signatures transcriptionnelles associées à la ploïdie des isolats. Des signatures transcriptionnelles spécifiques à certaines sous-populations de *S. cerevisiae* ont également été identifiées et mises en relation avec leur implication dans des processus industriels. Ce jeu de données considérable permet aussi d'analyser la régulation de la transcription des gènes. Par exemple, un mécanisme de compensation de l'expression des gènes est constaté en réponse aux variations du nombre de copies dans les génomes. Également, une étude d'association pangénomique a pu être réalisée en considérant la variation d'expression pour chaque gène dans 969 isolats comme un phénotype. Cette étude d'association a mis en évidence des variants génétiques régulateurs, ou eQTL, influençant l'expression des gènes à distance ou localement ainsi que leur contribution à la variance phénotypique. Ces travaux ont ainsi fourni pour la première fois une vision d'ensemble de la régulation génétique de l'expression des gènes au sein d'une population de plusieurs centaines de levures.

Dans l'exploration des relations génotype-phénotype, il a été mis en évidence une complexité sous-jacente même dans des cas dits simples ou monogéniques. En effet, alors que dans ces cas, une mutation est responsable d'un phénotype, l'impact du fonds génétique des individus avec la mutation peut entraîner une variation du phénotype résultant (sévérité, délais d'apparition pour une maladie, par exemple).

Cependant, la prévalence de ces effets du fonds génétique dans les génomes n'a jamais été estimée et caractérisée à l'échelle d'une espèce. Mon **second chapitre** explore ainsi cette problématique à l'aide d'une stratégie haut-débit de saturation des génomes en transposons, appliquée à plus de 100 isolats naturels de *S. cerevisiae* représentatifs de la diversité de l'espèce. Nous avons, dans ce cadre, pu étudier l'impact de plusieurs milliers d'interruptions de gènes par des transposons dans ces différents fonds génétiques. Le séquençage puis la comparaison des sites d'insertion des transposons dans les génomes révèlent des motifs particuliers dans les gènes et leurs régions adjacentes, selon l'impact phénotypique engendré par les insertions. Une absence de détection d'insertions de transposons dans un gène et son promoteur reflète une perte de fitness engendrée par l'insertion. À l'inverse, de nombreuses insertions sont détectées dans les gènes pour lesquels l'insertion de transposon n'a pas d'effet majeur sur le fitness. Un modèle logistique basé sur ces motifs particuliers a été mis en place afin de prédire l'impact d'une perte de fonction pour chaque gène sur le fitness du fonds génétique considéré. Ces valeurs prédictives ont ainsi permis d'estimer la proportion à l'échelle du génome de gain et de perte de fitness dans les différents fonds génétiques par rapport à la souche de référence S288C. L'analyse des gènes impactés par ces effets du fonds génétique révèle des évènements de variation de fitness liées à l'environnement, à savoir une compétition en milieu composé de galactose ainsi que des évènements plus rares (~1/3) strictement liés au fonds génétique.

CHAPITRE I

Species-wide exploration of the inherited gene expression variation in yeast

Introduction

In the last decade, population genomics allowed to explore the genetic diversity at a population-scale via high-throughput sequencing strategies. Large resequencing surveys including thousands of individuals from the same species were initiated in humans (Auton et al., 2015) as well as in different model organisms such as the plant *Arabidopsis thaliana* (Alonso-Blanco et al., 2016), the budding yeast *Saccharomyces cerevisiae* (Peter et al., 2018) and the worm *Caenorhabditis elegans* (Lee et al., 2021). The main goal of these population genomic studies was to compare and explore the genetic variants identified across large populations. The determination of such catalogues of variants has made it possible to better understand how genetic diversity has been elaborated and maintained over generations. Within the *S. cerevisiae* species, a maximum nucleotide diversity of 1.8% has been observed across a population of 1,011 natural isolates of various ecological origins around the world (Peter et al., 2018). These variants uncovered a strong population structure with 26 specific lineages, reflecting the ecological and geographical origins of the isolates (Peter et al., 2018). Besides SNP (Single-Nucleotide Polymorphism) divergence, other sources of genetic variations are responsible of the observed genetic diversity such as Copy Number Variants (CNVs) and ploidy level variations (Gallone et al., 2016; Gorkovskiy and Verstrepen, 2021; Otto, 2007; Peter et al., 2018).

The detection of genetic variants in thousands of individuals also offers the possibility of relating genotype and phenotype. By performing genome-wide association studies (GWAS), genetic variants present in the population can be statistically associated with a given trait (Visscher et al., 2012, 2017). In recent years, association studies have helped to dissect the genetic origins of complex traits such as diseases, growth measures or resistance to chemical compounds in humans as well as in different model organisms (Alonso-Blanco et al., 2016; Buniello et al., 2019; Cook et al., 2017; Gonzales et al., 2018; Peter et al., 2018; Read and Massey, 2014; Tam et al., 2019). In yeast, genome-wide association studies were performed on a collection of 1,011 diverse isolates for different traits by estimating strain growth under different conditions (Peter et al., 2018). Interestingly, it was found that the associated CNVs explained a higher proportion of the phenotypic variance observed within the population, with a median of 36.8% versus 4.49% for associated SNPs (Peter et al., 2018). However, despite the power of GWAS strategies, the genetic architecture of complex traits remains poorly understood, in particular because the

number of traits included in the association analysis for the same population is still limited.

The understanding of the genetic regulation of the different molecular intermediates leading to the final phenotypes is essential to uncover a greater fraction of the heritability of complex traits. Indeed, gene expression variation has been widely associated with diverse phenotypic variations (Albert and Kruglyak, 2015; Hill et al., 2021). Genetic variants associated with gene expression variation, or expression Quantitative Trait Loci (eQTL), could explain a part of the complex phenotypic variance. Genome-wide screen of expression variation through linkage analysis uncovered general mechanisms of transcription regulation in model organisms such as *S. cerevisiae* (Albert et al., 2018; Brem et al., 2002) or *C. elegans* (Rockman et al., 2010). Local and distant eQTL can affect gene expression levels (Albert and Kruglyak, 2015; Rockman and Kruglyak, 2006). While local or *cis*-regulatory variants directly influence gene expression through mutations in the promoter for example, distant or *trans*-eQTL impact distant non-coding or coding genomic regions which mediate expression regulation, such as transcription factors. The main results of these linkage analysis revealed a higher impact of local eQTL on phenotypic variance but more distant eQTL with low effect on the gene expression level (Albert et al., 2018). Association studies were also performed to explore the impact of the broad genetic diversity on expression regulation at the species-wide level. To date, only few studies took the advantage of a large natural population to study the genotype-phenotype relationship using gene expression level as a trait (Kawakatsu et al., 2016; Kita et al., 2017; Skelly et al., 2013; The GTEx Consortium, 2017, 2020). In yeast for instance, gene expression variation was explored on a small subset of 85 diverse *S. cerevisiae* isolates and unveiled local mechanisms of transcription regulation as well as negative selection of regulatory variants (Kita et al., 2017). The largest analysis focused on 49 human tissues over 838 individuals (The GTEx Consortium, 2017, 2020). Similarly, mostly local eQTL were detected across tissues to explain the genetic basis of expression variations in humans (The GTEx Consortium, 2017, 2020).

The genetic basis of gene expression regulation was overall dissected via linkage analysis and did not take advantage of the entire species-wide genetic diversity. A discrepancy was observed in the general genome-wide transcriptome regulation between linkage and association analysis results. While linkage analysis uncovered both distant and local regulatory variants, association studies mainly focused on the

local eQTL. The sample sizes of the species-wide analysis were limiting so far. Indeed, previous analysis showed that distant eQTL have a lower effect size compared to local eQTL (Albert et al., 2018; Gilad et al., 2008). As a result, the statistical power of GWAS to detect such low effect variants was limited by a too small sample size (The GTEx Consortium, 2020). In humans for instance, except for 3 tissues, less than 600 donors were included in association studies, with 282 donors per tissue on average (The GTEx Consortium, 2020). In order to overcome these limitations, a greater number of individuals must be included in the analysis of the transcriptional landscape. An organism such as *S. cerevisiae* is a suitable model due to the great genetic diversity and the ability to have an exhaustive dataset in order to study the regulation of expression for a large population.

Here, we took advantage of the well-described and completely sequenced set of 1,011 yeast isolates to explore the genetic origins of gene expression variation. This large sample size increased the power of GWAS, allowing for in-depth characterization of local and distant regulatory variants impacting gene expression variation. In addition, this comprehensive dataset provided a species-wide overview of the transcriptional landscape. Overall, a tight gene expression regulation was observed across the population with strong co-expressed networks. A discrepancy of the expression between accessory and core genome as well as a dosage compensation of expression levels in the context of CNVs and aneuploidies were observed. Specific transcriptional signatures within lineages were nevertheless associated with some specific domesticated processes. By considering the expression level of each gene as a trait, a total of 4,684 eQTL were detected. While the distant eQTL are preponderant (~83.5%), the local regulation explained a higher proportion of the gene expression variation with regulatory variants located in the promoter. To our knowledge, this transcriptional landscape analysis is the most comprehensive overview of the genome-wide expression regulation at the species-wide level, which could overcome the statistical limitations of GWAS, particularly in detecting distant eQTL.

Results

Overview of the strain collection and the generated transcriptomic dataset

To obtain a comprehensive overview of gene expression variation across an entire species, we sought to explore the transcriptional landscape across the 1,011 *Saccharomyces cerevisiae* isolates collection. These isolates were previously collected, sequenced and described to represent the wide genetic, ecological and geographical diversity of the species (Peter et al., 2018). The transcriptomes of 1,010 diverse isolates (Table S1) were obtained through a high-throughput RNA sequencing strategy (Albert et al., 2018). Thereby, mRNA were extracted from cells in exponential growth phase in synthetic complete (SC) medium with glucose as a carbon source and sequenced by Illumina short-read sequencing. Among the 1,010 isolates, we kept a total of 969 of them with at least 1 million of mapped reads for a more in-depth analysis. The mean coverage per isolate is 6.45 million mapped reads (Figure S1). The final dataset constitutes a broad picture of the species with a maximum genetic diversity of 1.4 %, distributed across the 26 well-defined clades (Peter et al., 2018) (Figure 1A). These lineages are related to ecological (Figure 1A) and geographical (Figure 1B) origins, with either domesticated or wild isolates.

Sequencing reads were aligned across a set of 6,680 ORFs, with 4,936 and 1,744 ORFs that are part of the core and accessory genome, respectively. The gene expression level was normalized from read counts to transcripts per million reads (TPM) and transformed in log2(TPM+0.5) for each isolate. A set of 196 genes was filtered out because the level of expression was too low in the population (see Methods). These genes are small with a median size of 347 bp and most of them are dubious or with unknown functions. Among those with established functions, we found genes involved in meiotic cell cycle process. The final set is thus composed of 6,484 ORFs (4,847 core genes and 1,637 accessory genes) and a set of 6,005 ORFs is on average expressed in each strain (Table S2).

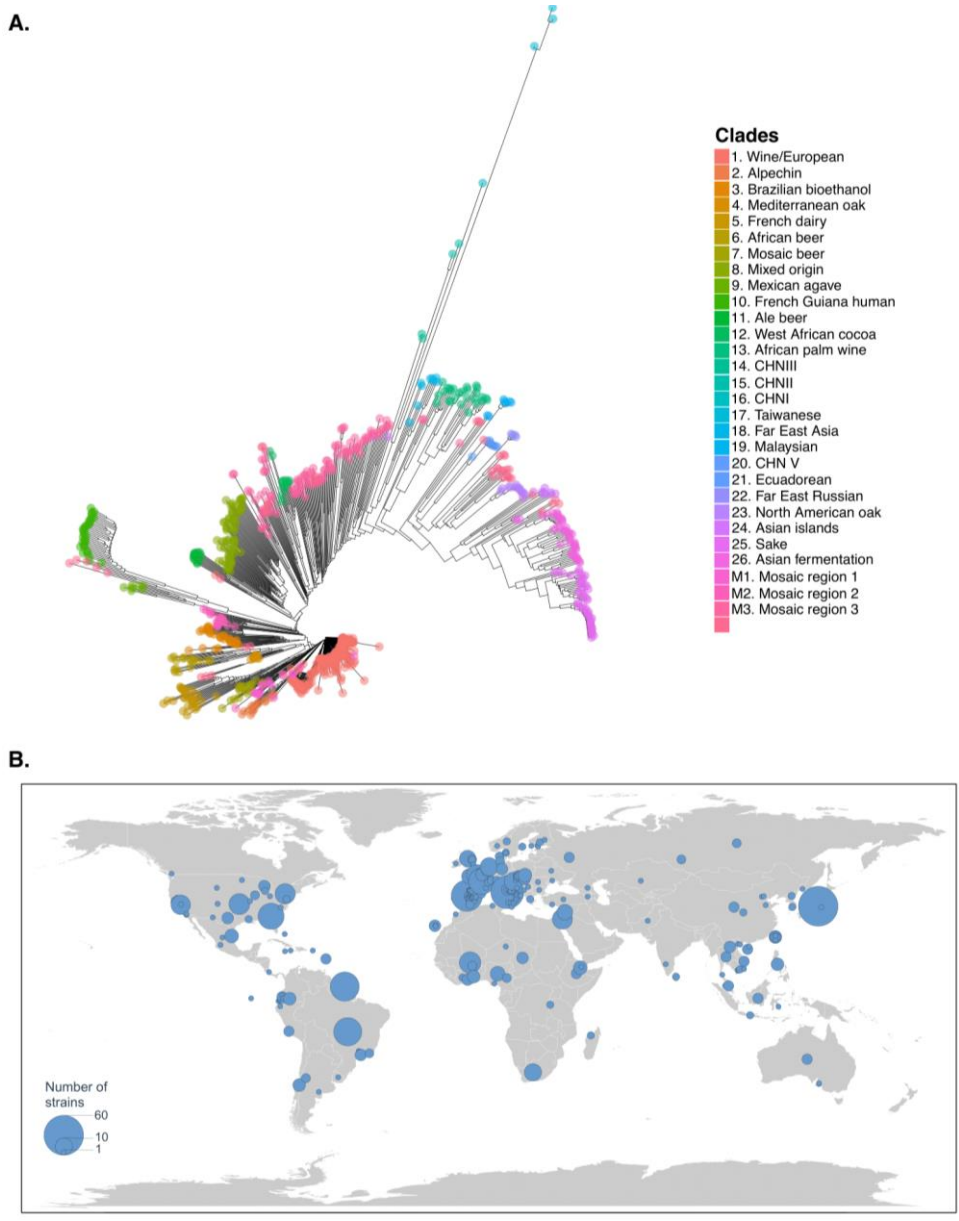
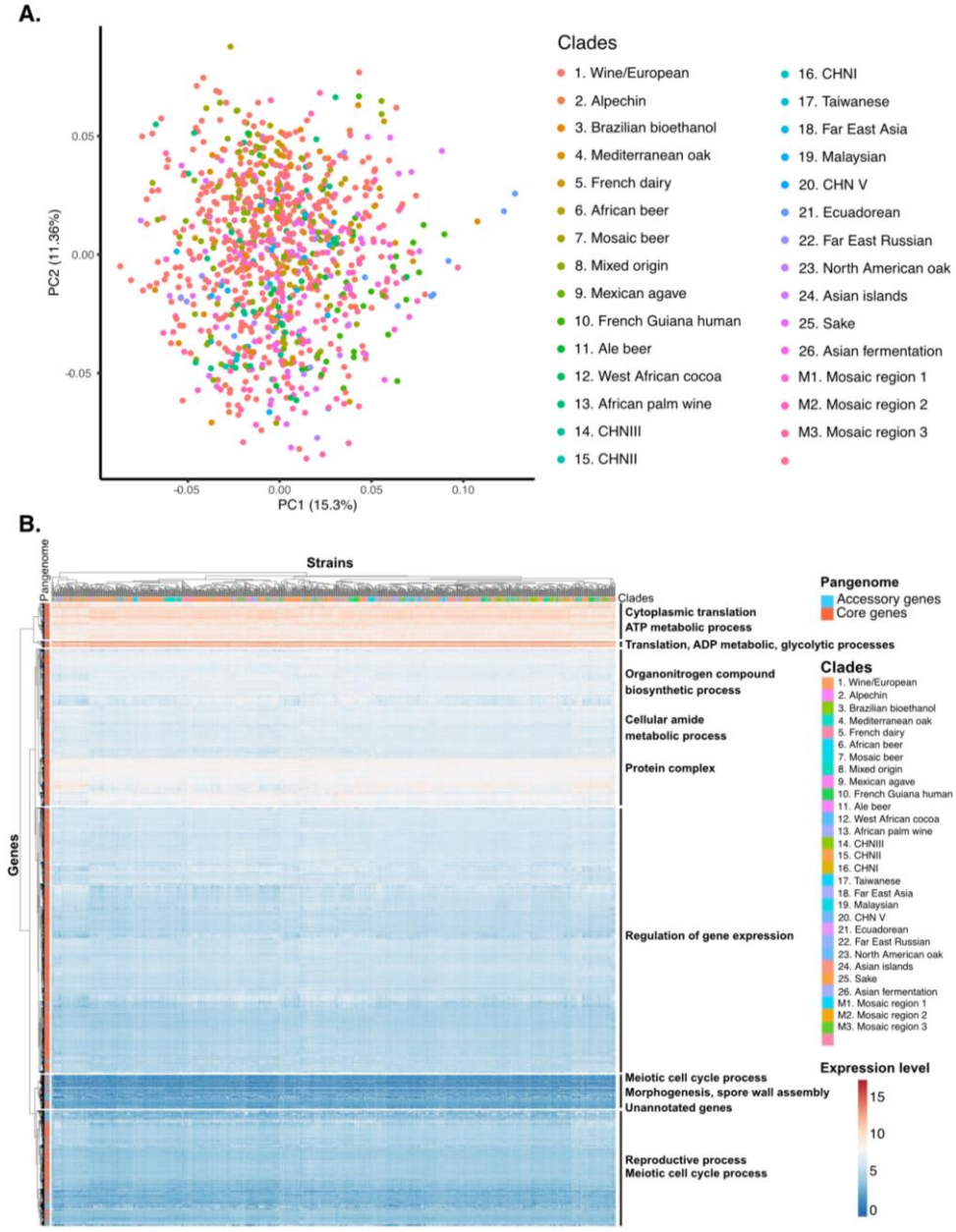


Figure 1. Overview of the 969 natural isolates. (A.) Genetic diversity of the 969 strains based on a neighbour-joining tree. This tree is constructed using biallelic SNPs identified in the RNA-seq reads. Each color represents one of the 26 lineages or one of the mosaic regions described in Peter et al., 2018. (B.) Worldwide geographical origins of the 969 isolates.

Exploration of the transcriptomic landscape and modules of co-expression

As previously mentioned, the *S. cerevisiae* population is well structured according to the genetic diversity. The determination of the transcriptome of a total of 969 isolates provides a strong dataset to explore the structure of the population according to gene expression level. To detect whether certain subpopulations show specific transcriptional patterns, a principal component analysis was first performed on the 6,484 expressed genes (Figure 2A). However, no obvious grouping according to subpopulations was observed but instead an admixture of the isolates based on a complex structure of gene expression variation. This complex structure was confirmed by a hierarchical clustering displaying a similar admixture of the isolates (Figure 2B). In yeasts, previous analysis using smaller datasets already reported this complex structure (Brion et al., 2015; Kita et al., 2017). Obviously, the growth conditions in the laboratory are not representative of the behavior of each isolate which is strongly linked to their living environment, in particular for domesticated strains.

Figure 2. Structure of the transcriptomic landscape across the species. (A.) Principal Component Analysis (PCA) of the expression variation across 969 natural isolates, each dot corresponds to an isolate colored according to the associated lineage. (B.) Hierarchical clustering of the expression variation of 6,089 genes in the population. Core and accessory genes were recognized by two colors as well as the affiliated clades for each isolate. The tree was constructed on the log2(TPM+0.5) expression values with *pheatmap* function and separated in 6 subgroups for which GO term enrichments were performed on SGD (p-value <= 0.01) to detect co-expression networks.


A.
PC2 (11.36%)
PC1 (15.3%)
0.05
0.00
-0.05
-0.05
0.00
0.05
0.10
Clades
1. Wine/European
2. Alpechin
3. Brazilian bioethanol
4. Mediterranean oak
5. French dairy
6. African beer
7. Mosaic beer
8. Mixed origin
9. Mexican agave
10. French Guiana human
11. Ale beer
12. West African cocoa
13. African palm wine
14. CHNIII
15. CHNII
16. CHNI
17. Taiwanese
18. Far East Asia
19. Malaysian
20. CHN V
21. Ecuadorean
22. Far East Russian
23. North American oak
24. Asian islands
25. Sake
26. Asian fermentation
M1. Mosaic region 1
M2. Mosaic region 2
M3. Mosaic region 3
B.
Strains
Pangenome
Genes
Clades
Cytoplasmic translation
ATP metabolic process
Translation, ADP metabolic, glycolytic processes
Organonitrogen compound biosynthetic process
Cellular amide metabolic process
Protein complex
Regulation of gene expression
Meiotic cell cycle process
Morphogenesis, spore wall assembly
Unannotated genes
Reproductive process
Meiotic cell cycle process
Pangenome
Accessory genes
Core genes
Clades
1. Wine/European
2. Alpechin
3. Brazilian bioethanol
4. Mediterranean oak
5. French dairy
6. African beer
7. Mosaic beer
8. Mixed origin
9. Mexican agave
10. French Guiana human
11. Ale beer
12. West African cocoa
13. African palm wine
14. CHNIII
15. CHNII
16. CHNI
17. Taiwanese
18. Far East Asia
19. Malaysian
20. CHN V
21. Ecuadorean
22. Far East Russian
23. North American oak
24. Asian islands
25. Sake
26. Asian fermentation
M1. Mosaic region 1
M2. Mosaic region 2
M3. Mosaic region 3
Expression level
15
10
5
0

The clustering analysis also allowed the detection of regulatory networks resulting in co-expressed genes among the 969 genetic backgrounds. Indeed, modules of co-expression were identified, highlighting specific functions of the upregulated and downregulated genes at the population level (Figure 2B). The gene hierarchical tree was divided into 6 modules to perform GO term enrichment analysis. Within the co-expression module of the upregulated genes (Table S3), an enrichment toward cytoplasmic translation process (p-value = $5{,}59.10^{-12}$) was found as well as for ADP metabolic (p-value = $1{,}78.10^{-11}$) and glycolytic processes (p-value = $2{,}09.10^{-09}$). The enrichments for cytoplasmic translation and, this time, ATP metabolic processes are even stronger in the second most expressed gene module (p-value = $1{,}30.10^{-85}$ and $1{,}64.10^{-14}$ respectively, Table S4). This enrichment for translation process was previously highlighted in this species (Gasch et al., 2000; Wodicka et al., 1997). A larger module of co-expression with an intermediate level of expression included genes involved in biosynthetic and metabolic processes, highlighting the importance of both translation and metabolism in cell growth. Conversely, accessory genes are among the less expressed genes. Indeed, the module containing the less expressed genes is significantly enriched for accessory genes (Fisher test, p-value < $2{,}2.10^{-16}$). This subgroup also displayed an enrichment for unannotated genes (p-value = $5{,}76.10^{-55}$, Table S5) and genes involved in component for morphogenesis and in the meiotic cell cycle process (p-values = $1{,}01.10^{-06}$ and $4{,}85.10^{-05}$ respectively, Table S5). As an example, the master meiosis regulator *IME1* (Tam and van Werven, 2020) is one of the less expressed genes as the induction of this gene and other sporulation genes are a response to nitrogen starvation in the presence of a poor carbon source (Freese et al., 1982; Neiman, 2011).

Interestingly, within the most variable genes (variance > 1 across 969 genetic backgrounds), GO term enrichments for specific biological networks were identified (Figure S2, Table S6) such as carbohydrate transmembrane transport genes (p-value = $4{,}11.10^{-09}$), thiamine-containing compound metabolic process (p-value = $4{,}95.10^{-05}$) or maltose metabolic process (p-value = $7{,}67.10^{-05}$). As described below, these genes reflect functional variations related to the metabolic process specific to certain lineages.

Transcriptional signatures related to domestication processes

To characterize the transcriptional signatures related to specific subpopulations, we plotted the fold change of gene expression in each clade compared to the rest of the population against a p-value associated with the difference of expression. Outlier expressed genes are determined by a threshold set on the corrected p-value (< 0.05 with Bonferroni correction). Among the different clades with significant results, a variable number of transcriptional signatures was detected, ranging from 2 to 1,552 gene signatures in the Asian islands (24.) and French Human Guiana (10.) lineages, respectively (Figure 3A, Table S7). Specific networks frequently associated with domestication processes are highlighted in these lineages. The transcriptional signatures in the clades indeed illustrate the selection pressure undergone by the domesticated strains to focus on pathways essential for industrial processes.

Interestingly, by comparing the wild (n=54) and the domesticated (n=516) strains, we found that genes involved in aerobic respiration are highly upregulated in the strains that are not involved in human-related activities (Figure 3B, Figure S3-1, Table S8). Indeed, a gene ontology analysis of upregulated genes in the wild strains revealed significant enrichment for mitochondrial translation, ATP metabolic and aerobic respiration processes (Table S9). By contrast, this set of genes is downregulated in subpopulations involved in fermentation processes such as the mixed origin clade (8.), for example (Figure S3-1, Table S10). These results are consistent with a previous study which revealed a significant variation in the expression level of these genes generally associated with the origins of the strains (Skelly et al., 2013). However, although the West African cocoa isolates (12.) are domesticated and used for cocoa fermentation, the 13 strains from this clade directly derived from wild populations (Peter et al., 2018) and genes involved in ATP metabolic process are upregulated in those strains (p-value = $1{,}08.10^{-20}$) (Figure S3-1, Table S11). These results are concordant with the divergence in genome evolution between wild and domesticated lineages. While the wild population evolved more rapidly without specific functional constraint, the domesticated subpopulations evolved under strong selection pressure, mostly in the framework of fermentation processes (Peter et al., 2018).

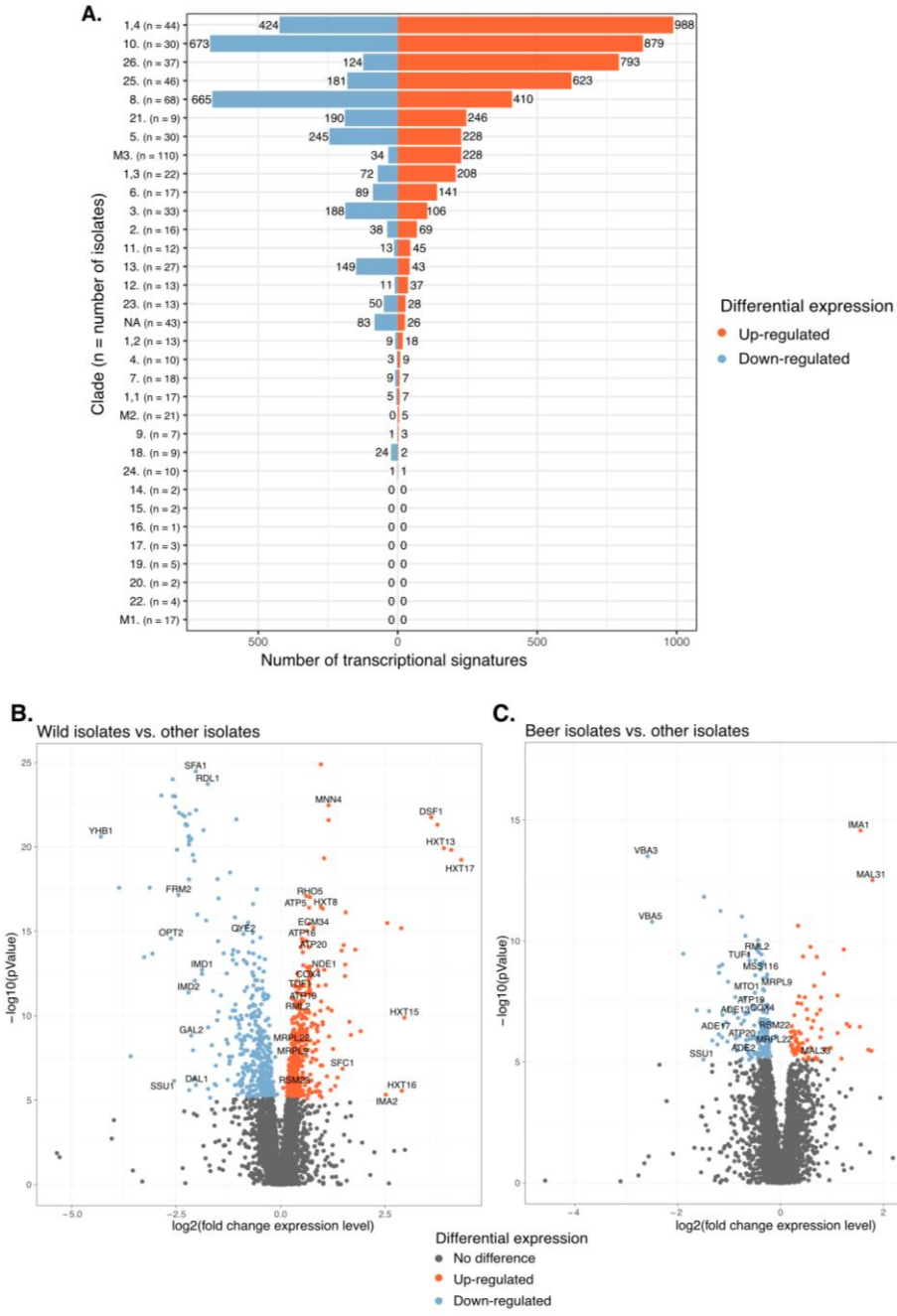

A.
Clade (n = number of isolates)
1,4 (n = 44)
10. (n = 30)
26. (n = 37)
25. (n = 46)
8. (n = 68)
21. (n = 9)
5. (n = 30)
M3. (n = 110)
1,3 (n = 22)
6. (n = 17)
3. (n = 33)
2. (n = 16)
11. (n = 12)
13. (n = 27)
12. (n = 13)
23. (n = 13)
NA (n = 43)
1,2 (n = 13)
4. (n = 10)
7. (n = 18)
1,1 (n = 17)
M2. (n = 21)
9. (n = 7)
18. (n = 9)
24. (n = 10)
14. (n = 2)
15. (n = 2)
16. (n = 1)
17. (n = 3)
19. (n = 5)
20. (n = 2)
22. (n = 4)
M1. (n = 17)
424 988
673 879
124 793
181 623
665 410
190 246
245 228
34 228
72 208
89 141
188 106
38 69
13 45
149 43
11 37
50 28
83 26
9 18
3 9
9 7
5 7
0 5
1 3
24 2
1 1
0 0
0 0
0 0
0 0
0 0
0 0
0 0
0 0
0 0
500
0
500
1000
Number of transcriptional signatures
Differential expression
Up-regulated
Down-regulated
B.
Wild isolates vs. other isolates
−log10(pValue)
log2(fold change expression level)
SFA1
RDL1
YHB1
MNN4
DSF1
HXT13
HXT17
FRM2
RHO5
ATP5
HXT8
EGM34
OPT2
QYE2
ATP16
ATP20
IMD1
NDE1
IMD2
HXT15
GAL2
MRPL22
MRPL9
SFC1
SSU1
DAL1
RSM25
HXT16
IMA2
C.
Beer isolates vs. other isolates
−log10(pValue)
log2(fold change expression level)
MAL32
MAL12
IMA1
VBA3
MAL31
VBA5
RML2
TUF1
MSS116
MRPL9
MTO1
ATP19
ADE13
ADE17
RSM22
ATP20
MRPL22
ADE2
MAL33
SSU1
Differential expression
No difference
Up-regulated
Down-regulated

Figure 3 (part 1). Transcriptional signatures associated with domestication processes and ploidy levels. (A.) Number of genes down- (blue) or up- (red) regulated detected in each clade, or subclade for the wine (1.) lineage. The clades are identified by their corresponding numbers and the strains (n) constituting the group. Transcriptional signatures were defined with the expression level fold change and a p-value cutoff of 0.05 with Bonferroni correction (B.) Volcano-plot comparing the gene expression levels between the 54 wild isolates and the rest of the population. GO term enrichments on SGD (p-value <= 0.01) revealed an upregulation of aerobic respiration process genes (red). Examples of downregulated genes linked to domestication processes and fermentation (blue) are also labeled. (C.) Volcano-plot comparing the gene expression levels in the 48 beer isolates to the rest of the population. The most upregulated genes (red) are involved in maltose metabolic pathway and GO term enrichments on SGD (p-value <= 0.01) revealed a downregulation of aerobic respiration process genes (blue).

In addition to this clear dichotomy between domesticated and wild strains, signatures more specific to certain subpopulations can be highlighted. In the strains isolated from French cheeses (French dairy clade – 5.), we found that the *GAL* transcriptional regulators *GAL3, GAL4* and *GAL80* genes were overexpressed as well as calcium ion transport genes (*VCX1*, *RCH1*, *YVC1* and *PMC1*) for example (Figure S3-1). In addition, transcriptional signatures are also frequently detected for isolates used in alcoholic fermentation such as for the production of wine, beer or sake. In wine isolates, especially the subclade 4, several membrane transporter genes involved in drug resistance, cell detoxification or mannoproteins are among the most upregulated genes (Figure S3-1). The expression of these genes is required to manage the different chemical compounds used during wine production. For example, the *SSU1* gene, which encodes a plasma membrane sulfite pump, is significantly upregulated in the wine subclade 4 (Figure S3-1). This overexpression results from the selection of strains that resist to sulfite excess during winemaking process. Indeed, different chromosomal rearrangements, translocations (Pérez-Ortín et al., 2002; Zimmer et al., 2014) as well as an inversion (García-Ríos and Guillamón, 2019) were identified in wine isolates as responsible of an overexpression of the *SSU1* gene. In the beer subpopulations, transcriptional signatures were also highlighted. Beer strains are polyphyletic and consequently spread over several lineages such as African beer (6.), mosaic beer (7.), mixed origin (8.) and ale beer (11.). By considering the 48 beer isolates or even each clade independently, an enrichment of genes associated in maltose metabolic process was observed for the upregulated genes (p-value = $3{,}32.10^{-05}$) (Figure 3C, Tables S12-S13). In previous population genomic studies, mutations and duplications of the *MAL* genes were indeed identified in beer isolates

(Gallone et al., 2016; Gonçalves et al., 2016) resulting as an overexpression of this specific pathway to adapt to the fermentation environment. In sake isolates (25.), a total of 14 genes associated with thiamine metabolic process were found to be upregulated (Figure S3-1). Interestingly, overexpression of these genes has been shown to be responsible for a high yield of ethanol during sake production (Oba et al., 2011; Shobayashi et al., 2007).

The statistical power of our analysis is nevertheless limited for some subgroups composed of too few individuals. Indeed, transcriptional signatures were not detected in a total of 8 clades containing at most 5 isolates or even 17 isolates in the case of the mosaic region 1 (M1.) (Figure 3A). As this latest clade includes isolates involved in diverse processes or from distinct ecological origins, we could not highlight any enrichment for specific metabolic pathways. Only the Ecuadorean clade (21.) with 9 wild isolates is an exception with 436 signature genes, among which 246 are upregulated and show an enrichment in mitochondrial translation processes (p-value = $2{,}21.10^{-43}$, Figure S3-1, Table S14).

Transcriptional signatures associated with ploidy levels

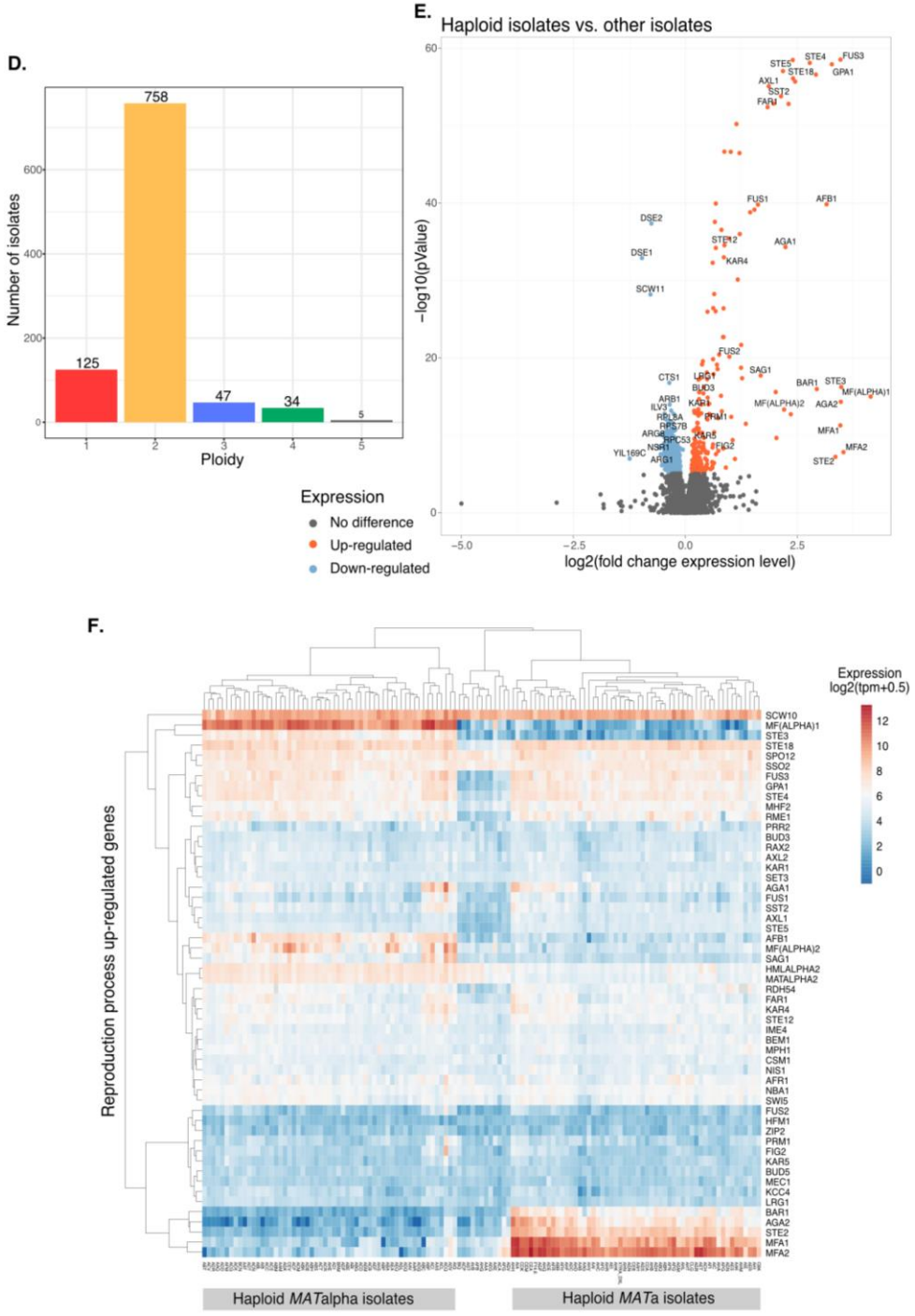
Besides transcriptional signatures related to specific subpopulations, we also explored the impact of the ploidy level on gene expression as previously investigated in the S288C reference isolate (Galitski et al., 1999). The diploid state is most common across the *S. cerevisiae* species (Peter et al., 2018). Among the 969 studied isolates, only 7 isolates are naturally haploid while 118 other haploid isolates were genetically manipulated (Figure 3D, Table S1). Higher ploidies (3n-5n) are more naturally represented within the population with 86 isolates (Figure 3D, Table S1). The set of polyploid strains is highly enriched in specific subpopulations such as the beer, mixed origin and African palm wine clades (Peter et al., 2018).

The transcriptional signatures in polyploids are therefore linked to those observed in these specific clades (Table S15). Indeed, the upregulated genes are strongly associated with maltose metabolism (Figure S3-1), corresponding to the signature observed for the beer isolates. In addition, gene ontology analysis also revealed enrichment similar to that of the mixed origin lineage, such as cellular amino acid biosynthesis process (p-value = $3{,}10.10^{-23}$) (Table S16). Finally, an enrichment for ATP metabolic (p-value = $9{,}25.10^{-09}$) and ergosterol biosynthetic processes (p-value = $1{,}82.10^{-12}$) is also detected as downregulated, which is mainly due to the fact that given polyploid isolates are domesticated (beer, baker and palm wine) (Tables S16-S17).

Unlike polyploid isolates, haploids are distributed within different subpopulations, and thus the transcriptional signatures associated with this set of 125 haploids highlight a specific impact of this state of ploidy (Table S15). Indeed, the set of upregulated genes is mainly associated with the mating pathway (Figure 3E). Within the most variable genes in the population, these genes were already observed as differentially expressed according to the ploidy and haploid specific (Figure S2). Gene ontology analysis of the upregulated genes highlights an enrichment in reproduction (p-value = $9{,}75.10^{-11}$) with different processes such as conjugation (p-value = $1{,}27.10^{-18}$) and response to pheromone (p-value = $3{,}77.10^{-16}$) (Table S18). Furthermore, genes with the higher fold change are the mating pheromone factor a and α genes (*MF(ALPHA)1* and *MFA2*) and the receptors for the a and α factor pheromone genes (*STE3* and *STE2*), which initiate the signalling response that leads to mating (Haber, 2012; Hagen et al., 1986) (Figure 3E). Interestingly, the upregulated genes involved in reproduction process displayed divergent patterns

depending on the mating type of the isolate (Figure 3F). A hierarchical grouping of these genes makes it possible to identify genes co-expressed with the mating pheromone factors a and α (Figure 3F). Finally, a strong enrichment in ribosome biogenesis was observed for the downregulated genes in this set of haploid strains (p-value = $6{,}90.10^{-119}$) (Figure 3E, Table S19).

Figure 3 (part 2). Transcriptional signatures associated with domestication processes and ploidy levels. (D.) Distribution of ploidy levels in the 969 isolates. (E.) Volcano-plot comparing the gene expression levels between the 125 haploid isolates and other ploidies. Genes involved in reproduction and mating processes are upregulated (red) in these isolates as highlighted by GO term enrichments on SGD (p-value <= 0.01) while ribosome biogenesis genes are downregulated (red). (F.) Hierarchical clustering of the 54 genes associated to "reproduction" process GO term in the 125 haploid isolates. Two subgroups of haploid strains can be defined with this clustering according to the expression of specific mating type genes.


D.
Number of isolates
Ploidy
125
758
47
34
5
E.
Haploid isolates vs. other isolates
−log10(pValue)
log2(fold change expression level)
Expression
No difference
Up-regulated
Down-regulated
F.
Reproduction process up-regulated genes
Expression
log2(tpm+0.5)
Haploid MATalpha isolates
Haploid MATa isolates

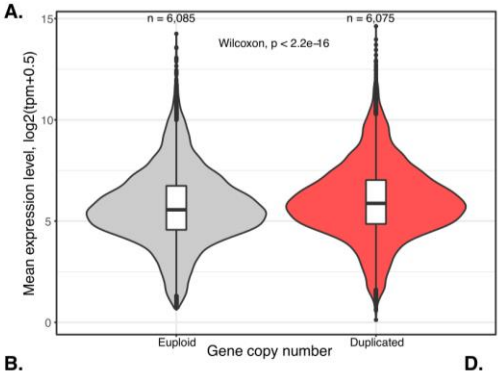
Dosage compensation of expression level at the genome-wide scale

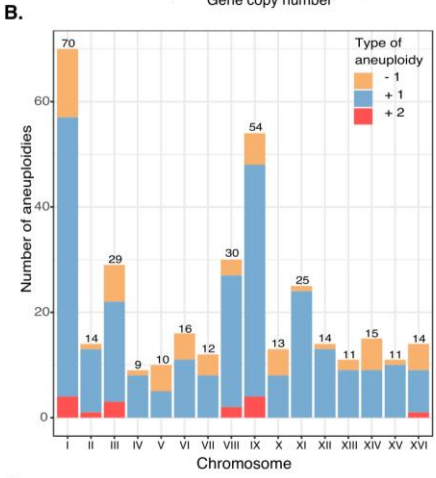
Among the various genetic variants found in the *S. cerevisiae* species, the copy number variants (CNVs) were shown to have a high impact on the phenotypic diversity (Peter et al., 2018). Genes are frequently duplicated across the 1,011 *S. cerevisiae* population. Indeed, almost each gene is involved in a genic, segmental or chromosomal duplication in at least one isolate. It was consequently interesting to explore the transcriptional response to CNVs in the extensive dataset we generated.

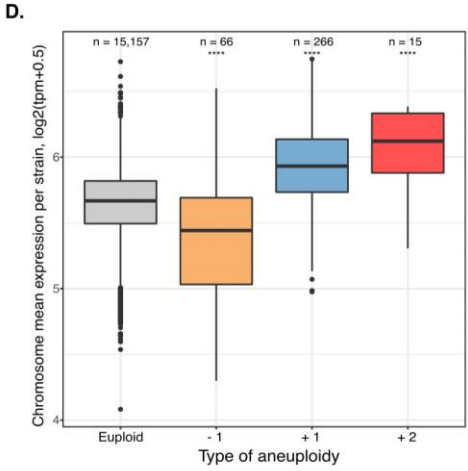
By looking at the entire set of duplicated genes in our sample, we found that the mean expression of this set of genes is slightly higher compared to non-duplicated genes (5.98 *vs.* 5.68), revealing an overall weak effect on the transcription level (Figure 4A). Indeed, the expression fold change is not directly correlated with the number of copies present in the genome. As an example, the presence of two copies versus one copy leads to a gene expression fold change of 1.41 on average, when considering TPM values (Figure S4A). This observation highlights the impact of gene dosage with a general transcriptional regulation of the copy number variation. In addition to this general trend, we also sought to know whether certain genes involved in specific functions were more or less impacted by this transcriptional buffering. Our comprehensive dataset is powerful enough to perform such a systematic comparison. In total, we found a set of 824 genes that are significantly differentially expressed when duplicated (Wilcoxon test, p-value with Bonferroni correction) (Table S20). Interestingly, among this set of genes, we identified a large subtelomeric region located on the chromosome XVI (Figure S4B), going from *AQY1* to YPR204W genes. This duplicated region is present in both wild and domesticated lineages (Bergström et al., 2014; Ogihara et al., 2008) and may have diverse evolutionary origins and functional outcomes. Among them, the arsenite resistance is commonly associated to the *ARR1-3* genes duplication in *S. cerevisiae* as well as in its sister species *Saccharomyces paradoxus* (Bergström et al., 2014; Peter et al., 2018). Interestingly, these genes are constitutively upregulated even in a stress-free environment.

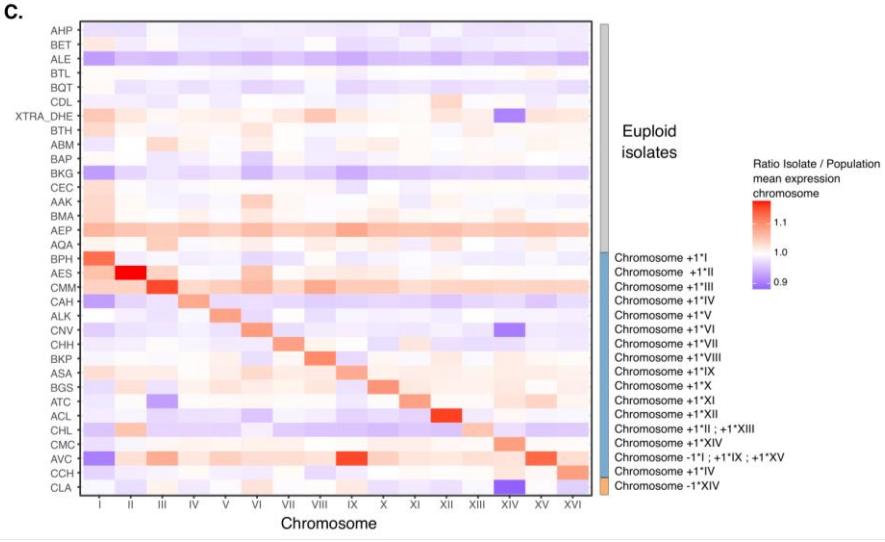
In the last decade, many studies highlighted a global fitness disadvantage in aneuploid isolates compared to euploid (Peter et al., 2018; Tan et al., 2013; Torres et al., 2007; Yona et al., 2012). In order to understand the genetic mechanisms underlying these phenotypic consequences, the impact of CNVs on gene expression or protein levels in aneuploid isolates was explored (Dephoure et al., 2014; Hose et al., 2015). However, these studies have focused either on laboratory aneuploid

strains only or on a very small subset of wild aneuploid isolates. Our large transcriptomic dataset is a major advantage in exploring the impact of aneuploidies on transcription at the species-wide level. Indeed, aneuploidies are widespread across the *S. cerevisiae* genomes (Peter et al., 2018). A low enrichment in aneuploidies for the shorter chromosomes (chromosomes I, III and IX) and a high enrichment in the sake, ale beer and mixed-origin subpopulations were observed. In our dataset, a total of 204 isolates carries at least one sub/supernumerary chromosome (Figure 4B). Similar to the trend observed for all of the duplicated genes, the genes affected by one or more negative or positive aneuploidies were downregulated or upregulated, respectively. These events can easily be detected genome-wide (Figure 4C). A correlation between the chromosome number variation and the gene expression level can be observed at the genome and chromosome scales (Figure 4D, Figure S4C). While the median expression level of the euploid chromosomes is 5.64, the median expression levels of aneuploidy -1, +1 and +2 are 5.35, 5.92 and 6.04, respectively (Figure 4D). These results clearly highlight an effect of the aneuploidies on the transcriptional level but again a transcriptional regulation of the CNVs can be observed.


A.
n = 6,085
n = 6,075
Wilcoxon, p < 2.2e-16
Mean expression level, log2(tpm+0.5)
Euploid
Duplicated
Gene copy number


B.
Type of aneuploidy
- 1
+ 1
+ 2
Number of aneuploidies
Chromosome


D.
n = 15,157
n = 66
n = 266
n = 15
Chromosome mean expression per strain, log2(tpm+0.5)
Euploid
- 1
+ 1
+ 2
Type of aneuploidy


C.
Euploid isolates
Ratio Isolate / Population mean expression chromosome
Chromosome +1*I
Chromosome +1*II
Chromosome +1*III
Chromosome +1*IV
Chromosome +1*V
Chromosome +1*VI
Chromosome +1*VII
Chromosome +1*VIII
Chromosome +1*IX
Chromosome +1*X
Chromosome +1*XI
Chromosome +1*XII
Chromosome +1*II ; +1*XIII
Chromosome +1*XIV
Chromosome -1*I ; +1*IX ; +1*XV
Chromosome +1*IV
Chromosome -1*XIV
Chromosome

Overall, our dataset corroborated at the species level the general dosage compensation of the gene transcription, already described in some aneuploid strains (Hose et al., 2015). Furthermore, different expression patterns were detected between euploid and aneuploid isolates, revealing the mechanisms underlying tolerance to aneuploidy in natural strains. Indeed, the mitochondrial genes involved in ATP metabolic processes (p-value = $8{,}09.10^{-20}$, Figure S4D, Table S21) are downregulated in aneuploid isolates. This trend was previously observed across 15 aneuploid strains (Hose et al., 2015) and interestingly also in the context of Down syndrome (Helguera et al., 2013; Liu et al., 2017). As previously mentioned, these downregulated transcriptional signatures are observed in domesticated subpopulations such as the beer clades, which are prone to aneuploidies. Nevertheless, the same transcriptional signature is also seen in aneuploid isolates not related to these specific subpopulations (Figure S4E, Table S22).

Figure 4. Impact of the gene copy number on gene expression. (A.) Mean expression levels for each gene in euploid (6,085 genes with n copies = 1) or duplicated (6,075 genes with n copies > 1) context. The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test. (B.) Distribution of aneuploidies and their types along the chromosomes of 204 isolates. (C.) Heat-map representing the ratio of the mean expression level on each chromosome between a selected isolate and the mean expression of the chromosome in the population. 16 euploid isolates randomly selected display homogeneous expression along the 16 chromosomes while the expression ratio allows the detection of positive (towards the red) or negative (towards the blue) aneuploidies in the 17 selected aneuploid isolates. New aneuploidies seem to be detectable by analyzing the expression levels such as a negative aneuploidy on chromosome XIV for the strains XTRA_DHE and CNV. (D.) Box plots of the mean expression level for each isolate on each chromosome according to the aneuploidy type. The p-values are calculated using a two-sided Mann–Whitney–Wilcoxon test between euploid subgroup and each aneuploidy type, **** p-value < $1e^{-04}$.

Pangenome and gene expression variation

The pangenome of *S. cerevisiae* was recently established from *de novo* genome assemblies and revealed many evolutionary events such as introgressions or horizontal gene transfers (HGT) (Peter et al., 2018). For example, Alpechin, Mexican agave and French Guiana clades exhibit many introgressions from diverse subpopulations of *S. paradoxus*. To explore the transcriptional landscape of the pangenome, RNA sequencing reads were aligned to 6,285 ORFs of the S288C reference genome as well as to 395 supplementary ORFs present in other isolates (Table S2). Among this set, a total of 4,936 ORFs are present in all the strains and constitute the core genome whereas 1,744 ORFs are part of the accessory genome. Variable ORFs are classified according to their origins, *i.e.* ancestral, introgression, horizontal gene transfer (HGT) and candidate HGT (Figure 5A). Accessory genes are found in 1 to 969 isolates and their frequency varies according to the ORF origins (Figure S5A). For instance, ancestral ORFs are mainly present in almost all the isolates while HGT and introgressions are only present in a small subset of isolates.

With our comprehensive transcriptomic dataset, we surveyed the gene expression levels across the 969 genetic backgrounds with variable genic content. Overall, accessory genes are less expressed than the core genome in the population, with a mean expression of 5.22 vs. 5.72 (Figure 5B). The gene expression variance is also higher for accessory ORFs, 6.11 vs. 4.23. These results are consistent with the discrepancy already described between core and accessory genes (Peter et al., 2018). Indeed, genes part of the core genome undergo more genetic constraints with a lower ratio of non-synonymous to synonymous polymorphisms compared to the accessory genes (Figure 5C). Essential genes which mostly belong to the core genome are also more expressed than non-essential genes and their expression is less variable (Figures S5B and S5C). Gene expression variation thus reflects distinct evolutionary trajectories of the ORFs of the pangenome. Moreover, the expression level is lower regardless of the origin of the accessory genes (Figure 5D). ORFs coming from horizontal gene transfer (HGT and candidate HGT) represent the least expressed genes. We have nevertheless highlighted certain highly expressed ORFs involved in the fermentation process found in *Zygosaccharomyces* species, wine contaminants (Marsit et al., 2015; Novo et al., 2009).

The majority of the introgressed ORFs comes from *S. paradoxus*, a *S. cerevisiae* sister species. These genes often have an ortholog, which retains the function of *S. cerevisiae* allele and replaces it or coexists in the heterozygous state. To study the

gene expression variation and adaptation of introgressed alleles, we focused on the homozygous cases. When comparing a total of 439 genes homozygous for the allelic version of *S. cerevisiae* or *S. paradoxus*, no significant difference in gene expression was observed overall (Figure 5E). Nevertheless, we found that 76 out of the 439 genes are differentially expressed when introgressed, with 29 and 47 of them being upregulated and downregulated, respectively (Figures S5D and S5E, Table S23). No specific function could have been highlighted but in many cases, several genes from *S. paradoxus* are introgressed together with a conserved synteny and the entire region is differentially expressed in a similar manner in the lineage containing the introgressions.

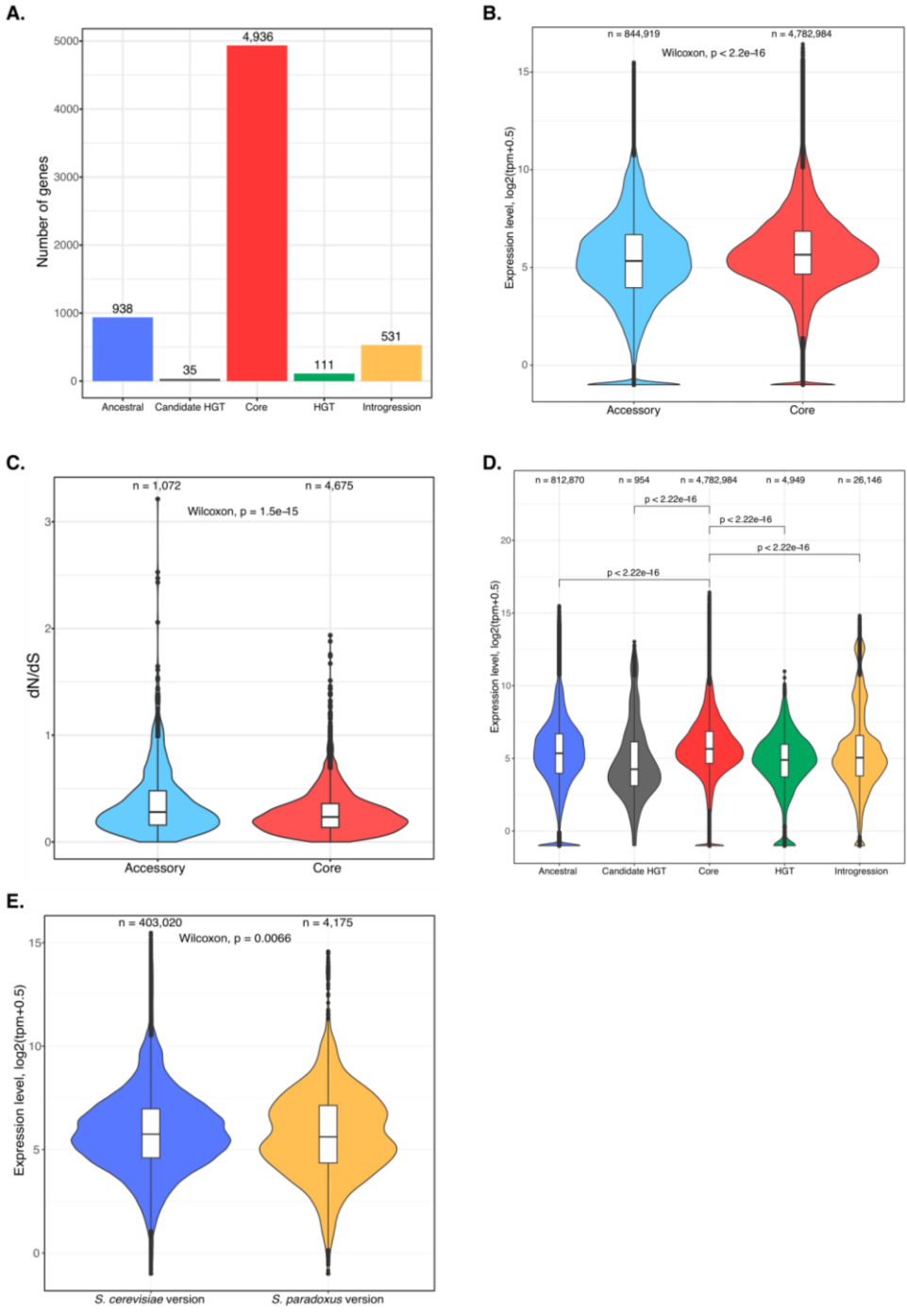

A.
Number of genes
5000
4000
3000
2000
1000
0
4,936
938
35
111
531
Ancestral
Candidate HGT
Core
HGT
Introgression
B.
n = 844,919
n = 4,782,984
Wilcoxon, p < 2.2e-16
Expression level, log2(tpm+0.5)
15
10
5
0
Accessory
Core
C.
n = 1,072
n = 4,675
Wilcoxon, p = 1.5e-15
dN/dS
3
2
1
0
Accessory
Core
D.
n = 812,870
n = 954
n = 4,782,984
n = 4,949
n = 26,146
p < 2.22e-16
p < 2.22e-16
p < 2.22e-16
p < 2.22e-16
Expression level, log2(tpm+0.5)
20
15
10
5
0
Ancestral
Candidate HGT
Core
HGT
Introgression
E.
n = 403,020
n = 4,175
Wilcoxon, p = 0.0066
Expression level, log2(tpm+0.5)
15
10
5
0
S. cerevisiae version
S. paradoxus version

Figure 5. Transcriptional landscape of *Saccharomyces cerevisiae* pangenome. (A.) Number of the different types of genes constituting the species pangenome included in our analysis. (B.) Gene expression levels in accessory and core genes in 969 isolates. The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test. (C.) A higher ratio of non-synonymous to synonymous polymorphisms, dN/dS, in accessory genes uncovers a lower selection pressure in these genes. The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test. (D.) Gene expression levels in different types of accessory genes and the core genome in 969 isolates. The p-values are calculated using a two-sided Mann–Whitney–Wilcoxon test between core genome expression and each accessory gene type. (E.) Gene expression level comparison between 439 genes homozygous either for *S. cerevisiae* or *S. paradoxus* allele. The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test.

Local eQTL as a major source of gene expression variation

Gene expression is a fundamental and intermediate step of a process that will lead to the phenotypic diversity of a population. Previous surveys explored the genetic basis of gene expression variation and regulation. These studies have led to the identification of expression quantitative trait loci (eQTL) via either linkage or association strategies (Hill et al., 2021). Linkage analysis highlighted a complex inheritance of gene expression levels and a pleiotropic effect of the eQTL (Albert et al., 2018; Brem et al., 2002; Schadt et al., 2003). However, even with an extensive mapping population, eQTL are still limited to more or less large genomic regions according to the number of recombination break points (Albert et al., 2018). On the other hand, genome-wide association studies (GWAS) had a higher resolution to specifically spot the causal nucleotides (The GTEx Consortium, 2015, 2017, 2020). The wide *S. cerevisiae* collection is a powerful resource to investigate the regulatory variations behind gene expression. The complete transcriptional landscape of 969 different individuals will increase the statistical power to find, among other, distant eQTL which explain a lower phenotypic variance and were not detected in a small set of 85 strains (Kita et al., 2017).

1- eQTL regulate genes far from their locations

We considered 5,868 traits (corresponding to the expression level of each gene) varying across the 969 isolates to perform GWAS. Between these isolates, 75,828 single-nucleotide polymorphic sites were found with a minor allele frequency (MAF) higher to 5%. These SNPs were used to perform a mixed-model association and map the eQTL that influence gene expression levels. Furthermore, CAVIAR (Hormozdiari et al., 2014) was used to filter the local eQTL and identify the causal genetic variant for each trait. In total, 4,684 expression quantitative trait loci (eQTL) were associated with the gene expression level of 2,023 different genes (Figure 6A, Table S24) and explained a variable part of the phenotypic variance, ranging from 13.6% to 46.8% (Figure S6A). These eQTL regulate the transcriptional level near or far from the target gene. Local or distant eQTL were defined according to the distance between the SNP and the influenced gene (see Methods). A total of 775 local or *cis*-eQTL were identified in a range of 25 kb on either side of the gene whereas 3,909 distant or *trans*-eQTL acted from further away. Interestingly, 83.5% of the genetic variants involved in gene expression variation were distant eQTL among which 88% (n=3,453) are located on a different chromosome. Only 199

unique genes are both locally and remotely regulated whereas 1,640 and 184 phenotypes are only regulated by distant and local eQTL respectively. However, despite the predominance of distant eQTL regulating gene expression, the phenotypic variance is more explained by local eQTL, with a mean of 18% against 15.4% for *trans*-eQTL (Figure 6B). Because of their pleiotropic effects, *trans*-regulatory variants would be more deleterious, which explains a lower contribution to the phenotypic variance (Schaefke et al., 2013). By contrast, *cis*-eQTL evolved more rapidly and are selected for more beneficial and straight effects (Coolon et al., 2014; Metzger et al., 2017). The detection of distant eQTL is consequently more challenging because of the reduced phenotypic effects. Even if a large number of *trans*-regulatory variants was detected using our transcriptomic dataset, a large part of the phenotypic variance is still not explained. Among the 2,023 genes regulated by eQTL, the expression level of more than half of them (n=1,168) was only controlled by a single genetic variant (Figure S6B).

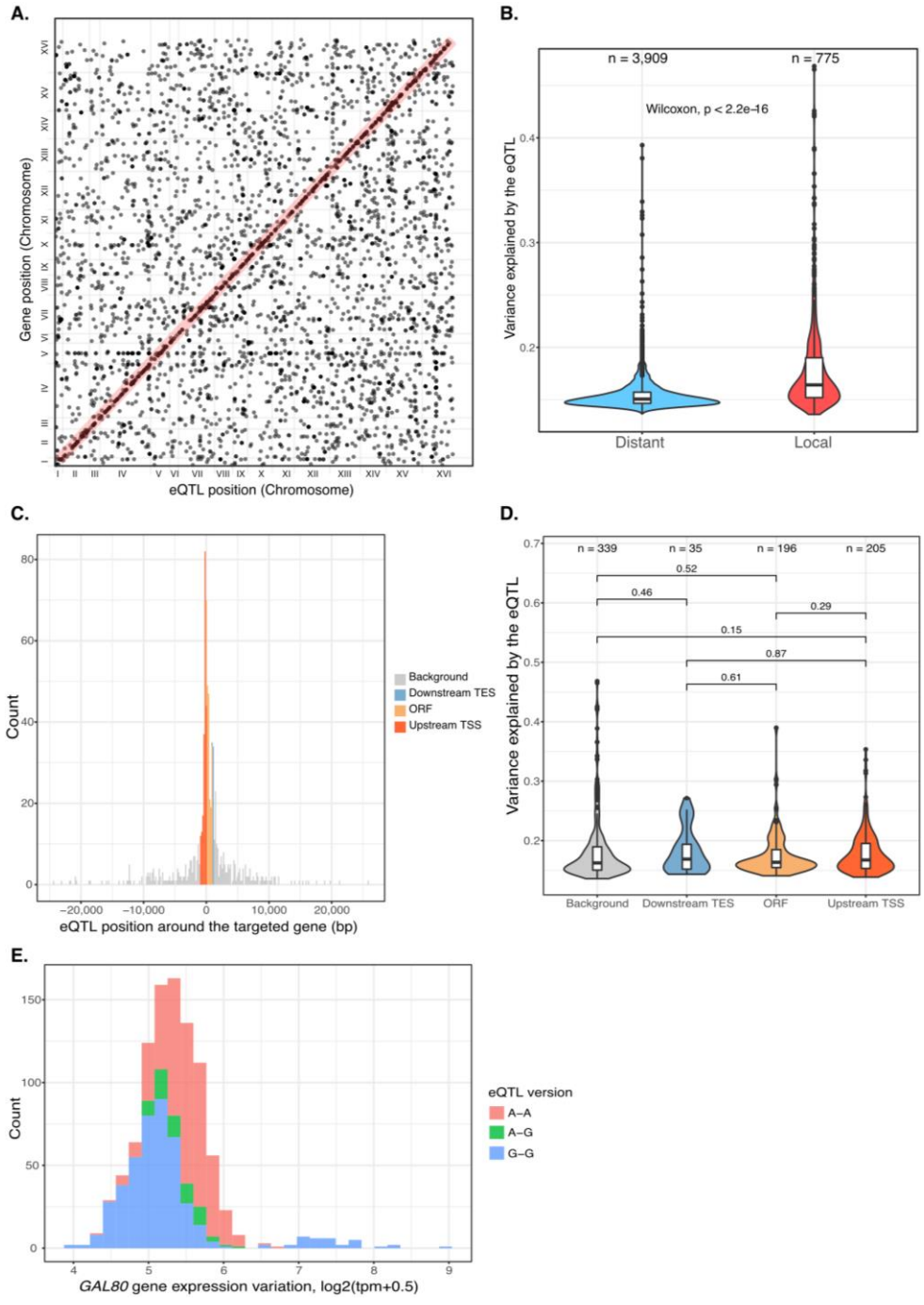

A.
Gene position (Chromosome)
eQTL position (Chromosome)
B.
n = 3,909
n = 775
Wilcoxon, p < 2.2e-16
Variance explained by the eQTL
Distant
Local
C.
Count
Background
Downstream TES
ORF
Upstream TSS
eQTL position around the targeted gene (bp)
D.
n = 339
n = 35
n = 196
n = 205
0.52
0.46
0.29
0.15
0.87
0.61
Variance explained by the eQTL
Background
Downstream TES
ORF
Upstream TSS
E.
Count
eQTL version
A–A
A–G
G–G
GAL80 gene expression variation, log2(tpm+0.5)

Figure 6. Genome-wide association studies highlight distant and local regulatory variants of gene expression. (A.) Location of the detected eQTL across the genome. Map of the genomic position of the 4,684 eQTL (x-axis) against the position of the regulated phenotype (y-axis). The diagonal (red) represents the local eQTL. Grey lines correspond to the chromosome limits. (B.) Distribution of the phenotypic variance explained by the 3,909 distant eQTL or the 775 local eQTL. The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test. (C.) Distribution of the local eQTL around the targeted gene. The upstream region (red) corresponds to the 1,000 bp before the transcription start site (TSS), the downstream region (blue) corresponds to the 300 bp after the stop. The position of eQTL located within the ORF (yellow) is reported in per 1,000. Hence, 1,000bp are added to the position of downstream eQTL. The background regions contain the remaining regions located 25 kb before and after the gene. (D.) Explained phenotypic variance of local eQTL according to its location from the targeted gene. No significant differences are found using a two-sided Mann–Whitney–Wilcoxon test. (E.) *GAL80* gene expression distribution according to the local eQTL allelic version that regulates its expression level. All the isolates which overexpressed *GAL80* carried a guanine and this group of strains is enriched in French Dairy (5.).

2 – Accessory genes are more locally regulated

While the gene expression variance is similar (Figure S6C), we found that the mean expression of genes with associated eQTL was lower than that of the other genes, 5.54 vs. 5.84 (Figure S6D). We therefore looked to what extent the genetic selection pressure affected the regulation of gene expression. First, we found that expression of essential genes (Dowell et al., 2010; Giaever et al., 2002) is less influenced by eQTL (Fisher test, p-value = 0.0053). Previous results already hypothesized a strong negative selection of the genetic mutations leading to gene misregulation that could have a deleterious effect on individual viability (Kita et al., 2017; Ronald and Akey, 2007). In line with a negative selection of eQTL, the impacted phenotypes have less genetic constraints, illustrated by a higher ratio of nonsynonymous to synonymous polymorphisms (dN/dS) in these genes (Figure S6E). This negative selection *de facto* induced a thin regulation of the core genome compared to accessory genes much more targeted by eQTL (410 accessory /2023; Fisher test, odds ratio = 1.26, p-value = 0.001). This enrichment in accessory genes is even higher in the 383 locally regulated phenotypes (111/383; Fisher test, odds ratio = 1.96, p-value = 4,216e^{-08}). As *cis*-eQTL have a higher phenotypic impact and evolve rapidly next to the targeted gene (Coolon et al., 2014; Metzger et al., 2017), this regulatory mechanism is privileged for accessory genes which are lately selected or acquired within the species. In contrast, *trans*-eQTL have a pleiotropic regulation which targets the core genome more. Isolates with similar origins own close genomes with conserved variable genes that can be linked to the environmental niche or domestication processes. In order to conserve the fitness level, local expression regulation of accessory genes is then the more rapid and efficient way.

3 – Regulatory variations widespread along the genome

Among the 75,828 single-nucleotide polymorphic sites found across the 969 isolates, only 5% (3,738 unique eQTL) are involved in gene expression regulation. These eQTL are widespread along the genome without any strong hotspot regions, *i.e.* regulating the expression level of many genes (Figures 6A and S6F). In total, 3,121 (67%) of the eQTL are located inside a CDS (coding DNA sequence), which corresponds to the proportion of genomic coding content and 1,598 genes hold at least one eQTL (Table S24). The genetic sequence of these regulatory genes also underwent slightly less genetic constraints, characterized by a higher dN/dS (Figure S6G). All of them are not annotated as involved in gene expression regulation but

almost half of the regulatory genes are implicated in regulation process and in binding activities (p-values = $4{,}7.10^{-04}$ and $1{,}32.10^{-01}$ respectively, GOrilla database, Table S25). We also highlighted gene network enrichment in regulation of chromosome organization and transcription by RNA polymerase II processes (p-values = $1{,}17.10^{-01}$ and $1{,}58.10^{-01}$ respectively, GOrilla database, Table S25). The phenotypic variance is slightly more explained by intergenic eQTL (Figure S6H). Indeed, intergenic regions contained regulatory elements such as promoters which are directly related to the transcription initiation and so to the gene expression level (Hahn and Young, 2011; Tirosh et al., 2009).

4 – Promoter region as a regulatory hotspot

As mentioned, gene expression variation is highly impacted by local and intergenic eQTL (Figures 6B, Figure S6H). We thus focused our analysis on the 775 *cis*-regulatory variants and more precisely on the distribution of these eQTL around the targeted gene. A higher number of eQTL is detected within the 500 bp before the transcription start site (TSS) of the regulated gene (Figure 6C). Besides, a smaller increase of eQTL number is spotted in the first 50% of the gene as well as in the 3'-untranslated transcribed region (3'-UTR) (Figure 6C). Local regulation hence occurred in the direct neighbour regulatory regions as previously described (Kita et al., 2017; Ronald et al., 2005; Skelly et al., 2013) and the promoter sequence variation effect on gene expression was deeply studied in yeast (Duveau et al., 2017; Lubliner et al., 2015). These genetic variations could modify the nucleosome conformation or the TATA element for instance. Recently, the promoter outcomes on transcriptional level could even be predicted to a certain extent (de Boer et al., 2020). However, despite the known broad effect of promoters on gene expression, the explained phenotypic variance was not higher for any of the regions (Figure 6D). This suggested a pleiotropy of local mechanisms to affect the gene expression level across the species.

5 – Mutation in promoter induces *GAL80* overexpression in dairy strains

Using the transcriptomic dataset of the 969 diverse *S. cerevisiae* isolates, we could identify specific transcriptional signatures related to the strain environment as well as regulatory variants involved in these processes by genome-wide association studies. For instance, transcriptional signatures linked to the dairy production were identified in the French dairy lineage constituted of 30 isolates. Genes involved in

the calcium ion transport and galactose pathways are overexpressed (Figure S3-1) in this subpopulation. The three overexpressed *GAL* genes are all involved in the transcriptional regulation of the pathway (Horák, 2013; Johnston, 1987). GAL3p is a repressor of inhibition, which binds the GAL80p inhibitor in presence of galactose. This repression allows the activation of the transcription factor GAL4p that triggers the galactose enzymatic reaction, usually inhibited by GAL80p. Our experiments were performed in a glucose environment which explained why the other *GAL* genes are not also overexpressed. The differences in expression levels highlighted a precise regulation of the regulatory elements of the *GAL* pathway in the dairy strains. Interestingly, at least one eQTL is detected for each of these genes (Table S24), the expression of the *GAL3* and *GAL4* genes are each regulated by 1 *trans*-regulatory variant, located in *UBP7* and *RIA1*, respectively. By contrast, *GAL80* expression is regulated by two distant eQTL located in the *PRM5* and YMR315W genes, and by one *cis*-eQTL located 71 bp before the TSS, i.e. in the promoter region. This local eQTL explained 25.4% of the phenotypic variance. The polymorphic site carried either a guanine (46.5% of the isolates), an adenine (46.5%) or both nucleotides in heterozygous cases (7%). Interestingly, all the isolates for which *GAL80* gene is upregulated carry a guanine (Figure 6E) suggesting a causal effect of this allele in the expression variation. Moreover, the 30 French dairy isolates bear this SNP. However, the phenotypic variance of *GAL80* expression is not completely explained by these detected *cis*- and *trans*-eQTL. The genetic basis of *GAL80* expression variation is difficult to dissect because barely 5% of the isolates overexpressed the gene. Hence, genetic variants involved in this overexpression could have a minor allele frequency lower than 5% and hence could not be detected via GWAS. This example still emphasized the genetic complexity underlying the architecture of traits, selected for industrial process for instance.

Allele Specific Expression (ASE) and *cis*-regulatory changes

Diverse ploidies were described within the *S. cerevisiae* species, with haploid to pentaploid isolates resulting in high level of heterozygosity across the genomes. While most of the natural isolates are diploid, 52.7% of them are heterozygous (Peter et al., 2018). The impact of heterozygosity on gene expression variation has been explored in hybrids and reveals a major role of *cis*-regulatory changes (Metzger et al., 2017; Wittkopp et al., 2004). A strong correlation between *cis*-acting eQTL and allele specific expression has been found in various human tissues (The GTEx Consortium, 2015, 2020). The extensive dataset we generated thus gives the opportunity to survey expression variation in 289 euploid diploid heterozygous isolates through allele specific expression (ASE) analysis.

Among the 214,551 heterozygous sites distributed in 3,750 unique genes across the 289 isolates, the expression of around 3% of the sites (n = 6,513) was significantly imbalanced in 1,100 unique genes (binomial test, p-value corrected with Benjamini and Hochberg method). The genetic regulation of these 1,100 differentially expressed alleles has been linked to the 775 *cis*-eQTL detected by GWAS. Interestingly, the expression of 0.60% of the genes influenced by a heterozygous *cis*-eQTL was significantly imbalanced while only 0.066% of the genes regulated by a homozygous *cis*-eQTL displayed an allelic imbalance. The differential expression of heterozygous alleles was thus significantly more regulated by heterozygous *cis*-eQTL (binomial test, odds ratio = 9.18). These results are consistent with previous analysis in humans, uncovering local bi-allelic regulation (The GTEx Consortium, 2015, 2020).

Over the 1,100 unique genes with heterozygous imbalanced sites, 625 unique genes carried all of their heterozygous sites imbalanced (Figure S7A) from 1 to 18 sites per gene (Figure S7B). A total of 186 out of the 625 genes displayed imbalanced expression in several isolates and we therefore looked to see if some of these unbalanced genes were specific to given subpopulations (Figure S7C). We found that 16 genes are significantly enriched in 8 clades (Fisher test, Table S26) with the highest enrichment for the *SSU1* gene in the wine subpopulation (1.). Interestingly, the *SSU1* gene was also detected as an upregulated transcriptional signature, with 2.3-fold change in gene expression in this given subpopulation. As previously mentioned, several chromosomal rearrangements influencing the expression of the *SSU1* gene have been described in the wine strains for sulfite tolerance (García-Ríos

and Guillamón, 2019; Pérez-Ortín et al., 2002). We then hypothesized that the 23 isolates with the *SSU1* allelic imbalance carried both one translocated allele and one wild type allele. This heterozygosity was indeed confirmed by polymerase chain reaction (PCR) with optimized primers (Table S27) that allowed to discriminate the chromosomal rearrangement from the wild type allele (Marullo et al., 2020; Pérez-Ortín et al., 2002). Different types of translocations were identified across the 23 isolates with imbalanced gene expression (Table S28). These chromosomal rearrangements induced a modification to the promoter of the *SSU1* gene and thus lead to an upregulation. Indeed, compared to other natural isolates, the 23 isolates heterozygous for the translocations significantly overexpressed *SSU1* gene (Figure S7D). The strong *SSU1* upregulation detected in the wine lineage nevertheless suggested the presence of these causal rearrangements in other wine genetic backgrounds, probably at the homozygous state. Overall, this example supported the impact of heterozygosity on the gene expression level as well as the genetic selection for environmental adaptation.

Finally, introgression events are frequent in certain *S. cerevisiae* isolates in homozygous and heterozygous states, as mentioned previously (Peter et al., 2018). Previous analysis revealed similar expression of the homozygous version of the *S. cerevisiae* and *S. paradoxus* alleles (Figure 5E). The study of differential expression of heterozygous alleles is possible via the ASE analysis. Different filters were first applied to reduce the mapping bias toward the reference alleles while the alternative alleles correspond to the introgressed genetic variants (see Methods). After these steps, no significant difference was found between non-introgressed and *S. paradoxus* introgressed heterozygous sites, suggesting a similar regulation of heterozygous alleles (Figure S7E). In a heterozygous context, *S. paradoxus* introgression events thereby didn't result in higher differential expression levels compared to other non-introgressed heterozygous alleles. Whatever the genomic context, homozygous and heterozygous, introgression events are thus not significantly differently expressed than the *S. cerevisiae* allele in the specific experimental conditions used. These events still have a central role in genome evolution of the species and are particularly conserved in certain subpopulations (D'Angiolo et al., 2020; Peter et al., 2018).

Discussion

The wide overview of the transcriptional landscape that we conducted across the *S. cerevisiae* species led to a deeper characterization of the regulatory variants impacting gene expression variation. The first result of our global survey revealed the absence of transcriptional patterns specific to certain subpopulations. However, specific transcriptional signatures associated to given lineages were detected and mainly highlighted the functional impact of the domestication processes. Indeed, these signatures were selected to facilitate human activities by the over- or underexpression of specific genes. As an example, the gene *SSU1*, responsible for resistance to sulfite excess during winemaking, was overexpressed in the wine lineage. This overexpression was associated to characterized structural variants, translocations as well as inversion (Marullo et al., 2020; Pérez-Ortín et al., 2002), selected in these isolates at the homozygous or heterozygous state as verified by an ASE analysis. The genetic regulation of gene expression was thus under positive selection especially for genes involved in metabolic pathways required for domestication.

Mechanisms of genome evolution implied both SNPs and SVs including CNVs commonly described in adaptation responses to environment as well as domesticated processes (Bergström et al., 2014; Gonçalves et al., 2016). As an example, genes involved in the metabolism of maltose, a carbon source specific to beer fermentation, are duplicated and overexpressed in the corresponding isolates (Gallone et al., 2016). Interestingly, a dosage compensation of expression was observed for duplicated genes, either at the genic or chromosomal level. Previous analysis already described this phenomenon in aneuploid laboratory isolates (Dephoure et al., 2014; Hose et al., 2015), and our extensive dataset provides for the first time a species-wide overview of the expression buffering effect of duplication. In addition, different steps occur in this process of gene expression regulation, from the initial mRNA to the final protein products. Strong dosage compensation of the protein abundance levels was characterized in aneuploid as well as single gene duplication context especially in essential and multiprotein complex genes (Ascencio et al., 2021; Chen et al., 2019; Dephoure et al., 2014). It would then be interesting to obtain the species-wide overview of these different steps such as the protein translation through the quantification of ribosome-protected mRNA fragments using ribosome profiling (Ingolia et al., 2009) as well as the protein abundance measures through mass spectrometry (Yates et al., 2009). The comparison of the different levels would allow

to understand the different mechanisms of post-transcriptional as well as post-translational buffering effects.

Establishing the complete transcriptome of hundreds of natural isolates has been a major advantage in dissecting the genotype-phenotype relationship. We were thus able to perform genome-wide association studies to detect eQTL. A large number of distant eQTL were detected but it was shown that they explained a lower part of the phenotypic variance compared to local eQTL. Interestingly, the expression of accessory genes was more impacted by local regulatory variants revealing a differential gene expression regulation between the core and the accessory genome. Nevertheless, a large proportion of the phenotypic variance remains to be explored within the species. Among the potential sources of missing heritability (Manolio et al., 2009), structural variants and CNVs in particular could be further investigated to have a global view of the impact of these genetic variants on gene expression variation. Indeed, SVs are frequently linked to phenotypic adaptations and a previous GWAS in yeast on different growth conditions revealed a greater impact of CNVs than SNPs on phenotypic variance (Peter et al., 2018). A precise characterization of SVs using long-read sequencing strategies would therefore be necessary to assess their impact on gene expression variation in the *S. cerevisiae* species (De Coster et al., 2021).

Methods

Description of the isolates

A total of 1,010 unique isolates were gathered from different studies (Table S1). 981 isolates came from the 1,011 strains collection (Peter et al., 2018), 26 strains from Marsit et al., 2015 and 3 lab strains were added: FY4 (a haploid S288C strain), Σ1278b and CEN.PK. These strains represent the wide genetic diversity of the species and diverse geographical and ecological origins (Figure 1). Genomes of all the strains were previously sequenced and the pangenome was established for all except the 26 strains from Marsit et al., 2015.

The ploidy of the 26 isolates from Marsit et al., 2015 was estimated by flow cytometry with a quantification of the cell DNA content. Defined *S. cerevisiae* isolates from known ploidies (n to 5n) were added to the samples as controls. Cells in exponential growth phase were successively washed in water, ethanol (70°) and sodium-citrate buffer (50 mM, pH 7.5) before a RNase A treatment (500 µg/mL). To avoid cell aggregates, each sample was sonicated followed by the DNA labelling with propidium iodide (16 µg/mL), a fluorescent intercalating agent. DNA content was then quantified using the 488 nm excitation laser of the Accuri C6 plus flow cytometer (BD Biosciences).

Growth culture and RNA sequencing

The growth rate of the 1,010 strains (Table S1) was estimated after measurements of the OD during a 48h-culture in synthetic complete (SC) medium with glucose as a carbon source with a microplate reader (Tecan Infinite F200 Pro). The strains were then grouped in 96-well plates according to their growth rates and grown in 1mL of SC with glucose until cells reached exponential growth phase and an OD of ~ 0.3. Hence, 750µL of cells are transferred in sterile 0.45µM 96-well filter plates (Norgen, #40008) on a vacuum manifold (VWR, #16003-836). We applied vacuum to remove all the SC medium, sealed with aluminum foil seals (Dutscher, #760213), and flash froze the entire plate in liquid N2 to store the plate at -80°C before mRNA extraction.

mRNA were extracted with the Dynabeads® mRNA Direct Kit (ThermoFisher, #61012) from an optimized protocol to work in 96-well plates (Albert et al., 2018). Using glass beads and lysis buffer, cells have been lysed with a bead-beater (VWR, #412-0167) before RNA denaturation for 2 minutes at 65°C. Two rounds of washes followed by mRNA isolation have been realized with magnetic beads coupled to

oligo $(dT)_{25}$ residues which can hybridize the polyA tails of mRNA. A final volume of 10µL of purified mRNA was obtained to prepare the sequencing library.

Sequencing libraries were prepared with the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (NEB, #E7765L) still in 96-well plates. The 3-step protocol was followed and adapted for around 10ng of intact RNA in 5µL. The first step consisted in cDNA synthesis from 5µL of purified mRNA. The purified mRNA were fragmented during 15 minutes at 94°C and the two cDNA strands were successively synthesized. The cDNA were then purified using NEBNext sample purification magnetic beads and eluted in 50µL of 0.1X Tris-EDTA buffer. Dual index duplex adapters were ligated to the cDNA in the second step. In total, 96 combinations of TS HT dual index duplex mixed adapters from IDT® (Integrated DNA Technologies®) were used and each prepped DNA was assembled to a unique barcode combination, 2.5µL at 2µM. The adaptor-ligated DNA were purified using NEBNext sample purification magnetic beads and eluted in 15µL of 0.1X Tris-EDTA buffer. A final PCR enrichment of the barcoded DNA was performed in a 9-cycle amplification using Illumina P5 and P7 universal primers (P5 IDT : 5'-AATGATACGGCGACCACCGA-3' ; P7 IDT : 5'-CAAGCAGAAGACGGCATACGA-3'). 21µL of final barcoded DNA were purified and eluted in 0.1X Tris-EDTA buffer.

For each sample, the final barcoded DNA was quantified using the Qubit™ dsDNA HS Assay Kit (Invitrogen™) in 96-well plate with a microplate reader (Tecan Infinite F200 Pro), excitation laser set at 485nm and emission laser at 528nm. All the samples from the 96-well plate with a concentration higher than 1ng/µL were grouped with the same amount of DNA, 20ng, for each. The DNA integrity of the pool was controlled on 1% agarose gel and quantified on Nanodrop and Qubit using the Qubit™ dsDNA HS Assay Kit (Invitrogen™).

The final pool of DNA was sequenced on Nextseq 550 high-output at the EMBL Genomics Core Facilities. In total, 1,046 samples were sequenced with duplicates for some of the 1,010 isolates because the number of final reads was too low. A mean of 6.45 million single-end reads of 75bp was obtained for each sample after demultiplexing (Figure S1).

Reads cleaning and strain assignment validation

Raw reads were cleaned with cutadapt (Martin, 2011) to remove adapter as well as low quality reads that were trimmed on the basis of a Phred score threshold of 30 and discarded if less than 40 nt long after this trimming step.

For each of the 1,046 samples, clean reads were mapped to the *S. cerevisiae* reference sequence using TopHat (v2.0.13) (Trapnell et al., 2009). Resulting bam files were sorted and indexed using SAMtools (v1.9). Duplicated reads were marked using Picard (v2.18.14). GATK (v4.1.0.0) HaplotypeCaller was used to call variants in each individual sample. The resulting SNPs were intersected with rare SNPs (defined as having a Minor Allele Frequency (MAF) less than 5%) described in Peter et al., 2018 using bcftools isec. That was done in a pairwise manner, so that the SNPs of each RNAseq sample could be intersected with the previously described rare SNPs of each strain.

The 1,046 samples were ranked based on the number of shared rare SNPs with each relevant strain described in the SNP matrix. This allows to automatically validate 940 unique isolates for which the expected strain was among the top 3 ranking strains. The remaining samples were manually investigated: 24 samples that were part of a large cluster of closely related strains could be validated as the expected strain and 19 samples could be unambiguously reassigned to the top 1 ranking strain. 14 samples out of the 1,046 could not be validated or reassigned and were discarded from the remaining analyses. After this step, a final set of 986 unique isolates was validated (Table S1).

Gene expression quantification

For each validated sample, clean reads were mapped to the *S. cerevisiae* reference sequence in which the SNPs of the corresponding strains were inferred (as described in Peter et al. 2018) plus the accessory genes that were not classified as ancestral or *S. paradoxus* orthologs in Peter et al. 2018 (n = 395). The mapping was achieved using STAR (Dobin et al., 2013) with the following parameters:

--outReadsUnmapped Fastx \

--outSAMtype BAM SortedByCoordinate \

--outFilterType BySJout \

--outFilterMultimapNmax 20 \

--outFilterMismatchNmax 4 \

--alignIntronMin 20 \

--alignIntronMax 2000 \

--alignSJoverhangMin 8 \

--alignSJDBoverhangMin 1

Isolates with more than 1 million reads mapped were kept for analysis, resulting in a final set constituted of 969 strains (Table S1).
Mapped reads counts were then obtained using the featureCounts function from the Subread package (Liao et al., 2014) with the genes described in the *S. cerevisiae* reference annotation (n = 6,285) and accessory genes (n = 395) as features. The following options were used in order to get multi-mapped reads counted as a fraction of the sites they mapped to:
-M \

--fraction

Finally, Transcripts Per Million (TPM) were calculated as a measure of transcription for each of those features and a log2(TPM+0.5) normalization was applied. From the set of 6,285 reference genes, 196 were filtered out because log2(TPM+0.5) was lower than 1 in 50% of the isolates. The final set is thus constituted of 6,484 ORFs which were used for downstream analyses (Table S2).

Neighbor joining tree

The variant calling files related to the 969 final strains generated through mapping to the reference sequence and limited to SNPs were combined with GenotypeGVCFs. The biallelic segregating sites were used to construct a neighbour-joining tree with the R packages *ape* and *SNPrelate*. Briefly, the .gvcf matrix was converted into a .gds file for individual dissimilarities to be estimated for each pair of individuals with the snpgdsDiss function. The bionj algorithm was then run on the obtained distance matrix.

General analysis

The final gene set used is composed of 6,484 genes among which 395 are non-reference genes. The statistics analysis and graphics were achieved using R software with *tidyverse* and *ggplot2* packages (R Core Team, 2020).

GO (Gene Ontology) term enrichments were performed on SGD (*Saccharomyces* Genome Database) website with the GO Term Finder tool (https://www.yeastgenome.org/goTermFinder). For each analysis, the list of outlier genes was compared to the list of 6,089 annotated genes from the final dataset, unless otherwise stated. Significant enrichment was considered under "Process" ontology with a p-value threshold of 0.01.

1 – Gene expression clustering analysis

To determine the structure of the species based on gene expression variation across the yeast natural population, a principal component analysis (PCA) was performed with *autoplot* function on the 6,484 expressed genes and the 969 isolates with their lineages associated (as described in Peter et al., 2018). A hierarchical clustering was then realized with *pheatmap* function on genes with a sufficient variance across the 969 isolates, i.e. the 6,089 genes from the reference genome to identify co-expression networks.

2 – Transcriptional signature detection

Transcriptional signatures were defined for different subgroups: the *S. cerevisiae* lineages as defined in Peter et al, 2018, the beer isolates, the wild, domesticated and clinical subgroups as well as for ploidy related subgroups (haploid, diploid and polyploid) and isolates carrying aneuploidy. Differentially expressed genes were identified for each subgroup by plotting the gene expression fold change between strains from the considered subgroup to the other isolates against a p-value associated to the expression variation. A p-value cutoff of 0.05 with Bonferroni correction was set to determine outlier expressed genes either up- or downregulated according to the fold change (Tables S7, S8, S12 and S15, Figures S3-1 and S3-2).

3 – Expression levels for CNV and aneuploidies

The number of copies of the 6,085 reference genes was estimated for the 969 isolates, based on the genomic DNA reads mapping (Marsit et al., 2015; Peter et al., 2018) and the estimation of the genome coverage through 1 kb sliding-windows with Control-FREEC (v10.6) (Boeva et al., 2011). A gene was considered as duplicated if more than half of its length was in a region detected as duplicated by Control-FREEC. 6,075 genes were observed as duplicated in at least one strain. A systematic comparison of the expression levels of these genes between duplicated and non-duplicated versions was performed using a Wilcoxon rank sum test. A significant

difference of gene expression was defined with a p-value cutoff of 0.05 with Bonferroni correction (Table S20).

Aneuploidies in strains from Marsit et al., 2015 and laboratory isolates were manually detected through the coverage plots of the genomic reads mapping. For all other isolates, the aneuploidy annotations from Peter et al., 2018 were considered. The mean expression along the 16 chromosomes was calculated for each isolate and plotted according to the chromosome copy number. Transcriptional signatures of the 204 aneuploid isolates subgroup were defined as previously described.

4 – Pangenome variation of expression

All the 6,680 mapped genes were considered to examine the pangenome expression variation across the species. We used the annotations of accessory gene origins described in Peter et al., 2018 and removed the unknown and plasmidic origins from analysis. For each reference gene, the mean ratio of non-synonymous to synonymous polymorphisms (dN/dS) was computed with the yn00 program in PAML software within the 1,011 yeast collection (Peter et al., 2018; Yang, 2007). Essential genes annotations were based on the deletion collection performed in S288C reference strain in complete medium supplemented in glucose (Giaever et al., 2002) for which the list is available on SGD.

Considering the reference genes for which an ortholog from *S. paradoxus* was detected as introgressed in some strains (Peter et al., 2018), 439 genes were found in a homozygous state for *S. cerevisiae* or for *S. paradoxus* alleles in at least one strain. A systematic comparison of the expression levels of these 439 genes between *S. cerevisiae* and *S. paradoxus* allele versions was performed using a Wilcoxon rank sum test. A significant difference of gene expression was defined with a p-value cutoff of 0.05 with Bonferroni correction (Table S23).

Genome-wide association studies (GWAS)

In order to reduce low-quality mapping and linkage, we removed sub-telomeric regions, 20kb each side of the chromosomes both from the SNPs matrix and from the expressed genes. A total of 75,828 single-nucleotide polymorphism sites between 969 strains with a minor allele frequency (MAF) lower than 5% were integrated in the matrix of genetic variants. Genome-wide association studies based on mixed-model association analysis were performed as described in Peter et al., 2018 using FaST-LMM (Lippert et al., 2011). In total, the expression variation (in TPM) of

5,868 genes was tested and a trait-specific p-value threshold was established for each gene by permuting phenotypic values between individuals 100 times. The significance threshold was the 5% quantile (the 5th lowest p-value from the permutations). The phenotypic variance explained by each SNP was computed with FaST-LMM.

Local and distant eQTL were distinguished according to the distance from the considered gene: local eQTL can be located 25 kb each side around the gene, all other being considered as distant eQTL. We used CAVIAR (CAusal Variants Identification in Associated Regions) to filter the causal local eQTL associated to a phenotype (Hormozdiari et al., 2014). CAVIAR is a fine-mapping method which accounts for linkage disequilibrium (LD) and effect sizes. The default parameters of the method were used with a maximum of 2 local eQTL detected per phenotype and a probability threshold of 95% to contain all the causal variants responsible for the phenotype.

Allele specific expression (ASE)

1 – ASE data generation

We selected all the isolates previously described as diploid, euploid and heterozygous (Peter et al., 2018) in order to perform ASE analysis on this population (n = 289). We quantified the biallelic expression of each of these isolates using the GATK tool ASEReadcounter (Castel et al., 2015) by providing it for each isolate a BAM file resulting from an alignment of RNA-seq reads on the reference genome and a VCF file containing all heterozygous positions of the corresponding isolate (data obtained from Peter et al., 2018). Heterozygous sites displaying a risk of allelic mapping bias were detected using their 75 bp mappability, with Genmap software (Pockrandt et al., 2020), and discarded. We used the allele count to calculate an alternative allele ratio (AAR):

$$\frac{alternative\ allele\ counts}{reference\ allele\ counts + alternative\ allele\ counts}$$

We finally excluded sites which did not have their heterozygosity supported by their alternative allele ratio (AAR = 0 or 1).

We detected imbalance in the allele expression using a simple binomial test corrected by FDR (Benjamini and Hochberg, 1995). In order to further compensate residual mapping bias in our results, we set the probability value of the binomial test to the

mean of the alternative allele ratio in all our 289 isolates instead of 0.5 (Lappalainen et al., 2013; Montgomery et al., 2010). Moreover, we performed the previous test on sites that were covered more than 29X in order to ensure enough statistical power to our binomial test. Finally, we limited our explorations of ASE to the heterozygous sites located in CDS. In total, a list of 214,551 heterozygous sites distributed in 3,570 unique genes was analyzed across our 289 isolates (median = 464 sites per isolate).

2 – ASE and GWAS

To explore the relation between allelic imbalance and local regulation, we considered the allelic versions of the 775 cis-eQTL in the 289 heterozygous isolates and their effect on heterozygous gene expression. Allelic versions of the eQTL were distinguished depending on whether they were homozygous (n = 196,509 cases) or heterozygous (n = 20,530 cases) in the isolate. The proportion of phenotypes displaying at least 1 significantly imbalanced heterozygous site among all the phenotypes impacted by these local eQTL was estimated for both homozygous and heterozygous categories. These proportions between homozygous and heterozygous cis-eQTL were then compared using a χ^2 test and a binomial test.

3 – Subpopulation ASE analysis and *SSU1* exploration

To identify allele specific expression signatures related to the established clades, we selected for each isolate the genes displaying significant allelic imbalance in all of their sites, corresponding to 1,279 cases with 625 unique genes. We controlled the concordance of the allelic expression of these genes by inferring 2 haplotypes using the alternative allele ratio, AAR (sites with AAR > 0.5 are on a haplotype, sites with AAR < 0.5 are on another).

For each gene for which at least 2 sites corresponding to two different inferred haplotypes were available, we calculated the mean AAR for each inferred haplotype and checked if they were anti-correlated. This means that if an allele corresponding to one haplotype is highly expressed, the other should be lowly expressed. We found that the mean AAR of the different sites revealed a significant anti-correlation between haplotypes (Figure S7F), supporting the reliability of the ASE of the previously selected genes.

In total, 186/625 genes displayed imbalanced expression in several isolates. By grouping the isolates into their respective clade, we calculated the number of occurrences of these genes in each clade. We then applied Fisher's exact tests

(corrected by FDR, Benjamini and Hochberg, 1995) in order to detect genes enriched in some specific clades using contingency tables (Table S29).
The region around *SSU1* gene was explored to detect translocations previously described (García-Ríos and Guillamón, 2019; Pérez-Ortín et al., 2002) and that could explain the allelic imbalance in the wine (1.) lineage if at heterozygous state (Table S28). PCR (Polymerase Chain Reaction) were performed on the 23 candidate isolates and 3 isolates as control for:

- the wild type *SSU1* locus – no translocation, FY4xFY5 diploid isolate
- the VIII-t-XVI translocation, ABD isolate, similar to Y9 (Pérez-Ortín et al., 2002)
- the XV-t-XVI translocation, GN isolate (Marullo et al., 2020).

DNA amplification was performed with primers described Table S27 (Marullo et al., 2020; Pérez-Ortín et al., 2002) from colony thermo-lysates. PCR fragments between 450 and 570 bp were obtained on 1.5% agarose gel. The structural variant type for each allele can be discriminated according to the fragment size (Table S28).

4 – ASE and *S. paradoxus* introgressions

Besides homozygous *S. paradoxus* introgressions, heterozygous cases of *S. cerevisiae* and *S. paradoxus* alleles were also identified in the species (Peter et al., 2018). The unfiltered VCF files from Peter et al., were corrected for coverage and mapping bias, allowing to get 3,338 sites related to heterozygous introgressed genes. A significant difference was found in terms of AAR between these 3,338 introgressed sites and the non-introgressed toward low values for introgressions (Figure S7G). However, among those sites, some were displaying aberrant genetic allele balance (AB tag in the VCF file) due to soft filtration. Thus, we iteratively performed several filtration steps of the genetic allele balance. In brief, at each step, the filtration value was set to exclude extreme genetic allele balance: for example, with a filtration value of 0.1, the site with a genetic allele balance higher than 0.9 or lower than 0.1 were discarded, for 0.2 the threshold was 0.8 and 0.2, *etc*. Ultimately, this led to select sites with genetic allele balance narrowed to 0.5 but also resulted in an important decrease in the number of sites (Figure S7H). In addition, at each filtration step, we compared the AAR between heterozygous introgressed sites and non-introgressed sites and found that the AAR difference between introgression related sites and the other sites decreased as the filtration value increased (Figure S7I). Because extreme genetic allele balance could be related to difference in terms of allele copy number and since our goal was to compare the allele balance in genes

with similar genetic organization in their alleles, we finally selected sites with a genetic allele balance between 0.33 and 0.66. This resulted in 356 sites distributed in 43 heterozygous introgressed genes.

Supplementary material

Supplementary tables

Supplementary tables are available at:
https://www.dropbox.com/sh/b3307r7u1cahb8c/AADCxC-N353oEBjP9S4_EFK4a?dl=0

Supplementary figures

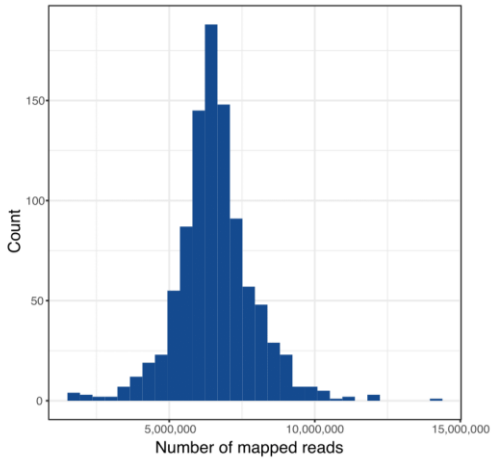


Figure S1. Distribution of the number of clean mapped reads in the final 969 isolates. A cutoff was established at 1 million of final mapped reads minimum to include an isolate in the final analysis.

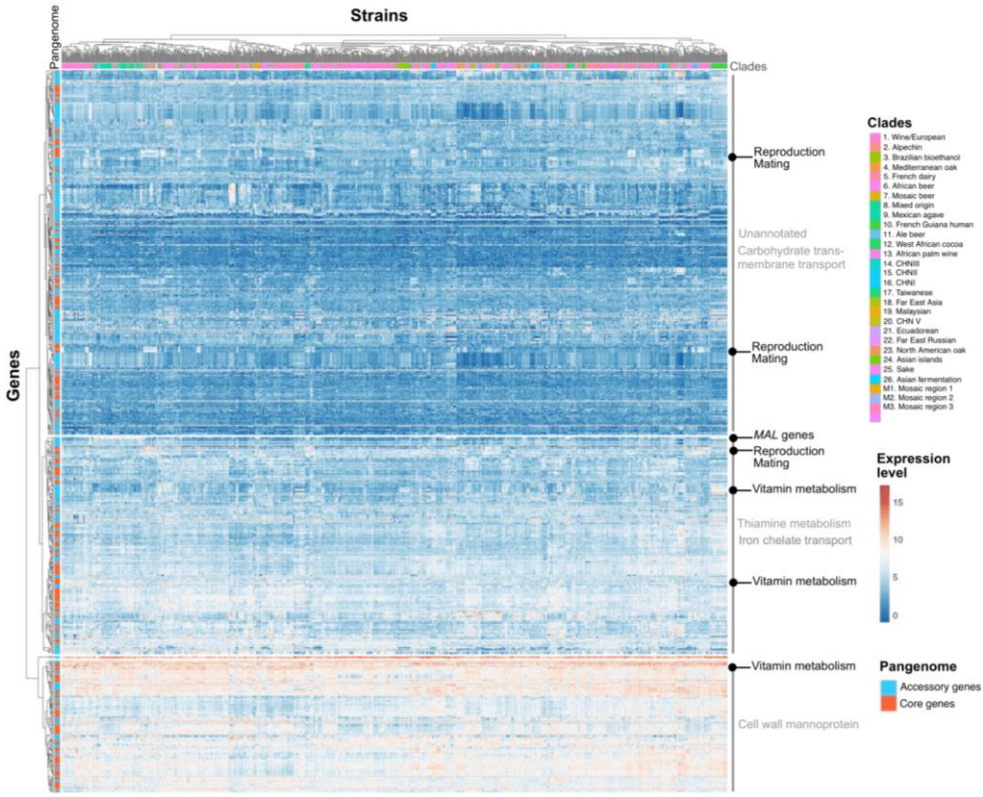
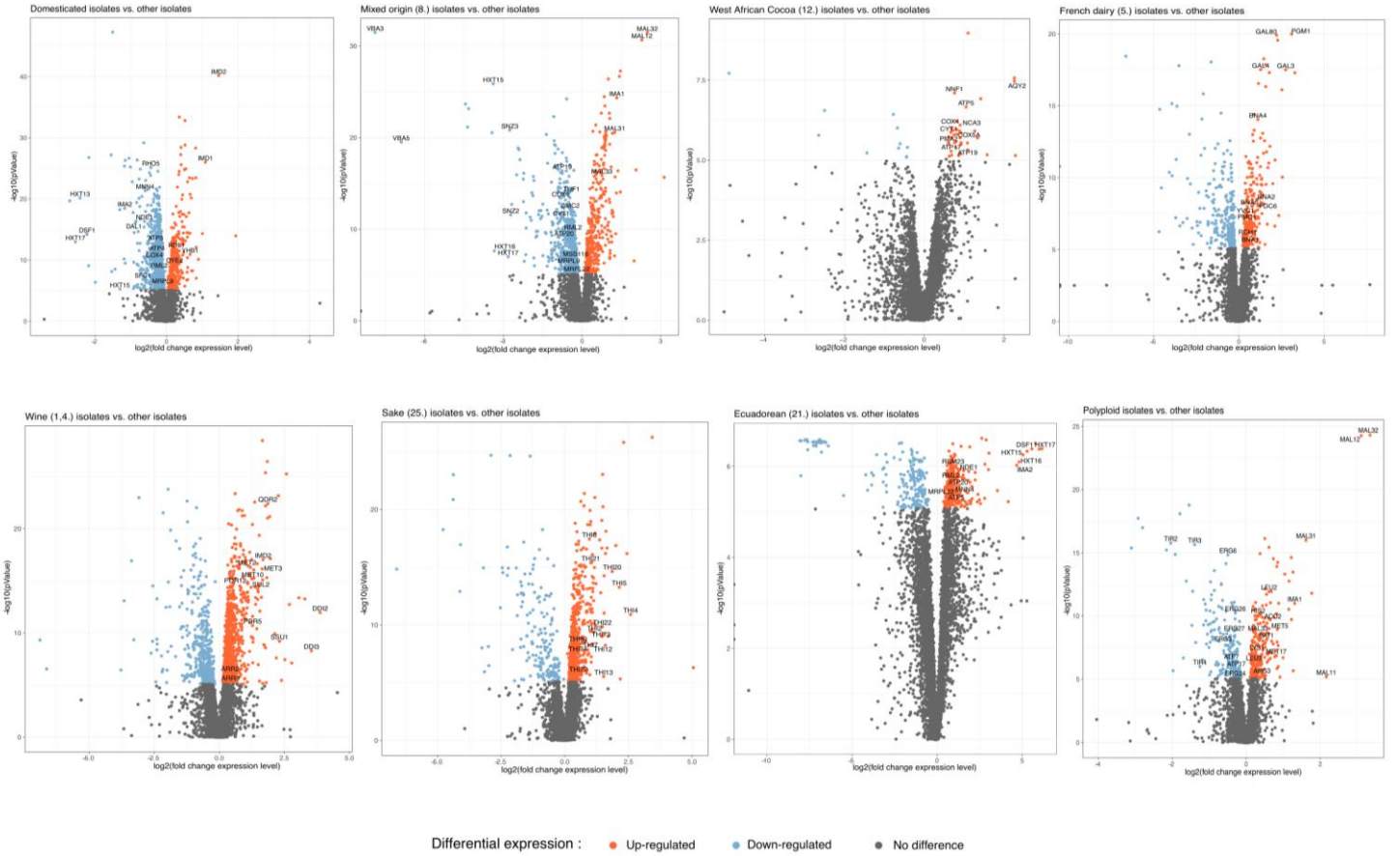


Figure S2. Structure of the most variable expressed genes across the species. Hierarchical clustering of the 650 expressed genes with a variance > 1 in the population. Core and accessory genes were identified by two colors as well as the affiliated clades for each isolate. The tree was constructed on the log2(TPM+0.5) values with *pheatmap* function and separated in 4 subgroups for which GO term enrichments were performed on SGD (p-value <= 0.01). Diverse metabolic processes or genes involved in mating and reproduction were detected within these variable genes.

Supplementary figure 3 - panel 1



Supplementary figure 3 - panel 2

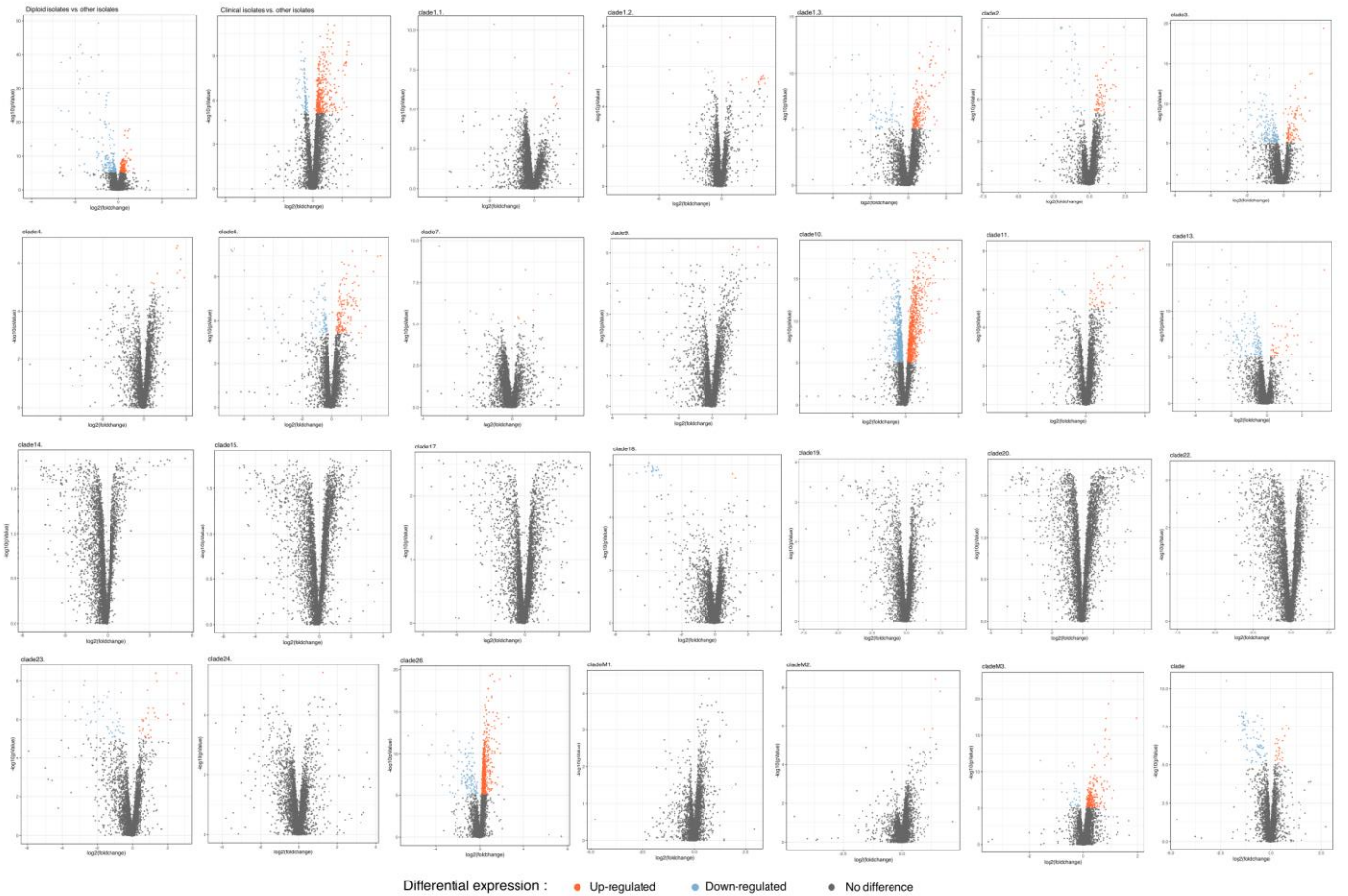
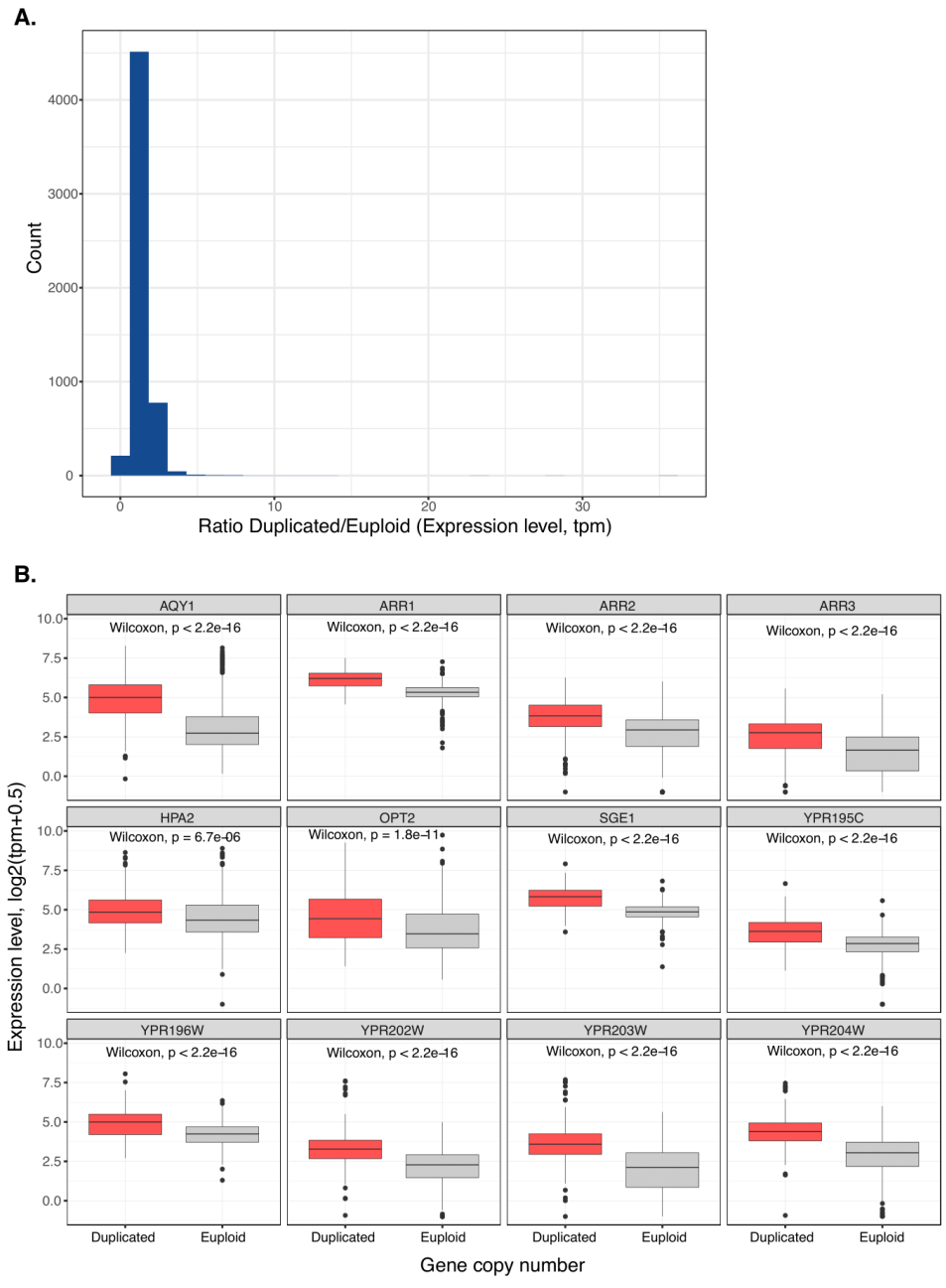


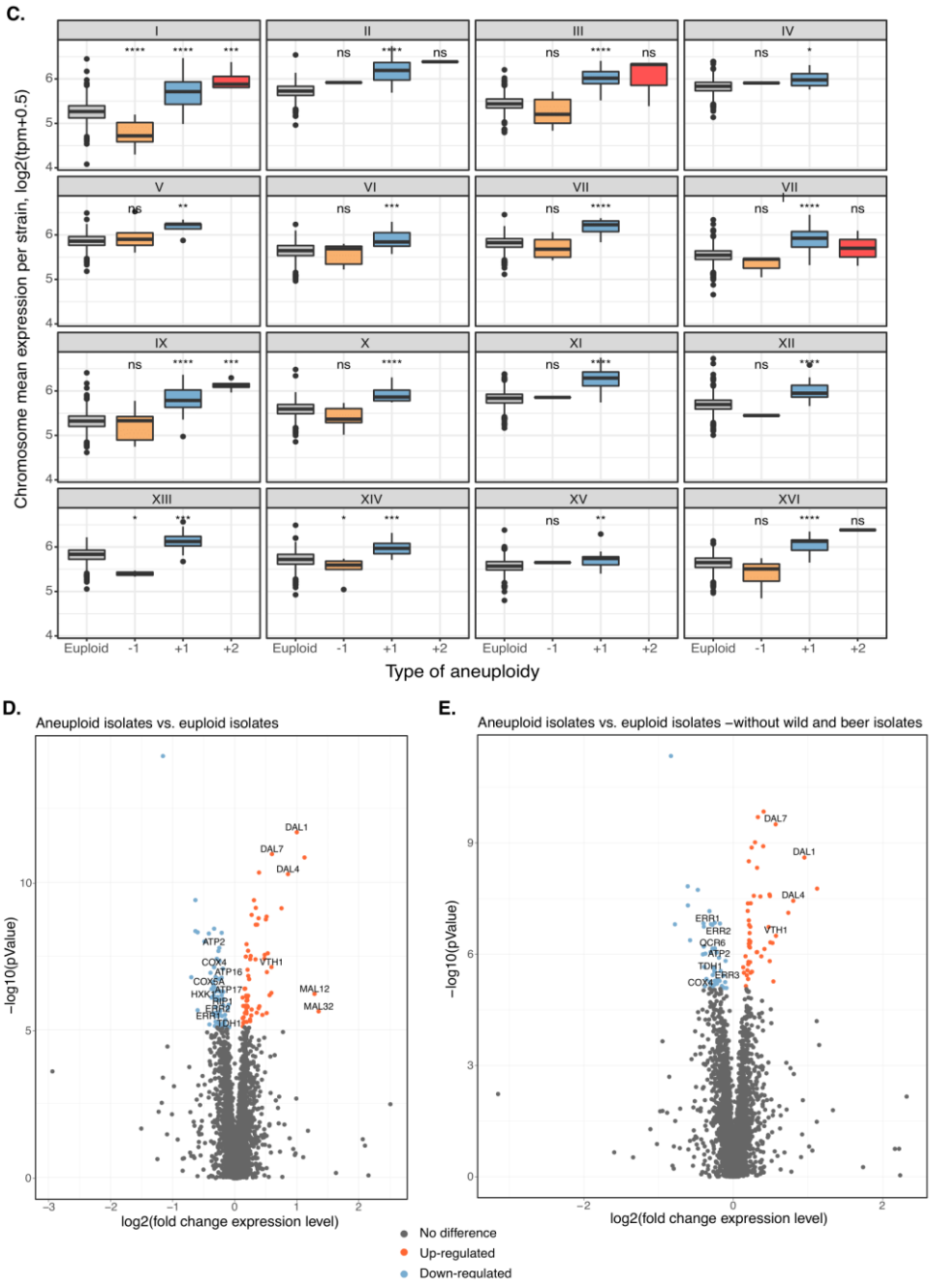
Figure S3. Volcano-plots comparing the gene expression levels between a defined subgroup and the rest of the population. Transcriptional signatures were defined with the expression level fold-change and a p-value cutoff of 0.05 with Bonferroni correction. The difference of expression (fold change) between the subgroups highlights down- (blue) or up- (red) regulated genes differentially expressed. The panel 1 represents subgroups with characterized signatures associated with the domestication impact or the ploidy level. The panel 2 represents other subgroups without strong networks differentially expressed.

Figure S4. Impact of the gene copy number on gene expression. (A.) Distribution of the ratio between expression level in TPM in duplicated genes (n copies = 2) and not duplicated (n copy = 1). The mean fold-change is 1.41 suggesting a dosage compensation of gene expression. (B.) Expression of a segmental duplication on a large subtelomeric region on the chromosome XVI containing 12 upregulated genes. Box plots of the expression level for each isolate according to the gene copy number. The p-values are calculated using a two-sided Mann–Whitney–Wilcoxon test. (C.) Box plots of the mean expression level for each isolate according to the aneuploidy type in each chromosome independently considered. The p-values are calculated using a two-sided Mann–Whitney–Wilcoxon test between euploid subgroup and each aneuploidy type, ns: not significant, * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001, **** p-value < $1e^{-04}$. (D. – E.) Transcriptional signatures in aneuploid isolates compared to euploid isolates (D.) and aneuploid isolates compared to euploid isolates without considering wild and beer isolates (E.). Outlier genes were defined with the expression level fold-change and a p-value cutoff of 0.05 with Bonferroni correction. The most upregulated genes (red) are involved in maltose metabolic pathway (D.) and in allantoin metabolism and GO term enrichments on SGD (p-value <= 0.01) revealed a downregulation of ATP metabolic process genes (blue).

Supplementary figure 4 - panel 1



Supplementary figure 4 - panel 2



Supplementary figure 5 - panel 1

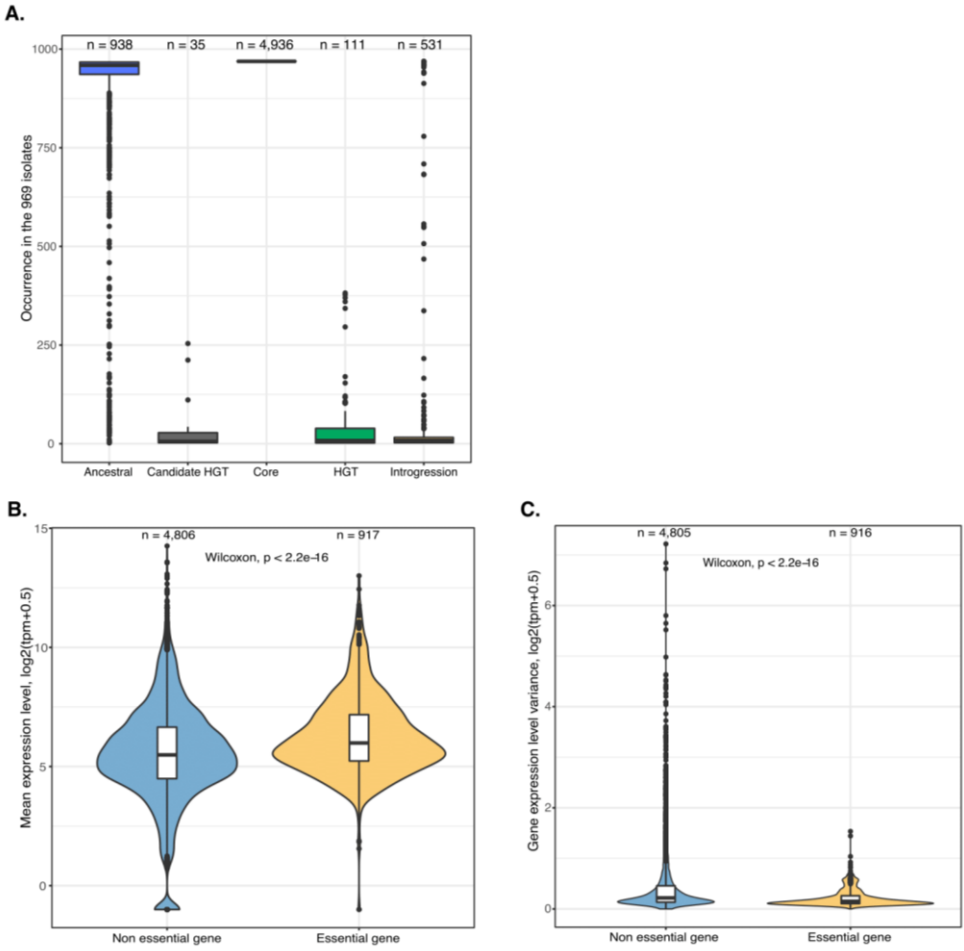
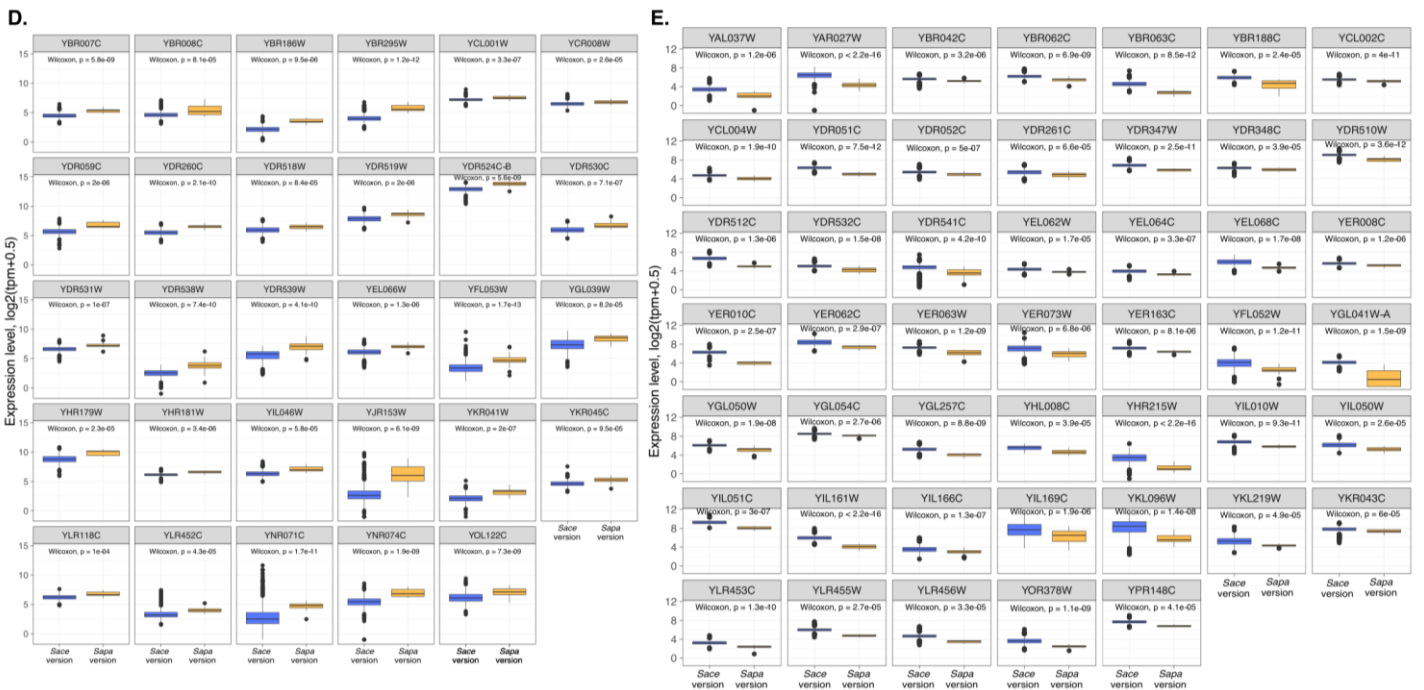


Figure S5. Gene expression in the pangenome. (A.) Distribution of the number of each ORF type in the 969 isolates. (B. – C.) Mean gene expression (B.) and gene expression variance (C.) in essential (yellow) and non-essential genes. The p-values are calculated using a two-sided Mann–Whitney–Wilcoxon test. (D. – E.) Outlier genes up- (D.) or down- (E.) regulated in the homozygous *S. paradoxus* introgression compared to its *S. cerevisiae* homozygous ortholog identified by a Wilcoxon-test, with Bonferroni correction. The p-values shown on the plots are calculated using a two-sided Mann–Whitney–Wilcoxon test.

Supplementary figure 5 - panel 2



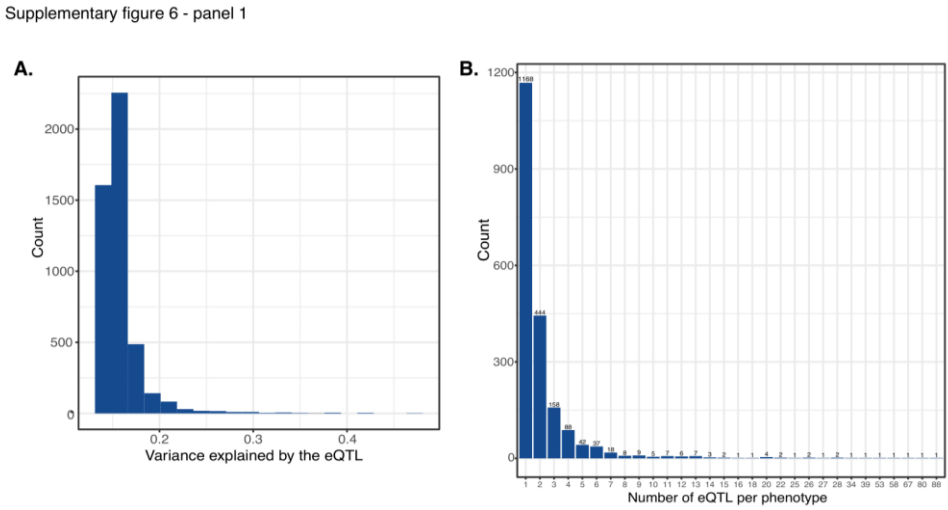
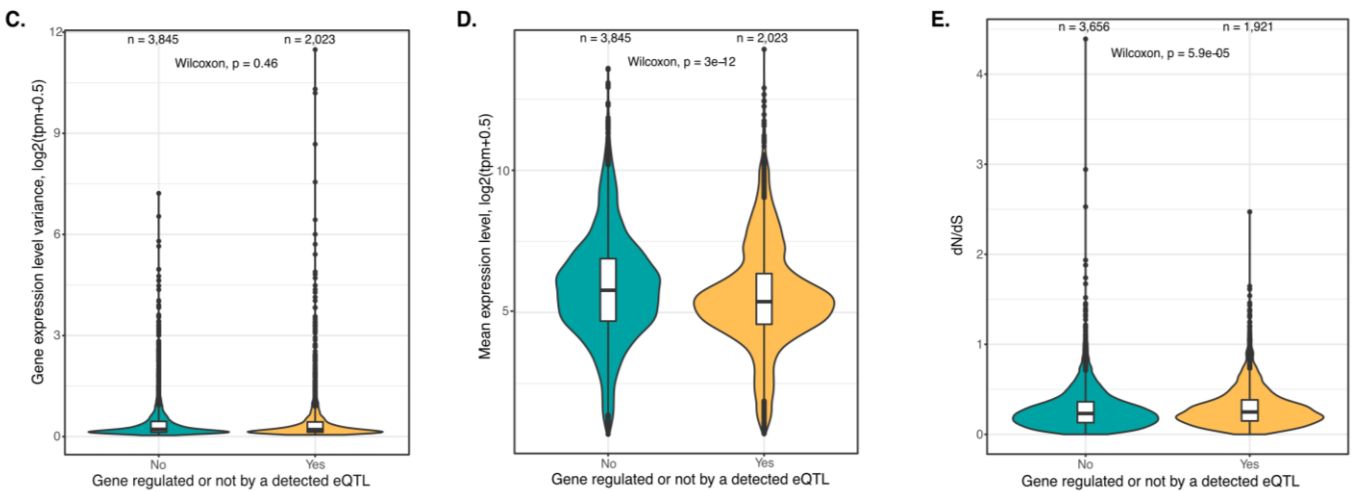


Figure S6. Overview of the GWAS results. (A.) Distribution of the phenotypic variance explained by the 4,684 eQTL associated with the expression level variation of 2,023 genes. (B.) Number of eQTL regulating a gene expression variation. (C. – E.) Exploration of the characterics of gene regulated by a detected eQTL (yellow) or without detected eQTL (green). Genes regulated by at least one eQTL are not more or less variable (C.) but are less expressed (D.) and have less constraints in their genetic sequence, illustrated by a higher dN/dS (E.). The p-values are calculated using a two-sided Mann–Whitney–Wilcoxon test. (F.) Distribution of the 4,684 eQTL position along the 16 chromosomes. (G.) dN/dS of genes in which we detected eQTL (yellow) or not (green). Genes with regulatory variants have thus less genetic constraints. The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test. (H.) Distribution of the phenotypic variance explained by the eQTL located in a gene (blue) or in an intergenic region (red). The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test.

Supplementary figure 6 - panel 2



Supplementary figure 6 - panel 3

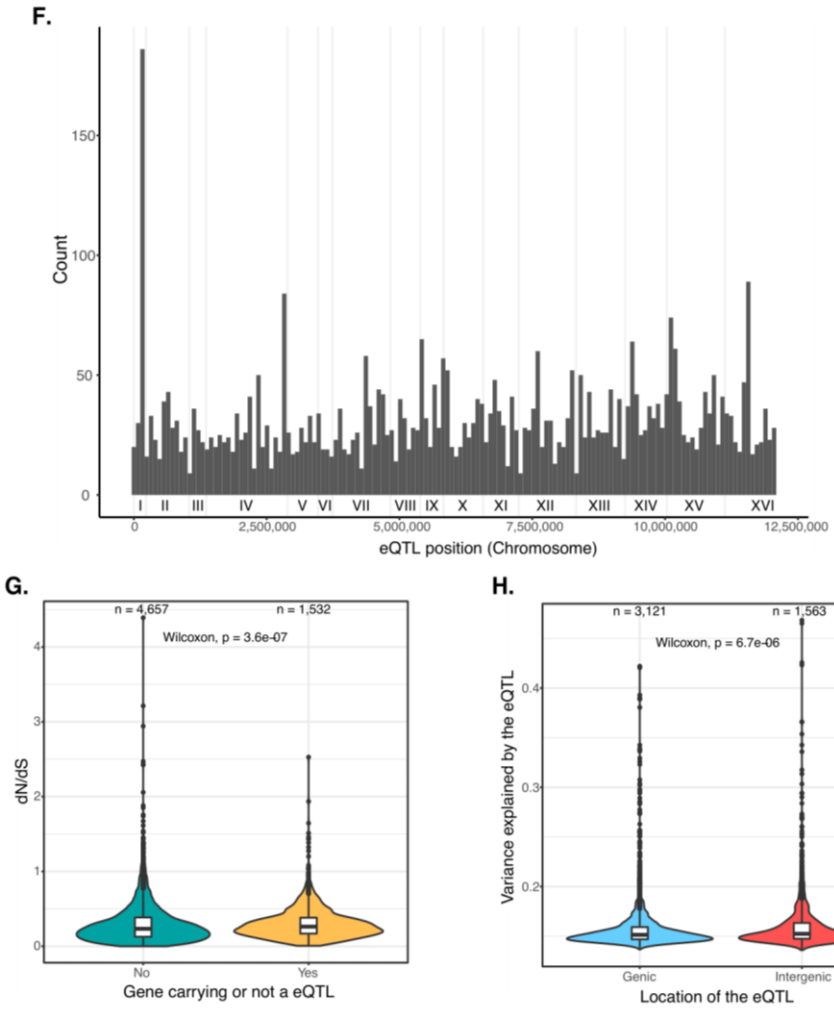
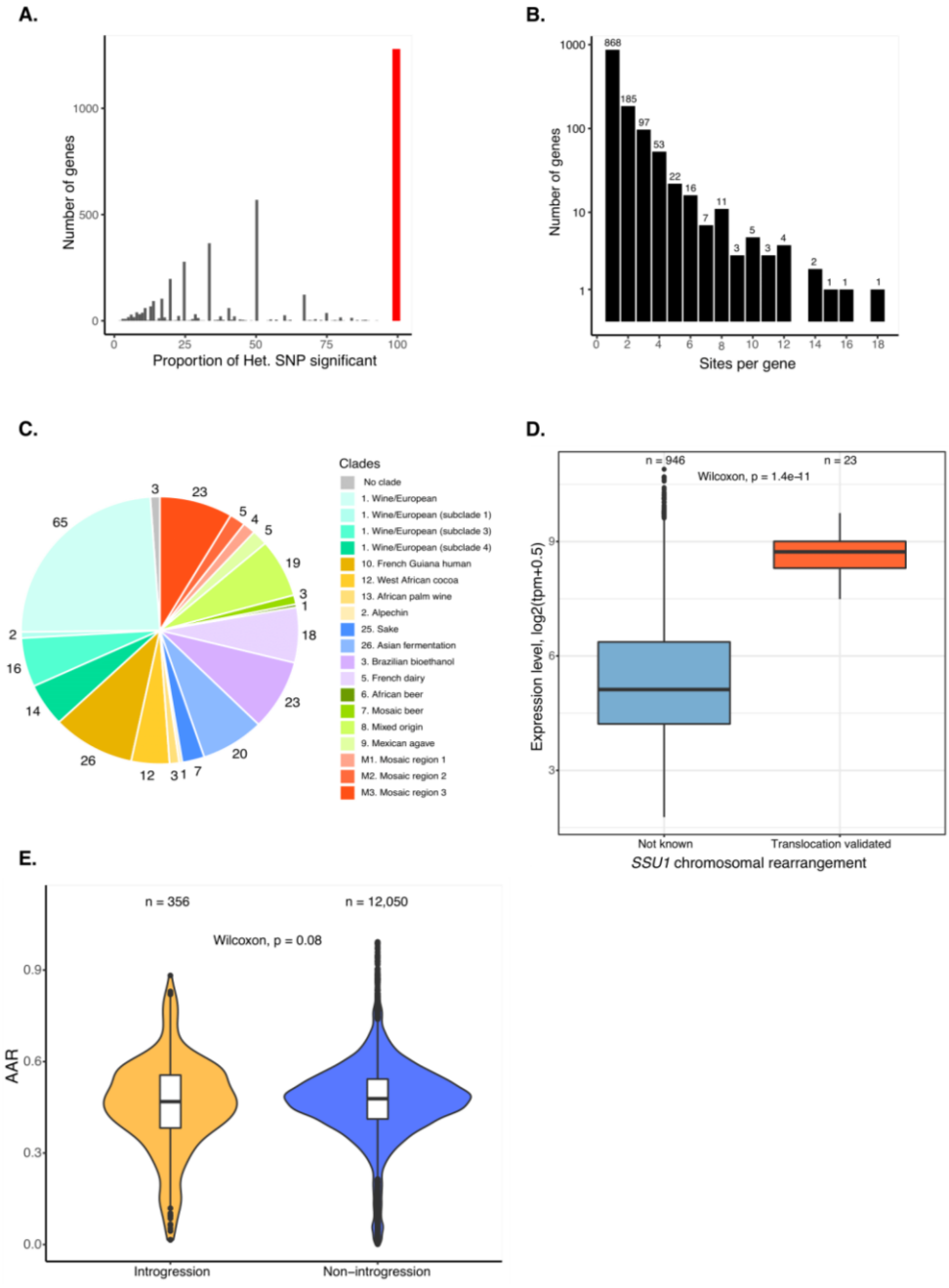
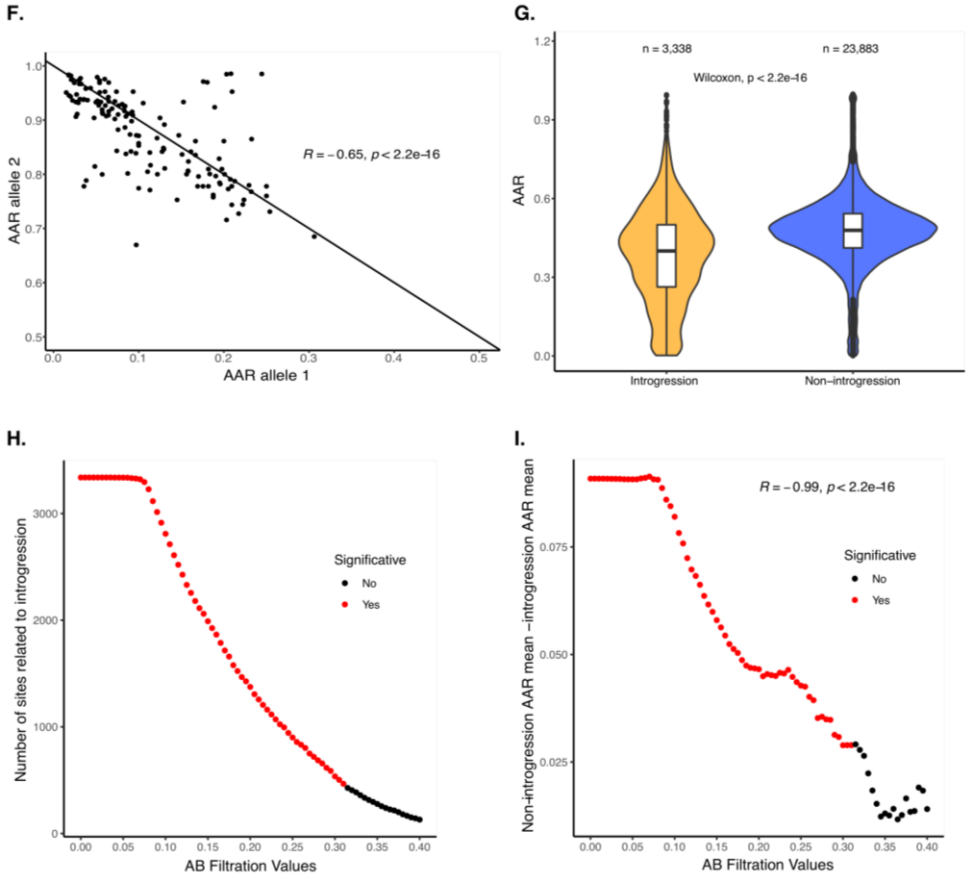


Figure S7. Allele Specific Expression (ASE) analysis. (A.) Proportion of heterozygous sites inside a gene displaying significant imbalance (x-axis) and the number of genes displaying the different proportion (y-axis). Genes having all their sites (100%) significantly imbalanced correspond to the red bar. (B.) Distribution of the number of sites per gene (after coverage and mapping bias correction) among 1,279 genes having all (100%) of their site significantly imbalanced. (C.) Distribution of the 270 heterozygous isolates with at least one gene carrying 100% of its sites significantly imbalanced across the different clades. (D.) *SSU1* gene expression variation between 23 isolates carrying a PCR-validated translocation on one of the two chromosomes and other isolates without a characterized *SSU1* region. The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test. (E.) Comparison between the alternative allele ratio (AAR) from heterozygous introgressed sites (after genetic allele balance filtration) and the AAR from non-introgressed heterozygous sites. The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test. (F.) Alternative allele ratio (AAR) correlation among the inferred haplotypes (see methods) of the genes having 100% of heterozygous sites significantly imbalanced. The correlation index and p-value are calculated using a Spearman correlation test. (G.) Comparison between the alternative allele ratio (AAR) from heterozygous introgressed sites (before genetic allele balance filtration) and the AAR from non-introgressed heterozygous sites. The p-value is calculated using a two-sided Mann–Whitney–Wilcoxon test. (H.) Difference between the mean alternative allele ratio (AAR) of the heterozygous introgressed sites and the other sites (y-axis) in function of the genetic allele balance (AB) filtration values (x-axis). The points' color highlights if the difference between the mean of introgressed and non-introgressed AAR is significantly different (using Wilcoxon test, p-value adjusted with Bonferroni correction). The correlation index and p-value were calculated using a Spearman correlation test. (I.) Number of remaining introgressed sites after each step of genetic allele balance values. The points' color highlights if the difference between the mean of introgressed and non-introgressed AAR is significant (using Wilcoxon test, p-value adjusted with Bonferroni correction).

Supplementary figure 7 - panel 1



Supplementary figure 7 - panel 2



References

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

Albert, F.W., Bloom, J.S., Siegel, J., Day, L., and Kruglyak, L. (2018). Genetics of trans-regulatory variation in gene expression. Elife *7*, 1–44.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K.M.M., Cao, J., Chae, E., Dezwaan, T.M.M., Ding, W., et al. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell *166*, 481–491.

Ascencio, D., Diss, G., Gagnon-Arsenault, I., Dubé, A.K., DeLuna, A., and Landry, C.R. (2021). Expression attenuation as a mechanism of robustness against gene duplication. Proc. Natl. Acad. Sci. U. S. A. *118*.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B *57*, 289–300.

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

de Boer, C.G., Vaishnav, E.D., Sadeh, R., Abeyta, E.L., Friedman, N., and Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat. Biotechnol. *38*, 56–65.

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

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science (80-.). *296*, 752–755.

Brion, C., Pflieger, D., Friedrich, A., and Schacherer, J. (2015). Evolution of intraspecific transcriptomic landscapes in yeasts. Nucleic Acids Res. *43*, 4558–4568.

Buniello, A., Macarthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.

Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biol. *16*, 195.

Chen, Y., Chen, S., Li, K., Zhang, Y., Huang, X., Li, T., Wu, S., Wang, Y., Carey, L.B., and

Qian, W. (2019). Overdosage of Balanced Protein Complexes Reduces Proliferation Rate in Aneuploid Cells. Cell Syst. *9*, 129-142.e5.

Cook, D.E., Zdraljevic, S., Roberts, J.P., and Andersen, E.C. (2017). CeNDR, the Caenorhabditis elegans natural diversity resource. Nucleic Acids Res. *45*, D650–D657.

Coolon, J.D., McManus, C.J., Stevenson, K.R., Graveley, B.R., and Wittkopp, P.J. (2014). Tempo and mode of regulatory evolution in Drosophila. Genome Res. *24*, 797–808.

De Coster, W., Weissensteiner, M.H., and Sedlazeck, F.J. (2021). Towards population-scale long-read sequencing. Nat. Rev. Genet. *30*.

D'Angiolo, M., Chiara, M. De, Yue, J., Irizar, A., Stenberg, S., Llored, A., Barré, B., Schacherer, J., Marangoni, R., and Gilson, E. (2020). A yeast living fossil reveals the origin of genomic introgressions. Nature *587*.

Dephoure, N., Hwang, S., O'Sullivan, C., Dodgson, S.E., Gygi, S.P., Amon, A., and Torres, E.M. (2014). Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. Elife *3*, e03023.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dowell, R.D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D.A., Alexander Rolfe, P., Heisler, L.E., Chin, B., et al. (2010). Genotype to phenotype: A Complex problem. Science (80-.). *328*, 469.

Duveau, F., Yuan, D.C., Metzger, B.P.H., Hodgins-Davis, A., and Wittkopp, P.J. (2017). Effects of mutation and selection on plasticity of a promoter activity in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. *114*, E11218–E11227.

Freese, E.B., Chu, M.I., and Freese, E. (1982). Initiation of yeast sporulation of partial carbon, nitrogen, or phosphate deprivation. J. Bacteriol. *149*, 840–851.

Galitski, T., Saldanha, A.J., Styles, C.A., Lander, E.S., and Fink, G.R. (1999). Ploidy regulation of gene expression. Science *285*, 251–254.

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., et al. (2016). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. Cell *166*, 1397-1410.e16.

García-Ríos, E., and Guillamón, J.M. (2019). Sulfur dioxide resistance in Saccharomyces cerevisiae: beyond SSU1. Microb. Cell (Graz, Austria) *6*, 527–530.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell *11*, 4241–4257.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387–391.

Gilad, Y., Rifkin, S.A., and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. *24*, 408–415.

Gonçalves, M., Pontes, A., Almeida, P., Barbosa, R., Serra, M., Libkind, D., Hutzler, M., Gonçalves, P., and Sampaio, J.P. (2016). Distinct Domestication Trajectories in Top-Fermenting Beer Yeasts and Wine Yeasts. Curr. Biol. *26*, 2750–2761.

Gonzales, N.M., Seo, J., Hernandez Cordero, A.I., St. Pierre, C.L., Gregory, J.S., Distler, M.G., Abney, M., Canzar, S., Lionikas, A., and Palmer, A.A. (2018). Genome wide association analysis in a mouse advanced intercross line. Nat. Commun. *9*.

Gorkovskiy, A., and Verstrepen, K.J. (2021). The Role of Structural Variation in Adaptation and Evolution of Yeast and Other Fungi. Genes (Basel). *12*.

Haber, J.E. (2012). Mating-type genes and MAT switching in Saccharomyces cerevisiae. Genetics *191*, 33–64.

Hagen, D.C., McCaffrey, G., and Sprague, G.F.J. (1986). Evidence the yeast STE3 gene encodes a receptor for the peptide pheromone a factor: gene sequence and implications for the structure of the presumed receptor. Proc. Natl. Acad. Sci. U. S. A. *83*, 1418–1422.

Hahn, S., and Young, E.T. (2011). Transcriptional regulation in saccharomyces cerevisiae: Transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. Genetics *189*, 705–736.

Helguera, P., Seiglie, J., Rodriguez, J., Hanna, M., Helguera, G., and Busciglio, J. (2013). Adaptive downregulation of mitochondrial function in down syndrome. Cell Metab. *17*, 132–140.

Hill, M.S., Vande Zande, P., and Wittkopp, P.J. (2021). Molecular and evolutionary processes generating variation in gene expression. Nat. Rev. Genet. *22*, 203–215.

Horák, J. (2013). Regulations of sugar transporters: insights from yeast. Curr. Genet. *59*, 1–31.

Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.

Hose, J., Yong, C.M., Sardi, M., Wang, Z., Newton, M.A., and Gasch, A.P. (2015). Dosage compensation can buffer copynumber variation in wild yeast. Elife *4*, 1–27.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Johnston, M. (1987). A model fungal gene regulatory mechanism: the GAL genes of Saccharomyces cerevisiae. Microbiol. Rev. *51*, 458–476.

Kawakatsu, T., Huang, S. shan C., Jupe, F., Sasaki, E., Schmitz, R.J.J., Urich, M.A.A., Castanon, R., Nery, J.R.R., Barragan, C., He, Y., et al. (2016). Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell *166*, 492–505.

Kita, R., Venkataram, S., Zhou, Y., and Fraser, H.B. (2017). High-resolution mapping of cis-

regulatory variation in budding yeast. Proc. Natl. Acad. Sci. U. S. A. *114*, E10736–E10744.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Lee, D., Zdraljevic, S., Stevens, L., Wang, Y., Tanny, R.E., Crombie, T.A., Cook, D.E., Webster, A.K., Chirakar, R., Baugh, L.R., et al. (2021). Balancing selection maintains hyper-divergent haplotypes in Caenorhabditis elegans. Nat. Ecol. Evol. *5*, 794–807.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods *8*, 833–835.

Liu, Y., Borel, C., Li, L., Müller, T., Williams, E.G., Germain, P.-L., Buljan, M., Sajic, T., Boersema, P.J., Shao, W., et al. (2017). Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. Nat. Commun. *8*, 1212.

Lubliner, S., Regev, I., Lotan-Pompan, M., Edelheit, S., Weinberger, A., and Segal, E. (2015). Core promoter sequence in yeast is a major determinant of expression level. Genome Res. *25*, 1008–1017.

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

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

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal; Vol 17, No 1 Next Gener. Seq. Data Anal.

Marullo, P., Claisse, O., Raymond Eder, M.L., Börlin, M., Feghali, N., Bernard, M., Legras, J.L., Albertin, W., Rosa, A.L., and Masneuf-Pomarede, I. (2020). SSU1 Checkup, a Rapid Tool for Detecting Chromosomal Rearrangements Related to the SSU1 Promoter in Saccharomyces cerevisiae: An Ecological and Technological Study on Wine Yeast. Front. Microbiol. *11*, 1–14.

Metzger, B.P.H., Wittkopp, P.J., and Coolon, J.D. (2017). Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species. Genome Biol. Evol. *9*, 843–854.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.

Neiman, A.M. (2011). Sporulation in the budding yeast Saccharomyces cerevisiae. Genetics *189*, 737–765.

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J.-

L., Wincker, P., Casaregola, S., et al. (2009). Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. Proc. Natl. Acad. Sci. U. S. A. *106*, 16333–16338.

Oba, T., Suenaga, H., Nakayama, S., Mitsuiki, S., Kitagaki, H., Tashiro, K., and Kuhara, S. (2011). Properties of a high malic acid-producing strains of Saccharomyces cerevisiae isolated from sake mash. Biosci. Biotechnol. Biochem. *75*, 2025–2029.

Ogihara, F., Kitagaki, H., Wang, Q., and Shimoi, H. (2008). Common industrial sake yeast strains have three copies of the AQY1-ARR3 region of chromosome XVI in their genomes. Yeast *25*, 419–432.

Otto, S.P. (2007). The Evolutionary Consequences of Polyploidy. Cell *131*, 452–462.

Pérez-Ortín, J.E., Querol, A., Puig, S., and Barrio, E. (2002). Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. Genome Res. *12*, 1533–1539.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature *556*, 339–344.

Pockrandt, C., Alzamel, M., Iliopoulos, C.S., and Reinert, K. (2020). GenMap: ultra-fast computation of genome mappability. Bioinformatics *36*, 3687–3692.

R Core Team (2020). R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria URL https://www.R-project.org/.

Read, T.D., and Massey, R.C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. Genome Med. *6*, 109.

Rockman, M. V, and Kruglyak, L. (2006). Genetics of global gene expression. Nat. Rev. Genet. *7*, 862–872.

Rockman, M. V, Skrovanek, S.S., and Kruglyak, L. (2010). Selection at linked sites shapes heritable phenotypic variation in C. elegans. Science *330*, 372–376.

Ronald, J., and Akey, J.M. (2007). The evolution of gene expression QTL in Saccharomyces cerevisiae. PLoS One *2*, e678.

Ronald, J., Brem, R.B., Whittle, J., and Kruglyak, L. (2005). Local regulatory variation in Saccharomyces cerevisiae. PLoS Genet. *1*, e25.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. Nature *422*, 297–302.

Schaefke, B., Emerson, J.J., Wang, T.-Y., Lu, M.-Y.J., Hsieh, L.-C., and Li, W.-H. (2013). Inheritance of gene expression level and selective constraints on trans- and cis-regulatory changes in yeast. Mol. Biol. Evol. *30*, 2121–2133.

Shobayashi, M., Ukena, E., Fujii, T., and Iefuji, H. (2007). Genome-wide expression profile

of sake brewing yeast under shaking and static conditions. Biosci. Biotechnol. Biochem. *71*, 323–335.

Skelly, D.A., Merrihew, G.E., Riffle, M., Connelly, C.F., Kerr, E.O., Johansson, M., Jaschob, D., Graczyk, B., Shulman, N.J., Wakefield, J., et al. (2013). Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Res. *23*, 1496–1504.

Tam, J., and van Werven, F.J. (2020). Regulated repression governs the cell fate promoter controlling yeast meiosis. Nat. Commun. *11*, 2271.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484.

Tan, Z., Hays, M., Cromie, G.A., Jeffery, E.W., Scott, A.C., Ahyong, V., Sirr, A., Skupin, A., and Dudley, A.M. (2013). Aneuploidy underlies a multicellular phenotypic switch. Proc. Natl. Acad. Sci. U. S. A. *110*, 12367–12372.

The GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-.). *348*, 648 LP – 660.

The GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

Tirosh, I., Barkai, N., and Verstrepen, K.J. (2009). Promoter architecture and the evolvability of gene expression. J. Biol. *8*, 95.

Torres, E.M., Sokolsky, T., Tucker, C.M., Chan, L.Y., Boselli, M., Dunham, M.J., and Amon, A. (2007). Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science *317*, 916–924.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2004). Evolutionary changes in cis and trans gene regulation. Nature *430*, 85–88.

Wodicka, L., Dong, H., Mittmann, M., Ho, M.H., and Lockhart, D.J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat. Biotechnol. *15*, 1359–1367.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yates, J.R., Ruse, C.I., and Nakorchevsky, A. (2009). Proteomics by mass spectrometry:

approaches, advances, and applications. Annu. Rev. Biomed. Eng. *11*, 49–79.

Yona, A.H., Manor, Y.S., Herbst, R.H., Romano, G.H., Mitchell, A., Kupiec, M., Pilpel, Y., and Dahan, O. (2012). Chromosomal duplication is a transient evolutionary solution to stress. Proc. Natl. Acad. Sci. U. S. A. *109*, 21010–21015.

Zimmer, A., Durand, C., Loira, N., Durrens, P., Sherman, D.J., and Marullo, P. (2014). QTL dissection of Lag phase in wine fermentation reveals a new translocation responsible for Saccharomyces cerevisiae adaptation to sulfite. PLoS One *9*, e86298.

CHAPITRE II

Exploration of the differential impact of loss-of-function mutations linked to genetic backgrounds using transposition saturation

Introduction

Among individuals, the same mutation can sometimes lead to different phenotypes due to standing genomic variations in different genetic backgrounds (Chandler et al., 2014; Chen et al., 2016; Chow, 2016; Cooper et al., 2013; Fournier and Schacherer, 2017; Hou et al., 2018; Mullis et al., 2018; Sackton and Hartl, 2016). Such background effect could have broad implications in phenotype-genotype correlation studies, including health and diseases. Indeed, background effects have been found in multiple human Mendelian disorders, where individuals carrying the same causal mutation can display a wide range of clinical symptoms, including variable severity, clinical outcomes and age-of-onset (Chen et al., 2016; Chow et al., 2016; Cooper et al., 2013; Cutting, 2010; Dorfman, 2012; Steinberg and Sebastiani, 2012). The underlying origins of this background effect can be both extrinsic, i.e. due to environmental factors (Cutting, 2010; Williams et al., 2008) and intrinsic, i.e. due to interactions between the causal variant and other genetic modifiers (Chow et al., 2016; Dorfman, 2012; Steinberg and Sebastiani, 2012). So far, a handful of examples of modifier genes have been identified associated with human disorders, most notably in cystic fibrosis (Cutting, 2010; Dorfman, 2012). However, such examples remain rare due to the low number of sample cases in most human Mendelian diseases.

In recent years, several large-scale surveys in different model organisms such as yeasts, nematodes or fruit flies highlighted the broad influence of genetic backgrounds on the phenotypic outcomes associated with gene loss-of-function mutations (Blomen et al., 2015; Boutros et al., 2004; Dowell et al., 2010; Hart et al., 2015; Kamath et al., 2003; Kim et al., 2010; Paaby et al., 2015; Vu et al., 2015; Wang et al., 2015). In yeast, a hallmark study comparing systematic gene deletion collections in two laboratory strains, Σ1278b and S288C, showed that ~1% of all genes (57/5,100) can display background-dependent gene essentiality, where the deletion of the same gene can be lethal in one background but not in the other (Dowell et al., 2010). Several mechanisms underlying such background-dependent gene essentiality have been identified, including cyto-nuclear interactions between the mitochondrial genome and/or viral elements with the nuclear genome (Edwards et al., 2014) as well as genetic interactions between the primary deletion gene and background-specific genetic modifiers (Hou et al., 2019). While gene essentiality can be the most severe manifestation of background-dependent phenotype associated with gene loss-of-function, other variation, such as gain- or loss-of-fitness related to

different genetic backgrounds and different environmental conditions are also common in yeast (Galardini et al., 2019; Mullis et al., 2018). For example, ~20% of yeast genes showed background-dependent fitness variation across a large panel of culture conditions including the presence of various drugs, osmotic stress and nutrient sources in 4 genetically diverse strains (Galardini et al., 2019). Nevertheless, all studies currently available only include a handful of genetic backgrounds in any model systems and therefore cannot accurately reflect the extend of background effect at the species level.

In the *Saccharomyces cerevisiae* yeast, a collection of 1,011 diverse isolates originated from various ecological and geographical sources has been completely sequenced (Peter et al., 2018). This collection represents an incomparable resource to systematically study the genetic background effect at the species level. Several strategies exploring gene loss-of-function phenotypes have been developed in *S. cerevisiae*, including PCR-based systematic gene deletions (Dowell et al., 2010), gene-disruption using CRSIPR-Cas9 based methods (Sadhu et al., 2018), repeated backcrosses (Galardini et al., 2019) and transposon mutagenesis (Gangadharan et al., 2010; Michel et al., 2017; Van Opijnen and Levin, 2020). Among these strategies, transposon mutagenesis based on random excision and insertion are particularly attractive to explore a large number of genetically diverse individuals in parallel. Such methods rely on transposition events through a carrier plasmid, which allow for the generation of millions of mutants carrying genome insertions leading to gene loss-of-functions (Van Opijnen and Levin, 2020). Due to the random insertion patterns in each genetic background, these methods are not reliant on sequence homology as it is the case for traditional PCR-based gene deletions and CRISPR-Cas9 related strategies (Sadhu et al., 2018; Sharon et al., 2018). They also do not present the risk of inadvertently introducing exogenous genomic regions as it might be the case for backcross-based strategies (Galardini et al., 2019).

In this study, we selected over one hundred natural isolates that are broadly representative of the diversity across *S. cerevisiae* species and performed transposon saturation analyses using the *Hermes* transposition system (Gangadharan et al., 2010). We sequenced and analyzed large pools of transposon insertion mutants and constructed a logistic model to predict the fitness effects of gene loss-of-function based on the insertion densities within and around each annotated gene. The modeled fitness was generally reflective of major cellular processes. Comparing the fitness prediction between the isolates and the reference S288C, we identified ~600 genes

with background-dependent fitness variation, corresponding to ~15% of the genome (632/4,469). A large fraction of these background-dependent fitness genes showed continuous variation across the population and highlighted a functional rewiring between mitochondrial functions and transcription & chromatin remodeling, as well as nuclear-cytoplasmic transport. Overall, background-dependent fitness genes are functionally coherent, with members of the same protein complex of biological process showing similar variability within a given genetic background. Background-dependent fitness genes tend to show an intermediate level of integration in genetic networks compared to non-essential and essential genes, and might be under positive or relaxed purifying selection at the population level.

Results

Transposon saturation using the *Hermes* system

To get an overview of fitness variation associated with loss-of-function mutations across different genetic backgrounds in *S. cerevisiae*, we performed transposon saturation analyses in various natural isolates using the *Hermes* transposon system. The *Hermes* transposon system was previously adapted in yeast to allow for the selection of random insertion events in liquid culture, which makes this system particularly suitable for parallel analyses of large numbers of genetically diverse individuals (Gangadharan et al., 2010). This system relies on a centromeric plasmid that contains the *Hermes* transposase under the control of a modified galactose inducible promoter *GalS*, together with a transposon carrying a selectable marker (Figure 1A). Briefly, for any strain of interest, the plasmid is first transformed into stable haploid cells, then propagated in media containing galactose to induce excision and reinsertion of the transposon at a random genome location, thereby generating a large pool of individuals that carries hundreds of thousands insertions along the genome (Figure 1A). After a recovery phase in rich media, the genomic DNA of this mutant pool is extracted, then fragmented and circularized (Figure 1A). Using PCR with outward facing primers targeting specifically the transposon, a library that contains exclusively the insertion sites can be constructed and subsequently sequenced using standard Illumina methods (Figure 1A). In principle, transposon insertions that cause severe fitness defect, for example those occurred in essential genes, will not be recovered due to the competitive disadvantage compared to events occurred in genes that are not essential. Analyzing the insertion patterns along the genomes of different individuals therefore provides a proxy for fitness variation of loss-of-function mutations.

In total, we selected 107 natural isolates originated from diverse ecological and geographical sources that are broadly representative of the species diversity (Figure 1B, Table S1). Stable haploid variants of this set of isolates has been generated previously (Fournier et al., 2019; Hou et al., 2016) and are all able to grow in galactose media. We adapted the published version of the *Hermes* transposon plasmid to carry a hygromycin resistance marker instead of nourseothricin to ensure compatibility with the selected strains, which may carry either a *KanMX* or *NatMX* marker at the *HO* locus. Transposon insertion profiles for each isolate were obtained as described (Figure 1A). We observed a marked variability in terms of insertion

efficiency across different genetic backgrounds, ranging from ~100 to ~300,000 unique insertion sites (Figure S1A, Table S1). No discernible correlation between the genetic origin of the isolates and the transposon insertion efficiency was observed (Table S1). We next compared the insertion preferences (Gangadharan et al., 2010) between the reference strain S288C and the remaining 106 isolates (Figure S1B). Insertion densities for known sequence motifs were conserved across different genetic backgrounds (Figure S1B).

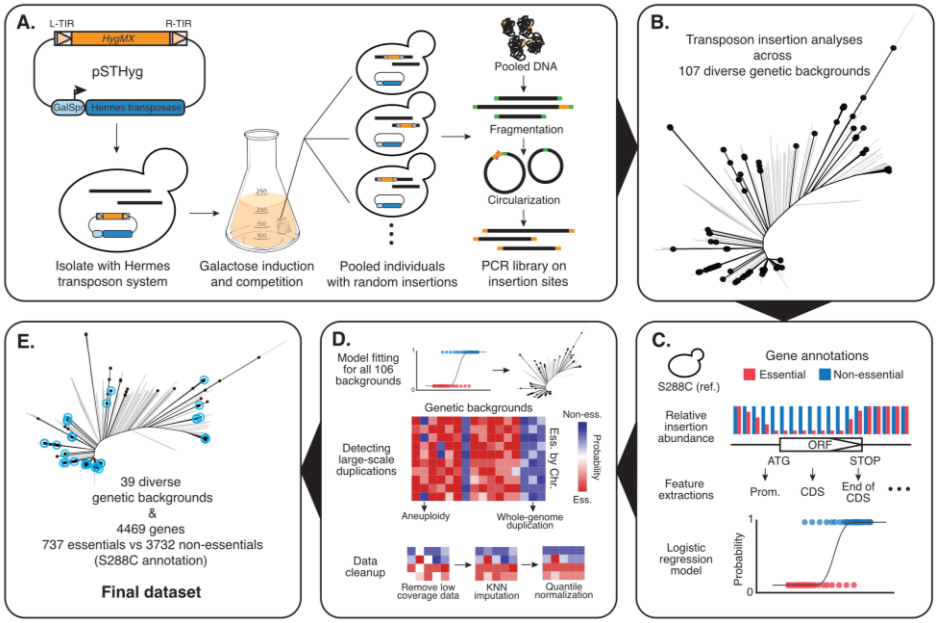


Figure 1. Summary of the *Hermes* transposon saturation procedure. (A.) A centromeric plasmid carrying the *Hermes* transposase and a transposon containing a hygromycin resistance marker (*HygMX*) is transformed into a haploid isolate background. Random transposon insertions are induced and selected. The mutant pool is then recovered and a PCR library that contains only the insertion sites is constructed and sequenced. (B.) Distribution of the selected 107 isolates across the species. The neighbor-joining tree was constructed using biallelic SNPs in the 1,011 yeast collection (Peter et al., 2018). Selected strains are highlighted in black. (C.) A logistic model was constructed using insertion profiles in the reference strain S288C. Gene essentiality annotations were used as a binary classifier, excluding those annotated as involved in galactose metabolism, respiration and slow growth. (D.) The logistic model was applied to insertion patterns in the remaining 106 isolates. Large-scale genome duplications were detected by looking at fitness predictions for all annotated essential genes along each chromosome. Low coverage regions were removed then imputed using k-nearest-neighbor method. The imputed fitness matrix was then quantile normalized. (E.) The final dataset after imputation consist of 39 isolates and 4,469 genes. Strains included in the final dataset are highlighted in blue.

Using insertion profiles and gene essentiality annotations in the reference S288C, we analyzed the average insertion patterns in the promoter (-500 bp to ATG, 100 bp window), the coding region (CDS), and terminator (STOP to +500 bp, 100 bp window) for all annotated essential vs. non-essential genes (Giaever et al., 2002) and found that insertion numbers drops from -100 bp prior to the CDS and extends to the entire CDS, with on average ~3 fold less insertions within the CDS in essential genes compared to non-essential genes (Figure S1C-E). This pattern is congruent with previous observations using the *Hermes* system (Gangadharan et al., 2010).

Modeling fitness using insertion patterns and machine learning

To leverage insertions in genes and surrounding regions, we constructed a logistic model that takes into account simultaneously insertions that occurred in the promoter, CDS, and terminator, using insertion profiles from the reference strain S288C and the corresponding gene essentiality annotations as a binary classifier (Figure 1C). After filtering steps to include only genes with a sufficient insertion coverage and without any known bias on the fitness effect due to our experimental conditions, i.e. galactose media, a total of 4,604 ORFs was included in the model, corresponding to 867 essential genes and 3,737 non-essential genes (Table S2). This model was statistically validated and applied to insertion profiles of all 107 diverse isolates (Figure 1D).

For each annotated ORF of *S. cerevisiae* genome (~6,300 in total), a probability was calculated based on the logistic model, ranging from a value of 1, corresponding to most likely to be non-essential, to 0, corresponding to most likely to be essential. As the transposon insertion saturation varies both between genomic regions and isolates, several cut-offs were applied to obtain a dataset that can be statistically comparable. Genomic regions with low coverage in non-discriminating regions for fitness prediction (Figure S2A-B) were first removed. Thus, only the most covered genetic backgrounds (n = 52) were conserved to include enough genes to compare (see Methods). Within these 52 isolates, the predicted fitness values unveiled non-essential complete genomes or entire chromosomes for a total of 13 and 5 isolates respectively (Figure 1D, Figure S3A). These signals suggested either whole-genome duplication or aneuploidies. Large-scale genome duplications including aneuploidies and endoreduplications are indeed frequently observed in experimental evolution in yeast (Harari et al., 2018; Johnson et al., 2021; Venkataram et al., 2016). Such events may hamper the accuracy of the modeled fitness effect in the context of transposon

insertion analysis, as genes within the duplicated region will all appear to be fit/non-essential due to insertions in only one of the two copies of the gene. The 13 strains with whole genome endoreduplication and the specific duplicated chromosomes in the 5 strains with aneuploidies were removed from the analysis.

In total, ~200 genes out of the 4,469 genes constituting the final set displayed divergent fitness predictions in the isolate S288C compared to the reference annotations (Giaever et al., 2002) (Table S3). More than 70% of annotated non-essential but predicted to be likely essential genes correspond to slow grow or galactose-specific fitness defect genes (Figure S3B). Conversely, among the 26 annotated essential genes predicted to be highly likely non-essential, with a predicted probability > 0.8, auxotroph-required genes were highlighted because the prototroph S288C isogenic strain was used in our study. Essential domains could also prevent from essential predictions using our model which consider the entire ORF and its adjacent regions (Figure S3C). Indeed, transposon saturation is powerful enough to detect essentiality at the domain-scale (Michel et al., 2017) in some cases such as background-specific essential genes between S288C and Σ1278b previously described (Dowell et al., 2010) and recaptured in our dataset (Figure S3D).

Overall, the predicted probability based on our logistic model can serve as a reasonable proxy for fitness variation for gene loss-of-function. Interestingly, domain-specific essentialities can be recaptured by the raw insertion patterns but not by the modeled fitness values (Figure S3C-D). However, as this effect is inherent to the transposon saturation system, it should not lead to differential fitness effect prediction in different strain backgrounds. The final dataset consists of 39 isolates from diverse origins and predicted fitness for 4,469 genes, which is further analyzed (Figure 1E, Table S3).

Environmental dependency of fitness variability across backgrounds

A hierarchical clustering based on the predicted fitness of 4,469 genes across the 39 different genetic backgrounds was performed (Figure 2). The profile similarity based on the predicted fitness effects was not correlated with the genetic diversity of the isolates (Figure 2). Genes that are consistently essential across different isolates clustered together and are enriched for essential biological processes including ribosome biogenesis, rRNA processing, DNA replication, protein transport and cell cycle (Figure 2).

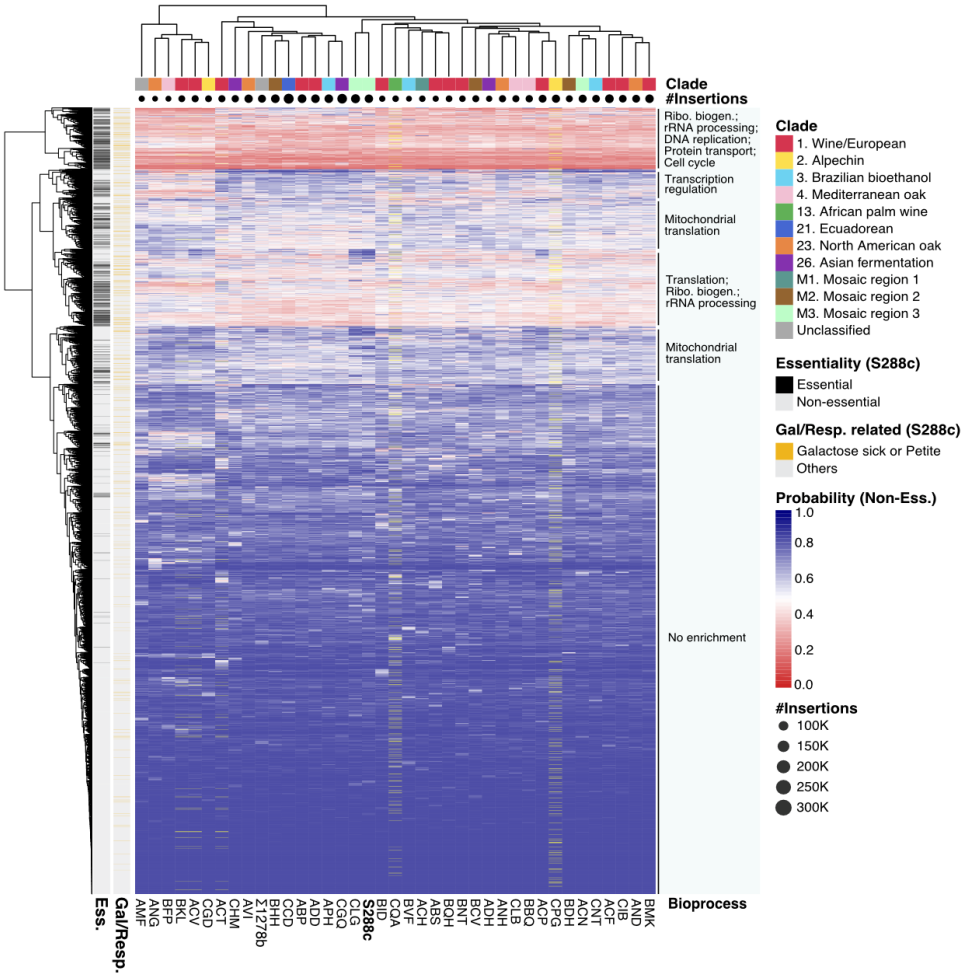


Figure 2. Hierarchical clustering of 4,469 fitness predictions across 39 genetic backgrounds. The distance matrix was calculated using the Euclidean distance method. The genetic origin of each isolate was color-coded, and the total insertion numbers per isolate was represented by dot size under the origin color code. Genes annotated as essential in the reference S288C are highlighted in black, and genes annotated as either galactose or respiration related are highlighted in yellow on the sidebars. Genes within duplicated chromosomes were removed (yellow bars on heatmap). Biological processes that are enriched in subclusters are annotated.

Genes that are consistently non-essential across backgrounds formed a large cluster with no significant enrichment for any specific biological process. Interestingly, several clusters of genes with variable fitness effects were identified, displaying modular switches from healthy to sick phenotypes across the entire population. Gene enrichment analyses highlighted genes involved in mitochondrial translation, transcription regulation and general translational processes (Figure 2). A large proportion of these genes with population-wide fitness variation consists of nuclear encoded mitochondrial genes involved in respiration, which were expected to show a selective disadvantage in our mutant pool. This observation could suggest that such general fitness variability may be environment-related rather than background-specific. However, other biological processes in addition to respiration/mitochondrial functions were also enriched, for which the impact of environment vs. genetic background on their fitness variability remains unclear.

For each gene, the predicted fitness in a given isolate was systematically compared with predictions in the reference strain S288C to identify the background-dependent fitness variation. In total, 632 unique genes were detected with a significant reverse prediction as being essential or non-essential compared to the reference. Among these genes, 458 show a loss-of-fitness (S288C healthy and background sick) and 174 a gain-of-fitness (S288C sick and background healthy) compared to the reference. The number of differential fitness genes identified in any single isolate ranges from 8 (ACP) to 88 (BQH), with a median of 61 for loss-of-fitness cases; and from 6 (CGD) to 42 (AMF), with a median of 16 for gain-of-fitness cases (Figure 3A). A total of 163 out of all 632 hits were related to respiration/mitochondrial function, representing ~20% to ~60% of loss-of-fitness hits depending on the genetic background (Figure 3A). Furthermore, these respiration-related genes tend to impact more backgrounds on average than non-respiration related hits (Figure 3B). These observations recaptured what was shown on the hierarchical clustering where mitochondrial related genes were highly enriched in clusters with modular fitness variation across multiple backgrounds (Figure 2). Again, due to the over-representation of these respiration-related genes and their continuous fitness variation in the population, we hypothesize that these hits are likely to be impacted by the environment (i.e. competition in galactose media) in addition to any specific genetic backgrounds.

To distinguish background-specific from environment-related cases previously hypothesized, the variation of fitness prediction in the population was estimated with

a z-statistics (Figure 3C, see Methods). Over the 632 unique hits, 179 background-specific hits were found and impact mainly a single genetic background (Table S4). Compared to the environment-related group, these background-specific cases are rare, with a median of 5 hits per isolate both loss- and gain-of-fitness types combined (Figure 3A). No significant enrichments for any biological processes or molecular functions were identified and especially respiration /mitochondrial-related genes are not over-represented (23/179 vs. 691/4,469 in the background, Fischer's exact test P-value = 0.82). By contrast, these respiration-related genes are significantly over-represented in the remaining group (140/453 vs. 691/4,469, Fischer's exact test P-value = 1.6e-10, odds ratio = 2). Each of these 453 environment-related hits impact on average 6 genetic backgrounds.

Environment-related fitness variation reveals potential functional rewiring

While an enrichment for respiration-related genes was detected in the environment-related cases, the majority of the genes composing this group is involved in other biological processes. Pairwise comparisons of fitness predicted values of the 453 environment-related genes across the 39 isolates highlighted two main subgroups of genes (Figure 4A). Based on the 292 environment-related hits significantly correlated or anti-correlated (see Methods), a network was constructed and confirmed the two main subnetworks that are correlated within the subgroup but are anti-correlated between subgroups (Figure 4B, Figure S4A). One subgroup contains mainly respiration-related genes, specifically genes involved in mitochondrial translation (Figure 4A, Figure S4A), which are anti-correlated with genes involved in transcription regulation and chromatin remodeling (*SPT7, SPT8, SWC4, SWC5, ARP6, ARP7, SIN3, RKR1, YAF9, UME1, NGG1, CHD1, STH1*, for example) as well as genes involved in nuclear-cytoplasmic protein transfer (*KAP120, KAP122, KAP123, NUP57, NUP100, NUP188, POM152, NIC96, MLP1*, for example) (Figure S4A). Many of these correlations were found between members of the same protein complexes. Several members of the transcription/nuclear transport subgroup are also annotated as respiration-related (deletion leads to absence of respiration) albeit not being directly involved in mitochondrial function, such as *SIN3*, a general chromatin remodeler and *KAP123*, a karyopherin responsible for nuclear import of ribosomal proteins. In addition to this large network, several small networks were also detected (Figure S4B-D), including *PMT1, PMT2* and *GET2*, which are involved in ER related glycosylation and are known to have physical interactions. The functional enrichments in the anti-correlated subgroups suggest a potential "rewire" between

mitochondrial translation and transcription regulation/nuclear transport, where modular switched of fitness effects associated with gene loss-of-function can occur in different strain backgrounds.

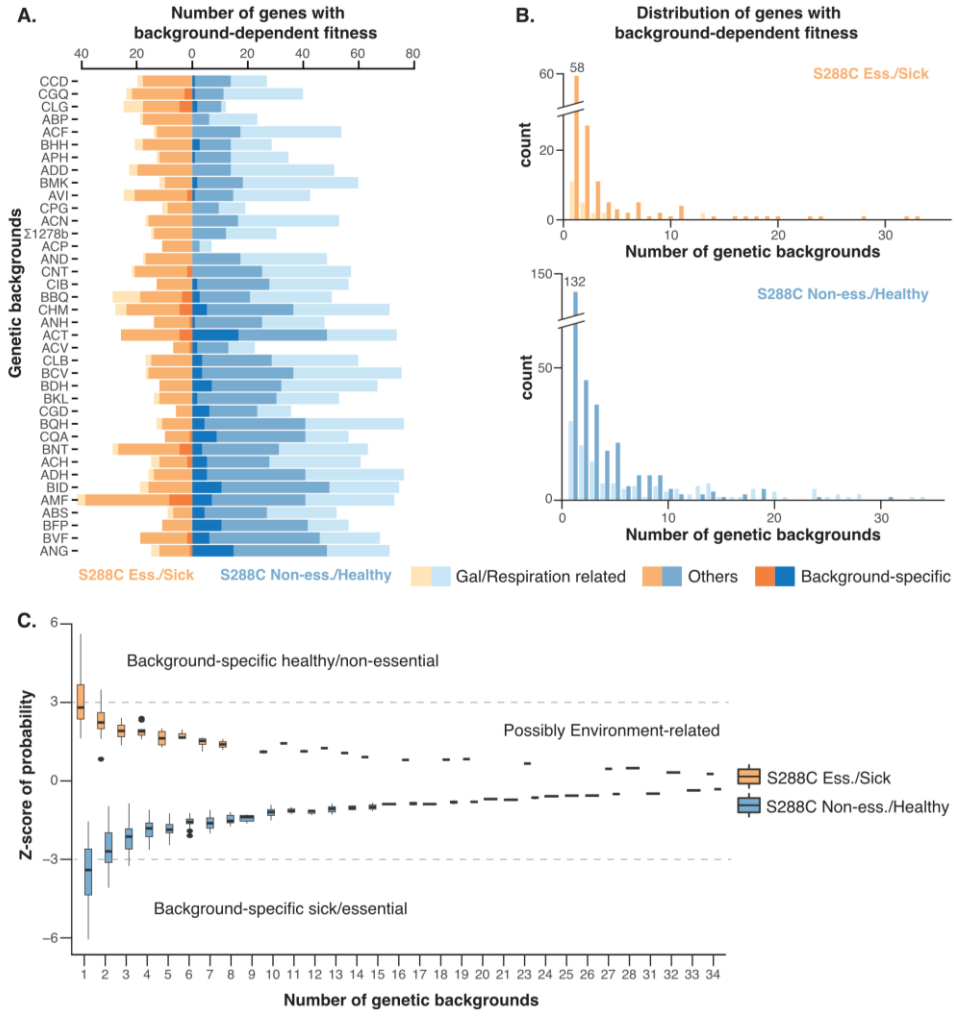


Figure 3. Number and distribution of background-dependent fitness variation genes. (A.) Number of hits detected in each genetic background. Genes annotated as galactose or respiration-related and genes that are background-specific are color-coded as indicated. Strains are sorted according to the total number of insertions. (B.) The number of genetic backgrounds impacted by the detected hits. Top panel, gain-of-fitness genes compared to S288C; bottom panel, loss-of-fitness genes compared to S288C. (C.) Z-statistic distribution for hits that impact different numbers of genetic backgrounds. A cut-off of |z-statistics| > 3 is indicated with dotted lines.

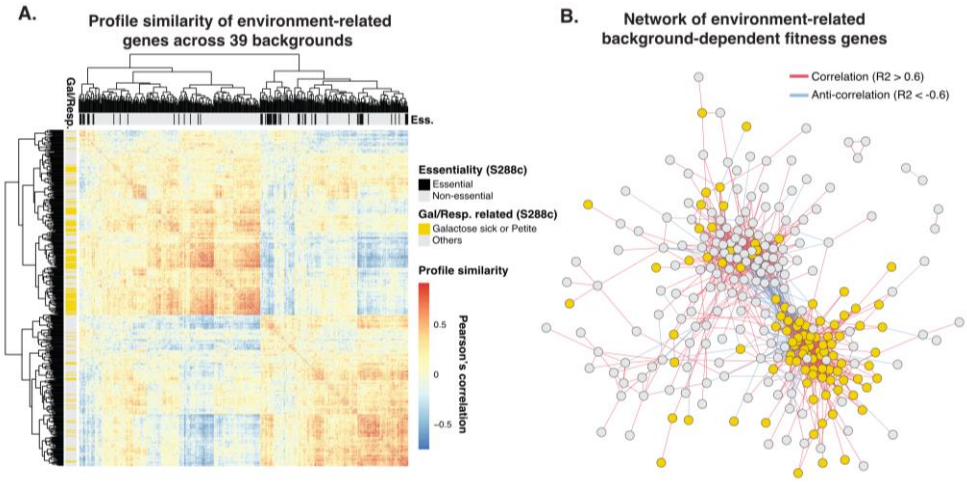
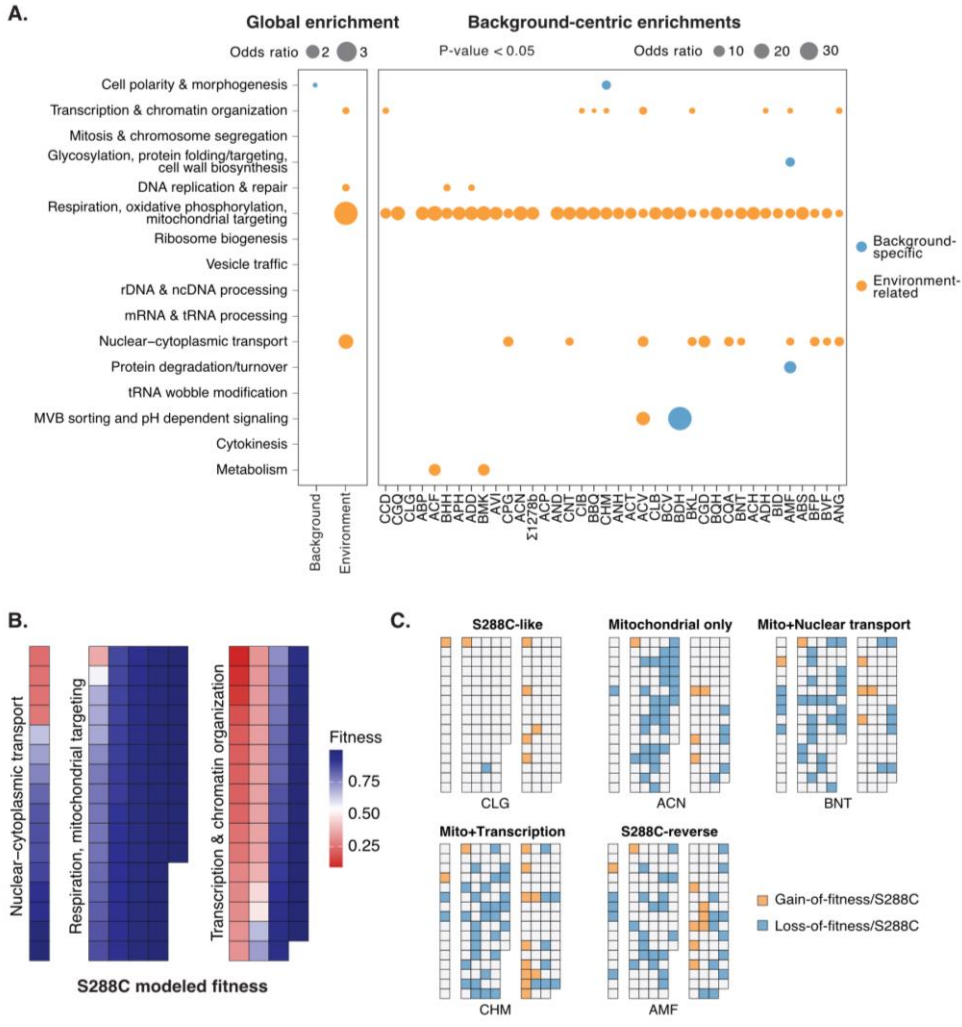


Figure 4. Correlation analyses for environment-related hits. (A.) Pairwise profile similarity based on predicted fitness across 39 backgrounds. Distance matrix was based on pairwise Pearson's correlation. Gene essentiality annotations are indicated on the upper sidebar and genes annotated as involved in galactose/respiration are indicated on the left sidebar. (B.) Network based on profile similarity among environment-related hits. Genes annotated as involved in galactose/respiration are colored in yellow. Positive correlation (> 0.6) are represented as red edges and negative correlation (< -0.6) are represented as blue edges. Complete network with annotated gene names can be found in Figure S4.

Functional insights into fitness variation genes

Based on functional annotations of the genes from our dataset (see Methods), enrichment analysis of fitness variation genes at the global level highlighted that background-specific hits are not enriched for most processes except for cell polarity (OR = 1.49, P-value = 0.026). Environment-related hits are enriched for Respiration/Mitochondrial functions (OR = 3.77, P-value = 4.16e-17), as well as Transcription & chromatin regulation (OR = 1.53, P-value = 0.002), Nuclear cytoplasmic transport (OR = 2.14, P-value = 0.004) and DNA repair (OR = 1.55, P-value = 0.01) (Figure 5A, Table S4).

Figure 5. Functional enrichments and rewiring for background-dependent fitness genes. (A.) Enrichments across 16 functional neighbourhoods defined by SAFE (Costanzo et al., 2016). Dot sizes represent odds ratios between the number of hits in a given neighbourhood vs. the total number of hits detected, with the size of the neighbourhood vs. the total number of genes in the dataset as background, using one-sided Fischer's exact test.



(Figure 5.) Global enrichment for background-specific (blue) and environment-related (orange) hits are presented on the left panel, and strain-centric enrichments are on the right panel. Enrichments with a p-value < 0.05 are shown. (B.) Predicted fitness for genes annotated in Respiration/mitochondrial targeting, Transcription and chromatin organization and Nuclear-cytoplasmic transport in the reference S288C. Genes in different processes are color-coded according to the predicted fitness probability, toward 1 (blue) for healthy/non-essential genes and 0 (red) for sick/essential genes. Detailed annotated version of this chart can be found in Figure S5A. (C.) Predicted fitness in other backgrounds compared to the reference S288C. Blue squares represent a switch from healthy to sick (loss-of-fitness) and orange squares represent a switch from sick to healthy (gain-of-fitness) for any given gene in a given background. All 38 isolates are shown in Figure S5B.

When looking at the same neighborhood enrichment at the strain level, environment-related hits are enriched for Mitochondrial functions in most genetic backgrounds except for ACP and CLG, the latter of which has a predicted fitness profile similar to the reference S288C (Figure 2). A large fraction of isolates showed significant enrichments for Transcription & chromatin regulation and Nuclear-cytoplasmic transport (Figure 5A). These enrichments are congruent with the rewiring hypothesis based on the profile similarity network analysis (Figure 4B). Indeed, when looking at genes annotated in these functional neighborhoods specifically, we observed various degrees of rewiring depending on the backgrounds (Figure 5B-C). In the reference S288C, loss-of-function for genes annotated in these three neighborhoods showed either high- or low-fitness predictions (Figure 5B, Figure S5A). Whereas in other genetic backgrounds, these predictions can be reversed as gain- or loss-of-fitness hits compared to S288C, with profiles that range from similar to S288C (i.e. CLG) to almost completely reversed (i.e. AMF) (Figure 5C). Such rewire could thus include either only mitochondrial-related genes, or with one or more processes related to either Transcription & chromatin regulation or Nuclear-cytoplasmic transport (Figure 5C). Depending on the genetic background, different sets of genes within the same functional neighborhood could be involved, highlighting the dynamics of such rewire (Figure S5B).

Compared to environment-related hits, a low number of background-specific cases were detected per isolate (Figure 3A) and thus tend to show little functional enrichment. However, in rare cases where multiple hits are detected in the same genetic background, some enrichments emerge (Figure 5A, Table S4). For example, in the strain BDH, 8 background-specific hits were detected with 3 annotated into one of the 16 functional neighborhoods, and two of which are involved in MVB sorting and pH-dependent signaling (*RIM8* and *RIM101*). Both genes are non-essential in S288C but predicted as loss-of-fitness in the BDH background (Figure 5A). In the strain AMF, 16 background-specific hits were detected with 11 annotated, among which 2 were involved in Protein degradation and turnover (*VID28* and *PRE3*) and 3 were involved in Glycosylation & cell wall biogenesis (*OST1, OPI3* and *FAB1*). These observations demonstrate that background-specific fitness variation genes, while rare, can be functionally coherent and involve multiple members of the same protein complex or biological process.

Finally, as previously posited (Hou et al., 2018), genes with background-dependent fitness variation tend to show an intermediate level of connectivity in terms of

genetic interactions (Figure 6A, Table S5) and an intermediate functional similarity between interacting gene pairs compared to genes that are consistently non-essential or essential (Figure 6B). Both environment-related and background-specific hits displayed an intermediate pattern. Interestingly, background-specific hits reveal higher non-synonymous to synonymous substitution rates (dN/dS) compared to both essential genes and non-essential genes (Figure 6C), indicating a potential positive selection or relaxed purifying selection on these genes at the population level. Overall, genes with background-dependent fitness variation are functionally coherent yet can be diverse within a single genetic background. Genes with environment-related fitness variation share general evolutionary features with background-specific cases.

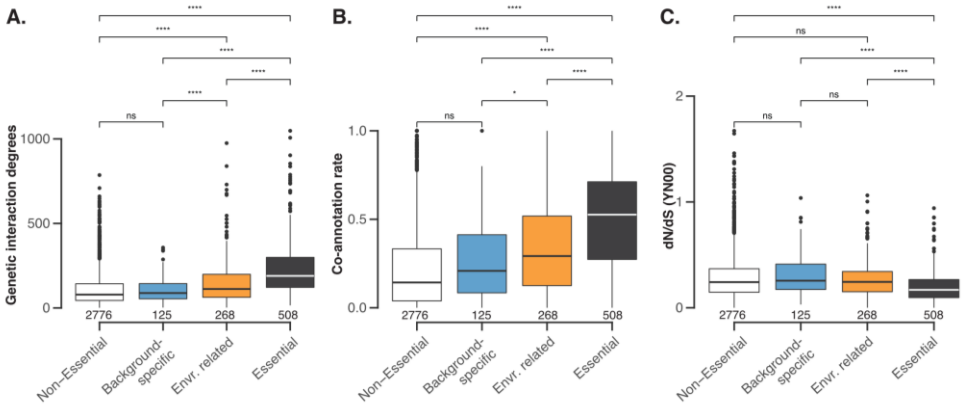


Figure 6. Evolutionary features associated with background-dependent fitness genes. (A.) Genetic interaction degrees derived from the yeast global genetic interaction network (Costanzo et al., 2016) for non-essential, background-specific, environment-related and essential gene categories. The number of genes annotated in each category are indicated. (B.) Functional co-annotation rates for different gene categories. The co-annotation rate corresponds to the fraction of interaction partners that are annotated in the same biological process as the primary gene (Costanzo et al., 2016). (C.) Mean non-synonymous vs. synonymous substitution rates (dN/dS) across 1,011 natural yeast isolates using the YN00 method (Peter et al., 2018). Comparisons between categories were performed using T-test, and significance levels are as indicated, with ns: p-value > 0.05, *: p-value < 0.05, **: p-value < 0.01, ***: p-value < 0.001 and ****: p-value < 0.0001.

Discussion

A large number of natural yeast isolates were surveyed for background-dependent fitness variations associated with gene loss-of-functions using a transposon saturation strategy. We modeled fitness by considering transposon insertion densities within gene coding sequence and surrounding regions. Comparison of the modeled fitness between different isolates and the reference S288C allowed for the identification of 632 genes displaying background-dependent phenotypes. The majority of these cases (453/632) showed continuous fitness variation across population and is at least partly related to the environment. By contrast, background-specific cases tend to be rare, with on average 5 hits per isolate. At the individual level, both environment-related and background-specific variable fitness genes are functionally coherent, meaning that genes involved in the same biological process tend to show similar fitness variation in the same genetic background.

A large fraction of the background-dependent fitness variation genes was classified as environment-related based on two factors. First, genes with background-dependent fitness variation were highly enriched for respiration/mitochondrial functions, which are expected to show a fitness loss under prolonged growth in media with galactose as the unique carbon source. Indeed, mitochondrial-related genes were also found to be background-dependent in a previous study involving 4 different isolates on conditions with non-fermentable carbon sources (Galardini et al., 2019). Second, most background-dependent fitness genes showed a continuous variation across the population. Further analyses showed that genes in two biological processes, namely Transcription & chromatin remodeling and Nuclear-cytoplasmic transport, are anticorrelated with genes involved in Mitochondrial translation in terms of their fitness profiles. These anticorrelations indicate a modular change in the relative fitness of genes involved in these processes compared to the reference strain S288C. However, whether such rewiring effect is exclusively related to respiration conditions or could represent a general background-dependency effect remains difficult to disentangle due to the experimental conditions necessary for transposon saturation analyses.

There are several indications that suggest the rewiring effect could have implications beyond a specific experimental condition. In a recent large-scale analysis of environment-dependent genetic interactions, it was shown that most interactions specific to an environmental condition are in fact part of the global genetic

interaction network that were exacerbated or attenuated in the tested condition (Costanzo et al., 2021). Comparing to genetic interactions between pairs of gene deletion mutants, background-dependent gene loss-of-function phenotype could be considered as interactions between the loss-of-function gene and modifier variants specific to the background, which should share general properties to genetic interactions with deletion mutants. Indeed, we tested the gene deletion phenotype for one of the environment-related hits involved in transcription & chromatin remodeling, *BMH1* (Figure S5C). This gene was identified as loss-of-fitness in multiple genetic backgrounds compared to S288C in our data. Interestingly, the loss-of-fitness phenotype was indeed confirmed on standard rich media YPD, suggesting the environment-related fitness variation genes could have a general effect independent of the experimental conditions. In addition, genes involved in chromatin remodeling were also found to display background-dependent fitness effects in a previous study comparing S288C and a natural isolate 3S (Mullis et al., 2018). Moreover, the environment-related hits shared the same evolutionary features compared to background-specific ones, suggesting the environmental effect may only exacerbate the observed functional rewiring in different genetic backgrounds.

Compared to environment-related hits with general functional rewiring effect across different individuals, background-specific cases are mostly unique to different genetic backgrounds. These cases are rare but can be coherent, involving members of the same protein complex or bioprocess. While we screened the largest number of diverse genetic backgrounds in yeast to date, it is still difficult to have an accurate estimation of the frequency of such cases as well as the landscape of the underlying genetic modifiers.

While transposon saturation strategy can be versatile to genetic diversity among isolates, this method also presents some limitations. Among all the isolates initially tested, only around half showed a reasonable level of insertion efficiency, highlighting the unexpected variability of transposon activity across different individuals. This variability results in an underestimate of the number of genes with background-dependent phenotypes. In addition, loss-of-function phenotypes that are related to specific protein domains but not the entire ORF are difficult to identify, unless the insertion efficiency is extremely high. The *Hermes* system, as is also the case for all transposon saturation systems currently available in yeast, requires selection of transposon insertion events in the presence of galactose induction (Van Opijnen and Levin, 2020). This competition effect in a non-fermentable carbon

source may complicate downstream analysis as the effects of environment vs. genetic background can be difficult to unravel. New strategies that takes into account these factors are still needed in order to get a more precise view of background-dependent gene loss-of-function phenotypes at the species level.

Methods

Strains and growth conditions

A total of 106 isolates were selected from the 1,011 *Saccharomyces cerevisiae* collection (Peter et al., 2018). A prototrophic haploid strain FY5, isogenic to the reference strain S288C was also included (Table S1). Haploid segregants derived from the 106 natural isolates were obtained after *HO* deletion and tetrad dissection (Fournier et al., 2019; Hou et al., 2016). Detailed descriptions of the strains can be found in Table S1. Strains were maintained at 30°C using YPD (1% Yeast extract; 2% Peptone, 2% Dextrose) in liquid culture or solid plates (2% of agar). Transposon activity was induced in YPGal (1% Yeast extract; 2% Peptone, 2% Galactose) with Hygromycin B (200 µg/mL). Sporulation was induced on solid plates containing 1% of potassium acetate and 2% of agar.

Ploidy control

Ploidy was estimated by flow cytometry. Cells in exponential growth phase were washed in water, then 70% ethanol and sodium-citrate buffer (50 mM, pH 7.5) followed by RNase A treatment (500 µg/mL). To avoid cell aggregates, each sample was sonicated then the DNA was labelled with propidium iodide (16 µg/mL), a fluorescent intercalating agent. DNA content was then quantified using the 488 nm excitation laser of the Accuri C6 plus flow cytometer (BD Biosciences).

Cell transformation

Cells in exponential growth phase were chemically transformed using the EZ-Yeast Transformation Kit (MP biomedicals). We incubated cells 30 minutes at 42°C with EZ-Transformation solution, carrier DNA and either 100 ng of pSTHyg plasmid or 1 µg of PCR fragment. After regeneration in YPD, cells were spread on solid YPD plate supplemented with Hygromycin B and incubated at 30°C until transformants appeared.

Construction of the pSTHyg plasmid

In order to be compatible with our isolates already carrying either a nourseothricin or a kanamycin resistance cassette, the nourseothricin cassette of the pSG36 plasmid (Gangadharan et al., 2010) was replaced by a hygromycin B resistance cassette. The

pSG36 plasmid was amplified in 2 fragments by PCR excluding the *NatMX* cassette, then assembled with the *HphMX* cassette amplified from p41 plasmid (Addgene #58547) with overlapping regions using Gibson assembly. The new plasmid, pSTHyg was amplified in *E. coli* and extracted using the GeneJET Plasmid Miniprep Kit (Thermo Scientific™). The construction was verified using enzymatic digestion with *KpnI* and *PvuI*.

Generation of transposon insertion mutant pools

Each natural isolate was grown in liquid YPD medium and chemically transformed with 100 ng of pSTHyg plasmid as described. From the selective transformation plates, a single clone was picked and grown in 30 ml of YPD supplemented in hygromycin B under agitation at 30°C until saturation (~ 24h). Cells were then diluted at an OD of 0.05 in 50 ml of YPGal supplemented with hygromycin B to activate the transposase and induce the transposition for 72h at 30°C. Two successive dilutions were then performed for 24h at an OD of 0.5 in 100 ml of YPD then YPD supplemented with hygromycin B to enrich for cells the transposon in their genome. The final 100ml culture was centrifuged, water-washed and 500 µL aliquots of cells were frozen at -20°C.

Sequencing library preparation

In order to sequence the genomic regions with a transposon insertion, the genomic DNA of the pool of cells carrying insertion events was extracted using the MasterPure™ Yeast DNA Purification Kit (Lucigen). Cells were lysed using a lysis solution supplemented in zymolyase 20T (1.5 mg/ml). Proteins and cellular debris were removed with the MPC Protein Precipitation Reagent and several RNase A treatments were realized to eliminate RNA. gDNA was then precipitate with ethanol. The pellet was washed twice with 70% ethanol and resuspended in 80 µl of water. The gDNA sample integrity was controlled on 1% agarose gel and quantified on Nanodrop and Qubit using the Qubit™ dsDNA BR Assay Kit (Invitrogen™). 2 x 2 µg of gDNA were digested in parallel with 50 units of *DpnII* (NEB #R0543L) and *NlaIII* (NEB #R0125L) in 50 µl for 16h at 37°C. The enzymatic reactions were inactivated for 20 min at 65°C and DNA fragments were ligated with 25 Weiss units of T4 Ligase (Thermo Scientific #EL0011) in a total volume of 400µLlfor 6h at 22°C. Circular DNA were then precipitated overnight at -20°C with ethanol, salt (NaOAc 3M pH5.2) and glycogen. After an 70% ethanol wash, the DNA pellet was

resuspended in 50µL of water. The junction between the genomic region and the transposon insertion site was amplified on both *DpnII* and *NlaIII* digested and re-circularized gDNA by PCR using outward-facing primers targeting the transposon. The PCR products were controlled on 1% agarose gel and displayed variable sizes centred around 750 bp. Nanodrop and Qubit using the Qubit™ dsDNA BR Assay Kit (Invitrogen™) quantifications were then performed to pool the same amount of *NlaIII*-digested and *DpnII*-digested PCR products. For each sample, at least 6 µg at minimum 30 ng/µl was then sent to the BGI (Beijing Genomics Institute) for sequencing. In total, each sequencing run provided 1 Gb of 100 bp paired-end reads using Illumina Hi-Seq 4000 or DNBseq technologies.

Determination of transposon insertion sites

The reads that contained the amplified part of the transposon were selected and the corresponding 57 bp sequence was trimmed with Cutadapt (Martin, 2011) and the reads corresponding to the plasmid were discarded. The cleaned reads were mapped to the S288C reference genome with the corresponding SNPs inferred for each isolate (Peter et al., 2018) with BWA (Li and Durbin, 2009). The genomic position of an insertion site was defined as the first base pair aligned on the genome after the transposon region. For each insertion site, the number of reads and their orientation were obtained.

Modelling the fitness effect of gene loss-of-function based on transposon insertion profiles

- **Model Construction**

The number of insertions in the promoter region (-100bp to ATG), beginning of the coding region (-100 to +100 from ATG), the coding region, end of the coding region (-100bp to +100bp from stop-codon) were normalized as insertion densities per 100bp. Gene essentiality annotations were obtained from SGD (phenotype “inviable”) exclusively for annotations with gene deletion in the S288C background. Respiration related gene annotations were obtained from SGD with the phenotype “respiration: absent” after gene deletion in S288C. Galactose-specific loss-of-fitness was determined in ref, with a stringent cut-off of < -0.2. A logistic model was constructed using the glm() function from the R package “stats”, using insertion densities in the reference strain S288C, in the promoter region (-100bp to ATG), beginning of the coding region (-100 to +100 from ATG), the coding region, end of

the coding region (-100bp to +100bp from stop-codon), raw insertion number in the coding region and gene sizes as predictors, and essentiality annotations as a binary classifier. To construct the model, we removed genes that are known to be non-essential but display a slow growth phenotype (Giaever et al., 2002), genes with differential fitness defect in galactose media (Costanzo et al., 2021), as well as genes showing respiration defects when deleted i.e. petite mutants, as these genes are likely to show a competitive disadvantage in the context of our experimental condition (Table S2). Genes that are localized in regions with low insertion densities, *i.e.* less than 3 insertions in the terminator region (STOP to +300 bp) and less than 50 insertions in a 10 kb region surrounding the gene (-5 kb before ATG and +5 kb after STOP) were also excluded. A total of 4,604 genes were included in the model (Table S2). 10-fold cross-validation was performed using the R package "caret", with trainControl() and train() functions, method = "glm", family = "binomial". Cross-validation results showed that the model has an average accuracy of 0.88 with a Kappa 0.57 (Table S2), which is fairly accurate considering the class imbalance in the set. The predictive value for non-essential labels is 0.91, contrasting to a lower predictive value of 0.70 for essential labels, indicating a better accuracy in predicting non-essential genes using this model.

- **Predictions based on the logistic model**

The logistic model estimated a probability for each annotated ORF (~6,300), ranging from a value of 1, corresponding to most likely non-essential, to 0, corresponding to most likely essential. Genomic regions with low insertion densities were removed from the analysis (Figure S2A-B). As the number of isolates that can be included in the analysis is directly correlated to the number of genes without low insertion densities in each background (Figure S6A), we used a k-nearest-neighbors algorithm to impute missing probability values after removing low insertion density regions. Imputations for missing values were performed using the function impute.knn() in the R package "impute", with k = 10, rowmax = 50% and colmax = 80%. Quantile normalization of the imputed matrix was performed using normalize.quantiles() function in the R package "preprocessCore". We calculated the number of genes that remained after imputation as a function of an initial cut-off of the number of genes without low insertion densities (Figure S6B). We chose a cut-off that leverages this trade-off between the number of backgrounds and the number of interpretable genes, leading to a final dataset of 4,469 genes in 52 isolates. All fitness prediction data can be found in Table S3.

- **Endoreduplication and aneuploidy detection using probability values**

High predicted probability values (tend to a value of 1, i.e. non-essential) for all essential genes were markers for endoreduplication of the genome (all chromosomes are duplicated). 7 out of 52 strains were detected as endoreduplicated and 6 others showed an intermediate to high probability prediction but was not high enough to be confidently classified as non-essential. These 6 strains were subsequently confirmed as a mixture of haploid and diploid cells using flow cytometry. we also detected aneuploidy of chromosome I for 3 strains (ACT, BKL and ACV), one strain with an aneuploidy of chromosome XII (CPG) and one strain with an aneuploidy of chromosome XIV (CQA). These aneuploidies were not present in the original isolate except for the chromosome I aneuploidies in ACT and ACV, highlighting the dynamics of genome instability in different genetic backgrounds. The 13 strains with whole genome endoreduplication and the specific duplicated chromosomes in the 5 strains with aneuploidies were removed from the analysis.

- **Estimation of the differential fitness score between each isolate and S288C**

From the final set of 4,469 genes in 39 isolates, a differential fitness score for each gene in each background was calculated by subtracting the predicted fitness value in a given strain by the corresponding fitness prediction in the reference S288C. A minimum of absolute value of the differential fitness score of 0.5 was considered significant, which corresponds to a bona fide reverse in the direction of being predicted as essential or non-essential according to our logistic model. In total, 632 unique genes were identified with marked fitness variation.

- **Distinction between background-specific and environment-related cases**

We calculated the z-statistics for all variable fitness hits to distinguish those that are background-specific from the ones that are possibly related to the environment (Table S4). Environment-related cases are more likely to vary continuously in the population with a low z-statistics. Cases that are truly specific to certain genetic backgrounds are detected with a high z-statistic score set at $|z| > 3$ (Figure 3C).

- **Correlation in environment-related cases**

Distance matrix was based on pairwise Pearson's correlation between predicted fitness values across the 39 strain backgrounds of the 453 environment-related cases. The network was based on the profile similarities where the edges correspond to a Pearson's correlation > 0.6 (correlation) or < -0.6 (anti-correlation).

- **Gene annotations in functional subgroups and enrichment**

We annotated genes in our dataset into 16 functional neighborhoods according to SAFE (Spatial Analysis of Functional Enrichment) (Costanzo et al., 2016) and looked for enrichment in different neighborhoods. For each neighborhood, we calculated the odds ratio of enrichment based on the number of hits annotated in the neighborhood vs. the total number of hits, with the size of the neighborhood and total number of genes as background (one-sided Fisher's exact test).

Validation of the phenotypic consequence of *BMH1* gene loss-of-function

Stable haploid isolates, FY5 and CIB were diploidized using the pHS2 plasmid (Addgene #81037) containing the *HO* gene encoding the endonuclease responsible for mating type switching and a hygromycin resistance cassette. The *BMH1* gene was replaced with a Hygromycin B resistance cassette in the diploid isolates. Sporulation was induced on AcK medium in diploid isolates heterozygous for *BMH1* gene deletion. Around 20 resulting tetrads were then dissected on YPD using a MSM 400 micromanipulator (Singer Instrument). Each spore grew for 48h at 30°C and the colony size was captured with the camera of the colony picker, PIXL (Singer Instrument). Colony size measurements were then analyzed using custom R scripts.

Data availability

All sequencing data related to this study were deposited to the European Nucleotide Archive (ENA) under the accession number PRJEB45777.

Supplementary material

Supplementary tables

Supplementary tables are available at:
https://www.dropbox.com/sh/61pgn2tvbo2ccdt/AADuaghBECbXSmzYWhOitSeZa?dl=0

TableS1. Description of isolates used in this study.

TableS2. Model construction and evaluations. This table contains 4 tables:

GenesInModel: 4,604 ORFs and their essentiality annotations used to construct the logistic model. Insertion numbers and densities within coding sequence and surrounding regions are included. Insertion numbers calculated from S288C insertion profile.

ModelSummary: Features included in the logistic model and their coefficient.

CrossValidation: Summary of the cross-validation results.

CMStat: Confusion matrix, prediction accuracy and precision for essential/non-essential labels.

TableS3. Raw and final dataset with predicted fitness. This table contains 3 tables:

Raw_data_pred: All raw predicted fitness based on the logistic model for 107 isolates.

Pred_final_39: Predicted fitness for 39 isolates included in the final dataset. Raw, imputed and quantile normalized predictions are shown.

Score_final_39: Differential fitness score by comparing the predicted fitness in a given isolate to S288C.

TableS4. Background-dependent fitness variation genes identified in this study. This table contains 4 tables:

Z-statistics: Z-statistics for each of the 632 hits, including the number of genetic backgrounds impacted for each hit.

Hits_SAFE_annotation: Annotations for each hit into the 16 functional neighborhoods according to SAFE (Costanzo et al., 2016).

Enrichment_global: Enrichment for all hits across 16 functional neighborhoods

Enrichment_Strain: Enrichment for hits in a given genetic background across 16 functional neighborhoods.

TableS5. Genetic interaction degree and dN/dS values and gene classifications.

Supplementary figures

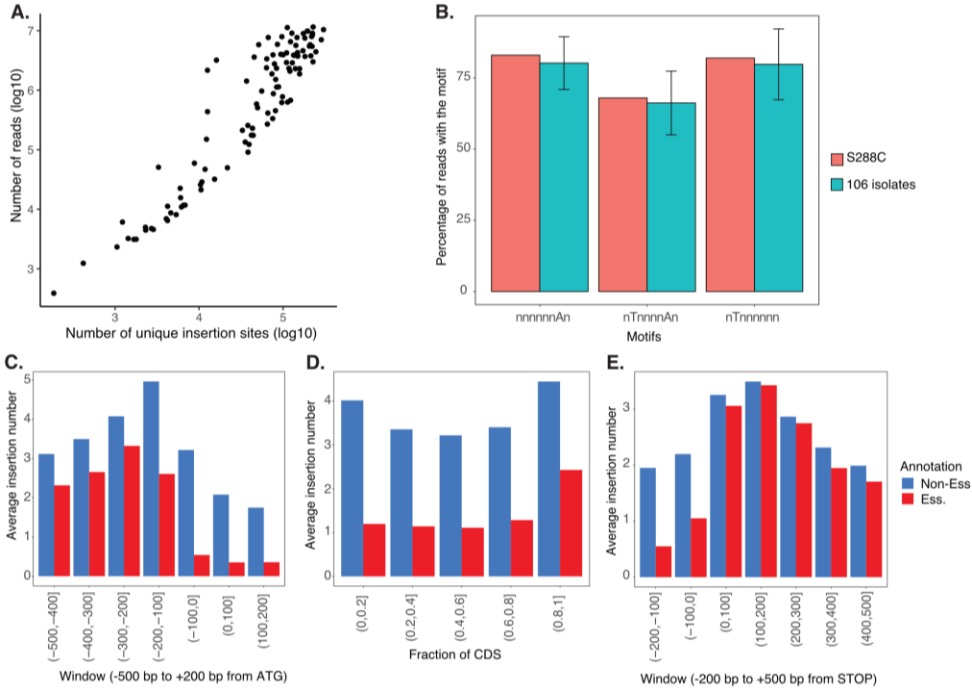


Figure S1. (A.) Number of reads (y-axis, log10 scale) vs. number of unique insertion sites (x-axis, log10 scale) across 107 diverse isolates. (B.) Insertion preference comparison between the reference S288C and the other 106 selected isolates. Sequence motifs (Gangadharan et al., 2010) are on the x-axis and the percentage of reads with a given motif are presented as color coded bars. Error-bars correspond to the standard deviation across different isolates. (C.) Insertion density comparison between essential and non-essential genes in S288C in the promoter region. Average insertion numbers in the -500bp to +200bp region relative to ATG are shown in 100bp windows. (D.) Insertion density comparison between essential and non-essential genes in S288C in the coding region (CDS). Average insertion numbers in the relative fractions of a given CDS are shown. (E.) Insertion density comparison between essential and non-essential genes in S288C in the terminator region. Average insertion numbers in the -200bp to +500bp region relative to the stop-codon are shown in 100bp windows.

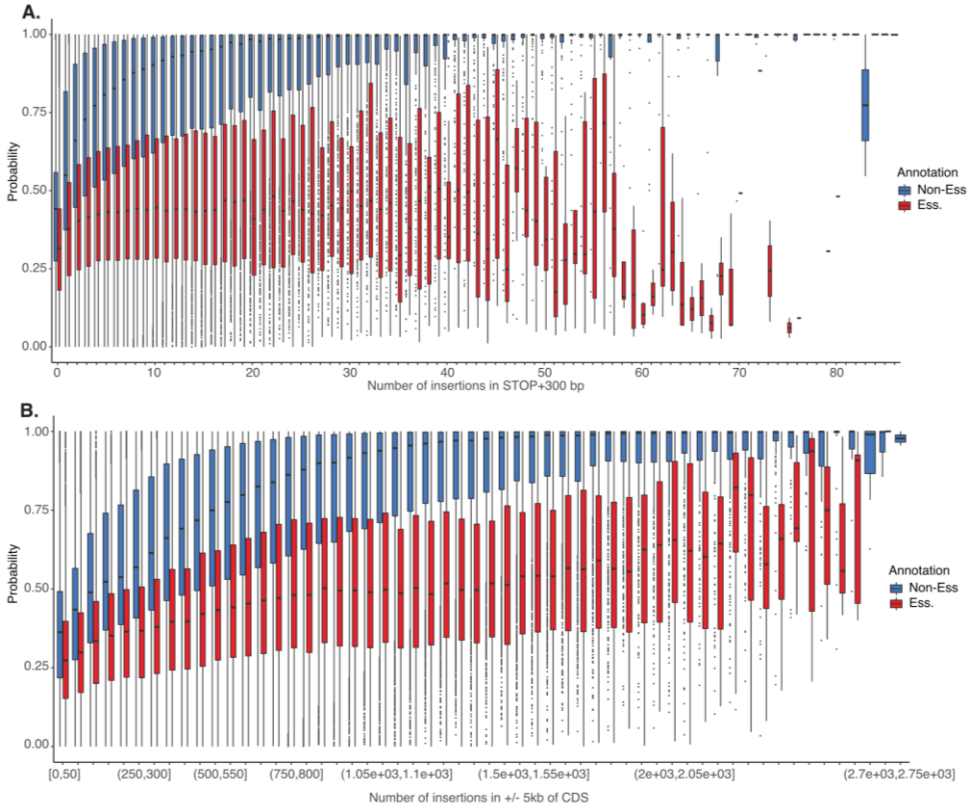


Figure S2. (A.) Predicted non-essential probabilities (y-axis) as a function of the number of insertions in the terminator region (300bp after stop-codon). Non-essential genes are in blue and essential genes in red. (B.) Predicted non-essential probabilities (y-axis) as a function of the number of insertions in a 10kb region surrounding the CDS (5kb before ATG and 5kb after stop-codon). Non-essential genes are in blue and essential genes in red.

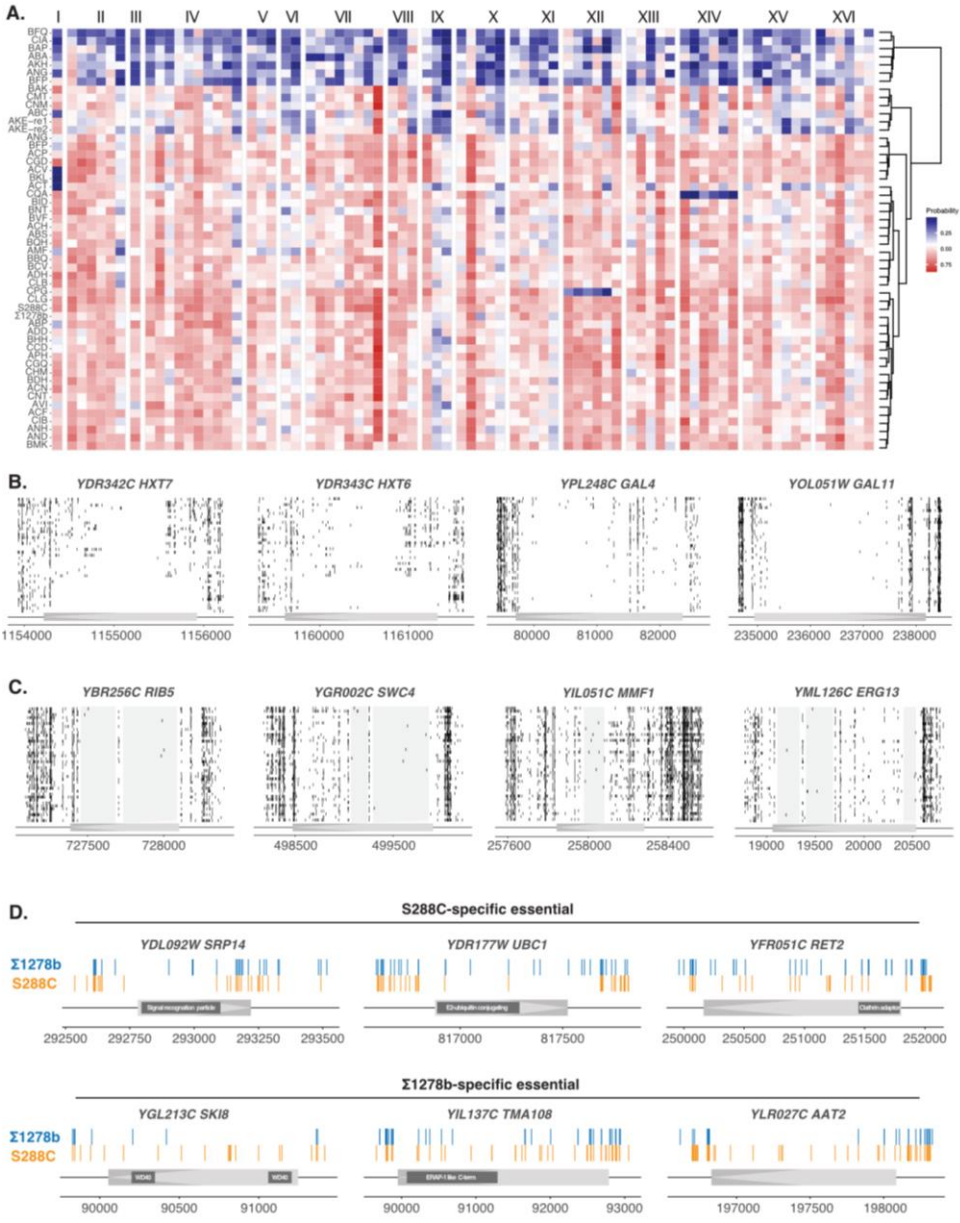

A.
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
Probability
B.
YDR342C HXT7
YDR343C HXT6
YPL248C GAL4
YOL051W GAL11
1154000
1155000
1156000
1160000
1161000
80000
81000
82000
235000
236000
237000
238000
C.
YBR256C RIB5
YGR002C SWC4
YIL051C MMF1
YML126C ERG13
727500
728000
498500
499500
257600
258000
258400
19000
19500
20000
20500
D.
S288C-specific essential
YDL092W SRP14
YDR177W UBC1
YFR051C RET2
Σ1278b
S288C
292500
292750
293000
293250
293500
817000
817500
250000
250500
251000
251500
252000
Σ1278b-specific essential
YGL213C SKI8
YIL137C TMA108
YLR027C AAT2
Σ1278b
S288C
90000
90500
91000
90000
91000
92000
93000
197000
197500
198000

Figure S3. (A.) Average non-essential probability or predicted fitness for every 10 successive essential genes along all 16 chromosomes for 52 strains that passed the coverage cut-offs. Strain-side clustering was based on predicted fitness for all genes. (B.) Insertion profiles for gene related to galactose metabolism that are annotated as non-essential in S288C but detected as essential/sick in all or a fraction of the 39 strains in the final dataset. Chromosomal positions and gene orientations are schematically presented on the x-axis and insertion profiles for each strain are presented as black vertical bars. (C.) Insertion profiles for essential genes predicted as non-essential in S288C. Shaded areas correspond to potential essential protein domains. (D.) Insertion profiles for genes previously shown background-specific essentiality between S288C and Σ1278b. Domain-specific essentiality regions are indicated.

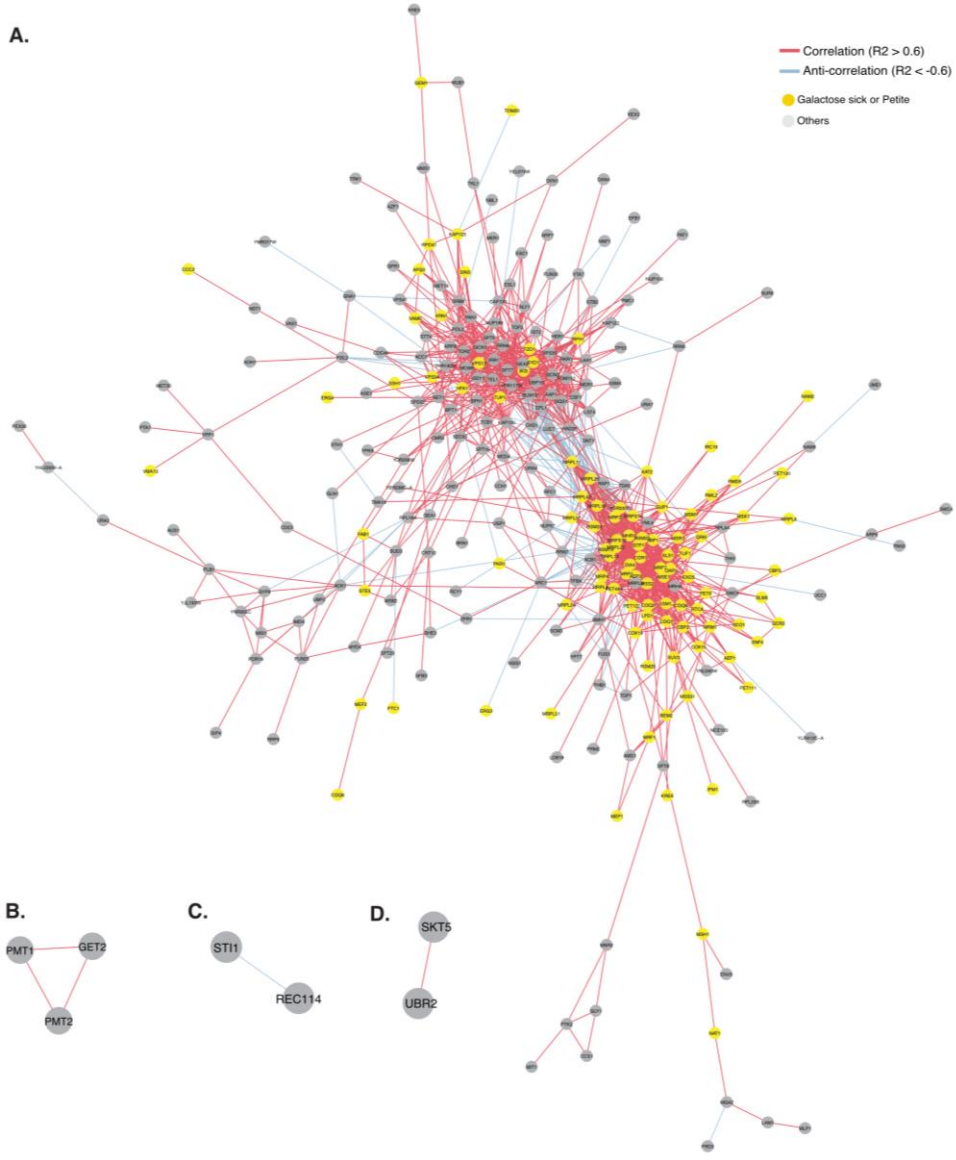


Figure S4. (A.) Annotated network based on profile similarity as shown in Figure 4B. (B.-D.) Subnetworks with significant correlations independent from the large subnetwork involving respiration-related hits.

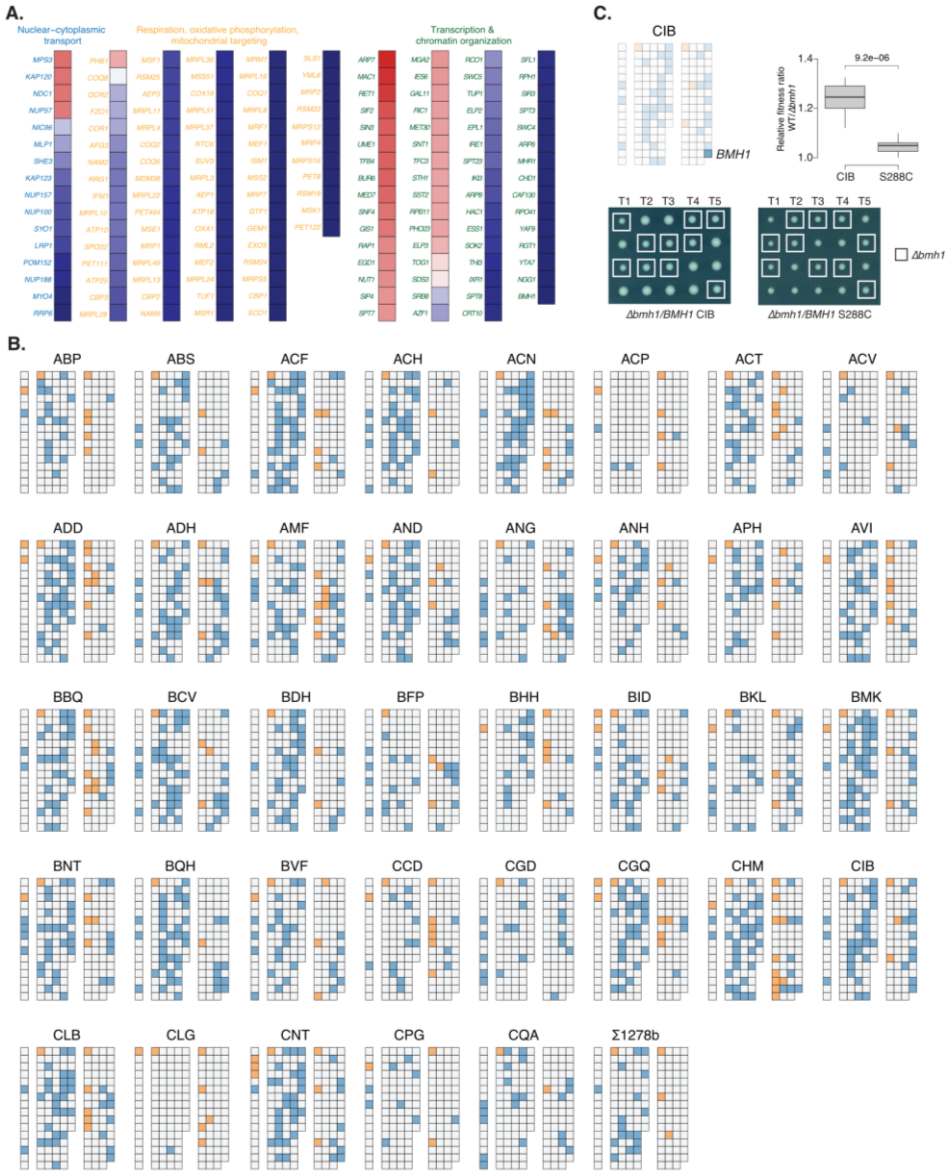


Figure S5. (A.) Predicted fitness for genes annotated in Respiration/mitochondrial targeting, Transcription and chromatin organization and Nuclear-cytoplasmic transport in the reference S288C with gene name annotations. Related to Figure 5B. (B.) Charts for all 38 isolates relative to the reference S288C. Related to Figure 5C. (C.) Example of functional rewire in a natural isolate CIB compared to the reference S288C for a transcription-related hit *BMH1*. Relative fitness ratio (colony size for WT divided by deletion of *BMH1*) is shown on the upper right panel. Colony sizes of *BMH1* deletion vs. WT were measured using tetrad dissection of hemizygous diploids. 5 tetrads are shown for each background.

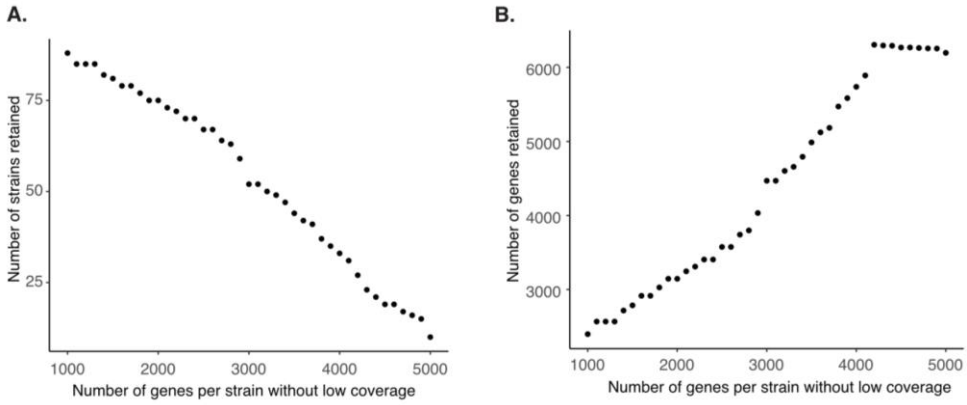


Figure S6. (A.) The number of strains retained as a function of cut-offs of the number of interpretable genes after removing low coverage regions (less than 50 insertions in the surrounding 10kb region and/or less than 3 insertions in the 300bp terminator region). (B.) Number of genes retained after imputation as a function of cut-offs of the number of interpretable genes after removing low coverage regions.

References

Blomen, V.A., Májek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., et al. (2015). Gene essentiality and synthetic lethality in haploid human cells. Science (80-.). *350*, 1092 LP – 1096.

Boutros, M., Kiger, A.A., Armknecht, S., Kerr, K., Hild, M., Koch, B., Haas, S.A., Paro, R., and Perrimon, N. (2004). Genome-wide RNAi analysis of growth and viability in Drosophila cells. Science *303*, 832–835.

Chandler, C.H., Chari, S., Tack, D., and Dworkin, I. (2014). Causes and consequences of genetic background effects illuminated by integrative genomic analysis. Genetics *196*, 1321–1336.

Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., Zhou, H., Tian, L., Prakash, O., Lemire, M., et al. (2016). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat. Biotechnol. *34*, 531–538.

Chow, C.Y. (2016). Bringing genetic background into focus. Nat. Rev. Genet. *17*, 63–64.

Chow, C.Y., Kelsey, K.J.P., Wolfner, M.F., and Clark, A.G. (2016). Candidate genetic modifiers of retinitis pigmentosa identified by exploiting natural variation in Drosophila. Hum. Mol. Genet. *25*, 651–659.

Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C., and Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum. Genet. *132*, 1077–1130.

Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*.

Costanzo, M., Hou, J., Messier, V., Nelson, J., Rahman, M., VanderSluis, B., Wang, W., Pons, C., Ross, C., Ušaj, M., et al. (2021). Environmental robustness of the global yeast genetic interaction network. Science *372*.

Cutting, G.R. (2010). Modifier genes in Mendelian disorders: the example of cystic fibrosis. Ann. N. Y. Acad. Sci. *1214*, 57–69.

Dorfman, R. (2012). Modifier gene studies to identify new therapeutic targets in cystic fibrosis. Curr. Pharm. Des. *18*, 674–682.

Dowell, R.D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D.A., Alexander Rolfe, P., Heisler, L.E., Chin, B., et al. (2010). Genotype to phenotype: A Complex problem. Science (80-.). *328*, 469.

Edwards, M.D., Symbor-Nagrabska, A., Dollard, L., Gifford, D.K., and Fink, G.R. (2014). Interactions between chromosomal and nonchromosomal elements reveal missing heritability. Proc. Natl. Acad. Sci. U. S. A. *111*, 7719–7722.

Fournier, T., and Schacherer, J. (2017). Genetic backgrounds and hidden trait complexity in natural populations. Curr. Opin. Genet. Dev. *47*, 48–53.

Fournier, T., Saada, O.A., Hou, J., Peter, J., Caudal, E., and Schacherer, J. (2019). Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. Elife *8*, 1–18.

Galardini, M., Busby, B.P., Vieitez, C., Dunham, A.S., Typas, A., and Beltrao, P. (2019). The impact of the genetic background on gene deletion phenotypes in Saccharomyces cerevisiae . Mol. Syst. Biol. *15*, 1–13.

Gangadharan, S., Mularoni, L., Fain-Thornton, J., Wheelan, S.J., and Craig, N.L. (2010). DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. Proc. Natl. Acad. Sci. *107*, 21966–21972.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387–391.

Harari, Y., Ram, Y., Rappoport, N., Hadany, L., and Kupiec, M. (2018). Spontaneous Changes in Ploidy Are Common in Yeast. Curr. Biol. *28*, 825--835.e4.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515–1526.

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

Hou, J., van Leeuwen, J., Andrews, B.J., and Boone, C. (2018). Genetic Network Complexity Shapes Background-Dependent Phenotypic Expression. Trends Genet. *34*, 578–586.

Hou, J., Tan, G., Fink, G.R., Andrews, B.J., and Boone, C. (2019). Complex modifier landscape underlying genetic background effects. Proc. Natl. Acad. Sci. U. S. A. *116*, 5045–5054.

Johnson, M.S., Gopalakrishnan, S., Goyal, J., Dillingham, M.E., Bakerlee, C.W., Humphrey, P.T., Jagdish, T., Jerison, E.R., Kosheleva, K., Lawrence, K.R., et al. (2021). Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations. Elife *10*.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature *421*, 231–237.

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., et al. (2010). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat. Biotechnol. *28*, 617–623.

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

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal; Vol 17, No 1 Next Gener. Seq. Data Anal.

Michel, A.H., Hatakeyama, R., Kimmig, P., Arter, M., Peter, M., Matos, J., De Virgilio, C., and Kornmann, B.T. (2017). Functional mapping of yeast genomes by saturated transposition. Elife *6*.

Mullis, M.N., Matsui, T., Schell, R., Foree, R., and Ehrenreich, I.M. (2018). The complex underpinnings of genetic background effects. Nat. Commun. *9*, 1–10.

Van Opijnen, T., and Levin, H.L. (2020). Transposon Insertion Sequencing, a Global Measure of Gene Function. Annu. Rev. Genet. *54*, 337–365.

Paaby, A.B., White, A.G., Riccardi, D.D., Gunsalus, K.C., Piano, F., and Rockman, M. V. (2015). Wild worm embryogenesis harbors ubiquitous polygenic modifier variation. Elife *4*, 1–17.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature *556*, 339–344.

Sackton, T.B., and Hartl, D.L. (2016). Genotypic Context and Epistasis in Individuals and Populations. Cell *166*, 279–287.

Sadhu, M.J., Bloom, J.S., Day, L., Siegel, J.J., Kosuri, S., and Kruglyak, L. (2018). Highly parallel genome variant engineering with CRISPR-Cas9. Nat. Genet. *50*, 510–514.

Sharon, E., Chen, S.-A.A., Khosla, N.M., Smith, J.D., Pritchard, J.K., and Fraser, H.B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. Cell 1–14.

Steinberg, M.H., and Sebastiani, P. (2012). Genetic modifiers of sickle cell disease. Am. J. Hematol. *87*, 795–803.

Venkataram, S., Dunn, B., Li, Y., Agarwala, A., Chang, J., Ebel, E.R., Geiler-Samerotte, K., Hérissant, L., Blundell, J.R., Levy, S.F., et al. (2016). Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast. Cell *166*, 1585-1596.e22.

Vu, V., Verster, A.J., Schertzberg, M., Chuluunbaatar, T., Spensley, M., Pajkic, D., Hart, G.T., Moffat, J., and Fraser, A.G. (2015). Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. Cell *162*, 391–402.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096–1101.

Williams, R.A., Mamotte, C.D.S., and Burnett, J.R. (2008). Phenylketonuria: an inborn error of phenylalanine metabolism. Clin. Biochem. Rev. *29*, 31–41.

CONCLUSION ET PERSPECTIVES

Améliorer la dissection de l'architecture génétique des traits complexes à l'échelle d'une espèce grâce à des stratégies à haut-débit

L'élucidation des bases génétiques des traits complexes se présente comme un objectif fondamental en médecine humaine notamment. En effet, l'identification des variants génétiques causaux et la compréhension des mécanismes sous-jacents sont un atout majeur pour comprendre des maladies, et potentiellement trouver de nouvelles cibles thérapeutiques. Cependant, l'exploration de l'origine génétique de la variance phénotype se révèle être extrêmement complexe. Afin de disséquer cette complexité, de nombreuses stratégies analytiques ont ainsi été mises en place sur différents organismes modèles (Cain et al., 2020; Doudna and Charpentier, 2014; MacKay et al., 2009; Segrè et al., 2006; Tam et al., 2019). Nos travaux se sont concentrés sur la levure *Saccharomyces cerevisiae*, une espèce pour laquelle de nombreux outils génétiques et moléculaires existent, et une population de plusieurs milliers d'individus a été séquencée (Duan et al., 2018; Gallone et al., 2016; Marsit et al., 2015; Peter et al., 2018). Deux techniques à haut-débit nous ont permis de franchir de nouvelles étapes pour mieux caractériser les relations entre le génotype et les phénotypes.

D'une part, une stratégie de mutagénèse par insertions de transposon a été réalisée dans une centaine d'individus représentant la diversité de l'espèce *S. cerevisiae*. Cette méthode permet de mettre en évidence la variation de la tolérance des pertes de fonction engendrées par l'insertion de transposon selon le fonds génétique. En effet, un modèle logistique basé sur la variation des profils d'insertions dans ~4500 gènes entre 39 isolats a permis de prédire l'effet des mutations de pertes de fonction sur le fitness des isolats. De cette manière, les pertes de fonction de 15% des gènes (n = 632) entraînent dans les différents fonds génétiques un gain ou une perte de fitness par rapport à la souche de référence, S288C. Parmi ces 632 cas, plus de 2/3 des gènes (n = 453) sont associés à des variations de fitness liées à l'environnement, à savoir une compétition en milieu composé de galactose. Des gènes impliqués dans les processus de respiration ont ainsi des conséquences variables sur le fitness dans différents fonds génétiques. L'étude de ces cas liés à l'environnement révèle également une anti-corrélation du fitness prédit entre des gènes de la respiration et des gènes impliqués dans la régulation de la transcription et le remodelage de la chromatine, ainsi que le transport nucléo-cytoplasmique. Dans un isolat, si la perte

de fonction des gènes relatifs à la respiration a des effets délétères sur le fitness de l'individu, alors aucun effet phénotypique ne sera observé pour les gènes associés à la régulation de transcription et / ou au transport nucléaire-cytoplasmique, et inversement. Par ailleurs, ~1/3 des gènes (n = 179) sont spécifiques aux effets du fonds génétique et sont majoritairement unique à un isolat. De manière intéressante, bien que ces cas soient rares, ces gènes sont fonctionnellement associés chez un même individu. Cela suggère donc une cohérence de l'effet du fonds génétique au sein de réseaux d'interactions génétiques. De plus, ces 15% de gènes présentant un effet dépendant du fonds génétique ont des caractéristiques évolutives communes et sont plus enclins à contenir des mutations.

D'autre part, une stratégie de séquençage des ARN messagers (RNA-seq) couplée à une étude d'association pangénomique ont fourni la base de l'exploration de la variation d'un millier de transcriptomes au sein de l'espèce *S. cerevisiae*. Une régulation importante et complexe du niveau d'expression des gènes a été mise en évidence dans la population. Alors que des évènements de domestication liés à certaines sous-populations ont apporté des signatures transcriptionnelles pour une voie métabolique précise, l'expression importante des processus métaboliques (glycolyse) et moléculaires (transcription, traduction) généraux est conservée de manière équivalente dans la population. Des mécanismes globaux de l'expression des gènes ont aussi été mis en évidence, tels que l'expression plus faible des gènes accessoires ou la compensation de l'expression des variants du nombre de copies, gènes ou chromosomes, par exemple. Grâce au jeu de données conséquent généré, une étude d'association pangénomique a été réalisée afin d'identifier les variants génétiques régulateurs de la variation de l'expression des gènes, ou eQTL, dans 969 isolats naturels. Au total, 4684 eQTL ont été associés à la variation d'expression de 2023 gènes, parmi lesquels environ 83,5% influencent l'expression des gènes à distance. Bien que minoritaires, les eQTL locaux, souvent localisés dans les promoteurs, expliquent une plus grande partie de la variance phénotypique observée dans la population (18% *vs*. 15,4%). De manière intéressante, tandis les eQTL distants ciblent à la fois le génome de base et les gènes accessoires, les eQTL locaux régulent majoritairement le génome accessoire, suggérant un mécanisme différent de régulation de la transcription pour ces gènes.

Plusieurs centaines de transcriptomes : ressource considérable pour démêler les relations génotype-phénotype

Ces premières analyses des 969 transcriptomes de *S. cerevisiae* constituent une première étape dans la dissection de l'origine génétique de la variation du niveau d'expression des gènes au sein d'une espèce. Cependant, le ou les eQTL associés à la variation d'expression de ~ 2000 gènes n'expliquent qu'une partie de la variance phénotypique (en moyenne 16%). Une importante héritabilité manquante subsiste encore (Manolio et al., 2009). Bien que les variants rares et à faible fréquence, c'est-à-dire avec une fréquence allélique inférieure à 5% dans la population étudiée, ne puissent pas être inclus dans les analyses d'association pangénomique (MacKay et al., 2009; Manolio et al., 2009; Tam et al., 2019), d'autres sources d'héritabilité manquante, comme les variants structurels par exemple, peuvent être caractérisées et étudiées. Différentes approches et perspectives sont ainsi envisagées avec le jeu de données généré.

Analyse approfondie du pangénome

Les méthodes d'extraction et de séquençage utilisées pour générer les transcriptomes permettent d'accéder à tous les ARNm transcrits au sein des isolats étudiés, à la fois correspondant aux gènes constituant le génome de base de l'espèce (core genome) ainsi que des gènes accessoires. Dans un premier temps, les gènes accessoires identifiés par séquençage de l'ADN génomique (Peter et al., 2018) ont servi pour étudier l'expression du génome accessoire par rapport au génome de base. De manière intéressante, des différences, à la fois du niveau d'expression et de la régulation, sont observées entre les gènes accessoires et le génome de base. Afin d'affiner ces résultats, il est possible de reconstruire le pangénome à partir des transcrits séquencés dans la population. Cette stratégie permettra alors de définir plus précisément les gènes présents et transcrits dans chaque génome, que ce soit des gènes accessoires déjà identifiés, mais aussi des gènes *de novo* exprimés dans certaines souches. Dans cette optique, un assemblage peut être réalisé à partir de toutes les lectures séquencées non alignées sur le génome de la souche de référence, S288C (Figure 1). À partir de ces assemblages, des recherches de type BLAST permettent de valider les gènes accessoires provenant d'évènements d'introgressions ou de transferts horizontaux de gènes. Ces analyses ont récemment été initiées dans l'équipe et révèlent une structure conservée des gènes accessoires au sein des différentes sous-populations. En effet, ces gènes accessoires sont introgressés de plus

d'une dizaine d'espèces différentes et pour chaque sous-population, l'origine, le nombre et l'expression de ces gènes définissent une signature de celle-ci. L'identification de gènes *de novo* est une tâche plus ardue à réaliser. En effet, ces gènes dérivent de régions intergéniques ou chevauchantes avec d'autres gènes et ne sont pas introgressés d'autres espèces (Vakirlis et al., 2018). Sans connaître l'orientation des lectures de RNA-seq, il est également difficile de discerner l'origine de la lecture entre gènes chevauchants. De plus, une majorité de gènes potentiels courts (< 100 pb) résultent de l'assemblage des lectures de RNA-seq non alignées sur le génome de référence, le bruit doit ainsi être différencié des réels gènes *de novo* (Blevins et al., 2021). L'étude du pangénome à travers le RNA-seq a un avantage principal car tous les gènes transcrits sont identifiés. La transcription de ces gènes met ainsi en évidence leur rôle fonctionnel, faisant office de marqueurs d'évolution et d'adaptation à l'environnement (Bergström et al., 2014; Blevins et al., 2021; Peter et al., 2018).

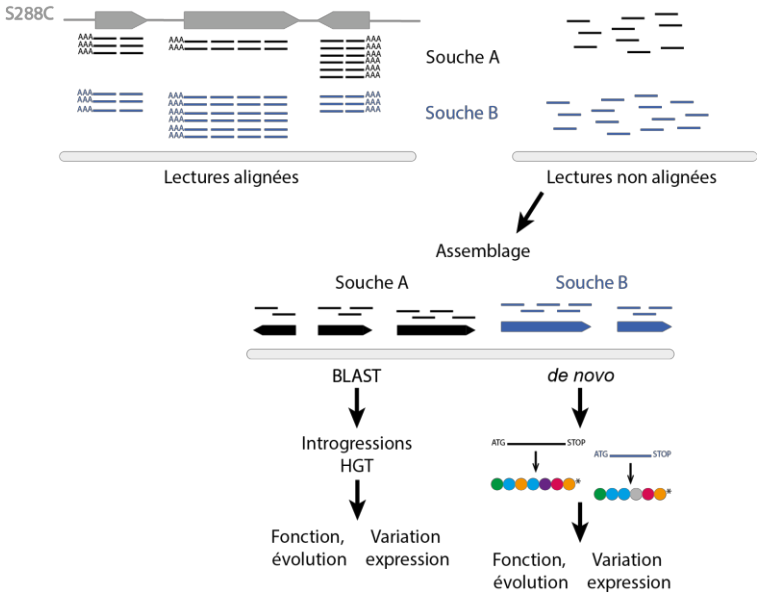


Figure 1. Etude de la composition et de l'expression du pangénome de *S. cerevisiae* à partir des lectures non alignées. Deux types de gènes accessoires peuvent être identifiés : les gènes provenant d'évènements d'introgressions ou de transferts horizontaux de gènes, ou des gènes *de novo* exprimés pour lesquels la séquence protéique peut prédire de leur fonction.

Caractérisation des variants structurels pour explorer l'héritabilité manquante

L'étude d'association pangénomique réalisée pour identifier les variants génétiques responsables de la variance du niveau d'expression des gènes prend uniquement en compte les sites de polymorphisme nucléotidique (SNP) de l'espèce. Cependant, les SNP ne représentent qu'une partie de la diversité génétique présente au sein de la population. Parmi les autres variants génétiques, les CNV ont pu être déterminés lors du séquençage des génomes des 1011 isolats naturels de *S. cerevisiae* (Peter et al., 2018). Ces CNV ont aussi été inclus dans les études d'association pangénomique pour 36 traits (croissance dans différentes conditions expérimentales) et ont un impact bien plus important que les SNP sur la variance phénotypique (Peter et al., 2018). Il est donc primordial d'inclure ces CNV dans nos études d'association pangénomique pour déterminer leur impact sur la variation d'expression des gènes. Cette perspective sera réalisée à court terme car la méthodologie et les données nécessaires sont connues. L'intégralité des variants structurels présents dans les génomes est néanmoins difficile à caractériser avec un simple séquençage des génomes par la stratégie Illumina. Un des objectifs de l'équipe, à moyen terme, consiste à reséquencer les génomes des 1011 isolats naturels de *S. cerevisiae* avec une stratégie de type Oxford Nanopore. Grâce à cette stratégie, des lectures longues sont générées (plusieurs kilobases). Ces lectures longues permettront ainsi d'identifier les variations incluant des centaines de paires de bases comme les translocations ou les inversions et leurs points de cassure sur les chromosomes. Grâce à ces données, une nouvelle matrice de variants génétiques incluant aussi la complexité structurelle des génomes sera la base de nouvelles études d'association pangénomique pour disséquer l'origine génétique de la variation d'expression des gènes.

Dissection des différents niveaux de régulation de l'expression des gènes comme intermédiaires des relations entre le génotype et les phénotypes

En parallèle de notre projet de RNA-seq, une collaboration avec l'équipe de Markus Ralser (The Francis Crick Institute-Londres / Charité University-Berlin) a permis d'établir le protéome pour une partie importante de la collection d'isolats de *S. cerevisiae*. Des valeurs d'abondance protéique ont été obtenues pour environ 2000 gènes dans 808 isolats naturels. De la même manière que pour le transcriptome, une étude d'association pangénomique a été réalisée et a révélé au total 1924 pQTL (pour *protein Quantitative Trait Loci*) impactant la variance phénotypique de 857 protéines traduites. Le chevauchement entre eQTL et pQTL détectés par les analyses d'association est peu important (n = 15/1924 QTL). Pour les gènes impliqués dans le métabolisme des acides aminés notamment, la régulation peut être affectée par les conditions de culture différentes entre les deux jeux de données, en SC+AA (milieu Synthétique Complet avec Acides Aminés) ou SC-AA (milieu Synthétique Complet sans Acides Aminés) pour le transcriptome ou le protéome, respectivement. Cependant, une étude antérieure, comparant transcriptomes et protéomes (générés dans les mêmes conditions) de la descendance d'un croisement entre 2 isolats, a aussi mis en évidence une différence importante entre variants génétiques associés à la variation de transcription et de traduction des gènes (Foss et al., 2007). Grâce aux données de protéomique récemment générées sur un grand échantillon d'isolats de *S. cerevisiae* (n = 808), nous constatons que la synthèse protéique est majoritairement régulée à distance (91,2% des pQTL) malgré un impact plus important des variants locaux (situés principalement dans les promoteurs) sur la variance phénotypique (Figure 2). Il est ainsi intéressant de noter que les règles de la régulation génétique de la transcription et de la traduction sont similaires dans ce sens. Cependant, les origines génétiques de cette régulation (à savoir les variants génétiques détectés) sont bien différentes.

Ces deux jeux de données fournissent également la possibilité d'explorer les mécanismes de régulation post-transcriptionnelle. De manière générale, nous constatons une atténuation post-transcriptionnelle illustrée par une diminution de la variance des niveaux d'abondance protéique entre les isolats naturels par rapport à la variance de l'expression des gènes (Figure 2). Plus spécifiquement, une compensation de la transcription, et encore plus de la synthèse protéique des gènes affectés par une aneuploïdie, est détectée chez les 200 individus concernés. Ces

résultats confirment, à l'échelle d'une population naturelle, les observations déjà établies dans des souches disomiques construites en laboratoire (Dephoure et al., 2014; Hose et al., 2015).

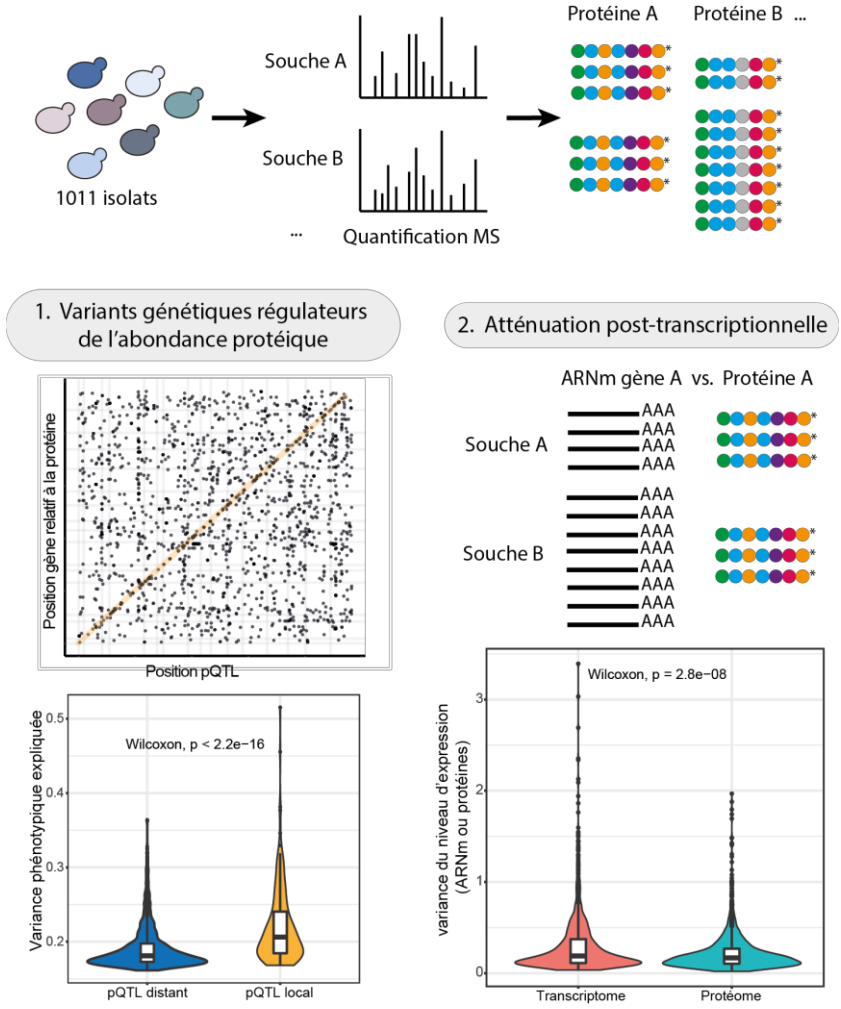


Figure 2. Exploration du protéome au sein de l'espèce de *S. cerevisiae*. Les valeurs d'abondance protéique dans la population sont générées par spectrométrie de masse et permettent de réaliser une étude d'association pangénomique (1.). Cette dernière révèle 91,2% de pQTL distants expliquant une plus faible proportion de la variance phénotypique que les pQTL locaux. La comparaison de la variance dans la population de chaque gène transcrit et traduit révèle une plus faible variance des valeurs d'abondance protéique (2.). Ceci suggère ainsi un mécanisme d'atténuation post-transcriptionnelle de l'expression des gènes.

L'ensemble de ces données illustre clairement différents niveaux de régulation de l'expression des gènes qui doivent être décrits avec précision pour comprendre l'implication des interactions entre transcription et traduction dans la diversité phénotypique observée au sein d'une population. Afin de compléter ces résultats, l'objectif, à moyen terme, sera de générer les protéomes de l'ensemble des isolats en milieu SC+AA, donc dans les mêmes conditions que les transcriptomes. Enfin, le cœur de la relation entre transcription et traduction peut aussi être exploré pour comprendre les mécanismes de régulation post-transcriptionnelle grâce à la stratégie de ribosome profiling (Ingolia et al., 2009). Cette technique consiste au séquençage des fragments d'ARNm pris en charge par les ribosomes pendant la traduction, révélant ainsi la dynamique de traduction des différents gènes. Un premier aperçu de cet aspect de la régulation post-transcriptionnelle a été établi avec la comparaison des niveaux de RNA-seq et de ribosome profiling pour ~ 4000 gènes dans 8 isolats naturels. De manière intéressante, ces travaux ont également montré une dynamique d'atténuation post-transcriptionnelle importante, notamment pour les gènes essentiels et / ou impliqués dans les complexes protéiques. De la même manière qu'avec le transcriptome et le protéome, il serait intéressant d'étudier un plus grand nombre d'individus afin de mettre en place une analyse d'association pangénomique. Cependant, le ribosome profiling reste encore difficilement applicable à grande échelle, notamment pour une population de levures de plusieurs centaines d'individus. Des avancées technologiques seront ainsi essentielles pour atteindre ces objectifs et disséquer la complexité derrière la régulation de l'expression des gènes.

La ressource des 1000 transcriptomes de *S. cerevisiae* constitue un atout majeur dans l'exploration des relations entre le génotype et le phénotype. De nombreuses perspectives découlent de ces travaux, et couplés à diverses stratégies expérimentales ou analytiques, permettront d'élucider la complexité génétique expliquant la variance phénotypique. À l'heure actuelle, grâce aux avancées scientifiques et technologiques, l'analyse de centaines et même de milliers de phénotypes à haut-débit se développe de manière croissante. Nous pouvons par exemple citer l'analyse de plus de 500 traits de morphologie des cellules (Ohya et al., 2005; Yvert et al., 2013) ou l'utilisation d'individus, identifiables par des code-barres moléculaires, en compétition dans des milliers de conditions expérimentales en parallèle (Piotrowski et al., 2017). Ces nouvelles techniques permettront ainsi de fournir des ressources pour des études d'association pangénomique supplémentaires.

Références

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

Blevins, W.R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J.L., Espinar, L., Díez, J., Carey, L.B., and Albà, M.M. (2021). Uncovering de novo gene birth in yeast using deep transcriptomics. Nat. Commun. *12*, 604.

Cain, A.K., Barquist, L., Goodman, A.L., Paulsen, I.T., Parkhill, J., and van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. Nat. Rev. Genet. *21*, 526–540.

Dephoure, N., Hwang, S., O'Sullivan, C., Dodgson, S.E., Gygi, S.P., Amon, A., and Torres, E.M. (2014). Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. Elife *3*, e03023.

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

Duan, S.F., Han, P.J., Wang, Q.M., Liu, W.Q., Shi, J.Y., Li, K., Zhang, X.L., and Bai, F.Y. (2018). The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. Nat. Commun. *9*.

Foss, E.J., Radulovic, D., Shaffer, S.A., Ruderfer, D.M., Bedalov, A., Goodlett, D.R., and Kruglyak, L. (2007). Genetic basis of proteome variation in yeast. Nat. Genet. *39*, 1369–1375.

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., et al. (2016). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. Cell *166*, 1397-1410.e16.

Hose, J., Yong, C.M., Sardi, M., Wang, Z., Newton, M.A., and Gasch, A.P. (2015). Dosage compensation can buffer copynumber variation in wild yeast. Elife *4*, 1–27.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

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

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

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

Ohya, Y., Sese, J., Yukawa, M., Sano, F., Nakatani, Y., Saito, T.L., Saka, A., Fukuda, T.,

Ishihara, S., Oka, S., et al. (2005). High-dimensional and large-scale phenotyping of yeast mutants. Proc. Natl. Acad. Sci. U. S. A. *102*, 19015 LP – 19020.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature *556*, 339–344.

Piotrowski, J.S., Li, S.C., Deshpande, R., Simpkins, S.W., Nelson, J., Yashiroda, Y., Barber, J.M., Safizadeh, H., Wilson, E., Okada, H., et al. (2017). Functional annotation of chemical libraries across diverse biological processes. Nat. Chem. Biol. *13*, 982–993.

Segrè, A. V, Murray, A.W., and Leu, J.-Y. (2006). High-resolution mutation mapping reveals parallel experimental evolution in yeast. PLoS Biol. *4*, e256.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484.

Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., and Lafontaine, I. (2018). A Molecular Portrait of De Novo Genes in Yeasts. Mol. Biol. Evol. *35*, 631–645.

Yvert, G., Ohnuki, S., Nogami, S., Imanaga, Y., Fehrmann, S., Schacherer, J., and Ohya, Y. (2013). Single-cell phenomics reveals intra-species variation of phenotypic noise in yeast. BMC Syst. Biol. *7*, 54.

ANNEXES

Articles scientifiques publiés

Fournier, T., Saada, O.A., Hou, J., Peter, J., **Caudal, E.**, Schacherer, J. (2019). Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. Elife *8*, 1–18.

Articles scientifiques en cours de rédaction

Caudal E., Hou J., Garin M., Dutreux F., Fournier T., Friedrich A., Schacherer J. Exploration of the differential impact of loss-of-function mutations linked to genetic backgrounds using transposition saturation. (en préparation)

Caudal E., Dutreux F., Teyssonnière E., Abou Saada O., Caradec C., Friedrich A., Schacherer J. Species-wide exploration of the inherited gene expression variation in yeast. (en préparation)

Communications scientifiques

Séminaire de Microbiologie de Strasbourg

Strasbourg, 29/03/2018, *communication orale*

Prix « Casden » de la seconde meilleure communication orale

RNA workshop, stRas rNa sAlon

Strasbourg, 11/06/2018, *communication orale*

13th international conference "Levures, Modèles & Outils"

Rheinau – Suisse, 13/09/2018, *communication orale*

Colloque iGénolevures

Strasbourg, 29/11/2018, *communication orale*

Enseignement

Mission complémentaire « OpenLAB »

2017-2019

L'opération OpenLAB consiste à faire découvrir le monde de la recherche à des élèves de première à travers 2h de travaux pratiques. Au cours de ces 2h, les élèves devaient répondre à une problématique biologique grâce à du matériel de laboratoire que nous apportions dans les lycées. L'objectif était également d'engager la discussion avec les élèves notamment sur leurs questions concernant l'orientation et les études supérieures et de leur expliquer notre quotidien de doctorant.

Mission complémentaire « Femmes en Sciences »

2019-2020

Les interventions « Femmes en Sciences », organisées par le Jardin des Sciences de l'Université de Strasbourg, ont pour objectif de rendre visible la place des femmes dans les différents domaines scientifiques. Des rencontres d'environ 2h avec une classe de collégiens ou lycéens permettent de discuter des *a priori* des élèves sur les domaines accessibles pour les femmes mais aussi de notre sujet de recherche à travers des échanges en petits groupes. Ces échanges ont été remplacés par une vidéo explicative de mon sujet de thèse accessible pour les élèves intéressés lors du confinement.

eLIFE elifesciences.org

RESEARCH ARTICLE

Extensive impact of low-frequency variants on the phenotypic landscape at population-scale

Téo Fournier, Omar Abou Saada, Jing Hou, Jackson Peter, Elodie Caudal, Joseph Schacherer*

Université de Strasbourg, CNRS, GMGM UMR 7156, Strasbourg, France

Abstract Genome-wide association studies (GWAS) allow to dissect complex traits and map genetic variants, which often explain relatively little of the heritability. One potential reason is the preponderance of undetected low-frequency variants. To increase their allele frequency and assess their phenotypic impact in a population, we generated a diallel panel of 3025 yeast hybrids, derived from pairwise crosses between natural isolates and examined a large number of traits. Parental versus hybrid regression analysis showed that while most phenotypic variance is explained by additivity, a third is governed by non-additive effects, with complete dominance having a key role. By performing GWAS on the diallel panel, we found that associated variants with low frequency in the initial population are overrepresented and explain a fraction of the phenotypic variance as well as an effect size similar to common variants. Overall, we highlighted the relevance of low-frequency variants on the phenotypic variation.
DOI: https://doi.org/10.7554/eLife.49258.001

***For correspondence:** schacherer@unistra.fr

Competing interests: The authors declare that no competing interests exist.

Funding: See page 15

Received: 12 June 2019
Accepted: 23 October 2019
Published: 24 October 2019

Reviewing editor: Christian R Landry, Université Laval, Canada

Introduction

Natural populations are characterized by an astonishing phenotypic diversity. Variation observed among individuals of the same species represents a powerful raw material to develop better insight into the relationship existing between genetic variants and complex traits (***Mackay et al., 2009***). The recent advances in high-throughput sequencing and phenotyping technologies greatly enhance the ability to determine the genetic basis of traits in various organisms (***Alonso-Blanco et al., 2016***; ***Auton et al., 2015***; ***Mackay et al., 2012***; ***Peter et al., 2018***). Dissection of the genetic mechanisms underlying natural phenotypic diversity is within easy reach when using classical mapping approaches such as linkage analysis and genome-wide association studies (GWAS) (***Mackay et al., 2009***; ***Visscher et al., 2017***). Alongside these major advances, however, it must be noted that there are some limitations. All genotype-phenotype correlation studies in humans and other model eukaryotes have identified causal loci in GWAS explaining relatively little of the observed phenotypic variance of most complex traits (***Eichler et al., 2010***; ***Hindorff et al., 2009***; ***Manolio et al., 2009***; ***Shi et al., 2016***; ***Stahl et al., 2012***; ***Wood et al., 2014***; ***Zuk et al., 2014***).

Despite the efforts made to find the genetic variants responsible for complex traits, the variants found explain only a small part of the heritability, that is of the fraction of the phenotypic variance explained by the underlying genetic variability. One of the most striking examples is observed with human height. This trait is estimated to be 60–80% heritable (***Speed et al., 2017***; ***Visscher et al., 2008***) but close to 700 variants found in an analysis based on more than 250,000 individuals only explain 20% of this total heritability (***Wood et al., 2014***). Multiple justifications for this so-called missing heritability have been suggested, including the presence of low-frequency variants, (***Gibson, 2012***; ***Hindorff et al., 2009***; ***Manolio et al., 2009***; ***Pritchard, 2001***; ***Walter et al., 2015***),

structural variants (*e.g.* copy number variants) (**Peter et al., 2018**), small effect variants, as well as the low power to estimate non-additive effects (**Cordell, 2009**; **Mackay, 2014**; **Zuk et al., 2012**).

Variants present in less than 5% of the individuals are coined as low-frequency variants and are known to be involved in a large number of rare Mendelian disorders (**Gibson, 2012**). However, implication of rare variants is also pervasive in common diseases and other complex traits. Assessing the impact and effect of low-frequency variants at a population scale and on a large phenotypic spectrum will allow to gain better insight into the genetic architecture of the phenotypic variation in a species. As GWAS cannot deal with low-frequency and rare variants due to statistical limitations, except for very large sample sizes, their effect has often been overlooked.

Among model organisms, the budding yeast *Saccharomyces cerevisiae* is especially well suited to dissect variations observed across natural populations (**Fay, 2013**; **Peter and Schacherer, 2016**). *S. cerevisiae* isolates can be found in a broad array of biotopes both human-associated (*e.g.* wine, sake, beer and other fermented beverages, food, human body) or wild (*e.g.* plants, soil, insects) and are distributed world-wide (**Peter et al., 2018**). Phenotypic diversity among yeast isolates is significant and the *S. cerevisiae* species presents a high level of genetic diversity ($\pi = 3\times10^{-3}$), much greater than that found in humans (**Lek et al., 2016**). Because of their small and compact genomes, an unprecedented number of 1,011 *S. cerevisiae* natural isolates has recently been sequenced (**Peter et al., 2018**). Yeast genome-wide association analyses have revealed functional Single Nucleotide Polymorphisms (SNPs), explaining a small fraction of the phenotypic variance (**Peter et al., 2018**). However, these analyses highlighted the importance of the copy number variants (CNVs), which account for a larger proportion of the phenotypic variance and have greater effects on phenotypes compared to the SNPs. Nevertheless, even when CNVs and SNPs are taken together, the phenotypic variance explained is still low (approximately 17% on average) and consequently a large part of it is unexplained.

Interestingly, much of the detected genetic polymorphisms in the 1011 yeast genomes dataset are low-frequency variants with almost 92.7% of the polymorphic sites associated with a minor allele frequency (MAF) lower than 0.05. This trend is similar to that observed in the human population (**Auton et al., 2015**; **Walter et al., 2015**) and definitely raised a question regarding the impact of low-frequency variants on the phenotypic landscape within a population and on the missing heritability (**Zuk et al., 2014**). Here, we investigated the underlying genetic architecture of phenotypic variation as well as unraveling part of the missing heritability by accounting for low-frequency genetic variants at a population-wide scale and non-additive effects controlled by a single locus. For this purpose, we generated and examined a large set of traits in 3025 hybrids, derived from pairwise crosses between a subset of natural isolates from the 1,011 *S. cerevisiae* population. This diallel crossing scheme allowed us to capture the fraction of the phenotypic variance controlled by both additive and non-additive phenomena as well as infer the main modes of inheritance for each trait. We also took advantage of the intrinsic power of this diallel design to perform GWAS and assess the role of the low-frequency variants on complex traits.

Results

Diallel panel and phenotypic landscape

Based on the genomic and phenotypic data from the 1,011 *S. cerevisiae* isolate collection (**Peter et al., 2018**), we selected a subset of 55 isolates that were diploid, homozygous, genetically diverse (**Figure 1a**), and originated from a broad range of ecological sources (**Figure 1b**) (*e.g.* tree exudates, *Drosophila*, fruits, fermentation processes, clinical isolates) as well as geographical origins (Europe, America, Africa and Asia) (**Figure 1c** and **Supplementary file 1**). A full diallel cross panel was constructed by systematically crossing the 55 selected isolates in a pairwise manner (**Figure 1d**). In total, we generated 3025 hybrids, representing 2970 heterozygous hybrids with a unique parental combination and 55 homozygous hybrids. All 3025 hybrids were viable, indicating no dominant lethal interactions existed between the parental isolates. We then screened the entire set of the parental isolates and hybrids for quantification of mitotic growth abilities across 49 conditions that induce various physiological and cellular responses (**Figure 1—figure supplement 1**, **Figure 1—figure supplement 2**, **Supplementary file 2**). We used growth as a proxy for fitness traits (see

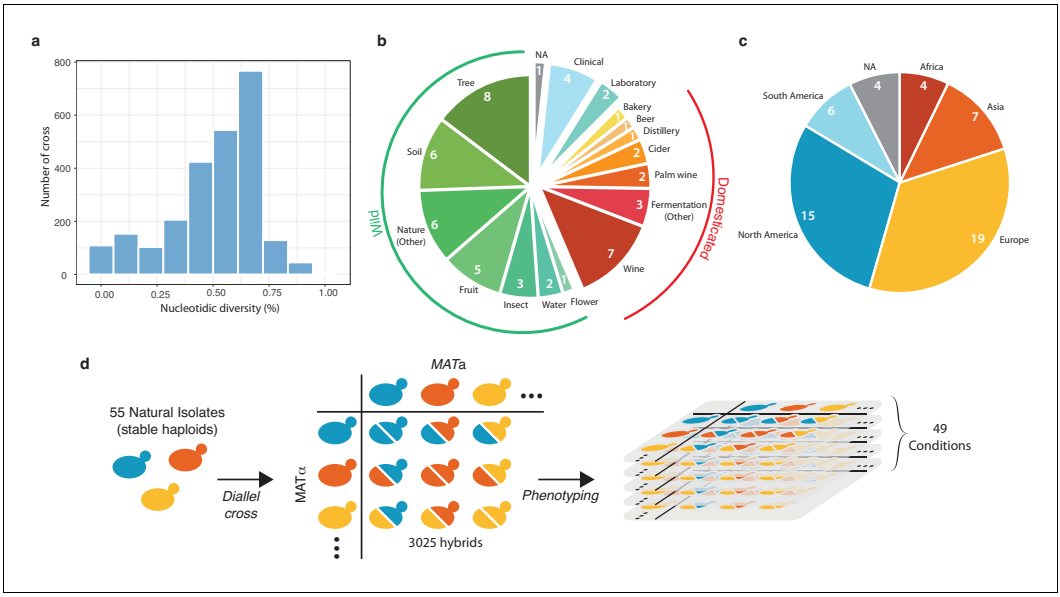


Figure 1. Diversity of the 55 selected natural isolates and diallel design. (**a**) Pairwise sequence diversity between each pair of parental strains. (**b**) Ecological origins of the selected strains. See also ***Supplementary file 1***. (**c**) Geographical origins of the selected strains. (**d**) Generation of the diallel hybrid panel. 55 natural isolates available as both mating types as stable haploids were crossed in a pairwise manner to obtain 3025 hybrids. This panel was then phenotyped on 49 growth conditions impacting various cellular processes.

DOI: https://doi.org/10.7554/eLife.49258.002

The following source data and figure supplements are available for figure 1:

Source data 1. Growth ratios for every hybrid and parental isolate on each growth condition.

DOI: https://doi.org/10.7554/eLife.49258.006

Figure supplement 1. Phenotypic variance in hybrids.

DOI: https://doi.org/10.7554/eLife.49258.003

Figure supplement 2. Correlation between conditions.

DOI: https://doi.org/10.7554/eLife.49258.004

Figure supplement 3. Phenotypic correlation between *MAT*a and *MAT*α isolate.

DOI: https://doi.org/10.7554/eLife.49258.005

Materials and methods). Ultimately, this phenotyping step led to the characterization of 148,225 hybrid/trait combinations.

Estimation of genetic variance components using the diallel panel (additive vs. non-additive)

The diallel cross design allows for the estimation of additive *vs.* non-additive genetic components contributing to the variation in each trait by calculating the combining abilities following Griffing's model (***Griffing, 1956***). For each trait, the General Combining Ability (GCA) for a given parent refers to the average fitness contribution of this parental isolate across all of its corresponding hybrid combinations, whereas the Specific Combining Ability (SCA) corresponds to the residual variation unaccounted for from the sum of GCAs from the parental combination. Consequently, the phenotype of a given hybrid can be formulated as $\mu + GCA_{parent1} + GCA_{parent2} + SCA_{hybrid}$, where μ is the mean fitness of the population for a given trait. We found a near perfect correlation (Pearson's r = 0.995, p-value<2.2e-16) between expected and observed phenotypic values, confirming the accuracy of

the model used (see Materials and methods). Using GCA and SCA values, we estimated both broad- (H^2) and narrow-sense (h^2) heritabilities for each trait (***Figure 1***). Broad-sense heritability is the fraction of phenotypic variance explained by genetic contribution. In a diallel cross, the total genetic variance is equal to the sum of the GCA variance of both parents and the SCA variance in each condition. Narrow-sense heritability refers to the fraction of phenotypic variance that can be explained only by additive effects and corresponds to the variance of the GCA in each condition (***Figure 2a***). The H^2 values for each condition ranged from 0.64 to 0.98, with the lowest value observed for fluconazole (1 µg.ml^{-1}) and the highest for sodium meta-arsenite (2.5 mM), respectively. The additive part (h^2 values) ranged from 0.12 to 0.86, with the lowest value for fluconazole (1 µg.ml^{-1}) and the highest for sodium meta-arsenite (2.5 mM), respectively. While broad- and narrow-sense heritabilities are variable across conditions, we also observed that on average, most of the phenotypic variance can be explained by additive effects (mean h^2 = 0.55). However, non-additive components contribute significantly to some traits, explaining on average one third of the phenotypic variance observed (mean H^2 - h^2 = 0.29) (***Figure 2b***). Despite a good correlation between broad- and narrow-sense heritabilities (Pearson's r = 0.809, p-value=1.921e-12) (***Figure 2c***), some

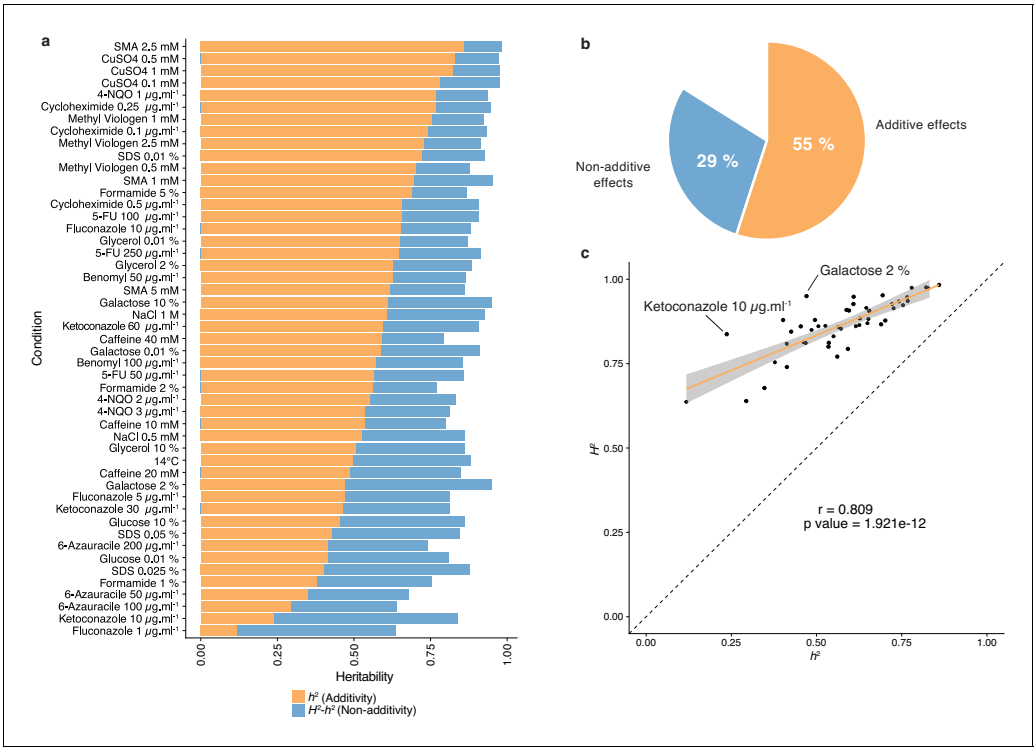


Figure 2. Heritability measurements. (**a**) The whole bar represents the overall heritability (H^2) for each condition tested. Orange part of the bars represents the narrow-sense heritability h^2, that is the fraction of phenotypic variance explained by additive effects, while blue part depicts the fraction of phenotypic variance explained by non-additive effects. (**b**) Overall mean additive and non-additive effects for every tested growth condition. (**c**) Representation of H^2 as a function of h^2 showing the relative additive versus non-additive effects for each condition. Outlier conditions in terms of non-additive variance will lie further away from the linear regression line. Person's r (95% confidence interval: 0.684–0.889) with the corresponding p-value is displayed.

DOI: https://doi.org/10.7554/eLife.49258.007

traits display a larger non-additive contribution, such as in galactose (2%) or ketoconazole (10 µg/ml). Interestingly, we revealed that these two conditions revealed to be mainly controlled by dominance (see below). Altogether, our results highlight the main role of additive effects in shaping complex traits at a population-scale and clearly show that this is not restricted to the single yeast cross where this trend was first observed (***Bloom et al., 2013***; ***Bloom et al., 2015***). Nonetheless, non-additive effects still explain a third of the observed phenotypic variance. This result also corroborates at a species-wide level the extensive impact of non-additive effects on phenotypic variance (***Forsberg et al., 2017***; ***Yadav et al., 2016***).

Relevance of dominance for non-additive effects

To have a precise view of the non-additive components, the mode of inheritance and the relevance of dominance for genetic variance, we focused on the deviation of the hybrid phenotypes from the expected value under a full additive model. Under this model, the hybrid phenotype is expected to be equal to the mean between the two parental phenotypes, hereinafter referred as Mean Parental Value or Mid-Parent Value (MPV). Deviation from this MPV allowed us to infer the respective mode of inheritance for each hybrid/condition combination (***Lippman and Zamir, 2007***), that is additivity, partial or complete dominance towards one or the other parent and finally overdominance or underdominance (***Figure 3a–b***, see Materials and methods). Only 17.4% of all hybrid/condition combinations showed enough phenotypic separation between the parents and the corresponding hybrid, allowing the complete partitioning in the seven above-mentioned modes of inheritance. For the 82.6% remaining cases, only a separation of overdominance and underdominance can be achieved

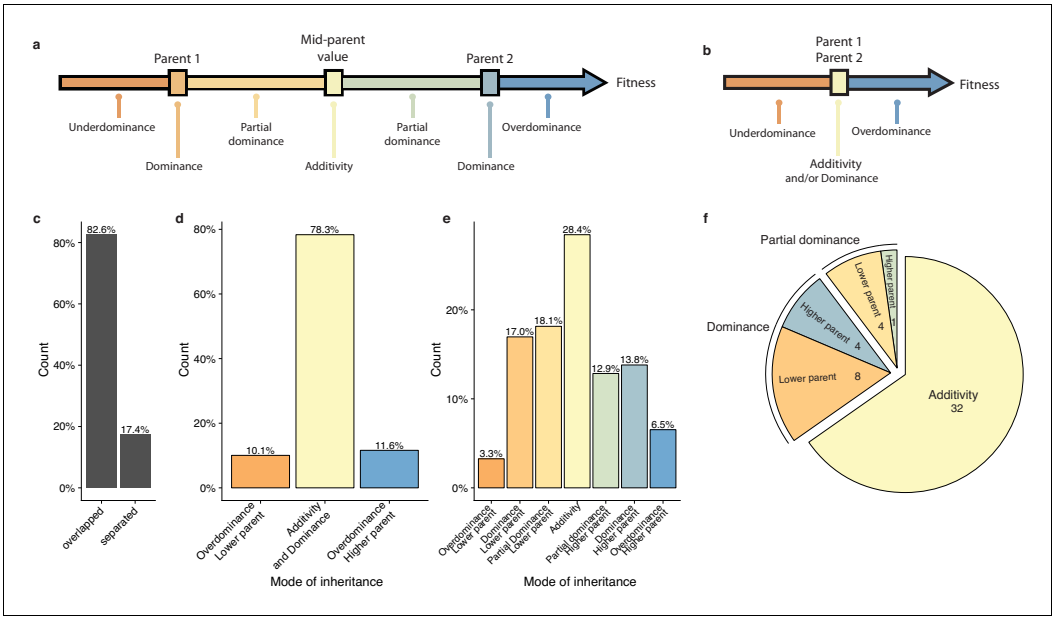


Figure 3. Mode of inheritance. (**a**) Representation of the different mode of inheritance depending on the hybrid value when a separation can be achieved between parental strains and (**b**) if a clear separation cannot be achieved between parental strains. (**c**) Percentage of parental phenotypes separated from each other for which a complete partition of different inheritance modes can be achieved. (**d**) Inheritance modes for every cross and condition where no separation can be achieved between the two homozygous parents. e. Inheritance modes for every cross and condition where a clear phenotypic separation can be achieved between the two homozygous parents. (**f**) The number of conditions in each main inheritance mode.
DOI: https://doi.org/10.7554/eLife.49258.008

(*Figure 3c*). Interestingly, these events are not as rare as previously described (*Zörgö et al., 2012*), with 11.6% of overdominance and 10.1% of underdominance (*Figure 3d*). When a clear separation is possible (*Figure 3e*), one third of the condition/cross combinations detected were purely additive whereas the rest displayed a deviation towards one of the two parents, with no bias (*Figure 3e*). When looking at the inheritance mode in each condition, most of the studied growth conditions (32 out of 49) showed a prevalence of additive effects (*Figure 3f*). However, 17 conditions were not predominantly additive throughout the population. Indeed, a total of 12 conditions were detected as mostly dominant with 4 cases of best parent dominance, including galactose (2%) and ketoconazole (10 $\mu g.ml^{-1}$), and 8 of worst parent dominance. The remaining five conditions displayed a majority of partial dominance (*Figure 3f*). These results confirm the importance of additivity in the global architecture of traits, but more importantly, they clearly demonstrate the major role of dominance as a driver for non-additive effects. Nevertheless, the presence of conditions with a high proportion of partial dominance combined with the cases of over and underdominance may indicate a strong and pervasive impact of epistasis on phenotypic variation.

Diallel design allows mapping of low-frequency variants in the population using GWAS

Next, we explored the contribution of low-frequency genetic variants (MAF <0.05) to the observed phenotypic variation in our population. Genetic variants considered by GWAS must have a relatively high frequency in the population to be detectable, usually over 0.05 for relatively small datasets (*Visscher et al., 2017*). Consequently, low-frequency variants are evicted from classical GWAS. However, the diallel crossing scheme stands as a powerful design to assess the phenotypic impact of low-frequency variants present in the initial population as each parental genome is presented several times, creating haplotype mixing across the matrix and preserving the detection power in GWAS.

To avoid issues due to population structure, we selected a subset of hybrids from 34 unrelated isolates in the original panel to perform GWAS (see Materials and methods, *Supplementary file 1*). By combining known parental genomes, we constructed 595 hybrid genotypes *in* silico, matching one half matrix of the diallel plus the 34 homozygous diploids. We built a matrix of genetic variants for this panel and filtered SNPs to only retain biallelic variants with no missing calls. In addition, due to the small number of unique parental genotypes, extensive long-distance linkage disequilibrium was also removed (see Materials and methods), leaving a total of 31,632 polymorphic sites in the diallel population. Overall, 3.8% (a total of 1,180 SNPs) had a MAF lower than 0.05 in the initial population of the 1,011 *S. cerevisiae* isolates but surpassed this threshold in the diallel panel, reaching a MAF of 0.32 (*Figure 4a–b*).

To map additive as well as non-additive variants impacting phenotypic variation, we performed GWA using two different models (*Seymour et al., 2016*) (see Materials and methods). We used a classical additive model, encoding for SNPs where linear relationship between trait and genotype is assessed, that is every locus has a different encoding for each genotype. To account for non-additive inheritance, we also used an overdominant model, which only considers differences between heterozygous and homozygous thus revealing overdominant and dominant effects. For each of these two models, we performed mixed-model association analysis of the 49 growth conditions with FaST-LMM (*Lippert et al., 2011*; *Widmer et al., 2015*). Overall, GWAS revealed 1723 significantly associated SNPs (*Figure 4—source data 1*) by detecting from 2 to 103 significant SNPs by condition, with an average of 39 SNPs per condition. Minor allele frequencies of the significantly associated SNPs were determined in the 1011 sequenced genomes, from which the diallel parents were selected (*Figure 4*). Interestingly, 16.3% of the significant SNPs (281 in total) corresponded to low-frequency variants (MAF <0.05), with 19.5% of them (55 SNPs) being rare variants (MAF <0.01). This trend is the same and maintained for both models, with 19.3% and 15.2% of low-frequency variants for the additive and overdominant models, respectively. Due to the scheme used, it is important to note that it is possible to increase the MAF of low-frequency variants at a detectable threshold in the diallel panel and to query their effects but it is still difficult for truly rare variants (MAF <0.01), probably leading to an underestimation. However, these results clearly show that low-frequency variants indeed play a significant part in the phenotypic variance at the population-scale. We then estimated the contribution of the significant variants to total phenotypic variation (see Materials and methods) in our panel and found that detected SNPs could explained 15% to 32% of the variance, with a median of 20% (*Figure 4d*). When looking at the variance explained by each variant over their

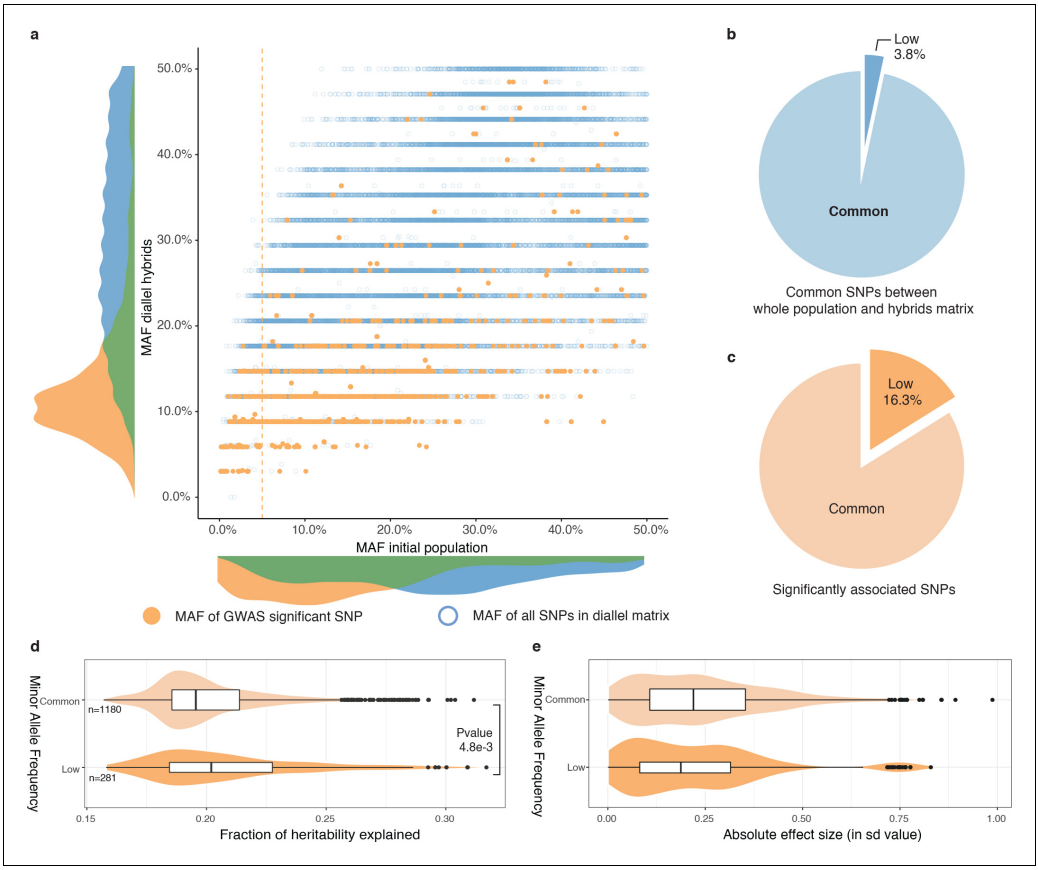


Figure 4. Rare and low-frequency variants detection. (**a**) Comparison of MAF for each SNP between the whole population (1011 strains) and the hybrid diallel matrix used for GWAS. Hollow blue circles represent the MAF of all SNPs common to the initial population and the diallel hybrids (31,632). Full orange circles show the MAF of significantly associated SNPs. Vertical orange line shows the 5% MAF threshold. (**b**) Proportion of SNPs with a MAF below 0.05. (**c**) Proportion of significantly associated SNPs with a MAF below 0.05. (**d**) Fraction of heritability explained for common and low-frequency variants. P-value was calculated using a two-sided Mann-Whitney-Wilcoxon test, difference in location of $-4.5e^{-3}$ (95% confidence interval $-7.9e^{-3}$ - $1.4e^{-3}$). (**e**) Absolute effect size of common and low-frequency variants.

DOI: https://doi.org/10.7554/eLife.49258.009

The following source data and figure supplement are available for figure 4:

Source data 1. Significantly associated SNPs SNPs without MAF are SNPs that were not biallelic in the initial population of 1011 isolates (***Peter et al., 2018***).

DOI: https://doi.org/10.7554/eLife.49258.011

Figure supplement 1. Significantly associated SNPs.

DOI: https://doi.org/10.7554/eLife.49258.010

respective allele frequency, it is noteworthy that low-frequency variants explained roughly the same proportion of the phenotypic variation (median of 20.2%) than the common SNPs (median of 19.6%) (***Figure 4d***). In addition, the variance explained by the associated rare variants were also higher on average than the rest of the detected SNPs (***Figure 4—figure supplement 1a***). It is noteworthy that

this trend was robust and conserved across the two encoding models implemented, accounting for additive and overdominant effects (***Figure 4—figure supplement 1a***). However, these results cannot be extrapolated to the whole population and only hold in the scope of our diallel population where these variants are now overrepresented compared to the natural population. Indeed, variance explained is related to the surveyed population because its value relies on the MAF of the variants. Therefore, in the whole natural population of 1011 isolates, their contribution to the phenotypic variance will be less important because of their lower MAF. To obtain a value that is unrelated to the studied population, we measured their respective effect size (***Figure 4e***). Here again we found that on average, low-frequency variant have about the same effect size (mean of 0.23 sd) than the common variants (mean of 0.25 sd).

To gain insight into the biological relevance of the set of associated SNPs, we first examined their distribution across the genome and found that 62.5% of them are in coding regions (with coding regions representing a total of 72.9% of the *S. cerevisiae* genome) (***Figure 4—figure supplement 1b***), with all of these SNPs distributed over a set of 546 genes. Over the last decade, an impressive number of quantitative trait locus (QTL) mapping experiments were performed on a myriad of phenotypes in yeast leading to the identification of 145 quantitative trait genes (QTG) (***Peltier et al., 2019***) and we found that 19 of the genes we detected are included in this list (***Figure 4—figure supplement 1c***). In addition, 22 associated genes were also found as overlapping with a recent large-scale linkage mapping survey in yeast (***Bloom et al., 2019***) (***Figure 4—figure supplement 1c***). We then asked whether the associated genes were enriched for specific gene ontology (GO) categories (***Supplementary file 3***). This analysis revealed an enrichment (p-value=5.39×10^{-5}) in genes involved in 'response to stimulus' and 'response to stress', which is in line with the different tested conditions leading to various physiological and cellular responses.

SGD1 and the mapping of a low-frequency variant

Finally, we focused on one of the most strongly associated genetic variant out of the 281 low-frequency variants significantly associated with a phenotype. The chosen variant was characterized by two adjacent SNPs in the *SGD1* gene and was detected in 6-azauracile (100 µg.ml^{-1}) with a p-value of 2.75e-8 with the overdominant encoding and 6.26e-5 with the additive encoding. Their MAF in the initial population is only 2.5% and reached 9% in the diallel panel with three genetically distant strains carrying it (***Figure 5a***). The SNPs are in the coding sequence of *SGD1*, an essential gene encoding a nuclear protein. The minor allele (AA) induces a synonymous change (TT**G** (Leu) → TT**A** (Leu)) for the first position and a non-synonymous mutation (**G**AA (Glu)→ **A**AA (Lys)) for the second position (***Figure 5a***). The phenotypic advantage conferred by this allele was observed with a significant difference between the homozygous for the minor allele, heterozygous and homozygous for the major allele (***Figure 5b***). To functionally validate the phenotypic effect of this low-frequency variant, CRISPR-Cas9 genome-editing was used in the three strains carrying the minor allele (AA) in order to switch it to the major allele (GG) and assess its phenotypic impact. Both mating types have been assessed for each strain. When phenotyping the wildtype strains containing the minor allele and the mutated strains with the major allele, we observed that the minor allele confers a phenotypic advantage of 0.2 in growth ratio compared to the major allele (***Figure 5c***) therefore validating the important phenotypic impact of this low-frequency variant. However, no assumptions can be made regarding the exact effect of this allele at the protein-level because no precise characterization has ever been carried out on Sgd1p and no particular domain has been highlighted.

Discussion

Understanding the source of the missing heritability is essential to precisely address and dissect the genetic architecture of complex traits. Over the years, the diallel hybrid panel design has proven its strength to dissect part of the genetic architecture of traits in populations. One of the main advantages of using such experimental design is the ability to precisely isolate the part of phenotypic variance that is controlled by additive effects from the one controlled by non-additive effects. While our analysis revealed that an important part of the phenotypic variance is linked to additive effects, about a third remains ruled by non-additive interactions encompassing dominance and epistasis. These results are in line with previous findings.

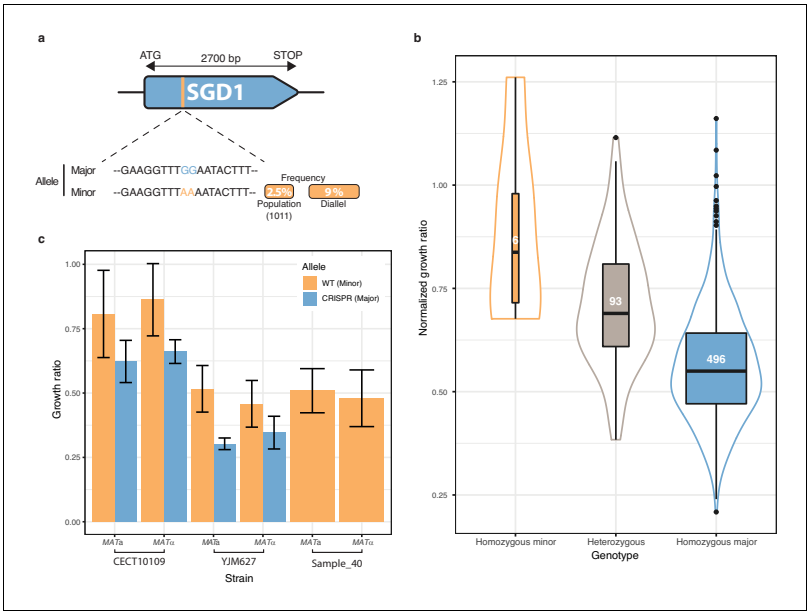


Figure 5. Low-frequency variant functional validation in 6-azauracil 100 µg.ml^{-1}. (**a**) Schematic representation of *SGD1* with the relative position of the detected SNPs. The minor allele is represented in orange with its MAF in the population and in the diallel cross panel. (**b**) boxplot and density plot of the normalized growth ratios for each genotype on 6-azauracil 100 µg.ml^{-1}. Number of observation is displayed in the boxplots. (**c**) Phenotypic validation after allele replacement of the minor allele with the major allele using CRISPR-Cas9 in the strains carrying the minor allele. Error bars represent median absolute deviation (four replicates).

DOI: https://doi.org/10.7554/eLife.49258.012

However, care should be taken with the classification of the mode of inheritance. Indeed, as we do not know how many loci are involved for each hybrid's phenotype, we can only assess the final phenotypic outcome of all the genetic variants involved and not on a locus by locus basis. This classification does not take into account their number, effect size and interactions. Consequently, the mode of inheritance that we described here solely reflects how the phenotype of the hybrid varies with respect to its parents. For example, several interactions could take place with opposite effect, leading to a final phenotype that appears as being controlled by an additive mode of inheritance (*i.e.* the hybrid phenotype equal to the mid parent value). However, in the cases where dominance was detected as a mode of inheritance, this might reflect the presence of a single locus having a strong phenotypic impact acting dominantly thus being responsible by itself for the phenotype. Yet, if two hybrids show a complete dominance in the same condition, it does not mean that the same alleles are involved in both.

Although few low-frequency and rare variants were considered in our GWAS (4%) due to stringent filtering conditions, a strong enrichment in these variants has been observed in the significantly associated ones (16%), demonstrating the ubiquity of low-frequency variants with important phenotypic impact. However, when looking at the population level, even though they do have effect sizes similar to common variants, they are not going to explain an important part of variance because it relies both on effect size and allele frequency. A good example of this phenomenon has been seen with a study of human height in more than 700,000 individuals. A total of 83 significantly associated rare and low-frequency variants with effect sizes up to 2 cm have been mapped (***Marouli et al., 2017***). On average, they explained the same amount of phenotypic variation as common variants,

which displayed much smaller effect sizes of about 1 mm. Our results suggest that a high number of low-frequency variants play a decisive role in the phenotypic landscape of a population both in term of number and effect size. Taken one by one, they do not explain a lot of phenotypic variance in a large population. Yet, altogether, they might actually explain a greater part of the variation than the one explained by common variants.

The contribution of rare and low-frequency variants to traits is largely unexplored. In humans, these genetic variants are widespread but only a few of them have been associated with specific traits and diseases (**Walter et al., 2015**). Recently, it has been shown that the missing heritability of height and body mass index is accounted for by rare variants (**Wainschtein et al., 2019**). We also recently found in yeast that most of the previously identified Quantitative Trait Nucleotides (QTNs) using linkage mapping were at low allele frequency in the 1,011 *S. cerevisiae* population (**Hou et al., 2016**; **Hou et al., 2019**; **Peltier et al., 2019**; **Peter et al., 2018**). A total of 284 QTNs were identified by linkage mapping and 150 of them are present at a low frequency in the population of 1011 isolates (**Peltier et al., 2019**; **Peter et al., 2018**). However, these QTNs were mapped with mostly closely related genetic backgrounds, encompassing a total of 59 strains with 30% of them coming from laboratory and 41% coming from the wine cluster, which has a very low genetic diversity (**Peter et al., 2018**). Moreover, experimentally validated QTNs are, most of the time, genetic variants with the most important phenotypic impact, which has been previously recognized as inducing an ascertainment bias (**Rockman, 2012**). It also raised the question of whether these rare and large effect size alleles discovered in specific crosses are really relevant to the variation across most of the population.

Here, we quantified the contribution of low-frequency variants across a large number of growth conditions and found that among all the genetic variants detected by GWAS on a diallel panel, 16.3% of them have a low-frequency in the initial population and explain a significant part of the phenotypic variance (21% on average). This particular diallel design also presents an intrinsic power to evaluate the additive vs. non-additive genetic components contributing to the phenotypic variation. We assessed the effect of intra-locus dominance on the non-additive genetic component and showed that dominance at the single locus level contributed to the phenotypic variation observed. However, other more complicated inter-loci interactions may still be involved. Altogether, these results have major implications for our understanding of the genetic architecture of traits in the context of unexplained heritability. In parallel to a recent large-scale linkage mapping survey in yeast (**Bloom et al., 2019**), our study highlights the extensive role of low-frequency variants on the phenotypic variation.

Materials and methods

Construction of the diallel panel

Selection of the *S. cerevisiae* isolates

Out of the collection of 1011 strains (**Peter et al., 2018**), a total of 53 natural isolates were carefully selected to be representative of the *S. cerevisiae* species. We selected isolates from a broad ecological origins and we prioritized for strains that were diploid, homozygous, euploid and genetically as diverse as possible, that is up to 1% of sequence divergence. All the isolate details, including ecological and geographical origins, are listed in **Supplementary file 1**. In addition to these 53 isolates, we included two laboratory strains, namely ∑1278b and the reference S288c strain.

Generation of stable haploids

For each selected parental strain, stable haploid strains were obtained by deleting the *HO* locus. The *HO* deletions were performed using PCR fragments containing drug resistance markers flanked by homology regions up and down stream of the *HO* locus, using standard yeast transformation method. Two resistance cassettes, *KanMX* and *NatMX*, were used for *MAT*a and *MAT*α haploids, respectively. The mating-type (*MAT*a and *MAT*α) of antibiotic-resistant clones was determined using testers of well-known mating type. For each genetic background, we selected a *MAT*a and *MAT*α clone that are resistant to G418 or nourseothricin, respectively.

Phenotyping of the parental haploid strains was performed to check for mating type-specific fitness effects. All *MAT*a and *MAT*α parental strains were tested on all 49 growth conditions (see

below) using the same procedure as the phenotyping assay of the hybrid matrix. The overall correlation between the *MAT*a and *MAT*α parental strains was 0.967 (Pearson, p-value<1e-324), with an average correlation per strain of 0.976 across different conditions (***Figure 1—figure supplement 3***). No significant mating type specificity was identified.

Diallel scheme

Parental strains were arrayed and pregrown in liquid YPD (1% yeast extract, 2% peptone and 2% glucose) overnight. Mating was performed with ROTOR (Singer Instruments) by pinning and mixing *MAT*a over *MAT*α parental strains on solid YPD. The parental strains, that is 55 *MAT*a *HO::ΔKanMX* and 55 *MAT*α *HO::ΔNatMX* strains were arrayed and mated in a pairwise manner on YPD for 24 hr at 30˚C. The mating mixtures were replicated on YPD supplemented with G418 (200 µg.ml^{-1}) and nourseothricin (100 µg.ml^{-1}) for double selection of hybrid individuals. After 24 hr, plates were replicated again on the same media to eliminate potential residuals of non-hybrids cells. In total, we generated 3025 hybrids, representing 2970 heterozygous hybrids with a unique parental combination and 55 homozygous hybrids.

High-throughput phenotyping and growth quantification

Quantitative phenotyping was performed using endpoint colony growth on solid media. Strains were pregrown in liquid YPD medium and pinned onto a solid SC (Yeast Nitrogen Base with ammonium sulfate 6.7 g.l^{-1}, amino acid mixture 2 g.l^{-1}, agar 20 g.l^{-1}, glucose 20 g.l^{-1}) matrix plate to a 1536 density format using the replicating ROTOR robot (Singer Instruments). Two biological replicates (coming from independent cultures) of each parental haploid strain were present on every plate and six biological replicates were present for each hybrid. As 27 plates were used in order to phenotype all the hybrids, 27 technical replicates (same culture in different plates) of the parents were present. The resulting matrix plates were incubated overnight to allow sufficient growth, which were then replicated onto 49 media conditions, plus SC as a pinning control (***Figure 1—figure supplement 1***, ***Supplementary file 2***). The selected conditions impact a broad range of cellular responses, and multiple concentrations were tested for each compound (***Figure 1—figure supplement 2***). Most tested conditions displayed distinctive phenotypic patterns, suggesting different genetic basis for each of them (***Figure 1—figure supplement 2***). The plates were incubated for 24 hr at 30˚C (except for 14˚C phenotyping) and were scanned with a resolution of 600 dpi at 16-bit grayscale. Quantification of the colony size was performed using the R package Gitter (***Wagih and Parts, 2014***) and the fitness of each strain on the corresponding condition was measured by calculating the normalized growth ratio between the colony size on a condition and the colony size on SC. As each hybrid is present in six replicates, the value considered for its phenotype is the median of all its replicates, thus smoothing the effects of pinning defect or contamination. This phenotyping step led to the determination of 148,225 hybrid/trait combinations (***Figure 1—source data 1***).

Diallel combining abilities and heritabilities

Combining ability values were calculated using half diallel with unique parental combinations, excluding homozygous hybrids from identical parental strains. For each hybrid individual, the fitness value is expressed using Griffing's model (***Griffing, 1956***):

$$z_{ij} = \mu + g_i + g_j + s_{ij} + e$$

Where z_{ij} is the fitness value of the hybrid resulting from the combination of i^{th} and j^{th} parental strains, μ is the mean population fitness, g_i and g_j are the general combining ability for the i^{th} and j^{th} parental strains, s_{ij} is the specific combining ability associated with the $i \times j$ hybrid, and e is the error term (i = 1...N, j = 1…N, N = 55). General combining ability for the i^{th} parent is calculated as:

$$g_i = \left(\frac{N-1}{N-2}\right) \times (\bar{z}_{i\cdot} - \mu)$$

Where N is the total number of parental types, $\bar{z}_{i\cdot}$ is the mean fitness value of all half sibling hybrids involving the i^{th} parent, and μ is the population mean. The error term associated with g_i is:

$$e_{g_i} = \sqrt{\frac{(N-1) \times \sigma^2 z_{ij\cdot}}{n \times N \times (N-2)}}$$

Where N is the total number of parental types, n is the number of replicates for the $i \times j$ hybrid, and $\sigma^2 z_{ij\cdot}$ is the variance of fitness values from a full-sib family involving the i^{th} and j^{th} parents, which is expressed as:

$$\sigma^2 z_{ij\cdot} = \sigma^2 z_i + \sigma^2 z_j + \sigma^2 z_{ij} + 2 \times cov(z_i, z_j)$$

Specific combining ability for the $i \times j$ hybrid combination therefore:

$$s_{ij} = \bar{z}_{ij\cdot} - g_i - g_j - \mu$$

The error term associated with s_{ij} is:

$$e_{s_{ij}} = \sqrt{\frac{(N-3) \times \sigma^2 z_{ij\cdot}}{n \times (N-1)}}$$

Using combining ability estimates, broad- and narrow-sense heritabilities can be calculated. Narrow sense heritability (h^2) accounts for the part of phenotypic variance explained only by additive variance, expressed as the additive variance (σ^2_A) over the total phenotypic variance observed (σ^2_P):

$$h^2 = \frac{\sigma^2_A}{\sigma^2_P} = \frac{\sigma^2_{(g_i+g_j)}}{\sigma^2_{(g_i+g_j)} + \sigma^2_{s_{ij}} + \sigma^2_e}$$

Where $\sigma^2_{(g_i+g_j)}$ is the sum of GCA variances, $\sigma^2_{s_{ij}}$ is the SCA variance and σ^2_e is the variance due to measurement error, which is expressed as:

$$\sigma^2_e = (N-2)\left(\bar{e}_{g_i} + \bar{e}_{g_j}\right)^2 + \frac{\left(\frac{(N^2-N)}{2} - 1\right)}{\left(\frac{(N^2-N)}{2} + N - 3\right)} \times \bar{e}^{\,2}_{s_{ij}}$$

On the other hand, broad-sense heritability (H^2) depicts the part of the phenotypic variance explained by the total genetic variance σ^2_G:

$$H^2 = \frac{\sigma^2_G}{\sigma^2_P} = \frac{\sigma^2_{(g_i+g_j)} + \sigma^2_{s_{ij}}}{\sigma^2_{(g_i+g_j)} + \sigma^2_{s_{ij}} + \sigma^2_e}$$

Phenotypic variance explained by non-additive variance is therefore equal to the difference between H^2 and h^2. All calculations were performed in R using custom scripts.

Calculation of mid-parent values and classification of mode of inheritance

Mid-Parent Value (MPV) is expressed as the mean fitness value of both diploid homozygous parental phenotypes:

$$MPV = \frac{P1 + P2}{2}$$

Comparing the hybrid phenotypic value (*Hyb*) to its respective parents' allows for an inference of the mode of inheritance for each hybrid/trait combination (***Figure 3a–b***). To obtain a robust classification, confidence intervals for each class were based on the standard deviation of hybrid (six replicates) and parents (54 replicates). *P2* is the phenotypic value of the fittest parent while *P1* is the phenotypic value of the least fit parent.

Inheritance mode	Formula
Underdominance	$Hyb1-(\sigma_{P1}+\sigma_{Hyb})$
Dominance P1	$P1-(\sigma_{P1}+\sigma_{Hyb})1+(\sigma_{P1}+\sigma_{Hyb})$
Partial dominance P1	$P1+(\sigma_{P1}+\sigma_{Hyb})-(\frac{\sigma_{P1}+\sigma_{P2}}{2}+\sigma_{Hyb})$
Additivity	$MPV+(\frac{\sigma_{P1}+\sigma_{P2}}{2}+\sigma_{Hyb})2-(\sigma_{P2}+\sigma_{Hyb})$
Partial dominance P2	$MPV-(\frac{\sigma_{P1}+\sigma_{P2}}{2}+\sigma_{Hyb})+(\frac{\sigma_{P1}+\sigma_{P2}}{2}+\sigma_{Hyb})$
Dominance P2	$P2-(\sigma_{P2}+\sigma_{Hyb})2+(\sigma_{P2}+\sigma_{Hyb})$
Overdominance	$P2+(\sigma_{P2}+\sigma_{Hyb})$

When a clear separation is possible between the two parental phenotypic values ($P1+\sigma_{P1}2-\sigma_{P2}$),the full decomposition in the seven above mentioned categories is possible (***Figure 3a***). However, in most of the cases, the two parental phenotypic values are not separated enough to achieve this but it is still possible to distinguish between overdominance and underdominance (***Figure 3b***, ***Figure 3d***). All calculations were performed in R using custom scripts.

Genome-wide association studies on the diallel panel

Whole genome sequences for the parental strains were obtained from the 1002 yeast genome project (***Peter et al., 2018***). Sequencing was performed by Illumina Hiseq 2000 with 102 bases read length. Reads were then mapped to S288c reference genome using bwa (v0.7.4-r385) (***Li and Durbin, 2009***). Local realignment around indels and variant calling has been performed with GATK (v3.3–0) (***McKenna et al., 2010***). The genotypes of the F1 hybrids were constructed in silico using 34 parental genome sequences. We retained only the biallelic polymorphic sites, resulting in a matrix containing 295,346 polymorphic sites encoded using the 'recode12' function in PLINK (***Chang et al., 2015***). Those genotypes correspond to a half-matrix of pairwise crosses with unique parental combinations, including the diagonal,that is the 34 homozygous parental genotypes. For each cross, we combined the genotypes of both parents to generate the hybrid diploid genome. As a result, heterozygous sites correspond to sites for which the two parents had different allelic versions. We removed long-range linkage disequilibrium sites in the diallel matrix due to the low number of founder parental genotypes by removing haplotype blocks that are shared more than twice across the population, resulting in a final dataset containing 31,632 polymorphic sites.

We performed GWA analyses with different encodings (***Seymour et al., 2016***). In the additive model, the genotypes of the F1 progeny were simply the concatenation of the genotypes from the parents. As homozygous parental alleles were encoded as 1 or 2, the possible alleles for each site in the F1 genotype were '11' and '22' for homozygous sites and '12' for heterozygous sites. We also used an overdominant genotype encoding, where both the homozygous minor and homozygous major alleles were encoded as '11' and the heterozygous genotype was encoded as '22'.

Mixed-model association analysis was performed using the FaST-LMM python library version 0.2.32 (https://github.com/MicrosoftGenomics/FaST-LMM) (***Widmer et al., 2015***). We used the normalized phenotypes by replacing the observed value by the corresponding quantile from a standard normal distribution, as FaST-LMM expects normally distributed phenotypes. The command used for association testing was the following: single_snp(bedFiles, pheno_fn, count_A1 = True), where bedFiles is the path to the PLINK formatted SNP data and pheno_fn is the PLINK formatted phenotype file. By default, for each SNP tested, this method excludes the chromosome in which the SNP is found from the analysis in order to avoid proximal contamination. Fast-LMM also computes the fraction of heritability explained for each SNP. The mixed model adds a polygenic term to the standard linear regression designed to circumvent the effects of relatedness and population stratification.

We estimated a condition-specific p-value threshold for each condition by permuting phenotypic values between individuals 100 times. The significance threshold was the 5% quantile (the 5th lowest p-value from the permutations). With that method, variants passing this threshold will have a 5% family-wise error rate. However, we do not have any estimation of the false positive rate. Taken together, GWA revealed 1723 significantly associated SNPs (***Figure 4—source data 1***), with 1273 and 450 SNPs for overdominant and additive model, respectively.

Variance explained and effect size

Variance explained by each SNP is calculated by PLINK. Care must be taken that in order to obtain the variance explained by all SNPs, it is not possible to sum up the variance explained by each individual SNP based on the fact that SNPs are not completely independent from one another.

The effect size was calculated using the formula for Cohen's d:

$$d = \frac{\bar{x_1} - \bar{x_2}}{sd_{Pooled}}$$

Where the pooled standard deviation is calculated with the following formula:

$$sd_{Pooled} = \sqrt{\frac{{sd_1}^2 + {sd_2}^2}{2}}$$

Under the additive model, the heterozygote phenotype is equidistant to both possible homozygote phenotypes (minor allele and major allele), so our calculation of the effect size could either compare the heterozygotes with the homozygotes in the minor allele, or the heterozygotes with the homozygotes in the major alleles. We chose to use the latter since the major allele grants us more statistical power. The formula we used to obtain the effect size for a given SNP under this model is the following:

$$Effect\ size = \frac{\bar{x}_{Heterozygous} - \bar{x}_{Major}}{sd_{Pooled}}$$

Under the overdominant model, the heterozygote phenotype is compared to the phenotype of the group of both homozygotes (minor and major), so the formula we used to obtain the effect size for a given SNP under this model is the following:

$$Effect\ size = \frac{\bar{x}_{Heterozygous} - \bar{x}_{Major\ and\ Minor}}{sd_{Pooled}}$$

Gene ontology analysis

GO term enrichment was performed using SGD GO Term Finder (https://www.yeastgenome.org/goTermFinder) with the 546 unique genes containing significantly associated SNPs (***Figure 4—source data 1*** and ***Supplementary file 3***). Significant enrichment is considered under 'Process' ontology with a p-value cutoff of 0.05.

CRISPR-Cas9 allele editing

pAEF5 plasmid containing Cas9 endonuclease and the guide RNA targeting *SGD1* was co-transformed with the repair fragment of 100 nucleotides containing the desired allele. Transformed cells were then plated on YPD supplemented with 200 µg.ml^{-1} hygromycin at 30˚C to select for transformants. Colonies were then arrayed on a 96 well plate with 100 µl YPD and grown for 24 hr to induce plasmid loss. The plate was then pinned back onto solid YPD for 24 hr then replica plated to YPD supplemented with 200 µg.ml^{-1} hygromycin to check for plasmid loss. Allele specific PCR was performed on colonies that lost the plasmid (***Wangkumhang et al., 2007***) to distinguish correctly edited allele from wildtype allele. Strains who showed amplification for the edited allele and no amplification for the wildtype allele were phenotyped (four technical replicates and four biological replicates) on the corresponding condition to measure differences with their wildtype counterparts.

Statistical tests

Person's correlation test was used to assess linear correlation between two sets.

Wilcoxon Mann Whitney was used to determine if two independent samples have the same distribution.

Correlogram of all tested growth conditions. Numbers in each cell represent 100 x Pearson's r value.

Acknowledgements

We thank Joshua Bloom and Leonid Kruglyak for insightful discussions, comments on the manuscript as well as for sharing their unpublished manuscript. We thank Maitreya Dunham and the members of the Schacherer laboratory for comments and suggestions. We also thank Gilles Fischer for providing the pAEF5 plasmid. This work was supported by a National Institutes of Health (NIH) grant R01 (GM101091-01) and a European Research Council (ERC) Consolidator grant (772505). TF is supported in part by a grant from the Ministère de l'Enseignement Supérieur et de la Recherche and in part by a fellowship from the medical association la Fondation pour la Recherche Médicale. JS is a Fellow of the University of Strasbourg Institute for Advanced Study (USIAS) and a member of the Institut Universitaire de France.

Additional information

Funding

Funder	Grant reference number	Author
National Institutes of Health	R01 GM101091-01	Joseph Schacherer
European Research Council	Consolidator grants (772505)	Joseph Schacherer
Fondation pour la Recherche Médicale	Graduate student grant	Téo Fournier
Institut Universitaire de France		Joseph Schacherer
University of Strasbourg Institute for Advanced Study		Joseph Schacherer
Ministère de l'Enseignement Supérieur et de la Recherche		Téo Fournier

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Téo Fournier, Conceptualization, Resources, Data curation, Software, Formal analysis, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Omar Abou Saada, Software, Formal analysis, Writing—review and editing; Jing Hou, Conceptualization, Software, Formal analysis, Methodology, Writing—review and editing; Jackson Peter, Software, Formal analysis; Elodie Caudal, Resources, Investigation, Writing—review and editing; Joseph Schacherer, Conceptualization, Supervision, Funding acquisition, Validation, Methodology, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Téo Fournier https://orcid.org/0000-0002-4860-6728
Joseph Schacherer https://orcid.org/0000-0002-6606-6884

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.49258.020
Author response https://doi.org/10.7554/eLife.49258.021

Additional files

Supplementary files

- Supplementary file 1. Strains used for the diallel cross with their ecological and geographical origins.

DOI: https://doi.org/10.7554/eLife.49258.013

- Supplementary file 2. Phenotyping conditions and their respective type of induced stress.

DOI: https://doi.org/10.7554/eLife.49258.014

• Supplementary file 3. GO Term associated with the 546 unique genes with a significantly associated SNPs.

DOI: https://doi.org/10.7554/eLife.49258.015

• Transparent reporting form DOI: https://doi.org/10.7554/eLife.49258.016

Data availability

All data generated or analysed during this study are included in the manuscript and supporting files. Source data files have been provided for Figures 1 and 4.

The following previously published dataset was used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Jackson Peter, Matteo De Chiara, Anne Friedrich, Jia-Xing Yue, David Pflieger, Anders Bergström, Anastasie Sigwalt, Benjamin Barre, Kelle Freel, Agnès Llored, Corinne Cruaud, Karine Labadie, Jean-Marc Aury, Benjamin Istace, Kevin Lebrigand, Pascal Barbry, Stefan Engelen, Arnaud Lemainque, Patrick Wincker, Gianni Liti, Joseph Schacherer	2018	Genome evolution across 1,011 Saccharomyces cerevisiae isolates	https://www.ncbi.nlm.nih.gov/sra?term=ERP014555	NCBI SRA, ERP014555

References

Alonso-Blanco C, Andrade J Becker C Bemm F Bergelson J Borgwardt JM Zhou X. 2016. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell* **166**:481–491. DOI: https://doi.org/10.1016/j.cell.2016.05.063, PMID: 27293186

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR, 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**:68–74. DOI: https://doi.org/10.1038/nature15393, PMID: 26432245

Bloom JS, Ehrenreich IM, Loo WT, Lite TL, Kruglyak L. 2013. Finding the sources of missing heritability in a yeast cross. *Nature* **494**:234–237. DOI: https://doi.org/10.1038/nature11867, PMID: 23376951

Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L. 2015. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nature Communications* **6**:8712. DOI: https://doi.org/10.1038/ncomms9712, PMID: 26537231

Bloom JS, Boocock J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, Kruglyak L. 2019. Rare variants contribute disproportionately to quantitative trait variation in yeast. *eLife* **8**:e49212. DOI: https://doi.org/10.7554/eLife.49212, PMID: 31647408

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**:7. DOI: https://doi.org/10.1186/s13742-015-0047-8, PMID: 25722852

Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics* **10**:392–404. DOI: https://doi.org/10.1038/nrg2579, PMID: 19434077

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11**:446–450. DOI: https://doi.org/10.1038/nrg2809, PMID: 20479774

Fay JC. 2013. The molecular basis of phenotypic variation in yeast. *Current Opinion in Genetics & Development* **23**:672–677. DOI: https://doi.org/10.1016/j.gde.2013.10.005, PMID: 24269094

Forsberg SK, Bloom JS, Sadhu MJ, Kruglyak L, Carlborg Ö. 2017. Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nature Genetics* **49**:497–503. DOI: https://doi.org/10.1038/ng.3800, PMID: 28250458

Gibson G. 2012. Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13**:135–145. DOI: https://doi.org/10.1038/nrg3118

Griffing B. 1956. Concept of general and specific combining ability in relation to diallel crossing systems. *Australian Journal of Biological Sciences* **9**:463–493. DOI: https://doi.org/10.1071/BI9560463
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS* **106**:9362–9367. DOI: https://doi.org/10.1073/pnas.0903103106, PMID: 19474294
Hou J, Sigwalt A, Fournier T, Pflieger D, Peter J, de Montigny J, Dunham MJ, Schacherer J. 2016. The hidden complexity of mendelian traits across natural yeast populations. *Cell Reports* **16**:1106–1114. DOI: https://doi.org/10.1016/j.celrep.2016.06.048, PMID: 27396326
Hou J, Tan G, Fink GR, Andrews BJ, Boone C. 2019. Complex modifier landscape underlying genetic background effects. *PNAS* **116**:5045–5054. DOI: https://doi.org/10.1073/pnas.1820915116, PMID: 30804202
Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**:285–291. DOI: https://doi.org/10.1038/nature19057, PMID: 27535533
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760. DOI: https://doi.org/10.1093/bioinformatics/btp324, PMID: 19451168
Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**:833–835. DOI: https://doi.org/10.1038/nmeth.1681, PMID: 21892150
Lippman ZB, Zamir D. 2007. Heterosis: revisiting the magic. *Trends in Genetics* **23**:60–66. DOI: https://doi.org/10.1016/j.tig.2006.12.006, PMID: 17188398
Mackay TF, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**:565–577. DOI: https://doi.org/10.1038/nrg2612, PMID: 19584810
Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, Richardson MF, Anholt RR, Barrón M, Bess C, Blankenburg KP, Carbone MA, Castellano D, Chaboub L, Duncan L, Harris Z, et al. 2012. The Drosophila Melanogaster genetic reference panel. *Nature* **482**:173–178. DOI: https://doi.org/10.1038/nature10811, PMID: 22318601
Mackay TF. 2014. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics* **15**:22–33. DOI: https://doi.org/10.1038/nrg3627, PMID: 24296533
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**:747–753. DOI: https://doi.org/10.1038/nature08494, PMID: 19812666
Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, Fine RS, Lu Y, Schurmann C, Highland HM, Rüeger S, Thorleifsson G, Justice AE, Lamparter D, Stirrups KE, Turcot V, Young KL, Winkler TW, Esko T, Karaderi T, et al. 2017. Rare and low-frequency coding variants alter human adult height. *Nature* **542**:186–190. DOI: https://doi.org/10.1038/nature21039, PMID: 28146470
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:1297–1303. DOI: https://doi.org/10.1101/gr.107524.110, PMID: 20644199
Peltier E, Friedrich A, Schacherer J, Marullo P. 2019. Quantitative trait nucleotides impacting the technological performances of industrial *Saccharomyces cerevisiae* Strains. *Frontiers in Genetics* **10**:683. DOI: https://doi.org/10.3389/fgene.2019.00683, PMID: 31396264
Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A, Cruaud C, Labadie K, Aury JM, Istace B, Lebrigand K, Barbry P, Engelen S, Lemainque A, Wincker P, Liti G, et al. 2018. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**:339–344. DOI: https://doi.org/10.1038/s41586-018-0030-5, PMID: 29643504
Peter J, Schacherer J. 2016. Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast* **33**:73–81. DOI: https://doi.org/10.1002/yea.3142, PMID: 26592376
Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics* **69**:124–137. DOI: https://doi.org/10.1086/321272, PMID: 11404818
Rockman MV. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* **66**:1–17. DOI: https://doi.org/10.1111/j.1558-5646.2011.01486.x
Seymour DK, Chae E, Grimm DG, Martín Pizarro C, Habring-Müller A, Vasseur F, Rakitsch B, Borgwardt KM, Koenig D, Weigel D. 2016. Genetic architecture of nonadditive inheritance in *Arabidopsis thaliana* hybrids. *PNAS* **113**:E7317–E7326. DOI: https://doi.org/10.1073/pnas.1615268113, PMID: 27803326
Shi H, Kichaev G, Pasaniuc B. 2016. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics* **99**:139–153. DOI: https://doi.org/10.1016/j.ajhg.2016.05.013, PMID: 27346688
Speed D, Cai N, Johnson MR, Nejentsev S, Balding DJ. 2017. Reevaluation of SNP heritability in complex human traits. *Nature Genetics* **49**:986–992 . DOI: https://doi.org/10.1038/ng.3865
Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FA, Kathiresan S, Wijmenga C, Gregersen PK, Alfredsson L, Siminovitch KA, Worthington J, de Bakker PI, Raychaudhuri S, Plenge RM, Diabetes Genetics Replication and Meta-analysis Consortium, Myocardial Infarction Genetics Consortium. 2012. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genetics* **44**:483–489. DOI: https://doi.org/10.1038/ng.2232, PMID: 22446960

Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era–concepts and misconceptions. *Nature Reviews Genetics* **9**:255–266. DOI: https://doi.org/10.1038/nrg2322, PMID: 18319743

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**:5–22. DOI: https://doi.org/10.1016/j.ajhg.2017.06.005, PMID: 28686856

Wagih O, Parts L. 2014. Gitter: a robust and accurate method for quantification of colony sizes from plate images. *G3: Genes|Genomes|Genetics* **4**:547–552. DOI: https://doi.org/10.1534/g3.113.009431

Wainschtein P, Jain DP, Yengo L, Zheng Z, Cupples LA, Shadyab AH, McKnight B, Shoemaker BM, Mitchell BD, Psaty BM, Kooperberg C, Roden D, Darbar D, Arnett DK, Regan EA, Boerwinkle E, Rotter JI, Allison MA, McDonald M-LN, Chung MK, et al. 2019. Recovery of trait heritability from whole genome sequence data. *Yearbook of Paediatric Endocrinology* **16**:14.15. DOI: https://doi.org/10.1530/ey.16.14.15

Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, Hendricks AE, Danecek P, Li R, Floyd J, Wain LV, Barroso I, Humphries SE, Hurles ME, et al. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526**:82–90. DOI: https://doi.org/10.1038/nature14962, PMID: 26367797

Wangkumhang P, Chaichoompu K, Ngamphiw C, Ruangrit U, Chanprasert J, Assawamakin A, Tongsima S. 2007. WASP: a Web-based Allele-Specific PCR assay designing tool for detecting SNPs and mutations. *BMC Genomics* **8**:275. DOI: https://doi.org/10.1186/1471-2164-8-275, PMID: 17697334

Widmer C, Lippert C, Weissbrod O, Fusi N, Kadie C, Davidson R, Listgarten J, Heckerman D. 2015. Further improvements to linear mixed models for Genome-Wide association studies. *Scientific Reports* **4**:6874. DOI: https://doi.org/10.1038/srep06874

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, Amin N, Buchkovich ML, Croteau-Chonka DC, Day FR, Duan Y, Fall T, Fehrmann R, Ferreira T, Jackson AU, Karjalainen J, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**:1173–1186. DOI: https://doi.org/10.1038/ng.3097, PMID: 25282103

Yadav A, Dhole K, Sinha H. 2016. Differential regulation of cryptic genetic variation shapes the genetic interactome underlying complex traits. *Genome Biology and Evolution* **8**:evw258. DOI: https://doi.org/10.1093/gbe/evw258

Zörgö E, Gjuvsland A, Cubillos FA, Louis EJ, Liti G, Blomberg A, Omholt SW, Warringer J. 2012. Life history shapes trait heredity by accumulation of loss-of-function alleles in yeast. *Molecular Biology and Evolution* **29**:1781–1789. DOI: https://doi.org/10.1093/molbev/mss019, PMID: 22319169

Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS* **109**:1193–1198. DOI: https://doi.org/10.1073/pnas.1119675109, PMID: 22223662

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. 2014. Searching for missing heritability: designing rare variant association studies. *PNAS* **111**:E455–E464. DOI: https://doi.org/10.1073/pnas.1322563111, PMID: 24443550

Résumé

L'élucidation des origines génétiques des traits complexes reste une problématique majeure en biologie. Ces dix dernières années, les projets de reséquençage de milliers d'individus d'une même espèce ont permis d'explorer la diversité génétique au sein des génomes. Des études d'association pangénomique ont alors été initiées pour identifier les variants génétiques impliqués dans la variance phénotypique observée dans des populations étudiées. Cependant, les variants détectés n'expliquent qu'une portion plus ou moins importante de la variance phénotypique. Mieux caractériser la relation entre génotypes et phénotypes a ainsi été la base de mes travaux de thèse. Pour cela, une collection de plus de 1000 isolats naturels de *Saccharomyces cerevisiae* divers et séquencés a été une ressource clé pour disséquer dans un premier temps les origines génétiques de la variation du niveau d'expression de plus de 6000 gènes au sein de cette espèce. Dans un second temps, une stratégie de mutagénèse par insertion de transposons réalisée dans une centaine d'isolats de levure a permis d'estimer dans quelles proportions le fonds génétique impacte la variation des phénotypes.

Mots-clés : relations génotype-phénotype, génomique des populations, levure, GWAS, transcriptome, mutagénèse par insertion de transposons

Elodie CAUDAL

Équipe Variation intra-spécifique et évolution des génomes
Laboratoire de Génétique Moléculaire, Génomique et Microbiologie
UMR 7156 / CNRS, Université de Strasbourg

Doctorat sous la direction de
Joseph SCHACHERER