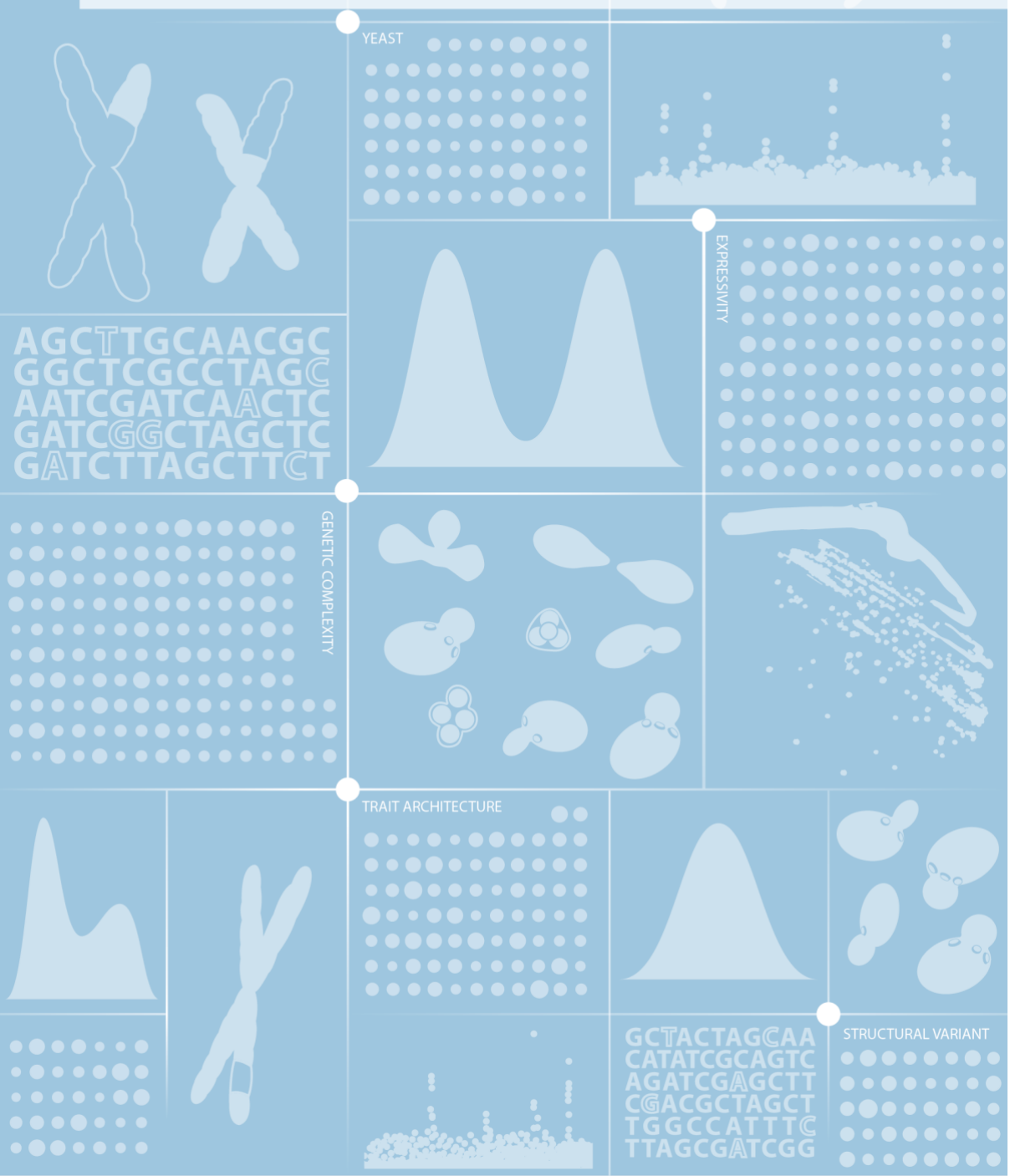
Unraveling the genetic architecture of traits in natural yeast populations

Téo Fournier
HaploTeam | UMR7156 | CNRS | Université de Strasbourg | 2019



Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

ÉCOLE DOCTORALE ED414

UMR7156 Génétique Moléculaire Génomique Microbiologie

THÈSE

présentée par :

Téo FOURNIER

soutenue le : **27 Septembre 2019**

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

Discipline/ Spécialité : Génétique

Élucidation de l'architecture génétique des traits dans une population naturelle de levures

THÈSE dirigée par :

Pr. SCHACHERER Joseph	Professeur, université de Strasbourg
Pr. DE MONTIGNY Jacky	Professeur, université de Strasbourg

RAPPORTEURS :

Pr. LANDRY Christian	Professor, université de Laval, Canada
Dr. ROCKMAN Matthew	Associate Professor, université de New York

---

AUTRES MEMBRES DU JURY :

Dr. BLANDIN Stéphanie	Chargée de recherche, université de Strasbourg
Pr. DUJON Bernard	Professeur, Institut Pasteur, Paris

Université de Strasbourg

UNIVERSITY OF STRASBOURG

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

DOCTORAL SCHOOL ED414

UMR7156 molecular genetics, genomics, microbiology

Doctoral dissertation presented by:

Téo FOURNIER

defended on: **September 27th 2019**

In partial fulfillment of the requirements of a degree of : **Doctor of Philosophy**

Discipline/ Specialty : Genetics

Unraveling the genetic architecture of traits in natural yeast populations

Ph.D. advised by:

Pr. SCHACHERER Joseph	Professor, university of Strasbourg
Pr. DE MONTIGNY Jacky	Professor, university of Strasbourg

REPORTER:

Pr. LANDRY Christian	Professor, university of Laval, Canada
Dr. ROCKMAN Matthew	Associate Professor, New York University

OTHER MEMBERS OF THE COMMITTEE:

Dr. BLANDIN Stéphanie	Researcher, university of Strasbourg
Pr. DUJON Bernard	Professor, Pasteur Institute, Paris

Abstract

Understanding the rules governing the astonishing diversity existing between individuals belonging to the same population has been one of the central role of biology. Recent years have seen the advent of genome-wide association studies to link genotype and phenotype at a population level. However, in most of the cases, an important amount of phenotypic variance remains unexplained and is called missing heritability. By combining the powerful model *Saccharomyces cerevisiae*, an elegant design borrowed to classical genetics and high-throughput strategies of genotyping and phenotyping, this work focused on increasing knowledge on the genetic architecture of traits and more precisely on some putative causes of this missing heritability at a species-wide level. Thus, we could quantify the effect of low frequency variants, obtain a global view of the genetic complexity spectrum as well as the impact of the genetic backgrounds on this complexity. Lastly, by using cutting edge long read sequencing strategies, a strong foundation for the identification of structural variants in natural population has been laid and allowed to a first view of their phenotypic effect.

Acknowledgements

I would first like to thank all the member of the committee, Pr. Christian Landry, Dr. Matthew Rockman, Pr. Bernard Dujon and Dr. Stéphanie Blandin for accepting to evaluate this work. I would also like to address special thanks to both Dr. Stéphanie Blandin and Pr. Bernard Dujon who were also part of my two mid-thesis committees and gave me great advices to finish my thesis in the best possible conditions.
The following work has been completed at the department of molecular genetics, genomics and microbiology, UMR7156/CNRS, University of Strasbourg, under the supervision of Pr. Joseph Schacherer and Pr. Jacky de Montigny.

On dit souvent que l'ambiance de travail dans le labo joue un grand rôle dans le bon déroulement d'une thèse. Je ne pense pas me tromper en disant que cela a été on ne peut plus vrai en ce qui me concerne. J'ai été vraiment chanceux de passer ces presque cinq années au sein de cette équipe. Vous êtes tous formidables et tous à votre manière m'avez permis d'arriver au bout de ce travail. Merci à tous, vraiment. Que ce soit dans lors de lab retreats ou à l'ekiden, l'haploteam est une équipe soudée où il fait bon être. J'espère que cela restera encore longtemps le cas.

Merci tout d'abord à mes co-directeurs pour l'encadrement dont j'ai pu bénéficier.
Jacky, malgré tes fonctions de doyens de la faculté des sciences de la vie, tu as réussi à trouver quelques moments pour prendre des nouvelles, prodiguer des bons conseils et toujours un petit mot pour finir avec le sourire.
Joseph, merci pour tout, pour m'avoir réellement donner goût à la science pendant ces cinq années, en partageant ta passion, tes connaissances, ta rigueur scientifique qui me serviront encore longtemps j'en suis sûr. Merci de ton encadrement, de m'avoir poussé quand il le fallait, d'avoir cru en moi et en mon travail. Moi qui avais pour plan d'arrêter après le stage de master, je n'ai jamais regretté le choix de continuer en doctorat. Ces années de thèse auront avant tout été une superbe expérience humaine et professionnelle. Les bons moments comme les moins bons, tout ça fait partie de la thèse. Merci de

m'avoir donné des opportunités de voyager pour présenter mon travail dans différents pays. Merci également pour tous ces bons moments hors du labos. Il y a tellement d'anecdotes à raconter mais je pense que les meilleures restent quand même à Prague entre le mystère du bar introuvable et le « nota bar ». Notre petit tour chez Singer était aussi une superbe expérience avec notre séjour au « beach hotel » de Minehead et un voyage en voiture mémorable …

Anne, merci pour tous les bons moments passés autour de la science ou non ainsi que pour ta patience quand je viens te voir toutes les 5 minutes en te demandant si le basecalling est terminé ou non. Au moins c'était l'occasion pour toi de demander des conseils « pour des amies ». J'espère que tu prendras conscience de la différence entre les margaritas. Un indice, l'une se mange, l'autre se boit et elle est encore meilleure « sweet and strong ».

Merci aux deux bigoudènes infiltrées de l'équipe, Claudia et Elodie. Claudia, merci de m'avoir appris la recette de la paillasse flambée, certainement une variante bretonne de la tarte flambée alsacienne. Un jour peut-être je porterai une blouse et mettrai des gants pour manipuler. En tout cas si ça arrive, ça sera grâce à toi ! Elodie, merci pour tous les fous rires et les bons moments qu'on a partagé même si certains t'ont coûté un Iphone… Aussi, et je t'en suis reconnaissant, grâce à toi je connais maintenant la marche à suivre pour installer correctement une machine à laver ! Il s'en est passé du temps depuis ton stage de M1 et je ne regrette pas d'avoir choisi « la moins dégueu », je te souhaite tout le meilleur pour ta fin de thèse ! N'oublie pas qu'en cas de coup de mou, il suffit de regarder la photo de Jing sur le silverstar pour que tout reparte !

Jing ! merci infiniment de m'avoir formé, appris R et aussi de m'avoir appris à disséquer des tétrades, il se trouve que ça m'a pas trop mal servi ! On fera un concours de dissection avec Andreas ! C'est pas tous les jours qu'on a la chance de croiser quelqu'un d'aussi passionné que toi, aussi bien par la science que par les lego ! Cette thèse c'est aussi en partie la tienne car ce projet c'est ton projet, enfin « c'est notre projet » comme dirait un dont je ne citerai pas le nom.

Les anciens du bureau, Jean-Séb et Jackson, Tic et Tac. Dire qu'on pouvait tenir des jours juste en se lançant des répliques de la cité de la peur, de Kaamelott ou encore de OSS117. Merci d'avoir toujours gardé la bonne humeur, on aura eu pas mal de fous rire et de délires plus ou moins alambiqués ! Ah la belle époque ! « *Tempora mori, tempora mundis recorda*. Voilà, eh ben ça par exemple, ça ne veut absolument rien dire, mais l'effet reste le même... ». Plein de bonnes choses pour vos aventures respectives !

Claudine, partenaire constante de bureau, merci pour toutes les discussions enrichissantes qu'on a pu avoir. Sache que je suis très heureux d'avoir pu découvrir les coins dans lesquels tu habitais près de WOOOERRRRRTH (je suis plus sûr de la prononciation) !

Sabrina, ma petite Bibitruc, merci pour tous ces fous rires. Je ne compte plus les fois où tu es venue paniquée dans le bureau parce que tes résultats étaient « du brun » alors que tu avais juste oublié une virgule ou une variable dans ton code R ! bon courage pour la fin ! On se recroisera peut-être du côté de chez les chtis ! ;)

Omar, désolé de t'avoir fait t'arracher les cheveux sur des questions d'héritabilité expliquée… Merci aussi pour les bons souvenirs, notamment la session jeux en amphi Maresquelle ! Plein de courage pour la fin de ta thèse.

Merci Fabien pour tous les bons moments, merci à toi qui s'assure qu'un café est toujours prêt peu importe l'heure et les circonstances ! Tout le meilleur pour la suite.

Emna, merci d'avoir toujours le sourire et je te souhaite bon courage dans tes aventures remplies de kombucha (et de plein d'autres choses aussi je l'espère) !

Abhishek, thank you for all the discussions, I'm sure that you will make good use of the Pixl when it will arrive ! And good luck with your lifting competition, hope that this excuse not to run this year's Ekiden will be worth it !

Chris, we're both united by the love of gin, that was a great pleasure to work with you for these past months. I hope that your little cat will stop giving a hard time and I wish you the very best for your future which I have no doubt will be bright ! Stay as you are Mr German !

Bon, Andreas, je te laisse les clés du rotor et de la phenobox en partant, prends en soin ! je suis certain que cette thèse que tu vas commencer t'emmènera très loin et qu'elle sera riche en succès ! Tu as tout ce qu'il faut pour réussir, profites-en.

Merci à tous ceux qui ne sont pas du labo mais qui ont joué un grand rôle dans mon bonheur personnel. Tout d'abord merci à Antoine, Emma, Damien et Thomas, vous êtes des amis en or avec qui on ne peut que passer des bons moments. Que ce soit au badminton, au RU, autour d'un verre ou à un concert déjanté d'Etienne de Crécy, c'est que du bonheur ! Merci à Marie, ma p'tite Gnomette parisienne qui a toujours su être là pour moi quand il le fallait. Tu ne m'as jamais quitté depuis le lycée et j'espère que demain n'est pas la veille !
Merci aussi à tous les autres qui ne sont pas cités mais qui ont partagés avec moi des aventures tout au long de ces années.

Merci bien sûr à ma famille qui a toujours su me montrer un soutien indéfectible tout au long de mes études. Vous avez su me motiver, me faire comprendre l'importance du travail et plus généralement m'inculquer des valeurs saines que cela soit au travail ou dans la vie de tous les jours, qui je l'espère vont encore me porter loin. Merci du fond du cœur ! La distance nous sépare mais c'est toujours un immense plaisir de tous vous retrouver : Papa, Maman autour d'un bon mojito fait maison près de la piscine à Marquefave ou dans les pentes escarpées des Pyrénées ; Boris, pour découvrir de nouveaux resto sur Paris ou pour jouer a des jeux débiles ; Lara, quand la nostalgie Vosgienne me gagne, je sais que je peux y retourner à tout moment.

Enfin, mon plus grand merci revient tout naturellement à toi Nathalie, qui, avec ta gentillesse et ton amour sans faille m'ont permis de survoler ou de mettre de côté tous les petits tracas et autres problèmes qui se sont posés à moi. Je suis conscient de la chance que j'ai de t'avoir à mes côtés et je ne te dirai jamais assez à quel point je t'en suis reconnaissant. Je t'aime mais ça tu le sais déjà.

Table of Content

STATE OF THE ART

Genotype-phenotype relationship

Highlighting the factors controlling the variation present in natural populations is a keystone that has been gathering efforts of several generations of biologists for more than 150 years. Indeed, in the mid 19th century, a Moravian friar, Gregor Mendel laid the foundation for the dissection of the underlying genetic basis of traits by setting out to understand the principles of heredity (Mendel, 1866). His work paved the way for modern genetics. Johannsen proposed in 1911 that genotype and phenotype are two distinct abstraction levels working together : '*the qualities of both ancestor and descendant are in quite the same manner determined by the nature of the "sexual substance"—i.e. the gametes—from which they have developed'* (Johannsen, 1911). He then coined this 'sexual substance' to the genotype and the 'qualities' of the individuals to the phenotype. However, despite several generations of geneticists having tackled the question, fully grasping the intricacies of the relationship between genotype and phenotype still remains strongly challenging. The astonishing amount of phenotypic variation observed between individuals of every natural population is tightly linked to the underlying genetic variation. A better comprehension on how the two are connected is one of the main drivers in a wide spectrum of domains, including human genetics, etiology but also diagnosis and prognosis of complex diseases, evolutionary biology, quantitative genetics and genomics.

'As the great botanist Bichat long ago said, if everyone were cast in the same mould, there would be no such thing as beauty.'

--- Charles Darwin, The Descent of Man, and Selection in Relation to Sex, 1871

Decomposition of a trait

In order to try and unravel the phenotype-genotype relationship, one might first understand how the two are related to each other and what defines them. The phenotypic variance is the result of the sum of several variances: the genetic, the environmental and the one generated by the interaction between genotype and environment. A good illustration of all those effects is seen with human skin color. The pigmentary phenotype of skin is complex both at the genetic and physiological level. However, a fair number of genes have been found to impact skin pigmentation level (Rees, 2003). Exposure to ultra-violet radiation is an environmental factor having a major impact on skin pigmentation level. I know for a fact that a PhD student writing his thesis manuscript indoor will have a lighter skin color than the one who spend several weeks under the sunrays. Finally, on a more serious note, one example allowing to illustrate how genes and environment can jointly contribute to phenotype is the susceptibility to skin cancer (Gupta et al., 2016). Indeed, individuals with naturally darker skin tones, although not being immune from it (Lozano et al., 2012), will have a reduced risk of developing skin cancer due to long and or repetitive UV-light exposure.

The genetic variance can in turn be broken down into additive and non-additive effects. Variants can act additively when the sum of their effects equals their combined effects. Conversely, variants that act in a non-additive way will have a different combined effect compared to the sum of their individual effect. Non-additive phenomena encompass interactions both intralocus *i.e.* dominance and interloci, *i.e.* epistasis. The environmental part is *de facto* very variable and difficult to take into account. One main advantage of using model organisms is that this becomes easy to control because all experiments can be carried out under standardized conditions thus allowing to disregard this variance. The only remaining source of variance, outside genetic factors is then the experimental errors *i.e.* the noise (Bloom et al., 2013).

The genetic architecture of traits

To fully uncover how traits are shaped by the genotype, it is of prime interest to fully unveil their genetic architecture. Genetic architecture of traits depicts the attributes linked to genetic variation which will induce a phenotypic variation in a population (Mackay, 2001; Timpson, Greenwood, Soranzo, Lawson, & Richards, 2018). More precisely, it encompasses the number of variants governing a given phenotype, the extent of their effect on the said phenotype, the frequency at which they are found in the population and finally the interactions they may have between each other and also with their environment. It seems obvious that each trait, depending on their complexity will have its very own genetic architecture. The simplest phenotypes are controlled by only one gene and follow a Mendelian inheritance. In this category can be found traits related to anatomy like shape of the ears (Gordon et al., 2013), metabolism such as lactose intolerance (Swallow, 2003) and in many genetic disorders, *e.g.* cystic fibrosis (Ratjen et al., 2015), Huntington's disease (Bates et al., 2015), Retinitis pigmentosa (Parmeggiani et al., 2011) and many other. In fact, more than 5,000 monogenic traits have been mapped to a single gene (https://www.omim.org/). Oppositely, other traits are highly polygenic with hundreds or even thousands of genes responsible for it such as height which is often used as the archetype of extreme polygenicity. However, studies highlighted the fact that a broad range of other traits going from diabetes to body mass index or autoimmune diseases (Boyle et al., 2017; Shi et al., 2016) are also highly polygenic thus deciphering the complete genetic architecture of those traits is highly challenging. Getting a deeper understanding of this genetic architecture is of prime interest to better grasp the biological foundations of those traits but also to improve diagnosis and even prognosis of certain diseases by giving a predictive power based on the genotype. However, this obviously raises some ethical questions and dilemma especially for uncurable diseases, as Wexler stated about Huntington's disease: "*Do*

you want to know how and when you are going to die, especially if you have no power to change the outcome? Should such knowledge be made freely available? How does a person choose to learn this momentous information? How does one cope with the answer?" (Wexler, 2018).

Mapping the causal variants

To unveil the genetic causes underlying the natural variation observe among traits, strategies have been developed both for human and model organisms. The aim of these techniques is very straightforward: mapping all the loci that induce a phenotypic variation for a given trait in a population. Quantitative Trait Loci (QTL) correspond to genomic regions that are involved in the variation of a quantitative phenotype in a population. Individuals used in a mapping population need to have segregating markers along their genomes. One prerequisite for QTL mapping is then to genotype and phenotype each individual. Although several techniques can be used like microarray-based genotyping, with the plummeting cost of DNA sequencing, genotyping by whole genome sequencing is now becoming the standard method and allows to detect every discriminating markers between each individual *e.g.* Single Nucleotide Polymorphisms (SNP) or small insertions/deletions (Indels). There are two main approaches to detect QTLs, either by linkage analysis or genome-wide association studies. Both approaches have several alternatives but not all of them will be discussed in the following sections. We will discuss the advantages but also the limits for both mapping strategies.

Linkage analysis

Linkage mapping aims to identify QTLs based on a mapping population coming from two genetically distinct parental lines. After crossing, the offspring is individually genotyped and phenotyped in a quantitative manner. Recombination events during meiosis shuffles parental genetic markers along the genomes of the resulting progeny. The following step results in testing for each marker if all the progeny having one of the parental alleles have a significantly different phenotype from the progeny carrying the other parental allele. If this is the case, a QTL will be present around the locus of this particular marker (Figure 1). A key factor

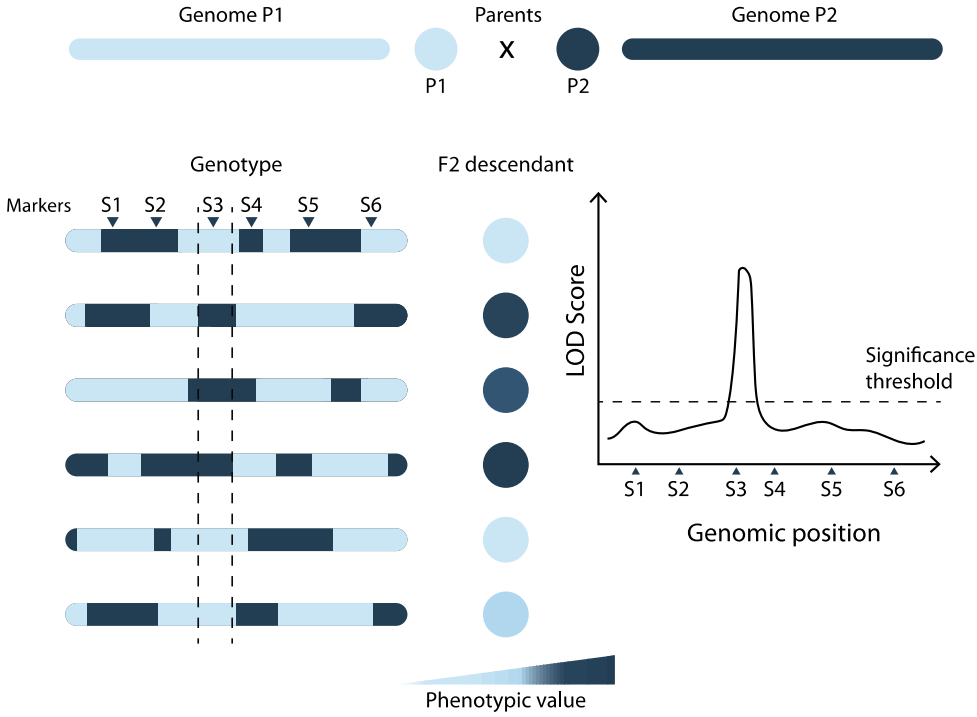


Figure 1. Principle of linkage mapping
To quantify to which extent a linkage is found between a marker and a phenotype, the logarithm of odds (LOD Score) is calculated and when above the significance threshold, a QTL is detected.

influencing the size of the mapped genomic regions is the number of markers and how they are distributed between the parental lines. The second important variable is the recombination rate in the QTL region. The higher the recombination rate is (Figure 2), the more crossing over there will be in the progeny, the shorter the genomic blocks between recombination points will be, the smaller number of individual markers will be in those blocks. Classical approaches use F2 generation as mapping population but depending on the organisms used, this can lead to the identification of QTL spanning several megabases if the recombination rate is low.

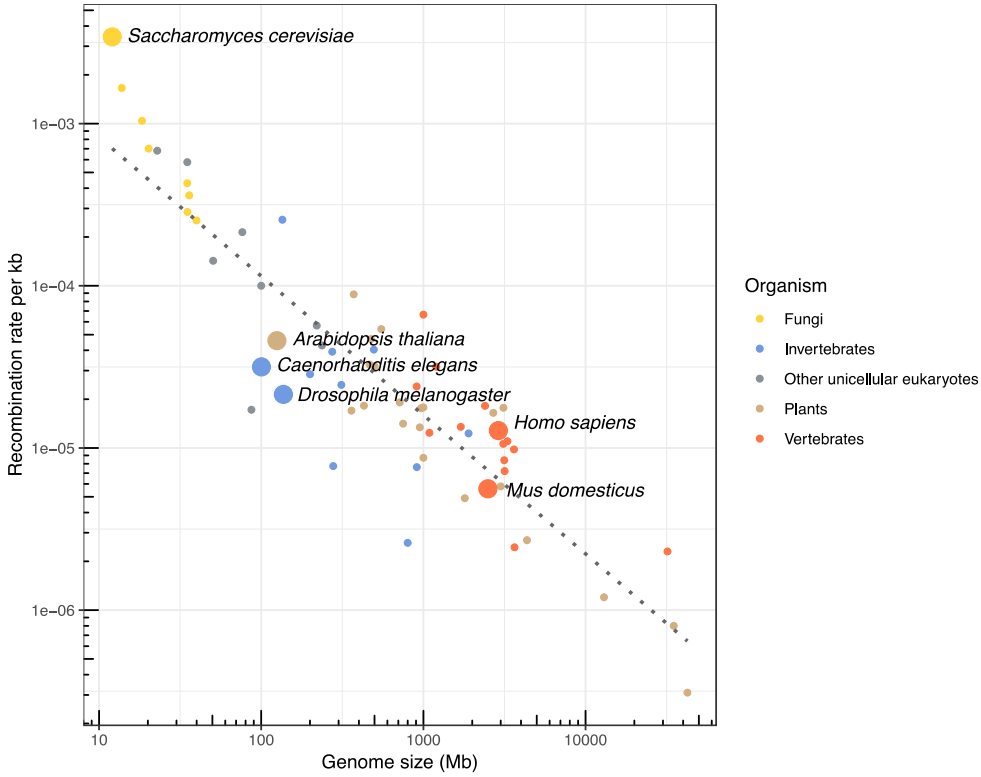


Figure 2. Recombination rate in multiple organisms
Recombination rate per kb is inversely correlated to the genome size (in log scale), Pearson'r =0.90, p-value=6.3e^{-25}. The name of different model organism are displayed and colors represent different phyla. *S. cerevisiae* stands as an outlier having a recombination rate above the expected one given his genome size. Adapted from Lynch (2006).

Alternatives to the classical QTL approach exists (Figure 3). One of them is the Bulk Segregant Analysis (BSA) (Magwene et al., 2011; Segrè et al., 2006). For BSA, individuals coming from a cross are still phenotyped separately and the one displaying extreme phenotypes *i.e.* in the tails of the distribution, are pooled in order to be genotyped as a bulk (Figure 3A). Because individuals with similar phenotype will tend to display similar genotypes, loci involved in the phenotype of interest will co-segregate in individuals displaying the same phenotype. Therefore, allele **f**requencies will be measured in this pool and strong deviation from random allele segregation (allele frequency of 0.5) will map the QTLs position (Figure 3A).

In order to reduce this extensive linkage disequilibrium between markers, successively backcrossed individuals or recombinant inbred lines can be used (van Swinderen et al., 1997) (Figure 3B). Another workaround to reduce the size of mapped region is to increase the size of the mapping population and/or using model organisms with high recombination rate *e.g. Saccharomyces cerevisiae* (Lynch, 2006) (Figure 2).

To go even further in scale and in resolution, x-QTL methods have been developed (Figure 3C) where a very large number - usually around a million - of individuals coming from recombinant inbred lines are phenotyped in bulk to keep only the extreme phenotype in a complex trait e.g. with flow cytometry. Then, as for BSA, sequencing of the pooled individuals will detect any deviation in allele frequencies to pinpoint causal QTLs. This method has now been used in yeast for a few years (Ehrenreich et al., 2010) but very recently, in a tour de force, has been successfully applied to *C. elegans* (Burga et al., 2019).

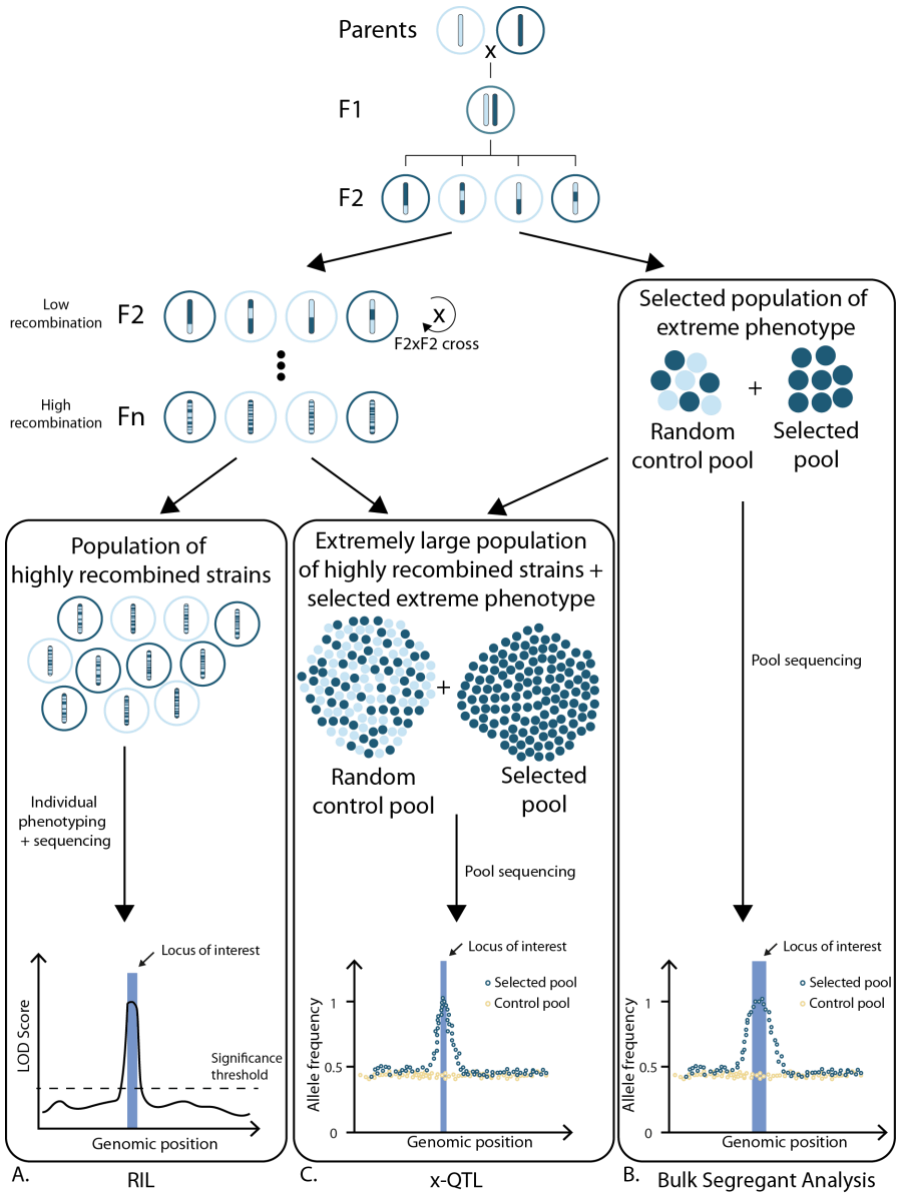


Figure 3. Alternative strategies to linkage mapping
Different type of mapping populations can be used. Either A. a population after numerous inbred crosses to generate highly recombinant individuals, B. a pool of F2 individuals with extreme phenotypes, or C. a combination of the two previous by selecting extreme phenotypes from highly recombined individuals.

Once QTLs are mapped, there is still a long-way to be able to pinpoint the exact genetic variants explaining the observed phenotypic variance. Indeed, to date, thousands of QTLs have been mapped but only a fraction has been narrowed down to a Quantitative Trait Gene (QTG) and even less at the nucleotide resolution *i.e.* Quantitative Trait Nucleotide (QTN). Moreover, because of the tedious process of functional validation, variants to validate a QTL are often prioritized leading to an ascertainment bias (Rockman, 2012): only the QTL with the largest effect and corresponding to missense or non-sense variants in candidate genes will be investigated and subsequently validated. However, thanks to high throughput genome editing technique, we know that a lot of functional variation is linked to non-coding regions (Jakobson and Jarosz, 2019; Sharon et al., 2018) and that synonymous variants happen to often have a strong phenotypic impact by modifying codon bias usage thus impeding the optimal translation speed ultimately leading to problem in the folding of the protein (Jakobson and Jarosz, 2019; She and Jarosz, 2018).

Linkage mapping has been applied to several organisms such as yeast (Steinmetz et al., 2002), worm (Rockman and Kruglyak, 2009) or plants (Brachi et al., 2010), but also using various phenotypes such as expression level to map eQTL (Brem and Kruglyak, 2005; Rockman et al., 2010): QTL that are involved in the modification of expression level of a transcript. Over the years, plethora of QTLs have been found. In yeast, to date, 284 QTNs have been functionally validated to the nucleotide level (Peltier et al., 2019) revealing the extensive work for bridging the gap between genotype and phenotype using this model organism for the past 20 years.

One argument demeaning the use of classical linkage analysis is that this technique tends to only map variants with a large phenotypic effect. However, the use of more powerful linkage techniques like x-QTL allow to map variants with smaller effects.

Another drawback of linkage mapping is that it can only highlight QTLs between two parental lines. With the extensive phenotypic and genetic diversity observed inside a population, QTLs highlighted by linkage mapping between several parental pairs taken across the population can differ significantly. Nevertheless, techniques do exist to go beyond this biparental model with joint analysis of multiple related biparental families (Jamann et al., 2015) but requires a lot of time and efforts to generate all the corresponding mapping populations. This allows to detect shared QTL between several lines but also to test for genetic interaction between QTL depending on the genetic background as shown by a massive QTL mapping between 16 lines in a round robin design with each of the cross having a mapping population of a thousand individuals (Bloom et al., 2019).

Genome-wide association studies

The target of Genome-Wide Association Studies (GWAS) is to assess the effect of an allele by finding statistical difference between the average phenotypic value of a group of individuals with and another without it and repeat this process for each

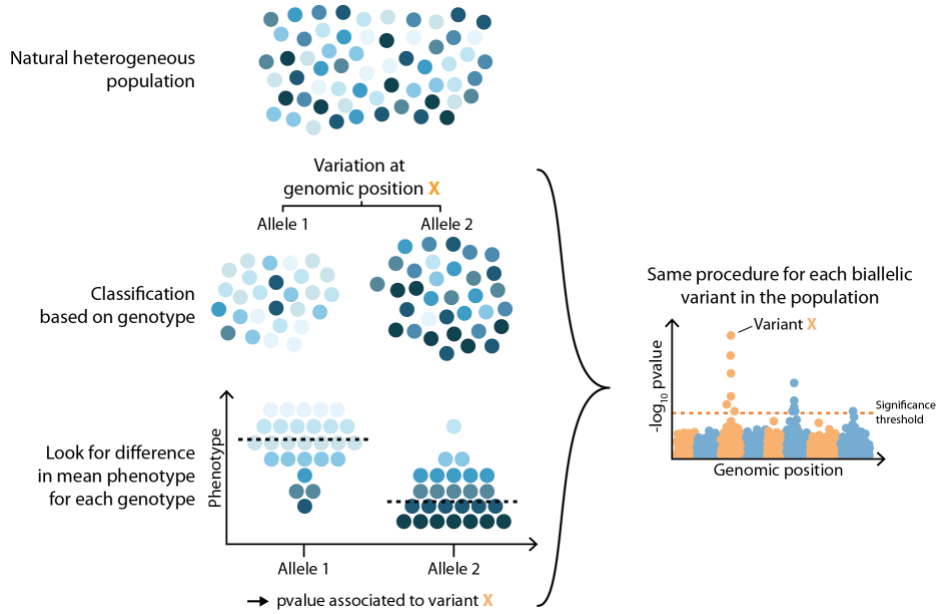


Figure 4. Principle of GWAS
For each discriminating variant of the genome between individuals of a natural population, statistical phenotypic difference will be assessed between the group of individual with one or the other allelic version. Permutations allow for a significance threshold to be decided. Any variant having a p-value above (smaller value) this threshold will be significantly associated with the phenotype.

discriminating polymorphic site along the genome. GWAS thus aim at finding trait associated variants among all the biallelic sites in a population (Figure 4). Unlike linkage mapping which uses genetic recombination during meiosis, association studies rely on one side to ancestral recombination but also more generally to the historical evolutionary forces that drives natural population *i.e* balance between mutations and selection. Therefore, it takes advantage of the natural genetic and phenotypic variation and does not require the need to generate a *de novo* mapping population. Moreover, GWAS can potentially access all variants present in natural

population and not be restricted to a small amount of genetic backgrounds, thus having the eagerness to shed light on the genetic architecture of traits. However, we will discuss how GWAS despite having greatly improved our global understanding of the genotype-phenotype relationship might still not be the goose that lays golden eggs.

GWAS has been first introduced in human genetics almost 15 years ago (Hirschhorn and Daly, 2005) but is now also applied in several model organisms such as yeast (Peter et al., 2018), plants (Alonso-Blanco et al., 2016; Atwell et al., 2010; Seymour et al., 2016), mouse (Flint and Eskin, 2012; Gonzales et al., 2018) or worm (Cook et al., 2017) as well as non-model organisms with crops (Wang et al., 2017) and cattle (Higgins et al., 2018). GWAS in human in the last 15 years has yielded plethora of variants strongly associated (p value threshold of $5x10^{-8}$) with complex traits. The GWAS Catalog (https://www.ebi.ac.uk/gwas/) (MacArthur et al., 2017) has inventoried to this date nearly 90 000 trait associated SNPs. The real power of GWAS resides in the fact that its detection power depends on the sample size used. Indeed, the more samples are drawn from the initial population, the higher the chances are of discovering more phenotypically relevant variants for some traits. GWAS allows to map common variants having small effect but also the few common variant responsible for an important variation in phenotype.

GWAS, albeit a powerful tool, also has a fair number of limitations and prerequisites for its use. Due to historical reasons, both geographic or demographic, a population can be divided in subpopulations and encounter new sources of stresses, environments or selection forces, thus evolving separately one from the other and retaining specific genotypic patterns. This phenomenon where subpopulations can be differentiated by comparing genotypes is referred to as population structure or population stratification (PS). It is the result of non-random mating, so allele

frequencies will differ from one subpopulation to another. One consequence of PS is that it might result in spurious associations between genotype and phenotype. Thus, when conducting GWAS, care must be taken to avoid or correct for PS (Hellwege et al., 2017).

Theoretically, GWAS aims at uncovering all the variant contributing to a phenotype. Yet, when comparing the phenotypic variance explained by all significantly associated variants with the genetic variance, also called broad-sense heritability (H^2), a significant discrepancy is found. This difference has been coined as "missing heritability" (Maher, 2008; Manolio et al., 2009). The textbook example for missing heritability is human height. This phenotype has long been estimated to be about 80% heritable using twin studies (Silventoinen et al., 2003) but recent reevaluation put the heritability at around 60% (Speed et al., 2017). In 2008, only 40 genetic variants had been significantly associated with height and explained only about 5% of the heritability (Manolio et al., 2009). Six years later, the number of associated SNPs was around 700 and explain roughly 20% of broad-sense heritability (Wood et al., 2014). In 2018, by using meta-analysis of several GWAS, with a sample size of 700,000 individuals, 3290 height associated SNPs still explain only around 25% of the broad-sense heritability (Yengo et al., 2018). Although reducing, the gap between explained and observed heritability is still significant. In the last decade, a lot of efforts have been put in investigating the potential sources of this missing heritability. This search is important because this would first yield a better understanding of the biology and etiology of traits. Besides, it would also allow for a gain in predictive power for personalized medicine. While we are able to infer with near perfect accuracy phenotype from genotype in some experimental crosses (Bloom et al., 2013; Hallin et al., 2016; Märtens et al., 2016), gaining this level of prediction within natural population is far from being easy. Looking at all the sources of missing heritability comes back to understanding and measuring all the parameters having a role in the genetic architecture of traits (Eichler et al., 2010).

Potential sources of missing heritability

The role of rare variants

One of the major points taken from the populations genomics studies is that when looking at the allele frequencies of all the variants, a strong bias towards lower allele frequencies is detected (Auton *et al.*, 2015; Peter *et al.*, 2018) implying that rare variants are (very) common. In human for example, about 90% of all the genetic variants detected have a Minor Allele Frequency (MAF) lower than 0.05 (Auton *et al.*, 2015), meaning that they are present in less than 5% of the individuals. This observation is not idiosyncratic to human as the same ascertainment is true for other model species such as yeast (Peter et al., 2018) (Figure 5). This implies the very

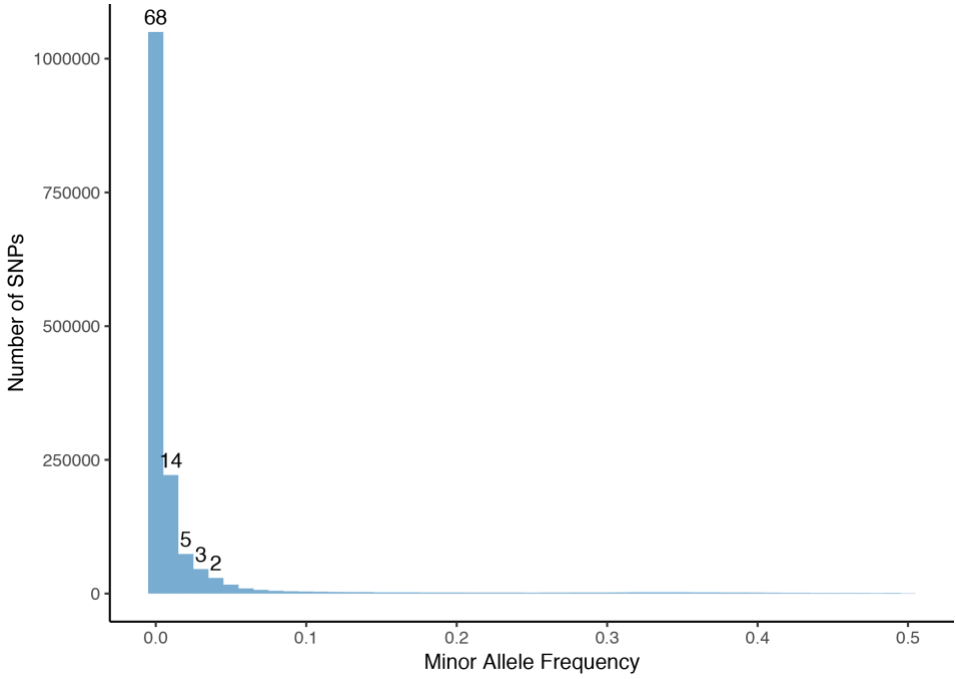


Figure 5. MAF of variants in 1,011 yeast isolates
92% of the nucleotidic variants in yeasts have a MAF lower than 0.05. Percentage of total number of variants is indicated above each bar for variants below a MAF of 0.05. Adapted from Peter et al. (2018).

counterintuitive fact that most of the genetic variation observed in a population is constituted by rare variants. As the definition of rare variants in terms of MAF is variable throughout literature, hereinafter, we will refer to common (MAF > 0.05), low frequency (MAF < 0.05) or rare (MAF < 0.01) variants.

Although low-frequency and rare variants are known for a long time to be sources for an important number of mendelian disorders (Gibson, 2012), they also play a role in numerous common diseases or other complex traits. Variants having a strong negative impact on the phenotype are expected to be found at low allele frequencies because of selection pressure against them. However, because of high mutation rates at the population level, purifying selection is not strong enough to remove all deleterious mutations. Moreover, selection might not act as strongly for recessive variants or mutations involved in cryptic variation *i.e.* has no phenotypic effect in one condition but can be deleterious (or beneficial) in another.

In yeast, out of the 284 QTNs previously detected by linkage mapping, 150 are present at a low frequency in the initial population of 1011 isolates (Peltier et al., 2019; Peter et al., 2018) (Figure 6). When looking at only one cross in linkage mapping, most of the phenotypic variance might be explained by few of those SNPs. However, when looking at the population level, even though they do have large effect size, they do not explain an important part of variance because heritability relies both on effect size and allele frequency. A good example can be 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.

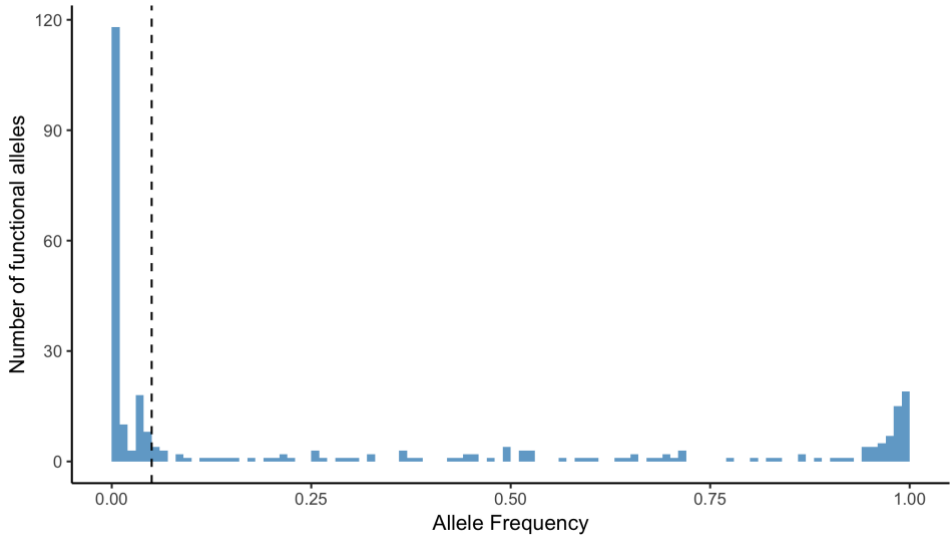


Figure 6. Allele frequency of the known QTNs in yeast
150 QTNs are found at a low frequency in the population. Each bar has a width of 0.01 allele frequency. Dotted line shows the 0.05 allele frequency threshold. Adapted from Peltier et al. (2019).

Unfortunately, rarer variants in the population are difficult to detect with GWAS (Gibson, 2012; Manolio *et al.*, 2009). This is because a strong relationship does exists between sample size and statistical power to detect alleles with lower allele frequencies (Gorlov et al., 2008), meaning that the rarer a variant is in the population, the bigger the sample size has to be for detecting it. Moreover, GWAS significance not only depends on the frequency of the variant but also on the effect size, so that small effect variants won't be detected easily even with very large sample sizes. On top of that, human genotyping method often relies on SNP arrays which by design can only capture variants down to a certain frequency in the population but this problem is now alleviated with whole exome sequencing becoming standard. Accounting for low frequency and rare variants becomes easier in human as with the always growing sample size and the possibility to perform meta-analysis using data coming from different datasets, sufficient detection power can be achieved. Moreover, whole exome or even whole genome sequencing allows to account for the entire allelic spectrum of a population.

Numerous efforts have been made towards accounting for rare variants with different experimental designs (Lee et al., 2014; Zuk et al., 2014). Recent studies really shed light on the real phenotypic impact that low frequency and rare variants have at the population-level and to which extent they contribute to missing heritability. Indeed, recent mapping studies focused on assessing the role of these variants and all of them converged to the same conclusion: the effect of rare variants is pervasive across the population with the presence of a high number of them having a strong phenotypic impact. This allows them to account for an important part of the missing heritability (Bloom et al., 2019; Fournier et al., 2019; Wainschtein et al., 2019). However, rare variants alone are not sufficient to completely bridge the gap between observed and explained heritability, suggesting that part of it is still hidden under the effect of other mechanisms.

Non-additive effects

GWAS mainly works with the assumption that phenotypes are under complete additive control. However, the phenotypic effects of the causal variants may not be necessarily combining their effect under an additive model but may instead interact in a synergistic manner resulting in phenotypes that are either higher or lower than expected under complete additivity. Genetic interactions can either take place between two different loci (epistasis) or inside the same locus (dominance) in a diploid context. Study of yeast crosses showed that although traits were mostly controlled by additivity, about a third of the genetic variance is linked to non-additive effects (Bloom et al., 2013, 2015; Fournier et al., 2019).

First introduced by Mendel in the mid 19th century with its third law of inheritance (Mendel, 1866), dominance and recessiveness are genetic concepts that nowadays seem obvious. Most of the mutation inducing loss of function are recessive for the reason that most of the time, the activity of the gene product from the other functional allele is sufficient to ensure proper functioning of the biological pathway. Exception is made in the case of haploinsufficiency where the absence of one of the two copies leads to abnormal phenotypes (Deutschbauer et al., 2005). However, these loss of function mutations represents only a fraction of the mutational effect spectrum.

To decompose genetic variance and more precisely the part played by non-additive effect, a successive intercrossing of two divergent yeast strains over 12 sexual generations has been achieved in order to first reduce linkage between markers of the two strains. Then, 86 *MAT*a and 86 *MAT*α haploids coming from this twelfth generation were crossed in a systematic pairwise manner to generate a large hybrid diploid panel of 6,642 individuals. After phenotyping this set on 9 growth conditions, a complete decomposition of traits variance in diploid has been performed. They found that on average, dominance accounted for 10% of the total phenotypic variance and 9% was accounted for by epistatic interactions (Hallin et al., 2016).

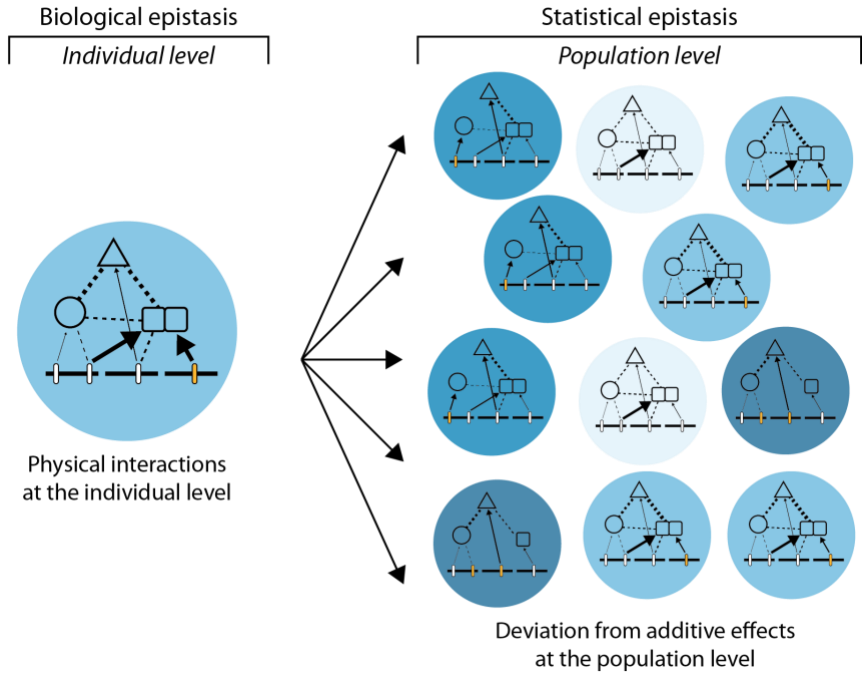


Figure 7. Biological and statistical epistasis
Biological epistasis arises from interaction in an individual whereas statistical epistasis comes from a population phenomenon with varying genotypes modifying interactions in a background dependent manner. DNA variants are represented by vertical bars, biomolecules by circle, triangle and squares, the interactions by dashed lines. Final phenotype is represented as background color Adapted from Moore & Williams (2005).

Bateson first coined the term "epistasis" in 1909 to explain observed deviation from the expected Mendelian inheritance (Bateson, 1909). From an etymological point of view, epistasis means standing upon suggesting characters stacked on each other from which you had to remove the top one *i.e.* the epistatic character, to reveal the underlying one *i.e.* the hypostatic character. This definition is now known as the biological epistasis. Since then the definition has seen a lot of remodeling and a widely used definition is the one of Fisher who takes a statistical approach to define epistasis. He explained that epistasis accounts for the deviation from additivity in a

linear model (Fisher, 1918). However, both *i.e.* biological and statistical definition of epistasis are important because of the core difference between the two. On the one hand, biological epistasis is taking place at the cellular level and reflects the phenotypic effect of physical molecular interactions between proteins and/or molecules (Figure 7A). On the other hand, statistical epistasis can only happen in a population because it compares the relationship between genotypes and phenotypic variation to a linear model (Figure 7B), although other contexts of non-linear genotype-phenotype maps can exist (Sailer and Harms, 2017). This main difference is also what is linking them (Moore and Williams, 2005): biological epistasis in several individuals will lead to statistical epistasis at the population level (Figure 7). However, two molecules that do not have a direct physical interaction may also exhibit epistatic interactions *e.g.* via different biological pathways.

Although examples of epistasis are known to have a phenotypic impact since Bateson's work, it is only in this last decade that we fully grasp the prevalence of such interactions at a genomic level by having a view of the global genetic interaction networks. Genome-scale characterization of genetic interactions has been performed on model organisms using either knock-down by RNA interference for *Caenorhabditis elegans* (Byrne et al., 2007; Lehner et al., 2006) and *Drosophila melanogaster* (Billmann et al., 2016; Horn et al., 2011) or knock-out approach with synthetic genetic array technology for *S. cerevisiae* (Costanzo et al., 2010, 2016). The pairwise cross of the yeast deletion mutant collection lead to the generation of about 23 million double mutants. This achievement allowed to detect almost one million genetic interactions with about 550,000 negative and 350 000 positive interactions (Costanzo et al., 2016), revealing how important genetic interactions are in shaping the phenotypic landscape of a cell. This digenic interaction screening has also been useful to understand the functional wiring of the cell. However, this gave no information about higher order epistasis that include three or more genes. Recently, a glimpse into the global trigenic interaction network has been revealed

(Kuzmin et al., 2018) with a subset of 200,000 triple mutants and estimated that trigenic interactions might be as much as 100 times more prevalent than digenic interactions and is highly interconnected with it. Extensive higher-order epistasis *i.e.* interactions between three, four or even more loci, have been characterized in yeast (Mullis et al., 2018; Taylor and Ehrenreich, 2014, 2015) as well as in other species (Hanlon et al., 2006; Pettersson et al., 2011).

When investigating non-additivity effects at the population level, only variation that naturally occur within this population can be taken into account. As already discussed before, in most of the natural populations, a strong bias towards lower allele frequencies is observed. However, having a sufficiently high MAF is critical when it comes to detecting interactions in a population in order to achieve sufficient statistical value. Moreover, if a given variant shows epistatic or dominance effect at a given locus but is at a very small allele frequency, the phenotypic variance at the population level might just appear as additive. So non-additive effect might in fact just add a "random perturbation" without converging towards positive or negative interactions (Mackay, 2014; Paixão & Barton, 2016) unless they are under strong selection and gain sufficient allele frequency in the population to be detected through GWAS. Conclusions about the relative contribution of non-additive effects on complex traits variation suggests that in an human outbred population, epistasis and dominance do not play a significant role in phenotypic variation, except for some cases in the ABO locus for effects on factor VIII and Willebrand factor (Mäki-Tanila and Hill, 2014; Zhu et al., 2015).

Structural variants

SNPs are far from being the only form of genetic variation present in the genome. Structural variants (SV) come in all shapes and sizes: this includes deletions, insertions, copy number variation (CNV), inversions and translocations (Figure 8). Altogether those large-scale variants make up for an important part of genetic variation between individuals from the same species. Indeed, human genomes are constituted only of ~0.1% of SNPs but of ~1.5% of structural variants (Pang *et al.*, 2010). Thus, their phenotypic effect at both the individual and population level can be extensive. SVs contribute to human genomic variation leading to fitness advantages (Radke and Lee, 2015) but are obviously well known for being the cause of rare and common diseases (Conrad et al., 2010; Sudmant et al., 2015; Weischenfeldt et al., 2013) as well as drivers of oncogenesis (Beroukhim et al., 2010) by activating oncogenes and inactivating tumor suppressors (Zack et al., 2013). SVs can be on one hand copy number neutral with inversions and balanced translocations and copy number variant with insertions, deletion or imbalanced translocations. Although being known for a long time, the role of structural variants in the phenotypic landscape has been overlooked for a very simple reason: detecting them remains challenging. Nevertheless, accessing CNVs is easily achievable through coverage analysis of short-read sequencing strategies like Illumina. Gross chromosomal rearrangements such as translocations can be directly detected by looking at karyotypes. Having a systematic and precise detection is hard to achieve even through the use of the most recent advances in sequencing technologies. Detecting SVs relies heavily on the ability to map the exact start and end breakpoints for the variant in question. Although SV detection through the use of short read length is feasible, recent advances in sequencing technologies with long reads such as Oxford Nanopore (Jain et al., 2016) facilitates breakpoints detection with the adapted detection tools (Cretu Stancu et al., 2017; Gong et al., 2018; Sedlazeck et al., 2018). Moreover, as longer reads also allow for highly contiguous *de novo*

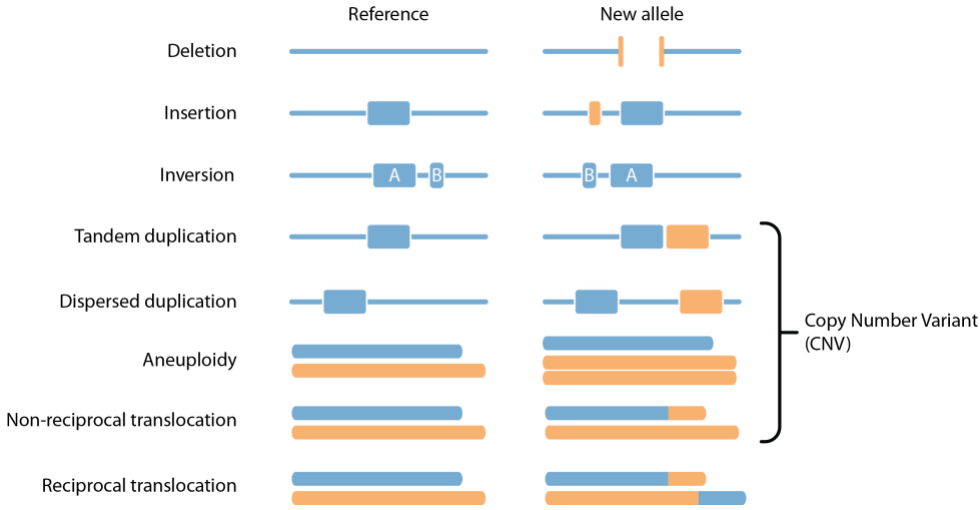


Figure 8. Types of structural variants
A schematic overview of the structural characteristics of SVs with balanced SVs encompassing Deletion, Insertion, Inversion and reciprocal translocations. On the other hand, unbalanced SVs also called CNVs arise with tandem or dispersed duplications, aneuploidies and non-reciprocal translocations.
Modified from Hurles et al. (2008)

genome assemblies, it is then possible to accurately map the SVs after assembly (Biederstedt *et al.*, 2018) and not directly with the raw reads mapped against a reference genome. However, as fast as this sequencing technologies is advancing, to this date, it is still prone to systematic errors especially in homopolymers (Jain et al., 2018) and requires polishing steps by more accurate reads such as Illumina to produce good quality assemblies (Fournier et al., 2017; Istace et al., 2017).

Out of the many phenotypic outcomes that structural variants can have (Weischenfeldt et al., 2013), chronic myeloid leukemia stands as a good illustration on how translocations can have dramatic phenotypic effects. Patients diagnosed with chronic myeloid leukemia carry in 95% of the cases the Philadelphia chromosomes which is characterized by a reciprocal translocation between chromosomes 9 and 22 (Druker et al., 1996). At the breakpoint, genomic recombination results in the

juxtaposition of *BCR* and *ABL1* which generates the *BCR-ABL1* fusion gene. The Bcr-Abl protein has a tyrosine kinase activity that is deregulated which in turn will lead to oncogenesis (Lugo et al., 1990). Although many reciprocal translocations can remain phenotypically silent in an individual, problems may arise during meiosis where only a fraction of the gametes will be euploid. Thus, translocation carriers will display important problems of fertility and/or risk of transmitting severe phenotypic conditions to their progeny who will carry partial aneuploidies.

A reciprocal translocation between chromosome VIII and XVI in *S. cerevisiae* confers a sulfite resistance (Pérez-Ortín et al., 2002). Interestingly, this translocation is restricted to a subpopulation of strains used in winemaking, due to a selection pressure on this phenotypic advantage as sulfite is an antioxidant and antimicrobial compound widely used in winemaking processes. It has been shown that the translocation happened in micro-homology regions in the two promoters of the *SSU1* and *ECM34* genes (Pérez-Ortín et al., 2002). Ssu1p is involved in sulfite efflux (Park and Bakalinsky, 2000), however, this simple change of promoter is not sufficient to induce the observed sulfite resistance. As a matter of fact, by using CRISPR-Cas9, a recent study engineered this translocation in the strain BY which normally does not have a strong resistance to sulfite (Fleiss et al., 2019). Surprisingly, this even had a negative impact on sulfite resistance. The cause for this was that more than a simple promoter change, the *ECM34* promoter in the wine strains also has tandem repeat of a 76 bp and that the number of these repeats is directly correlated to the sulfite resistance level in the translocated strains (Pérez-Ortín et al., 2002). Indeed, the introduction of the repeated sequence in the promoter induced an increase in sulfite resistance (Fleiss et al., 2019). This example depicts how complex structural variants *i.e.* a combination of translocation and copy-number variant can have an important phenotypic impact that would thrive through natural selection and be part of *S. cerevisiae* evolution.

The previous examples were cases of extreme phenotypic outcome for translocations where this type of event could disrupt genes, change their regulation or create fusion genes. However, the same study (Fleiss et al., 2019) also randomly introduced non-gene disrupting translocations by targeting TY transposable elements as translocation breakpoints and phenotyped the translocated strains on various conditions. The translocated strains displayed a surprising amount of phenotypic variance on several growth conditions. This led to the conclusions that the sole modification of the 3D architecture of the genome is enough to generate a range of fitness effect thus widening the phenotypic landscape of an individual.

Variation in the number of copies can be either in direct repeat *i.e.* tandem duplication, or somewhere else in the genome *i.e.* dispersed duplication (Figure 8), they also can display a wide distribution of sizes and fitness effects, from few base pairs in direct repeats as shown in the previous example of sulfite resistance in yeast, up to complete chromosomal duplications as in down syndrome for humans. CNV formation can occur via different mechanisms, both through recombination *e.g.* Nonhomologous End Joining, and replication-based mechanisms *e.g.* fork stalling and template switching (Stankiewicz and Lupski, 2010). They are extremely prevalent in with more than 80% of all ORFs belonging to the yeast pangenome having a CNV in at least one strain (Peter et al., 2018).
CNV has been proven to be a fast way of adapting to new stringent environments and displayed extensive dynamic in the way they appear and are maintained (Lauer *et al.*, 2018). Although adaptation through CNV often leads to direct fitness gain for a particular phenotype, as large size CNV can be energetically costly for the cell (Tang and Amon, 2013), it can also represent a case of fitness trade off under other conditions (Sunshine et al., 2015). 1,834 strains with telomeric amplifications (amplification from one point in the chromosome up to the telomere) were

engineered. Competition experiment with all strains pooled in three different nutrient limited conditions revealed that 175 of these CNVs induced fitness variation but they were mostly condition specific. Indeed, only four regions increased the fitness in all three conditions and seven decreased it in all three conditions (Sunshine et al., 2015).

CNV have already been taken into account in some GWAS (Craddock et al., 2010; Marshall et al., 2017). In yeast for example, GWAS with both CNVs and SNPs revealed that on average, variation in copy number explained a larger part of the heritability than SNP (Peter et al., 2018). Complex genetic disorders have SNP based heritability but are also possibly caused by multiple if not all types of structural variants. For instance, in Autism Spectrum Disorder (ASD), most of the liability is explained by common variants (Gaugler et al., 2014). However, CNVs are significantly associated in 10% of the individuals presenting ASD (Sebat et al., 2007), indels in 1% (Weiss et al., 2008), as well as cases of gene-disrupting balanced translocations (Kim et al., 2008).

Other causes for missing heritability

- **Non-coding variants**

A fair number of studies on human traits are based on exome SNP arrays. This implies that only gene-coding regions are taken into account which in human consist in only ~1.5% of the whole genome. However, most of the functional variation can be due to regions outside of exons (Hindorff et al., 2009; The ENCODE project consortium, 2012). Indeed, by elegantly developing a high-efficiency and scalable CRISPR-Cas9 genome editing technique entitled CRISPEY, more than 16,000 individual genetic variants were introduced in a yeast genetic background. Among them, 572 were found to significantly modify the fitness on rich glucose medium. These phenotypically relevant variants were particularly enriched in transcription binding sites or other regulatory sequences with only a fifth directly located in amino acid sequences (Sharon et al., 2018). A common intronic variant recently mapped through GWAS affects the regulation of *EDN1* through trans-effects. This gene is implicated in vasoconstriction and the variant strongly upregulates its expression level, leading to multiple vascular diseases, suggesting pleiotropic effects (Gupta et al., 2017). These observations imply that most of the variation actually comes from modification of gene expression levels.

- **Mitochondrial effects**

In humans, plethora of diseases are associated with mitochondrial variants. Mitochondrial DNA is, as its nuclear counterpart, prone to variation between different individuals as shown in humans (Diroma et al., 2014), fly (Bevers et al., 2018), fission yeast (Tao et al., 2019) or in worm (Zhu et al., 2019). In *S. cerevisiae*, a recent study on 96 natural isolates highlighted the importance role played by cyto-nuclear interactions in both respiratory and non-respiratory phenotypes (Vijayraghavan et al., 2019). Interactions between nuclear and mitochondrial

genomes can have substantial phenotypic effects (Joseph et al., 2013) sometimes leading up to a lethal phenotype, called cytonuclear incompatibilities (Chou and Leu, 2010; Hou et al., 2015). Although being very often overlook in GWAS, some recent associations studies based on mitochondrial DNA start to arise (Guyatt et al., 2019).

- **Pangenome**

All individuals belonging to the same population do not have the same exact gene content. The totality of the genes present in a population is called the pangenome (Tettelin et al., 2005). It encompasses the core genome referring to the genes present in every individual and the accessory genome accounting for genes present in a subset of all the individuals. Part of the reasons why gene content might differ between individuals is the presence of introgressed regions resulting from an hybridization event with another species but also the presence of horizontal gene transfers (HGT). The study of 1,011 natural *S. cerevisiae* isolates emphasized the extensive size of the yeast pangenome with close to 7,800 ORFs containing 2,856 accessory ORFs (Peter et al., 2018). About one third of the accessory genes originated from hybridization events with *Saccharomyces paradoxus*, a sister species of *S. cerevisiae*. This accessory genome is more prone to variation both in terms of nucleotidic sequence and in copy number. Moreover, some of these variable ORFs were significantly associated with a phenotype (Peter et al., 2018). In addition, introgressions and HGT giving specific domestication related functions in cheese and flor populations of yeast have been detected (Legras et al., 2018). Together, these studies confirm the role of the accessory genome in the phenotypic landscape of a species.

- **Environmental factors**

With human diseases, A problem might also arise from the fact that most of the studied diseases suffer from a categorical and dichotomic classification: people are either cases if they display the symptoms or control if not. However, being controlled by multiple genes, complex diseases appear either with a range of symptoms and severity or by a threshold dependent response masking a continuum of other molecular traits and risk factors (Gibson, 2012). Genetic disorders can also have different states with several development stages. Moreover, factors like environmental exposure or tissue type likely change regulatory mechanisms of the cells. Indeed, Trans eQTLs appear more highly tissue-specific than cis-eQTLs (GTEx Consortium, 2017). They can also induce epigenetic changes that will in turn modify gene expression (Finucane et al., 2018; Yengo et al., 2018). Post-transcriptional regulation mechanisms (Wu et al., 2013) are also playing an important role in the phenotypic landscape. However, in order to have the best understanding and quantification of all those mechanisms, they have to be considered at the same time by the use of a systemic approach by linking SNPs, gene expression level and protein level.

Genetic background effect

Despite the importance of understanding the genetic basis of complex traits, we currently lack complete knowledge of the relevant genetic components, even in scenarios where environment and other non-heritable contributing elements are well controlled (Mackay et al., 2009). The impact of genetic backgrounds, *inter alia*, on the phenotypic expression are still poorly understood to date. However, a better understanding of background-specific effect on phenotypic expression variation would lead to a greater perception of the genotype–phenotype relationship.

Behind the simplicity of a Mendelian inheritance, there is a clear hidden complexity of how variants exert a functional impact among individuals of the same species. Although this has been known for decades, the continuous level of the underlying phenotypic spectrum is overlooked. It is evident that most monogenic mutations do not always strictly follow Mendelian inheritance (Antonarakis et al., 2010). Many genetic disorders are referred as Mendelian that is caused by monogenic mutations. However, people inheriting the same mutation often display variation in phenotypic expression. This has come to be described by two words: 'penetrance' and 'expressivity' (Jarvik and Evans, 2017; Zlotogora, 2003). First, a mutation can exhibit incomplete penetrance, meaning that an individual may have this particular mutation but may not express the expected phenotype because of modifiers, epistatic interactions or suppressors present in the genome or because of the environment (Figure 9). An example is the *BRCA1* alleles, which predispose to breast and ovarian cancer in humans. Individuals with a mutation in the *BRCA1* gene have more or less 80% risk to develop this disease, therefore showing incomplete penetrance (Mavaddat et al., 2013). Second, the penetrance of a mutation is sometimes 100%, meaning that all the individuals present the expected trait (Figure 9), but they exhibit

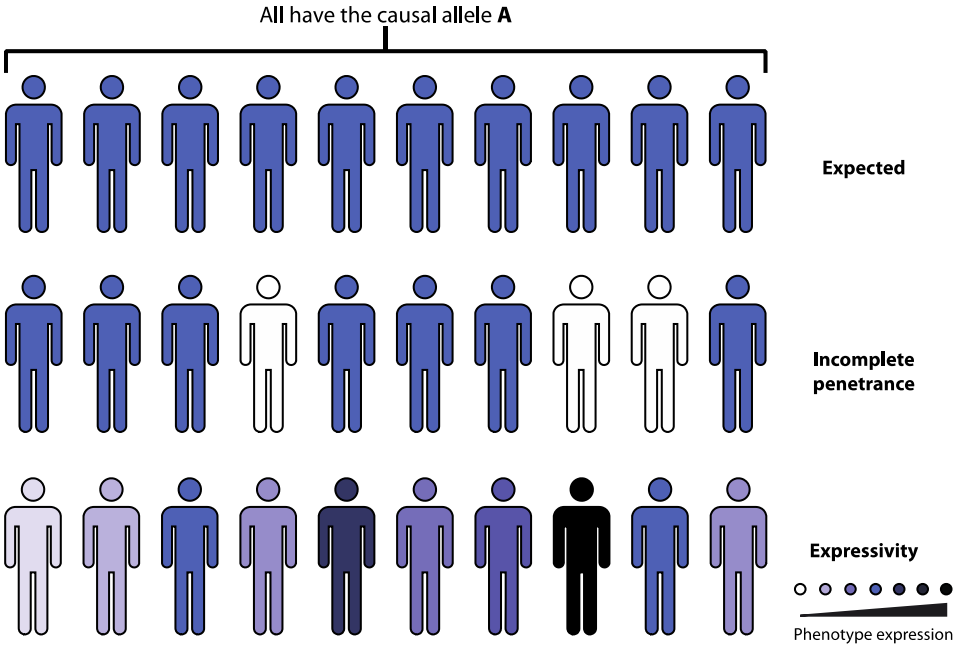


Figure 9. Penetrance and expressivity of traits
In the case of a monogenic disease, all individual carrying the causal allele are expected to develop the same trait. However, in some cases, individual with the causal allele do not express the expected phenotype, resulting in incomplete penetrance. For other traits, the phenotype will be expressed differentially depending on the individuals: some will develop more severe symptoms while other display milder symptoms thus representing the phenotypic expressivity.

different degrees of expressivity. Neurofibromatosis type I, a Mendelian disorder, is a notorious example of large variable expressivity. The disease is caused by dominant mutations in the *NF1* gene (Pasmant et al., 2012) and individuals carrying a mutation show a significant phenotypic heterogeneity. In fact, this is the case of a large number of diseases referred as caused by mutations occurring in single genes such as cystic fibrosis, Huntington's disease, and Fragile X (Arning, 2016; Cutting, 2010; Garber et al., 2008). In the case of cystic fibrosis, there is even evidence that modifiers, mutations in other genes, impact the phenotype (Emond et al., 2012; Rosendahl et al., 2013). Even for Down Syndrome, a whole chromosome disorder, there is evidence of phenotypic expression variation due to genetic background

differences (Ackerman et al., 2012; Li et al., 2012). More broadly, the phenotypic expression can be modified by various factors with the two most reported being age (Mavaddat et al., 2013) and sex (Tai et al., 2007). However, phenotypic expression can also be impacted by genetic background with the presence of genetic interactions and modifiers as already mentioned, mutation type (Thauvin-Robinet et al., 2009) and environment (Lachance et al., 2013). The distinction between penetrance and expressivity reflects an overly simplified view for several reasons. First, the full breadth of expression is not systematically characterized for any monogenic mutation in humans. Second, considerable uncertainty is introduced at the phenotypic level, because it is difficult to accurately characterize a trait measurement

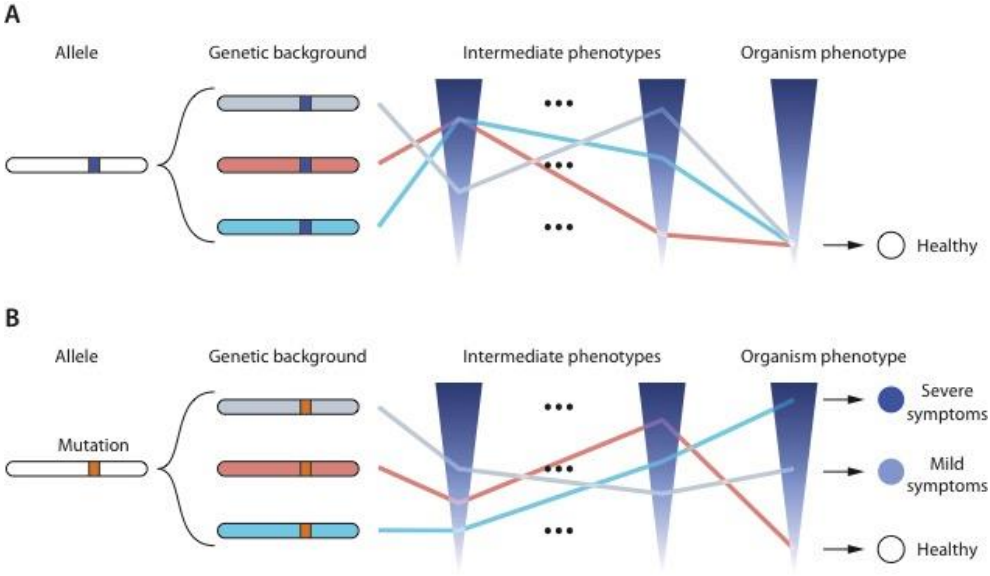


Figure 10. Phenotypic impact of the genetic background
A. An allele present in different genetic background could result in the same phenotypic outcome at the organismal level. However, this does not mean that intermediate phenotypes such as molecular traits (*e.g.* gene expression level) will be the same. Each layer of intermediate phenotype acts as a lens that can deviate the phenotype in a specific way with the organism phenotype as the focal point of all those superimposed lenses.
B. If a mutation occurs in this gene, incident intermediates phenotypes can be altered and completely change the organism phenotypic outcome.

for most genetic disorders. Most diseases are obviously a complex layering of intermediate molecular traits, for example gene expression, methylation, protein and metabolite levels. Several layers of intermediate molecular traits account for the global phenotype at the individual level. Thus, two individuals can display the same trait at the organism level but exhibit completely different intermediate phenotypes at the molecular level, or vice versa (Figure 10). To better understand the genetic basis of diseases, a more precise estimation of the phenotypic value as well as a more complete picture of the genetic architecture of the molecular traits are probably essential.

Genetic backgrounds, natural populations and model organisms

Variation among individuals of natural populations provides useful raw material to dissect the relationship between genetic variants and phenotypes (Alonso-Blanco et al., 2016; Auton et al., 2015; Durbin et al., 2010; Walter et al., 2015). Moreover, high-throughput genotyping and phenotyping technologies have greatly enhanced the power to dissect the genetic complexity hidden behind traits in model as well as in non-model organisms (Ellegren, 2014). A focus on the effects of the genetic backgrounds in natural populations is timely given several recent technological developments. Besides classical examples in human diseases, variation of phenotypic expressivity of monogenic mutations were also observed in model organisms at a genome-wide scale such as in yeast (Dowell et al., 2010; Hou et al., 2016), mouse (Doetschman, 2009; Montagutelli, 2000; Percival et al., 2017; Yoshiki and Moriwaki, 2006) and worm (Paaby et al., 2015; Vu et al., 2015).

High-throughput experiments are very useful to quantify the prevalence of the genetic background effects on functional variants between individuals. As an example, model organisms allow for systematic testing of loss-of-function phenotypes. In this context, systematic gene deletion collections were obtained for two closely related yeast *Saccharomyces cerevisiae* laboratory isolates (S288c and

S1278b) (Giaever et al., 2002; Montagutelli, 2000). An extensive difference of gene essentiality was found by comparing those two gene knockout libraries. In fact, nearly 5% of the genes identified as essential in one isolate are dispensable for survival in the other. In addition, rescue of the viable phenotype generally is of high order of complexity, requiring several modifier genes to counter the effect of a conditionally lethal deletion (Dowell et al., 2010). The genetic basis behind the disparity observed between these genetic backgrounds is still unknown. A similar study has been conducted using the nematode *Caenorhabditis elegans* by knocking down ~1400 genes with RNAi in the two canonical N2 Bristol and CB4856 Hawaiian isolates (Vu et al., 2015). Reduced expression of ~20% of the tested genes led to a trait that varied considerably across the lines. In parallel, the same conclusion was reached by targeting 29 maternal-effect genes in 55 wild *C. elegans* strains from around the world (Paaby et al., 2015). By perturbing known embryonic genes, the variability of the embryonic lethality expressivity across genetic backgrounds was clearly highlighted. Finally, the same mutation has also been recently expressed in a large number of *Drosophila* genetic backgrounds (Chow et al., 2016; Park et al., 2014). The Rh1G69D allele, which is a model for retinitis pigmentosa (RP), was crossed in multiple isolates of the *Drosophila* Genetic Reference Panel representing roughly 200 wild-derived strains (Chow et al., 2016; Mackay et al., 2012). It turns out that the retinal phenotype of Rh1G69D varies in a quantitative manner throughout the population, suggesting strong background effects. Using genome-wide association followed by functional validation with RNAi knock-down, the authors identified 10 modifier loci involved in the expressivity of RP (Chow et al., 2016). Many of these modifiers have human orthologs and most have not yet been implicated in the onset of retinitis pigmentosa. All together, these examples highlight that the phenotypic expression of a specific mutation varies tremendously and heritably, depending on the interacting alleles present in each genetic background.

The hidden complex inheritance of simple Mendelian cases is a continuum

By performing a species-wide survey of monogenic variants in the yeast S. cerevisiae, it has been recently shown that genes and alleles underlying the onset of Mendelian traits are variable in terms of their type, frequency and genomic distribution at the population level (Hou et al., 2016). The effect of a rare monogenic mutation of the *PDR1* gene, which confers resistance to cycloheximide and anisomycin, was explored and highlighted a continuum of the phenotypic spectrum. The Pdr1p protein is a transcription factor regulating the expression of various multidrug resistance ATP-Binding Cassette (ABC) transporters. In a yeast clinical isolate (YJM326), the presence of a nonsynonymous mutation in the sequence of the inhibitory domain of Pdr1p leads to constitutive expression of the downstream transporter coding genes, conferring the drug resistance trait. Twenty sensitive natural isolates were crossed with the resistant YJM326 isolate and the fitness distribution as well as the segregation of the drug resistance in the offspring were evaluated (Figure 11). Seventy percent of the cases displayed a classic Mendelian inheritance. But more interestingly, increased genetic complexity was observed in 30% of the cases, with significant and continuous deviations from the Mendelian expectation (Figure 11). In five cases, a slight deviation from Mendelian inheritance was observed. The level of genetic complexity was low and the variation of expressivity observed in these cases was due to the presence of one or two modifiers and/or gene interactions. Finally, the fitness distribution appeared to be normal for one given cross, which is characteristic of a complex trait. This study clearly demonstrated that the genetic complexity of traits could be dynamic, transitioning from clear Mendelian to diverse complex inheritance patterns depending on various genetic backgrounds. The power of this study lies in the fact that assumptions regarding the number of modifiers involved can be made by looking at the phenotypic distribution and segregation patterns in the offspring (Figure 11). Consequently, it is possible to more accurately estimate the genetic complexity of

traits. Deeper dissection of the transition between simple and complex traits in natural populations might therefore lead to new insights into the genetic architecture of traits.

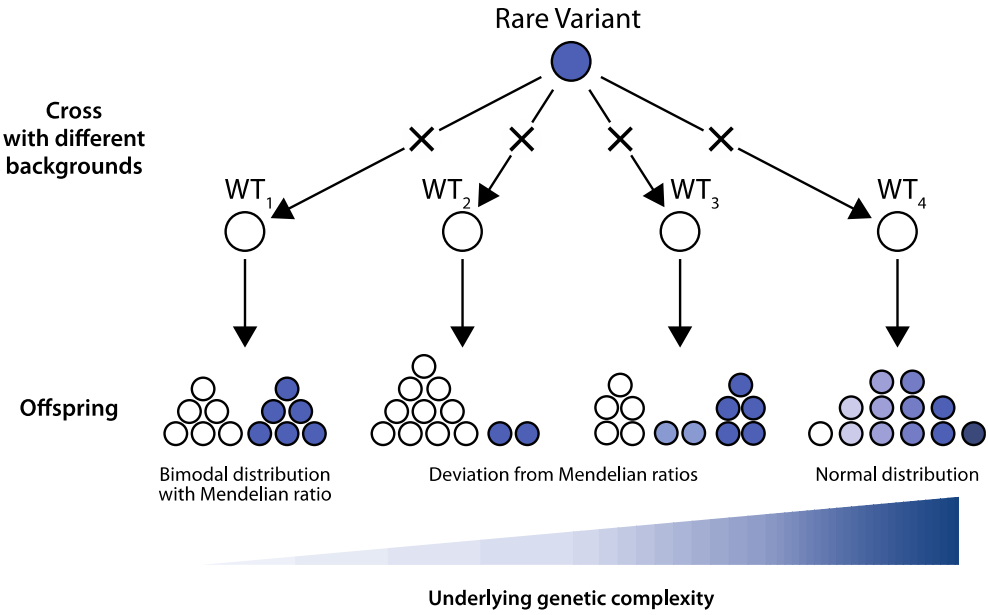


Figure 11. Trait complexity acts as a continuum at the species level
When crossing a rare variant with other genetic backgrounds, underlying genetic complexity of trait displays a continuum ranging from Mendelian or monogenic trait up to a complex trait. Genetic complexity underlying the trait can be assessed by looking at the offspring phenotypic distribution. A bimodal distribution following Mendelian ratios (2:2 for haploids and 3:1 for diploids) suggests a monogenic trait. Deviations from these ratios are signs of higher but intermediate level of complexity. Ultimately, a normal phenotypic distribution depicts a complex phenotype.

Conclusion

Understanding the phenotypic effects of natural genetic variants remains a major challenge in biology. This is obviously clear in the case of personalized medicine, with the hope to predict an individual's disease risk from his genetic data. The advances of high-throughput sequencing technologies hold the promise that whole-genome sequencing will be routine in medical care and will enhance the power to determine the genetic basis of traits. Comprehensive dissection of the genetic mechanisms underlying natural phenotypic diversity seems to be within reach. Since the rise of high-throughput sequencing technologies, a lot of effort has been put into genome-wide association and linkage mapping strategies to dissect the genotype–phenotype relationship. Nevertheless, limitations have been clearly highlighted by all association studies in humans, where all causal variants found fail to explain the entirety of the observed phenotypic variance. This unexplained variance is better known as the missing heritability (Manolio et al., 2009; Zuk et al., 2014). Because of this missing heritability, predictions about phenotypic variation remain limited. Possible reasons for this grey zone are the presence of a high number of rare variants, which are background specific, in natural populations and the intricate pattern of genetic interaction between all the genes that cannot be detected using these methods. Rare variants and genetic interactions clearly contribute to phenotypic expressivity variation. Deeper characterization of the inheritance, expressivity and genetic interactions hidden behind the phenotypic landscape of natural populations will bring further valuable insight into the conversion of genetic into phenotypic variation.

Yeast as a powerful tool to dissect the genotype-phenotype relationship.

Studying the causes for the missing heritability and more generally genetic architecture of traits require the use of an adapted model organism. Yeast stands as a powerful model to study it. They have long been used as a model in several domains of biology because of being a unicellular eukaryote that is easily laboratory amendable, with a short generation time, a small and compact genome (Goffeau et al., 1996) and numerous genetic or molecular tools available. As proved with the various examples discussed in the previous sections, yeast stands as an exceptional asset to dissect all the aspects of the genetic architecture of traits. We now have access to more than a thousand of natural isolates presenting a wide range of phenotypic and genetic diversity (Figure 12).

As with their haplodiplobiontic life cycle, crossing and sporulating isolates to generate successive generations is relatively easy. Obtaining a large progeny of hundreds or even thousands of segregants from a large number of crosses between natural isolates can be done quickly. A major advantage of yeast is the fact that once genotyped, one can phenotype them as far as imagination goes thanks to their clonal replication by budding. This comes in handy for applying mapping strategies on a large panel of phenotypes either related to growth, morphology or molecular traits.

Yeast also are convenient for scaling up experiment in large-scale high-throughput studies and can efficiently be paired with automated handling both in liquid or solid to manipulate thousands of isolates at the same time. These high-throughput methods can uncover the full potential of yeast especially when paired with classical yeast genetic techniques such as crosses, tetrad dissections or genome editing. The very fact that yeast sporulates is what makes them very unique for genetic studies. Indeed, through sporulation and tetrad dissection, it is possible to access the outcome of individual meiosis events and chromosomal segregation in the four resulting spores.

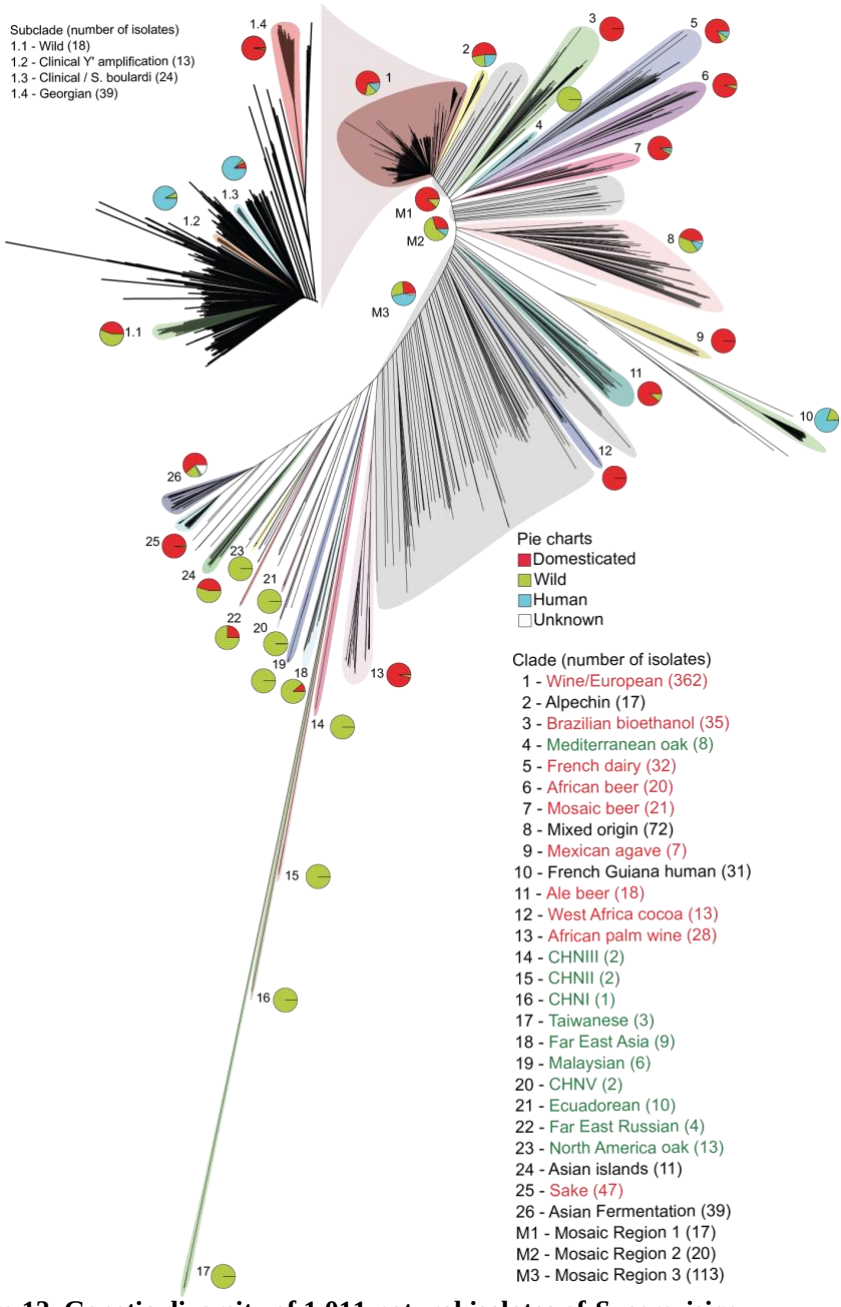


Figure 12. Genetic diversity of 1,011 natural isolates of *S. cerevisiae*
Tree based on pairwise nucleotidic diversity between isolates. 26 clades based on geographical or ecological origins are depicted by colors. The top left inset represents a magnification of the Wine/European cluster. Adapted from Peter *et al.* (2018).

Bibliography

Ackerman, C., Locke, A.E., Feingold, E., Reshey, B., Espana, K., Thusberg, J., Mooney, S., Bean, L.J.H., Dooley, K.J., Cua, C.L., et al. (2012). An excess of deleterious variants in VEGF-A pathway genes in down-syndrome-associated atrioventricular septal defects. Am. J. Hum. Genet. 91, 646–659.

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

Antonarakis, S.E., Chakravarti, A., Cohen, J.C., and Hardy, J. (2010). Mendelian disorders and multifactorial traits: the big divide or one for all? Nat. Rev. Genet. 11, 380–384.

Arning, L. (2016). The search for modifier genes in Huntington disease – Multifactorial aspects of a monogenic disorder. Mol. Cell. Probes 30, 404–409.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., et al. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465, 627–631.

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

Bates, G.P., Dorsey, R., Gusella, J.F., Hayden, M.R., Kay, C., Leavitt, B.R., Nance, M., Ross, C.A., Scahill, R.I., Wetzel, R., et al. (2015). Huntington disease. Nat. Rev. Dis. Prim. 1, 15005.

Bateson, W. (1909). Mendel's principles of heredity (Cambridge: Cambridge University Press,).

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.

Bevers, R.P.J., Litovchenko, M., Kapopoulou, A., Braman, V.S., Robinson, M.R., Auwerx, J., Hollis, B., and Deplancke, B. (2018). Extensive mitochondrial population structure and haplotype-specific phenotypic variation in the Drosophila Genetic Reference Panel. BioRxiv 466771.

Biederstedt, E., Oliver, J.C., Hansen, N.F., Jajoo, A., Dunn, N., Olson, A., Busby, B., and Dilthey, A.T. (2018). NovoGraph: Human genome graph construction from multiple long-read de novo assemblies. F1000Research 7, 1391.

Billmann, M., Horn, T., Fischer, B., Sandmann, T., Huber, W., and Boutros, M. (2016). A genetic interaction map of cell cycle regulators. Mol. Biol. Cell 27, 1397–1407.

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. BioRxiv 607291.
Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic (Cell Press).
Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J., and Roux, F. (2010). Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS Genet. 6, 40.
Brem, R.B., and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. 102, 1572–1577.
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.
Byrne, A.B., Weirauch, M.T., Wong, V., Koeva, M., Dixon, S.J., Stuart, J.M., and Roy, P.J. (2007). A global analysis of genetic interactions in Caenorhabditis elegans. J. Biol. 6, 8.
Chou, J.Y., and Leu, J.Y. (2010). Speciation through cytonuclear incompatibility: Insights from yeast and implications for higher eukaryotes. BioEssays 32, 401–411.
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.
Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature 464, 704–712.
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.
Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science (80-.). 327, 425–431.
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 (80-.). 353, aaf1420.
Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713–720.
Cretu Stancu, M., Van Roosmalen, M.J., Renkens, I., Nieboer, M.M., Middelkamp, S., De Ligt, J., Pregno, G., Giachino, D., Mandrile, G., Espejo Valle-Inclan, J., et al. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat. Commun. 8, 1326.
Cutting, G.R. (2010). Modifier genes in Mendelian disorders: The example of cystic fibrosis. Ann. N. Y. Acad. Sci. 1214, 57–69.
Deutschbauer, A.M., Jaramillo, D.F., Proctor, M., Kumm, J., Hillenmeyer, M.E., Davis, R.W., Nislow, C., and Giaever, G. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics 169, 1915–1925.

Diroma, M.A., Calabrese, C., Simone, D., Santorsola, M., Calabrese, F.M., Gasparre, G., and Attimonelli, M. (2014). Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data. BMC Genomics 15, S2.

Doetschman, T. (2009). Influence of genetic background on genetically engineered mouse phenotypes. Methods Mol. Biol. 530, 423–433.

Dowell, R.D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D.A., Rolfe, P.A., Heisler, L.E., Chin, B., et al. (2010). Genotype to phenotype: a complex problem. Sci. (New York, NY) 328, 469.

Druker, B.J., Tamura, S., Buchdunger, E., Ohno, S., Segal, G.M., Fanning, S., Zimmermann, J., and Lydon, N.B. (1996). Effects of a selective inhibitor of the Ab1 tyrosine kinase on the growth of Bcr-Ab1 positive cells. Nat. Med. 2, 561–566.

Durbin, R.M., Altshuler, D.L., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De La Vega, F.M., Donnelly, P., et al. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.

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.

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

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. Trends Ecol. Evol. 29, 51–63.

Emond, M.J., Louie, T., Emerson, J., Zhao, W., Mathias, R.A., Knowles, M.R., Wright, F.A., Rieder, M.J., Tabor, H.K., Nickerson, D.A., et al. (2012). Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nat. Genet. 44, 886–889.

Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. 50, 621–629.

Fisher, R.. (1918). The correlation between relatives on the supposition of mendelian inheritance. Trans. R. Soc. Edinburgh 52, 399–433.

Fleiss, A., O'Donnell, S., Fournier, T., Agier, N., Delmas, S., Schacherer, J., and Fischer, G. (2019). Reshuffling yeast chromosomes with CRISPR/Cas9. BioRxiv 415349.

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

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. BioRxiv 609917.

Fournier, T.T., Gounot, J.-S.J.-S., Freel, K., Cruaud, C., Lemainque, A., Aury, J.-M., Wincker, P., Schacherer, J., and Friedrich, A. (2017). High-Quality de Novo

Genome Assembly of the Dekkera bruxellensis Yeast Isolate Using Nanopore MinION Sequencing. G3 (Bethesda). 7, g3.300128.2017.
Garber, K.B., Visootsak, J., and Warren, S.T. (2008). Fragile X syndrome. Eur. J. Hum. Genet. 16, 666–672.
Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., et al. (2014). Most genetic risk for autism resides with common variation. Nat. Genet. 46, 881–885.
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.
Gibson, G. (2012). Rare and common variants: Twenty arguments. Nat. Rev. Genet. 13, 135–145.
Goffeau, A., Barrell, G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science (80-.). 274, 546–567.
Gong, L., Wong, C.H., Cheng, W.C., Tjong, H., Menghi, F., Ngan, C.Y., Liu, E.T., and Wei, C.L. (2018). Picky comprehensively detects high-resolution structural variants in nanopore long reads. Nat. Methods 15, 455–460.
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, 5162.
Gordon, C.T., Petit, F., Kroisel, P.M., Jakobsen, L., Zechi-Ceide, R.M., Oufadem, M., Bole-Feysot, C., Pruvost, S., Masson, C., Tores, F., et al. (2013). Mutations in endothelin 1 cause recessive auriculocondylar syndrome and dominant isolated question-mark ears. Am. J. Hum. Genet. 93, 1118–1125.
Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. (2008). Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. Am. J. Hum. Genet. 82, 100–112.
GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature 550, 204–213.
Gupta, A.K., Bharadwaj, M., and Mehrotra, R. (2016). Skin Cancer Concerns in People of Color: Risk Factors and Prevention. Asian Pac. J. Cancer Prev. 17, 5257–5264.
Gupta, R.M., Hadaya, J., Trehan, A., Zekavat, S.M., Roselli, C., Klarin, D., Emdin, C.A., Hilvering, C.R.E., Bianchi, V., Mueller, C., et al. (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. Cell 170, 522-533.e15.
Guyatt, A.L., Brennan, R.R., Burrows, K., Guthrie, P.A.I., Ascione, R., Ring, S.M., Gaunt, T.R., Pyle, A., Cordell, H.J., Lawlor, D.A., et al. (2019). A genome-wide association study of mitochondrial DNA copy number in two population-based cohorts. Hum. Genomics 13, 6.
Hallin, J., Märtens, K., Young, A.I., Zackrisson, M., Salinas, F., Parts, L., Warringer, J., and Liti, G. (2016). Powerful decomposition of complex traits in a diploid model. Nat. Commun. 7, 13311.
Hanlon, P., Lorenz, W.A., Shao, Z., Harper, J.M., Galecki, A.T., Miller, R.A., and Burke,

D.T. (2006). Three-locus and four-locus QTL interactions influence mouse insulin-like growth factor-I. Physiol. Genomics 26, 46–54.
Hellwege, J.N., Keaton, J.M., Giri, A., Gao, X., Velez Edwards, D.R., and Edwards, T.L. (2017). Population Stratification in Genetic Association Studies. Curr. Protoc. Hum. Genet. 95, 1.22.1-1.22.23.
Higgins, M.G., Fitzsimons, C., McClure, M.C., McKenna, C., Conroy, S., Kenny, D.A., McGee, M., Waters, S.M., and Morris, D.W. (2018). GWAS and eQTL analysis identifies a SNP associated with both residual feed intake and GFRA2 expression in beef cattle. Sci. Rep. 8, 14301.
Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. 106, 9362–9367.
Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6, 95–108.
Horn, T., Sandmann, T., Fischer, B., Axelsson, E., Huber, W., and Boutros, M. (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. Nat. Methods 8, 341–346.
Hou, J., Friedrich, A., Gounot, J.-S., and Schacherer, J. (2015). Comprehensive survey of condition-specific reproductive isolation reveals genetic incompatibility in yeast. Nat. Commun. 6, 7214.
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.
Istace, B., Friedrich, A., d'Agata, L.L., Faye, S.S., Payen, E., Beluche, O., Caradec, C., Davidas, S., Cruaud, C., Liti, G., et al. (2017). de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. Gigascience 6, 1–13.
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, 239.
Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36, 338–345.
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.
Jamann, T.M., Balint-Kurti, P.J., and Holland, J.B. (2015). QTL mapping using high-throughput sequencing. In Plant Functional Genomics: Methods and Protocols: Second Edition, pp. 257–285.
Jarvik, G.P., and Evans, J.P. (2017). Mastering genomic terminology. Genet. Med. 19, 491–492.
Johannsen, W. (1911). The Genotype Conception of Heredity. Am. Nat. 45, 129–159.
Joseph, B., Corwin, J.A., Li, B., Atwell, S., and Kliebenstein, D.J. (2013). Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation

in the metabolome. Elife 2.
Kim, H.G., Kishikawa, S., Higgins, A.W., Seong, I.S., Donovan, D.J., Shen, Y., Lally, E., Weiss, L.A., Najm, J., Kutsche, K., et al. (2008). Disruption of Neurexin 1 Associated with Autism Spectrum Disorder. Am. J. Hum. Genet. 82, 199–207.
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, eaao1729.
Lachance, J., Jung, L., and True, J.R. (2013). Genetic Background and GxE Interactions Modulate the Penetrance of a Naturally Occurring Wing Mutation in Drosophila melanogaster. G3: Genes|Genomes|Genetics 3, 1893–1901.
Lauer, S., Avecilla, G., Spealman, P., Sethia, G., Brandt, N., Levy, S.F., and Gresham, D. (2018). Single-cell copy number variant detection reveals the dynamics and diversity of adaptation. PLoS Biol. 16, e3000069.
Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. Am. J. Hum. Genet. 95, 5–23.
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.
Lehner, B., Crombie, C., Tischler, J., Fortunato, A., and Fraser, A.G. (2006). Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nat. Genet. 38, 896–903.
Li, H., Cherry, S., Klinedinst, D., DeLeon, V., Redig, J., Reshey, B., Chin, M.T., Sherman, S.L., Maslen, C.L., and Reeves, R.H. (2012). Genetic modifiers predisposing to congenital heart disease in the sensitized down syndrome population. Circ. Cardiovasc. Genet. 5, 301–308.
Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S.Y., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. Lancet 380, 2095–2128.
Lugo, T.G., Pendergast, A.M., Muller, A.J., and Witte, O.N. (1990). Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. Science (80-.). 247, 1079–1082.
Lynch, M. (2006). The origins of eukaryotic gene structure. Mol. Biol. Evol. 23, 450–468.
MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901.
Mackay, T.F.C. (2001). The Genetic Architecture of Quantitative Traits. Annu. Rev. Genet. 35, 303–339.
Mackay, T.F.C.C. (2014). Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. Nat. Rev. Genet. 15, 22–33.
Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D.,

Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature 482, 173–178.
Mackay, T.F.C.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 Publishing Group).
Mäki-Tanila, A., and Hill, W.G. (2014). Influence of gene interaction on complex trait variation with multilocus models. Genetics 198, 355–367.
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.
Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine, R.S., Lu, Y., Schurmann, C., Highland, H.M., et al. (2017). Rare and low-frequency coding variants alter human adult height. Nature 542, 186–190.
Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D., et al. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat. Genet. 49, 27–35.
Märtens, K., Hallin, J., Warringer, J., Liti, G., and Parts, L. (2016). Predicting quantitative traits from genome and phenome with near perfect accuracy. Nat. Commun. 7, 11512.
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.
Mendel, G. (1866). Versuche über Plflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865.
Montagutelli, X. (2000). Effect of the genetic background on the phenotype of mouse mutations. J. Am. Soc. Nephrol. 11 Suppl 1, S101–S105.
Moore, J.H., and Williams, S.M. (2005). Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. BioEssays 27, 637–646.
Mullis, M.N., Matsui, T., Schell, R., Foree, R., and Ehrenreich, I.M. (2018). The complex underpinnings of genetic background effects. Nat. Commun. 9, 3548.
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.
Paixão, T., and Barton, N.H. (2016). The effect of gene interactions on the long-term response to selection. Proc. Natl. Acad. Sci. 113, 4422–4427.
Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 11, R52.

Park, H., and Bakalinsky, A.T. (2000). SSUI mediates sulphite efflux in Saccharomyces cerevisiae. Yeast 16, 881–888.

Park, S.Y., Ludwig, M.Z., Tamarina, N.A., He, B.Z., Carl, S.H., Dickerson, D.A., Barse, L., Arun, B., Williams, C.L., Miles, C.M., et al. (2014). Genetic complexity in a Drosophila model of diabetes-associated misfolded human proinsulin. Genetics 196, 539–555.

Parmeggiani, F., S. Sorrentino, F., Ponzin, D., Barbaro, V., Ferrari, S., and Di Iorio, E. (2011). Retinitis Pigmentosa: Genes and Disease Mechanisms. Curr. Genomics 12, 238–249.

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

Peltier, E., Friedrich, A., Schacherer, J., and Marullo, P. (2019). Quantitative Trait Nucleotides Impacting the Technological Performances of Industrial Saccharomyces cerevisiae Strains. Front. Genet. 10, 683.

Percival, C.J., Marangoni, P., Tapaltsyan, V., Klein, O., and Hallgrímsson, B. (2017). The Interaction of Genetic Background and Mutational Effects in Regulation of Mouse Craniofacial Shape. G3 Genes, Genomes, Genet. 7, 1439–1450.

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

Pettersson, M., Besnier, F., Siegel, P.B., and Carlborg, Ö. (2011). Replication and explorations of High-Order epistasis using a large advanced intercross line pedigree. PLoS Genet. 7, e1002180.

Radke, D.W., and Lee, C. (2015). Adaptive potential of genomic structural variation in human and mammalian evolution. Brief. Funct. Genomics 14, 358–368.

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, J.L. (2003). Genetics of Hair and Skin Color. Annu. Rev. Genet. 37, 67–90.

Rockman, M. V (2012). The QTN program and the alleles that matter for evolution: All that's gold does not glitter. Evolution (N. Y). 66, 1–17.

Rockman, M. V., and Kruglyak, L. (2009). Recombinational landscape and population genomics of caenorhabditis elegans. PLoS Genet. 5, e1000419.

Rockman, M. V, Skrovanek, S.S., and Kruglyak, L. (2010). Selection at linked sites shapes heritable phenotypic variation in C. elegans. Science (80-.). 330, 372–376.

Rosendahl, J., Landt, O., Bernadova, J., Kovacs, P., Teich, N., B̈odeker, H., Keim, V., Ruffert, C., M̈ossner, J., Kage, A., et al. (2013). CFTR, SPINK1, CTRC and PRSS1 variants in chronic pancreatitis: Is the role of mutated CFTR overestimated? Gut 62, 582–592.

Sailer, Z.R., and Harms, M.J. (2017). Detecting high-order epistasis in nonlinear genotype-phenotype maps. Genetics 205, 1079–1088.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B.,

Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science (80-.). 316, 445–449.
Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods 15, 461–468.
Segrè, A. V, Murray, A.W., and Leu, J.Y. (2006). High-resolution mutation mapping reveals parallel experimental evolution in yeast. PLoS Biol. 4, 1372–1385.
Seymour, D.K., Chae, E., Grimm, D.G., Martín Pizarro, C., Habring-Müller, A., Vasseur, F., Rakitsch, B., Borgwardt, K.M., Koenig, D., and Weigel, D. (2016). Genetic architecture of nonadditive inheritance in Arabidopsis thaliana hybrids. Proc. Natl. Acad. Sci. U. S. A. 113, E7317–E7326.
Sharon, E., Chen, S.A.A., Khosla, N.M., Smith, J.D., Pritchard, J.K., and Fraser, H.B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. Cell 175, 544-557.e16.
She, R., and Jarosz, D.F. (2018). Mapping Causal Variants with Single-Nucleotide Resolution Reveals Biochemical Drivers of Phenotypic Change. Cell 172, 478-490.e15.
Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. 99, 139–153.
Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D.I., Cornes, B.K., Davis, C., Dunkel, L., De Lange, M., Harris, J.R., Hjelmborg, J.V.B., et al. (2003). Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. Twin Res. 6, 399–408.
Speed, D., Cai, N., Johnson, M.R., Nejentsev, S., and Balding, D.J. (2017). Reevaluation of SNP heritability in complex human traits. Nat. Genet. 49, 986–992.
Stankiewicz, P., and Lupski, J.R. (2010). Structural Variation in the Human Genome and its Role in Disease. Annu. Rev. Med. 61, 437–455.
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.
Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81.
Sunshine, A.B., Payen, C., Ong, G.T., Liachko, I., Tan, K.M., and Dunham, M.J. (2015). The Fitness Consequences of Aneuploidy Are Driven by Condition-Dependent Gene Effects. PLoS Biol. 13, e1002155.
Swallow, D.M. (2003). Genetics of Lactase Persistence and Lactose Intolerance. Annu. Rev. Genet. 37, 197–219.
van Swinderen, B., Shook, D.R., Ebert, R.H., Cherkasova, V.A., Johnson, T.E., Reis, R.J.S., and Crowder, C.M. (1997). Quantitative trait loci controlling halothane sensitivity in Caenorhabditis elegans. Proc. Natl. Acad. Sci. 94, 8232–8237.
Tai, Y.C., Domchek, S., Parmigiani, G., and Chen, S. (2007). Breast cancer risk among male BRCA1 and BRCA2 mutation carriers. J. Natl. Cancer Inst. 99, 1811–1814.

Tang, Y.C., and Amon, A. (2013). Gene copy-number alterations: A cost-benefit analysis. Cell 152, 394–405.
Tao, Y.-T., Suo, F., Wang, Y.-K., Huang, S., and Du, L.-L. (2019). Intraspecific diversity of fission yeast mitochondrial genome. BioRxiv 624742.
Taylor, M.B., and Ehrenreich, I.M. (2014). Genetic interactions involving five or more genes contribute to a complex trait in yeast. PLoS Genet. 10, e1004324.
Taylor, M.B., and Ehrenreich, I.M. (2015). Higher-order genetic interactions and their contribution to complex traits. Trends Genet. 31, 34–40.
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. 102, 13950–13955.
Thauvin-Robinet, C., Munck, A., Huet, F., Génin, E., Bellis, G., Gautier, E., Audrézet, M.P., Férec, C., Lalau, G., Des Georges, M., et al. (2009). The very low penetrance of cystic fibrosis for the R117H mutation: A reappraisal for genetic counselling and newborn screening. J. Med. Genet. 46, 752–758.
The ENCODE project consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.
Timpson, N.J., Greenwood, C.M.T.T., Soranzo, N., Lawson, D.J., and Richards, J.B. (2018). Genetic architecture: The shape of the genetic contribution to human traits and disease (Nature Publishing Group).
Vijayraghavan, S., Kozmin, S.G., Strope, P.K., Skelly, D.A., Lin, Z., Kennell, J., Magwene, P.M., Dietrich, F.S., and McCusker, J.H. (2019). Mitochondrial genome variation affects multiple respiration and nonrespiration phenotypes in saccharomyces cerevisiae. Genetics 211, 773–786.
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.
Wainschtein, P., Jain, D.P., Yengo, L., Zheng, Z., TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., et al. (2019). Recovery of trait heritability from whole genome sequence data. BioRxiv 588020.
Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature 526, 82–90.
Wang, X., Pang, Y., Zhang, J., Wu, Z., Chen, K., Ali, J., Ye, G., Xu, J., and Li, Z. (2017). Genome-wide and gene-based association mapping for rice eating and cooking characteristics and protein content. Sci. Rep. 7, 17203.
Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. Nat. Rev. Genet. 14, 125–138.
Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A.R., Green, T., et al. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. Obstet. Gynecol. Surv.

63, 361–363.
Wexler, N.S. (2018). The Tiresias complex: Huntington's disease as a paradigm of testing for late-onset disorders. FASEB J. 6, 2820–2825.
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.
Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. Nature 499, 79–82.
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.
Yoshiki, A., and Moriwaki, K. (2006). Mouse phenome research: Implications of genetic background. ILAR J. 47, 94–102.
Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. 45, 1134–1140.
Zhu, Z., Bakshi, A., Vinkhuyzen, A.A.E., Hemani, G., Lee, S.H., Nolte, I.M., Van Vliet-Ostaptchouk, J. V., Snieder, H., Esko, T., Milani, L., et al. (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. Am. J. Hum. Genet. 96, 377–385.
Zhu, Z., Han, X., Wang, Y., Liu, W., Lu, Y., Xu, C., Wang, X., Hao, L., Song, Y., Huang, S., et al. (2019). Identification of specific nuclear genetic loci and genes that interact with the mitochondrial genome and contribute to fecundity in caenorhabditis elegans. Front. Genet. 10, 28.
Zlotogora, J. (2003). Penetrance and expressivity in the molecular age. Genet. Med. 5, 347–352.
Zuk, O.O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. U. S. A. 111, E455-464.

PROJECT SUMMARY

Understanding how genetic variance controls the amount of phenotypic variation in a natural population is a central question in biology. This allows individuals to adapt to ever changing environments and the presence of other selective pressures. Each trait is shaped by the genetic makeup of an individual and thus can be idiosyncratic to this individual. The main theme of my project is to use the awesome power of yeast genetics to investigate as much as possible the genetic architecture of traits through the lens of the missing heritability and its different main components.

Based on the recently published collection of whole genome resequencing of more than 1,000 natural isolates of *S. cerevisiae*, our lab currently has the best understanding of both genetic and phenotypic natural diversity of any eukaryotic model. By taking advantage of this raw material, we selected a panel of 55 natural strains that are diploid, homozygous and representative of the natural diversity of the species. From these, stable founder haploid were generated in both mating types and crossed in an all by all manner, also called a diallel cross scheme, generating more than 3,000 hybrids that were then phenotyped on a wide range of conditions impacting various cellular pathways. This dataset proved to be a goldmine to investigate genetic architecture of traits. **The first chapter** of this project aimed at uncovering several aspects of this genetic architecture (Figure 1A). First, the diallel hybrid panel allowed us to precisely measure and separate the part of the phenotypic variance explained by additive phenomena from the part due to non-additive phenomena at a population-scale. This allowed to confirm that most of the phenotypic variance is controlled by additivity while non-additivity still represent about a third of the total variance. Next, we sought to infer mode of inheritance for each cross/trait combination. There again, most of the conditions were predominantly controlled by additivity but we also found an important role for dominance as well as over- and underdominance suggesting the presence of genetic interactions with large effects.

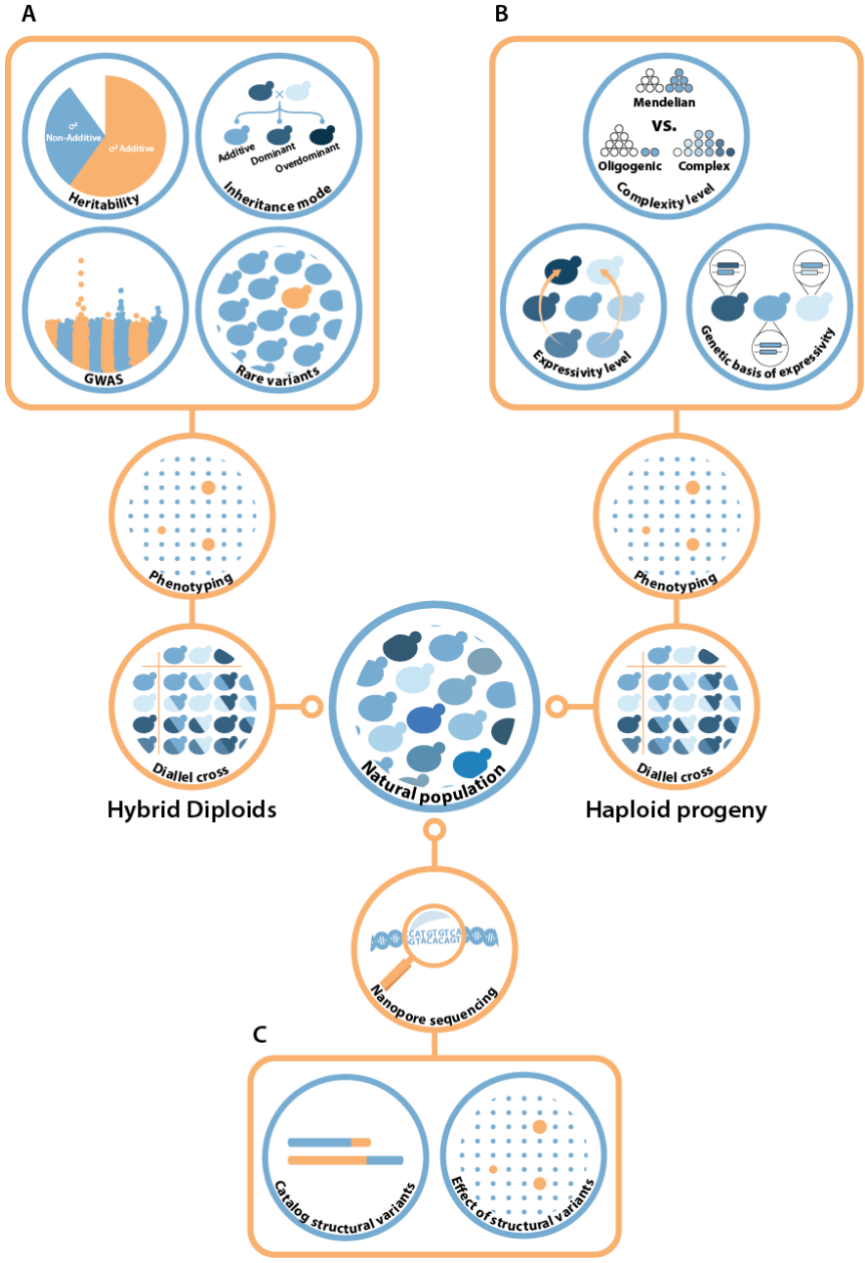


Figure 1. Schematic overview of the project
Starting from populations of natural yeast isolates, we wanted to investigate the putative cause of missing heritability. For that, Several aspects of genetic architecture of traits have been investigated using: **A.** a diallel hybrid diploid panel, **B.** the haploid progeny of these hybrids and **C.** natural structural variants.

The last aim of this study was to uncover the part played by low frequency and rare variants in the phenotypic landscape of the population. By performing GWAS, we could see an enrichment of significantly associated variants having lower allele frequency and that on average, they explained as much heritability as the common variants.

To go further with this first analysis of the architecture of traits in diploid, **in the second chapter**, we then aimed at obtaining a global view of the genetic complexity of traits by looking at the phenotypic distribution of large progenies (Figure 1B). To do so, we subsampled the previously generated diallel panel of hybrid and proceeded with 190 crosses coming from a diallel of 20 parental strains. For each of these crosses, a large offspring was generated and then phenotyped on the same conditions as the hybrids. Phenotypic distribution of the offspring allows to differentiate between monogenic traits, oligogenic traits and complex traits. Although most of the cross/trait combinations were complex, some conditions showed an extensive number of monogenic inheritance patterns thus revealing the presence of high impact variants. We then could follow the effect of these high effect variants across different genetic backgrounds to trace their expressivity level throughout the population. We uncovered a large number of cases of such expressivity with increases in the genetic complexity in some crosses carrying one of these variants.

In the last chapter, our efforts focused on another important part of both genetic architecture and missing heritability which is the impact of structural variation in the phenotypic landscape (Figure 1C). The goal here was first to introduce the tools needed to perform such studies, that is taking advantage of new generation long read sequencing technologies such as Oxford Nanopore to facilitate genome assemblies and structural variant detection in natural populations. The first aim of this study has been to focus on a non-model yeast species, namely *Brettanomyces bruxellensis*.

This species features extensive large-scale structural variation as demonstrated by pulsed field gel electrophoresis in several natural isolates, making it a model of choice for this type of study. However, the lack of a good quality reference genome pushed us to generate one on our own. Thus, we completed the assembly of the UMY321 with 8 chromosomes and a total of 13Mb by combining the length of nanopore reads with the quality of Illumina short reads. In order to highlight structural variants, two other divergent strains of *B. bruxellensis* have also been sequenced using the same method, which allowed to highlight the presence of 11 gross chromosomal rearrangements. Another part of this work was to obtain a much broader view of the overall landscape of structural variation in *S. cerevisiae*. To do so, 100 natural isolates coming from the 1,002 yeast genomes project have been sequenced both with Illumina and Oxford nanopore. *De novo* assembly of their genome has just been finalised. This dataset will allow us to better grasp the prevalence of structural variants in a natural population as well as in a near future to assess their putative association with phenotypic variation.

Altogether, this project focused on obtaining a better view of several aspects of the genetic architecture of traits in natural populations and more precisely to some causes for missing heritability. Thanks to an adapted experimental design and the use of a powerful model organism, it allowed us to study these mechanisms underlying missing heritability by taking advantage of the very large phenotypic and genotypic diversity present in yeast. Indeed, we could quantify the effect of low frequency variants, to determine the complexity spectrum of genetic complexity as well as its background dependent dynamic. Finally, we showed the prevalence of structural variation between individuals belonging to the same population, thus laying the foundation for further experiments assessing their phenotypic outcome.

CHAPTER 1

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

Summary

Genome-wide association studies (GWAS) allow to dissect the genetic basis of complex traits at the population level. However, despite the extensive number of trait-associated loci found, they often fail to explain a large part of the observed phenotypic variance. One potential reason for this discrepancy could be the preponderance of undetected low-frequency genetic variants in natural populations. To increase the allele frequency of those variants and assess their phenotypic effects at the population level, we generated a diallel panel consisting of 3,025 hybrids, derived from pairwise crosses between a subset of natural isolates from a completely sequenced 1,011 *Saccharomyces cerevisiae* population. We examined each hybrid across a large number of growth traits, resulting in a total of 148,225 cross/trait combinations. Parental *versus* hybrid regression analysis showed that while most phenotypic variance is explained by additivity, a significant proportion (29%) is governed by non-additive effects. This is confirmed by the fact that a majority of complete dominance is observed in 25% of the traits. By performing GWAS on the diallel panel, we detected 1,723 significantly associated genetic variants, 16.3% of which are low-frequency variants in the initial population. These variants, which would not be detected using classical GWAS, explain 21% of the phenotypic variance on average. Altogether, our results demonstrate that low-frequency variants should be accounted for as they contribute to a large part of the phenotypic variation observed in a population.

Introduction

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 continuous search to unravel the intricate relationship existing between genotype and phenotype in natural populations has seen tremendous progress in the last 10 years with the rise of high-throughput sequencing and GWAS (Alonso-Blanco et al., 2016; Auton et al., 2015; Mackay et al., 2012; Peter et al., 2018; Visscher et al., 2017). However, this major leap forward also comes with some limitations. As discussed before, very often in association studies, variants associated with complex traits fail to explain a large part of the observed phenotypic variation (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). This missing heritability has been attributed to several phenomena, one of which is the presence of low frequency and rare variants (Gibson, 2012; Hindorff et al., 2009; Manolio et al., 2009; Pritchard, 2002; Walter et al., 2015). Therefore, measuring the effect of rare and low frequency variants in natural populations is of prime interest in order to better understand the genetic architecture of traits as well as to unravel part of the missing heritability. 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 3,025 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 Table S1). Haploid isogenic lines of both mating types were generated for each of the diploid homozygous lines (Figure S1). A full diallel cross panel was constructed by systematically crossing the 55 selected isolates in a pairwise manner (Figure 1D). In total, we generated 3,025 hybrids, representing 2,970 heterozygous hybrids with a unique parental combination and 55 homozygous hybrids. All 3,025 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 S2, Figure S3, Table S2). We used growth as a proxy for fitness traits (see Methods). Ultimately, this phenotyping step led to the characterization of 148,225 hybrid/trait combinations.

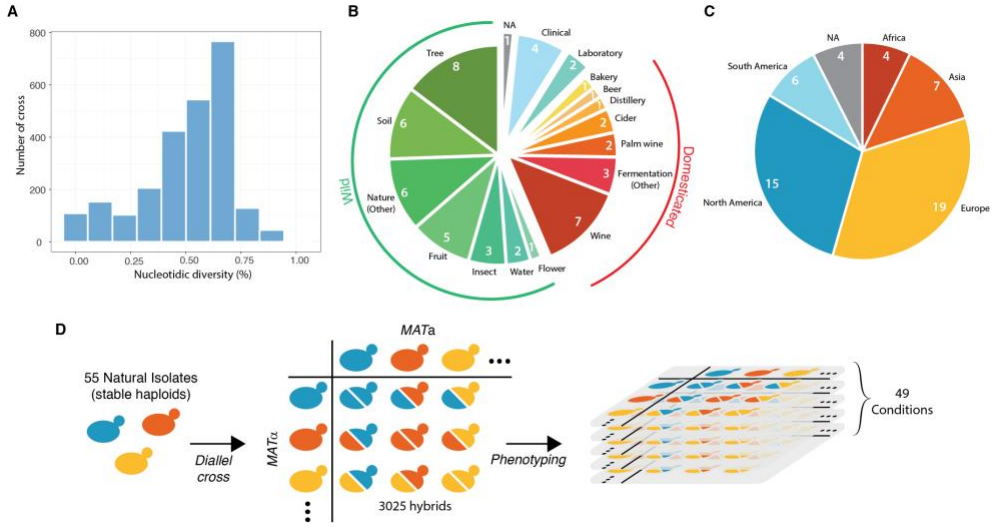


Figure 1. Diversity of the 55 selected natural isolates and diallel design
A. Pairwise sequence diversity between each pair of parental strains. **B.** Ecological origins of the selected strains. **C.** Geographical origins of the selected strains.
D. Generation of the diallel hybrid panel. 55 natural isolates available as both mating types as stable haploids were crossed in a pairwise manner to obtain 3,025 hybrids. This panel was then phenotyped on 49 growth conditions impacting various cellular processes.

Estimation of genetic variance components using the diallel panel

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 Methods). Using GCA and SCA values, we estimated both broad- (H^2) and narrow-sense (h^2) heritabilities for each trait (Figure 2). 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 $\mu g.ml^{-1}$) 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 $\mu g.ml^{-1}$) 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 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, 2015). Nonetheless, non-additive effects still plain a third of the observed phenotypic variance.

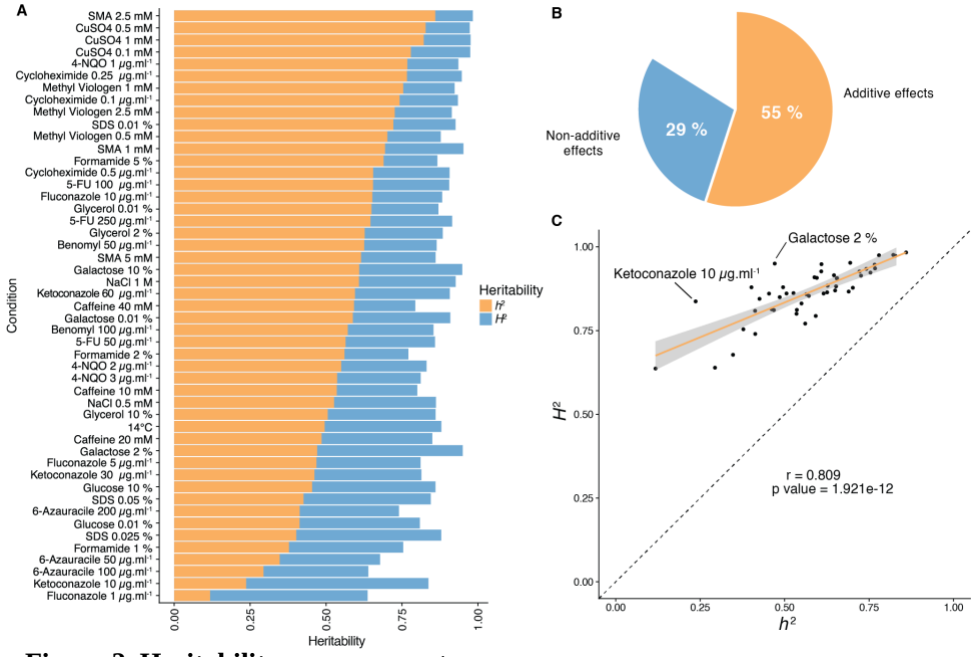


Figure 2. Heritability measurements

A. Orange bars represent the narrow-sense heritability h2 for each condition while blue bars represent broad-sense heritability H2. The difference between H2 and h2 depicts the part of variance due to non-additive effects. **B.** Overall mean additive and non-additive effects for every tested growth condition. **C.** Representation of H2 as a function of h2 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.

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 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/trait combination (Lippman and Zamir, 2007), *i.e.* additivity, partial or complete dominance towards one or the other parent and finally overdominance or underdominance (Figure 3A-B, see Methods).

Only 17.4% of all hybrid/trait 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 3C). Interestingly, these events are not as rare as previously described (Zorgo 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 trait/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 traits (33 out of 49) showed a prevalence of additive effects (Figure 3F).

However, 17 traits were not predominantly additive throughout the population. Indeed, a total of 12 traits 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 5 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 impact of epistasis on phenotypic variation.

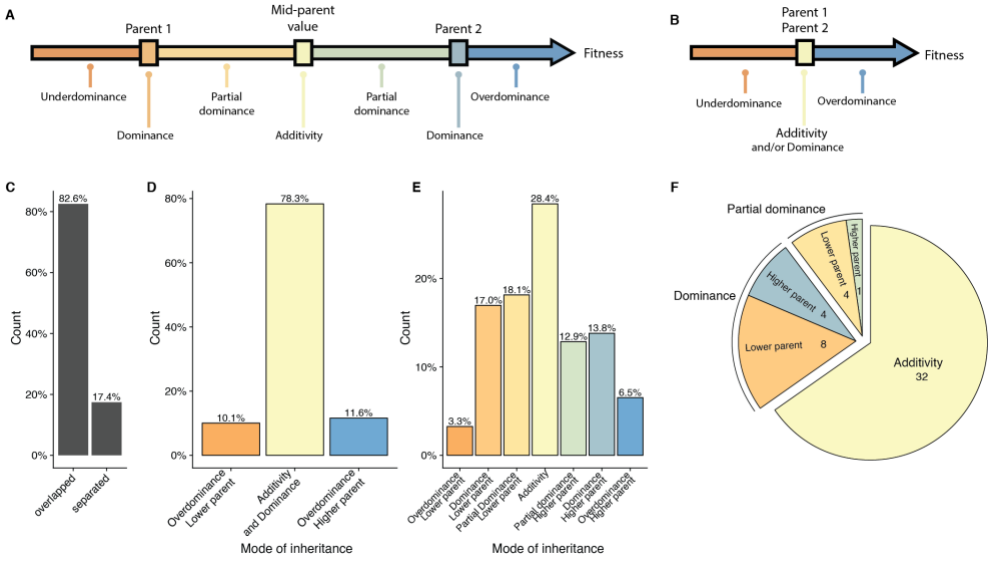


Figure 3. Mode of inheritance

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.

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.

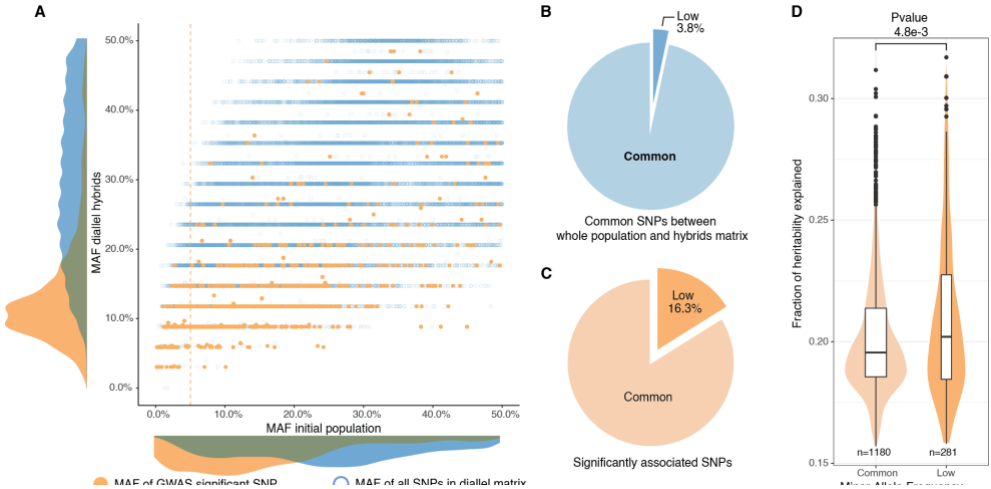


Figure 4. Rare and low-frequency variants detection
A. Comparison of MAF for each SNP between the whole population (1,011 strains) and the hybrid diallel matrix used for GWAS. Hollow blue circles represent the MAF of all SNPs common to the initial population and the diallel hybrids (31,632). Full orange 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 $-7.9e^{-3}$ $-1.4e^{-3}$).

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 Methods, Table S1). 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 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 Methods). We used a classical additive model, encoding for SNPs where linear relationship between trait and genotype is assessed, *i.e.* 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 traits with FaST-LMM (Lippert et al., 2011; Widmer et al., 2014). Overall, GWAS revealed 1,723 significantly associated SNPs (Table S4) by detecting from 2 to 103 significant SNPs by condition, with an average of 39 SNPs per trait. Minor allele frequencies of the significantly associated SNPs were determined in the 1,011 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 Methods) 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 respective allele frequency, it is noteworthy that low-frequency variants explained a slightly (but significantly) higher 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 S4A). It is noteworthy that this trend was robust and conserved across the two encoding models implemented, accounting for additive and overdominant effects (Figure S4A).

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 S4), 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 178 quantitative trait genes (QTG) (Peltier et al., 2019) and we found that 27 of the genes we detected are included in this list (Figure S4C). In addition, 23 associated genes were also found as overlapping with a recent large-scale linkage mapping survey in yeast (Bloom et al., 2019) (Figure S4C). We then asked whether the associated genes were enriched for specific gene ontology (GO) categories (Table S3). 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 (TTG (Leu) → TTA (Leu)) for the first position and a non-synonymous mutation (GAA (Glu)→ AAA (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.

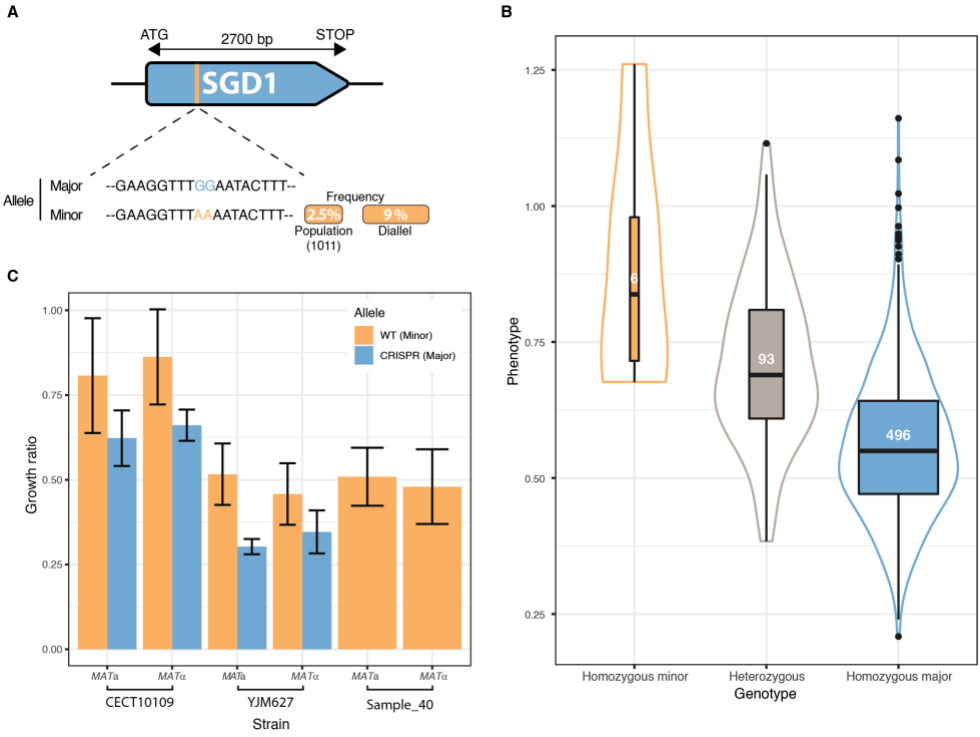


Figure 5. Low-frequency variant functional validation in 6-azauracil 100µg.ml^{-1}
A. Schematic representation of SGD1 with the relative position of the detected SNPs. The minor allele is represented in orange with its MAF in the population and in the diallel cross panel. **B.** boxplot and density plot of the normalized phenotypes for each genotype on 6-azauracil 100 µg.ml-1. Number of observations 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 (4 replicates).

Conclusion

Understanding the source of the missing heritability is essential to precisely address and dissect the genetic architecture of complex traits. 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, 2019; Peter et al., 2018). This observation was corroborated by additional mapped loci via linkage mapping and analyses (Bloom et al., 2019). 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 traits 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.

Publication related to this chapter:

Fournier, T., Abou-Saada, O., Hou, J., Peter, J., Caudal, E., and Schacherer, J. Extensive impact of low-frequency variants on the phenotypic landscape at

Supplementary material

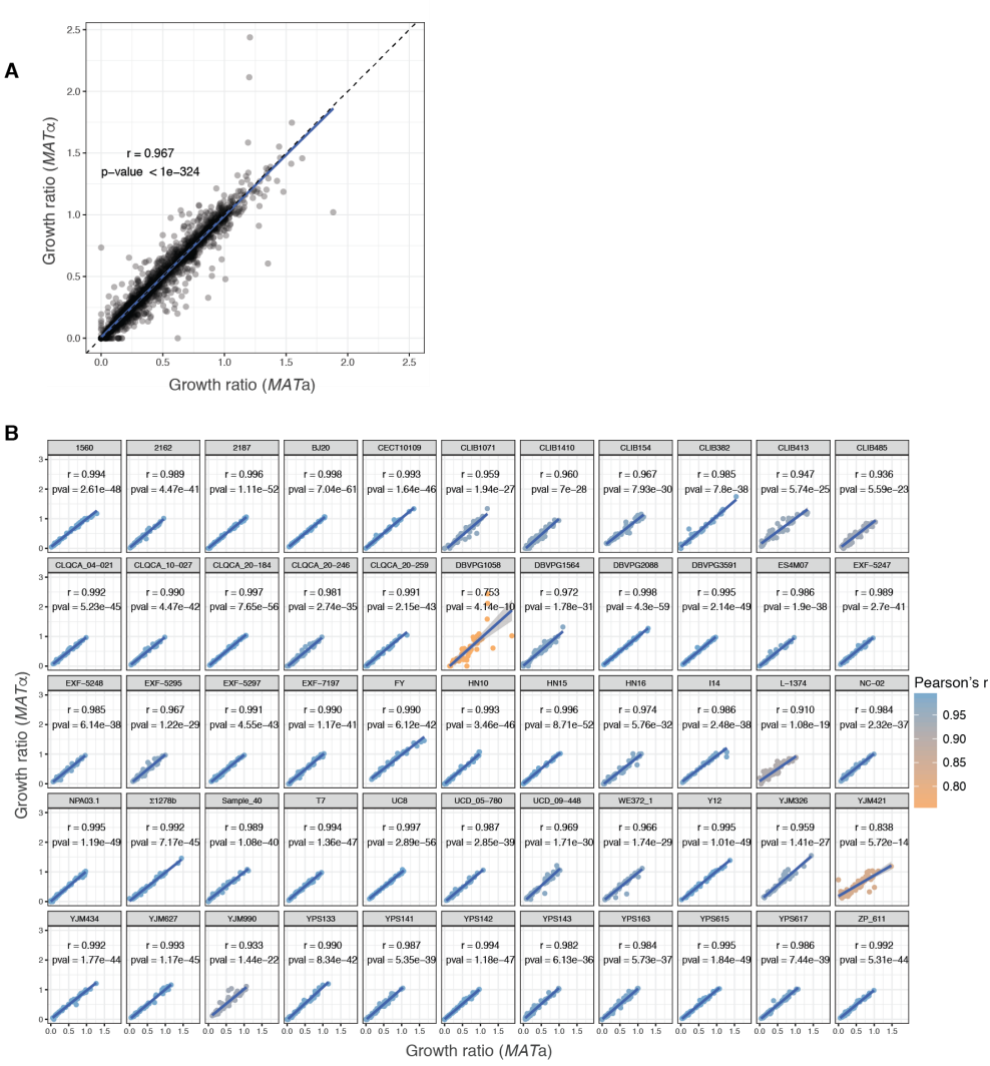


Figure S1. Phenotypic correlation between *MAT*a and *MATα* isolates
A. Correlation between growth ratio of different mating types for all parental strains across all conditions. **B.** Correlation between mating types by strain. Pearson's r and corresponding p-values are indicated for each strain. The growth ratio used is the median of 54 replicates for each strain.

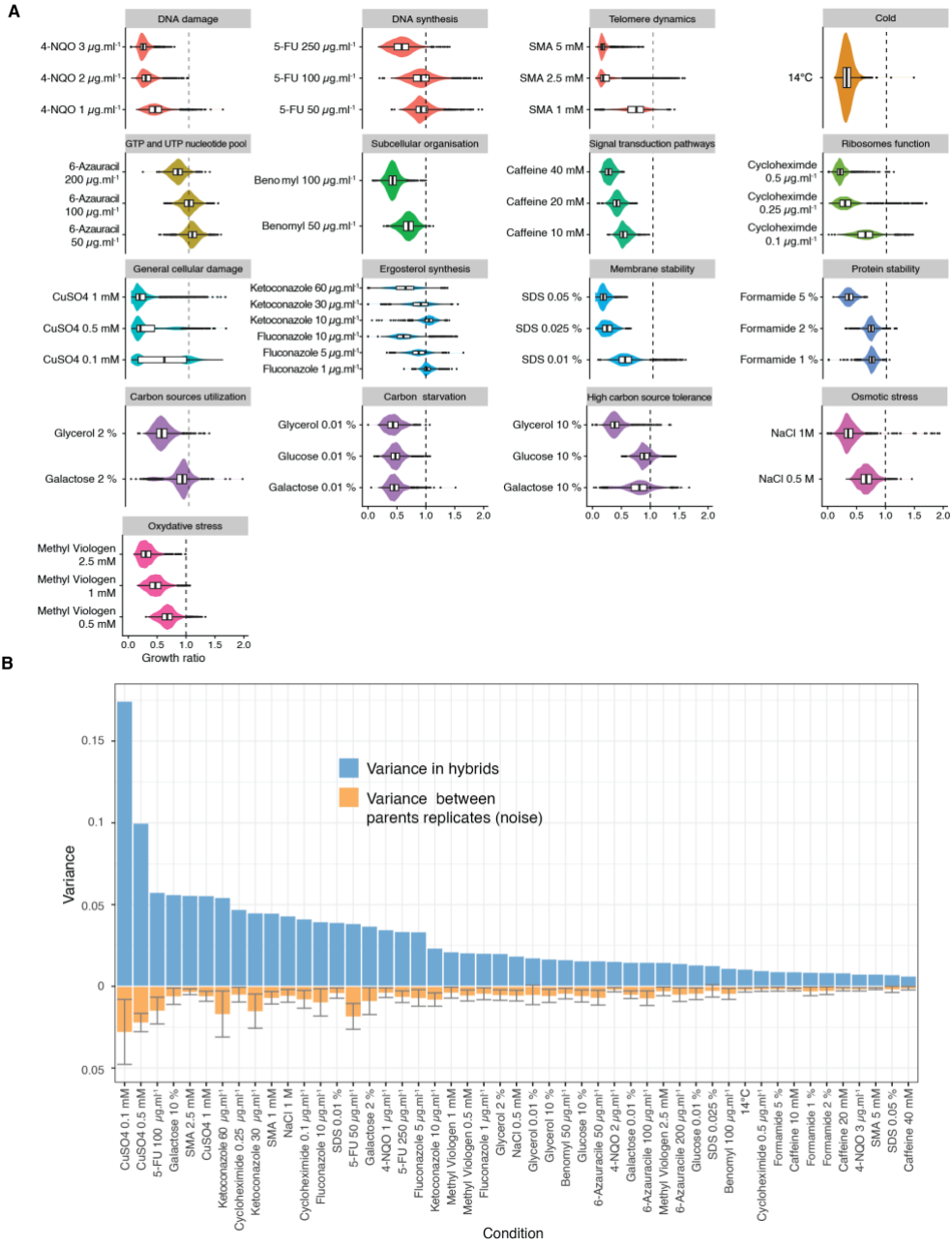

A
DNA damage
4-NQO 3 μg.ml-1
4-NQO 2 μg.ml-1
4-NQO 1 μg.ml-1
DNA synthesis
5-FU 250 μg.ml-1
5-FU 100 μg.ml-1
5-FU 50 μg.ml-1
Telomere dynamics
SMA 5 mM
SMA 2.5 mM
SMA 1 mM
Cold
14°C
GTP and UTP nucleotide pool
6-Azauracil 200 μg.ml-1
6-Azauracil 100 μg.ml-1
6-Azauracil 50 μg.ml-1
Subcellular organisation
Beno myl 100 μg.ml-1
Benomyl 50 μg.ml-1
Signal transduction pathways
Caffeine 40 mM
Caffeine 20 mM
Caffeine 10 mM
Ribosomes function
Cycloheximde 0.5 μg.ml-1
Cycloheximde 0.25 μg.ml-1
Cycloheximde 0.1 μg.ml-1
General cellular damage
CuSO4 1 mM
CuSO4 0.5 mM
CuSO4 0.1 mM
Ergosterol synthesis
Ketoconazole 60 μg.ml-1
Ketoconazole 30 μg.ml-1
Ketoconazole 10 μg.ml-1
Fluconazole 10 μg.ml-1
Fluconazole 5 μg.ml-1
Fluconazole 1 μg.ml-1
Membrane stability
SDS 0.05 %
SDS 0.025 %
SDS 0.01 %
Protein stability
Formamide 5 %
Formamide 2 %
Formamide 1 %
Carbon sources utilization
Glycerol 2 %
Galactose 2 %
Carbon starvation
Glycerol 0.01 %
Glucose 0.01 %
Galactose 0.01 %
High carbon source tolerance
Glycerol 10 %
Glucose 10 %
Galactose 10 %
Osmotic stress
NaCl 1M
NaCl 0.5 M
Oxydative stress
Methyl Viologen 2.5 mM
Methyl Viologen 1 mM
Methyl Viologen 0.5 mM
0.0 0.5 1.0 1.5 2.0
Growth ratio
B
Variance
0.15
0.1
0.05
0
0.05
Variance in hybrids
Variance between parents replicates (noise)
CuSO4 0.1 mM
CuSO4 0.5 mM
5-FU 100 μg.ml-1
Galactose 10 %
SMA 2.5 mM
CuSO4 1 mM
Ketoconazole 60 μg.ml-1
Cycloheximide 0.25 μg.ml-1
Ketoconazole 30 μg.ml-1
SMA 1 mM
NaCl 1 M
Cycloheximide 0.1 μg.ml-1
Fluconazole 10 μg.ml-1
SDS 0.01 %
5-FU 50 μg.ml-1
Galactose 2 %
4-NQO 1 μg.ml-1
5-FU 250 μg.ml-1
Fluconazole 5 μg.ml-1
Ketoconazole 10 μg.ml-1
Methyl Viologen 1 mM
Methyl Viologen 0.5 mM
Fluconazole 1 μg.ml-1
Glycerol 2 %
NaCl 0.5 mM
Glycerol 0.01 %
Glycerol 10 %
Benomyl 50 μg.ml-1
Glucose 10 %
6-Azauracile 50 μg.ml-1
4-NQO 2 μg.ml-1
Galactose 0.01 %
6-Azauracile 100 μg.ml-1
Methyl Viologen 2.5 mM
6-Azauracile 200 μg.ml-1
Glucose 0.01 %
SDS 0.025 %
Benomyl 100 μg.ml-1
14°C
Cycloheximide 0.5 μg.ml-1
Formamide 5 %
Caffeine 10 mM
Formamide 1 %
Formamide 2 %
Caffeine 20 mM
4-NQO 3 μg.ml-1
SMA 5 mM
SDS 0.05 %
Caffeine 40 mM
Condition

Figure S2. Phenotypic variance in hybrids

A. Phenotypic distribution for all hybrids in the different growth conditions. Conditions are organized by type of stress in each panel. **B.** Blue bars show the phenotypic variance of the growth ratio for the hybrids in each condition (mean = 0.027). Orange bars represent the variance due to noise between each plate (mean = 0.006). Noise has been measured as the mean variance of every parental replicates across all plates for each condition (2 replicates per plate, 27 plates, *i.e.* 54 replicates per parental isolate). Error bars represent interquartile range.

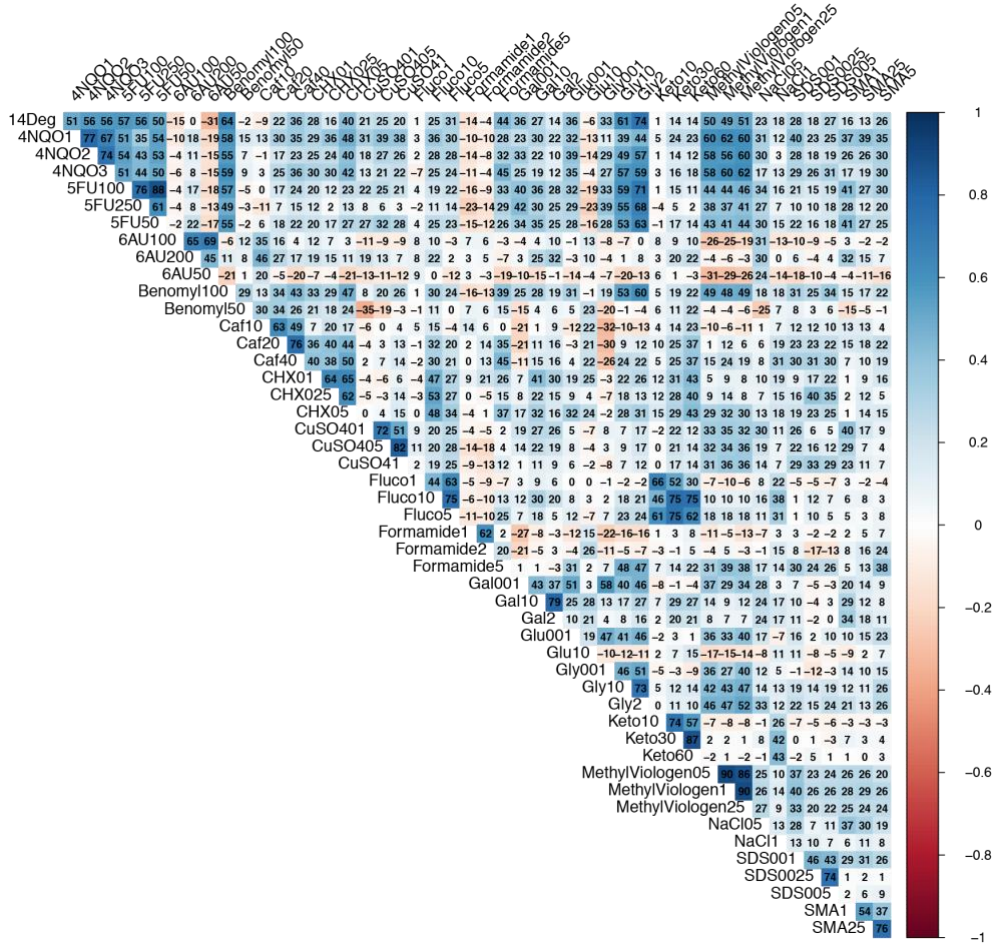


Figure S3. Correlation between conditions

Correlogram of all tested growth conditions. Numbers in each cell represent 100 x Pearson's r value.

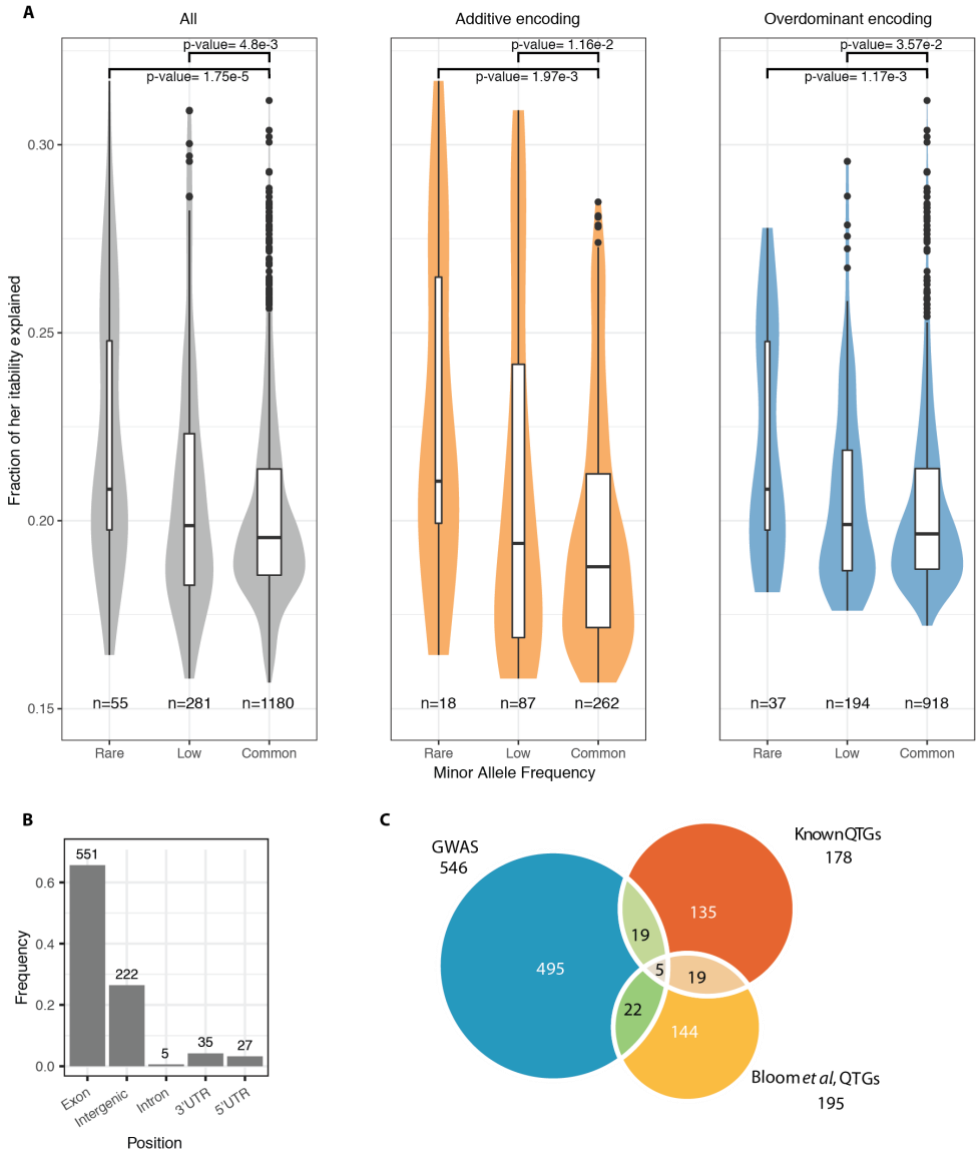


Figure S4. Significantly associated SNPs

A. Variance explained for each significantly associated SNPs, for rare (MAF<1%), low-frequency (MAF<5%) and common (MAF>5%) variants for both encoding models (in grey), additive encoding only (in orange) and overdominant encoding (in blue). All p-values are calculated using a two-sided Mann-Whitney-Wilcoxon test. **B.** Position of the unique significantly associated SNPs. **C.** Venn diagram comparing the overlap between the 546 unique genes in our dataset with the 178 known QTGs (Peltier et al., 2019) and 195 QTGs recently highlighted (Bloom et al., 2019).

Table S1. Strains used for the diallel cross

Strain Name	Isolation	Ecological Origin	Continent	GWAS included
∑1278b	NA	Laboratory	NA	Yes
1560	Manzanilla-Alorena, olive (Noe)	Nature	Europe	Yes
2162	Forest soil, 30C	Soil	Europe	
2187	Forest soil, 30C	Soil	Europe	Yes
BJ20	Bark from Quercus wutaishanica	Tree	Asia	Yes
CECT10109_1b	Prickly pear	Fruit	Europe	Yes
CLIB1071	Cider brewery, dry cider	Cider	Europe	Yes
CLIB1410	Rice wine. Oenology	Fermentation	Asia	
CLIB154_1b	Wine	Wine	Europe	
CLIB382_1b	Beer	Beer	Europe	Yes
CLIB413_1b	Fermenting rice beverage	Fermentation	Asia	Yes
CLIB485	Cider brewery	Cider	Europe	
CLQCA_04-021	Beetle	Insect	South America	Yes
CLQCA_10-027	Grass	Nature	South America	
CLQCA_20-184	Flower from Heliconia sp.	Flower	South America	Yes
CLQCA_20-246	Termite mound	Insect	South America	Yes
CLQCA_20-259	Decaying fruit	Fruit	South America	
DBVPG1058	Baker's yeast	Bakery	Europe	
DBVPG1564	Grape must	Wine	Europe	Yes
DBVPG2088	Cognac	Distillery	Europe	
DBVPG3591_1b	Cocoa beans	Nature	NA	Yes
ES4M07	Fruiting body of Geastrum sp.	Fruit	Asia	Yes
EXF-5247	Seawater in harbour	Water	Europe	Yes
EXF-5248	Seawater in harbour	Water	Europe	Yes
EXF-5295	Kefyr	Fermentation	Europe	
EXF-5297	Mashed pears	Fruit	Europe	
EXF-7197	Quercus sp.	Tree	Europe	Yes
FY4	NA	Laboratory	NA	Yes
HN10	Rotten wood	Nature	Asia	Yes
HN15	Rotten wood	Nature	Asia	Yes
HN16	Soil	Soil	Asia	Yes
I14_1b	Vineyard soil	Wine	Europe	Yes
L-1374	Wine	Wine	South America	
NC_02_b	Exudate from Quercus sp.	Tree	North America	Yes
NPA03.1	Palm wine	Palm wine	Africa	Yes
sample 40	Tree leaves	Tree	Europe	Yes
T7_b	Exudate from Quercus sp.	Tree	North America	
UC8_1b	Wine	Wine	Africa	
UCD_05-780	Beetle	Insect	North America	
UCD_09-448	Olives	Fruit	North America	Yes
WE372_1b	Wine	Wine	Africa	
Y12_1b	Palm wine	Palm wine	Africa	Yes
YJM326_b	Human, clinical	Human. clinical	North America	Yes
YJM421_b	Ascites fluid	Human. clinical	North America	Yes
YJM434_1b	Human, clinical	Human. clinical	NA	
YJM627	Seg, Y55	NA	Europe	Yes
YJM990	Clinical	Human, clinical	North America	
YPS133	Soil beneath Quercus alba	Soil	North America	
YPS141	Soil beneath Quercus velutina	Soil	North America	
YPS142	Surface of Tuber magnatum	Nature	North America	
YPS143	Banana wine	Wine	North America	Yes
YPS163	Soil beneath Quercus rubra	Soil	North America	Yes
YPS615	Quercus sp.	Tree	North America	
YPS617	Quercus sp.	Tree	North America	Yes
ZP_611	Quercus robur	Tree	North America	Yes

Table S2. Phenotyping conditions and their respective type of induced stress

Categories	Sub-categories	Conditions	Abbreviation
Reference		SC	
Cell wall	Membrane stability	SC SDS 0.01%	SDS001
		SC SDS 0.025%	SDS0025
		SC SDS 0.05%	SDS005
	Ergosterol synthesis	SC fluconazole 1 µg/ml	Fluco1
		SC fluconazole 5 µg/ml	Fluco5
		SC fluconazole 10 µg/ml	Fluco10
	Erg synthesis + multiple targets	SC ketoconazole 10 µg/ml	Keto10
		SC ketoconazole 30 µg/ml	Keto30
		SC ketoconazole 60 µg/ml	Keto60
Cold		SC 14°C	14Deg
DNA metabolism	Telomere dynamics	SC sodium (meta)arsenite 1 mM	SMA1
		SC sodium (meta)arsenite 2.5 mM	SMA25
		SC sodium (meta)arsenite 5 mM	SMA5
	DNA damage	SC 4-NQO 1 µg/ml	4NQO1
		SC 4-NQO 2 µg/ml	4NQO2
		SC 4-NQO 3 µg/ml	4NQO3
	DNA synthesis	SC 5-FU 50 µg/ml	5FU50
		SC 5-FU 100 µg/ml	5FU100
		SC 5-FU 250 µg/ml	5FU250
General cellular damage		SC CuSO4 0.1 mM	CuSO401
		SC CuSO4 0.5 mM	CuSO405
		SC CuSO4 1 mM	CuSO41
Metabolism	Carbon sources utilization	SC galactose 2%	Gal2
		SC glycerol 2%	Gly2
	Carbon starvation	SC glucose 0.01%	Glu001
		SC galactose 0.01%	Gal001
		SC glycerol 0.01%	Gly001
	High carbon source tolerance	SC glucose 10%	Glu10
		SC galactose 10%	Gal10
		SC glycerol 10%	Gly10
Osmotic stress		SC NaCl 0.5 M	NaCl05
		SC NaCl 1 M	NaCl1
Oxydative stress		SC methyl viologen 0.5 mM	MV05
		SC methyl viologen 1 mM	MV1
		SC methyl viologen 2.5 mM	MV25
Protein stability		SC formamide 1%	Form1
		SC formamide 2%	Form2
		SC formamide 5%	Form5
Signal transduction pathways		SC caffeine 10 mM	Caf10
		SC caffeine 20 mM	Caf20
		SC caffeine 40 mM	Caf40
Subcellular organisation	Microtubules function	SC benomyl 50 µg/ml	Beno50
		SC benomyl 100 µg/ml	Beno100
Translation	Ribosomes function	SC cycloheximide 0.1 µg/ml	CHX01
		SC cycloheximide 0.25 µg/ml	CHX025
		SC cycloheximide 0.5 µg/ml	CHX05
Transcription	GTP and UTP nucleotide pools	SC 6-azauracil 50 µg/ml	6AU50
		SC 6-azauracil 100 µg/ml	6AU100
		SC 6-azauracil 200 µg/ml	6AU200

Table S3. GO Term associated with the 546 unique genes with a significantly associated SNPs

GOID	TERM	Pvalue	NUM LIST ANNOTATIONS	LIST SIZE	TOTAL NUM ANNOTATIONS	FDR RATE
GO:0050896	response_to_stimulus	5.40E-05	147	546	1277	0.00%
GO:0051716	cellular_response_to_stimulus	6.00E-04	128	546	1112	0.00%
GO:0065007	biological_regulation	1.50E-03	205	546	2026	0.00%
GO:0006950	response_to_stress	2.00E-03	96	546	787	0.00%
GO:0033554	cellular_response_to_stress	4.17E-03	90	546	736	0.00%
GO:0010646	regulation_of_cell_communication	8.23E-03	28	546	147	0.00%
GO:0051179	localization	1.23E-02	162	546	1568	0.00%
GO:0048583	regulation_of_response_to_stimulus	1.74E-02	33	546	195	0.25%
GO:0050794	regulation_of_cellular_process	2.14E-02	159	546	1547	0.22%
GO:0050789	regulation_of_biological_process	2.36E-02	168	546	1656	0.20%
GO:0009966	regulation_of_signal_transduction	2.88E-02	26	546	140	0.18%
GO:0035556	intracellular_signal_transduction	3.28E-02	39	546	255	0.17%
GO:0007154	cell_communication	4.07E-02	61	546	472	0.15%

Table S4. Significantly associated SNPs

This table is available on the online version of the paper on BioRXiv:

https://www.biorxiv.org/content/10.1101/609917v1

References

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

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

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. BioRxiv 607291.

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

Gibson, G. (2012). Rare and common variants: Twenty arguments. Nat. Rev. Genet. 13, 135–145.

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

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

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

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.

Lippman, Z.B., and Zamir, D. (2007). Heterosis: revisiting the magic. Trends Genet. 23, 60–66.

Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature 482, 173–178.

Mackay, T.F.C.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.

Peltier, E., Friedrich, A., Schacherer, J., and Marullo, P. (2019). Quantitative Trait Nucleotides Impacting the Technological Performances of Industrial Saccharomyces cerevisiae Strains. Front. Genet. 10, 683.

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.

Pritchard, J.K. (2002). Are Rare Variants Responsible for Susceptibility to Complex Diseases? Am. J. Hum. Genet. 69, 124–137.

Seymour, D.K., Chae, E., Grimm, D.G., Martín Pizarro, C., Habring-Müller, A., Vasseur, F., Rakitsch, B., Borgwardt, K.M., Koenig, D., and Weigel, D. (2016). Genetic architecture of nonadditive inheritance in Arabidopsis thaliana hybrids. Proc. Natl. Acad. Sci. U. S. A. 113, E7317–E7326.

Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. 99, 139–153.

Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A.S., et al. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. 44, 483–489.

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.

Wainschtein, P., Jain, D.P., Yengo, L., Zheng, Z., TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., et al. (2019). Recovery of trait heritability from whole genome sequence data. BioRxiv 588020.

Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature 526, 82–90.

Widmer, C., Lippert, C., Weissbrod, O., Fusi, N., Kadie, C., Davidson, R., Listgarten, J., and Heckerman, D. (2014). Further improvements to linear mixed models for genome-wide association studies. Sci. Rep. 4, 6874.

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.

Zorgo, E., Gjuvsland, A., Cubillos, F.A., Louis, E.J., Liti, G., Blomberg, A., Omholt, S.W., and Warringer, J. (2012). Life History Shapes Trait Heredity by Accumulation of Loss-of-Function Alleles in Yeast. Mol. Biol. Evol. 29, 1781–1789.

Zuk, O.O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. U. S. A. 111, E455-464.

CHAPTER 2

Species-wide survey of genetic complexity and phenotypic expressivity of traits

Summary

Understanding the genetic basis of traits with the underlying level of complexity and how it varies depending on the genetic background is of prime interest to gain better insights into the genetic architecture of traits. The classical dichotomy existing between monogenic and complex traits is overly simplistic as the inheritance of the trait complexity behaves in a dynamic way depending on the considered genetic background. Indeed, variation of a given trait can be controlled by only one gene in a specific genetic background and have several modifiers in another one. However, no systematic and species-wide assessment of this phenotypic expressivity has been performed yet. To dissect the prevalence of expressivity and the overall genetic complexity of traits at a population-scale, we first generated a large diallel hybrid panel composed of 190 unique hybrids coming from 20 natural isolates representative of the *S. cerevisiae* genetic diversity. For each of these hybrids, a large progeny of 160 individuals (corresponding to 40 full tetrads) was obtained, leading to a total of 30,400 offspring individuals. Their mitotic growth has been assessed on 40 growth conditions inducing various cellular stress. As the phenotypic distribution of the offspring of a given cross allows to infer the inheritance patterns to a trait, we assessed the inheritance patterns for 3,841 cross/trait combinations and revealed that while complex inheritance were the most common, 11% of the cross/traits combinations had their phenotypic variation controlled by a single gene with a large effect and 4% displayed digenic interactions. We identified 26 major effect loci on various traits and parental genetic backgrounds. Measurement of the extent of expressivity was performed by investigating the variation of inheritance patterns throughout all the crosses having one parent who carries one of these loci. We found that trait complexity was highly dynamic and tightly linked to the genetic background. Indeed, 22 out of the 26 major effect loci were subjected to various level of expressivity with one to nine crosses showing departure from monogenic inheritance.

Introduction

The year 1900 has been a keystone for modern genetics with the independent rediscovery of Mendel's law by De Vries, Correns and Tschermak (Correns, 1900; Tschermak-Seysenegg, 1900; De Vries, 1900). This was followed 20 years later by the work of Altenburg and Muller who first dissected a complex traits in *Drosophila* (Altenburg and Muller, 1920). Since then, phenotypes were usually classified as either monogenic if they follow a Mendelian inheritance pattern or complex if the phenotypic diversity is explained by the combined effect of multiple genes. However, this classical view is overly simplistic and does not reflect the true nature of genetic complexity of traits. One lesson learned from both model organism and human genetic studies is that the effect of a given variant can be highly variable across several genetic backgrounds and can be modulated by the combined action of other variants (Antonarakis et al., 2010; Chow et al., 2016; Fournier and Schacherer, 2017; Hou et al., 2016a; Paaby et al., 2015).

Yet, we still lack a comprehensive and complete view of the inheritance patterns of phenotypes in different genetic backgrounds but also and more importantly to understand the dynamic of specific genetic variants in different genetic backgrounds. The underlying genetic complexity of a trait can be assessed by looking at inheritance patterns. To do so, one might first access an important number of descendant and assess their phenotypic distribution. A unique feature of yeast is tetrad analysis. Indeed, when crossing two haploid yeast strains, the resulting diploid cell can then undergo meiosis leading to the generation of a tetrad constituted of four haploid spores, enclosed in an ascus. Therefore, by harvesting and analyzing each segregant of the tetrad it is possible to access the result of each independent meiosis event. Then, the pattern of phenotypic distribution of this offspring relative to its parents as well as the tetrad segregation information can be used to construe the underlying genetic complexity of a phenotype.

By crossing a single *S. cerevisiae* lab strain (namely, ∑1278b) to 41 natural isolates

and phenotyping their offspring on 30 growth conditions impacting various cellular pathways, a first estimation of the monogenic compared to complex inheritance has been previously carried out and highlighted that 8.9 % of the studies cross/trait combinations displayed Mendelian inheritance (Hou et al., 2016a). On top of that, this studies also demonstrated the dynamic of trait complexity depending on the genetic background that a particular variant lies in. Indeed, an isolate containing a variant conferring resistance to cycloheximide and anisomycin was crossed with 20 isolates sensitive to these compounds. Offspring analysis showed that in 30% of the cross, a deviation from a Mendelian inheritance was observed. This expressivity reflected the presence of genetic modifiers in some of the explored genetic backgrounds. However, this study suffered from several biases. First, for the estimation of the prevalence of Mendelian inheritance, as only one strain was systematically crossed to the others, the full breadth of genetic diversity hasn't been explored. Moreover, strong allelic effects that are specific to this particular background might impact several crosses in a similar way thus inducing a bias. Extending this study by performing a "many by many" cross instead of a "one by many" will allow to obtain a systematic and unbiased view of the genetic complexity of traits as well as measuring expressivity for variants with important phenotypic effect. We already showed in the previous chapter how a diallel cross design can help in many ways to understand more about the genetic architecture of traits in a diploid panel.

Here, we combined the power of classical yeast genetic techniques with high throughput phenotyping and machine learning algorithm to get the first species-wide view of genetic complexity of traits but also to investigate expressivity through the lens of genetic complexity in a high number of cross/trait combinations. Twenty *S. cerevisiae* natural isolates that are representative of the entire species diversity were crossed in a pairwise manner to obtain 190 unique hybrids. Then we obtained a large progeny of 160 individual for each of these crosses leading to 30,400

individuals. The phenotyping of this diallel offspring panel on 40 growth conditions impacting different physiological pathways allowed us to analyze the phenotypic distribution and segregation patterns of the progenies of 7,600 cross/trait combinations. We could confidently infer the inheritance pattern for 3,841 of those combinations and found that 11% were following a monogenic inheritance pattern, 4% had an oligogenic inheritance with examples of recessive epistasis and suppressor genes. Most of the cross/trait combinations surveyed (80%) displayed a complex inheritance pattern. Moreover, we could assess the prevalence of expressivity at the population level by following the deviation of monogenic inheritance for strains carrying a major locus.

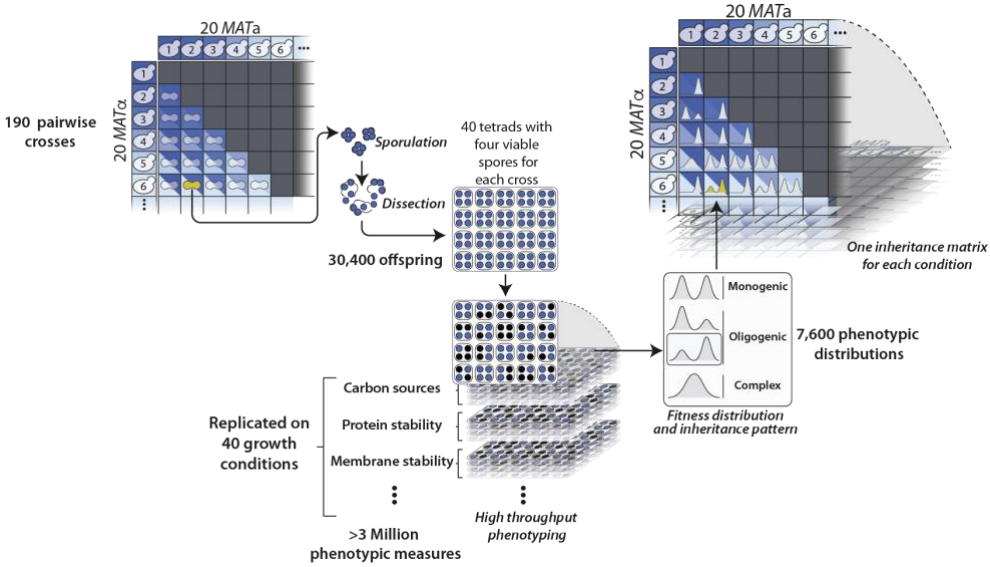


Figure 1. Experimental design of the diallel offspring panel
Experimental design followed to infer inheritance pattern to each cross/trait combination coming from a pairwise cross of 20 natural isolates

Results

Generation of a large offspring at a species-wide level

Out of the 55 natural isolates selected in the diallel hybrid panel used in chapter 1, we wanted to select 20 isolates that would still be representative of *S. cerevisiae* genetic diversity. However, we also wanted to make sure that the selected strains had collinear genomes *i.e.* devoid of any gross chromosomal translocations, as this would strongly impede the offspring viability of the hybrids (Hou et al., 2014). This selection is also necessary because our analysis pipeline is based on segregation patterns of the offspring phenotypes. Segregation analysis is only possible if information for all four spores of a tetrad is available as loci are expected to segregate in a 2:2 ratio during meiosis. To assess their genome collinearity, all 55 haploid strains were crossed with the reference strain S288C and five tetrads (a total of 20 spores) were dissected (Figure S1). Number of viable spores per tetrad has been obtained. All crosses are expected to show mostly tetrads containing four viable spores. However, for example, translocated strains are characterized by a predominance of tetrads with only three or two viable spores. Surprisingly, 28 strains showed viability profiles deviated from full viable tetrads (Figure S1). In these strains, putative translocations are suspected, highlighting the importance of such structural variation in the phenotypic landscape of *S. cerevisiae*. The 20 selected isolates (Table 1) were then crossed in an all by all manner without reciprocal crosses or homozygous crosses leading to a half diallel cross of 190 hybrids. For each of these hybrids, a large progeny of 160 haploids coming from 40 tetrads with four viable spores were obtained, summing up to 30,400 spores. In total, 66,992 spores had to be manually dissected to obtain the expected progeny from fully viable tetrads.

Table 1: Strains used in this study

Strain Name	Isolation	Ecological Origin	Continent	Abbreviated name
YJM627	Seg, Y55	NA	Europe	00
YPS141	Soil beneath Quercus velutina	Soil	North America	05
DBVPG1564	Grape must	Wine	Europe	09
UCD_09-448	Olives	Fruit	North America	11
CLIB1071	Cider brewery, dry cider	Cider	Europe	17
WE372	Wine	Wine	Africa	18
DBVPG1058	Baker's yeast	Bakery	Europe	42
YJM421	Ascites fluid	Human. clinical	North America	53
YJM434	Human, clinical	Human. clinical	NA	54
CECT10109	Prickly pear	Fruit	Europe	60
Y12	Palm wine	Palm wine	Africa	65
CLIB154	Wine	Wine	Europe	67
NPA03.1	Palm wine	Palm wine	Africa	70
HN16	Soil	Soil	Asia	74
EXF-7197	Quercus sp.	Tree	Europe	76
CLQCA_20-184	Flower from Heliconia sp.	Flower	South America	78
YPS615	Quercus sp.	Tree	North America	80
1560	Manzanilla-Alorena, olive	Nature	Europe	82
FY	NA	Laboratory	NA	83
∑1278b	NA	Laboratory	NA	84

Offspring viability and reproductive isolation

One of the first phenotypes that we analysed was the viability of the offspring to look for reproductive isolation. Reproductive isolation can either take place before mating (pre-zygotic) preventing formation of a viable zygote or after mating (post-zygotic) leading to reduced offspring viability. All the crosses performed were viable suggesting no pre-zygotic reproductive isolation in the studied population. Under normal circumstances, with no post-zygotic reproductive isolation, viability of the progeny of a given cross should lie between 85 and 100% with lethality possibly being attributed to experimentation (*e.g.* zymolyase digestion or spore manipulation with the micro-needle) or to random errors in chromosome segregation during meiosis (Chu and Burgess, 2016). Any deviation from this implies the presence of post-zygotic reproductive isolation between two parental isolates. Many factors can lead to a drop in offspring viability (Hou et al., 2016b) such as chromosomal

rearrangements (Charron et al., 2014; Hou et al., 2014) as well as genetic incompatibilities (Bikard et al., 2009; Hou et al., 2015; Seidel et al., 2008). However, the parental strains were selected based on their genome collinearity, chromosomal rearrangements are theoretically out of the picture. In our 190 crosses, we observed a mean viability of 78% with levels ranging from 48% to 96% (Figure 2A, Figure S2). In total, 72% of the crosses displayed viability below 85% suggesting a strong prevalence of reproductive isolation in our panel. Genetic divergence between parental isolates goes from 0.04% up to 0.95%. Interestingly, we could observe a moderate but significant anti-correlation between offspring viability and genetic distance between the parental isolates (Figure 2B). This result suggests that the genetic divergence level between parental strains is a driver of intraspecific reproductive isolation. A possible explanation for this phenomenon is the result of the mismatch repair (MMR) system through its anti-recombination effect which detects mismatches between two homeologous chromosomes and will prevent formation of crossover (Iyer et al., 2006). Anti-recombination leads to poor chromosomal segregation during meiosis and ultimately to lethal aneuploidies (Chu and Burgess, 2016). We also observed this phenomenon in other yeast species with higher level of intraspecific genetic diversity such as *Lachancea kluyveri* which goes up to 3% of divergence between the most diverged strains (Figure 2C). More importantly, the slope of the curve is almost exactly the same between the two species, suggesting that the effect of the MMR on anti-recombination depending on the level of heterozygosity is linear and conserved across species (Figure 2C). However, crosses with viability being strongly deviated from this linear regression are also observed indicating the presence of other viability impeding mechanisms in our panel.

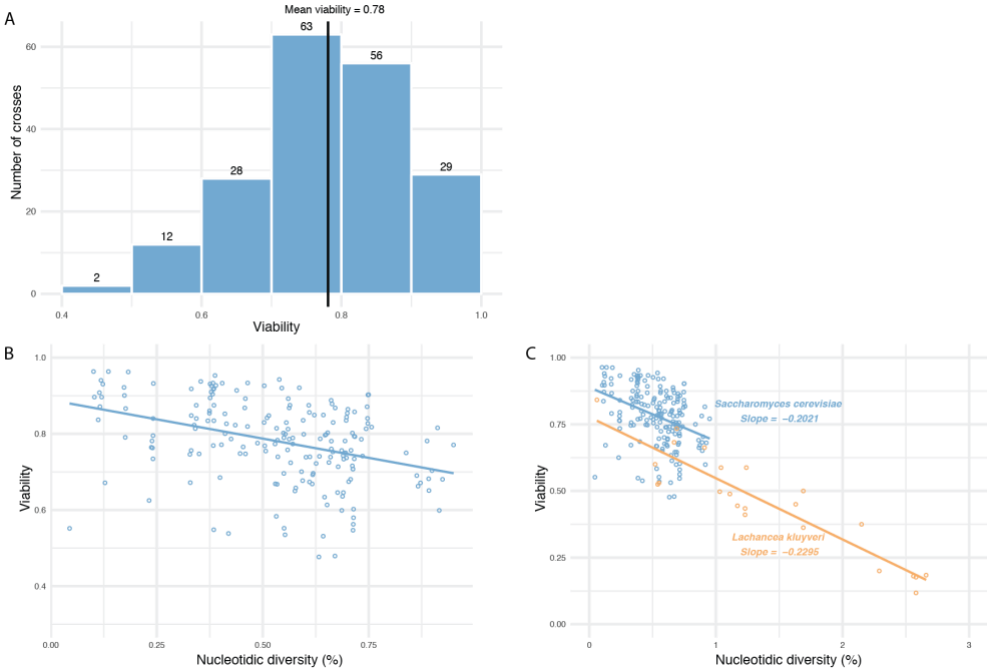


Figure 2. Reproductive isolation in the diallel panel

A. Viability of the offspring coming from the 190 crosses. **B.** Relationship between genetic distance separating the two parents and the offspring viability in the 190 crosses. Correlation assessed by Pearson's r, r = -0.38, p-value = 8.35e-8. Blue line is the fitted linear model with a slope of -0.2. **C.** Relationship between genetic distance separating the two parents and the offspring viability in both *S. cerevisiae* (blue) and in *Lachancea kluyveri* (orange) which has a much wider genetic diversity (up to 3%). Regression line for *L. kluyveri* is -0.22, Pearson's r = -0.92, p-value=3.12e-14.

Table 2: Conditions used for phenotyping the diallel offspring panel

Categories	Sub-categories	Conditions	Abbreviation
Reference		SC	
Cell wall	Membrane stability	SC SDS 0.01%	SDS001
		SC SDS 0.025%	SDS0025
		SC SDS 0.05%	SDS005
	Ergosterol synthesis	SC fluconazole 1 µg/ml	Fluco1
		SC fluconazole 5 µg/ml	Fluco5
		SC fluconazole 10 µg/ml	Fluco10
	Erg synthesis + multiple targets	SC ketoconazole 10 µg/ml	Keto10
		SC ketoconazole 30 µg/ml	Keto30
		SC ketoconazole 60 µg/ml	Keto60
Cold		SC 14°C	14Deg
DNA metabolism	Telomere dynamics	SC sodium (meta)arsenite 1 mM	SMA1
		SC sodium (meta)arsenite 2.5 mM	SMA25
		SC sodium (meta)arsenite 5 mM	SMA5
	DNA synthesis	SC 5-FU 50 µg/ml	5FU50
		SC 5-FU 100 µg/ml	5FU100
		SC 5-FU 250 µg/ml	5FU250
General cellular damage		SC CuSO4 0.1 mM	CuSO401
		SC CuSO4 0.5 mM	CuSO405
		SC CuSO4 1 mM	CuSO41
Metabolism	Carbon sources utilization	SC galactose 2%	Gal2
		SC glycerol 2%	Gly2
	Carbon starvation	SC glucose 0.01%	Glu001
		SC galactose 0.01%	Gal001
		SC glycerol 0.01%	Gly001
	High carbon source tolerance	SC glucose 10%	Glu10
		SC galactose 10%	Gal10
		SC glycerol 10%	Gly10
Osmotic stress		SC NaCl 0.5 M	NaCl05
Oxydative stress		SC methyl viologen 0.5 mM	MV05
		SC methyl viologen 1 mM	MV1
		SC methyl viologen 2.5 mM	MV25
Protein stability		SC formamide 1%	Form1
		SC formamide 2%	Form2
		SC formamide 5%	Form5
Signal transduction pathways		SC caffeine 10 mM	Caf10
		SC caffeine 20 mM	Caf20
		SC caffeine 40 mM	Caf40
Translation	Ribosomes function	SC cycloheximide 0.1 µg/ml	CHX01
		SC cycloheximide 0.25 µg/ml	CHX025
		SC cycloheximide 0.5 µg/ml	CHX05

Inferring inheritance patterns

The main objective of this work has been to infer complexity level of traits at a population scale and assess its dynamic across multiple genetic backgrounds. To do so, we first conducted a large-scale phenotyping of the whole panel of 30,400 haploid progeny coming from 190 hybrids. We selected 40 conditions impacting various cellular pathways (Table 2) and measured their mitotic growth ability on solid media by assessing colony sizes. From more than three million phenotypic measurements grouped for each cross and condition (trait), we obtained 7,600 phenotypic distributions of haploid progenies *i.e.* one distribution for each cross/trait combination. The inheritance pattern reflects the genetic complexity of a trait in a given cross between two specific genetic backgrounds (Figure 3).

The simplest case of genetic complexity is the mendelian inheritance. It can be seen for a cross where the phenotypic variation is controlled by only one locus with each of the two parental strains bearing a different allelic version generating distinct phenotypes. After going through meiosis, as loci segregate randomly, half of the offspring will inherit the allele of one parent while the other half will inherit the other allele. This will translate on the phenotypic level as a bimodal distribution of the trait with each mode enclosing half of the progeny and centered on one parental phenotypic value. This can be confirmed by the analysis of the segregation pattern in the tetrad where all tetrads should have two spores in each mode (Figure 3B).

One main advantage of looking at the progeny's phenotypic distributions is the ability to distinguish between various types of digenic interactions. Indeed, distinction between the presence of modifier genes such as suppressor (a gene masking the effect of another) and recessive epistatic interactions can be readily ascertained. Any deviation from a Mendelian 2:2 segregation of the phenotype depicts a change of the underlying genetic complexity with the presence of a genetic

interaction. In both cases of suppressor and epistasis, a bimodal distribution of the progeny's phenotype will still be observed but each cluster won't be equally represented. A main cluster will encompass roughly 75% of the offspring while the other retains about 25%. Moreover, when looking at the segregation of the phenotype in the tetrads, a maximum of tetrads with three spores in the main cluster is expected as this will represent the tetratype (Figure 3C-D). Distinction between recessive epistatic interaction and the specific case of suppressor is achieved by the position of the parental isolates relative to the two modes of the distribution. Indeed, the scenario where the two parents are centered on the main cluster is characteristic of an epistatic interaction with the auxiliary cluster being formed with the offspring carrying the two interacting alleles (Figure 3D). In the case of a suppressor, one parent is centered on the main cluster and the other parent displays a phenotype similar to the auxiliary cluster (Figure 3C).

On the other end of the complexity spectrum is the complex inheritance pattern when a phenotypic variance in a given cross is controlled by multiple loci. Under the assumption that one trait is governed by several (more than two) genes with small effects, the parental combinations will be shuffled in the progeny leading to a normal distribution of the phenotype (Figure 3A). In the case of a complex inheritance, no particular clustering of the phenotype can be done with a unimodal phenotypic variation of the offspring so tetrad analysis is uninformative.

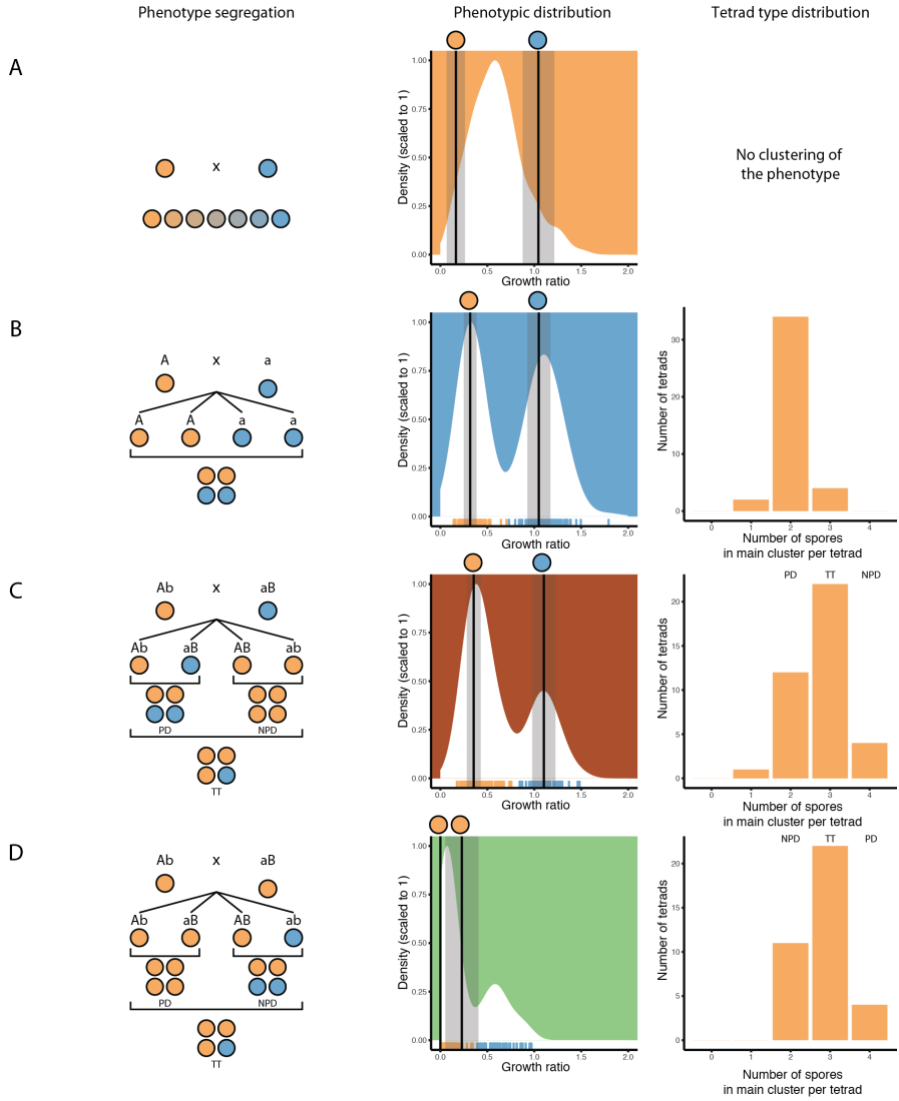


Figure 3. Inferring inheritance based on offspring phenotypic segregation
Left panel shows the expected phenotypic segregation in each type of tetrad, Parental Ditype (PD), Non-Parental Ditype (NPD) and Tetratype (TT). The middle panel is the density of the phenotypic distribution of the offspring. Black lines are the median of the parental phenotypes. Grey boxes show the median absolute deviation of the parental phenotypic values. Bars in the bottom are the phenotypic value for each descendant and are color coded depending on the cluster they belong to. Right panel is the segregation pattern of the spores in each tetrad with their respective tetrad type PD, NPD or TT. **A.** Example of a complex inheritance. **B.** Monogenic inheritance. **C-D.** oligogenic inheritance with **C.** a suppressor and **D.** a recessive epistasis

Framework of the analysis of inheritance patterns

To infer the complexity level for each of the cross/trait combination, we based our analysis on a manually constructed decision tree to classify distributions into different inheritance categories based on their underlying genetic complexity (Figure S3). Yet, the first step of this process is the determination of unimodality vs. bimodality of the distribution. This distinction is far from trivial and required to be assessed in a very specific manner. To do so, we used a machine learning algorithm, more precisely, we build a random forest classifier.

This method belongs to the class of supervised machine learning meaning that it requires an *a priori* learning. Applied to our problem, this meant that a subset of the distributions first had to be manually annotated as bimodal or unimodal. These manually annotated distributions then served as a training set. The construction of the training set is explained in Methods. During training, we give the model the true (expected) output (bimodal or unimodal) for a representative subset of our dataset (*i.e.* the training set). The model then learned to distinguish between the two types of distributions based on the available variables. A decision tree computes the best predictors to achieve the most accurate estimates. A single decision tree is not very accurate by itself, especially for data with important variances like ours. The idea behind a random forest is then to combine a large number of all those decision trees (*i.e.* a forest) having a weak prediction power by themselves to obtain a far better prediction power in the end.

Concretely, with our dataset, as random forest expects only one observation with multiple descriptive variables, our distributions had first to be summarized by a certain number of values. We first fitted a mixture model of two gaussian distribution with an Expectation Maximization (EM) algorithm which estimated the mean, variance and proportion of each of the two clusters of the mixture model (Figure 4). From these values, 13 different variables were computed to facilitate the distinction between unimodal and bimodal distributions (See methods). These were then fed

into the random forest. A majority voting on the output was done to attribute the final class, *i.e.* the modality of the input distribution (Figure 4).

Once the modality of the distribution has been assessed, information about the phenotypic value of the parental isolates and analysis of segregation of the phenotype in each tetrad allowed for a final classification of each cross/trait combination to a complexity level (Figure S3). Combining all of these parameters allowed us to infer the complexity level for each cross/traits combination by making a difference between a monogenic inheritance pattern, several types of oligogenic phenotypic distributions and complex inheritance patterns (Figure S3).

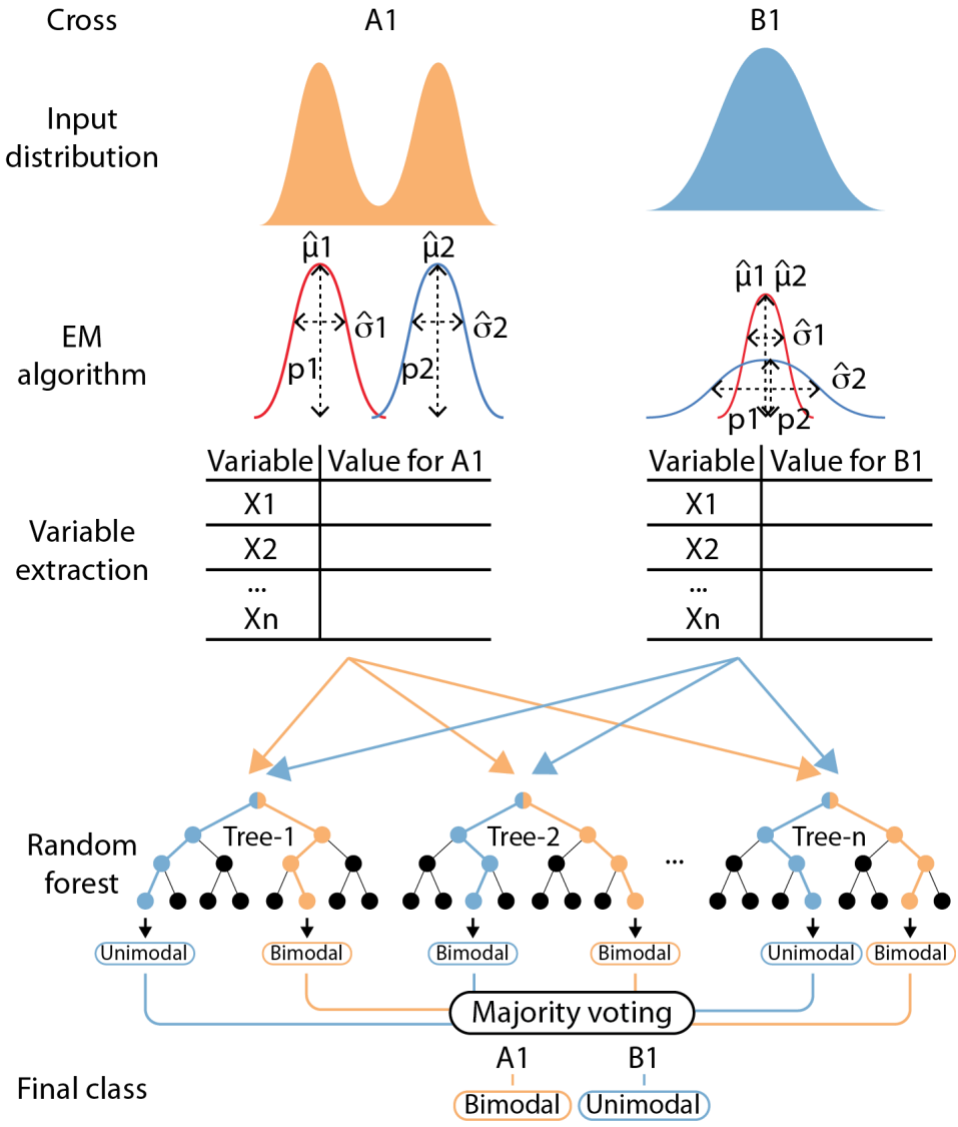


Figure 4. Workflow to classify a distribution as bimodal or unimodal
The input phenotypic distribution of the offspring from one cross is first fed into a EM algorithm to fit two normal distributions on the distribution and estimate the parameters of each. From these parameters, 13 variables are computed which will then be used to run the random forest. Several trees are randomly created and each will output a result regarding the modality of the input distribution. The final classification will be achieved by majority voting of all the trees.

To assess the effectiveness and precision of our random forest model, we manually annotated for bimodality or unimodality seven conditions (caffeine 10 $mg.ml^{-1}$, cycloheximide 0.5 $mg.ml^{-1}$, $CuSO_4$ 0.1 mM, $CuSO_4$ 0.5 mM, galactose 2%, methylviologen 1 $mg.ml^{-1}$ and NaCl 0.5 M) with various levels of bimodality. We then compared the output of the random forest with the manually annotated distributions. This resulted in one confusion matrix for each condition (Figure S4A). We then computed the Negative Predictive Value (NPV), the precision, the sensitivity and specificity for each confusion matrix (Figure S4B). While NPV, precision and specificity are always very high with means of 92.7%, 96.2% and 95.5% respectively, sensitivity is more variable with a mean of 79.1%. This indicates that although there are few false positive, there is sometimes a significant amount of false negative meaning that distributions that are actually bimodal are not detected as such. Although not being extremely biased, our model still has flaws that would need to be addressed by improving sampling method for the training test (See methods) or adding new descriptive variables such as p-values of multimodality tests such as the Hartigan's Dip test (Hartigan and Hartigan, 1985) or the Silverman test (Silverman, 1981).

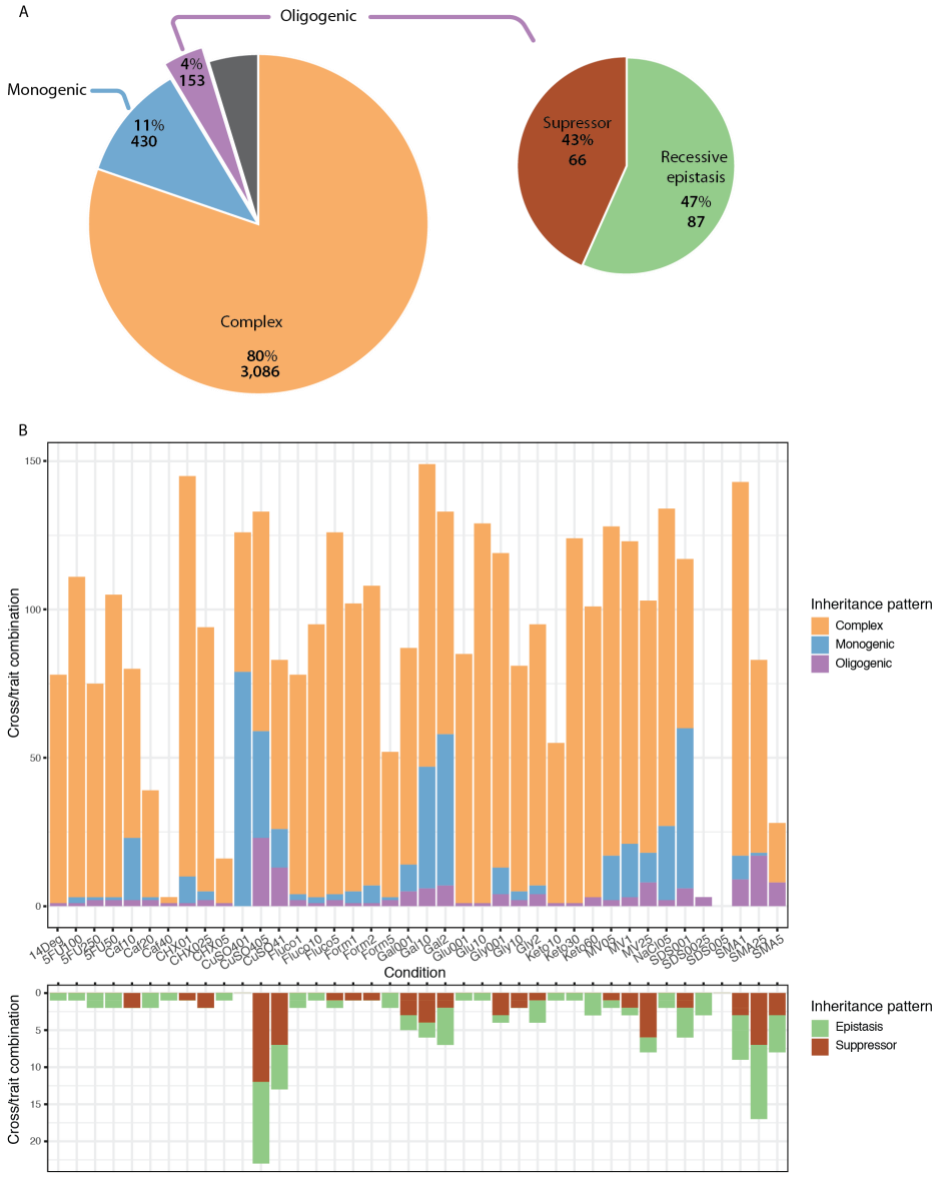


Figure 5. Overview of the inheritance patterns

A. Global repartition of the inheritance patterns for the 3,841 cross/trait combinations. **B.** Condition-wise repartition of the inheritance patterns. Bottom panel shows the separation of oligogenic inheritance patterns between epistasis and suppressors.

Global picture of inheritance patterns

In order to automatically infer genetic complexity in the 7,600 distributions, some criteria first needed to be fulfilled. The two parental isolates from which the cross is made need to have distinct phenotypic values. Indeed, in this case, when looking at the phenotypic distribution of the offspring, no distinction could be ascertained between a bimodal distribution with each mode centered on a parent and a normal distribution (Figure S3). To ensure a good separation between the two parents of each cross, we filtered out all those cross/trait combinations that had an absolute difference between the two parental phenotypic value smaller than the experimental noise (see Methods). Although this filtering step removed 50% of the overall distributions, it allowed for a more robust extrapolation of the underlying genetic complexity. In some conditions (SDS 0.05% and 0.025%, caffeine 40 mM and cycloheximide 0.5 mg/ml), this filter removed most if not all of the distributions because these conditions were very stringent and just a few of the offspring were actually fit enough to grow but all parents and most of the offspring were dead. However, this illustrates how reshuffling of loci can create rare but strong allelic combinations leading to extreme phenotypes, very far from parental phenotypic values.

We classified the remaining 3,841 distributions in one of three complexity level: monogenic, oligogenic and complex (Figure 5, Figure S3). Overall, 80.3% of the considered distributions displayed inheritance patterns corresponding to a complex inheritance pattern (Figure 5A). In the meantime, 11.2% appear as monogenic and only 4% as oligogenic (Figure 5A). The remaining 4.5% failed to be sorted into one of the previous categories for various reasons, either the parents could not be confidently attributed to one cluster or the tetrad segregation phenotype could not result in a confident classification. These results confirm the fact that inheritance patterns are mainly complex but also that in a non-negligible number of cases, one gene is actually responsible for most of the observed genetic variance. However, this

overview can be completed by the fact that this repartition of the complexity is highly dependent on the surveyed conditions. Indeed, extensive variation in the complexity repartition was observed between the conditions (Figure 5B). For example, conditions like copper sulfate, SDS or galactose show high proportion of monogenic inheritance (up to 58% in $CuSO_4$ 0.1 mM). Contrarily, the resistance to ketoconazole or growth at 14°C showed no such trend, suggesting either that phenotypic variation for some traits have a simpler genetic basis or that high effect variants are common in these particular traits. However, these results were expected because of previous population-scale phenotyping (Fournier et al., 2019; Peter et al., 2018) which already showed a bimodal distribution of the phenotype in the whole natural population of *S. cerevisiae* for traits such as copper sulfate, galactose or NaCl. In copper sulfate, we also know from GWAS analysis (Peter et al., 2018) that the main component of phenotypic variation is a copy number variation of the *CUP1* gene.

We further dissected the 153 distributions corresponding to an oligogenic inheritance and could highlight several types of digenic interactions (Figure 5). First, we detected 87 cases of recessive epistasis. In 66 cross/trait combinations, inheritance patterns suggest the presence of a suppressor. However, our model might underestimate the real number of oligogenic distributions because distribution with two clusters of unequal repartitions are more delicate to detect if they are close to each other. Indeed, the two distributions might easily merge if their standard deviation is high due to the combined effect of multiple genes with small effects.

Condition dependent major effect loci

One of the main advantages of using a diallel design is that we can follow the effect of a genetic variant across several genetic backgrounds spread across the genetic diversity of the whole population. When looking at genetic complexity in all the crosses sharing one parent, we could infer the presence of a major locus with high phenotypic impact. Such variant is expected to mostly lead to a monogenic inheritance in the offspring of each cross involving this particular strain. For that, we introduced a "high effect score". By taking all crosses sharing a common parent in a given condition, we computed this score as the proportion of crosses displaying a monogenic inheritance. For instance, out of 19 crosses sharing a given parent under the same growth condition, if 18 display a monogenic inheritance, then the score of the parent would be 18/19 = 0.95. The closer it is to 1, the higher the probability of having a major locus in this strain. By looking at scores above 0.5 (suggesting that at least half of the crosses with the same parental strain display Mendelian inheritance in their offspring) a total of 26 high effect variants were found (Figure 6). They were spread throughout nine conditions and present in 17 strains suggesting that almost each strain has at least one high impact variant. Conditions such as CuSO4 (0.1 mM) or SDS (0.01%) display respectively 9 and 7 high effect variants. With these information alone, it is not possible to say if the causal variant are the same between all the strain or not. For copper sulfate, as stated before, one reason for this high number of strains carrying variant with large phenotypic effect might just be a higher number of *CUP1* copies. Opposite to that, we can see conditions such as Methyl Viologen (MV) or NaCl where only one strain seems to carry a high phenotypic effect variant. This suggests the presence of a low frequency variant. In these cases, mapping of the causal variant should be performed by bulk segregant analysis. Once mapped, minor allele frequency of the variant could be easily determined within the 1,011 reference panel.

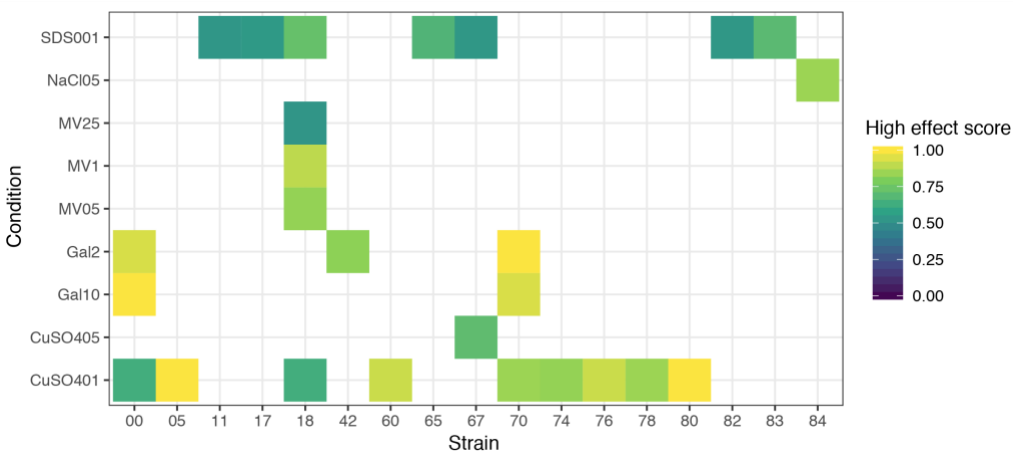


Figure 6. Variants with high phenotypic impact
Each tile represents the presence of a detected major locus in a given parental strain and condition. Color of the tile represents the high effect score.

Genetic background and expressivity

As stated, if a variant with a high phenotypic impact is present in one strain, most of the crosses involving this strain as a parent should display a monogenic inheritance pattern. Interestingly, only four strains had a high impact score of 1 (Figure 6) suggesting that the complexity level in crosses carrying such variants remains unchanged regardless of the background it lies in. This highlights that most of the high effect variants display various modifications of the genetic complexity depending on the genetic backgrounds they are in. The presence of such expressivity implies the presence of other modifier genes contributing to the phenotype and modifying the expected phenotypic outcome of the initial causal variant. We followed the effect of each of the 26 variant across all the 20 genetic background to assess the fraction of changes in inheritance patterns. These changes can be from monogenic to oligogenic with the effect of a suppressor, or even from monogenic to complex with several other loci combining their effect with the variant of interest.
The crosses involving the strains carrying high effect variants display between zero and nine cases with inheritance patterns deviating from monogenic (Figure 7A).

If we follow the expressivity of the major effect locus present in the strain WE372_1 (number 18, Figure 6-7) in different concentrations of methylviologen, we detect 2, 2 and 9 cases of phenotypic distribution displaying deviation from the expected monogenic inheritance in media supplemented with 0.5, 1 and 2.5 $mg.ml^{-1}$ of methylviologen, respectively (Figure 7B). However when looking at their inheritance pattern, we can see that in the first concentration, two distributions are classified as complex, in the second the two increases in complexity correspond to suppressors and in the higher concentration, six inheritance pattern indicated the presence of suppressors and three complex distributions were detected (Figure 7B). The crosses classified as complex in the lowest and highest concentration might just be misclassifications. When only focusing at the suppressors, we can see a progression as none are detected in the first concentration, then two in methylviologen 1 $mg.ml^{-1}$ and finally up to six in the highest concentration. This suggests a threshold dependent effect for the suppressor (or suppressors) of this phenotype.

This example also allows to understand that expressivity level described here might be overestimated in some cases because of spurious classifications of distributions as complex. As in the lowest concentration, parents are relatively close to each other, which impacts the accurate detection of bimodal distribution in the offspring. Repeating the phenotyping with more replicates per offspring for the cases displaying high expressivity might help improving confidence in the detection of complexity modifications by reducing the noise in the phenotype measurement.

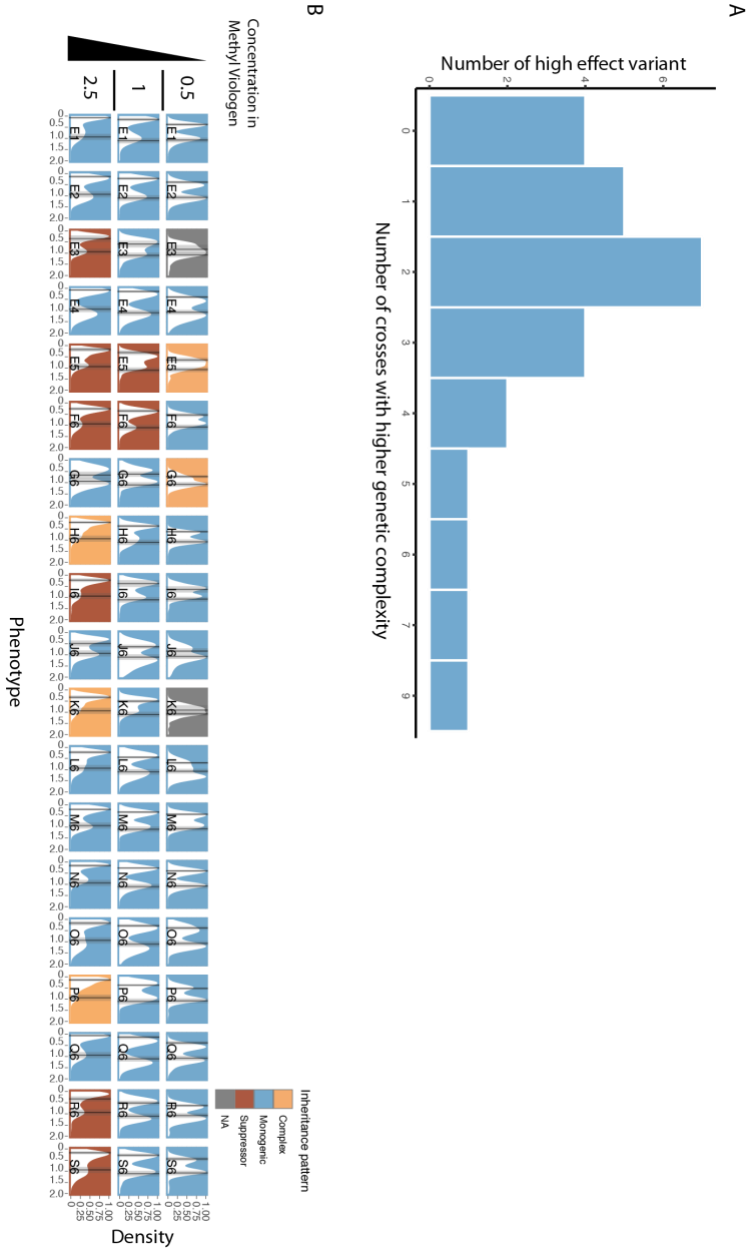


Figure 7. Expressivity of high effect variants

A. Number of inheritance pattern showing deviations from monogenic inheritance for each major locus.

B. All offspring phenotypic distribution for the 19 crosses involving the isolate WE372_1 as a parent in the three media supplemented with different concentrations of methyl-viologen. Background color represent the inheritance pattern associated

Conclusion

By performing a species-wide screen of genetic complexity of traits in *S. cerevisiae* with 190 crosses coming from 20 natural isolates, we were able to assess the complexity level of 7,600 cross/trait combinations. Moreover, we highlighted the prevalence of expressivity with most of the followed variants displaying departure from monogenic inheritance patterns.

The study of the phenotypic distribution allows to reflect the underlying genetic complexity of a trait up to a certain degree, proving to be a powerful tool to detect strong allelic effects for monogenic and low complexity genetic interactions. Nevertheless, it remains very limited when it comes to figuring out the number and effect size of genetic variants involved in complex traits. Indeed, simulations of traits with only two loci acting additively already resulted in a normal distribution of the phenotype. Another important limitation of this method is the lack of power to resolve small effects. Indeed, we can only assume the genetic complexity of traits that show really contrasting phenotypes between the different alleles. If two alleles show a relatively small phenotypic difference between them, no differentiation will be possible due to experimental and biological noise masking the true allelic effect. For this purpose, we measured experimental noise to obtain confidence in the results.

Genetic complexity is variable between traits

It clearly appears that some traits have a more simple genetic basis than others. When for some trait, almost half of the crosses show monogenic inheritance, others only display complex inheritance patterns.

Because of the way we detect them, suppressors are modifier genes that mask or counter the effect of an otherwise monogenic trait, it thus makes sense to find them only in crosses with one strain carrying a monogenic variant. However, strong

digenic recessive epistatic interactions, although being found in almost all conditions, are quite rare in terms of proportion compared to other type of traits. This underlies that these interactions are strongly related to specific and rare allelic combinations between precise genetic backgrounds suggesting little selection acting. There again, our experimental design only allows us to focus on the interactions having an important phenotypic impact, which represent only a fraction of all digenic interaction that can potentially exist (Costanzo et al., 2016).

The dynamic nature of trait complexity

We highlighted the dynamic nature of trait complexity with the effect of monogenic variants transitioning from monogenic inheritance pattern up to a complex one depending on the genetic background it lies in. This suggests that the number of gene controlling a phenotype is highly dependent on specific and sometimes rare allelic combinations modifying the "expected" phenotype. Although such a continuum of genetic complexity has already been observed with a different variant (Hou et al., 2016a), almost all the detected high effect variants in this study with only 20 different natural isolates showed variation in their trait complexity level across several genetic backgrounds. We can easily predict that with more genetic backgrounds, other modification of the genetic complexity and thus of monogenic distribution will arise. There are two possible ways to test this. First, following the same design as the one used here, one can cross the strain carrying a particular variant with other strains and look for modifications in the inheritance pattern in their progeny. Another solution which could be applied at a much higher scale would be to first map the variant via bulk segregant analysis. Then, we could introduce it in an important number of strains via CRISPR/Cas9 directed mutagenesis and assess the induced phenotypic change.

Based on our results, we can assert that expressivity is pervasive, as seen for variants with a strong phenotypic effect. This observation questions the existence of monogenic traits at the population level. Indeed, Mendelian inheritance seems to be mostly cross/trait specific rather than conforming to a simple trait related pattern for every individual. This might be because of the intricacies of genetic interactions and metabolic pathways combined with the extensive genetic variation which yields important number of allelic combinations with potential epistatic effect. This would allow in some cases to widen the phenotypic and complexity landscape of a trait. The dynamic nature of trait complexity also raises the fact that gaining strong phenotype prediction power solely based on genotype is very unlikely even for traits thought to be monogenic.

Altogether, this works lays the ground for a more complete and in detail exploration of variants displaying different levels of expressivity by testing their effects in a wider number of genetic backgrounds. This would allow to obtain a picture of the diversity of the modifier landscape.

Supplementary material

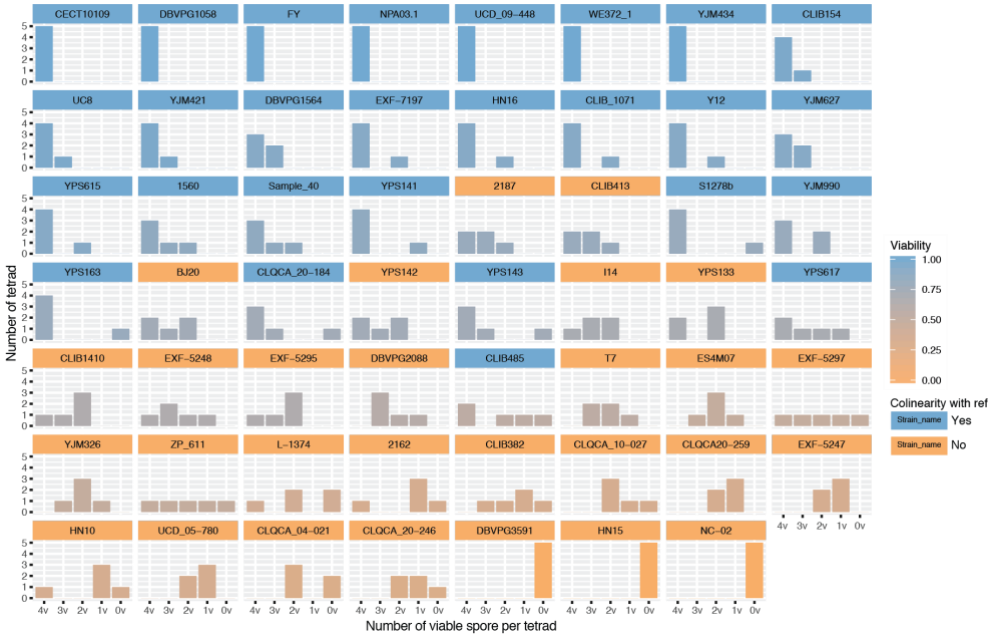


Figure S1. Assessing the collinearity of the genomes
Each of the 55 strains were crossed with the reference strain to assess the collinearity. Here is represented the number of viable spores per tetrad for each cross. Color of the bars represent the viability of the offspring and label color represent the genome collinearity or not between the tested isolates and the reference strain.

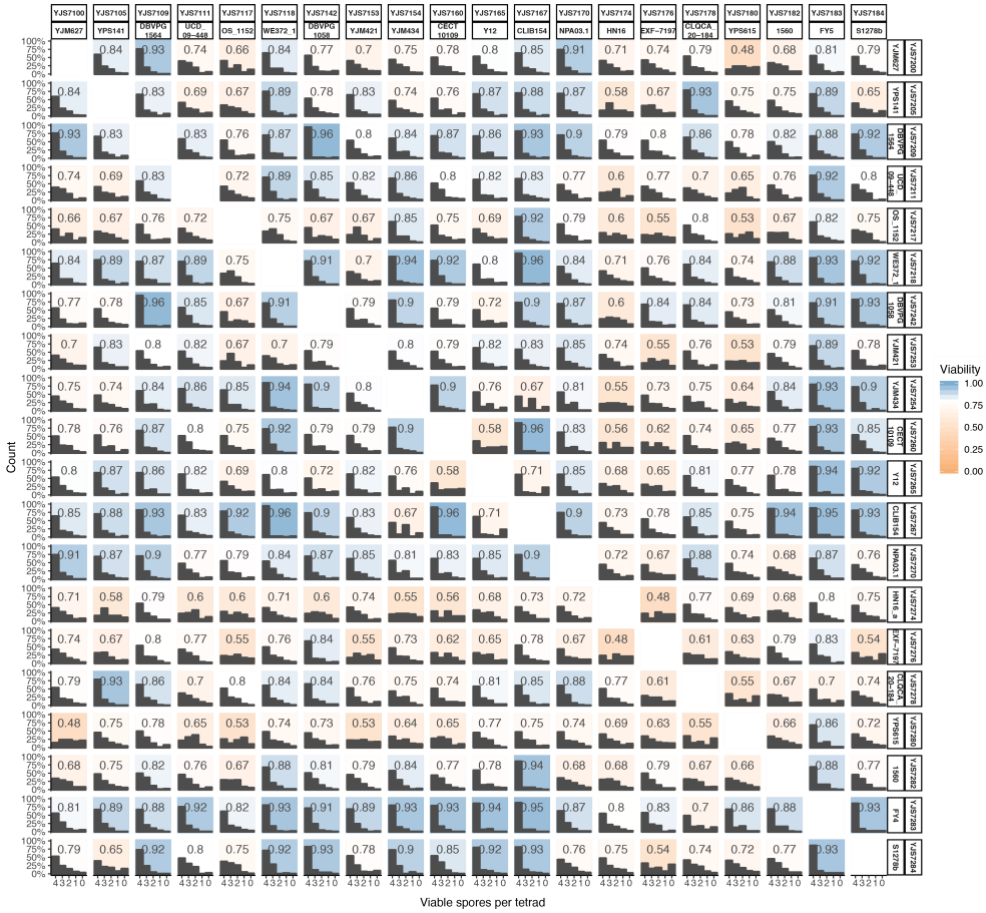


Figure S2. Viability per tetrad in the diallel offspring panel

Each panel represent the number of tetrads having 4,3,2,1 or 0 viable spores for each cross. Background color and number represent the viability of the offspring. Only the viability of the offspring coming from 190 crosses in the upper triangle has been assessed, lower triangle of this matrix has just been copy pasted to facilitate the reading when following one particular isolates. The matrix of panel is facetted with *MAT*α isolates on the x axis and *MAT*a isolates on the y axis.

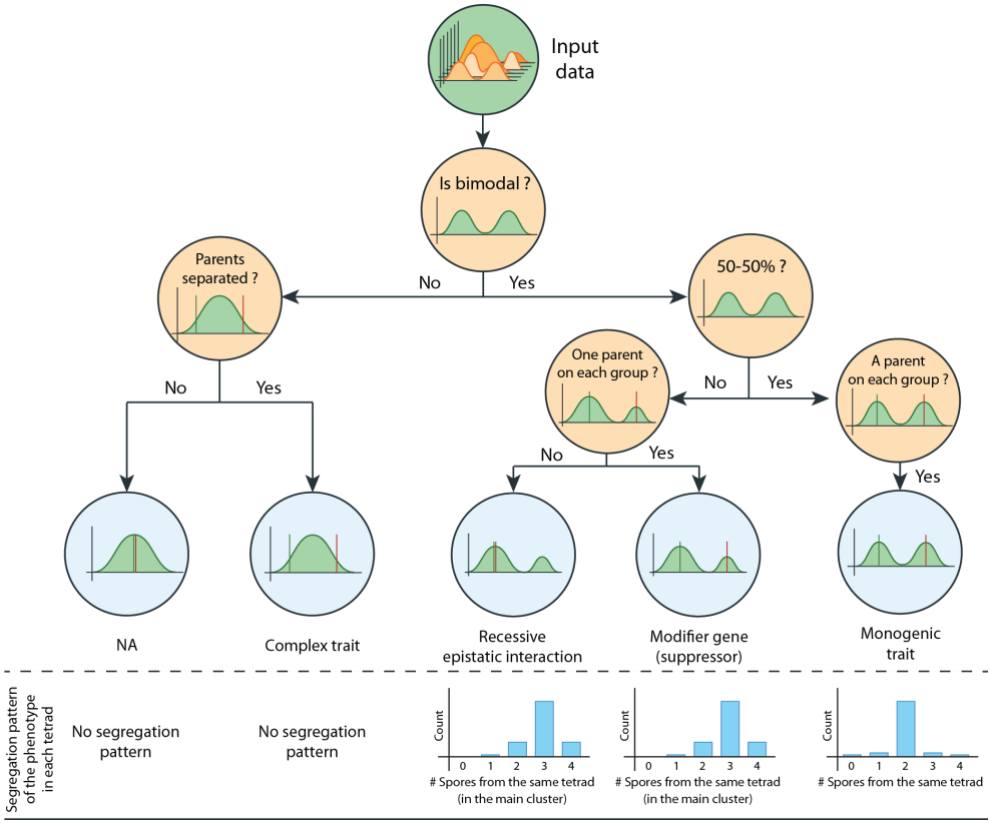


Figure S3. Decision tree for the classification of inheritance modes
The phenotypic distribution of the progeny of each cross on each condition is extracted. The first step is to determine the unimodality or bimodality of the distribution. If the distribution is bimodal, the parental phenotypic values and the segregation of the phenotype in the tetrads is assessed to further classify the distribution between oligogenic and monogenic inheritance. If a normal distribution is observed, the good separation of the parental phenotype is assessed to ensure the presence of a complex trait.
Courtesy of Andreas Tsouris.

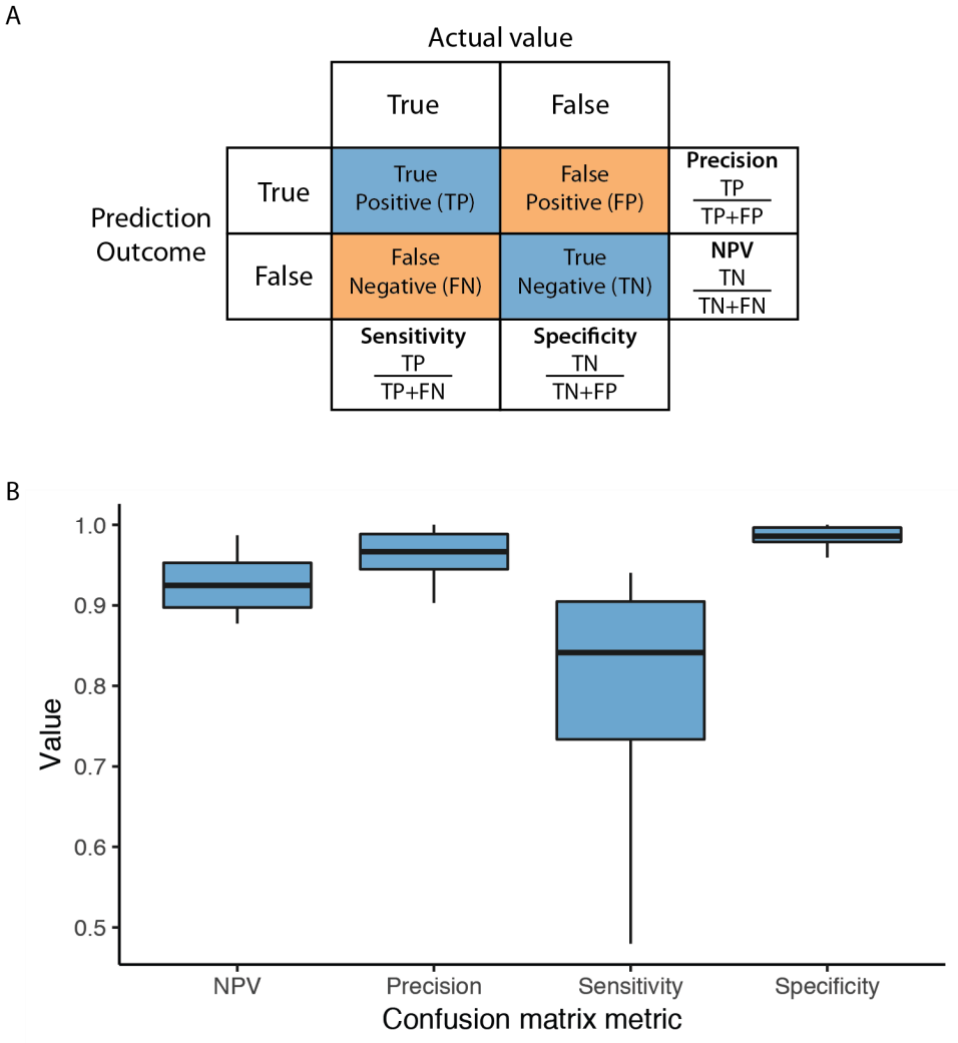


Figure S4. Confusion matrix to assess the power of the random forest
A. Example of a confusion matrix with the metrics associated to it (Precision, Negative Predictive Value (NPV), Sensitivity and specificity) and how to compute them. **B.** Metrics of the seven confusion matrices to assess the reliability of the constructed random forest.

References

Altenburg, E., and Muller, H.J. (1920). The genetic basis of truncate wing - An inconstant and modifiable character in Drosophila. Genetics 5, 1–59.

Antonarakis, S.E., Chakravarti, A., Cohen, J.C., and Hardy, J. (2010). Mendelian disorders and multifactorial traits: the big divide or one for all? Nat. Rev. Genet. 11, 380–384.

Bikard, D., Patel, D., Le Metté, C., Giorgi, V., Camilleri, C., Bennett, M.J., and Loudet, O. (2009). Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. Science (80-.). 323, 623–626.

Charron, G., Leducq, J.B., and Landry, C.R. (2014). Chromosomal variation segregates within incipient species and correlates with reproductive isolation. Mol. Ecol. 23, 4362–4372.

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.

Chu, D.B., and Burgess, S.M. (2016). A Computational Approach to Estimating Nondisjunction Frequency in Saccharomyces cerevisiae. G3 (Bethesda). 6, 669–682.

Correns, C. (1900). Gregor Mendels Regel über das Verhalten der Nachkommenschaft der Bastarde. Berichte Des Dtsch. Bot. Gesellschaft 158–168.

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 (80-.). 353, aaf1420.

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. BioRxiv 609917.

Hartigan, J.A., and Hartigan, P.M. (1985). The Dip Test of Unimodality. Ann. Stat. 13, 70–84.

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., Friedrich, A., Gounot, J.S., and Schacherer, J. (2015). Comprehensive survey of condition-specific reproductive isolation reveals genetic incompatibility in yeast. Nat. Commun. 6, 7214.

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

Hou, J., Fournier, T.T., and Schacherer, J. (2016b). Species-wide survey reveals the various flavors of intraspecific reproductive isolation in yeast. FEMS Yeast Res. 16.

Iyer, R.R., Pluciennik, A., Burdett, V., and Modrich, P.L. (2006). DNA mismatch repair: Functions and mechanisms. Chem. Rev. 106, 302–323.

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.

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.

Seidel, H.S., Rockman, M. V, and Kruglyak, L. (2008). Widespread genetic incompatibility in C. elegans maintained by balancing selection. Science (80-.). 319, 589–594.

Silverman, B.W. (1981). Using Kernel Density Estimates to Investigate Multimodality. J. R. Stat. Soc. Ser. B 43, 97–99.

Tschermak-Seysenegg, A. Von (1900). Ueber künstliche Kreuzung bei Pisum sativum. Berichte Des Dtsch. Bot. Gesellschaft 18, 232–239.

De Vries, H. (1900). Sur la fécondation hybride de l'albumen. Biol. Zent. Bl. 20, 129–130.

CHAPTER 3

Exploring the structural variation landscape using long read sequencing

Summary

The extensive genetic variation between individuals of a population is not restricted to Single Nucleotide Polymorphisms (SNPs). Structural variation is a key player of this variation and is likewise causative of various phenotypes. However, the nature and number of these events at a population-scale is complex. Small reads sequencing technologies fail to accurately detect Structural Variants (SVs) such as translocations and inversions and yield a high number of false positive. Recent years have seen the advent of long read sequencing technologies which pledge to allow for precise SV detection. This chapter summarizes various projects and collaborations I was involved in, all aiming at building a stronger knowledge of structural variation with particular emphasis on translocations. In the first part, we set out the interest of using a non-conventional yeast species, namely *Brettanomyces bruxellensis* which harbors extensive genomic rearrangements. In order to carry out systematic detection of SVs in this species, we first had to generate a high-quality reference sequence that could also be used for any other population genomics studies. This sequence holds its promises by being of high contiguity and completeness and will set the ground for mapping structural variation when comparing different individuals of this species.
The second part focuses on building an exhaustive species-wide catalog of structural variation in *S. cerevisiae* by the sequencing and *de novo* assembly of roughly 100 natural isolates using long read sequencing (out of which I sequenced 30). Initial results mapped 42 translocations and 148 inversions but finer analysis are required.
Finally, the third part of this chapter is dedicated to understand the phenotypic outcome of translocations in a fixed genetic background. Through the creation of an efficient and precise genome editing tool, dozens of independents translocated strains were generated in a single genetic background. Those strains only varied by their chromosomal organization, my role has been to phenotype them. It revealed that the sole effect of tridimensional reorganization due to translocations was sufficient to generate an important phenotypic diversity.

Introduction

Current studies and knowledge about genetic variation are strongly biased towards single nucleotide polymorphisms. Nonetheless, SNPs are far from being the only source of genetic variation. Genomes can also differ by the presence of structural variation that can be both balanced if the number of copies is not changed or unbalanced if the variant result in a change of the copy number of a portion of the genome. We already know from population scale studies that Copy Number Variants (CNVs) play a significant role in the phenotypic diversity of a population as GWAS encompassing CNVs in 1,011 yeast natural isolates revealed that they outnumbered SNPs both by number but also by their effect size (Peter et al., 2018). However, CNVs are far from being the only type of SVs in a population. Moreover, unbalanced SVs have been more extensively studied than balanced SVs for the simple reason that they are easier to detect, especially at a population scale. Indeed, short-read sequencing approaches do not allow for precise detection of SVs because they tend to yield a high number of false positives as well as false negatives. With the recent advances in sequencing technologies and the rise of long read sequencing, detection of such variants at a large scale is now much more attainable.

Long reads facilitate SV detection for two main reasons: on the one hand, as the reads are longer, there are more chances to detect a read that either spans the entire length of the variant plus its flanking regions or contains the breakpoint of a translocation or an inversion. On the other hand, longer reads mean easier assembly. Assembly can be assimilated to a puzzle with pieces that fit together. Assembling a puzzle with 10,000 pieces is more difficult and error-prone than assembling a 10-pieces puzzle. If we can cover the entire genome with less reads, it will be easier to piece them together. Nowadays with long read strategies, we can obtain read length with a mean of 20 kb or higher which means that theoretically for a typical yeast genome of 12 Mb, only 600 reads are needed to cover its entire length. Once

assembled, genomes can be compared to each other to reveal structural variations. This strategy of mapping structural variation through *de novo* assembly has recently been applied to 22 strains of *S. cerevisiae* using Oxford Nanopore long read sequencing (Istace et al., 2017). This study unveiled part of the structural variation landscape in this species by detecting a total of 29 translocations and 4 inversions (Istace et al., 2017). Similarly, using PacBio sequencing in 7 *S. cerevisiae* natural isolates and 5 isolates of its sister species *Saccharomyces paradoxus*, 28 inversions and 6 reciprocal translocations were mapped with most of them being in *S. paradoxus* (Yue et al., 2017).

In this chapter, we focused on laying the basis for a systematic exploration of SVs at a species-wide level both in *S. cerevisiae* and in the non-model yeast *Brettanomyces bruxellensis* as well as gaining knowledge on the phenotypic impact of such variants. To do so, we sequenced isolates with Oxford Nanopore Technologies long read solution. Unlike other sequencing methods based on the synthesis of DNA molecules such as Illumina or PacBio, the Oxford Nanopore approach consists in an array of proteins (pores) that detects consecutive 6-mer of a native DNA molecule sensing a change in electrical signal as DNA is fed through the pore.

Part 1 : High-quality *de novo* genome assembly of the *Brettanomyces bruxellensis* yeast using nanopore MinION sequencing

Our first aim has been to focus on structural variation in a non-model yeast species. We chose to work on the yeast species *Brettanomyces bruxellensis*. This species is isolated from different fermented beverages. It is of high industrial interest because of its association with wine spoilage where it produces volatile phenolic compounds that are very odorant with smells described as barnyard or horse sweat (Chatonnet et al., 1992). However, *B. bruxellensis* is also responsible for specific organoleptic properties of spontaneously fermented Belgian beers (hence its name) such as lambic or gueuze (Spitaels et al., 2014). What makes this species of high interest for structural variant exploration is its genomic plasticity. Indeed, natural isolates show different ploidy levels (Avramova et al., 2018; Borneman et al., 2014; Curtin and Pretorius, 2014) and extensive chromosomal rearrangements, which were observed through electrophoretic karyotypes (Hellborg and Piskur, 2009). The exploration of structural variants such as large indels, inversions and translocations at the species level would help to provide insights into the forces that shape genomic architecture and evolution. However, to conduct a population genomic survey, the availability of a high-quality reference sequence with a completeness level allowing to cover most of the genomic variation and a contiguity level to efficiently detect structural variants, is a prerequisite. The lack of such reference genome pushed us to generate a *de novo* and high-quality genome assembly of the UMY321 isolate with the combination of long reads coming from Oxford Nanopore and short reads from Illumina sequencing.

Three *B. bruxellensis* isolates (UMY321, UMY315, and 133) were sequenced in this study (Table 1). These strains were determined to be diploid based on flow cytometry analysis and were all isolated from wine or grape must in Italy or South Africa. The genome of the UMY321 isolate was sequenced using a combination of Oxford Nanopore long-read and Illumina short-read sequencing data to obtain a high-quality assembly. By contrast, the UM315 and 133 isolates were only sequenced using a short-read strategy. In addition, these genomes were compared to previously genome sequences of six other *B. bruxellensis* isolates (Table 1) (Borneman et al., 2014; Crauwels et al., 2014; Curtin et al., 2012; Olsen et al., 2015; Piškur et al., 2012).

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

Strain	Ploidy	Ecological Origin	Geographical Origin	Reference
AWRI1499	3n	Wine	Australia	Curtin *et al.*(2012)
AWRI1608	3n	Wine	Australia	Borneman *et al.*(2014)
AWRI1613	2n	Wine	Australia	Borneman *et al.*(2014)
CBS11270	2n	Industrial ethanol	Sweden	Olsen *et al.*(2015)
CBS2499	2n	Wine	France	Piškur *et al.*(2012)
ST05_12_22	2n	Lambic beer	Belgium	Crauwels *et al.*(2014)
UMY315	2n	Must	Italy	This study
UMY321	2n	Red wine	Italy	This study
133	2n	Merlot wine	South Africa	This study

De novo genome assembly construction and comparison

Sequencing of the UMY321 isolate required three MinION runs using the R7.3 chemistry and 2D libraries with a DNA fragmented to 8 kb. Briefly, 2D sequencing works by linking a hairpin to one end of a double stranded DNA fragment which allows to sequence both strands of the DNA thus yielding better sequencing accuracy than 1D alone (where only one of the two strands is sequenced). A total of 1.15 Gb

was generated with 41,686 2D reads having an average quality greater than nine (phread score). For our assembly, we focused on these reads which represented 376.8 Mb with a mean read length of 9,033 bp and a median of 8,676 bp. Four subsets representing different coverage (10x, 15x, 20x and the total dataset representing roughly 25x) were submitted to four different assemblers: ABruijn (Lin et al., 2016) Canu (Berlin et al., 2015), miniasm (Li, 2016), and SMARTdenovo (https://github.com/ruanjue/smartdenovo). One known flaw of Oxford Nanopore sequencing is the high error rate associated with it (around 10% for 2D reads with R7.3 chemistry) (Jain et al., 2016). In order to counter this, the assemblies were subsequently polished with Illumina reads (around 100x of paired-end reads) using Pilon (Walker et al., 2014). This hybrid strategy allows to take advantage of both sequencing technologies, combining the ease of assembly given by Oxford Nanopore long reads with the precision of the Illumina paired-end reads.

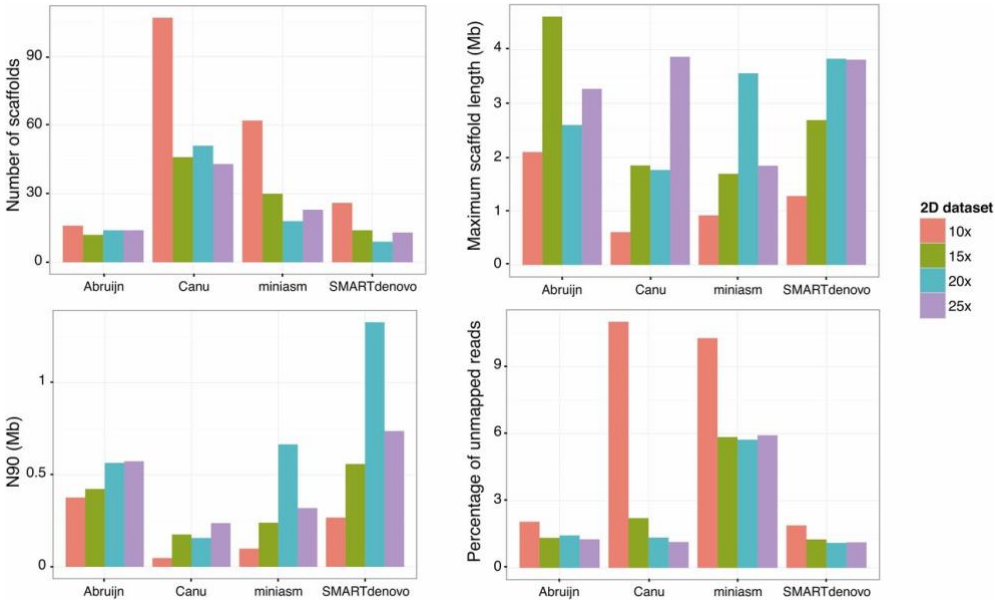


Figure 1. Metrics related to the constructed assemblies
Metrics are displayed per assembler and per dataset used.

Comparing the results from the four assemblers with the four datasets, we decided to make our final assembly based on the results of SMARTdenovo with a 20x coverage because it resulted in the best contiguity metrics (Figure 1): the final assembly contained eight scaffolds for a size of 12,965,163 bp, revealing near chromosome scale resolution. This level of contiguity is essential for the detection of structural variants. The completeness of our assembly has also been assessed by running CEGMA (Parra et al., 2007). It revealed that out of the 248 most conserved genes in all eukaryotic genomes, 242 displayed complete alignment and only 3 were not detected in our assembly. This result confirmed the high level of completeness of this assembly.

Comparison with available assemblies of *B. bruxellensis*

To date, several assemblies of the *B. bruxellensis* species have already been released (Borneman et al., 2014; Crauwels et al., 2014; Curtin et al., 2012; Olsen et al., 2015; Piškur et al., 2012). These assemblies are related to isolates from different ecological and geographical origins (Table 1). They were mostly constructed by combining several sequencing methods, such as 454, PacBio, and Illumina, as well as optical mapping in the most recently published assembly (Olsen et al., 2015). The assemblies have very variable metrics associated with each of them (Table 2). In terms of contiguity, our assembly and the assembly generated for the CBS11270 isolate are close, and reach a chromosome-scale resolution. However, the CBS11270 assembly is much larger than the others (17.3 Mb *vs*. 12.7–13.4 Mb), although it does also contain ~2.5 Mb of undetermined (N) residues.

Table 2. Metrics associated to *B. bruxellensis* publicly available assemblies

Strain	# Scaffolds	Assembly Size (Mb)	Maximum Scaffold Size	N50	N90	# N
AWRI1499 (Curtin *et al.* 2012)	324	12.7	170,307	65,420	22,583	57
CBS11270 (Olsen *et al.* 2015)	15	17.3	4,993,495	3,706,654	944,992	2,497,785
CBS2499 (Piškur *et al.* 2012)	84	13.4	2,877,306	1,792,735	190,560	586,105
ST05_12_22 (Crauwels *et al.*2014)	85	13.1	1,439,423	732,210	177,142	218,317
UMY321 (this study)	8	13	3,829,289	1,917,156	1,329,398	2708

By comparing the assembly metrics, we determined that our assembly is closer to that for CBS11270, which was generated by combining PacBio and Illumina sequencing methods as well as optical mapping, and much better than the other three available for comparison, which were much more fragmented and comprised at least 84 scaffolds.

A MUMmer comparison of our UMY321 assembly to that of CBS11270 indicates that 91 and 99.6% of the assemblies aligned, respectively, with one another and revealed that the scaffolds are mostly collinear (Figure 2). However, some large repetitive regions can be observed in the CBS11270 assembly, *e.g.* on chromosome 1, between chromosomes 1 and 6, and between chromosomes 4 and 5 (Figure 2) that are absent in our assembly, and could explain the size differences between the assemblies (17.3 Mb vs. 12.97 Mb). Moreover, some synteny breaks can be observed, at the level of scaffolds, specifically between three and four. All the inconsistencies between the assemblies could be related either to structural rearrangements between the isolates or to assembly errors and would require further investigations to reach a conclusion as to their most likely source.

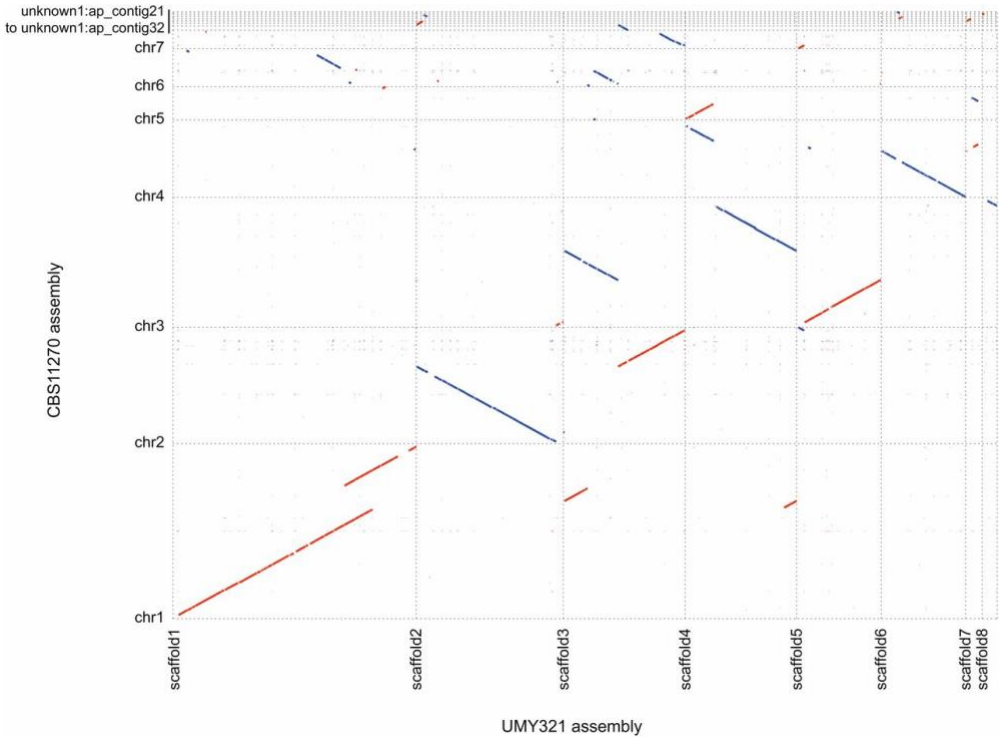


Figure 2. Comparison of the CBS11720 and UMY321 assemblies
The alignments and the plot were generated with the MUMmer software suite. Red lines: sequences aligning in the same direction. Blue lines: sequences aligning in the opposite.

Suitability of our assembly for population genomics studies

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

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

A total of 83,006 SNPs was detected with GATK (McKenna et al., 2010), among which 374 were homozygous and 82,632 were heterozygous. The 374 homozygous SNPs could be considered as false positives. Although not completely negligible, this number is very low and could be related to the high error rate of the MinION technology, which is not completely compensated by using Illumina short reads (Istace et al., 2017).

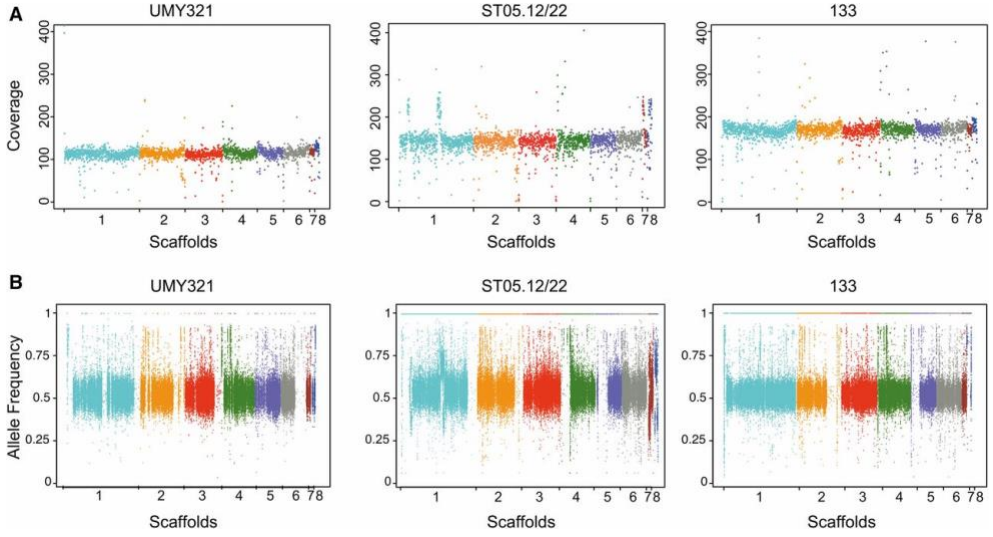


Figure 3. Mapping of the Illumina reads vs. the UMY321 reference assembly. **A.** Illumina reads coverage along the reference genome. **B.** Frequency of the reference allele at heterozygous sites along the genome. (Each color corresponds to a scaffold).

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

B. bruxellensis is a yeast species of great importance in fermented beverage industries, largely thought of as a contaminant organism (Masneuf-Pomarede et al., 2016; Schifferdecker et al., 2014). This species is also an interesting model to study genome evolution and dynamics as it is characterized by a large genomic plasticity. For these reasons, we sought to generate a high-quality genome assembly and ultimately obtain a suitable reference genome for population genomics. Our analyses show that the *B. bruxellensis* assembly that we generated with a combination of moderate coverage (20x) MinION long-reads in addition to a higher coverage (100x) of Illumina reads utilized for sequence polishing purposes, is highly valuable for population genomic studies and outperforms previously available sequences. To obtain a species-wide view of the genetic variability of *B. bruxellensis*, many more isolates should be surveyed using both short-read as well as long-read sequencing techniques, which will to qualify and quantify the extensive structural variation happening in this species. In the laboratory, dozens of natural isolates are currently being sequenced using the Oxford Nanopore MinION.

This part is a modified version of the publication:

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

*: Equal contribution

Part 2: Generation of a population-wide catalog of structural variation in 95 natural *S. cerevisiae* isolates

Despite the good knowledge of the genetic diversity based on SNPs and its implication on the phenotypic landscape that we acquired recently in *S. cerevisiae* (Peter et al., 2018), we still lack a complete and unbiased view of structural variation in this species. Nevertheless, a previous study sequenced 22 strains of *S. cerevisiae* with long read technologies (Istace et al., 2017). Although this study allowed to understand the important prevalence of structural variation in natural populations, the number of considered strains were not enough to really grasp the full diversity of SVs and understand the global repartition of these types of variants throughout the whole species. Questions still remain such as the presence of conserved SVs across the population that potentially have a phenotypic advantage to some isolates. Therefore, we wanted to obtain a much broader view which could lead to the generation of an exhaustive catalog of the structural variation in *S. cerevisiae* by sequencing 95 natural isolates, representative of the species diversity, both in term of ecological origins but also in terms of genomic features. Indeed, we selected 15 haploid isolates, 35 heterozygous diploids and 2 homozygous diploids. To these, 43 monosporic isolates coming from heterozygous diploids were also added (Table3).
All the strains have been sequenced using Oxford Nanopore sequencing technology to provide long reads thus facilitating *de novo* genome assembly and subsequent detection of SVs based on these assemblies. A first glance at the results already detected that about a third of the natural isolates are carrier of a translocation. Moreover, 174 insertions of 5 kb or more have been detected with one particular insertion of 35 kb in chromosome 14 appearing in 75% of the sequenced isolates proving that some SVs are under strong selection allowing them to raise to an important frequency in the population. More in-depth analysis are needed to obtain a complete picture of SVs in *S. cerevisiae*.

Table 3. Strains sequenced and *de novo* assembled

Standardized name	Strain	Ecological origin	Geographical origin	Genome
AAB	CBS422a	Beer	Odessa, Ukraine	Haploid
AAC	CBS2165a	Beer	England	Hetero/aneu/poly
AAR	CLIB382_1b	Beer	Ireland	Haploid
ABA	YJM326_b	Human, clinical	California, USA	Haploid
ACH	CLIB483_1b	Cider	Brittany, France	Haploid
ADA	Y55	Lab	NA	Homozygous
ADE	PW5_b	Palm wine	Aba, Abia State, Nigeria	Hetero/aneu/poly
ADI	YJM981_b	Human, clinical	Italy	Haploid
AEH	CBS7964	Industrial	Brazil	Haploid
AEL	CBS1394	Distillery	NA	Hetero/aneu/poly
AFH	CBS1509	Distillery	NA	Haploid
AFI	CBS2183	Wine	Chateau Chalon France	Haploid
AGA	CBS3012	Wine	Cadiz, Spain	Hetero/aneu/poly
AGK	CBS2361	Nature	UK	Haploid
AHG	CBS1586	Fruit	NA	Haploid
AHL	CBS3081	Industrial	Spain	Haploid
AIC	CBS2807	Wine	Slovakia	Haploid
AIE	CBS2910	Human	Portugal	Hetero/aneu/poly
AIF	CBS457	Wine	Italy	Hetero/aneu/poly
AIG	CBS1463	Beer	NA	Haploid
AIS	MC9	wine	AP, Italy	Hetero/aneu/poly
AKH	NPA02-1	Palm wine	Nigeria	Monosporic
ALH	CLQCA_19-011	Nature	Napo, Ecuador	Monosporic
ALI	CLQCA_20-060	Water	Ecuador	Hetero/aneu/poly
ALS	21-4-0116	Tree	Male levare, Slovakia	Monosporic
AMH	EN14S01	Soil	Sinyi, Nantou, Taiwan	Monosporic
AMM	SJ5L12	Tree	Beinan, Taitung, Taiwan	Monosporic
AMP	SJ5L14	Fruit	Taian, Miaoli, Taiwan	Monosporic
ANL	A-6	Beer	Ghana	Hetero/aneu/poly
ANM	A-18	Beer	Ghana	Hetero/aneu/poly
APG	VF8_(6)	Bioethanol	Araras, S‹o Paulo, Brazil	Haploid
AQG	CBS7539	Beer	Plovdiv, Bulgaria	Monosporic
ARN	CBS2246	Human, clinical	Netherlands	Hetero/aneu/poly
ASB	CBS4255	Human, clinical	NA	Hetero/aneu/poly
ASG	CBS1489	Human, clinical	Italy	Homozygous
ATM	UC10	Wine	California, USA	Monosporic
ATV	CECT1462	Beer	UK	Hetero/aneu/poly
AVI	YPS163	Soil	Pennsylvanian	Monosporic
AVN	CH02	Beer	Ivory Coast	Hetero/aneu/poly
BAD	DJ71	Palm wine	Yoboki, Djibouti	Hetero/aneu/poly
BAF	DJ74	Palm wine	Yoboki, Djibouti	Hetero/aneu/poly
BAG	SX1	Tree	Shaanxi province, China	Monosporic
BAI	BJ6	Fruit	Changping, Beijing, China	Monosporic
BAK	BJ20	Tree	Beijing, China	Monosporic
BAL	HN6	Nature	Hainan province, China	Monosporic
BAP	HN16	Soil	Hainan province, China	Homozygous
BAQ	HN19	Tree	Hainan province, China	Monosporic
BBF	CCY_21-4-102	Water	Slovakia	Hetero/aneu/poly
BBM	908	Distillery	Jalisco, Mexico	Monosporic
BBT	2281	Wine	Spain	Hetero/aneu/poly
BCE	HE006	Human	French Guiana	Monosporic
BDC	#36	Nature	Israel	Hetero/aneu/poly
BDF	#57	Nature	Israel	Hetero/aneu/poly
BDH	#59	Nature	Israel	Monosporic

BDM	MAJ_A	Nature	Majunga, Madagascar	Monosporic
BDN	MAJ_G	Nature	Majunga, Madagascar	Monosporic
BEM	CLIB653	Beer	Chad	Hetero/aneu/poly
BFH	EXF-5871	Dairy	Slovenia	Hetero/aneu/poly
BFP	EXF-7197	Tree	Montenegro	Monosporic
BGN	CLIB561	Dairy	Normandy, France	Monosporic
BGP	CLIB562	Dairy	Normandy, France	Monosporic
BLD	DBVPG1608	Wine	La Mancha, Spain	Monosporic
BMC	UWOPS03-459.1	Tree	Malaysia	Monosporic
BPG	DBVPG1841	NA	Ethiopia	Hetero/aneu/poly
BPK	DBVPG1861	Water	Rajamaki River, Finland	Hetero/aneu/poly
BTE	YS8(E)	Bakery	NA	Hetero/aneu/poly
CAS	B-17	Wine	Georgia	Monosporic
CBK	1	Insect	Schleswig-Holstein	Haploid
CCC	CLQCA_20-156	Flower	Yasuni, Orellana	Monosporic
CCQ	Ksc2-2B	Tree	Japan	Monosporic
CCT	S11F3-6B	Tree	Sri Lanka	Monosporic
CDA	S8BM-32-4D(a)	Tree	Sri Lanka	Haploid
CDG	N3.00-7A	Wine	Blagoveshchensk, Russia	Monosporic
CDN	UCD_61-190-6A	Insect	California, USA	Hetero/aneu/poly
CEI	GE14S01-7B	Soil	Taiwan	Monosporic
CEL	JCM_3529-7B	Fermentation	Tailand	Monosporic
CEQ	MUCL_30909-2C	Fermentation	Burundi	Monosporic
CFC	4.5_WLP530	Beer	Westmalle, Belgium	Hetero/aneu/poly
CFS	UCD_40-255	Wine	Walnut Creek, California	Hetero/aneu/poly
CGH	VNL3	NA	Vietnam	Hetero/aneu/poly
CHS	SC 32 F. Dromer IP	Human, clinical	France	Monosporic
CIH	PB12	Human, clinical	Netherlands	Hetero/aneu/poly
CKB	malade 98 1655/125391	Human, clinical	Paris, France, H3	Hetero/aneu/poly
CLL	K10	Sake	Japan	Hetero/aneu/poly
CLN	K14	Sake	Japan	Homozygous
CMF	RIB6001	Sake	Japan	Hetero/aneu/poly
CNB	SM.8.2.C13	Bioethanol	Brazil	Monosporic
CPA	906	Nature	Mexico	Monosporic
CPG	1560	Nature	Aceituna, Spain	Monosporic
CPI	LCBG-3D6	Distillery	Tamaulipas, Mexico	Monosporic
CPS	FTPW4	Palm wine	Burkina Faso	Hetero/aneu/poly
CQI	MTF2552	Fermentation	West Africa	Monosporic
CQS	CEY647	Nature	French Guiana	Monosporic
CRB	SC2-37	Wine	Italy	Monosporic
CRE	CLQCA_17-111	Insect	Ecuador	Hetero/aneu/poly

As for *B. bruxellensis*, the first goal has been to sequence the genomic DNA of all these isolates using Oxford Nanopore. To do so, a total of 30 MinION and one PromethION flowcells have been used. My implication in this project has been to extract the DNA, prepare the librairies and run the samples on MinION flowcells for 30 of the isolates. After basecalling with Guppy, the 30 MinION flowcells yielded from 330 Mb up to 21.7 Gb (mean of 6 Gb) and 87 Gb for the PromethION flowcell. The sequencing yield was quite heterogeneous across the runs mostly because several generations of library preparation chemistry were used as this technology is moving forward and evolving quite fast. Once demultiplexed, a total of 196 Gb was available. We measured the mean and median read length as well as the N50 (shortest read to get half of the total bases sequenced) for each strain. Mean read length for each strain went from 3.2 kb to 25 kb (mean 10.5 kb), median read length from 1.7 kb to 21.8 kb (mean 6.8 kb) and N50 from 4.6 kb up to 38 kb (mean 18.5 kb) (Figure 4A). Overall these results were satisfying and as the amount of data per strain was sufficient (median coverage of 77x) (Figure 4B) we wanted to find the best possible set of reads for assembly. To do so, we used Filtlong (https://github.com/rrwick/Filtlong), to select the equivalent of 40x coverage of the best reads available for each strain (See methods) which allowed to improve the metrics of the dataset used (Figure 4A).

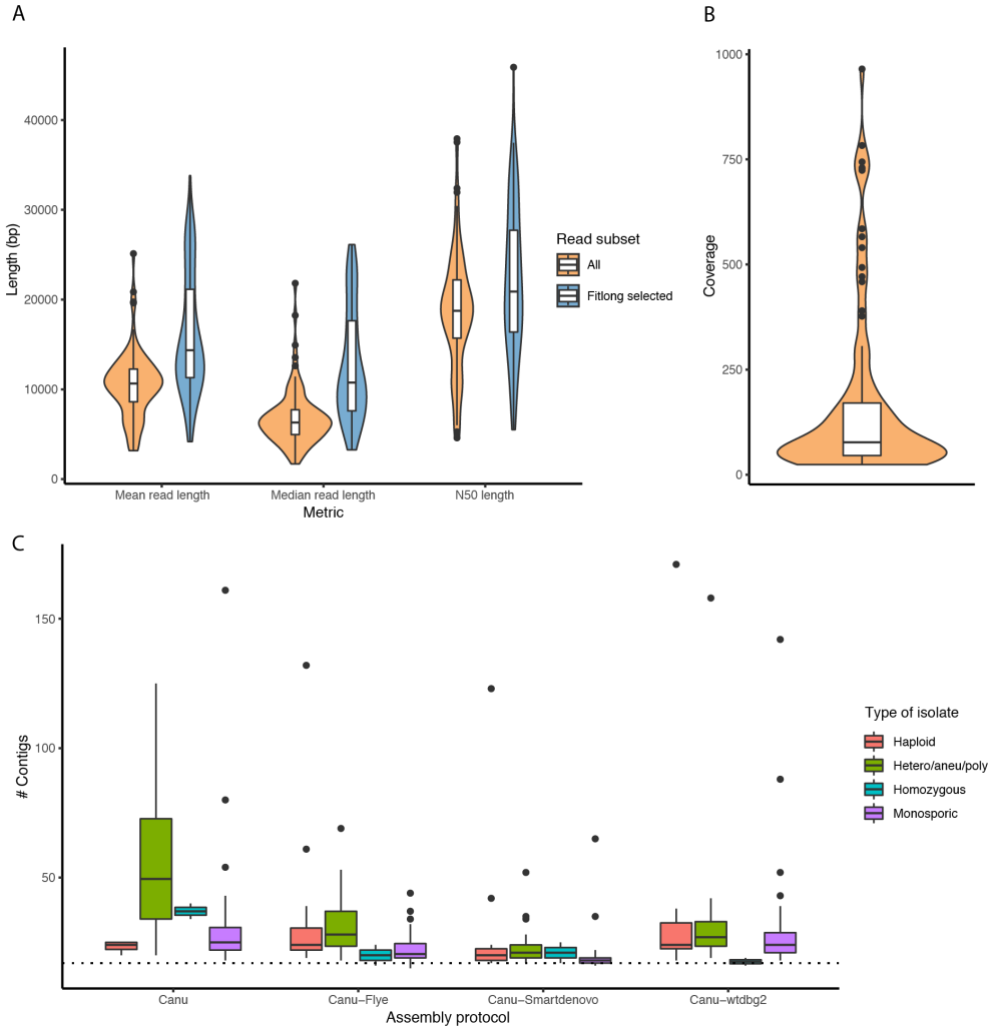


Figure 4. Metrics of the sequencing and assembly of the 95 strains.
A. Mean read length, median read length and N50. All reads without filtering are in orange, 40x coverage selection of best set of reads using Fitlong are represented in blue. **B.** Coverage distribution for the 95 strains. **C.** Number of contigs after assembly with different assembly pipelines. Colors represent the different type of isolates used in this study: Haploid, Heterozygous or aneuploid or polyploid isolates, Homozygous diploids and finally Monosporic diploid isolates. Dotted line represents a number of 17 contigs for the 16 genomic chromosomes and the mitochondrial contig.

Assembly of the genomes has been performed using several assembly pipelines with different assemblers *e.g.* Canu (Koren et al., 2017) as a standalone, or coupled with SMARTdenovo (https://github.com/ruanjue/smartdenovo), flye (Kolmogorov et al., 2019) and wtdbg2 (Ruan and Li, 2019), in order to find the best combination. Overall, SMARTdenovo offers better assembly quality with fewer contigs, especially for isolates that are heterozygous, aneuploid or polyploid (Figure 4C). Although the results are still fresh and would require further validations, with the available *de novo* assemblies coming from SMARTdenovo, a first catalog of structural variants can be put together. While an automated script is currently being developed to detect structural variants from *de novo* assemblies, a first rough estimation can be made by looking at the dotplot output given by MUMmer (Kurtz et al., 2004). This allows to visually see translocation events as well as inversions by comparing the contigs of the *de novo* assembly with the reference assembly.

This first overview of the assemblies allowed to detect 42 translocations with 18 being reciprocal and 24 being non-reciprocal in 31 strains. Overall, the number of translocations by isolate ranged from 0 to 4 (Figure 5A). A reciprocal translocation in the strain CECT10266 between the chromosomes VII and XII detected here has already been characterized twice (Hou et al., 2014; Istace et al., 2017) as being mediated by homologous recombination between two Ty2 retrotransposons. Extreme cases of translocations occur in the isolate UWOPS03-459.1 harboring four translocations with multiple successive translocation events happening in the same chromosome (Figure 5B) with the initial chromosome VII now being spread across four contigs in this strain suggesting three translocation events.

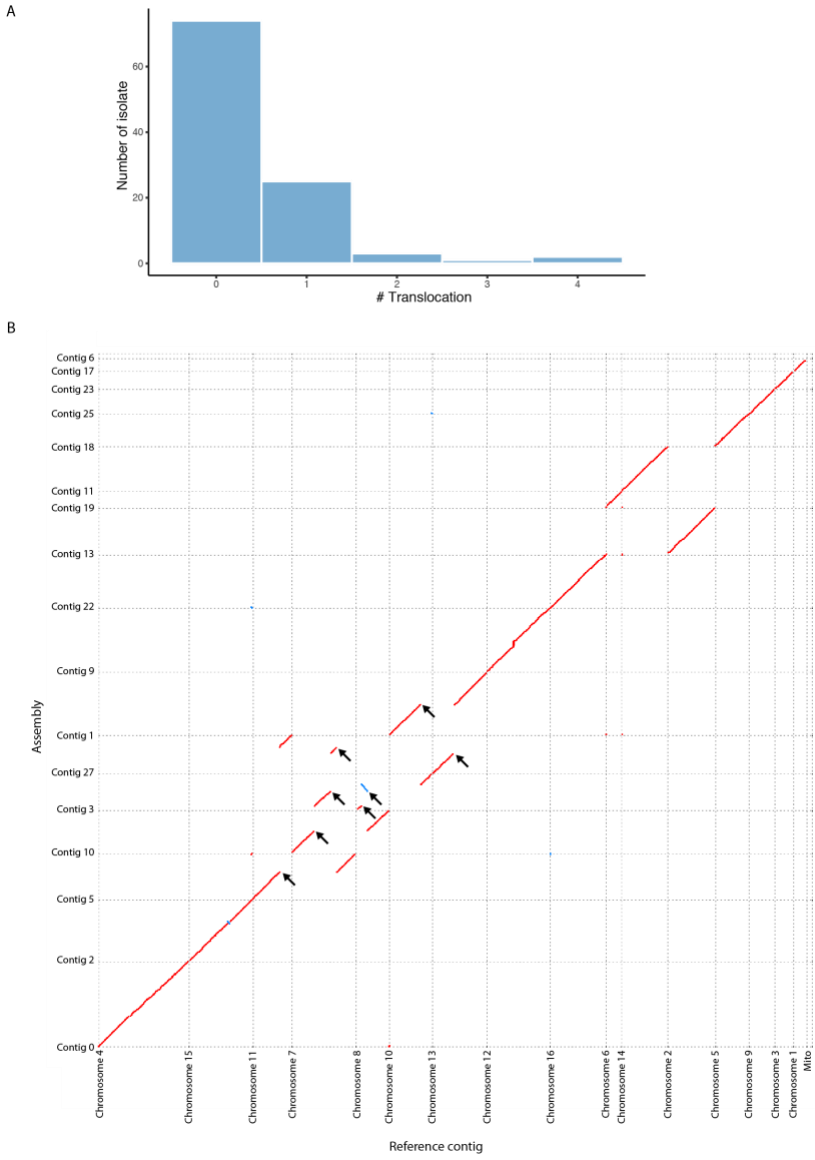


Figure 5. First overview of the translocations in *S. cerevisiae*
A. Number of translocations per strains. **B.** MUMmerplot of the UWOPS03-459.1 isolate. Red lines: sequences aligning in the same direction. Blue lines: sequences aligning in the opposite (Inversion). Black arrows indicate breakpoints of translocations.

A total of 148 inversions (of size >5kb) have been detected, ranging from 0 to 9 inversions per strain (Figure 6A). Interestingly, one inversion on the right arm of chromosome XIV and spanning approximately 35 kb has been detected in 75 strains suggesting that this particular inversion has been selected. It is located in a region that is flanked by Ty elements (Figure 6B) that could have recombined with each other which might explain the origin of this inversion. When looking at the position of the strains with and without this inversion on the tree of nucleotidic diversity of the species, no clear clustering of strains with and without this inversion seems to appear. However, all the strains belonging to the wine cluster and more generally strains belonging to a fermentation process harbor this inversion (Figure 6C).

Several limitations come from this analysis. The number of non-reciprocal translocations might be overestimated because translocations tend to happen next to telomeric regions so it is possible that reciprocal translocations happening within a few kilobases of the telomeres might not be visually detected as such. Another reason for the putative overestimation of non-reciprocal translocations and underestimation of reciprocal translocation events would be the poor contiguity of the assembly for some strains which complicates breakpoints detection. Indeed, if a translocation breakpoint is not resolved by the assembly, it will lie on a contig end and thus won't be detectable.

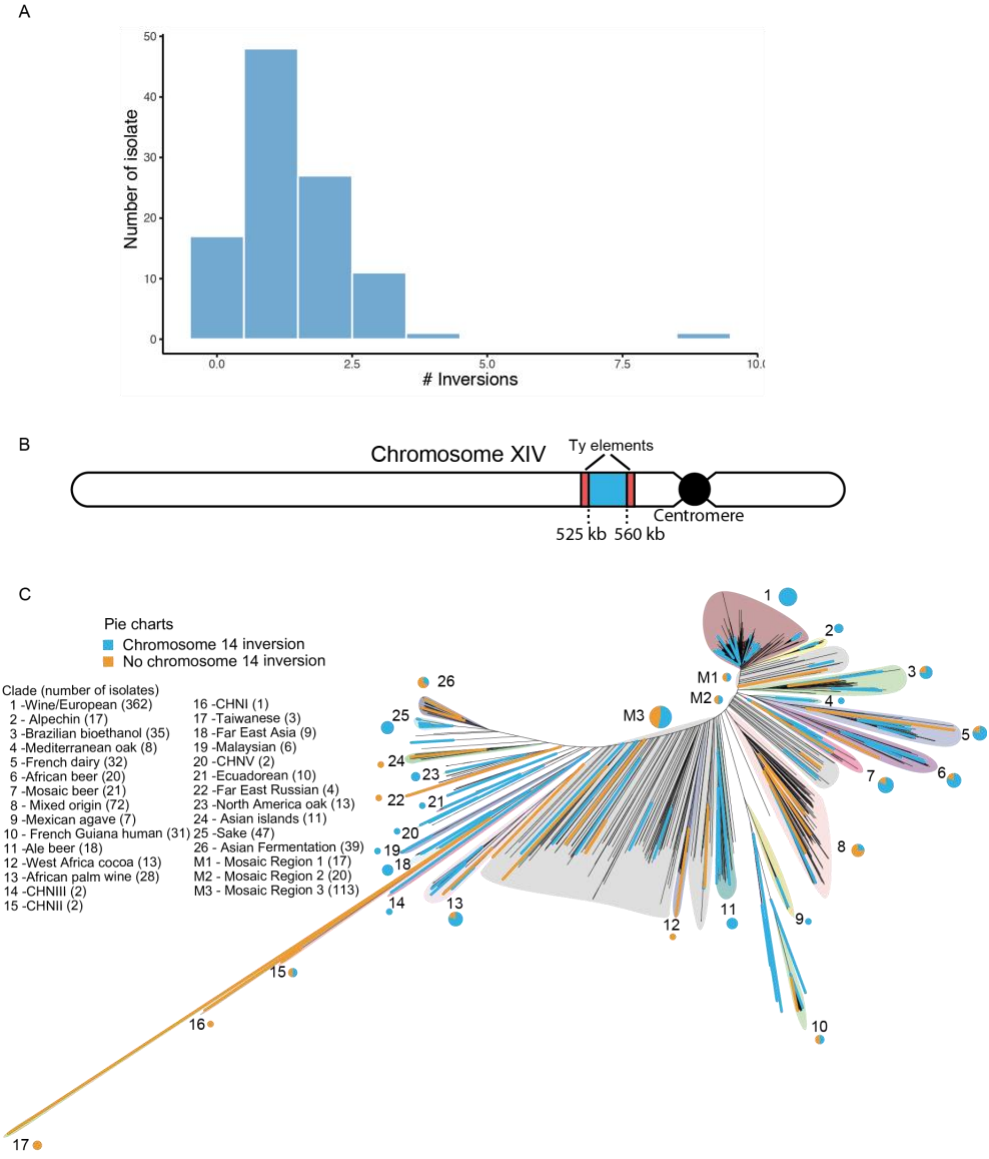


Figure 6. Overview of inversions.
A. Number of Inversions per strain. **B.** Schematic view of the inversion in chromosome 14. **C.** Neighbor-joining tree of 1,011 *S. cerevisiae* isolates. The 95 resequenced and assembled strains are color coded in blue where the inversion in chromosome 14 was detected and in orange for the strains were this inversion was not detected. Pie charts indicate the repartition of strains with and without this inversion in each cluster. Size of the pie charts reflects the number of strains.

This first glance at structural variation in the genomic landscape of *S. cerevisiae* allows to understand the important prevalence of chromosomal rearrangements with about a third of the strains displaying at least one translocation event. This result is lower than what has been previously reported by the study of 22 natural isolates of *S. cerevisiae* which suggested that 16 of these strains were carrying translocation events. However, as the present work is still in its premises, no real conclusions can be drawn at that time. First, all the assemblies need to be polished with Illumina reads to increase its accuracy. Then, the exact number of translocations and inversions can be refined and the exact breakpoints determined. Also, determination of the number and size of insertions and deletions has to be performed. An important point to take into account is that the 34 heterozygous strains still need to be phased in order to map potential heterozygous SVs in their genomes.

Once a catalog of structural variants will be available, the next step would be to gain knowledge about the phenotypic impact of structural variants. This is quite challenging especially for balanced variants. Indeed, in order to assess their phenotypic impact on the phenotypic landscape of the species, one solution would be to perform GWAS by considering the translocations or inversions for example. Once the catalog will be completed and polished, a genotyping of the detected SVs could be done in the entire population of 1,011 strains based on the Illumina reads already available so that frequencies of each SV can be determined. However, gaining enough detection power for obtaining genome-wide significance requires a variant to be present in at least 5% of the population which is very unlikely for such type of structural variants. Moreover, even if two events are similar, they might have completely distinct phenotypic outcome. Indeed, even the smallest difference in breakpoint position might be enough to change the phenotypic outcome of such variants. For example, a SV with a breakpoint inside a promoter region might not

have the same impact as if it was outside the promoter or inside a coding region although being only few dozens of base pairs apart. Although exact breakpoints were not determined yet, the example of the 35 kb inversion on chromosome 14 stands as a good example on how SVs can be positively selected for and reach high frequency in the population. Further validation is needed to know if such variation is linked to a phenotypic advantage or not.

Part 3: Assessing the phenotypic impact of structural variation through yeast chromosome reshuffling using CRISPR/Cas9

Both balanced and unbalanced SVs are known to have a phenotypic impact. However, the fitness effect of balanced SVs (Colson et al., 2004; Naseeb et al., 2016) has been less documented than CNVs, partly because they are much more challenging to map than CNVs. One way to assess the effect of balanced structural variants is to generate isogenic strains which differ only by chromosomal rearrangements. Previous studies demonstrated that double strand breaks (DSB) in dispersed repetitive elements such as the Ty retrotransposons can lead to chromosomal rearrangements (Argueso et al., 2008). Thus, generating multiple of those DSBs simultaneously in Ty elements might result in genome reshuffling with multiple translocations without being a gene or promoter disruptive event. For this purpose, a CRISPR-Cas9 based system to shuffle the yeast genome has been engineered by Aubin Fleiss in the group of Gilles Fischer. This system allows to generate DNA double strand breaks in long terminal repeats (LTR) regions of the yeast Ty3 retrotransposons which would then be randomly repaired with other homolog LTR regions. A gRNA with a target sequence is inserted in a pGZ110 plasmid containing the sequence of Cas9 endonuclease (Figure 7A). The target sequence can induce DSB in only 5 copies of Ty3 LTR located in chromosomes IV, VII, XV and XVI. Throughout the genome, 30 other copies of solo LTR display too much mismatches for the gRNA thus no DSB should theoretically occur at these sites. Chances are that after the Cas9 induced DSB, the homology region used for the repair comes from a different region of the genome thus leading to the generation of a balanced translocation (Figure 7B). This technique allows to study the effect of non-gene-disrupting translocations, meaning that the sole effect of the change in tridimensional configuration of the genome is assessed.

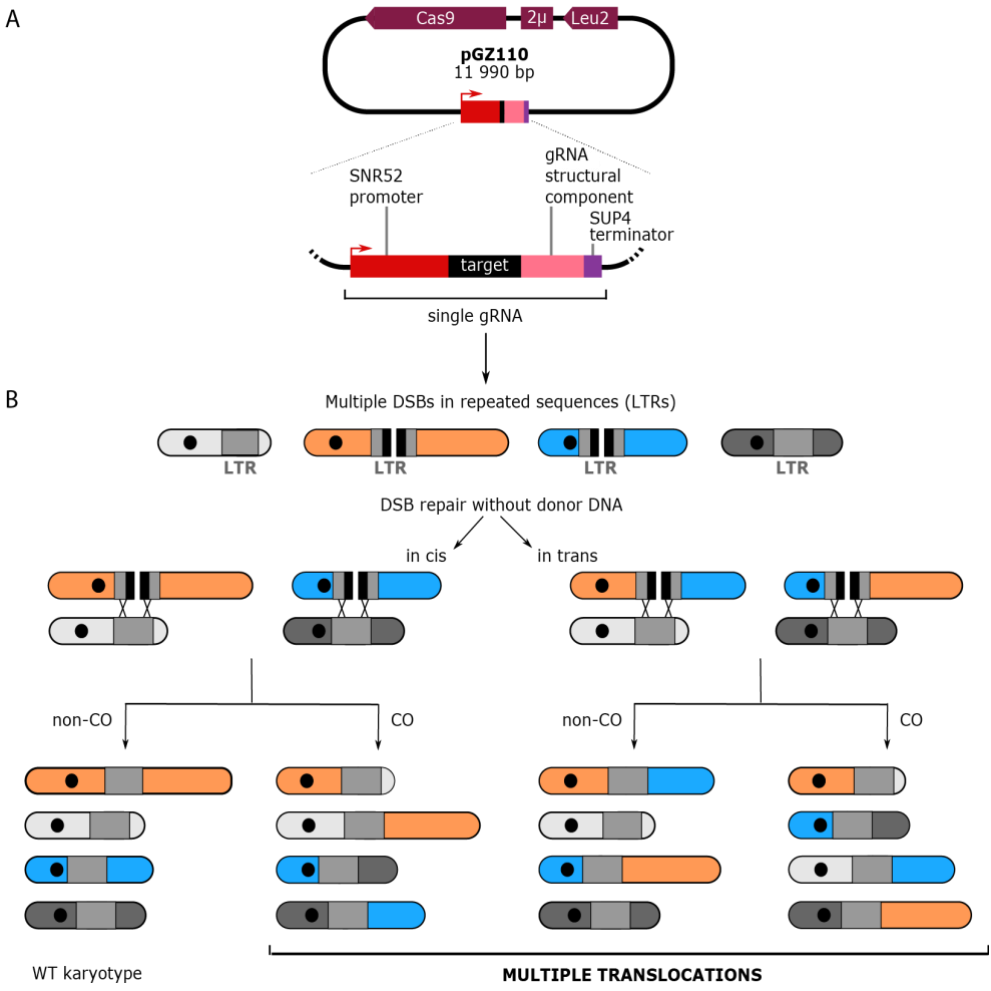


Figure 7. Strategy to reshuffle the yeast genome.
A. Cloning gRNA target sequences in the pGZ110 vector with a 20 bp oligonucleotide corresponding to the target sequence of a unique gRNA targeting LTRs of interest. **B.** Frequency of the reference allele at heterozygous sites along the genome. (Each color corresponds to a scaffold).

BY4741 and BY4742 cells were transformed with the Cas9/gRNA plasmid targeting the five Ty3 LTRs. In total, 211 and 159 transformants have been obtained, respectively. To assess the efficiency of the genome reshuffling, 69 transformants (37 BY4741 and 32 BY4742) were karyotyped with pulse field gel electrophoresis.

Among them, 30 showed clear chromosomal rearrangements on the gels, representing 18 different karyotypes in total over a total of 23 predicted of viable combinations of rearranged chromosomes (Figure 8A). This result demonstrated that genomes are efficiently reshuffled via this strategy. In total, 23 strains displayed the predicted karyotypic profiles (Figure 8B).

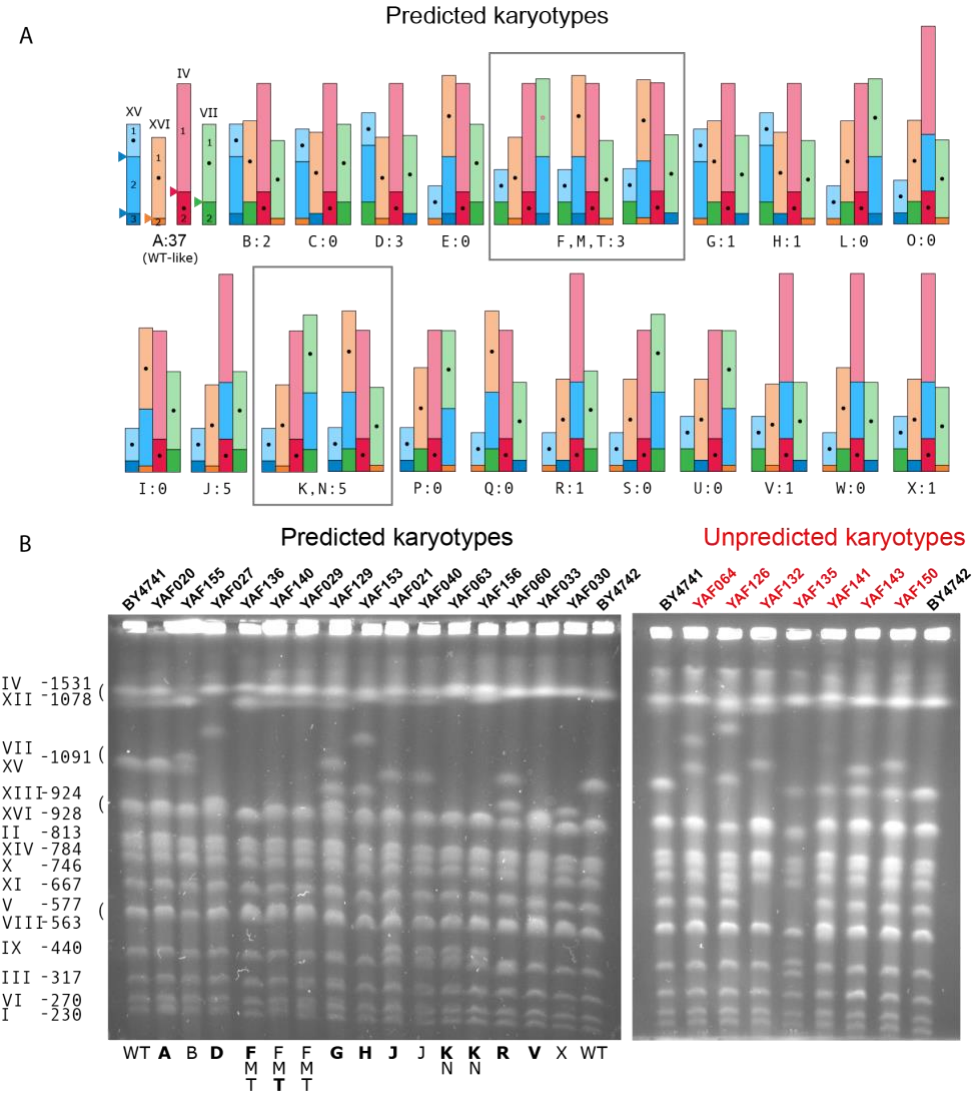


Figure 8. Induction of multiple rearrangements.
A. Predicted rearranged karyotypes (types A to X). Chromosomes are represented proportionally to their size in kb. Centromeres are represented by black dots. The chromosomal location of the 5 cutting sites corresponding to the 5 best matches to the gRNA are indicated by colored triangles on the type A profile. The number of strains of each type that were characterized by PFGE is indicated below each drawing. Types B to F have only 2 chimerical junctions resulting from a single reciprocal translocation between 2 chromosomes. Types G to M have 3 chimerical junctions resulting from translocations between 3 chromosomes (G, H, L, M) or the transposition of the chromosomal fragment XV.2 (I, J, K). Types N to V have 4 chimerical junctions resulting from a combination of translocations and transpositions. Types W and X have all 5 chimerical junctions. **B.** PFGE of 22 strains with predicted (left) and unpredicted (right) karyotypes. The control WT strains (BY4741 and BY4742) are located on the external lanes of each gel and their chromosome size is indicated on the left. The predicted types are in bold when all chimerical junctions were validated by PCR.

However, colony-PCR failed to validate the expected junctions for 16 of them. To investigate these junctions, Oxford Nanopore sequencing of the genomic DNA and *de novo* assembly of five strains allowed to characterize all rearrangements happening in these strains. Surprisingly, only one strain (YAF129) had the expected genome organization and no supplemental rearrangement. The four other strains had various additional rearrangements including simple duplication up to complex rearrangements involving multiple events (Figure 9). The karyotyping step did not allow to pinpoint those additional rearrangements because they resulted in very similar karyotypes. These unexpected rearrangements all happened in direct vicinity of transposable elements or solo LTRs (Figure 9). Moreover, seven strains did not display the expected karyotypes (Figure 8B) with translocations involving other chromosomes than the 4 initially targeted by the unique PAM sequence of the gRNA. Using once again a *de novo* assembly from Oxford Nanopore reads, the genome of the YAF064 exposed a reciprocal translocation between chromosome VII and XV. Although breakpoint on chromosome XV matches the targeted CRISPR cut site, the one on chromosome VII does not. Moreover, an important triplication of 110 kb flanked by LTRs regions is also present. All of these observations suggest that the unexpected rearrangements observed are most likely due to crossovers with uncut LTRs during search for homology and repair (Payen et al., 2008).

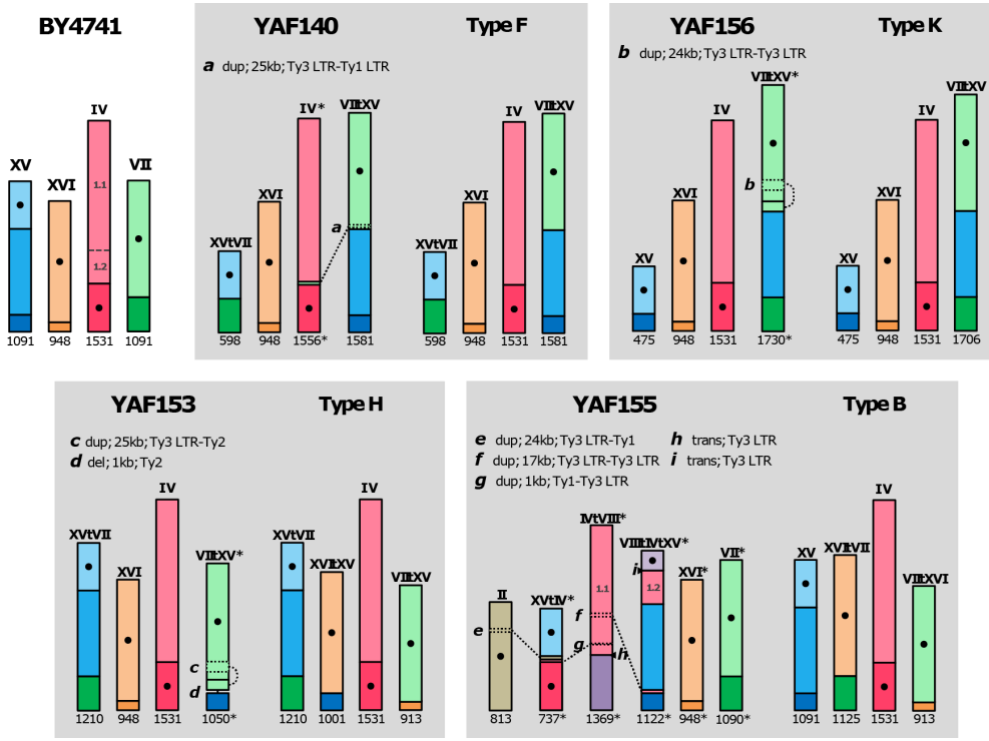


Figure 9. Long read sequencing resolved SVs in four strains with karyotype predicted rearrangements.

The wild type structure in BY4741 shows the 4 targeted chromosomes with black dots for centromeres. Each chromosome is fragmented by black lines representing DSB targeted Ty3 LTRs. Within chromosome IV a grey dotted line and sections named 1.1 and 1.2 represent an unpredicted position used for a reciprocal translocation in YAF155. Each shaded block contains both the chromosomal structure discovered using long read sequencing and the karyotype predicted by the CHEF gels (Type) for a single YAF strain. Below each chromosome is the size in kb. Stars on the chromosome name and size represent deviations from the corresponding karyotype predictions. Lower case italicized letters (a-i) represent unanticipated SVs captured by long reads. The key denotes from left to right, the type of SV (dup=duplication, del=deletion, trans=translocation), the size in kb and the repetitive element associated at the border of the element. For translocations only the SV type and repetitive element associated with the event is noted. For duplications, dotted lines represent the region duplicated and its new position. The SV d represents a deletion of 1kb from XV followed by recombination within full length Ty2 elements. For YAF155 two additional chromosomes, II (brown) and VIII (purple), were involved.

My role in this project has been to assess the phenotypic impact of such genome reshuffling. To do so, 22 rearranged strains were selected with 15 having a predicted karyotype and 7 having an unpredicted karyotype. As for the other phenotypic screens, these strains were phenotyped on 40 growth conditions (see Methods) which were compared to the growth in complete synthetic media. In total, 943 phenotypic measurements were performed. This led to the identification of reshuffled strains having a significant difference in phenotype compared to the original unshuffled strain. We identified 91 strain/trait combinations displaying superior growth and 48 displaying slower growth compared to their wild type counterpart.

The strongest phenotypic advantages of all corresponded to the strain with the unpredicted karyotype harboring the 110 kb triplication (YAF064, see above) when DNA synthesis is impaired (in the presence of the pyrimidine analog 5-fluorouracile) and in starvation (low carbon concentration 0.01% of galactose or glycerol) (Figure 10). However, none of 36 triplicated genes with a known function is directly involved either in DNA synthesis or starvation, suggesting that the other 18 uncharacterized genes present in this region could be involved in these phenotypes.

More generally, for the conditions that produce the greatest effects, most of the strains tended to react in a similar way. For instance, in the presence of 6-azauracil (6AU), 4NQO and high glucose concentration all the strains that showed a significant phenotypic variation grew slower than the WT while in the presence of galactose, caffeine, cycloheximide and fluconazole all the strains that showed a significant variation grew faster than the WT (Figure 10). Most strains (19 out of 22) showed variations in at least 2 different conditions showing that genome shuffling is efficiently broadening the phenotypic diversity.

The most variable strain, YAF132, presented significant growth variations in 17 out of the 40 conditions (faster and slower than the WT in 15 and 2 conditions, respectively). By opposition, the 2 type J strains (YAF021 and YAF040) as well as one strain with an unpredicted karyotype (YAF135) showed no phenotypic variation in nearly all the 40 conditions. The strain devoid of additional rearrangement as validated by sequencing (YAF129, see above) had significant growth variations in 13 conditions, including fitness advantage in many environmental conditions.

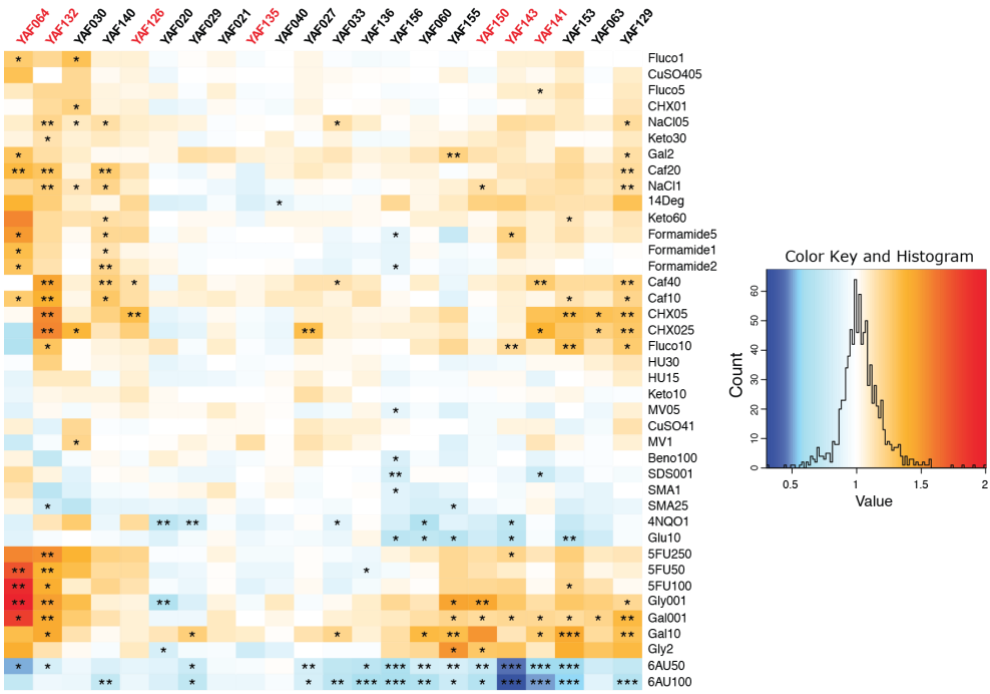


Figure 10. Phenotypic diversity of reshuffled strains.
Phenotypic variation among reshuffled strains. The heatmap represents the growth ratio of each strain (i.e. the colony size on the tested conditions divided by its size on SC) divided by the growth ration of BY4741 or BY4742, depending on the origin of the shuffled strain (Methods). The stars indicate the significant phenotypic effects (*, ** and *** indicate pval<10^{-2}, 10^{-3} and 10^{-4}, respectively). The strain names in red correspond to the unpredicted karyotypes in figure 8B.

We showed that the generation of scarless and markerless SVs in a fixed genetic background widens the phenotypic landscape accessible by a strain with many cases of fitness advantages. We could prove at least for one strain that the sole effect of changing the chromosomal configuration through balanced SVs such as reciprocal translocations without any gene or regulatory element disruption was sufficient to generate strong phenotypic diversity. As Ty3 LTRs, which are used as breakpoints in this study, are known to contain regulatory elements (Bilanchone et al., 1993), it is also possible that their transcriptional activity might result in the observed phenotypic diversity in the reshuffled strains. It is worthy to note that the strongest growth defect is observed in presence of 6-AU, a GTP depleting compound. Sensitivity to this compound is associated with mutations affecting transcriptional elongations (Exinger and Lacroute, 1992; Malagon et al., 2006; Mason and Struhl, 2005; Powell and Reines, 1996) suggesting global and important changes in the regulation of expression levels due to modification of the tridimensional conformation of the genome (Spielmann et al., 2018).

This part is a modified version of the publication:

Fleiss, A.*, O'Donnell, S.*, **Fournier, T.**, Lu. W, Agier, N., Delmas, S., Schacherer, J. and Fischer, G. Reshuffling yeast chromosomes with CRISPR/Cas9 (PloS Genetics, in press)

*: Equal contribution

References

Argueso, J.L., Westmoreland, J., Mieczkowski, P.A., Gawel, M., Petes, T.D., and Resnick, M.A. (2008). Double-strand breaks associated with repetitive DNA can reshape the genome. Proc. Natl. Acad. Sci. U. S. A.

Avramova, M., Cibrario, A., Peltier, E., Coton, M., Coton, E., Schacherer, J., Spano, G., Capozzi, V., Blaiotta, G., Salin, F., et al. (2018). Brettanomyces bruxellensis population survey reveals a diploid-triploid complex structured according to substrate of isolation and geographical distribution. Sci. Rep. 8, 4136.

Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M., and Phillippy, A.M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33, 623–630.

Bilanchone, V.W., Claypool, J.A., Kinsey, P.T., and Sandmeyer, S.B. (1993). Positive and negative regulatory elements control expression of the yeast retrotransposon Ty3. Genetics.

Borneman, A.R., Zeppel, R., Chambers, P.J., and Curtin, C.D. (2014). Insights into the Dekkera bruxellensis genomic landscape: comparative genomics reveals variations in ploidy and nutrient utilisation potential amongst wine isolates. PLoS Genet. 10, e1004161.

Chatonnet, P., Dubourdie, D., Boidron, J., and Pons, M. (1992). The origin of ethylphenols in wines. J. Sci. Food Agric. 60, 165–178.

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

Crauwels, S., Zhu, B., Steensels, J., Busschaert, P., De Samblanx, G., Marchal, K., Willems, K.A., Verstrepen, K.J., and Lievens, B. (2014). Assessing genetic diversity among Brettanomyces yeasts by DNA fingerprinting and whole-genome sequencing. Appl. Environ. Microbiol. 80, 4398–4413.

Curtin, C.D., and Pretorius, I.S. (2014). Genomic insights into the evolution of industrial yeast species Brettanomyces bruxellensis. FEMS Yeast Res. 14, 997–1005.

Curtin, C.D., Borneman, A.R., Chambers, P.J., and Pretorius, I.S. (2012). De-novo assembly and analysis of the heterozygous triploid genome of the wine spoilage yeast Dekkera bruxellensis AWRI1499. PLoS One 7, e33840.

Exinger, F., and Lacroute, F. (1992). 6-Azauracil inhibition of GTP biosynthesis in Saccharomyces cerevisiae. Curr. Genet.

Hellborg, L., and Piskur, J. (2009). Complex nature of the genome in a wine spoilage yeast, Dekkera bruxellensis. Eukaryot. Cell 8, 1739–1749.

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.

Istace, B., Friedrich, A., d'Agata, L.L., Faye, S.S., Payen, E., Beluche, O., Caradec,

C., Davidas, S., Cruaud, C., Liti, G., et al. (2017). de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. Gigascience 6, 1–13.

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, 239.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37, 540–546.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 27, 722–736.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. 5, R12.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32, 2103–2110.

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M., and Pevzner, P.A. (2016). Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl. Acad. Sci. U. S. A. 113, E8396–E8405.

Malagon, F., Kireeva, M.L., Shafer, B.K., Lubkowska, L., Kashlev, M., and Strathern, J.N. (2006). Mutations in the Saccharomyces cerevisiae RPB1 gene conferring hypersensitivity to 6-azauracil. Genetics.

Masneuf-Pomarede, I., Bely, M., Marullo, P., and Albertin, W. (2016). The genetics of non-conventional wine yeasts: Current knowledge and future challenges. Front. Microbiol. 6, 1563.

Mason, P.B., and Struhl, K. (2005). Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. Mol. Cell.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

Naseeb, S., Carter, Z., Minnis, D., Donaldson, I., Zeef, L., and Delneri, D. (2016). Widespread Impact of Chromosomal Inversions on Gene Expression Uncovers Robustness via Phenotypic Buffering. Mol. Biol. Evol.

Olsen, R.-A., Holmberg, K., Lötstedt, B., Käller, M., Vezzi, F., Bunikis, I., Pettersson, O.V., Tiukova, I., and Passoth, V. (2015). De novo assembly of Dekkera bruxellensis: A multi technology approach using short and long-read sequencing and optical mapping. Gigascience 4, 56.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067.

Payen, C., Koszul, R., Dujon, B., and Fischer, G. (2008). Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative

replication-based mechanisms. PLoS Genet.
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.
Piškur, J., Ling, Z., Marcet-Houben, M., Ishchuk, O.P., Aerts, A., LaButti, K., Copeland, A., Lindquist, E., Barry, K., Compagno, C., et al. (2012). The genome of wine yeast Dekkera bruxellensis provides a tool to explore its food-related properties. Int. J. Food Microbiol. 157, 202–209.
Powell, W., and Reines, D. (1996). Mutations in the second largest subunit of RNA polymerase II cause 6-azauracil sensitivity in yeast and increased transcriptional arrest in vitro. J. Biol. Chem.
Ruan, J., and Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. BioRxiv 530972.
Schifferdecker, A.J., Dashko, S., Ishchuk, O.P., and Piškur, J. (2014). The wine and beer yeast Dekkera bruxellensis. Yeast 31, 323–332.
Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. 19, 453–467.
Spitaels, F., Wieme, A.D., Janssens, M., Aerts, M., Daniel, H.-M., Van Landschoot, A., De Vuyst, L., and Vandamme, P. (2014). The Microbial Diversity of Traditional Spontaneously Fermented Lambic Beer. PLoS One 9, e95384.
Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963.
Yue, J.-X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., Bergström, A., Coupland, P., Warringer, J., Lagomarsino, M.C., et al. (2017). Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat. Genet. 49, 913–924.

METHODS

Wet lab procedures

Selection of the *Saccharomyces cerevisiae* isolates

Out of the collection of 1,011 strains (Peter et al., 2018), a total of 53 natural isolates were carefully selected to be representative of the *Saccharomyces cerevisiae* species. We selected isolates from broad ecological origins and we prioritized for strains that were diploid, homozygous, euploid and genetically as diverse as possible, *i.e.* up to 1% of sequence divergence. All the isolate details, including ecological and geographical origins, are listed in table 1. In addition to these 53 isolates, we included two laboratory strains, namely ∑1278b and the reference S288c strain (Table 1).

Table 1. Strains used in chapter 1 and 2. Strains in bold are used in chapter 2.

Strain Name	Isolation	Ecological Origin	Continent	GWAS included
∑1278b	NA	Laboratory	NA	Yes
1560	Manzanilla-Alorena, olive (Noe)	Nature	Europe	Yes
2162	Forest soil, 30C	Soil	Europe	
2187	Forest soil, 30C	Soil	Europe	Yes
BJ20	Bark from Quercus wutaishanica	Tree	Asia	Yes
CECT10109_1b	Prickly pear	Fruit	Europe	Yes
CLIB1071	Cider brewery, dry cider	Cider	Europe	Yes
CLIB1410	Rice wine. Oenology	Fermentation	Asia	
CLIB154_1b	Wine	Wine	Europe	
CLIB382_1b	Beer	Beer	Europe	Yes
CLIB413_1b	Fermenting rice beverage	Fermentation	Asia	Yes
CLIB485	Cider brewery	Cider	Europe	
CLQCA_04-021	Beetle	Insect	South America	Yes
CLQCA_10-027	Grass	Nature	South America	
CLQCA_20-184	Flower from Heliconia sp.	Flower	South America	Yes
CLQCA_20-246	Termite mound	Insect	South America	Yes
CLQCA_20-259	Decaying fruit	Fruit	South America	
DBVPG1058	Baker's yeast	Bakery	Europe	
DBVPG1564	Grape must	Wine	Europe	Yes
DBVPG2088	Cognac	Distillery	Europe	
DBVPG3591_1b	Cocoa beans	Nature	NA	Yes
ES4M07	Fruiting body of Geastrum sp.	Fruit	Asia	Yes
EXF-5247	Seawater in harbour	Water	Europe	Yes
EXF-5248	Seawater in harbour	Water	Europe	Yes
EXF-5295	Kefyr	Fermentation	Europe	

EXF-5297	Mashed pears	Fruit	Europe	
EXF-7197	Quercus sp.	Tree	Europe	Yes
FY4	NA	Laboratory	NA	Yes
HN10	Rotten wood	Nature	Asia	Yes
HN15	Rotten wood	Nature	Asia	Yes
HN16	Soil	Soil	Asia	Yes
I14_1b	Vineyard soil	Wine	Europe	Yes
L-1374	Wine	Wine	South America	
NC_02_b	Exudate from Quercus sp.	Tree	North America	Yes
NPA03.1	Palm wine	Palm wine	Africa	Yes
sample 40	Tree leaves	Tree	Europe	Yes
T7_b	Exudate from Quercus sp.	Tree	North America	
UC8_1b	Wine	Wine	Africa	
UCD_05-780	Beetle	Insect	North America	
UCD_09-448	Olives	Fruit	North America	Yes
WE372_1b	Wine	Wine	Africa	
Y12_1b	Palm wine	Palm wine	Africa	Yes
YJM326_b	Human, clinical	Human. clinical	North America	Yes
YJM421_b	Ascites fluid	Human. clinical	North America	Yes
YJM434_1b	Human, clinical	Human. clinical	NA	
YJM627	Seg, Y55	NA	Europe	Yes
YJM990	Clinical	Human, clinical	North America	
YPS133	Soil beneath Quercus alba	Soil	North America	
YPS141	Soil beneath Quercus velutina	Soil	North America	
YPS142	Surface of Tuber magnatum	Nature	North America	
YPS143	Banana wine	Wine	North America	Yes
YPS163	Soil beneath Quercus rubra	Soil	North America	Yes
YPS615	Quercus sp.	Tree	North America	
YPS617	Quercus sp.	Tree	North America	Yes
ZP_611	Quercus robur	Tree	North America	Yes

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 (Table 2) 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). No significant mating type specificity was identified.

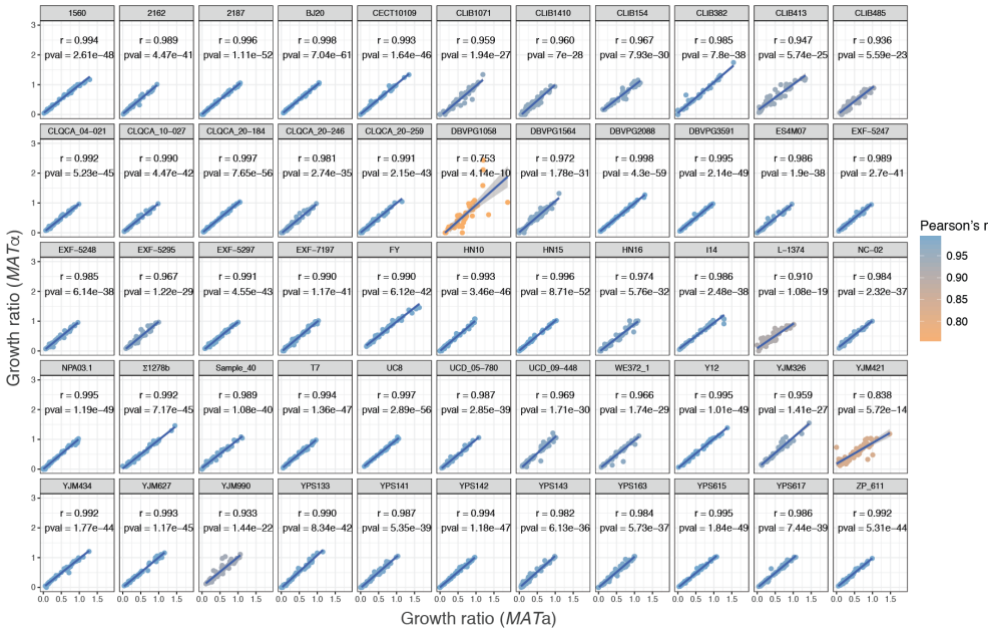


Figure 1. Phenotypic correlation between *MAT*a and *MAT*α isolates
Correlation of the phenotypes between mating types by strain. Pearson's r and corresponding p-values are indicated for each strain. Blue line indicates fitted linear model and grey envelope represents 95% confidence interval

Diploid 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, *i.e.* 55 *MAT*a *HO::*Δ*KanMX* and 55 *MAT*α *HO::*Δ*NatMX* strains were arrayed and mated in a pairwise manner on YPD for 24 hours 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 hours, plates were replicated again on the same media to eliminate potential residuals of non-hybrids cells. In total, we generated 3,025 hybrids, representing 2,970 heterozygous hybrids with a unique parental combination and 55 homozygous hybrids.

Selection of collinear strains

In chapter 2, as we wanted to maximize the chance of obtaining viable progenies, only considering colinear strains was a way to first remove important bias in offspring viability as translocations drastically impede viability by 25 or 50% with a majority of tetrads with three or two viable spores (Hou et al., 2014). Obtaining fully viable tetrads was of prime interest for us as we wanted to perform segregation analysis of the phenotype in tetrads.

All 55 strains were crossed with FY4 (reference strain), sporulated and 5 tetrads were dissected for each hybrid to check their collinearity. Colinear strains would show a majority of tetrads with four viable spores. This step lead to the characterization of 27 strains being colinear out of the 55 initially selected for the construction of the diallel hybrid panel (Figure S1 in Chapter 2). In chapter two, 20 of the strains that are colinear with the reference strain were selected to be as representative of the whole diversity of the species (Table 1).

Generation of large haploid progenies for 20x20 diallel cross

The 20 parental strains are a subset of the previous parental strains forming the diallel panel of 3,025 hybrids. The resulting 20 by 20 diallel cross encompassed 190 hybrids without reciprocal crosses. Each hybrid was sporulated for two days on a medium containing only 1% of potassium acetate. For each of the 190 crosses, enough spores were dissected in order to obtain 160 haploid progenies originating from 40 fully viable tetrads *i.e.* with all four spores viable. The manual dissection of 66,992 spores using the Singer SporePlay micromanipulator lead to a total panel of 30,400 haploid individuals coming from complete tetrads. In order to facilitate dissections, tetrads were incubated 15 minutes in a solution with 1.5e-2 $mg.ml^{-1}$ of zymolyase to gently digest the ascus wall. After dissection of the ascus wall, each spore of a tetrad is arrayed on solid YPD to retain the tetrad information. After incubation of 48h at 30°C, viable spores are determined by the formation of a colony.

Spore viability analysis

Spore viability has been assessed for each cross as the number of colony forming spores divided by the total number of dissected spores. Information about the number of viable spores per tetrad was also retained as this gives indications for inferring reproductive isolation mechanisms. The nucleotidic diversity between the parental isolates was computed as the number of SNPs differentiating the two genomes divided by the overall genome length.

High-throughput phenotyping and growth quantification

Quantitative phenotyping was performed using 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 1,536 density format using the replicating ROTOR™ robot (Singer Instruments). The plates were incubated for 24 hours at 30°C (except for 14°C phenotyping) and picture was taken at a 12Mpx resolution.

To correct for unevenness of cell spotting during plate replication, a picture of each plate was taken right after replication and spot size measured for each colony. This initial measurement was then subtracted to the final colony size after growth to get corrected colony sizes, thus strongly reducing experimental noise. Negative corrected values were adjusted to 0. 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 corrected colony size on a condition and the corrected colony size on SC. The value considered as the phenotype is the median of all the replicates, thus smoothing the effects of pinning defect or contamination.

- **Phenotyping of the diallel hybrids panel (Chapter 1)**

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 (Table 2). The selected conditions impact a broad range of cellular responses, and multiple concentrations were tested for each compound. Most tested conditions displayed distinctive phenotypic patterns, suggesting different genetic basis for each of them. This phenotyping step led to the determination of 148,225 hybrid/trait combinations.

Table 2. Phenotyping conditions and their respective type of induced stress

Categories	Sub-categories	Conditions	Abbreviation	Chapter
Reference		SC		
Cell wall	Membrane stability	SC SDS 0.01%	SDS001	1,2,3
		SC SDS 0.025%	SDS0025	1,2
		SC SDS 0.05%	SDS005	1,2
	Ergosterol synthesis	SC fluconazole 1 µg/ml	Fluco1	1,2,3
		SC fluconazole 5 µg/ml	Fluco5	1,2,3
		SC fluconazole 10 µg/ml	Fluco10	1,2,3
	Erg synthesis + multiple targets	SC ketoconazole 10 µg/ml	Keto10	1,2,3
		SC ketoconazole 30 µg/ml	Keto30	1,2,3
		SC ketoconazole 60 µg/ml	Keto60	1,2,3
Cold		SC 14°C	14Deg	1,2,3
DNA metabolism	Telomere dynamics	SC sodium (meta)arsenite 1 mM	SMA1	1,2,3
		SC sodium (meta)arsenite 2.5 mM	SMA25	1,2,3
		SC sodium (meta)arsenite 5 mM	SMA5	1,2
	DNA damage	SC 4-NQO 1 µg/ml	4NQO1	1,3
		SC 4-NQO 2 µg/ml	4NQO2	1
		SC 4-NQO 3 µg/ml	4NQO3	1
	DNA synthesis	SC 5-FU 50 µg/ml	5FU50	1,2,3
		SC 5-FU 100 µg/ml	5FU100	1,2,3
		SC 5-FU 250 µg/ml	5FU250	1,2,3
		SC Hydroxyurea 15 mg/ml	HU15	3
		SC Hydroxyurea 30 mg/ml	HU30	3
General cellular damage		SC CuSO4 0.1 mM	CuSO401	1,2
		SC CuSO4 0.5 mM	CuSO405	1,2,3
		SC CuSO4 1 mM	CuSO41	1,2,3
Metabolism	Carbon sources utilization	SC galactose 2%	Gal2	1,2,3
		SC glycerol 2%	Gly2	1,2,3
	Carbon starvation	SC glucose 0.01%	Glu001	1,2
		SC galactose 0.01%	Gal001	1,2,3
		SC glycerol 0.01%	Gly001	1,2,3
	High carbon source tolerance	SC glucose 10%	Glu10	1,2,3
		SC galactose 10%	Gal10	1,2,3
		SC glycerol 10%	Gly10	1,2
Osmotic stress		SC NaCl 0.5 M	NaCl05	1,2,3
		SC NaCl 1 M	NaCl1	1,3
Oxydative stress		SC methyl viologen 0.5 mM	MV05	1,2,3
		SC methyl viologen 1 mM	MV1	1,2,3
		SC methyl viologen 2.5 mM	MV25	1,2
Protein stability		SC formamide 1%	Form1	1,2,3
		SC formamide 2%	Form2	1,2,3
		SC formamide 5%	Form5	1,2,3
Signal transduction pathways		SC caffeine 10 mM	Caf10	1,2,3
		SC caffeine 20 mM	Caf20	1,2,3
		SC caffeine 40 mM	Caf40	1,2,3
Subcellular organisation	Microtubules function	SC benomyl 50 µg/ml	Beno50	1
		SC benomyl 100 µg/ml	Beno100	1,3
Translation	Ribosomes function	SC cycloheximide 0.1 µg/ml	CHX01	1,2,3
		SC cycloheximide 0.25 µg/ml	CHX025	1,2,3
		SC cycloheximide 0.5 µg/ml	CHX05	1,2,3
Transcription	GTP and UTP nucleotide pools	SC 6-azauracil 50 µg/ml	6AU50	1,3
		SC 6-azauracil 100 µg/ml	6AU100	1,3
		SC 6-azauracil 200 µg/ml	6AU200	1

- **Phenotyping of haploid offspring panel (Chapter 2)**

We measured two biological replicates of each haploid progeny. For the parental strains, four biological replicates per plate and 19 technical replicates were made as each parent was present in 19 crosses. The 30,400 haploid spores as well as the 20 parental strains in both mating types were phenotyped on 40 growth conditions (Table 2) leading to more than 2,500,000 phenotypic measurements.

- **Phenotyping of CRISPR reshuffled strains (Chapter 3)**

Each reshuffled strain had six biological replicates and the reference strains BY4741 and BY4742 were present 96 times on each condition. Raw sizes were corrected using two successive corrections: a spatial smoothing was applied to the colony size (Baryshnikova et al., 2010). This allowed to account for variation of the plate thickness. Another correction was then applied to rescale colony size by row and column (Baryshnikova et al., 2010) which is important for colonies lying at the edges of the plate thus having easier access to nutrients compared to strains in the center. All calculations were performed using R. Once the corrected sizes were obtained, the growth ratio of each colony was computed as the colony size on the tested conditions divided by its size on SC. To detect the phenotypic effect of the engineered translocations, each growth ratio has been normalized by the growth ratio of BY4741 or BY4742, depending on the origin of the shuffled strain, on the 40 tested condition (Table 2). Each experiment was repeated 2 times independently in the 40 growth conditions.

CRISPR-Cas9 genome editing

- **gRNA cloning**

pGZ110 plasmid containing Cas9 endonuclease coding sequence and a backbone of gRNA is first linearized using *LguI* restriction enzyme, creating 5' and 3' overhangs of 3 bp. In order to clone a single 20 bp target sequence in the gRNA backbone, two 23 bp complementary oligonucleotides with 3bp overhangs on 5' and 3' complementing the *LguI* restriction site are annealed. Annealing is achieved by mixing equimolar mix of both oligonucleotides, heating at 95°C for 5 minutes and cool at room temperature to allow for slow and correct annealing. The resulting double stranded insert is then ligated to the linearized plasmid.

- **Chromosomal reshuffling**

pGZ110 plasmid containing a 20bp insert targeting the Ty3-LTRs of interest has been transformed without any repair fragment to allow for translocation events to happen. After transformation, cells were plated on YPD to check for viability and on synthetic medium depleted of leucine to select for transformants. This was done by Aubien Fleiss.

- **Allele editing of *SGD1***

The pAEF5 plasmid is the same as pGZ110 except for the *LEU2* cassette that has been replaced by a *HygMX* cassette providing resistance to hygromycin. This allows to use the plasmid in prototrophic genetic backgrounds. 20 bp gRNA targeting *SGD1* has been cloned into the pAEF5 plasmid as explained in the previous section (gRNA cloning). This plasmid 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 hours to induce plasmid loss. The plate was then pinned back onto solid YPD for

24h 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 (4 technical replicates and 4 biological replicates) on the corresponding condition to measure differences with their wildtype counterparts.

Karyotyping yeast by Pulsed Field Gel Electrophoresis

Karyotyping of the reshuffled yeast were performed by preparing agarose plugs allowing to keep yeast chromosomes intact. This step has been performed following standard procedure (Török et al., 1993). Plugs were then placed in a 1% Seakem GTC agarose and 0.5x TBE gel. PFGE was conducted with the CHEF-DRII (BioRad) system following: 6 V/cm for 10 hours with a switching time of 60 seconds followed by 6 V/cm for 17h with switching time of 90 seconds. The included angle was 120° for the whole duration of the run.

Computational analysis

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 ($i = 1...N$, $j = 1 \ldots N$, $N = 55$). General combining ability for the i^{th} parent is calculated as:

$$\widehat{g_i} = \left(\frac{N-1}{N-2}\right) \times (\overline{z_{i\cdot}} - \mu)$$

Where N is the total number of parental types, $\overline{z_{i\cdot}}$ 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 $\sigma^2 z_{ij.}$ is the variance of fitness values from a full-sib family involving the *i*th and *j*th parents, which is expressed as:

$$\sigma^2 z_{ij.} = \sigma^2 z_i + \sigma^2 z_j + \sigma^2 z_{ij} + 2 \times cov(z_i, z_j)$$

Specific combining ability for the $i \times j$ hybrid combination therefore:

$$\widehat{s_{ij}} = \overline{z_{ij.}} - \widehat{g_i} - \widehat{g_j} - \mu$$

The error term associated with $\widehat{s_{ij}}$ is:

$$e_{s_{ij}} = \sqrt{\frac{(N-3) \times \sigma^2 z_{ij.}}{n \times (N-1)}}$$

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_A^2) over the total phenotypic variance observed (σ_P^2):

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} = \frac{\sigma^2_{(g_i+g_j)}}{\sigma^2_{(g_i+g_j)} + \sigma^2_{s_{ij}} + \sigma_e^2}$$

Where $\sigma^2_{(g_i+g_j)}$ is the sum of GCA variances, $\sigma^2_{s_{ij}}$ is the SCA variance and σ_e^2 is the variance due to measurement error, which is expressed as:

$$\sigma_e^2 = (N-2)\overline{\left(\overline{e_{g_i}} + \overline{e_{g_j}}\right)}^2 + \frac{\left(\frac{(N^2-N)}{2} - 1\right)}{\left(\frac{(N^2-N)}{2} + N - 3\right)} \times \overline{e_{s_{ij}}}^2$$

On the other hand, broad-sense heritability (H^2) depicts the part of the phenotypic variance explained by the total genetic variance σ_G^2:

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2} = \frac{\sigma^2_{(g_i+g_j)} + \sigma^2_{s_{ij}}}{\sigma^2_{(g_i+g_j)} + \sigma^2_{s_{ij}} + \sigma^2_e}$$

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.

Computation 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 . To obtain a robust classification, confidence intervals for each class were based on the standard deviation of hybrid (6 replicates) and parents (54 replicates) (Table 3). *P2* is the phenotypic value of the fittest parent while *P1* is the phenotypic value of the least fit parent.

Table 3. Confidence intervals for the classification in different inheritance mode

Inheritance mode	Formula
Underdominance	$Hyb < P1 - (\sigma_{P1} + \sigma_{Hyb})$
Dominance P1	$P1 - (\sigma_{P1} + \sigma_{Hyb}) < Hyb < P1 + (\sigma_{P1} + \sigma_{Hyb})$
Partial dominance P1	$P1 + (\sigma_{P1} + \sigma_{Hyb}) < Hyb < MPV - \left(\frac{\sigma_{P1} + \sigma_{P2}}{2} + \sigma_{Hyb}\right)$
Additivity	$MPV - \left(\frac{\sigma_{P1} + \sigma_{P2}}{2} + \sigma_{Hyb}\right) < Hyb < MPV + \left(\frac{\sigma_{P1} + \sigma_{P2}}{2} + \sigma_{Hyb}\right)$
Partial dominance P2	$MPV + \left(\frac{\sigma_{P1} + \sigma_{P2}}{2} + \sigma_{Hyb}\right) < Hyb < P2 - (\sigma_{P2} + \sigma_{Hyb})$
Dominance P2	$P2 - (\sigma_{P2} + \sigma_{Hyb}) < Hyb < P2 + (\sigma_{P2} + \sigma_{Hyb})$
Overdominance	$P2 + (\sigma_{P2} + \sigma_{Hyb}) < Hyb$

When a clear separation is possible between the two parental phenotypic values ($P1 + \sigma_{P1} < P2 - \sigma_{P2}$), the full decomposition in the seven above mentioned categories is possible (Table 3). 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. 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, 2010). 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, *i.e.* 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., 2014). 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 trait-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. Taken together, GWA revealed 1,723 significantly associated SNPs (Figure 4-Source Data 1), with 1,273 and 450 SNPs for overdominant and additive model, respectively.

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. Significant enrichment is considered under "Process" ontology with a p-value cutoff of 0.05.

Random forest classifier

Assessing bimodality on high number of distributions is very challenging. Although multiple statistical methods exist, they would often fail to precisely detect cases of bimodality vs unimodality in our very diverse dataset. To counter that while still retaining systematical and unbiased assessment of bimodality in our large dataset, we developed a random forest classifier approach. A random forest is a machine learning algorithm that works by building a large number of decision trees *i.e.* a forest, to cluster different observations in different groups. Each tree is different from the other as each node branching the tree is determined by randomly chosen variables among all available variables in the dataset. All decision trees are run independently and the majority is voting (ensemble method).

Variables used for the random forest

We first performed expectation maximization (EM) to fit 2 gaussian distributions over every distribution and extracted the following 5 parameters: $\hat{\pi}$,the proportion of observations in the main cluster; $\hat{\mu}_1$ and $\hat{\mu}_2$, the estimated means of each mode; $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, the estimated variances of each mode. From these five parameters, 13 variables were computed to use as input for the random forest (Table 4).

Table 4. Variables used in the random forest

Variable	Formula
Proportion in bigger cluster	$X_1 = \hat{\pi}$
Difference of means	$X_2 = \lvert\hat{\mu}_1 - \hat{\mu}_2\rvert$
Smallest variance	$X_3 = min(\hat{\sigma}_1^2, \hat{\sigma}_2^2)$
Highest variance	$X_4 = max(\hat{\sigma}_1^2, \hat{\sigma}_2^2)$
Ratio of variances	$X_5 = \frac{max(\hat{\sigma}_1^2, \hat{\sigma}_2^2)}{min(\hat{\sigma}_1^2, \hat{\sigma}_2^2)}$
Unbiased estimator of variance	$X_6 = \widehat{Var}(X) = S^2 = \frac{n}{n-1} S_n^2$
p-value of EM test applied to the distribution	$X_7 = \mathbb{P}(EM_n^{(K)} > q_{((\chi_{(2)})^2_{.095})}$
Asymmetry coefficient estimation	$X_8 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)(\hat{\sigma}^2)^{3/2}}$
Kurtosis estimation	$X_9 = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)(n-2)(n-3)(\hat{\sigma}^2)^2} - 3\frac{(n-1)^2}{(n-2)(n-3)}$
d value (Holzmann and Vollmer, 2008)	$X_{10} = \frac{\lvert\hat{\mu}_1 - \hat{\mu}_2\rvert}{2\sqrt{\hat{\sigma}_1\hat{\sigma}_2}}$
u value (Behboodian, 1970)	$X_{11} = \frac{\lvert\hat{\mu}_1 - \hat{\mu}_2\rvert}{2(\min(\hat{\sigma}_1, \hat{\sigma}_2)}$
$\Delta\mu$ value (Ashman et al., 1994)	$X_{12} = \frac{\sqrt{2}\lvert\hat{\mu}_1 - \hat{\mu}_2\rvert}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}$
bimodality coefficient (Pfister et al., 2013)	$X_{13} = \frac{X_8^2 + 1}{X_9 - 3 + \frac{3(n-1)^2}{(n-2)(n-3)}}$

Determination of a training set

Random forest first need a training set with already pre-annotated observations to train and find the tree that will be the most representative of the rest of the data. One drawback of this is that the final tree might be overfitted to the training set. Therefore, building a good training set is of prime interest to generate the forest, as this set will serve as the foundation from which all trees will be generated. A good training set must fulfill two main criteria to prevent overfitting: it needs to be representative of the sample and needs to be equilibrated between classes in order to avoid the predictor to have a stronger power towards one or the other class.

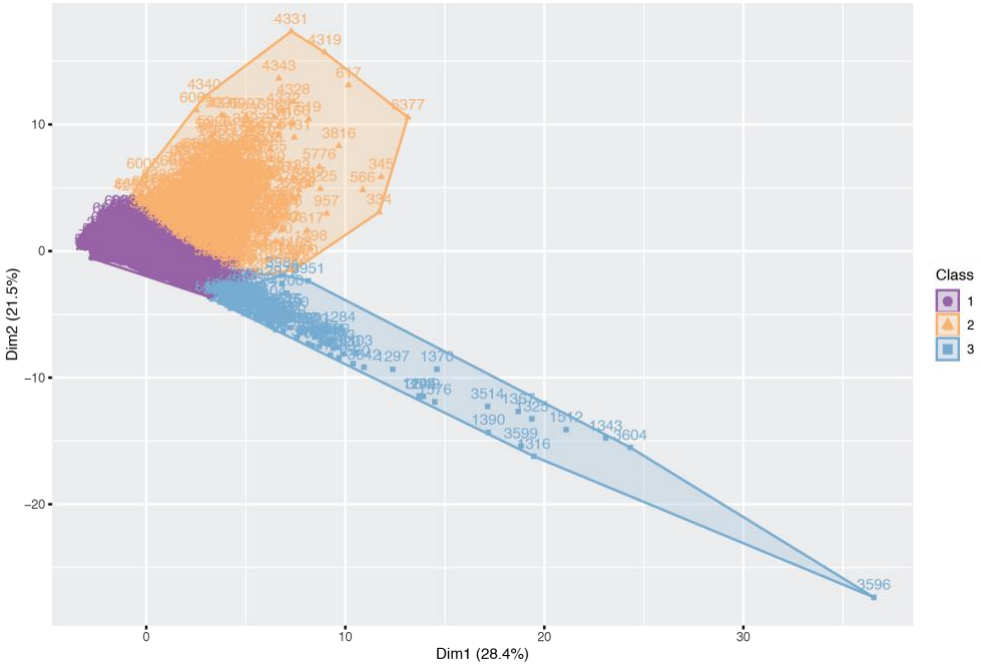


Figure 2. Hierarchical clustering based on principal component of all distributions

Separation on the first two dimensions on the 13 descriptive variables of each phenotypic distribution. Number represent the index of each distribution. Hierarchical clustering then finds the best number of classes (here three) and classified each distribution.

Once all the values described in table 4 have been computed for each distribution, we used hierarchical clustering based on principal component with the FactoMineR R package (Lê et al., 2008) to classify the distribution based on their parameters. We found three classes (Figure 2) with very different proportion : 88% of the total observations was in class1, 10% in class 2 and only 2% in class 3. To construct the training dataset while remaining representative of the initial sample, we iteratively selected a proportional amount of observations from each of those classes. Each iteration adds 30 observations and computes the precision based on the resulting confusion matrix:

$$precision = \frac{TP + TN}{TP + FP + TN + FN}$$

Where TP is the number of true positive, TN the number of true negative, FP the number of false positive and FN the number of false negative.

This sampling algorithm stops after two successive iterations showed less than 0.01% of improvement. With this sampling method, the final training set contained 510 phenotypic distributions coming from all tested conditions.

Experimental noise measurement

A measure of experimental noise was required to make distinction between significantly different measurements and variation due to our experimental design. In order to assess this noise in a systematic manner, we compared two identical but independent experiment of measuring colony size on standard complete growth medium of our entire dataset. The ratio between the two measurement is expected to be one. Any deviation would be imputable to noise. Thus we defined the noise as the mean standard deviation of this ratio for each cross.

Cluster assignment for parental strains

For each cross/trait combinations displaying bimodality, position of the parental phenotypes relative to the two modes of the offspring phenotypes has to be assessed. To infer the belonging of a parental phenotype to one or the other cluster, Wilcoxon test was performed between all phenotypic values coming from the replicates of a parent and all the offspring coming from the same cluster. Values of the test *i.e.* the difference in median of the two samples is extracted. The parent is inferred to the cluster that is closer to him *i.e.* with the smallest difference in median.

Decision Tree

Once modality of the distribution has been assessed. Other parameters still need to be checked to infer genetic complexity of a given cross/trait combination. Two parameters allow to precise the genetic complexity if a bimodal distribution is detected: the parental phenotype and the segregation of the tetrad regarding the two clusters. This tree is represented as Figure S3 in chapter 2. It allowed for a differentiation between complex traits, monogenic traits and oligogenic traits (Figure S3, chapter 2). Furthermore, as we benefit from the parental phenotype, based on their position relative to the two modes of a bimodal distribution, we can differentiate between a recessive epistatic interaction and the special case of a modifier gene acting as a suppressor (Figure S3, chapter 2). All traits categorized as oligogenic were manually curated to correct for misclassifications by the decision tree.

Filtering step

Although the noise in our experimental design is relatively contained, if two parents have the same phenotypic value, it is impossible to make a difference between an unimodal distribution and a bimodal distribution with two modes centered on each parent as they would have the same theoretical mean. With our analysis pipeline, such cases would be detected as a complex trait because of the unimodal nature of the offspring distribution thus leading to an overestimation of complex traits and underestimation of monogenic and/or oligogenic traits. Therefore, we only kept distributions with parents having distinct phenotypes, *i.e.* the difference between the parental phenotype has to be greater than the noise (|P1-P2| > noise). This filtering step left only 3,841 out of 7,600 phenotypic distributions. Although drastically reducing the number of cross/trait combination considered, this allowed to have a more robust estimation of trait complexity across the population.

Sequencing and *de novo* assembly

DNA preparation

Yeast cell cultures were grown overnight at 30° in 20 ml of YPD medium to early stationary phase before cells were harvested by centrifugation. Total genomic DNA was than extracted using the QIAGEN Genomic-tip 100/G according to the manufacturer's instructions. DNA quality was assessed on 1% agarose gel to check DNA integrity, quantified on Qubit and Nanodrop to assess quantity and purity of the sample.

Illumina sequencing

Genomic Illumina sequencing libraries were prepared with a mean insert size of 280 bp and were subjected to paired-end sequencing (2 × 100 bp) on Illumina HiSeq2000 sequencers.

Minion library preparation and sequencing

As MinION technology was launched in 2017 and quickly evolved with a lot of changes both in hardware, software and chemistry, all the project involving this technology had different protocols for library preparation but also different analysis and assembly pipelines.

- **Chapter 3 – Part 1:**

For this project, we used two-dimensional (2D) library preparation with the R7.3 pores as it yielded more accurate results after basecalling. 2 µg of genomic DNA was sheared to ~8,000 bp with g-TUBE. Sequencing libraries were prepared according to the SQK-MAP005-MinION gDNA Sequencing Kit protocol.

- **Chapter 3 – Part 2:**

Genomic DNA was first sheared to ~20,000 bp using g-TUBE. The sequencing of the 95 strains using Oxford Nanopore technology has been performed using R9.4 and R9.4.1 versions of the pores. Similarly two chemistries generations have been used for the library preparation, SQK-LSK108 and SQK-LSK109 respectively, which generated 1D sequences by ligation of adapters. Barcoding of the libraries by ligation of barcodes EXP-NBD113 (for SQK-LSK108) and EXP-NBD114 (for SQK-LSK109) allowed to multiplex up to 12 libraries on the same flowcell.

- **Chapter 3 – Part 3:**

Similarly, strains YAF019 and YAF064 were sequenced using 1D ligation library preparation using the SQK-LSK108 kit. Strains YAF129, YAF140, YAF153, YAF155 and YAF156 were barcoded with EXP-NBD114 and library prepared with SQK-LSK109 kit.

***De novo* genome assembly**

- **Chapter 3 – Part 1:**

For the assembly of the strain UMY321, raw reads were first basecalled with Albacore. Basecalled reads were then trimmed of their adapter using Porechop (https://github.com/rrwick/Porechop). Four different assemblers have been used and compared: ABruijn (v0.3b) (Lin et al., 2016), Canu (v1.1)(Berlin et al., 2015), miniasm (v0.2-r137-dirty) (Li, 2016), and SMARTdenovo (https://github.com/ruanjue/smartdenovo). These assemblers were assessed on different subset of 2D reads (10x, 15x, 20x or 25x). To cope with the high error rate of Oxford Nanopore reads, assemblies were further polished with 100x of Illumina reads using Pilon (v1.18) (Walker et al., 2014). SSPACE-LongRead (v1.1) (Boetzer and Pirovano, 2014) was finally used to scaffold the selected assembly using long-reads information.

- **Chapter 3 – Part 2:**

The 95 strains were assembled using the LRSDAY pipeline (Yue and Liti, 2018). Briefly, raw reads were basecalled and demultiplexed using Guppy (v2.3.5). Selection of the best set of reads was done with Fitlong (https://github.com/rrwick/Filtlong), removing all reads below 1,000 bp and having a mean Q-score (PHRED score) below 10. Final subsample represented a coverage of 40x. Subsampled reads were corrected using Canu (v1.8) (Koren et al., 2017). For the assembly, four different assemblers were benchmarked: Canu (Koren et al., 2017), Flye (Kolmogorov et al., 2019), SMARTdenovo (https://github.com/ruanjue/smartdenovo) and wtdbg2 (Ruan and Li, 2019).

- **Chapter 3 – Part 3:**

Similarly, CRISPR reshuffled strains were assembled using LRSDAY pipeline (Yue and Liti, 2018) with SMARTdenovo as assembler. This was done by Samuel O'Donnell.

Illumina reads mapping

Reads coming from Illumina HiSeq2000 were mapped with BWA (v0.7.4) (Li and Durbin, 2010). GATK (v3.3) (McKenna et al., 2010) was used for calling polymorphic positions as well as for local realignment of the reads around indels.

Assembly completeness evaluation

Assessing the completeness of the *B. bruxellensis* was needed to ensure the quality of the reference. The first parameter checked has been the proportion of unmapped short reads through the use of Samtools (v0.1.19) (Li et al., 2009) using the option "view -f 4 -c". Then, CEGMA (v2.5) (Parra et al., 2007) was run in order to recover and assess the number of ultraconserved eukaryotic genes as they are expected to be present in every eukaryotic assembly.

Whole genome comparison

Whole genome comparisons has been used either to assess assembly quality relative to another one (Chapter 3 part 1) but mostly to unveil gross structural rearrangements (Chapter 3 part 2). These comparisons were performed with the MUMmer suite (v3.0) (Kurtz et al., 2004). nucmer was used to align the sequences (with –maxmatch option). The alignments coordinates were extracted to determine the proportion of non-N residues of each assembly that were covered. The delta files were filtered for alignments <5 kb and plots were generated with mummerplot.

References

Ashman, K.A., Bird, C.M., and Zepf, S.E. (1994). Detecting bimodality in astronomical datasets. Astron. J.

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.Y., Ou, J., San Luis, B.J., Bandyopadhyay, S., et al. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat. Methods.

Behboodian, J. (1970). On the Modes of a Mixture of Two Normal Distributions. Technometrics 12, 131–139.

Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M., and Phillippy, A.M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33, 623–630.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 15, 211.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7.

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

Holzmann, H., and Vollmer, S. (2008). A likelihood ratio test for bimodality in two-component mixtures with application to regional income distribution in the EU. AStA Adv. Stat. Anal. 92, 57–69.

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.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37, 540–546.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 27, 722–736.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. 5, R12.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR : An R Package for Multivariate Analysis. J. Stat. Softw. 25, 1–18.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32, 2103–2110.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M., and Pevzner, P.A. (2016). Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl. Acad. Sci. U. S. A. 113, E8396–E8405.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067.

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.

Pfister, R., Schwarz, K.A., Janczyk, M., Dale, R., and Freeman, J.B. (2013). Good things peak in pairs: A note on the bimodality coefficient. Front. Psychol. 4, 700.

Ruan, J., and Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. BioRxiv 530972.

Seymour, D.K., Chae, E., Grimm, D.G., Martín Pizarro, C., Habring-Müller, A., Vasseur, F., Rakitsch, B., Borgwardt, K.M., Koenig, D., and Weigel, D. (2016). Genetic architecture of nonadditive inheritance in Arabidopsis thaliana hybrids. Proc. Natl. Acad. Sci. U. S. A. 113, E7317–E7326.

Török, T., Rockhold, D., and King, A.D. (1993). Use of electrophoretic karyotyping and DNA-DNA hybridization in yeast identification. Int. J. Food Microbiol.

Wagih, O., and Parts, L. (2014). gitter: A Robust and Accurate Method for Quantification of Colony Sizes From Plate Images. G3; Genes|Genomes|Genetics 4, 547–552.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963.

Wangkumhang, P., Chaichoompu, K., Ngamphiw, C., Ruangrit, U., Chanprasert, J., Assawamakin, A., and Tongsima, S. (2007). WASP: A Web-based Allele-Specific PCR assay designing tool for detecting SNPs and mutations. BMC Genomics 8, 275.

Widmer, C., Lippert, C., Weissbrod, O., Fusi, N., Kadie, C., Davidson, R., Listgarten, J., and Heckerman, D. (2014). Further improvements to linear mixed models for genome-wide association studies. Sci. Rep. 4, 6874.

Yue, J.X., and Liti, G. (2018). Long-read sequencing data analysis for yeasts. Nat. Protoc. 13, 1213–1231.

CONCLUSION & PERSPECTIVES

Studies aiming at building the link between genotype and phenotype often suffer from different limitations. Either they do not focus on a representative sample of the global variation at the population level or fail to explain an important part of this phenotypic variation. Here we wanted to investigate the full breadth of the genetic architecture of traits by combining a classical elegant and powerful crossing scheme with more advanced high-throughput techniques. The concept of the diallel cross is based on a scheme where a given set of individuals is crossed in all pairwise combinations. Consequently, all haplotype combinations are represented. This design has been extensively used by breeders for crops and cattle for decades with the aim to improve agronomic traits such as yield and to further dissect the underlying genetic components (Griffing, 1956). We applied the same technique but combined it with the powerful *S. cerevisiae* model and high-throughput techniques of both phenotyping and genotyping. Overall, this workflow allowed us to expose and decompose several aspects of traits' genetic makeup in a powerful and unbiased manner. Yet, the power of this design can be pushed even further and would allow other important discoveries leading to an even better understanding of how the genetic makeup of an individual can contribute to its phenotype.

The diallel panel as a framework for elucidating the genetic architecture of traits

By performing a diallel cross between a wide variety of natural isolates in chapter 1, corresponding to representatives of different populations representing almost the complete species, we generated a large panel of 3,025 diploids with all possible haplotype combinations. With this panel, we could precisely measure the relative part of phenotypic variation induced by additive as well as non-additive genetic effects. We were able to take advantage of the allele frequencies reshuffling of the alleles that were initially below the threshold of 5% in the initial population. Due to the pairwise crossing, we ended up with an increase of their frequency in the diallel population and retained enough power to perform Genome-Wide Association Studies. Classical GWAS approach would have filtered out these candidates due to their low frequency, thus overriding potential variants that have a substantial contribution to the phenotypic variation in the whole natural population.

As one of the main points of performing a diallel cross was to study the potential causes for missing heritability *e.g.* the low frequency variants and the non-additive effects, this design could also allow to investigate at a species-wide level other putative causes for this missing heritability such as the role of mitochondria in the phenotypic landscape. Here, the advantage of the diallel design lies in the fact that in a full diallel cross, crosses involving the same two parents are present twice and are called reciprocal crosses: theoretically, the genome of the cross between parent X of *MAT*a and parent Y of *MAT*α is the same as parent Y of *MAT*a and parent X of *MAT*α. Although this is true for the nuclear genome, this could differ for the mitochondrial genome. Indeed, as yeast are homoplasmic *i.e.* all copies of the mitochondrial DNA is the same in a cell, nothing says if the reciprocal cross would inherit its mitochondrion from the parent of *MAT*a , the parent of *MAT*α or if recombination between the two occurred (Dujon et al., 1974; Fritsch et al., 2014;

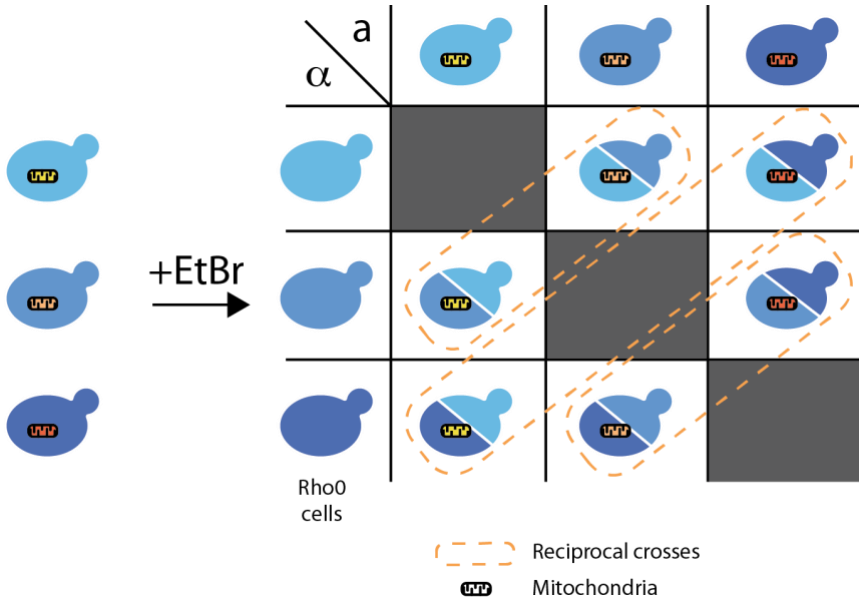


Figure 1. Assessing the mitochondrial effect in reciprocal crosses
All isolates from one of the mating type is grown in presence of ethidium bromide to remove the mitochondria from the cells. Then, a standard diallel cross can be done with the isolates from the opposite mating type still retaining their mitochondria. This results in reciprocal crosses differing only by their parental mitotype.

Leducq et al., 2017). Assessing the effect of the mitotype would require to have reciprocal crosses with one diploid having the mitotype of one parent and the other diploid having the mitotype of the other parent. Then, a phenotypic comparison of both reciprocal crosses differing only by their mitotype could be performed. Yet, selecting for a specific mitotype is challenging. One way to do this selection would be to remove the mitochondria from all parents from the same mating type so that only one mitotype can be passed on the diploid and no recombination event between the two parental mitochondrial DNA could happen (Figure 1) (Paliwal et al., 2014; Wolters et al., 2018).

Diallel offspring panel to assess the genetic complexity and phenotypic expressivity

In chapter 2, we dissected the genetic complexity of traits by using the offspring of 190 hybrids coming from the diallel panel. This allowed to reveal that at the species level, some conditions are controlled by few major effect genes compared to other conditions controlled by a larger number of loci. Interestingly, major effect loci often showed a departure from monogenic inheritance in the offspring indicating an increase of the genetic complexity of some cross/trait combinations. These findings highlight the presence of modifier genes in some genetic backgrounds. In conclusion, our results could assess the pervasive nature of expressivity at a population level with a range of intensity depending on the variant and/or condition considered.

We assumed according to the Mendelian dogma of inheritance that each of the two parental alleles will be transmitted to the progeny with equal probabilities *i.e.* 0.5/05. However, examples of deviation from this dogma exist, as seen in genomic regions that deviate from the expected frequency of parental alleles (0.5/0.5) in F2 progeny. This phenomenon, known as Transmission Ratio Distortion (Dunn and Bennett, 1968 ; TRD), occurs if one of the two parental allele is preferentially passed to its offspring. General mechanisms and concepts of unequal allelic transmission are known, with meiotic drive systems (Sandler and Novitski, 1957), segregation distorters (Charlesworth and Hartl, 1978) and deleterious or lethal genetic interactions (Bateson-Dobzhansky-Muller Incompatibilities ; BDMIs) (Dobzhansky, 1937; Muller, 1942). As BDMIs can lead to reproductive isolation, they can impact the evolution of some populations and are believed to be one of the drivers of speciation. TRDs have been highlighted in a wide range of cases both inter- and intra- species (Hou et al., 2015; Leppälä et al., 2013; Lyon, 2003; Seidel et al., 2008) but few information is available regarding the prevalence, repartition and nature of TRDs at a species-wide level (Seymour et al., 2019). Conducting a population-wide mapping of such events would also allow to understand the impact

of such events as evolutionary forces shaping populations and subpopulations in a species.

Using the offspring coming from a diallel panel has the potential to answer questions in a systematic manner. A crossing scheme as done in chapter 2 would be representative of almost the entire genetic diversity of the species. For each cross, an offspring with a very large number of individuals (~1500 individuals) has to be generated. Mapping TRDs in this context would consist in a regular bulk segregant analysis where all segregants are pooled and sequenced as a bulk. Then, allele frequency is assessed for each discriminating genomic position between the two parental strains. Any deviation from an allele frequency of 0.5 will pinpoint a locus involved in TRD. With the high number of offspring used and the high recombination rate of *S. cerevisiae*, we can obtain a good mapping resolution and resolve TRD to causal loci.

Obtaining a global and unbiased view of the overall population of *S. cerevisiae*

Despite the use of diallel panels, both in a diploid context and with their respective haploid progenies, part of the genetic architecture is still unexplained. An easy but tedious follow up would be to consider a larger number of natural isolates to construct the diallel panel *e.g.* a 100 x 100 diallel panel to encompass even more genetic diversity and be more representative of the natural population. As a large part of the *S. cerevisiae* natural population is constituted of heterozygous diploids (416 out of 694 diploid isolates) (Peter et al., 2018), this would first require to generate monosporic homozygous isolates then delete the *HO* endonuclease in order to generate stable haploid isogenic lines of both mating types. A drawback would then be that we "lose" the heterozygosity present in the natural diploid background, thus the monosporic isolate is no longer representative of the natural isolate and two monosporic isolates from the same heterozygous background will not display the same pattern of genetic variation.

Widening the accessible phenotypic range

Taking advantage of the asexual reproduction of the budding yeast, we can genotype and phenotype a given strain as many times as we want. Going further in deciphering the genotype-phenotype relationship would therefore benefit from investigating other phenotypes. Not only could we try other conditions with different type of stress such as other toxins or changing carbon sources, etc. But more conceptually, not focusing on a growth phenotype could bring more information or at least a different one. For example, phenotyping at the cellular and subcellular level has seen tremendous improvements and interest. Possibilities exist to generate extremely high quantity of data with high-throughput imaging systems that could then be combined to analysis software such as CalMorph (Ohya et al., 2005; Okada et al., 2015) or machine learning algorithms such as neural networks to do image analysis. For example, the CalMorph software is able to estimate 501 morphological parameters based on images of yeast cells whose cell wall, nuclear DNA and actin filaments have been stained with specific dyes. Both the parameters such as cell size, cell roundness, bud shape, nuclei size are measured but also their variance corresponding to a biological phenotypic noise. Moreover, it is possible to work with asynchronous cells as it is possible to categorize them depending on their phase of cellular cycle. Applying this technique to our diallel panel would reflect phenotypes way beyond colony size. Yet, the same methodology of looking for ratio between growth in a specific condition over growth in a standard condition can also be applied with those morphological traits thus increasing again the number of potential traits that we can look at.

Phenotyping goes as far as imagination goes…

References

Charlesworth, B., and Hartl, D.L. (1978). Population dynamics of the segregation distorter polymorphism of Drosophila melanogaster. Genetics 89, 171–192.

Dobzhansky, T. (1937). Genetics and the origin of species (New York: Columbia Univ. Press).

Dujon, B., Slonimski, P.P., and Weill, L. (1974). Mitochondrial genetics. IX. A model for recombination and segregation of mitochondrial genomes in Saccharomyces cerevisiae. Genetics 78, 415–437.

Dunn, L.C., and Bennett, D. (1968). A new case of transmission ratio distortion in the house mouse. Proc. Natl. Acad. Sci. 61, 570–573.

Fritsch, E.S., Chabbert, C.D., Klaus, B., and Steinmetz, L.M. (2014). A genome-wide map of mitochondrial DNA recombination in yeast. Genetics 198, 755–771.

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

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

Leducq, J.B., Henault, M., Charron, G., Nielly-Thibault, L., Terrat, Y., Fiumera, H.L., Shapiro, B.J., and Landry, C.R. (2017). Mitochondrial recombination and introgression during speciation by hybridization. Mol. Biol. Evol. 34, 1947–1959.

Leppälä, J., Bokma, F., and Savolainen, O. (2013). Investigating incipient speciation in Arabidopsis lyrata from patterns of transmission ratio distortion. Genetics.

Lyon, M.F. (2003). Transmission Ratio Distortion in Mice. Annu. Rev. Genet.

Muller, H.J. (1942). Isolating mechanisms, evolution and temperature. Biol. Symp.

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. 102, 19015–19020.

Okada, H., Ohnuki, S., and Ohya, Y. (2015). Quantification of cell, actin, and nuclear DNA morphology with high-throughput microscopy and calmorph. Cold Spring Harb. Protoc.

Paliwal, S., Fiumera, A.C., and Fiumera, H.L. (2014). Mitochondrial-nuclear epistasis contributes to phenotypic variation and coadaptation in natural isolates of Saccharomyces cerevisiae. Genetics 198, 1251–1265.

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.

Sandler, L., and Novitski, E. (1957). Meiotic Drive as an Evolutionary Force. Am. Nat. 91, 105–110.

Seidel, H.S., Rockman, M. V, and Kruglyak, L. (2008). Widespread genetic incompatibility in C. elegans maintained by balancing selection. Science (80-.). 319, 589–594.

Seymour, D.K., Chae, E., Arioz, B.I., Koenig, D., and Weigel, D. (2019). Transmission ratio distortion is frequent in Arabidopsis thaliana controlled crosses. Heredity (Edinb). 122, 294–304.

Wolters, J.F., Charron, G., Gaspary, A., Landry, C.R., Fiumera, A.C., and Fiumera, H.L. (2018). Mitochondrial recombination reveals mito–mito epistasis in yeast. Genetics 209, 307–319.

APPENDICES

List of publications

Fournier, T., Abou-Saada, O., Hou, J., Peter, J., Caudal, E., and Schacherer, J. (2019). Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. (**Elife**, In revision) – Available as a preprint in BioRxiv

Fleiss, A.*, O'Donnell*, S., **Fournier, T.**, Agier, N., Delmas, S., Schacherer, J., Fischer, G. (2019). Reshuffling yeast chromosomes with CRIPSR/Cas9. (**PLoS Genetics**, In press) – Available as a preprint in BioRxiv
*: Authors contributed equally to this work

Fournier, T.*, Gounot, J.-S.*, Freel, K., Cruaud, C., Lemainque, A., Aury, J.-M., Wincker, P., Schacherer, J., and Friedrich, A. (2017). High-Quality de Novo Genome Assembly of the Dekkera bruxellensis Yeast Isolate Using Nanopore MinION Sequencing. **G3** (Bethesda). *7*, g3.300128.2017.
*: Authors contributed equally to this work

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

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., **Fournier, T.**, and Schacherer, J. (2016). Species-wide survey reveals the various flavors of intraspecific reproductive isolation in yeast. **FEMS Yeast Res**. *16*.

List of communications

Yeast Genetics Meeting

Stanford, USA (2018)

Population-scale diallel cross reveals the impact of rare variants on the phenotypic landscape (Oral)

iGenolevures

Paris, France (2017)

Towards a species-wide view of the genetic architecture of traits (Oral)

International Conference on Yeast Genetics and Molecular Biology

Praha, Czech Republic (2017)

Towards a species-wide view of the genetic architecture of traits (Oral and Poster)

Séminaire de Microbiologie de Strasbourg

Strasbourg, France (2017)

Towards a species-wide view of the genetic architecture of traits (Oral)

Levures, Modèles et Outils

Brussels, Belgium (2016)

Unraveling the genetic complexity of traits in natural yeast population. (Oral)

Séminaire de Microbiologie de Strasbourg

Strasbourg, France (2016)

Unraveling the genetic complexity of traits in natural yeast population. (Poster)

Teaching and scientific popularization

OpenLAB

2016-2018

Bringing the lab to high-school pupils.

The goal of this operation is to make them discover the world of research through a two-hours practical work where they learn how to handle lab hardware and perform experiments. We also engaged discussion with them and answered their questions about research but more importantly about their future. This was very fulfilling to share my experience both with pupils and my fellow PhD students also enrolled in this project.