Elucidating the genetic basis of variation in populations by large scale phenomics and genomics

Johan Henning Hallin

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES DE LA VIE ET DE LA SANTÉ

IRCAN

THÈSE DE DOCTORAT

Elucider les facteurs génétiques à l'origine de la variabilité des populations par phénomique et génomique de masse

Johan HALLIN

CNRS UMR7284 - Population genomics and complex traits - Inserm U1081

Présentée en vue de l'obtention du grade de docteur en Sciences **d'**Université Côte d'Azur
Dirigée par : Gianni Liti
Soutenue le : 22 Mars 2018

Devant le jury, composé de :
Bertrand Llorente, Directeur de Recherche,
Le Centre de Cancérologie de Marseille

Daniela Delneri, Proffeseur,
The University of Manchester

Etienne Danchin, Directeur de Recherche,
Institut National de la Recherche Agronomique,

Gianni Liti, Directeur de Recherche,
Institut de Recherche sur le Cancer et le Vieillissement, Nice

Elucider les facteurs génétiques à l'origine de la variabilité des populations par phénomique et génomique de masse

Jury :

Président du jury
Etienne Danchin, Directeur de Recherche, Institut National de la Recherche Agronomique

Rapporteurs
Bertrand Llorente, Directeur de Recherche, Le Centre de Cancérologie de Marseille
Daniela Delneri, Proffeseur, The University of Manchester

Examinateur
Etienne Danchin, Directeur de Recherche, Institut National de la Recherche Agronomique

Résumé

La variabilité phénotypique existante au sein d'une population est d'une importance cruciale ; elle permet l'adaptation à de nouvelles conditions par la sélection naturelle de traits bénéfiques. La variabilité phénotypique est le résultat du polymorphisme génétique de chaque individu, couplé à l'influence de divers facteurs environnementaux. Ces travaux ont pour objectif d'élucider quels sont les facteurs génétiques responsables de la variabilité phénotypique de chaque individu afin de comprendre comment celle-ci évolue de génération en génération et peut s'accentuer au-delà des prédispositions parentales. Finalement, les résultats obtenus seront utilisés pour prédire un phénotype à partir d'un génotype inconnu. Nous avons utilisé des techniques de phénomique et de génomique de haut débit pour décomposer avec une précision inédite la variabilité phénotypique d'une large population de souches diploïdes de *Saccharomyces cerevisiae*. Le génotype exact de plus de 7000 souches uniques a ainsi été obtenu via le croisement et le séquençage de souches haploïdes distinctes. Nous avons mesuré la capacité de croissance de ces souches et identifié les composants génétiques influant sur ce trait. De plus, nous avons identifié des «loci de caractères quantitatifs» additifs et non-additifs, et étudié la fréquence du phénomène d'hétérosis et ses mécanismes. Enfin, en utilisant les données phénotypiques et génotypiques de la même population de levures, nous avons pu prédire les traits de chaque individu avec une presque parfaite exactitude. Ces travaux ont ainsi permis d'identifier avec précision les facteurs génétiques modulant la variation phénotypique d'une population diploïde, et de prédire un trait à partir du génotype et de l'ensemble des données phénotypiques. En plus de ce projet, nous travaillons aussi sur l'identification des bases génétiques à l'origine de la non-viabilité des gamètes, ainsi que sur la compréhension des caractères complexes chez des souches hybrides intra-espèce. De par l'étude de 9000 gamètes séquencés issus de six hybrides différents, nous avons pour objectif de caractériser leur profil de recombinaison et d'observer quelle est l'influence du fond génétique sur ce dernier. De plus, nous avons caractérisé la capacité de croissance de ces gamètes dans neuf conditions environnementales différentes et nous prévoyons de disséquer l'architecture génétique de ces traits dans différents fonds génétiques.

Mots-clés. Caractères quantitatifs, variation génétique, épistasie, hétérosis, prédictions.

Abstract

The phenotypic variation between individuals in a population is of crucial importance. It allows populations to evolve to novel conditions by the natural selection of beneficial traits. Variation in traits can be caused by genetic or environmental factors. This work endeavors to study the genetic factors that underlie phenotypic variation in order to understand how variation can be created from one generation to the next; to know what genetic mechanisms are most prominent; to learn how variation can extend beyond the parents; and finally, to use this in order to predict phenotypes of unknown genetic constellations. We used large scale phenomics and genomics to give an unprecedented decomposition of the phenotypic variation in a large population of diploid *Saccharomyces cerevisiae* strains. Constructing phased outbred lines by large scale crosses of sequenced haploid strains allowed us to infer the genetic makeup of more than 7,000 colonies. We measured the growth of these strains and decomposed the phenotypic variation into its genetic components. In addition, we mapped additive and nonadditive quantitative trait loci, we investigated the occurrence of heterosis and its genetic basis, and using the same populations we used phenotypic and genetic data to predict traits with near perfect accuracy. By using the phased outbred line approach, we succeeded in giving a conclusive account of what genetic factors define phenotypic variation in a diploid population, and in accurately predicting phenotypes from genetic and phenotypic data. Beyond the phased outbred line project, I am currently investigating the genetic basis of gamete inviability and complex traits in intraspecies yeast hybrids. Using 9,000 sequenced gametes from six different hybrids we aim to characterize their recombination landscape and how the genetic background influences it. Furthermore, we have phenotyped these gametes in nine conditions and will dissect the genetic architecture of these traits across multiple genomic backgrounds.

Keywords. Quantitative traits, genetic variation, epistasis, heterosis, predictions

Acknowledgments

I am of the opinion that we are all simply the result of our genome and its interactions with the environment. I was fortunate enough to be born with a genome and into an environment where I could pursue my academic career all the way to the end of this PhD. The list of contributors to both my genetic makeup and environment interactions could be made comically long. However, here I will stick to the most obvious genetic contributors and the more prominent human-human interactions.

First and foremost, I would like to thank Gianni for taking me in in 2014 for what started as a six month internship and ended four years later in this PhD. Thank you for leading me through my PhD, for trusting me to manage large projects, and finally for your open door and open mind.

Jonas, thank you for introducing me to the world of research, and for introducing me to Gianni. But most of all, thank you for your continued mentoring and guidance.

For the work presented in this thesis, I would especially like to thank Leo and Kaspar for the QTLs and predictions, Alex for the variance decomposition and Martin for Scan-o-matic.

Thank you Labex Signalife for the funding (ANR-11-LABX-0028-01).

To all my fantastic team members, I think you know that I would, literally, not have survived here in Nice without you. A special mention has to go to Agnès, who has helped me so much during my years in Nice that it borders on embarrassing. And thank you Ben for helping me with the French parts of this thesis.

I'd also like to thank my friends: the old who tirelessly saves a place for me in Sweden, and the new who have created a home for me in France.

María Isabel Acósta Lopez, thank you for your unwavering love and support. And, equally important, thank you for challenging me, complementing me and for somehow being exactly what I need regardless of what it might be and whether I know it or not.

To my family, min familj som jag älskar, min familj som jag litar på, min familj som jag stödjer mig mot, min familj som inspirerar mig, min familj som jag saknar, min familj som jag har saknat och min familj som jag kommer sakna tills jag får vett nog att komma hem igen.

Contents

Figures

Foreword

It is easy to imagine how the great variation on earth has long been appreciated by mankind. From the vast amount of species of plants, each with a unique flower, to all animals and insects. From mammal to microbe, there is scarcely an inch on this earth not inhabited, from smoldering volcanoes to the icy plains of Greenland.

Variation between species is all well and good, but the variation within a species was not always held in such a high regard as it is today. In Plato's dialogs, Socrates argued that there was but one true form and the variation around it was but cheap and ill-fated attempts to capture the true form. However, in the days of Charles Darwin and his contemporaries, variation was given a clear function.

The study of variation has increased in popularity over the years. In fact, a quick search on Pubmed reveals that the fraction of publications mentioning variation in either the title or abstract has been going up steadily since the 70's. (Fig. 1). This does, however, refer to variation in general, not exclusively to the study of variation in genotypes and phenotypes. Nevertheless, variation seems to be gaining recognition in the scientific community.

In this thesis, the first three chapters will give you a general introduction to the field of quantitative genetics, a powerful tool to study the link between phenotypic and genotypic variation. I will cover the basics of inheritance of complex traits, how we can find loci in the genome that contribute to the variation in a population, and how to predict the phenotypes of individuals. We will also look at how phenotypes are modulated by genetic mechanisms, and how they can give phenotypes that are more extreme from one generation to the next.

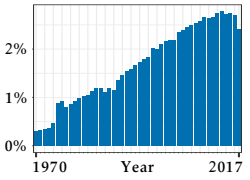


Figure 1. Propagation of variation in publications. The mention of variation has been going up steadily over the years

Accompanying this thesis are two articles both published in 2016 in *Nature Communications*. These are the main articles of this thesis and constitute the main body of my work. Each of them contain their own introduction, putting the work into a more specific context than I do in the first three chapters. After this there is a chapter on the work I am currently doing and have been doing since my articles were published. This is an ongoing project and mostly discusses methodology as well as some preliminary results. Lastly, I extend the discussion from the articles, adding some thoughts about the mapping population and heterosis analysis.

But before we get started, a quote from Alfred Russel Wallace's "On the Law Which Has Regulated the Introduction of New Species", worthy of thought when we are quick to divide and slow to unite.

The great gaps that exist between fishes, reptiles, birds, and mammals would then, no doubt, be softened by intermediate groups, and the whole organic world would be seen to be an unbroken and harmonious system.

Chapter 1

Genotype to phenotype

This chapter will introduce the concept of phenotypic and genetic variation, and how they are integrally related in a complex way. The complexity of this relationship is what my thesis tries to unravel. I will discuss how the genome can influence, as well as predict, phenotypes.

Variation is a key feature in natural populations, it gives natural selection something to act on, making it completely essential for life as we know it to have evolved. Although Charles Darwin (Fig. 1.1) makes no mention of it in the title of his iconic book *On the origin of species - by means of natural selection*, Alfred Russel Wallace (Fig. 1.2), independent co-discoverer of evolution by natural selection, gave credit to variation in naming his paper *On the tendency of varieties to depart indefinitely from the original type*. Understanding the genetic basis of complex phenotypic variation has been – and still is– a goal of the natural sciences. Knowledge of this relationship will aid in, for example, predicting disease risk and breeding desirable traits in crops (Mackay et al., 2009). It has, however, been difficult to assess the genetic contributions to variation.

1.1 Heritability

Variation in a phenotypic trait can have different sources, which constitutes one of the difficulties in knowing the genetic contribution. The amount of a population's variation in a trait that can be attributed to the genome is called **heritability**. Heritability is important, for example, when you want to select for a specific trait; if you want to select (through breeding) your cows to have more milk production, you would first want to know how much of the variation in the milk production is due to differences in the genome. If most of the variation is due to the environment, your efforts would be of more use optimizing that.

The heritability of a trait is not static, it can vary over time and place (Klug et al., 2009). If the environmental variation for a population is low, then the genome will have a stronger relative effect. The total phenotypic variance in a population (V_t) can be expressed as the environmental variance (V_e) and the genetic variance (V_g).

$$V_t = V_e + V_g$$

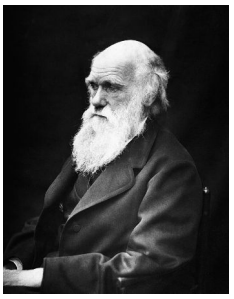


Figure 1.1. Charles Robert Darwin, author of the groundbreaking book *On the origin of species - by means of natural selection*, laying the foundation for modern biology. Image source.

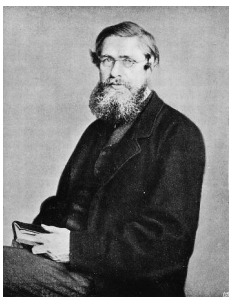


Figure 1.2. Alfred Russel Wallace, independent co-discoverer of evolution by natural selection. Image source.

Heritability. The fraction of phenotypic variance in a population that can be attributed to genetic variation.

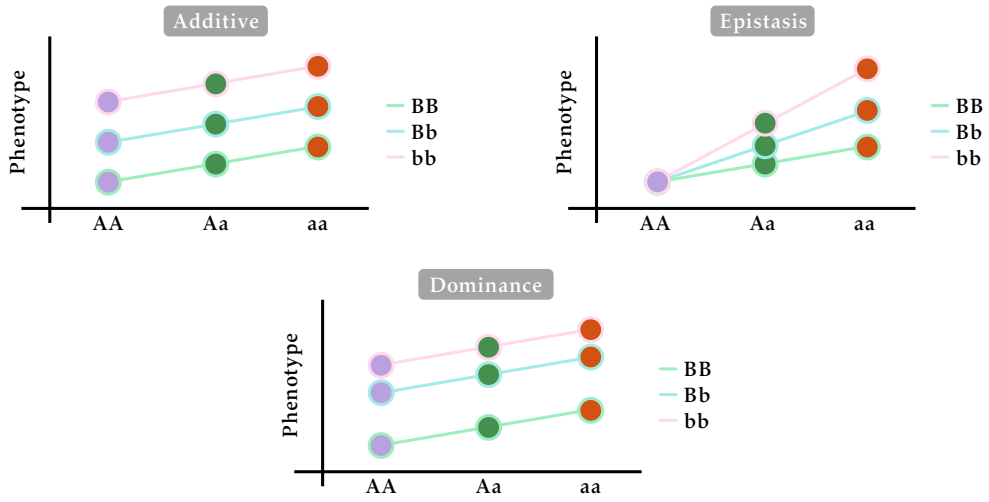


Figure 1.3. Components of genetic variance. This schematic shows the concept behind the different kinds of genetic components that make up the total genetic contribution (V_g) to the phenotypic variance of a population (V_t). **Additive** effects are the fixed effects that alleles contribute with, which are independent of the allele compositions at other loci. I.e. for a completely additive trait, the heterozygote (Aa or Bb) will have a phenotypic effect that equals the mean of the two homozygotes (AA and aa, or BB and bb), and the effect of either locus is independent of the genotype at the other locus. **Epistasis** or an epistatic interaction is when the effect of a locus is dependent on the genotype at a second locus. The figure represents the most simple interaction containing only two loci, where the effect of the a allele is enhanced with increasing numbers of the b allele at another locus. Finally, **Dominant** effects are the deviations from the additive within a locus, such that the heterozygote does not equal the mean of the two homozygotes. In this example the B locus has a dominant effect while the A locus is completely additive.

Monozygotic. Twins that spawn from the same zygote.

Dizygotic. Twins that spawn from different zygotes.

The genetic contribution to the phenotypic variation is called broad sense heritability, and can similarly be decomposed, such that:

$$V_g = V_{g,a} + V_{g,d} + V_{g,i}$$

Complex traits. An observable trait that has two or more genes modulating it.

where $V_{g,a}$ is additive variance, $V_{g,d}$ is dominant variance and $V_{g,i}$ is interaction variance. These concepts are explained in more detail in figure 1.3. The contribution of the additive component to phenotypic variation is called narrow sense heritability.

In humans, narrow sense heritability is commonly estimated by twin studies. In a classical twin study, comparing the phenotypic similarity of **monozygotic** twins (MZ) and **dizygotic** (DZ) twins gives an estimate of the heritability of that trait. The heritability is calculated as twice the difference between the correlation of the MZ twins and the DZ twins (Boomsma et al., 2002). E.g. if the correlation of MZ twins for a given trait is 0.6, and 0.3 for the DZ twins, then the heritability would be $(2(0.6 - 0.3))$ 60%.

1.2 Complex traits

Complex traits (or quantitative traits) lie at the heart of quantitative genetics, the field of genetics which is concerned with explaining the genetic background to varia-

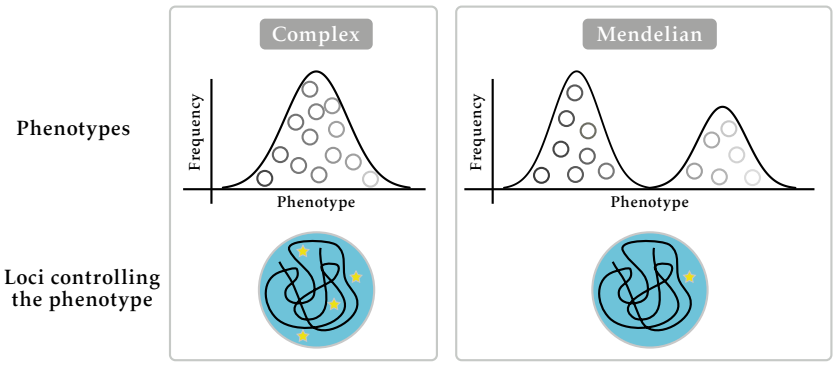


Figure 1.4. Mendelian and complex traits. A complex trait is controlled by several loci (★), creating a continuous distribution of phenotypes in the population. In contrast, a Mendelian trait is controlled by one gene, and because of this the phenotypic distribution of the population will be bimodal, given that there are two variants of this gene. This occurs in diploid populations when one variant is dominant over the other. I.e. even though there are four different possible genotypes at the given locus, there are only two possible phenotypes.

tion in traits. A complex trait is any observable trait that has a large variation within individuals of a population and is modulated by two or more genes, giving the trait a continuous distribution. A classical example of a complex trait is height in humans. A trait (or **phenotype**) that, by the way, has a quite high heritability (around 80% (Silventoinen et al., 2003)).

This is in contrast to Mendelian traits (or monogenic traits) (Fig. 1.4), where the phenotype is modulated by a single gene. However, monogenic traits may not be as simple as they seem. Sirr et al. (2015) find that even a seemingly monogenic trait can have genetic and/or nongenetic modifiers.

Any given complex trait is modulated by many genes in intricate networks and heritability can answer what portion of variation in the complex trait is defined by different genetic components, but we also want to know what specific sites in the genome affect the phenotype.

1.3 Quantitative trait loci

Phenotype. The phenotype of an organism is its collected set of traits, however, it is often used synonymously with trait, and will be used as such throughout this thesis.

A **Quantitative Trait Locus (QTL)** is, as its name suggests, a place in the genome that contributes quantitatively to a particular trait. The field of genetics has come a long way in locating these loci thanks to the fact that factors controlling phenotypes (genes) co-segregate with the phenotypes. QTLs can be located using two different methods, *i*) association-based mapping, or *ii*) linkage-based mapping, which will be discussed in section 1.3.2 and section 1.3.3

QTL. A locus in the genome that contributes to the variation of a trait

The power and resolution with which we locate QTLs are highly dependent on the sample size of your mapping population, and on the amount of recombination that you have between the individuals of the population.

1.3.1 Recombination

QTL mapping is wholly dependent on recombination. Recombination was pro-

Crossover. Reciprocal exchange of genetic material between the two homologous chromosomes (Whitby, 2005).

Non-crossover. nonreciprocal short length exchange of genetic material between the two homologous chromosomes (Whitby, 2005).

Marker. An identifiable position in the genome that differs between the parents of a cross. This can be genotyped to know which parent contributed with the stretch of DNA that covers the marker.

GWAS. A method to find locations in the genome that contribute to the variation of a trait in natural populations.

posed by Morgan (1911), to explain the mystery of some traits being coupled and others segregating randomly. His student, Sturtevant (1913) went on to create the first ever genetic map, using the theory laid down by Morgan (1911).

Recombination is initiated by double-strand breaks during prophase of meiosis I (Keeney et al., 1997), these breaks can subsequently be repaired by using the sister chromatid or homologous chromosome as a template. The recombination results in gene conversion associated to either a **crossover** or a **non-crossover** (Whitby, 2005).

During meiosis, there is a bias for using the homologous chromosome for repairing the double-strand breaks (Haber et al., 1984). Repairing double-strand breaks during mitosis, however, is biased to using the sister chromatid as a template which does not result in any change of genetic material since the two sister chromatids are identical (Fabre et al., 1984; Kadyk and Hartwell, 1992). The different bias in mitotic and meiotic recombination could be explained by the use of different recombination pathways (Schwacha and Kleckner, 1997).

The recombination landscape of *S. cerevisiae* was described in great detail by Mancera et al. (2008), mapping both crossovers and non-crossovers genome wide. They genotyped the four haploid spores from 51 meioses resulting from the sporulation of a hybrid between S288C (a lab strain) and YJM789 (derived from a clinical isolate (Wei et al., 2007)). By using ~52,000 **markers** they could give a detailed view of crossovers and non-crossovers since every event would likely be covered by several markers (median marker distance, 78bp).

They find on average 90.5 crossovers and 46.2 non-crossovers, however, they estimate that they missed ~30% of the non-crossovers and increase the number to 66.1.

Mancera et al. (2008) defined recombination hotspots as regions involved in more recombination events than expected by chance. 179 such regions were found, and corroborating previous studies, 84% of them overlap a promotor. Promotor regions are known to host most of the double-strand breaks during meiosis (Baudat and Nicolas, 1997; Gerton et al., 2000), and correlate well with recombination events even between different strains (Buhler et al., 2007; Mancera et al., 2008).

QTL mapping makes use of recombination to break the linkage between markers and loci in the genome that contribute to the variation of a trait.

1.3.2 Association-based mapping

An association-based QTL mapping experiment makes use of recombination that has occurred through-out history. It is used for natural populations and is generally called **Genome Wide Assocation Studies (GWAS)**. GWAS takes advantage of historical recombination events within a population. Due to the large amount of recombination events that have occurred in a natural population through-out evolution, the genome has been shuffled to a very high

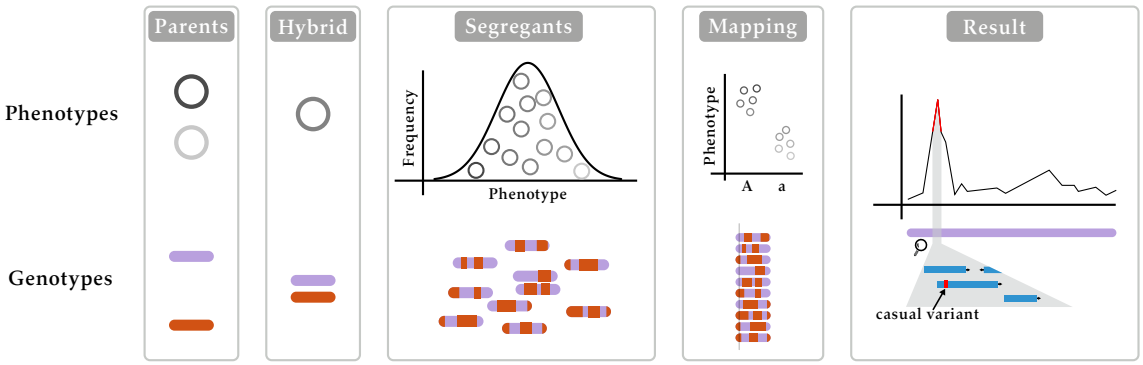


Figure 1.5. QTL mapping in practice. Linkage-based QTL mapping starts with a cross between two (or more) parents, creating a hybrid that has a phenotype which is (usually) intermediate of the two parents. In the case of yeast, the hybrid is sporulated and haploid segregants are isolated. These segregants are phenotyped and genotyped, once this is done the QTL mapping can start by using the genetic markers in the genome and sorting the segregants' phenotypes according to their genotype at the given marker. This is done at every marker in the genome to create a QTL map where some regions of the genome give a significant signal, meaning that those regions have an effect on the phenotype that reaches above the noise. These regions can then be further investigated to find the casual variant(s).

degree, unlinking all but the closest markers from the **causal** locus. Thanks to this, GWAS can locate causal loci with high precision (Mackay et al., 2009).

The association-based QTL mapping does, however, suffer from a few limitations. One being that GWAS experiments have low power to detect rare variants that have an effect on the phenotype (Visscher et al., 2012). And in the context of *S. cerevisiae*, GWAS studies will be severely hampered by the strong population structure (Liti and Louis, 2012; Strope et al., 2015).

Population structure results in spurious associations of variants to phenotypes due to stratification of the mapping population (Hamer and Sirota, 2000). In other words, if within your mapping population there are subpopulations, these subpopulations may differ in their allele frequencies and also, coincidentally, differ in their phenotype levels. This means that variants not actually contributing to the variation in a phenotype can be associated to it (Hamer and Sirota, 2000; Marchini et al., 2004). Linkage-based QTL mapping generally does not have this caveat, since they are based on experimental crosses. However, in (Hallin et al., 2016) we experience population structure due to the crossing scheme of the phased outbred lines (see chapter 4).

1.3.3 Linkage-based mapping

A linkage-based QTL mapping experiment will start with a cross (Fig. 1.5). It uses the same underlying theory as association-based mapping, but since there is no historical recombination in the mapping population (since it does not use natural populations), it has to be created through the crossing. You choose two parents that differ in your phenotype of interest (e.g. pollen

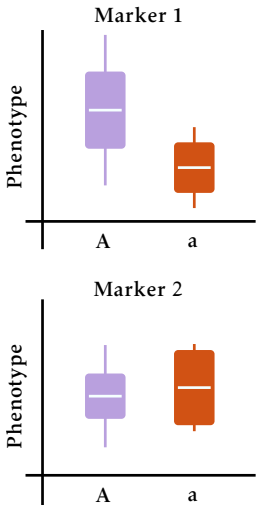


Figure 1.6. QTL mapping and significance testing. Significance tests between the two populations, one for each genotype, distinguishes between markers with (marker 1) and without (marker 2) association to a causative locus.

Causal. Refers to something that gives the actual effect, for example, causal locus, causal SNP or casual marker.

shape in the sweet pea) and that contain differences in the genome. You cross them together and use their progeny to locate regions in the genome that contribute to the differences between the progeny phenotypes. This is the type of mapping that I will be focusing on through-out the thesis, and I will from here on use the term linkage-based mapping interchangeably with QTL mapping.

Since linkage-based QTL mapping constructs its own mapping population, classically from a two-parent cross, it does not suffer from the problem of rare variants. All alleles are expected to be at a 50% frequency, and can thus be detected, even though they may represent a low frequency allele in the natural variation of the species as a whole (Parts, 2014). Additionally, using model organisms to construct the mapping population means that the phenotypes can be measured under controlled conditions with little environmental variation confounding the results. However, the mapping population will not contain as many recombination events, and so the resolution of the mapping suffers.

The statistical methods used in linkage-based mapping to find QTLs can vary in complexity. The most simple one, and the one used in the paper Powerful decomposition of complex traits in a diploid model, is the marker regression method. The simplicity of this method lies in the fact that it only uses the positions in the genome where you have marker data (Broman and Sen, 2009). At each marker it sorts the phenotypes of your samples depending on their genotype at the marker, as in figure 1.6. QTL mapping is a constantly evolving technique, for example, a recent development has made use of the Crispr Cas9 system but it was more than one hundred years ago that the theoretical foundation of QTL mapping was laid.

In 1904, Bateson, Saunders and Punnett (Bateson et al., 1904) publish their findings from experiments in the sweet pea (*Lathyrus odoratus*). They find deviations from expected Mendelian segregation of traits, and they propose that the factors controlling the two phenotypes they are investigating (pollen shape and color) are coupled. They write: *"There is, therefore, some coupling of pollen shape and colours"*. The nature of this coupling would remain unknown until 1911 when Thomas Hunt Morgan suggests that the factors (or genes) controlling traits are physically located on chromosomes (Morgan, 1911). With this, 32 years after Walther Flemming had discovered the chromosome (Flemming, 1878; Paweletz, 2001), there was no question as to the function of chromosomes; propagating genes to the next generation. A theory that had been outlined by Theodor Boveri and Walter Sutton in the *chromosome theory of inheritance* a few years earlier (Sutton, 1903). Other accounts of non-Mendelian segregation of traits are later attributed to this linkage between factors that control certain traits (Yuzo, 1915; Sax, 1923).

Capitalizing on the wealth of knowledge that had been built up, Andrew Paterson and his colleagues use Restriction Fragment Length Polymorphisms (RFLPs) to get

a map of the genome and locate at least 15 QTLs for three phenotypes in an interspecific backcross of tomato (Paterson et al., 1988). By doing this, they set the stage for QTL mapping with the entire genome covered by markers.

In order to locate the specific regions in the genome that are causally affecting the phenotype you need markers that are close enough to the causal locus so that they co-segregate. A dense grid of markers over the entire genome will increase the likelihood that you cover the area with the causal locus. Whole genome sequencing allows you to use all the **Single Nucleotide Polymorphisms (SNPs)** between the two parents you have chosen, if these parents are sufficiently genetically diverged you will end up with a distribution of markers over the entire genome (Bloom et al., 2013; Hallin et al., 2016).

A dense grid of markers must be complemented with a large sample size in order to detect weak effect loci (Bloom et al., 2013). The larger the sample size of your segregating population, the more power you will have to detect loci that do not have a very big effect on the phenotype you are investigating. Steps have been taken to increase the power of studies without necessarily increasing the genotyping and laboring cost; such as using bulk segregant analysis coupled with experimental evolution, where selection pressure is inflicted on a large pool of segregant strains and the changes in allele frequencies are measured to find regions that contribute to the adaptation of the pool (Ehrenreich et al., 2010; Parts et al., 2011); or constructing large cross grids where the parents are sequenced and the progeny mapping populations genotypes are inferred from the parents (Threadgill et al., 2002; Zou et al., 2005; Tsaih et al., 2005; Hallin et al., 2016).

SNP. A nucleotide position in the genome that differs between two given individuals.

1.3.4 Sample size

A limitation common for both association- and linkage-based QTL mapping is the sample size. In order for GWAS to find small effect loci, they continuously increase their sample size, doing meta-analyses creating ever growing mapping populations. For the classical trait of human height, sample size started out at between 10,000 to 20,000 individuals in 2008 (Sanna et al., 2008; Lettre et al., 2008). In 2010, a meta-analysis increased this number to 183,727 (Lango Allen et al., 2010), and the sample size race culminated with a staggering sample size of 253,288 individuals in 2014 (Wood et al., 2014).

In linkage-based QTL mapping, the importance of sample size was efficiently shown by Bloom et al. (2013), where increasing the sample size from 100 to 1,005 increased the amount of QTLs from two to fifteen. The fifteen QTLs found increased the amount of narrow-sense heritability from 21% to 78%, showing that sample size can account for missing heritability (discussed further in section 1.5).

QTL mapping concerns itself with finding the genotypes that are linked to a specific phenotype, but how about finding the phenotype that is linked to a specific genotype?

1.4 Predicting phenotypes from genotypes

For a thought-provoking prediction experiment, read Lippert et al. (2017) using genome-wide data to place an individual within the top ten candidates from a 100 person cohort with 88% accuracy. I.e., rather accurately using the genome to identify (or at least narrow down) individuals.

A goal for biology and medicine is to be able to predict the phenotype of an individual given his or her genetic makeup. A natural start to this is to locate the most important regions of the genome, as with the QTL mapping. In humans, linkage-based QTL mapping is not done due to ethical and practical limitations. What can be done are Genome Wide Association Studies.

The simple prediction of looking at ones parents vs. the genomic method was tested in a clever study by Aulchenko et al. (2009), where Sir. Francis Galton (Galton, 1886) came out on top.

A large number of phenotypes have had an even larger number of loci associated to them using GWAS. And have yielded important insight into the human biology and diseases (Visscher et al., 2012). However, in explaining the phenotypic variation of a population, GWAS often comes up short. If we move back the height in humans. The huge study from 2014 (Wood et al., 2014) used data from over 250,000 individuals and identified a staggering 697 variants in the genome that were significantly associated to height. However, these almost 700 variants only explain 16% of the heritability.

1.5 Missing heritability

The fact that detected variants have only been able to explain a very small amount of the total genetically determined variation has been called the **missing heritability** problem (Maher, 2008). For example, the variants that have been detected for human height do no more in predicting your height than glancing at your parents does (paraphrasing from Joel Hirschhorn in (Maher, 2008)). This missing heritability has been elusive and many different –non-mutually exclusive– explanations have been suggested (Maher, 2008; Manolio et al., 2009; Zuk et al., 2012).

The large lack of heritability explained by loci that have been found to have a detectable effect in height (a very well studied complex trait) highlights the difficulty in using these GWAS results to predict phenotypes. An approach that holds more promise is to use all the genetic information available for the population, not only the significantly causal loci. Being able to predict traits without prior knowledge of causal loci can revolutionize many aspects of biology (Ober et al., 2012). In the study by Ober et al. they perform the first attempt at predicting phenotypes using whole-genome-sequencing data. Although their predictive power was rather weak, they make a case for using whole-genome information rather than causal variants. This opinion has been enforced by Makowsky et al. (2011) who evaluated the ability of whole-genome data to aid in predictions, and similarly find that it can indeed improve them. They do, however, show that increasing the training sample or including related individuals may be a better way of improving the predictive ability.

Before this, Yang et al. (2010) try to explain the missing heritability in human height by using a large set of SNPs, and find that 45% of the variation in height can be attributed to these almost 300,000 variants. They go on to conclude that the missing heritability

can be explained by variants that have effects too small to be significantly detected, and by incomplete linkage disequilibrium between the loci with an effect and the loci that they have genotyped. The rationale here is that, if there is an incomplete linkage between the genotyped locus and the locus giving the effect, then a variant of the genotyped locus can be associated with many variants of the causal locus, diluting the effect (Visscher et al., 2010).

All in all, it is not clear to what extent phenotypes can be predicted, or where exactly the problem lies. Luckily, the theoretical limits of predicting traits can be tested using model systems where all variation not arising from the underlying genetic makeup can be controlled (Märtens et al., 2016).

References

Aulchenko Yurii S, Struchalin Maksim V, et al. **Predicting human height by Victorian and genomic methods.** *European journal of human genetics : EJHG*, 17(8):1070–1075, 2009.

Bateson William, Punnett Reginald, and Saunders Edith. **Experimental studies in the physiology of heredity**. *Reports to the evolution committee of the Royal Society*, pages 1–154, 1904.

Baudat F and Nicolas A. **Clustering of meiotic double-strand breaks on yeast chromosome III.** *Proceedings of the National Academy of Sciences of the United States of America*, 94(10):5213–5218, 1997.

Bloom Joshua S, Ehrenreich Ian M, Loo Wesley T, Lite Thúy-Lan Võ, and Kruglyak Leonid. **Finding the sources of missing heritability in a yeast cross.** *Nature*, 494(7436):234–237, 2013.

Boomsma Dorret, Busjahn Andreas, and Peltonen Leena. **Classical twin studies and beyond.** *Nature reviews. Genetics*, 3(11):872–882, 2002.

Broman Karl W and Sen Saunak. ***A Guide to QTL Mapping with R/qtl***. Statistics for Biology and Health. Springer New York, New York, NY, 2009.

Buhler Cyril, Borde Valérie, and Lichten Michael. **Mapping meiotic single-strand DNA reveals a new landscape of DNA double-strand breaks in Saccharomyces cerevisiae.** *PLoS biology*, 5(12):e324, 2007.

Ehrenreich Ian M, Torabi Noorossadat, et al. **Dissection of genetically complex traits with extremely large pools of yeast segregants.** *Nature*, 464(7291):1039–1042, 2010.

Fabre F, Boulet A, and Roman H. **Gene conversion at different points in the mitotic cycle of Saccharomyces cerevisiae.** *Molecular & general genetics : MGG*, 195(1-2):139–143, 1984.

Flemming Walther. **Zur Kenntnis der Zelle und ihrer Teilung-Erscheinungen**. *Schriften des naturwissenschaftlichen vereins für Schleswig-Holstein*, pages 23–27, 1878.

Galton Francis. **Regression Towards Mediocrity in Hereditary Stature.** *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246, 1886.

Gerton J L, DeRisi J, et al. **Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae.** *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11383–11390, 2000.

Haber J E, Thorburn P C, and Rogers D. **Meiotic and mitotic behavior of dicentric chromosomes in Saccharomyces cerevisiae.** *Genetics*, 106(2):185–205, 1984.

Hallin Johan, Märtens Kaspar, et al. **Powerful decomposition of complex traits in a diploid model.** *Nature Communications*, 7:13311, 2016.

Hamer D and Sirota L. **Beware the chopsticks gene**. *Molecular Psychiatry*, pages 1–3, 2000.

Kadyk L C and Hartwell L H. **Sister chromatids are preferred over homologs**

as substrates for recombinational repair in Saccharomyces cerevisiae. *Genetics*, 132(2):387–402, 1992.

Keeney Scott, Giroux Craig N, and Kleckner Nancy. **Meiosis-Specific DNA Double-Strand Breaks Are Catalyzed by Spo11, a Member of a Widely Conserved Protein Family**. *Cell*, 88(3):375–384, 1997.

Klug William S, Cummings Michael R, Spencer Charlotte A, and Palladino Michael A. ***Essentials of genetics***. Pearson, 7 edition, 2009.

Lango Allen Hana, Estrada Karol, et al. **Hundreds of variants clustered in genomic loci and biological pathways affect human height**. *Nature*, 467(7317):832–838, 2010.

Lettre Guillaume, Jackson Anne U, et al. **Identification of ten loci associated with height highlights new biological pathways in human growth**. *Nature genetics*, 40(5):584–591, 2008.

Lippert Christoph, Sabatini Riccardo, et al. **Identification of individuals by trait prediction using whole-genome sequencing data.** *Proceedings of the National Academy of Sciences of the United States of America*, 114(38):10166–10171, 2017.

Liti Gianni and Louis Edward J. **Advances in quantitative trait analysis in yeast.** *PLoS Genetics*, 8(8):e1002912, 2012.

Mackay Trudy F C, Stone Eric A, and Ayroles Julien F. **The genetics of quantitative traits: challenges and prospects.** *Nature Reviews Genetics*, 10(8):565–577, 2009.

Maher Brendan. **Personal genomes: The case of the missing heritability.** *Nature*, 456(7218):18–21, 2008.

Makowsky Robert, Pajewski Nicholas M, et al. **Beyond missing heritability: prediction of complex traits.** *PLoS Genetics*, 7(4):e1002051, 2011.

Mancera Eugenio, Bourgon Richard, Brozzi Alessandro, Huber Wolfgang, and Steinmetz Lars M. **High-resolution mapping of meiotic crossovers and non-crossovers in yeast**. *Nature*, 454(7203):479–485, 2008.

Manolio Teri A, Collins Francis S, et al. **Finding the missing heritability of complex diseases.** *Nature*, 461(7265):747–753, 2009.

Marchini Jonathan, Cardon Lon R, Phillips Michael S, and Donnelly Peter. **The effects of human population structure on large genetic association studies**. *Nature genetics*, 36:512 EP –, 2004.

Märtens Kaspar, Hallin Johan, Warringer Jonas, Liti Gianni, and Parts Leopold. **Predicting quantitative traits from genome and phenome with near perfect accuracy.** *Nature Communications*, 7:11512, 2016.

Morgan Thomas Hunt. **Random segregation versus coupling in Mendalian inheritance**. *Science (New York, N.Y.)*, pages 1–1, 1911.

Ober Ulrike, Ayroles Julien F, et al. **Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster.** *PLoS Genetics*, 8(5):e1002685, 2012.

Parts Leopold. **Genome-wide mapping of cellular traits using yeast.** *Yeast (Chichester, England)*, 31(6):197–205, 2014.

Parts Leopold, Cubillos Francisco A, et al. **Revealing the genetic structure of a trait by sequencing a population under selection.** *Genome research*, 21(7):1131–1138, 2011.

Paterson A H, Lander E S, et al. **Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms.** *Nature*, 335(6192):721–726, 1988.

Paweletz Neidhard. **Walther Flemming: pioneer of mitosis research**. *Nature reviews molecular cell biology*, 2(1):72–75, 2001.

Sanna Serena, Jackson Anne U, et al. **Common variants in the GDF5-UQCC region are associated with variation in human height**. *Nature genetics*, 40(2):198–203, 2008.

Sax K. **The Association of Size Differences with Seed-Coat Pattern and Pigmentation in PHASEOLUS VULGARIS.** *Genetics*, 8(6):552–560, 1923.

Schwacha Anthony and Kleckner Nancy. **Interhomolog Bias during Meiotic Recombination: Meiotic Functions Promote a Highly Differentiated Interhomolog-Only Pathway**. *Cell*, 90(6):1123–1135, 1997.

Silventoinen Karri, Sammalisto Sampo, et al. **Heritability of adult body height: a comparative study of twin cohorts in eight countries.** *Twin research : the official journal of the International Society for Twin Studies*, 6(5):399–408, 2003.

Sirr Amy, Cromie Gareth A, et al. **Allelic variation, aneuploidy, and nongenetic mechanisms suppress a monogenic trait in yeast.** *Genetics*, 199(1):247–262, 2015.

Strope Pooja K, Skelly Daniel A, et al. **The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen.** *Genome research*, 25(5):762–774, 2015.

Sturtevant Alfred Henry. **the linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association**. *Journal of Experimental Zoology*, 14:43–59, 1913.

Sutton Walter. **The chromosomes in heredity**. *Biological Bulletin*, (4):231–251, 1903.

Threadgill David W, Hunter Kent W, and Williams Robert W. **Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort.** *Mammalian genome : official journal of the International Mammalian Genome Society*, 13(4):175–178, 2002.

Tsaih Shirng-Wern, Lu Lu, Airey David C, Williams Robert W, and Churchill Gary A. **Quantitative trait mapping in a diallel cross of recombinant inbred lines.** *Mammalian genome : official journal of the International Mammalian Genome Society*, 16(5):344–355, 2005.

Visscher Peter M, Brown Matthew A, McCarthy Mark I, and Yang Jian. **Five years of GWAS discovery.** *American journal of human genetics*, 90(1):7–24, 2012.

Visscher Peter M, Yang Jian, and Goddard Michael E. **A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010).** *Twin research and human genetics : the official journal of the International Society for Twin Studies*, 13(6):517–524, 2010.

Wei Wu, McCusker John H, et al. **Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789.** *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12825–12830, 2007.

Whitby M C. **Making crossovers during meiosis.** *Biochemical Society transactions*, 33(Pt 6):1451–1455, 2005.

Wood Andrew R, Esko Tonu, et al. **Defining the role of common variation in the genomic and biological architecture of adult human height.** *Nature genetics*, 46(11):1173–1186, 2014.

Yang Jian, Benyamin Beben, et al. **Common SNPs explain a large proportion of the heritability for human height.** *Nature genetics*, 42(7):565–569, 2010.

Yuzo Hoshino. **On the inheritance of the flowering time in peas and in rice**. *The journal of the college of agriculture, Tohoku Imperical University, Sapporo, Japan*, pages 1–76, 1915.

Zou Fei, Gelfond Jonathan A L, et al. **Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations.** *Genetics*, 170(3):1299–1311, 2005.

Zuk Or, Hechter Eliana, Sunyaev Shamil R, and Lander Eric S. **The mystery of missing heritability: Genetic interactions create phantom heritability.** *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–1198, 2012.

Chapter 2

Hybrids and heterosis

Hybrids are the result of crossing any two individuals. Hybridization is a vital mechanism in the biological world, it creates variation by combining alleles in configurations that have not been seen before. A hybrid might get the best of both parents, or the worst, or will perhaps come out as a perfect midpoint in the continuous phenotypic distribution that lies between the two parents (Fig. 2.1).

This chapter will give an account of the phenotypes of hybrids and how they relate to their parents. The concept of heterosis (hybrid vigor) will be discussed as well as what possible mechanisms could cause this.

2.1 The use of hybrids

Perhaps the most famous hybrid is the mule. A cross between a female horse and male donkey, the mule has been known for centuries for its longevity and ability to work on less food (although these traits may have been overrated (Garrett, 1990)). Nevertheless it is a hybrid with a lot of character and history to go with it.

It is not only the mule that is an important agricultural hybrid, the use of maize hybrids between inbred lines went up from 10 to 90% between 1935 and 1939 in Iowa, USA. The increase in yield and uniformity of the plants that came with using hybrids led them to represent the bulk of the maize produced in the USA by 1950 (Crow, 1998).

Hybridization can also be a force in speciation. Leducq et al. (2016) finds an example of a hybrid between two lineages of ***Saccharomyces paradoxus***. This hybrid had a mosaic genome composed of mostly one parent with interspersed islands of the other. It is found in the contact zone of the two parents and exhibits intermediate phenotypes, as well as partial reproductive isolation. Leducq et al. (2016) hypothesize that the two parentals had come in contact when the glacial ice retreated approximately 10,000 years ago which is when the hybridization would have taken place.

The three examples above highlight a very interesting aspect of hybrids: their ability to outperform, or at least perform differently than, their parents. This can be to the benefit of humans, as in the example of increased yield of maize. As for the

S. paradoxus. *S. cerevisiae's* closest wild relative.

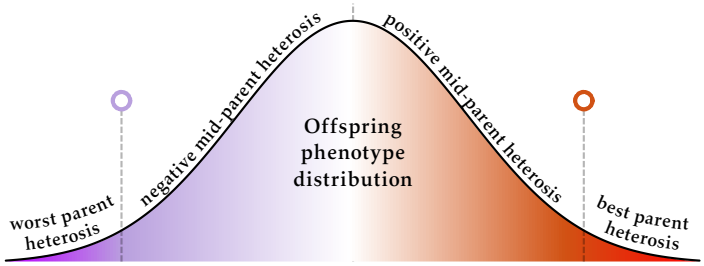


Figure 2.1. Offspring phenotype distribution. How the phenotypes of offspring behave are not always straight forward. Most often, they stay within the range of their two parents. In this figure, you see the distribution of phenotypes of offspring from the purple and orange parent. The parent phenotypes are flagged in the distribution. When the phenotype of an offspring is more extreme than that of either parent, we call it heterosis. Traditionally heterosis has been used only as defining traits where the offspring exceeds the two parents, however, I will use it to designate any deviation from the expected middle of the distribution.

mule, Charles Darwin eloquently described his admiration for this famous hybrid in his Diary of the Voyage of H.M.S. Beagle:

> *The mule always strikes me as a most surprising animal: that a Hybrid should possess far more reason, memory, obstinacy, powers of digestion & muscular endurance, than either of its parents. – One fancies art has here out-mastered Nature.*

2.2 Heterosis

Charles Darwin found that progeny from hybrid maize were performing better than that of progeny from self-pollinated plants. He stated that offspring from hybrid plants have a "*greater innate constitutional vigour*" (Duvick, 2001; Darwin, 1876). Traditionally, heterosis signifies an offspring that has a superior phenotype compared to its two parents, as it was coined by George Harrison Shull in 1914 (Larièpe et al., 2012; Shull, 1914). However, in this thesis and the accompanied article (Hallin et al., 2016) it will signify any deviation of offspring phenotypes from the immediately intermediate of the two parents (Shapira et al., 2014) (Fig. 2.1).

We designate four different categories of heterosis as seen in figure 2.1. Worst parent heterosis occurs when the offspring has a phenotype that is weaker than the weakest parent, while best parent heterosis is when an offspring has a stronger phenotype than that of the strongest parent. Positive and negative mid-parent heterosis are when the offspring have a phenotype that lies above or below the exact midpoint of the parents' phenotypes.

Heterosis has generally been investigated by looking at heterotic heterozygous offspring from crosses of highly homozygous (inbred) parents (Shapira et al., 2014). Unfortunately, this does not likely reflect how heterosis can impact natural evolution, as organisms are rarely homozygous to such an extent (at least humans (Joshi et al., 2015)). Albeit, *Saccharomyces cerevisiae* presents an exception here, as they are

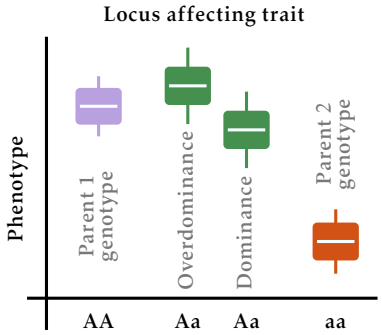


Figure 2.2. Dominance and overdominance. A locus can contribute to heterosis in different ways. A heterozygous locus may cause the phenotype to resemble one of the parents (dominance) or may exceed the phenotype of both parents (overdominance)

often highly homozygous in their natural state (Hansson and Westerberg, 2002; Magwene et al., 2011; Wang et al., 2012).

Heterosis can come as a consequence of dominance, overdominance and epistatic interactions (Shapira et al., 2014; Lippman and Zamir, 2007). These mechanisms are not mutually exclusive, but it is not clear which is most prominent. In the section below you will read about the two mechanisms that holds the focus of my thesis, dominance and overdominance.

2.3 Dominance & overdominance

We defined dominance previously in the heritability section, the only difference is that we are now inspecting how dominance can contribute to heterosis, rather than how it contributes to the overall phenotypic variation. Dominance comes from one allele masking the effect of another and has been shown to be quite prevalent in man-made yeast hybrids (Zörgö et al., 2012).

The phenomenon was discovered and coined as dominance by Gregor Mendel (Mendel, 1866). In this landmark paper he designates dominance as one parental character completely masking (dominant) the character of the other parent (recessive). Although Mendel was referring to traits being either dominant or recessive, the terms and their definitions are now used for alleles. His observation of dominance was that of complete dominance, i.e. where the trait of the hybrid was indistinguishable from that of one of the parents. Consequently, his definition of dominance only extended so far, but in this work the definition is extended to include any deviation from the mid-point of the two parents.

The **dominance hypothesis** posits that a hybrid offspring will contain many loci in the genome that has one strong allele and one weak, and that these two would be dominant and recessive, respectively. The strong dominant allele would complement the weak recessive allele, resulting in an offspring that is outperforming both parents (Bruce, 1910; Crow, 1948). When Bruce (1910) wrote this there were no experimental evidence to strengthen his assumptions, but since then numerous studies have found how dominance can contribute to heterosis (Xiao et al., 1995; Graham et al., 1997; Charlesworth and Willis, 2009).

While dominance relies on a number of different loci being complemented by the two different parents, overdominance only needs one occurrence to give a heterotic phenotype (Crow, 1948; Shapira et al.,

2014). Overdominance contributing to heterosis was proposed by East (1908) and it requires a positive interaction between two alleles at the same locus. I.e. the heterozygous state of a particular locus is more beneficial than the homozygous states of either allele. This is generally called the **overdominance hypothesis**.

Overdominance is a tempting explanation to heterosis as it only requires a few or one locus, while dominance requires several loci and, additionally, it depends on each parent having beneficial dominant variants at different loci that can complement the detrimental variants of the other parent.

However, the detection of true overdominant contributions to heterosis can be troublesome due to pseudo-overdominance. Pseudo-overdominance occurs when loci linked with the seemingly overdominant locus are in fact the loci that contribute to the phenotype. These loci are linked to the pseudo-overdominant locus and are in repulsion, i.e. the beneficial dominant alleles are coming from different parents, so combining them can result in a heterotic phenotype, and can give the impression of a locus having an overdominant effect (Charlesworth and Willis, 2009).

Several studies in plants have shown overdominance to be the mechanism by which heterosis occurs, it has been found in for example maize (Stuber et al., 1992), tomato (Semel et al., 2006) and rice (Li et al., 2001; Luo et al., 2001).

The study by Semel et al. (2006) is based on the inbred *Solanum lycopersicum* strain M82. They use 76 strains with the M82 background but each with a segment of *Solanum pennelli* (introgressed lines (ILs)) to create heterozygous strains by backcrossing these to M82. By doing this, they have strains that are, in a give segment of the genome, homozygous for *S. lycopersicum*, homozygous for *S. pennelli* or heterozygous. They use this population of strains in order to find how overdominance contributes to heterosis, and they measure 35 traits which they divide into reproductive, intermediate and non-reproductive. They find that traits that are associated to reproduction have a higher amount of overdominant QTLs than do traits that are non-reproductive. Although their study focuses on overdominance (20% of QTLS), QTLs with a dominant contribution to heterosis (27%) is more prevalant.

Semel et al. (2006) discard pseudo-overdominance as a confounding factor in their study, but the strongest QTL with a suggested overdominant effect found in the work by Stuber et al. (1992) was found to be due to pseudo-overdominance by a subsequent study that fine-mapped this QTL (Graham et al., 1997; Charlesworth and Willis, 2009).

It is still not clear what the relative contributions of dominance and overdominance are, but the amounting data on the subject seem to be favoring the masking of deleterious recessive alleles, i.e. dominance (Charlesworth and Willis, 2009).

References

Bruce A B. **The mendelian theory of heredity and the augmentation of vigor**. *Science (New York, N.Y.)*, 32(827):627–628, 1910.

Charlesworth Deborah and Willis John H. **The genetics of inbreeding depression.** *Nature Reviews Genetics*, 10(11):783–796, 2009.

Crow J F. **Alternative Hypotheses of Hybrid Vigor.** *Genetics*, 33(5):477–487, 1948.

Crow J F. ***90 years ago: the beginning of hybrid maize.***, volume 148. Genetics Society of America, Laboratory of Genetics, University of Wisconsin, Madison 53706, USA., 1998.

Darwin Charles. **The effects of cross and self fertilisation in the vegetable kingdom**. John Murray, 1876.

Duvick D N. **Biotechnology in the 1930s: the development of hybrid maize.** *Nature reviews. Genetics*, 2(1):69–74, 2001.

East E M. **Inbreeding in corn**. *Reports of the Connecticut Agricultural Experiment Station for Years*, pages 419–428, 1908.

Garrett Martin A. **The Mule in Southern Agriculture: A Requiem**. *The Journal of Economic History*, 50(4):925–930, 1990.

Graham Geoffrey I, Wolff David W, and Stuber Charles W. **Characterization of a Yield Quantitative Trait Locus on Chromosome Five of Maize by Fine Mapping**. *Crop Science*, 37(5):1601–1610, 1997.

Hallin Johan, Märtens Kaspar, et al. **Powerful decomposition of complex traits in a diploid model.** *Nature Communications*, 7:13311, 2016.

Hansson Bengt and Westerberg Lars. **On the correlation between heterozygosity and fitness in natural populations.** *Molecular ecology*, 11(12):2467–2474, 2002.

Joshi Peter K, Esko Tonu, et al. **Directional dominance on stature and cognition in diverse human populations.** *Nature*, 523(7561):459–462, 2015.

Larièpe A, Mangin B, et al. **The genetic basis of heterosis: multiparental quantitative trait loci mapping reveals contrasted levels of apparent overdominance among traits of agronomical interest in maize (Zea mays L.).** *Genetics*, 190(2):795–811, 2012.

Leducq Jean-Baptiste, Nielly-Thibault Lou, et al. **Speciation driven by hybridization and chromosomal plasticity in a wild yeast.** *Nature microbiology*, 1(1):15003, 2016.

Li Z K, Luo L J, et al. **Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass and grain yield.** *Genetics*, 158(4):1737–1753, 2001.

Lippman Zachary B and Zamir Dani. **Heterosis: revisiting the magic.** *Trends in genetics : TIG*, 23(2):60–66, 2007.

Luo L J, Li Z K, et al. **Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. II. Grain yield components.** *Genetics*, 158(4):1755–1771, 2001.

Magwene Paul M, Kayıkçı Ömür, et al. **Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in Saccharomyces cerevisiae.** *Proceedings of the National Academy of Sciences of the United States of America*, 108(5):1987–1992, 2011.

Mendel Gregor. **Versuche über plflanzenhybriden**. *Verhandlungen des naturforschenden Vereines in Brünn*, 1866.

Semel Yaniv, Nissenbaum Jonathan, et al. **Overdominant quantitative trait loci for yield and fitness in tomato.** *Proceedings of the National Academy of Sciences of the United States of America*, 103(35):12981–12986, 2006.

Shapira R, Levy T, Shaked S, Fridman E, and David L. **Extensive heterosis in growth of yeast hybrids is explained by a combination of genetic models.** *Heredity*, 113(4):316–326, 2014.

Shull George Harrison. **Duplicate genes for capsule-form in *Bursa bursa-pastoris* .** *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, 12(1):79–149, 1914.

Stuber C W, Lincoln S E, Wolff D W, Helentjaris T, and Lander E S. **Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers.** *Genetics*, 132(3):823–839, 1992.

Wang Qi-Ming, Liu Wan-Qiu, Liti Gianni, Wang Shi-An, and Bai Feng-Yan. **Surprisingly diverged populations of Sac-**

charomyces cerevisiae in natural environments remote from human activity. *Molecular ecology*, 21(22):5404–5417, 2012.

Xiao J, Li J, Yuan L, and Tanksley S D. **Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers.** *Genetics*, 140(2):745–754, 1995.

Zörgö Enikö, Gjuvsland Arne, et al. **Life history shapes trait heredity by accumulation of loss-of-function alleles in yeast.** *Molecular Biology and Evolution*, 29(7):1781–1789, 2012.

Chapter 3

On the use of yeast

"Kärt barn har många namn" is a Swedish proverb meaning: "a beloved child has many names". That certainly holds true for my model organism. Known to some by its latin name, ***Saccharomyces cerevisiae***, to others by **budding yeast**, but surely, to most by **bakers' yeast**, or simply, **yeast**. *S. cerevisiae* was the first eukaryote to have its genome sequenced in 1996, and has since then firmly asserted its position as a leading model system for genetics and genomics studies. Beyond that, it has made great contributions to practically every field of biology.

In this chapter you will read about yeast as a model in the different aspects of my work. And also about different approaches taken in order to dissect the genetic architecture of complex traits using yeast. All concepts discussed here have been explained in the previous chapters.

3.1 The model

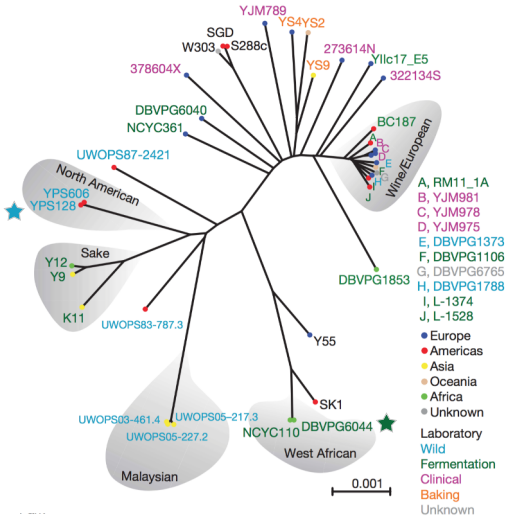
Yeast is an umbrella term that contains fungi who reproduce by budding or fission, and do not enclose their sexual states in fruiting bodies (Kurtzman et al., 2011). *S. cerevisiae* is a yeast and a unicellular fungus belonging to the *ascomycota* clade. Other ascomycetes are the molds belonging to the *Penicillium* genus, famously used to make antibiotics. The ascomycota is one of the two clades that the fungi have been divided into. The other one, *basidiomycota*, is the one that you might think of when you hear the word fungus (Fig. 3.1).

Figure 3.1. The facets of fungi. Fungi come in many different shapes and sizes. Here you find a fungi from the *basidiomycota* (above) as well as *S. cerevisiae* with the cell wall stained with calcoflour-white and expressing red fluorescent protein, representing the ascomycota (below, taken from Liti (2015)).

In the 1930's, the Danish geneticist Øjvind Winge at the Carlsberg laboratory in Copenhagen, Denmark, arguably initiated the field of yeast genetics with his work on alternation of generations and ascospores (Barnett, 2007).

On the 24th of April, 1996, *S. cerevisiae* was the first eukaryote to have its whole genome sequence released. The associated publication found a genome that is much more condensed than in other eukaryotes (e.g. the nematode *C. elegans* and humans) potentially containing 5885 protein coding genes (Goffeau et al., 1996).

But yeast has long had a life outside the lab. Possibly originating in China, it can be found all around the world and exists in a multitude of ecological niches (Liti et al., 2009; Liti, 2015). *S. cerevisiae* prefer-



In this figure, the West African strain DBVPG6044 (★) and the North American strain YPS128 (★) are the two strains used for the large cross grid in the two papers of this thesis.

Figure 3.2. ***S. cerevisiae*** **species tree.** This is the species tree of *S. cerevisiae* published by Liti et al. (2009). By sequencing the genome of 38 yeast strains from different geographical and ecological origin. They found five distinct major clades (in gray) interspersed by mosaic strains composed of the major clades.

ably exists in its diploid form and reproduces by asexual budding. Its sexual cycle can be induced by environmental triggers (which is exploited in the laboratory setting), when this happens it produces four haploid spores, segregating between the two mating types, **a** and α (Liti, 2015). The ease with which researchers can control the sexual cycle of yeast is one of its many benefits.

Other strengths of yeast as a model lie in its large population sizes, fast generation time, ease and low cost of cultivation, and the fact that it is a single cell eukaryotic organism with a relatively small genome size. A genome that, in spite of diverging from humans about 1 billion years ago (Douzery et al., 2004), shares around one third of its genes with humans (O'Brien et al., 2005). In one study, Kachroo et al. (2015) found that when they replaced 469 essential *S. cerevisiae* genes with their human orthologs, 200 of them could be funcionally replaced. This is a strong case for the shared ancestry of all organisms on the earth, and of using yeast as a model organism.

The research on *S. cerevisiae* has long focused on the model strain S288C or of strains derived from it (Liti, 2015). Naturally, only one strain (which, in fact, is a phenotypic outlier compared to other yeast strains (Warringer et al., 2011)) cannot represent the entire *S. cerevisiae* species, and certainly not the eukaryotic kingdom as a whole. Work on this model strain has been invaluable, but using the large reservoir of natural genetic and phenotypic variation in the *S. cerevisae* species tree will bring out new facets of population genetics.

3.2 Natural variation

In the last decade, the interest in the natural variation of yeast has increased among researchers (Liti, 2015). In 2009, two studies published in the same issue of Nature investigated large samples of yeast strains from diverse niches. Liti et al. (2009) and Schacherer et al. (2009) inspected the genome of 38 and 63 *S. cerevisiae* strains respectively and both found the species to have a strong population structure where a few well-defined lineages make out the back-bone of the species, with hybrids in between (Fig. 3.2).

Since then, quite a few studies have been done to bring more knowledge to the evolution of *S. cerevisiae* (Wang et al., 2012; Almeida et al., 2015; Strope et al., 2015; Gallone et al., 2016). Soon, the "1002 Yeast Genomes Project", a large collaborative project between Gianni Liti's team in Nice and Joseph Schacherer's team in Strasbourg, will reveal the largest collection of *S. cerevisiae* strains to date, along with extensive analysis of, among other things, their phylogenetic relationships.

These types of collections are important for further elucidating the genetics of complex traits, as different strains can be used for, for example, QTL mapping. Or in the case of these large collections, perhaps even for genome wide association studies. Using different strains with different genetic variants can reveal genes that are implicated in certain phenotypes, genes that could not be found with another set of strains (Treusch et al., 2015).

3.3 Phenotyping yeast

With the ever decreasing cost of sequencing (Wetterstrand, 2017), larger and larger mapping populations are feasible for which we can have a dense marker distribution. However, this means an ever larger amount of strains that need phenotyping, and our characterization of the **phenome** unfortunately lags behind our ability to characterize the genome (Houle et al., 2010).

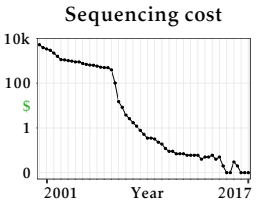


Figure 3.3. Sequencing cost has been decreasing rapidly over the years (Wetterstrand, 2017)

For more information on the 1002 Yeast Genomes Project, visit their website at 1002genomes.u-strasbg.fr.

Phenome. All conceivable phenotypes for an organism

Yeast has a range of different phenotypes that can be measured, from colony morphology (Taylor and Ehrenreich, 2015), to gene and protein expression (Brem et al., 2002; Albert et al., 2014; Parts et al., 2014). The most important phenotype for this thesis, however, is population growth.

Yeast population growth is measured either in liquid or on solid media, both of which, of course, have limitations. Liquid media has been (Warringer and Blomberg, 2003; Perlstein et al., 2007) and is being (Gallone et al., 2016; Yue et al., 2017) used to accurately measure the growth of yeast colonies. It is based on optical density measurements at one time-point (Gallone et al., 2016) or through-out the growth of the population at regular intervals (Warringer and Blomberg, 2003; Levy et al., 2012; Shapira et al., 2014). For large scale phenotyping however, liquid based methods can be difficult to scale up due to their time consuming nature (Zackrisson et al., 2016).

Phenotyping on solid media allows large scale monitoring of population growth, but has suffered from a lower precision and accuracy than its liquid counterparts (Zack-

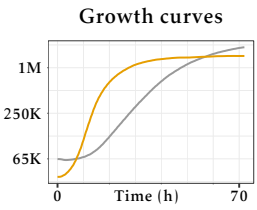


Figure 3.4. Growth curves. Different growth curves can give the same population density at a given time point, these two growth curves (taken from data from Hallin et al. (2016)) have different dynamics, but at ~60 hours, they have the same population density. *x-axis*: time in hours, *y-axis*: cell number.

risson et al., 2016). Furthermore, many large scale population growth experiments in yeast have used a single time point rather than a temporal monitoring of the growth (Sadhu et al., 2016; Kim et al., 2012; Gallone et al., 2016; Strope et al., 2015).

Measuring a single time point during growth is a gross oversimplification of the population growth dynamics. Since many different types of growth curves can amount to the same population size at any given time, the arbitrary choice of time point could have major consequences for the results of the study (Fig. 3.4) (Zackrisson et al., 2016).

A microbial growth curve can be characterized by different phases. Monod (1949) defined these phases as:

i Lag phase

ii Acceleration phase

iii Exponential phase

iv Retardation phase

v Stationary phase

vi Phase of decline

The most reproducible and easily defined of these phases is the exponential phase (or log phase) (Schaechter, 2015; Neidhardt, 2006). However, if accurately measured, the different phases can reveal different aspects of the genotype to phenotype map (Ibstedt et al., 2015). When the exponential phase is used to describe population growth it is generally reduced to growth rate (or generation time), which is calculated as a local regression of the steepest slope during the exponential phase (Warringer et al., 2011; Zackrisson et al., 2016). This value can then be used as a proxy for fitness.

The era of single time point measurements should have ended decades ago (Schaechter, 2015; Neidhardt, 2006), however, its very high through-put is alluring. In Zackrisson et al. (2016) we present a novel high through-put, high-accuracy phenotyping methodology for precise defining of microbial growth curves.

3.4 QTL mapping in yeast

The major challenges in QTL mapping is to have high enough power to detect small effect QTLs and to have high enough resolution to narrow down the QTL region to include as few non-causal markers as possible. The ultimate goal is to be able to find QTLs that explain all the variation in the phenotype and that these QTLs are small enough to identify the exact gene (or exact nucleotide) that contributes to the phenotype.

Different teams have come at these problems from different angles which will be discussed later on, but we will start off with some classical QTL mapping experiments which phenotype and genotype individual segregants.

3.4.1 Classical QTL mapping

Classical QTL mapping is based on genotyping and phenotyping individual segre-

gants, generally from a two-parent cross with the F1-segregants as the mapping population. Steinmetz et al. (2002) accurately identified and dissected the genetics behind high temperature growth (Htg) in a cross between a derivative of a clinical isolate (YJM145, Htg^+) and a lab strain (S288C, Htg^-). Interestingly, the hybrid between these two strains was heterotic, and the study would go on to define the underlying genetics of the phenomenon. The haploid progeny of the hybrid was phenotyped and 19 segregants with a strong high temperature growth phenotype were analysed at 3,444 genetic markers. They focused on the strongest QTL on chromosome XIV and using **reciprocal-hemizygosity** they found three genes within the QTL region that had an effect on the phenotype. The YJM145 allele of two of these genes was, expectedly, conferring a Htg^+ phenotype. For one, however, it was the S288C allele (i.e. the Htg^- strain) that conferred resistance. Although the beneficial S288C allele cannot fully explain the heterosis seen in the hybrid, it presents itself as an elegant contributor. While they manage to dissect the genetics of the trait in this specific QTL, they state that traits may be more complex, with more genes contributing; genes that may be closely linked.

The difficulties in identifying small effect QTLs, and the fact that they likely constitute a large portion of the variation that contributes to a complex trait (Mackay et al., 2009), led Lorenz and Cohen (2012) to further investigate their established model complex trait: sporulation efficiency (Gerke et al., 2006, 2009). In 2009, using a two-parent cross and 225 markers in a mapping population of 374 haploid segregants they located five QTLs, three of which had a large effect. However, all three large-effect QTLs had large confidence intervals of 50, 100 and 100kb. From these large-effect QTLs they located four nucleotide changes explaining 80% of the variation, and found extensive interactions between them, such that the combined effect of all of them exceed their individual effects. However, the small-effect QTLs were still elusive, and in 2012 they publish their findings on these less easily characterized QTLs (Lorenz and Cohen, 2012).

Reciprocal-hemizygosity. A method to look at the effect of different alleles in the same genetic background, it was developed by Steinmetz et al. (2002) in this article, and was later used in large scale to map QTLs (Wilkening et al., 2014).

Building on the knowledge from their previous articles, Lorenz and Cohen (2012) constructed crosses where they fixed the four large-effect variants found previously (Gerke et al., 2009). This eliminates their effect and allowed them to detect small-effect variants. Using 164 segregants they located four QTLs, interestingly, and in contrast to the large effect QTLs, the high sporulating strain contributed with two alleles increasing sporulation, and two alleles decreasing it. Another interesting observation was that the QTLs they found were highly dependent on which parental variant of the large-effect QTL was fixed in the strain, indicating QTL-QTL interactions between large- and small-effect QTLs. Also here, they conclude that they are not likely to have completely dissected this trait, and that more small-effect QTLs are still undetected.

Increasing power to detect small-effect

QTLs can be achieved, not by sequencing individual segregants, as is classically done, but instead by phenotyping and genotyping large pools of segregants.

3.4.2 Bulk segregant analysis

Bulk segregant analysis was developed by Michelmore et al. (1991) and was implemented on lettuce to showcase it as a fast and efficient way of detecting regions of the genome that are associated to genes of interest. The method consists in comparing two bulks of segregants originating from one cross. The two bulks will differ in the phenotype of interest and will be scored for a number of markers. The marker associated to the gene giving the phenotype should segregate between the two bulks and will thus be detected.

The article by Michelmore et al. actually used the term bulk**ed** segregant analysis, but many (myself included) seem to prefer the term bulk segregant analysis.

Ehrenreich et al. (2010) developed an extension of the bulk segregant analysis method for yeast which they termed Extreme QTL mapping or X-QTL. Ehrenreich et al. describes it as being composed of three key steps. *i*) Generating populations of segregants from a cross (in line with Michelmore et al.), *ii*) selecting for extreme values in these populations to recover segregants with values in the tail of the initial phenotype distribution, and finally *iii*) scoring these populations for their allele frequencies.

They selected BY4716, a lab strain derived from S288C, and the wine strain RM11-1a, to be the parental strains. The diploid hybrid from this cross was sporulated to create haploid segregants which constituted the populations used for the X-QTL. They investigated, among other phenotypes, resistance to the DNA damaging agent 4-nitroquinoline. A phenotype for which they had only found one significant QTL on chromosome XII, when studying it using conventional QTL mapping with 123 segregants (Demogines et al., 2008).

Ehrenreich et al. (2010) subjected some segregant populations to 4-nitroquinoline while others were grown under permissive conditions. With this method they successfully detect fourteen QTLs that reach above the significance threshold. These QTLs were detected by comparing the allele frequencies of the populations that had selected for 4-nitroquinoline resistance and those that had not.

A similar approach was taken by Parts et al. (2011), but with a few differences. They used two natural strains with pools that were both haploid and diploid. The pools were created by several rounds of crossing increasing the amount of recombination events. These populations were then subjected to high temperature stress (40°C), or permissive temperature (23°C) for twelve days. Using whole population DNA sequencing, they investigated the differences in allele frequencies between the selected and control populations to locate regions that had been selected for, i.e. regions that have an effect on the resistance to heat stress.

Similarly to Ehrenreich et al. (2010), a previous study using a conventional QTL mapping approach had located only one significant QTL (Cubillos et al., 2011). In-

stead, Parts et al. (2011) now find 21 QTLs by using the F12 haploid pool after 192 hours of selective growth. They find that prolonged selection will increase the power of locating small effect regions, at least up to a certain point and that the use of populations with a high amount of recombination results in narrow peaks. Narrow peaks harboring only a few possible causative genes (median interval size 6.4 kb, median number of 4 genes), and in some cases only one gene. Their peaks being narrower than that of, for example, Ehrenreich et al. (2010), who used the F1 segregants.

3.4.3 Crossing schemes

The mapping populations used by Parts et al. (2011) were created by several rounds of intercrossing. I.e. they made the yeast cells undergo several rounds of mating and sporulation. Each time the yeast cells sporulate and go from the diploid to the haploid state, recombination events occur between the two parental chromosomes. Increasing the amount of recombination events will decrease the size of linkage blocks (Darvasi and Soller, 1995), meaning that there will be smaller segments in the genome that belongs to either parent. This will increase the mapping resolution by unlinking variants that may or may not have an effect on the phenotype. These populations are called **advanced intercrossed lines** (Darvasi and Soller, 1995; Parts et al., 2011).

Instead of making several rounds of crosses within the same population, Treusch et al. (2015) designed a so called round-robin approach, in which they used twelve diverged strains (Schacherer et al., 2009) and crossed each strain to two others, creating twelve hybrids. They performed X-QTL analysis (as discussed previously) on each of these crosses in line with Ehrenreich et al. (2010). Although the round-robin approach as such does not increase power or resolution of the mapping, It does give a broader view of the natural variation in traits in contrast to when a single cross is used. The natural variation can be used in order to narrow down the potential list of causative loci once the mapping has been done, by comparing non-synonymous variants between the strains with and without the QTL.

In strains that do not readily go through meiosis, different strategies need to be used. Laureau et al. (2016) used Return To Growth (RTG) to achieve recombination between SK1 and S288C, although these two strains are not reproductively isolated, they serve as a proof of concept for the method. RTG takes advantage of yeasts ability to abort meiosis after the occurrence of double-strand breaks and recombination. When this return to growth happens, the resulting diploid strain has acquired recombined chromosomes between the two parents. The mother and first daughter cell can be isolated to catch all recombination events. Although they mostly describe the recombination landscape of RTG strains, they also map QTLs, and for a polygenic trait, arsenite resistance, they map a QTL including the ARR gene cluster, known to control arsenite resistance (Cubillos et al., 2011). The size of the QTL region is rather

Advanced intercrossed lines. Individual segregants from this F12 mapping population is what was used in the two papers of this thesis. Advanced intercrossed lines were first deviced by (Darvasi and Soller, 1995).

large at 106kb, but that is to be expected with the rather small sample size.

Loss of heterozygosity. In a diploid hybrid between two strains, all variants will be segregating, i.e. the diploid hybrid will be heterozygous at all markers between the parents. If the hybrid looses this heterozygosity in a region of the genome, that is called loss of heterozygosity, and renders that region homozygous for one of the parents

A recent study left crossing behind all together and used the genome editing technique CRISPR/Cas9 (Sadhu et al., 2016). This method goes back to genotyping and phenotyping individual segregants. However, the segregants have not undergone recombination during meiosis, but have rather had mitotic recombination induced through the double-strand breaks induced by the Cas9 protein (Doudna and Charpentier, 2014). This creates segregants with **loss of heterozygosity**, which are homologous for either parent. Sadhu et al. (2016) concentrated on the left arm of chromosome VII, and created a panel of 384 segregants with loss of heterozygosity in this region. They compared this panel with a panel of 768 classical cross-based segregants, and find that they have a higher resolution with the CRISPR/Cas9 panel (1kb).

Phased. The "phased" of phased outbred lines comes from the fact that the genomes of the hybrids are phased, i.e. we know which genotypes in the hybrid come from which parent.

They use the same method to increase the resolution of a QTL region found to contribute to manganese tolerance. Being able to target the recombination events they reach a very high density of breaks near and within the 2.9kb wide QTL region. Using this method they successfully locate the causal variant at nucleotide resolution.

Using traditional QTL mapping with individual segregants has the drawback of large costs in time and money for phenotyping and sequencing large enough numbers to have high enough power to detect low effect QTLs, as well as QTL-QTL interactions. Bulk segregant analysis increases the power but cannot be used to look at genetic interactions. In Hallin et al. (2016) we address these limitations and propose and apply a methodology based on large scale crossing of sequenced haploid strains (creating **phased** outbred lines) for decomposing the genetics of phenotypic variation.

3.5 Decomposition of genetic components

Bloom et al. (2013) used 1,008 haploid strains from a cross between a wine strain and a lab strain in order to investigate where they missing heritability problem has its solution.

The genetic contributions to phenotypic variation can be partitioned into additive effects, dominance effects, gene-environment interactions and gene-gene interactions. The use of haploid strains and phenotyping in controlled environments by Bloom et al. (2013) reduced the possible partitions to additive and gene-gene interactions.

Among 46 different measured traits, broad sense heritability was estimated from the repeatability of the trait measurements while narrow sense heritability was estimated by comparing the phenotypic similarities among individuals with their relatedness calculated from genotype data. Since only additive and gene-gene interactions exists in this experimental setup, the difference between broad and narrow heritability is an estimate of the contribution of gene-gene interactions to the phenotypic variation (Bloom et al., 2013).

The contribution of additive variance is of-

ten found to be higher than that of interactions. A finding that they reinforce in a later study (Bloom et al., 2015) where they use a larger panel of 4,390 segregants from the same cross.

A study using the panel of segregants from Bloom et al. (2013) by Young and Durbin (2014) further partition the phenotypic variance into pairwise genetic variation and higher order genetic interactions. They conclude there that pairwise interactions are not likely sufficient to explain the difference between narrow and broad sense heritability.

These articles give a good view of the genetic contributions to phenotypic variation, however, they are conducted in one cross and using haploid segregants. This excludes the dominance effects seen in diploids and as such complicates the drawing of conclusions about higher organisms. In Hallin et al. (2016) we use diploid strains, allowing us to detect dominance contributions and we successfully detect the contribution of even third-order interactions and have nearly no missing heritability.

3.6 Predicting traits in yeast

Predictions are closely related to QTLs and to heritability. The heritability of a trait sets the upper limit of what we can predict using the genome. If only 20% of the phenotypic variation in a trait is due to genotype difference between strains, then we cannot hope to predict the phenotype very accurately using the genome.

QTLs are connected to predictions since they are the loci in the genome with the most impact on the phenotypic variation. I.e. knowing the genotype of different individuals at a large effect QTL can give a good indication of what phenoype that individual will have.

Jelier et al. (2011) use conservation data of coding sequences to estimate the impact of variation in protein coding genes on the function of that protein. In short, they predicted the impact that sequence variation in a protein coding gene would have on the protein. Then they estimated the compounded effect that variation in all the genes associated to the trait in question has on the phenotype. Lastly, they compared their prediction results with real phenotype data.

Using this method they could predict the phenotypic variation, i.e. they predicted in relative terms how much a given strain would be affected by its genome in the environment measured, but they did not predict the actual phenotype.

Bloom et al. (2013) use detected QTLs to predict phenotypes. They show how, in their experiment, missing heritability can be explained by an insufficient sample size. Using an additive QTL model they can explain on average 88% of narrow sense heritability.

In our study (Märtens et al., 2016), we evaluate the theoretical limits for predictions and what information is most valuable when predicting traits.

3.7 Heterosis in yeast

Local neutrality hypothesis. Genes not under selection gain loss of function mutations, shaping the phenotypic variation between distinct natural yeast populations.

The population structure of *S. cerevisiae* makes it a good model for the study of heterosis. Having distinct populations arising mostly from clonal expansion with local adaptations (Liti and Louis, 2012; Warringer et al., 2011) is beneficial since crossing such distinct populations will result in highly heterozygous hybrids, i.e. loci in the genome that can contribute to heterosis through dominance or overdominance.

The following three papers (***i-iii***) investigate strains from Liti et al. (2009), and are quite interesting as they do so with varying results.

i) Zörgö et al. (2012) find that heterosis is exceedingly rare in natural yeast hybrids constructed in the lab. They propose that natural variation in traits in yeast comes in large part from loss of function mutations in genes that are locally not selected for, something they call the **local neutrality hypothesis**. They further hypothesize that a hybrid from two such strains will only ever reach the fitness of the best parent, not exceed it. Yeast strains that have one functional allele generally shows no phenotype (Deutschbauer et al., 2005), and so, natural strains that mainly differ between each other by loss of function mutations will be completely rescued when their genomes are brought together. Consistent with their local neutrality hypothesis, the distribution of heterosis is centered around phenotypes being completely dominant.

For the low amount of best parent heterotic hybrids that they do find, they propose that it is due to reciprocal masking of the loss of function mutations. They base this on the fact that there is an inverse relationship ($r = -0.51$) between the fitness average of the two parents, and the strength of the best parent heterosis. Basically, this means that parents that are weak in a specific environment likely carry a number of loss of function mutations which will be masked in the resulting hybrid, giving it a proportionally strong phenotype. In an environment where the parents are strong, however, there are less loss of function mutations between the two and the hybrid will be less likely to outcompete them.

Zörgö et al. (2012) conclude among other things that the genotype-phenotype landscape is most likely defined by genetic drift. However, it seems they did not test the potential benefit of the loss of function mutation in the yeast strains natural habitat, and can therefore not confidently say that the loss of function mutations are not actually adaptive.

ii) Plech et al. (2014) expand on the Zörgö et al. (2012) study and find that heterosis is uncommon among wild strains of *S. cerevisiae*, however, they find that domesticated strains exhibit a high occurrence of heterosis. Heterosis being defined as any positive deviation from the average of the two parents. They attribute their finding to the fact that they had a larger amount of strains as compared to the study from Zörgö et al. (2012).

The general conclusion of heterosis be-

ing more prevalent among domesticated strains seems to hold (and has since then been enforced by a study on *S. cerevisiae* and *S. paradoxus*), but the calling of heterosis was not completely to my liking due to the apparent lack of statistical inspection of the data. However, it would fit with the local neutrality hypothesis, given that the domestication would bring with it a relaxed selection pressure and therefore domesticated strains would harbor more detrimental variants. Again, favoring the dominance hypothesis.

iii) Another study on a similar (and somewhat overlapping) set of natural strains by Shapira et al. (2014) get rather different results. They find that in their panel of natural strains, an average of 35% were best parent heterotic, ranging from 23 to 47%. This is radically different from the low occurrence of best parent heterosis observed by Zörgö et al. (less than 5%). They suggest that this discordance is due to three components: *i*) only partial overlap of the strains used, *ii*) difference in heterosis calculation and *iii*) the environments used were more complex than in Zörgö et al. (2012). I would like to add a fourth possible explanation for this discrepancy: the seemingly lacking significance test of true best parent heterotic hybrids in Shapira et al. (2014). They call best parent heterosis simply when a hybrid exceeds the value of the best performing parent, while Zörgö et al. (2012) call best parent heterosis only when there is a significant difference between the hybrid and the best performing parent, as designated by a one-sided Student's t-test. The lack of a significance threshold would likely include false positive best parent heterotic hybrids and thus inflate the values.

Shapira et al. (2014) uses overdominance synonymously with best parent heterosis, good thing to keep in mind if reading it.

Similarly to Zörgö et al. (2012), Shapira et al. (2014) find that less fit parents tend to have more heterotic hybrids, giving merit to the dominance hypothesis. However, they also find heterotic hybrids between parents with high fitness, and, following the same logic, propose that this is due to overdominance or epistasis.

By backcrossing the hybrids to one parent, Shapira et al. tests the dominance hypothesis. The rationale is the following: since backcrossing will remove on average half of the heterozygosity, the resulting population should loose half of the phenotype if it is determined by dominance complementation. They find this, but also examples of when more or less of the phenotype is lost. They conclude that heterosis is complex and that its causes include dominance complimentation, overdominance, and epistasis. And that these causes differ in their pervasiveness between different hybrids, but that can they also exist simultaneously in the same hybrid.

A more fine-grained view of heterosis has been hindered by a common feature of these articles, which is that none of them look at individual variants' contribution to heterosis. Instead they infer the general genetic contribution to heterosis by looking solely at the phenotypes. Further hindering the genetic decomposition of heterosis is a common feature of QTL mapping experiments in *S. cerevisiae*: they have

largely been performed in haploids (with an exception of Parts et al. (2011)), making it impossible to look at contributions to diploid heterosis. With our methodology in Hallin et al. (2016), we take advantage of our large mapping population consisting of diploid hybrids and look at the dominance and overdominance contribution to heterosis of the QTLs found during linkage analysis.

References

Albert Frank W, Treusch Sebastian, Shockley Arthur H, Bloom Joshua S, and Kruglyak Leonid. **Genetics of single-cell protein abundance variation in large yeast populations.** *Nature*, 506(7489):494–497, 2014.

Almeida Pedro, Barbosa Raquel, et al. **A population genomics insight into the Mediterranean origins of wine yeast domestication.** *Molecular ecology*, 24(21):5412–5427, 2015.

Barnett James A. **A history of research on yeasts 10: foundations of yeast genetics.** *Yeast (Chichester, England)*, 24(10):799–845, 2007.

Bloom Joshua S, Ehrenreich Ian M, Loo Wesley T, Lite Thúy-Lan Võ, and Kruglyak Leonid. **Finding the sources of missing heritability in a yeast cross.** *Nature*, 494(7436):234–237, 2013.

Bloom Joshua S, Kotenko Iulia, et al. **Genetic interactions contribute less than additive effects to quantitative trait variation in yeast.** *Nature Communications*, 6:8712–6, 2015.

Brem Rachel B, Yvert Gaël, Clinton Rebecca, and Kruglyak Leonid. **Genetic dissection of transcriptional regulation in budding yeast.** *Science (New York, N.Y.)*, 296(5568):752–755, 2002.

Cubillos Francisco A, Billi Eleonora, et al. **Assessing the complex architecture of polygenic traits in diverged yeast populations.** *Molecular ecology*, 20(7):1401–1413, 2011.

Darvasi A and Soller M. **Advanced intercross lines, an experimental population for fine genetic mapping.** *Genetics*, 141(3):1199–1207, 1995.

Demogines Ann, Smith Erin, Kruglyak Leonid, and Alani Eric. **Identification and dissection of a complex DNA repair sensitivity phenotype in Baker's yeast.** *PLoS Genetics*, 4(7):e1000123, 2008.

Deutschbauer Adam M, Jaramillo Daniel F, et al. **Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast.** *Genetics*, 169(4):1915–1925, 2005.

Doudna Jennifer A and Charpentier Emmanuelle. **Genome editing. The new frontier of genome engineering with CRISPR-Cas9.** *Science (New York, N.Y.)*, 346(6213):1258096–1258096, 2014.

Douzery Emmanuel J P, Snell Elizabeth A, Bapteste Eric, Delsuc Frédéric, and Philippe Hervé. **The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils?** *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15386–15391, 2004.

Ehrenreich Ian M, Torabi Noorossadat, et al. **Dissection of genetically complex traits with extremely large pools of yeast segregants.** *Nature*, 464(7291):1039–1042, 2010.

Gallone Brigida, Steensels Jan, et al. **Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts.** *Cell*, 166(6):1397–1410.e16, 2016.

Gerke Justin, Lorenz Kim, and Cohen Barak. **Genetic interactions between transcription factors cause natural variation in yeast.** *Science (New York, N.Y.)*, 323(5913):498–501, 2009.

Gerke Justin P, Chen Christina T L, and Cohen Barak A. **Natural isolates of Saccharomyces cerevisiae display complex genetic variation in sporulation efficiency.** *Genetics*, 174(2):985–997, 2006.

Goffeau A, Barrell B G, et al. **Life with 6000 genes.** *Science (New York, N.Y.)*, 274(5287):546–563–7, 1996.

Hallin Johan, Märtens Kaspar, et al. **Powerful decomposition of complex traits in a diploid model.** *Nature Communications*, 7:13311, 2016.

Houle David, Govindaraju Diddahally R, and Omholt Stig W. **Phenomics: the next challenge.** *Nature Reviews Genetics*, 11(12):855–866, 2010.

Ibstedt Sebastian, Stenberg Simon, et al. **Concerted evolution of life stage performances signals recent selection on yeast nitrogen use.** *Molecular Biology and Evolution*, 32(1):153–161, 2015.

Jelier Rob, Semple Jennifer I, Garcia-Verdugo Rosa, and Lehner Ben. **Predicting phenotypic variation in yeast from individual genome sequences**. *Nature genetics*, 43(12):1270–1274, 2011.

Kachroo Aashiq H, Laurent Jon M, et al. **Systematic humanization of yeast genes reveals conserved functions and genetic modularity.** *Science (New York, N.Y.)*, 348(6237):921–925, 2015.

Kim Hyun Seok, Huh Juyoung, Riles Linda, Reyes Alejandro, and Fay Justin C. **A noncomplementation screen for quantitative trait alleles in saccharomyces cerevisiae.** *G3 (Bethesda, Md.)*, 2(7):753–760, 2012.

Kurtzman Cletus, Fell J W, and Boekhout Teun. ***The Yeasts***. A Taxonomic Study. Elsevier, 2011.

Laureau Raphaëlle, Loeillet Sophie, et al. **Extensive Recombination of a Yeast Diploid Hybrid through Meiotic Reversion.** *PLoS Genetics*, 12(2):e1005781, 2016.

Levy Sasha F, Ziv Naomi, and Siegal Mark L. **Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant.** *PLoS biology*, 10(5):e1001325, 2012.

Liti Gianni. **The fascinating and secret wild life of the budding yeast S. cerevisiae.** *eLife*, 4, 2015.

Liti Gianni, Carter David M, et al. **Population genomics of domestic and wild yeasts.** *Nature*, 458(7236):337–341, 2009.

Liti Gianni and Louis Edward J. **Advances in quantitative trait analysis in yeast.** *PLoS Genetics*, 8(8):e1002912, 2012.

Lorenz Kim and Cohen Barak A. **Small- and large-effect quantitative trait locus interactions underlie variation in yeast sporulation efficiency.** *Genetics*, 192(3):1123–1132, 2012.

Mackay Trudy F C, Stone Eric A, and Ayroles Julien F. **The genetics of quantitative traits: challenges and prospects.** *Nature Reviews Genetics*, 10(8):565–577, 2009.

Märtens Kaspar, Hallin Johan, Warringer Jonas, Liti Gianni, and Parts Leopold. **Predicting quantitative traits from genome and phenome with near perfect accuracy.** *Nature Communications*, 7:11512, 2016.

Michelmore R W, Paran I, and Kesseli R V. **Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations.** *Proceedings of the National Academy of Sciences of the United States of America*, 88(21):9828–9832, 1991.

Monod Jacques. **The growth of bacterial cultures**. *Annual review of microbiology*, 3:371–394, 1949.

Neidhardt Frederick C. **Apples, oranges and unknown fruit.** *Nature reviews. Microbiology*, 4(12):876–876, 2006.

O'Brien Kevin P, Remm Maido, and Sonnhammer Erik L L. **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic acids research*, 33(Database issue):D476–80, 2005.

Parts Leopold, Cubillos Francisco A, et al. **Revealing the genetic structure of a trait by sequencing a population under selection.** *Genome research*, 21(7):1131–1138, 2011.

Parts Leopold, Liu Yi-Chun, et al. **Heritability and genetic basis of protein level variation in an outbred population.** *Genome research*, 24(8):1363–1370, 2014.

Perlstein Ethan O, Ruderfer Douglas M, Roberts David C, Schreiber Stuart L, and Kruglyak Leonid. **Genetic basis of individual differences in the response to small-molecule drugs in yeast.** *Nature genetics*, 39(4):496–502, 2007.

Plech Marcin, de Visser J Arjan G M, and Korona Ryszard. **Heterosis is prevalent among domesticated but not wild strains of Saccharomyces cerevisiae.** *G3 (Bethesda, Md.)*, 4(2):315–323, 2014.

Sadhu Meru J, Bloom Joshua S, Day Laura, and Kruglyak Leonid. **CRISPR-directed mitotic recombination enables genetic mapping without crosses**. *Science (New York, N.Y.)*, pages 1–5, 2016.

Schacherer Joseph, Shapiro Joshua A, Ruderfer Douglas M, and Kruglyak Leonid. **Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae.** *Nature*, 458(7236):342–345, 2009.

Schaechter Moselio. **A brief history of bacterial growth physiology.** *Frontiers in microbiology*, 6:289, 2015.

Shapira R, Levy T, Shaked S, Fridman E, and David L. **Extensive heterosis in growth of yeast hybrids is explained by a combination of genetic models.** *Heredity*, 113(4):316–326, 2014.

Steinmetz Lars M, Sinha Himanshu, et al. **Dissecting the architecture of a quantitative trait locus in yeast.** *Nature*, 416(6878):326–330, 2002.

Strope Pooja K, Skelly Daniel A, et al. **The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen.** *Genome research*, 25(5):762–774, 2015.

Taylor Matthew B and Ehrenreich Ian M. **Transcriptional Derepression Uncovers Cryptic Higher-Order Genetic Interactions.** *PLoS Genetics*, 11(10):e1005606, 2015.

Treusch Sebastian, Albert Frank W, Bloom Joshua S, Kotenko Iulia E, and Kruglyak Leonid. **Genetic mapping of MAPK-mediated complex traits Across S. cerevisiae.** *PLoS Genetics*, 11(1):e1004913, 2015.

Wang Qi-Ming, Liu Wan-Qiu, Liti Gianni, Wang Shi-An, and Bai Feng-Yan. **Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity.** *Molecular ecology*, 21(22):5404–5417, 2012.

Warringer Jonas and Blomberg Anders. **Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in Saccharomyces cerevisiae.** *Yeast (Chichester, England)*, 20(1):53–67, 2003.

Warringer Jonas, Zörgö Enikö, et al. **Trait variation in yeast is defined by population history.** *PLoS Genetics*, 7(6):e1002111, 2011.

Wetterstrand Kris. **DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)**. *www.genome.gov/sequencingcostsdata/*, 2017.

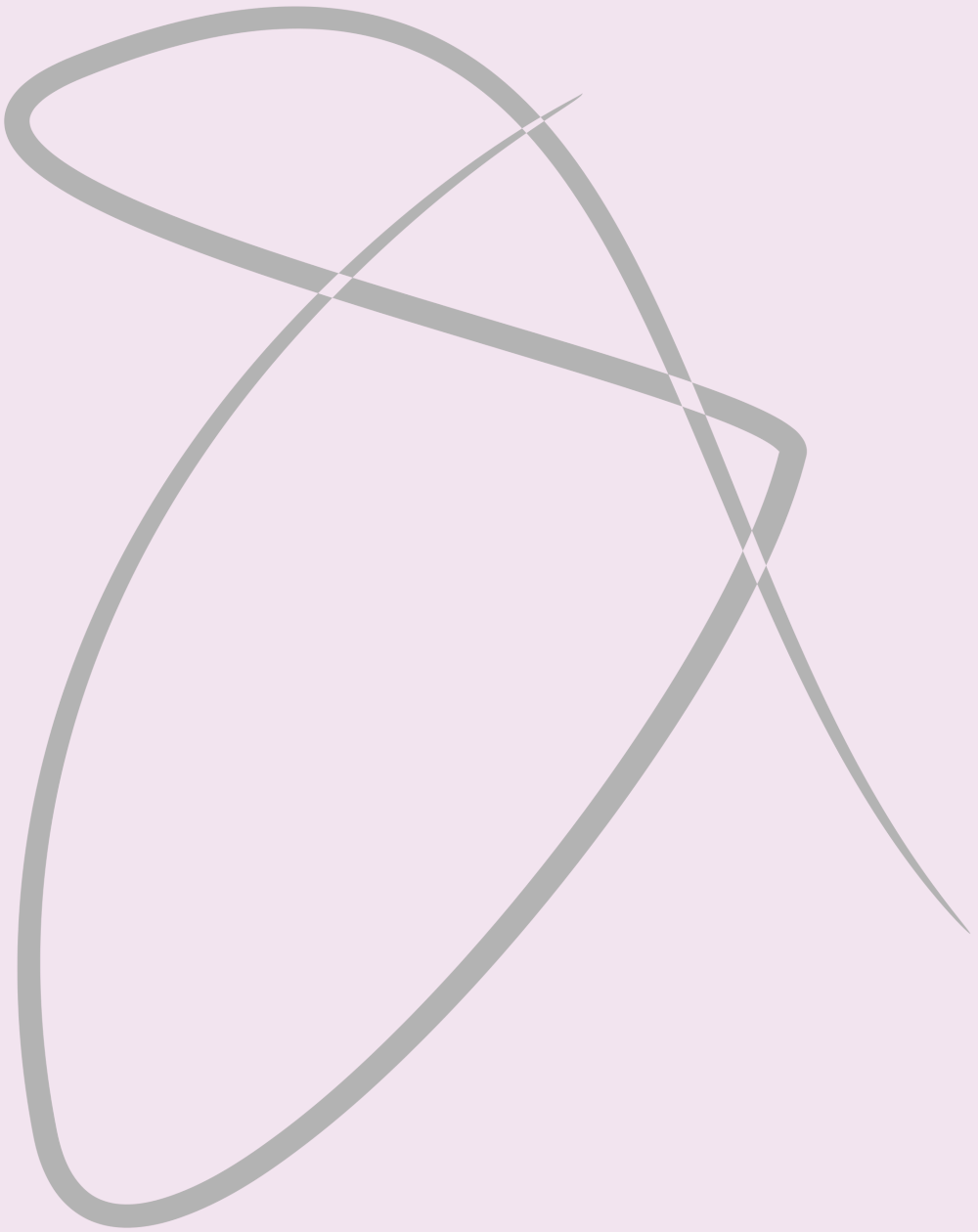
Wilkening Stefan, Lin Gen, et al. **An evaluation of high-throughput approaches to QTL mapping in Saccharomyces cerevisiae.** *Genetics*, 196(3):853–865, 2014.

Young Alexander I and Durbin Richard. **Estimation of epistatic variance components and heritability in founder populations and crosses.** *Genetics*, 198(4):1405–1416, 2014.

Yue Jia-Xing, Li Jing, et al. **Contrasting evolutionary genome dynamics between domesticated and wild yeasts.** *Nature genetics*, 49(6):913–924, 2017.

Zackrisson Martin, Hallin Johan, et al. **Scan-o-matic: High-Resolution Microbial Phenomics at a Massive Scale**. *G3 (Bethesda, Md.)*, 6:1–12, 2016.

Zörgö Enikö, Gjuvsland Arne, et al. **Life history shapes trait heredity by accumulation of loss-of-function alleles in yeast.** *Molecular Biology and Evolution*, 29(7):1781–1789, 2012.



Articles and ongoing project

During my undergraduate and PhD studies, I was lucky enough to have been a part of developing a high-throughput growth phenotyping methodology (Zackrisson et al., 2016). This phenotyping platform, Scan-o-matic, set the stage for the experiments that I would do during my PhD.

My two main publications are based on a large-scale crossing experiment that I set up and performed as a visiting researcher in Dr. Jonas Warringer's lab at the University of Gothenburg, producing more than 7,000 diploid hybrids which I phenotyped with Scan-o-matic and used to investigate the connection between genotype and phenotype.

During my time in Dr. Warringer's lab, I also created a smaller cross that I also phenotyped in different environments. This smaller set of hybrids was used in Yue et al. (2017) (see chapter 8 for the abstract).

In the first publication (chapter 4) I mapped QTLs using an approach set up by Kaspar Märtens and myself, as well as looking into the occurrence and genetic basis for heterosis with novel technique. Alexander Young contributed greatly with his compartmentalizing of the phenotypic variation into its additive, dominance, and second and third order epistasis components. During this project I was a visiting researcher in Dr. Leopold Parts lab at the Wellcome Trust Sanger Institute (United Kingdom) where I finalized the analysis for the article.

In (chapter 5) the focus was on testing the limits of prediction complex traits using genetic and phenotypic information from distant and close relatives. Kaspar used the phenotype data that I produced to spearhead the prediction analysis.

In (chapter 6) I describe my ongoing project. As it is ongoing, it will mostly focus on methodology and on some preliminary data. In this project, as in the other two, I use large scale phenomics and genomics to investigate the genotype to phenotype map, but this time with a focus on meiosis and gametes.

Chapter 4

Powerful decomposition of complex traits in a diploid model

Explaining trait differences between individuals is a core and challenging aim of life sciences. Here, we introduce a powerful framework for complete decomposition of trait variation into its underlying genetic causes in diploid model organisms. We sequence and systematically pair the recombinant gametes of two intercrossed natural genomes into an array of diploid hybrids with fully assembled and phased genomes, termed Phased Outbred Lines (POLs). We demonstrate the capacity of this approach by partitioning fitness traits of 6,642 *Saccharomyces cerevisiae* POLs across many environments, achieving near complete trait heritability and precisely estimating additive (73%), dominance (10%), second (7%) and third (1.7%) order epistasis components. We map quantitative trait loci (QTLs) and find nonadditive QTLs to outnumber (3:1) additive loci, dominant contributions to heterosis to outnumber overdominant, and extensive pleiotropy. The POL framework offers the most complete decomposition of diploid traits to date and can be adapted to most model organisms.

Johan Hallin*, Kaspar Märtens*, Alexander I. Young, Martin Zackrisson, Francisco Salinas, Leopold Parts, Jonas Warringer & Gianni Liti

Published in Nature Communications (2016)
doi:10.1038/ncomms13311

Introduction Decomposing the trait variation within natural populations into its genetic components is a fundamental goal of biology that has proven to be challenging (Visscher et al., 2012; Eichler et al., 2010). Environmental and gene-by-environment influences are difficult to control and alleles accounting for trait variation tend to have frequencies that are too low for their mostly weak effects to be reliably called (Yang et al., 2010). Compounding matters, many alleles are believed to influence each other within (dominance) and between (epistasis) loci (Lehner, 2011). Consequently, one trait can be the result of many different allele combinations, each combination being exceedingly rare in the population. This makes the individual contributions of most alleles near impossible to assess (Zuk et al., 2012).

Model organisms offer more complete dissection of complex traits because they can be analysed in controlled contexts, minimizing environmental and gene-by-environment variation, and in populations derived from a few founders, ensuring high frequencies of all alleles and allele combinations (Abney et al., 2000; Lehner, 2013). Because of their ease of use in genomics (Liti and Schacherer, 2011) and phenomics (Hancock et al., 2014), large panels of haploid yeast segregants have allowed for fine-grained dissection of complex traits (Bloom et al., 2015, 2013; Young and Durbin, 2014). Unfortunately, exhaustive trait decomposition in haploid crosses requires the costly genotyping of thousands of genomes, disregards dominance and provides much simplified estimates of epistasis. A more complete partitioning of trait variation that is relevant to a diploid context has remained elusive.

Inspired by previous thinking and theoretical work on recombinant inbred intercrosses in other model organisms (Threadgill et al., 2002; Tsaih et al., 2005; Zou et al., 2005), we here introduce a powerful and cost-effective framework for tracking the covariation through genome and phenome that allows accurate estimates of dominance and epistasis in diploid models. The framework is based on intercrossing two natural genomes over many sexual generations to reduce linkage (Parts et al., 2011; Cubillos et al., 2013) followed by sequencing and systematic pairing of the resulting haploid recombinant segregants to generate a very large array of diploid hybrids with fully assembled and phased genomes, termed Phased Outbred Lines (POLs). We validate the capacity of the POLs approach by genetic decomposition of growth trait variation across 6,642 diploid yeast genomes in nine distinct environments, and our results provide the most complete decomposition of diploid traits to date.

Results

An experimental framework for diploid complex trait analysis To accurately decompose diploid trait variation, we first isolated and sequenced the full genomes of 86 *MAT***a** and 86 *MATα* haploid *Saccharomyces cerevisiae* strains. These haploids were randomly drawn from a twelfth

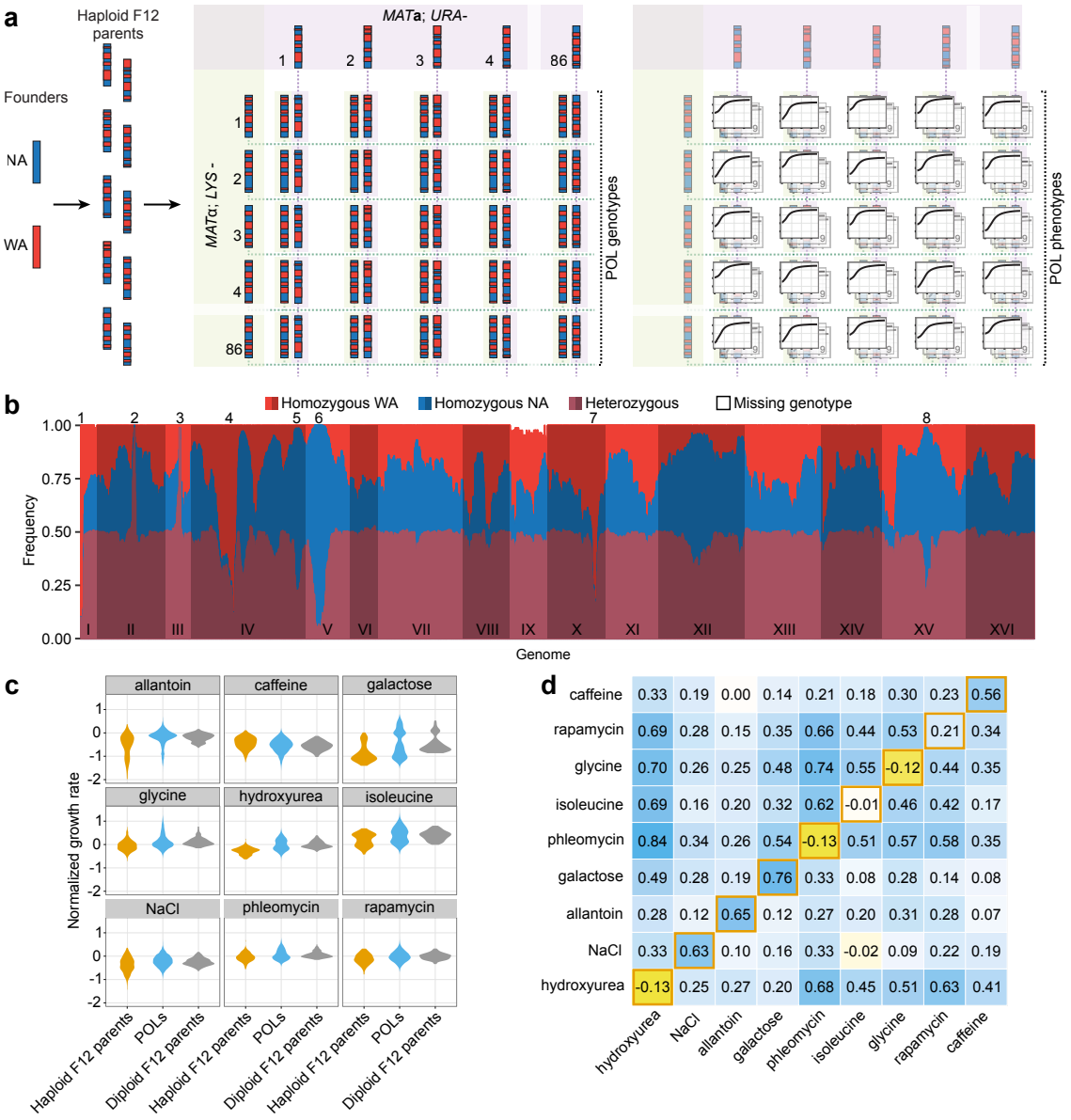


Figure 4.1. An experimental framework for analysis of diploid traits. (a) Experimental design. Left panel: Advanced intercrossed lines were constructed by multiple rounds of random mating and sporulation of North American (NA) and West African (WA) genomes. Middle panel: We sequenced 172 of the resulting segregants and paired these to generate an array of 7,310 diploid hybrids (POLs). Right panel: The POLs and their F12 haploid parents were growth phenotyped in nine environments, providing high resolution growth curves. (b) Frequency of homozygotes (red: WA/WA, blue: NA/NA), heterozygotes (purple: NA/WA) and missing genotypes (white, mostly attributed to chr. IX aneuploidies) at each segregating site among the 7,310 POLs. Deviations from 50% heterozygosity are explained by selection (numbers 1, 4–8) against one allele in the F12 haploid parent construction, or by forced heterozygosity at the *LYS2* (number 2) and *MAT* (number 3) loci. (c) Growth rate distributions of POLs (blue), their haploid F12 parents (orange) and the diploid parent estimates (grey, Methods). (d) Correlations (Pearson's r) between the growth rate and mean growth for POLs within environments (lower left to upper right diagonal; orange borders), between growth rates (above diagonal) and mean growth (below diagonal) in pairs of environments. Colour intensity (3-colour scale: dark yellow to white to dark blue) and number indicate the degree of correlation r.

generation two-parent intercross pool, constructed using highly diverged (0.53% nucleotide difference) wild strains, here termed North American (NA) and West African (WA). Only two alleles segregate at each polymorphic site, with on average equal representation in the pool (Parts et al., 2011). The sequenced haploids of opposite mating types were systematically crossed in all possible pairwise combinations to generate 7,396 genetically distinct diploid hybrids, retaining 6,642 POLs used for all downstream analysis (Fig. 4.1a, Methods).

With only a modest number of 172 haploid genomes sequenced (Illingworth et al., 2013), we could accurately infer the genomes of our large set of POLs. Notably, these genomes are fully phased, that is, we know the parent-of-origin for each allele and their combination into diplotypes. Furthermore, a very small fraction of genotype information is missing (max: 6.5%; mean: 0.5%; median: 0.1%; min 0%) and there are no confounding effects from segregating auxotrophies that contribute to trait variation (Supplementary Data 1). The hybrids showed remarkable uniformity, with heterozygote frequencies close to 50% (Fig. 4.1b). The few strong deviations (eight deviations>30%) from 50% heterozygosity were either due to selection for one parental allele during the intercross (overrepresentation of homozygous sites) or from the crossing design, the latter resulting in regions of fixed heterozygosity at the *MAT* and *LYS2* loci (Fig. 4.1b). Hybrid pairs sharing one haploid parent will be genetically more similar than two POLs that do not share a parent (expected fraction of loci with identical genotypes = 0.5 and 0.375, respectively), resulting in a bimodal distribution of the genetic relationship matrix entries (Märtens et al., 2016).

We precisely phenotyped the complete set of 6,642 designed POLs (median CoV = 10%, mean CoV = 14%), their F12 haploid parents, the diploid NA and WA founders and their hybrid in a well replicated ($n \leq 4$) manner, using a high resolution growth phenomics platform designed to minimize noise and bias(Zackrisson et al., 2016). We selected nine physiologically distinct environmental conditions (Supplementary Table 1) that challenged growth to different extents (Supplementary Fig. 1a), and we obtained >50 million population size estimates, organized into circa 250,000 growth curves (Fig. 4.1a, right panel). Extracting the (maximum) growth rate and a mean growth phenotype (Methods) from each growth curve (Supplementary Data 2), we found phenotype distributions across the POLs to be mostly monomodal (Fig. 4.1c; Supplementary Fig. 1a,b). Given the near absence of environmental variation, this implies complex traits with multiallelic influences. Growth in galactose and allantoin was bimodally distributed, in agreement with large effect sizes for the GAL3 (WA premature stop codon) and DAL (linked loci, WA loss-of-function SNPs in DAL1 and DAL4) genes respectively (Warringer et al., 2011; Ibstedt et al., 2015). Correlation between growth rate and mean growth ranged from −0.13 to 0.76 (Pearson's r; Fig. 4.1d, orange borders) but was overall low (mean r: 0.27; median r: 0.21). This

agrees with the hypothesis that distinct genetic factors control population expansion in different growth phases (Warringer et al., 2011, 2008). Correlations across environments were positive in all but one case (r = −0.02) and often of moderate or large magnitude (max r = 0.84, median r = 0.29; Fig. 4.1d). We cannot completely exclude a small influence of shared error on correlations, but the extensive standardization, randomization and normalization (Methods), and the large variation in pairwise correlations argue compellingly in favour of extensive positive pleiotropy.

Near complete variance decomposition of diploid traits Based on the *in silico* constructed diploid genomes, we used a random effects model to partition the variance in growth traits into components arising from additive (no interaction), dominance (intralocus interaction) and pairwise and third order epistatic effects (interlocus interactions) (Supplementary Note 1). We first evaluated whether the model could accurately estimate variance components as well as their uncertainty via simulation (Supplementary Note 1). The simulations showed that the model could accurately decompose the variance into additive, dominance, and pairwise epistatic components, and that s.e. estimates were well calibrated (Supplementary Data 3 and 4). When adding a component for third order interactions, the overall variance decomposition became somewhat biased, possibly due to introducing non-convexity into the optimization problem. However, the variance from third order interactions was estimated accurately (Supplementary Data 4). Due to the biasing effect, the variance decomposition for third order interactions was performed and reported separately.

The large sample size, known large variation in relatedness and absence of environmental variation allowed us to estimate nonadditive variance components with unprecedented accuracy. Thus, additivity, dominance and pairwise epistasis accounted for almost all trait variation (broad sense heritability, H^2 = 80–99% depending on the trait, median 91%, Fig. 4.2, upper panel). On average, the proportion of phenotypic variance explained by additive effects was 73% (50–87%), for dominance effects this was 10% (2–45%), and for pairwise interactions this was 7% (1–15%). Complete dominance of the functional NA over the nonfunctional WA cluster of DAL genes (Ibstedt et al., 2015) ensured a considerable dominance component for the variation in the two allantoin phenotypes, growth rate and mean growth. Otherwise, the large variance contributions of additive genetic influences were consistent across environments (Fig. 4.2, upper panel).

The trait with the largest estimated variance from pairwise epistasis was growth rate on glycine (15%); this epistasis variance contribution equalled one third of the largest dominance variance estimate (45% for allantoin growth rates). We estimated that third order interactions accounted for 1.7% of the trait variation on average (Fig. 4.2, lower panel). However, only growth rates on isoleucine, glycine

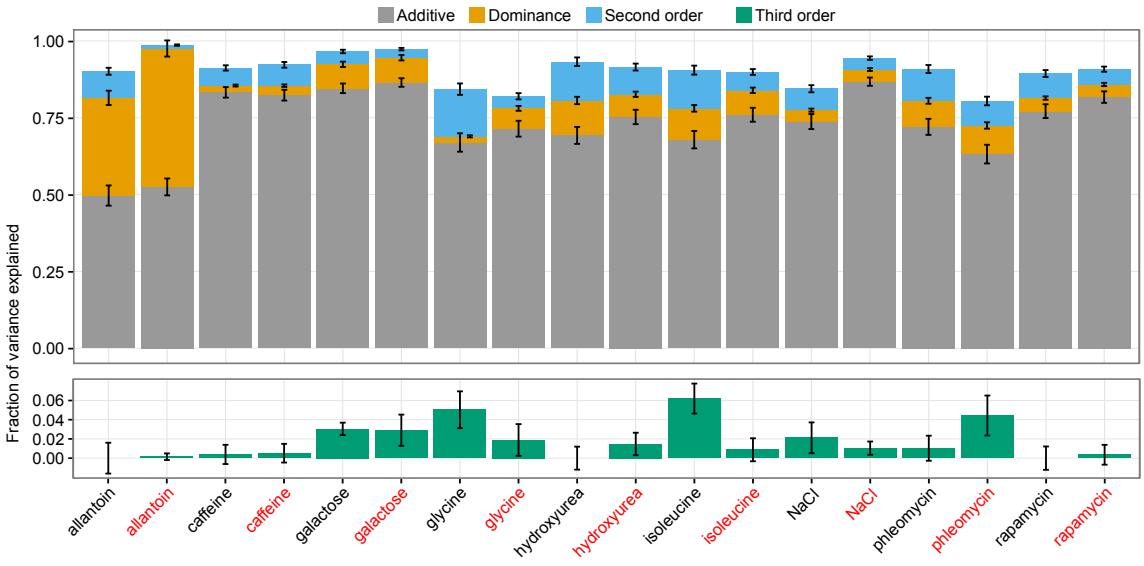


Figure 4.2. Near complete variance decomposition of diploid traits. Decomposing the total variance in growth traits across 6,642 diploids into additive (grey, upper panel), dominance (yellow, upper panel), second order epistatic (blue, upper panel) and third order epistatic (green, lower panel) genetic contributions. Black label = growth rate, red label = mean growth. Error bars = s.e.m.

and galactose, and mean growth in the presence of phleomycin were significantly (>2 s.e.m. from 0) affected by third order epistasis. Variation in genome wide levels of homozygosity had no detectable influence on yeast fitness traits (Supplementary Fig. 2). This is in stark contrast to its substantial negative effect on human traits, for example, height (Joshi et al., 2015). Thus, the data suggest that there is no general inbreeding depression in yeast, consistent with natural populations being largely homozygous (Magwene et al., 2011; Wang et al., 2012).

Cost-efficient QTL mapping in yeast POL diploid hybrids Our crossing design resulted in that one haploid genome of each POL is kept constant across the 86 POLs that are derived from any one of its haploid F12 parents (Fig. 4.1a). This sharing of half a genome accounted for surprisingly much of the overall variation in traits, which somewhat restricted our capacity to distinguish contributions from individual alleles and allele pairs from the effect of the genetic background. Nevertheless, our platform provided a cost-efficient framework for calling both additive and nonadditive (dominance and epistasis) QTLs in diploid models. We mapped QTLs using 52,466 markers, the inferred parent phenotypes (for additive effect of genetic background) and the hybrids' deviations from the average of the inferred parental phenotypes (for nonadditive effects; Methods). Both QTL mapping approaches accounted for the population structure. We called a total of 145 unique QTLs at 10% false discovery rate (FDR) with high resolution (median 1.8-LOD support interval = 3.67Kbp,

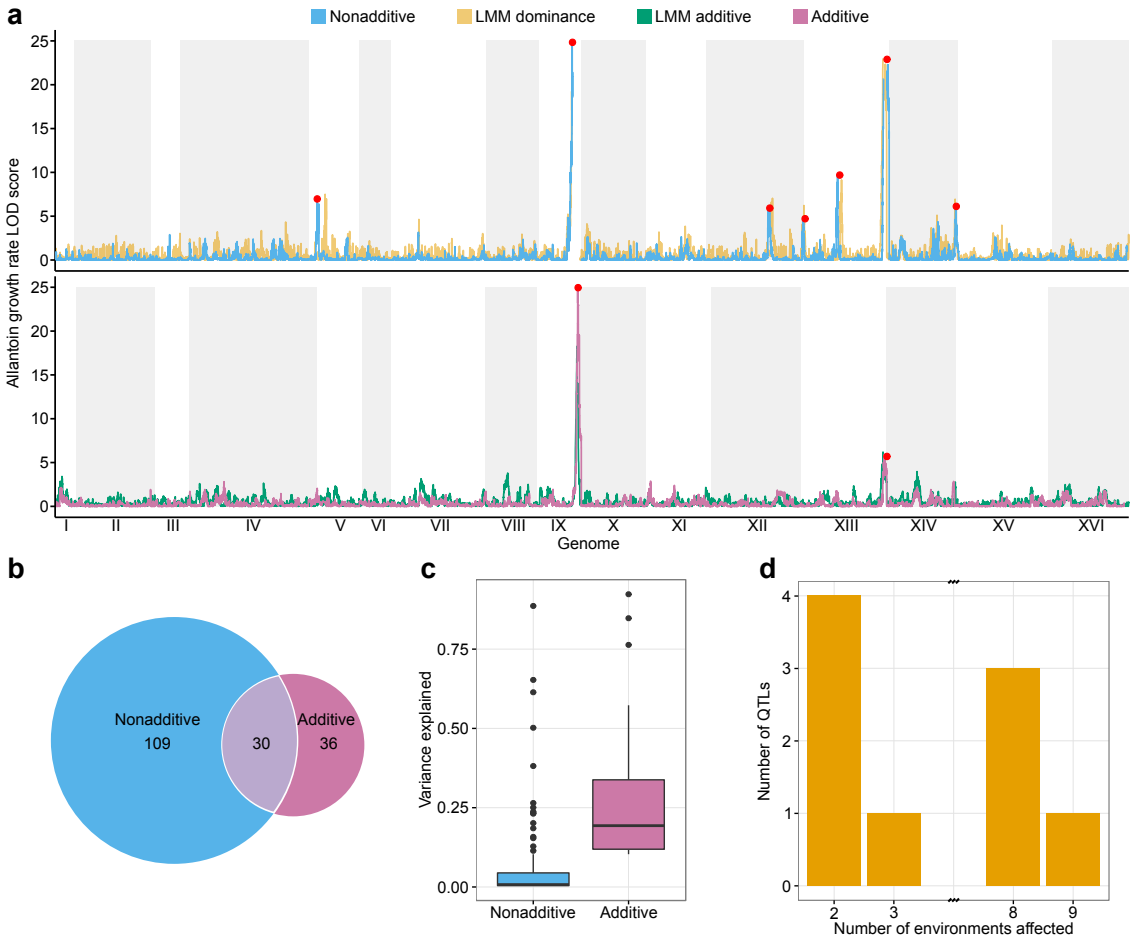


Figure 4.3. Cost-efficient QTL mapping in yeast POLs. QTLs were mapped across 6,642 genomes and 18 traits based on additive and nonadditive contributions. QTLs were validated as additive or dominant genetic contributions using Linear Mixed Models (LMM). (a) QTL signal strength (LOD score, *y*-axis) as a function of genomic position (*x*-axis), for growth rate on allantoin as sole nitrogen source, using additive (LMM and non-LMM; lower panel) and nonadditive (non-LMM and LMM only capturing dominance; upper panel) models. Red dots indicate significant (FDR, q = 10%) QTL calls. White/grey fields indicate chromosome spans. (b) Venn diagram of significant QTLs capturing additive and nonadditive genetic contributions. All 18 phenotypes (growth rate and mean growth over nine environments) were considered, with pleiotropic QTLs counted multiple times. (c) Tukey boxplot showing the fraction of variance explained by additive (purple) and nonadditive (blue) significant QTLs (non-LMM models). (d) Histogram of pleiotropic QTLs. A QTL was counted as shared across environments if peaks were within 10 kb of each other. No QTLs were significant in 4, 5, 6 or 7 environments.

Supplementary Data 5). These included the GAL3 stop codon variant, as well as the DAL1 and DAL4 non-synonymous and stop codon mutations, known to account for most of the variation in galactose and allantoin growth respectively (Fig. 4.3a, Supplementary Figs 3 and 4 and Supplementary Data 5).

Some (21%) of the QTLs contributed significantly to both additive and nonadditive phenotype components, but the majority were private to one of them (Fig. 4.3b). The nonadditive (75%) outnumbered the additive (25%) QTLs, but explained on average less of the variation (6 versus 28%, Student's t-test: $P = 2 \times 10^{6}$, Fig. 4.3c, Supplementary Fig. 5). Thus, significant nonadditive trait contributions were more common but weaker. The QTLs were confirmed using linear mixed models that separated additive, dominant and epistatic effects (Methods). In almost all cases, nonadditive QTLs coincided with dominance effects (Fig. 4.3a). The complete recessiveness of the WA GAL3 allele for galactose growth and of the WA DAL alleles for allantoin growth recapitulated established knowledge (Warringer et al., 2011; Ibstedt et al., 2015)(Supplementary Fig. 6a)

Only 32 of 145 (22%) additive and nonadditive QTLs called were mapped in a single environment, reflecting that extensive pleiotropy is the rule rather than the exception (Fig. 4.3d). Almost half (50 of 113, 44%) of the pleiotropic QTLs affected at least five environments, with universal growth QTLs on chr. XIII penetrating regardless of the environment and one QTL on each of chr. IX, X and XV penetrating in all but one environment (Fig. 4.3d). Given the wide span of environmental effects on growth and cellular physiology in our set of environments, this prevalence and penetrance of universal growth QTLs is remarkable. A surprisingly large number of QTLs (69%) were shared between growth rate and mean growth, given that the overall correlation between these growth variables was low (mean r = 0.27, Fig. 4.1d). This was to a large extent explained by the near universal chr. IX QTL affecting the two fitness components antagonistically: NA homozygotes grew slower but reached higher mean growth (Supplementary Fig. 6b). This profound fitness tradeoff penetrated regardless of environment and may therefore have had a large influence on natural selection on the ancestral wild strains. Finally, we note that disproportionately many (28 versus 9% expected, Fisher's exact test, P<0.0001) QTLs were subtelomeric; almost all (84%) of these were pleiotropic. This agrees with previous haploid studies, and adds credibility to the suggestion that hypervariable subtelomere structures and ORF compositions account for much of the remarkably large trait variation in yeast (Bergström et al., 2014; Cubillos et al., 2011).

Explaining heterosis by intralocus interactions The degree to which offspring phenotypes deviate from the mean of the parent phenotypes, heterosis, and which genetic factors that account for this difference are central questions in breeding. Capitalizing on the scale of our screen

(120,000 offspring traits), we established the phenotype discordance of the POLs from those inferred for their diploid parents (Methods) with previously unattainable completeness. Hybrid offspring where the inferred parents differed significantly from each other were retained for discordance analysis (Supplementary Fig. 7a). The majority of such offspring (89 to 95%, depending on threshold) that could be unambiguously called deviated significantly from the midparent and were thus midparent heterotic (Methods). Depending on the threshold 23–41% of these cases corresponded to the offspring being either superior (best parent heterosis) or inferior (worst parent heterosis) to both parents, with equal prevalence of best parent and worst parent heterosis (Fig. 4.4a). This is surprising given that earlier studies on non-recombined F1 diploids have indicated much higher prevalence of best parent heterosis than worst parent heterosis (Plech et al., 2014; Zörgö et al., 2012). In these earlier studies, all recessive loss-of-function alleles are compensated for and can contribute to best parent heterosis because diploid hybrids are complete heterozygotes. In our POLs, however, polymorphic sites are often homozygotic and recessive negative effects are therefore not always compensated for, explaining at least part of the difference.

Overdominance (heterozygotes at a locus being superior to both homozygotes), dominance (heterozygotes at a locus differing from the mean of the homozygotes) and epistasis can all contribute to best parent heterosis. However, calling such contri-

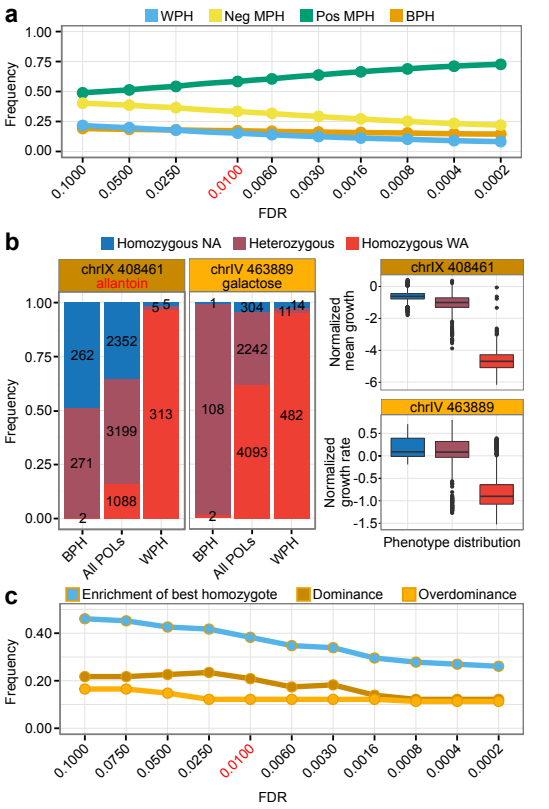


Figure 4.4. Explaining heterosis by intralocus interactions. (a) Frequencies of the heterotic POLs (*y*-axis) as a function of a range of FDR significance cut-off (*q*) values (*x*-axis). Line colour = type of heterosis. Red text = FDR *q*-value chosen for downstream analysis (a,c). (b) Left panel: example of QTLs called as contributing to best parent heterosis by dominance (dark orange) and by overdominance (light orange) respectively. Dominance was called as enrichment of strongest homozygote and overdominance as enrichment of heterozygous state among BPH POLs as compared with all POLs (left panel). Right panel: phenotype (top: allantoin, bottom: galactose) distribution depending on genotype composition at the same QTLs. (c) The frequency of QTLs called as contributing by enrichment of the best homozygote, dominance and overdominance respectively (*y*-axis) as a function of FDR significance cutoff (*q*) values (*x*-axis). The dominance contribution is a subfraction of the contributions from enrichment of the best homozygote. Note: we show the outcomes of a range of FDR cut-off values to illustrate the robustness of conclusions; the cut-offs used for downstream analysis was set beforehand and not influenced by the results. Best parent heterosis (BPH); mid parent heterosis (MPH); worst parent heterosis (WPH).

butions is challenging because multiple effects often act in parallel. In particular, overdominance may be modified by epistasis such that it only manifests in a minority of genetic backgrounds (Shapira et al., 2014). Thus, a QTL may not be overdominant in the average genetic background, but could nevertheless account for best parent heterosis in some lineages. Comparing the mean phenotypes for heterozygous and homozygous genotypes is therefore a blunt tool for detecting overdominant contributions to best parent heterosis. We devised an alternative approach, which consists of comparing the relative proportions of the genotypes among best parent heterotic POLs and the entire population of POLs. Overdominance contributions to best parent heterosis should manifest as overrepresentation of heterozygotes among best parent heterotic POLs, with no overrepresentation of either of the homozygotes. Similarly, dominance contributions should manifest as overrepresentation of the best homozygote, coupled with an unchanged or overrepresented heterozygote. Using the 115 QTLs unique to either the additive or nonadditive scan, we called overdominance contributions as more heterozygotes than expected among best parent heterotic POLs coupled with expected or less homozygotes χ^2 test, P<0.01; Fig. 4.4b, light orange, left panel), and dominance contributions as more of the better homozygote than expected coupled with expected or more heterozygotes (Fig. 4.4b, dark orange, left panel). We found 44 QTLs (38%) enriched for the best homozygote genotype, and in 24 of these the heterozygote genotype was either enriched or unchanged, suggesting dominance at these 24 loci. For 14 QTLs (12%) we found overdominance contributions. These proportions were consistent across a wide range of significance cut-offs (Fig. 4.4c). For the remaining 50% of QTLs, no significant contributions to the best parent heterosis were detected.

The dominance/overdominance contributions of QTLs to best parent heterotic POLs were often notably different from their contributions to the population as a whole (Fig. 4.4b). Only two of the 14 QTLs for which we detected overdominance in the best parent heterotic POLs had, on average, a significantly superior heterozygote state when the entire POL population was considered (Student's t-test, P<0.01). This suggests that dominance-by-dominance or dominance-by-additive interactions potentiate the best parent heterosis by shifting dominant or additive loci to overdominant, creating best parent heterosis, in a minority of backgrounds. For the chr. IX QTL with a near universal fitness trade-off, NA/WA eterozygotes were consistently enriched among offspring with superior growth rate, implying overdominance (Supplementary Fig. 7b). This was not the case for offspring with superior mean growth, where we instead found strong enrichment of the NA/NA homozygote, but near depletion of the NA/WA heterozygote. Finally, we called underdominant contributions to worst parent heterosis as more heterozygotes than expected among worst parent heterotic POLs. Overall, we found 7% of QTLs to contribute underdominantly to worst parent heterosis

(Supplementary Fig. 7c). We also called 39% of QTLs with dominant contributions to worst parent heterosis as more of the worst homozygote state than expected coupled with an enriched or unchanged fraction of the heterozygote state. To our knowledge, this is the most exhaustive dissection of heterosis to date.

Discussion

Traits have been exhaustively mapped and decomposed in haploid models (Bloom et al., 2015, 2013; Young and Durbin, 2014; Ehrenreich et al., 2010; Lorenz and Cohen, 2012) but extrapolation from haploid screens to the biology of diploids is precarious. Haploid designs cannot be used to measure intralocus interactions in the form of dominance, further, they only capture additive-by-additive epistasis. Moreover, ploidy has a fundamental impact on traits (Zörgö et al., 2013), both due to its influence on cell size and the masking of recessive alleles in diploids (Gerstein and Berman, 2015; Gerstein and Otto, 2009). The Phased Outbred Lines (POLs) presented here circumvent the shortcomings of haploid screens by offering decomposition of diploid traits with previously unattainable exhaustiveness. The capacity of the approach follows from generating a very large array of fully phased diploid genomes based on short read sequencing of only a moderate number of haploids. The alternative, acquiring phased genomes from direct sequencing of diploids, would require long-read sequencing of thousands of isolates and will remain economically unfeasible even in model organisms for years to come (Chaisson et al., 2015). As a direct consequence of our experimental design, each POL shares one haploid genome with siblings spawned from the same haploid parent. This sharing of half a genome had surprisingly large effects on trait similarity, greatly aiding both trait prediction from relatives (Märtens et al., 2016) and the partitioning of trait variation into its additive, dominant and epistatic components. In contrast, it somewhat restricted our ability to distinguish the weaker effects of individual loci and the calling of those QTLs. The large impact that sharing one haploid genome has on trait similarity among diploids, and the associated benefits and drawbacks, may or may not manifest in other model organisms. Beyond the removal of the sex-switch (HO gene) and introduction of sex-specific auxotrophic markers, POLs impose no requirements on the yeast genotypes used; the design is lineage agnostic. However, removal of the yeast sex-switch renders the cross directional and prevents the construction of a full diallel cross, something that is otherwise possible in for example monoecious plants where individuals express both sexes. The diploid hybrids have identical marker composition, avoiding growth effects derived from artificial auxotrophies that confound many haploid crossing designs (Perlstein et al., 2007; Mülleder et al., 2012).

The framework allowed partitioning diploid trait variation into its major components with little room for confounding effects, due to nearly all trait variation being accounted for. Additive effects

explained the vast majority of phenotypic variation, with approximately equal variance contributions from dominance and pairwise interactions at around 10% and 7%, respectively. The large explanatory power of additive genetics is well in line with findings in haploid screens (Bloom et al., 2015; Lorenz and Cohen, 2012). Third order epistasis explained <2% of the trait variation, comparable to, or somewhat less than, estimated for third (Bloom et al., 2013), or third and higher (Young and Durbin, 2014) order interactions in haploid yeast. Thus, although examples where three-way interactions affect trait variation can be found (Young and Durbin, 2014; Gerke et al., 2009; Taylor and Ehrenreich, 2015), and can explain extreme phenotypic outliers (Forsberg et al., 2016) they generally account for little trait variation. Despite the lower overall contribution of nonadditive compared with additive genetics to trait variation, we found nonadditive QTLs to outnumber additive QTLs. The weaker mean effect of nonadditive QTLs partially explains this discrepancy. In addition, differences in how QTLs were called means that we cannot completely exclude that we detected nonadditive effects with somewhat better power.

A stable haploid phase, indefinite storage as frozen stocks and easy mating will remain distinct advantages of yeast. Nevertheless, POLs can be employed in most higher model organisms, with only slight modifications to the approach. Panels of extensively recombined offspring can be generated using two or more founder parents in mouse, plants, flies and worms (Nordborg and Weigel, 2008; Mackay, 2014). Successive inbreeding or selfing is common practice to produce recombinant inbred lines (RILs). The gametes of these sequenced RILs can be paired by designed mating to generate the final array of POLs to be phenotyped. Somewhat analogous approaches exploiting near isogenic lines, or immortalized F2 populations, have been used in plants (Melchinger et al., 2007; Tang et al., 2010; Hua et al., 2003), although few individuals, genetic markers and recombination events and remaining segregating heterozygosity prevented both powerful decomposition of trait variation and highly resolved mapping of QTLs. Furthermore, genome phasing information in POLs derived from higher organisms is ideal for investigating parent-of-origin contributions to complex trait variation (Mott et al., 2014). To attain exhaustiveness while avoiding confounding effects from uncontrolled environmental variation, the cost-effectiveness of the genotyping needs to be matched by a phenotyping approach that achieves both scale and accuracy. The here reached broad sense heritability, with a lower bound mean estimate of 91%, may remain challenging to match in most species. Nevertheless, phenomics is advancing on broad fronts and simultaneous high throughput and accuracy is on the horizon in most model organisms (Hancock et al., 2014).

Methods

Generation of phased outbred lines F12 outbred lines were derived from a multi-generation two way intercross between an-

cestors of the North American (YPS128) and West African (DBVPG6044) populations, as described (Parts et al., 2011). Ancestral strains differed at 0.53% of nucleotide sites (Liti et al., 2009). Following random sporulation of F12 diploids, 86 stable haploids of each mating type were randomly isolated and their mating type and auxotrophies determined. Haploid genotypes were selected to allow systematic crossing: *MAT***a**, *ura3::KanMX*, *ho::HygMX* and *MATα; ura3::KanMX; ho::HygMX; lys2::URA3*. Haploids of different mating types were robotically mated on rich medium (1% yeast extract, 2% peptone, 2% glucose, 2% agar) in all pairwise combinations combining their complementary *LYS* and *URA* auxotrophies using a RoToR HDA robot (Singer Ltd, UK). Haploid cells of the same mating type do not mate and this feature prevents the construction of a full diallel cross (e.g., *MATα*/*MATα* and *MAT***a**/*MAT***a** diploid hybrids cannot be constructed). Diploid hybrids were selected twice on Synthetic Minimal (SM) medium (0.14% Yeast Nitrogen Base, 0.5% ammonium sulphate, 2% (w/v) glucose and pH buffered to 5.8 with 1% (w/v) succinic acid, 2% agar). The theoretical maximum amount of POLs from our experimental design was 7,396 (86 × 86); however, one F12 haploid strain (*MATα*, number 45) was contaminated prior to mating and all 86 hybrids spawning from this cross were therefore discarded (86 *MAT***a** × 85 *MATα* = 7,310 were retained). Furthermore, 8 F12 haploids were identified as having chr. IX aneuploidy (see Genotype construction below), the hybrids spawning from these haploids were included in the phenotyping in order to investigate the aneuploidy's effect on the phenotype. They were, however, excluded in all downstream analysis since they could interfere with the QTL mapping and they have a large fraction of missing genotypes on chr. IX. We do find a possible effect of the chr. IX aneuploidy mainly on the mean growth phenotype (see Supplementary Fig. 1a, bottom panel).

Genotype construction The haploid F12 parents were previously sequenced by short read sequencing, and mapped to the S288C reference genome in order to call segregating sites, infer genotypes and characterize the recombination landscape (Illingworth et al., 2013). All segregants were homoplasmic, carrying the same non-recombined WA mtDNA genome. This excludes confounding mtDNA inheritance effects since this is inherited randomly in a yeast hybrid from only one of the two parents. Chr. IX aneuploidy was identified based on higher sequencing coverage and higher fraction of heterozygous polymorphic sites compared with the genome as described in Cubillos et al. (2013). The following eight haploid F12 parents carried the aneuploidy: *MATα* 41, 53, 67 and *MAT***a** 206, 222, 223, 253, 258. Contaminated diploid hybrids and hybrids with chr. IX aneuploidies were excluded. Phased genomes of the 6,642 diploid hybrid offspring (81 *MAT***a** × 82 *MATα*) retained for the genetic analysis was constructed *in silico* using custom R code.

High resolution growth phenotyping High resolution growth phenotyping on

solid agar medium was performed using a 1536-colony plate layout. Each plate (Plus plate, Singer Ltd, UK) was cast with exactly 50ml of Syntetic Complete medium at 50°C (as SM above with added 0.077% Complete Supplement Mixture (CSM, Formedium)). Casting was performed on an absolutely leveled surface with drying for ~1 day. The base medium was supplemented with additional stressors or alternative carbon or nitrogen sources as indicated (Supplementary Table 1). The 7,310 POLs were distributed over 1,152 positions across eight plates. We used $n = 4$ replicates for each experimental plate, with replicates initiated from two different precultures and run in different instruments and plate positions to minimize bias. Their 172 haploid F12 parents ($n = 6$ replicates on each plate, two plates) and their diploid NA and WA ancestral lineages ($n = 72$ replicates on each plate, two plates) were phenotyped separately. Every 4th position was reserved for internal controls (diploid NA ancestral strains). These 384 controls were interleaved with experiments on preculture plates, ensuring equal treatment of controls and experiments. High resolution population size growth curves were obtained using Epson Perfection V700 PHOTO scanners (Epson corporation, UK) and the Scan-o-matic framework (Zackrisson et al., 2016). Scanners were maintained in a 30°C, high humidity environment that minimized light influx and evaporation. Experiments were run for 72h, with automated transmissive scanning and signal calibration in 20min intervals. Calibrated pixel intensities were transformed into population size measures by reference to cell counts obtained by optical density measurements on diluted samples. Raw population growth curves were slightly smoothed using a median (size = 5) and a Gaussian (width $\sigma = 1.5$) filter to remove noise. Poor quality curves (1%, descending from, for example, positions lacking colonies) were rejected following manual inspection (Zackrisson et al., 2016). Retained population growth curves were broken down into two growth phenotypes: (i) growth rate, extracted using linear regression from the steepest slope of the population's exponential phase, and (ii) mean growth, extracted as the area under the curve relative to its starting point but excluding the three first time points. To counter spatial bias on each 1,536 plate, the two growth phenotypes were normalized to the internal controls using the Scan-o-matic principle (Zackrisson et al., 2016). The final phenotypes used were the average phenotype across all replicates. Detailed protocols are available for the entire phenotype acquisition (Zackrisson et al., 2016). To circumvent the problem of calculating Coefficients of Variation (CoV) for normalized growth phenotypes spanning over both negative and positive values, these were reverted back into actual doubling times and yields, before CoV calculations. This reversion was performed by multiplying each normalized value with the median control trait value and reversion of the log transformation.

Phenotype variance partitioning We estimated additive relatedness from genotypes. We derived formulae for efficient

computation of the covariance due to dominance, pairwise and third order interaction effects (Supplementary Note 1). We fitted the model using restricted maximum likelihood, as in Yang et al. (2011). The variance decomposition and its associated standard errors were found to be accurate and close to unbiased in simulations when fitting additive, dominance, and pairwise interaction components (Supplementary Note 1). However, when adding a component for third order interactions, the overall variance decomposition became biased, even though the estimates of the third order component did not. We believe this may be the result of non-convexity in the optimization problem, as evidenced by bimodality in the distribution of estimates of pairwise interaction variance in simulations including the third order component. We therefore report estimates of the variance from third order interactions separately from the decomposition into additive, dominance and pairwise interaction components.

QTL mapping QTL calling was performed using the `scanone` function with the `marker regression` method in R/qtl (Broman and Sen, 2009) with estimated diploid parent phenotypes (additive genetic background contribution to traits) and POL deviations from the estimated diploid parents values (variation not explained by additive effects of parental background) respectively using the full set of 52,466 markers (including redundant markers). Diploid parental phenotypes were estimated as the median of all hybrids that descended from that parent. Using the deviations from expected midparent phenotype for the POLs has the additional critical benefit of effectively accounting for population structure by removing the additive effect of the more similar genetic composition due to shared parents. Significance thresholds were given by permutations (× 1,000), 1.8-LOD support intervals were calculated for each QTL using the `lodint` function in R/qtl, this corresponds to the LOD support interval stated as the preferred one for intercrosses in A guide to QTL Mapping by Broman et al. Broman and Sen (2009). QTL calling by linear mixed models, also accounting for population structure, was performed and used as verification. For these, in order to test each QTL, we constructed the realized genetic relationship matrix by discarding the SNPs within the 50kb neighbourhood of the SNP under consideration; these models were fitted with LIMIX (Lippert et al., 2014). Consecutive markers having the same genotype across all individuals were removed for increased computation speed, leaving 10,726 segregating sites (Märtens et al., 2016). We accounted for population structure in the LIMIX analysis by using the genetic relationship matrix defined by $K = \frac{1}{c}XX^T$ where X is a centred and standardized genotype matrix, and the normalizing constant c is the average diagonal value of XX^T. This is in contrast to the mapping in R/qtl where we instead modified the phenotype used, as stated at the beginning of this section. QQ-plots (Supplementary Fig. 8) confirm that the linear mixed models appropriately account for population structure: apart from

the locus with the strongest effect (*DAL* and *GAL* loci, in allantoin and galactose respectively), the distribution of the rest of P values follows the expected uniform distribution under the null.

Heterosis We used a Student's t-test to detect POLs significantly deviating (α<0.01) from the mean parent phenotype, either overperforming (positive mid parent heterosis) or underperforming (negative mid parent heterosis). The parent phenotypes used were estimated from all POLs descending from the given parent as described under QTL mapping in Methods, the variance of the mean parent phenotype was set to equal that of the most variable parent. POLs deviating from the mean parent were then tested using a Student's t-test (α<0.01) for positive deviations from the strongest parent (best parent heterosis, BPH) and for negative deviations from the weakest parent (worst parent heterosis, WPH). Hybrids deviating significantly from the two parents, but not from the estimated mid-parent, was called as not deviating from the mid parent expectation. Hybrids not falling into any of the stated categories were set as ambiguous and not considered, this might manifest as for example a hybrid not being significantly different from either parent.

Genetic contributions to heterosis To test for overdominance contributions to best parent heterosis we compared the expected and observed number of heterozygous genotypes among best parent heterotic POLs (defined as above). Calling overdominance as overrepresentation of the heterozygous state with no overrepresentation of either homozygous state. This was performed for each QTL separately using a χ^2 test, 115 QTLs were used, corresponding to all unique QTLs between the additive and nonadditive QTL scan. Entries to the χ^2 test were: observed number of heterozygotes and observed number of homozygotes (summed) among BPH POLs and the corresponding expected numbers, given distributions among all POLs. A range of cut-offs for significance was tested and the stability of results across cut-offs ascertained. We cannot completely exclude that pseudo-overdominance, that is, tightly linked loci with dominance of opposite parental alleles, confuse some assignments of overdominance. However, given the small linkage regions, we expect pseudo-overdominance to be rare and the associated overestimation of overdominance to be small. We tested for dominance similarly, but pooling the weaker homozygote state with the heterozygote state and calling significant enrichment of the better homozygote among BPH POLs. If the better homozygote was enriched, and the weaker was not, cases where the fraction of heterozygous was unchanged or enriched were called as dominance. Underdominance contributions to worst parent heterosis were called as for overdominance, but as enrichments of the heterozygous genotype among worst parent heterotic POLs. Finally, dominance contributions to worst parent heterosis were called as for dominance in best parent heterosis, but as enrichment of the weaker homozygote.

Data availability All data associated with this study is available in Supplementary Information of this publication. We used R, complemented with various packages (R Core Team, 2015; Wickham, 2009, 2007, 2015, 2011; Solymos and Zawadzki, 2016), for the analyses. The associated code can be found at `https://github.com/j-hallin/y10k`, and is available upon request.

Acknowledgements

J.H. was supported by the Labex SIGNALIFE (ANR-11-LABX-0028-01) and K.M. was supported by the European Regional Development Fund through the BioMedIT project, F.S. was supported by ATIP-Avenir (CNRS/INSERM), Becas Chile, CONICYT/FONDECYT (3150156) and MN-FISB (NC120043) postdoctoral fellowships. This study was funded by the Swedish Research Council (325-2014-6547 and 621-2014-4605), the Research Council of Norway (222364/F20) to J.W.; by a Marie Curie International Outgoing Fellowship, the Wellcome Trust, and Estonian Research Council (IUT34-4) to L.P.; ATIP-Avenir (CNRS/INSERM), ARC (grant number PJA20151203273), FP7-PEOPLE-2012-CIG (grant number 322035), ANR (ANR-13-BSV6-0006-01 and Labex SIGNALIFE ANR-11-LABX-0028-01), Cancéropôle PACA (AAP émergence 2015) and DuPont Young Professor Award to G.L.

Author information

Johan Hallin & Kaspar Märtens
These authors contributed equally to this work

Francisco Salinas
Present address: Millennium Nucleus for Fungal Integrative and Synthetic Biology (MN-FISB); Departamento de Genética Molecular y Microbiología, Pontificia Universidad Católica de Chile, Casilla 114-D, 8331150 Santiago, Chile

Affiliations
Institute for Research on Cancer and Aging, Nice (IRCAN), CNRS UMR7284, INSERM U1081, University of Nice Sophia Antipolis, 06107 Nice, France
Johan Hallin, Francisco Salinas & Gianni Liti

Institute of Computer Science, University of Tartu, 50090 Tartu, Estonia
Kaspar Märtens & Leopold Parts

Wellcome Trust Centre for Human Genetics, University of Oxford, OX3 7BN Oxford, UK
Alexander I. Young

Department of Chemistry and Molecular Biology, Gothenburg University, 405 30 Gothenburg, Sweden
Martin Zackrisson & Jonas Warringer

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, CB10 1SA Hinxton, UK
Leopold Parts

Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, 1430 Ås, Norway
Jonas Warringer

Contributions
J.H., J.W. and G.L. conceived and realised the crossing design. J.H. established the resource and generated data with help from

M.Z. and F.S. J.H., K.M. analysed the data and A.I.Y. performed the variance decomposition. L.P., J.W. and G.L. supervised the project. All authors wrote and approved the manuscript.

Competing interests

The authors declare no competing financial interests.

Corresponding authors

Correspondence to Leopold Parts or Jonas Warringer or Gianni Liti.

References

Abney M, McPeek M S, and Ober C. **Estimation of variance components of quantitative traits in inbred populations.** *American journal of human genetics*, 66(2):629–650, 2000.

Bergström Anders, Simpson Jared T, et al. **A high-definition view of functional genetic variation from natural yeast genomes.** *Molecular Biology and Evolution*, 31(4):872–888, 2014.

Bloom Joshua S, Ehrenreich Ian M, Loo Wesley T, Lite Thúy-Lan Võ, and Kruglyak Leonid. **Finding the sources of missing heritability in a yeast cross.** *Nature*, 494(7436):234–237, 2013.

Bloom Joshua S, Kotenko Iulia, et al. **Genetic interactions contribute less than additive effects to quantitative trait variation in yeast.** *Nature Communications*, 6:8712–6, 2015.

Broman Karl W and Sen Saunak. ***A Guide to QTL Mapping with R/qtl***. Statistics for Biology and Health. Springer New York, New York, NY, 2009.

Chaisson Mark J P, Wilson Richard K, and Eichler Evan E. **Genetic variation and the de novo assembly of human genomes.** *Nature Reviews Genetics*, 16(11):627–640, 2015.

Cubillos Francisco A, Billi Eleonora, et al. **Assessing the complex architecture of polygenic traits in diverged yeast populations.** *Molecular ecology*, 20(7):1401–1413, 2011.

Cubillos Francisco A, Parts Leopold, et al. **High-resolution mapping of complex traits with a four-parent advanced intercross yeast population.** *Genetics*, 195(3):1141–1155, 2013.

Ehrenreich Ian M, Torabi Noorossadat, et al. **Dissection of genetically complex traits with extremely large pools of yeast segregants.** *Nature*, 464(7291):1039–1042, 2010.

Eichler Evan E, Flint Jonathan, et al. **Missing heritability and strategies for finding the underlying causes of complex disease.** *Nature Reviews Genetics*, 11(6):446–450, 2010.

Forsberg Simon K G, Bloom Joshua S, Sadhu Meru, Kruglyak Leonid, and Carlborg Örjan. **Accounting for genetic interactions is necessary for accurate prediction of extreme phenotypic values of quantitative traits in yeast**. *bioRxiv*, page 059485, 2016.

Gerke Justin, Lorenz Kim, and Cohen Barak. **Genetic interactions between transcription factors cause natural variation in yeast.** *Science (New York, N.Y.)*, 323(5913):498–501, 2009.

Gerstein Aleeza C and Berman Judith. **Shift and adapt: the costs and benefits of karyotype variations.** *Current opinion in microbiology*, 26:130–136, 2015.

Gerstein Aleeza C and Otto Sarah P. **Ploidy and the causes of genomic evolution.** *The Journal of heredity*, 100(5):571–581, 2009.

Hancock John M, Robinson Peter N, et al. ***Phenomics***. CRC Press, 2014.

Hua Jinping, Xing Yongzhong, et al. **Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid.** *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2574–2579, 2003.

Ibstedt Sebastian, Stenberg Simon, et al. **Concerted evolution of life stage performances signals recent selection on yeast nitrogen use.** *Molecular Biology and Evolution*, 32(1):153–161, 2015.

Illingworth Christopher J R, Parts Leopold, Bergström Anders, Liti Gianni, and Mustonen Ville. **Inferring genome-wide recombination landscapes from advanced intercross lines: application to yeast crosses.** *PloS one*, 8(5):e62266, 2013.

Joshi Peter K, Esko Tonu, et al. **Directional dominance on stature and cognition in diverse human populations.** *Nature*, 523(7561):459–462, 2015.

Lehner Ben. **Molecular mechanisms of epistasis within and between genes.** *Trends in genetics : TIG*, 27(8):323–331, 2011.

Lehner Ben. **Genotype to phenotype: lessons from model organisms for human genetics.** *Nature Reviews Genetics*, 14(3):168–178, 2013.

Lippert Christoph, Casale Francesco Paolo, Rakitsch Barbara, and Stegle Oliver. **LIMIX: genetic analysis of multiple traits**. *bioRxiv*, page 003905, 2014.

Liti Gianni, Carter David M, et al. **Population genomics of domestic and wild yeasts.** *Nature*, 458(7236):337–341, 2009.

Liti Gianni and Schacherer Joseph. **The rise of yeast population genomics.** *Comptes rendus biologies*, 334(8-9):612–619, 2011.

Lorenz Kim and Cohen Barak A. **Small- and large-effect quantitative trait locus interactions underlie variation in yeast sporulation efficiency.** *Genetics*, 192(3):1123–1132, 2012.

Mackay Trudy F C. **Epistasis and quantitative traits: using model organisms to study gene-gene interactions.** *Nature Reviews Genetics*, 15(1):22–33, 2014.

Magwene Paul M, Kayıkçı Ömür, et al. **Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in Saccharomyces cerevisiae.** *Proceedings of the National Academy of Sciences of the United States of America*, 108(5):1987–1992, 2011.

Märtens Kaspar, Hallin Johan, Warringer Jonas, Liti Gianni, and Parts Leopold. **Predicting quantitative traits from genome and phenome with near perfect accuracy.** *Nature Communications*, 7:11512, 2016.

Melchinger A E, Piepho H P, et al. **Genetic Basis of Heterosis for Growth-Related Traits in Arabidopsis Investigated by Testcross Progenies of Near-Isogenic Lines Reveals a Significant Role of Epistasis**. *Genetics*, 177(3):1827–1837, 2007.

Mott Richard, Yuan Wei, et al. **The architecture of parent-of-origin effects in mice.** *Cell*, 156(1-2):332–342, 2014.

Mülleder Michael, Capuano Floriana, et al. **A prototrophic deletion mutant collection for yeast metabolomics and systems biology.** *Nature biotechnology*, 30(12):1176–1178, 2012.

Nordborg Magnus and Weigel Detlef. **Next-generation genetics in plants.** *Nature*, 456(7223):720–723, 2008.

Parts Leopold, Cubillos Francisco A, et al. **Revealing the genetic structure of a trait by sequencing a population under selection.** *Genome research*, 21(7):1131–1138, 2011.

Perlstein Ethan O, Ruderfer Douglas M, Roberts David C, Schreiber Stuart L, and Kruglyak Leonid. **Genetic basis of individual differences in the response to small-molecule drugs in yeast.** *Nature genetics*, 39(4):496–502, 2007.

Plech Marcin, de Visser J Arjan G M, and Korona Ryszard. **Heterosis is prevalent among domesticated but not wild strains of Saccharomyces cerevisiae.** *G3 (Bethesda, Md.)*, 4(2):315–323, 2014.

R Core Team. **R: A Language and Environment for Statistical Computing**. *R Foundation for Statistical Computing, Vienna, Austria*, 2015.

Shapira R, Levy T, Shaked S, Fridman E, and David L. **Extensive heterosis in growth of yeast hybrids is explained by a combination of genetic models.** *Heredity*, 113(4):316–326, 2014.

Solymos Peter and Zawadzki Zygmunt. **pbapply: Adding Progress Bar to '*apply' Functions [R package version 1.2-1]**. 2016.

Tang Jihua, Yan Jianbing, et al. **Dissection of the genetic basis of heterosis in an elite maize hybrid by QTL mapping in an immortalized F2 population.** *Theoretical and Applied Genetics*, 120(2):333–340, 2010.

Taylor Matthew B and Ehrenreich Ian M. **Transcriptional Derepression Uncovers Cryptic Higher-Order Genetic Interactions.** *PLoS Genetics*, 11(10):e1005606, 2015.

Threadgill David W, Hunter Kent W, and Williams Robert W. **Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort.** *Mammalian genome : official journal of the International Mammalian Genome Society*, 13(4):175–178, 2002.

Tsaih Shirng-Wern, Lu Lu, Airey David C, Williams Robert W, and Churchill Gary A. **Quantitative trait mapping in a diallel cross of recombinant inbred lines.** *Mammalian genome : official journal of the International Mammalian Genome Society*, 16(5):344–355, 2005.

Visscher Peter M, Brown Matthew A, McCarthy Mark I, and Yang Jian. **Five years of GWAS discovery.** *American journal of human genetics*, 90(1):7–24, 2012.

Wang Qi-Ming, Liu Wan-Qiu, Liti Gianni, Wang Shi-An, and Bai Feng-Yan. **Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity.** *Molecular ecology*, 21(22):5404–5417, 2012.

Warringer Jonas, Anevski Dragi, Liu Beidong, and Blomberg Anders. **Chemogenetic fingerprinting by analysis of cellular growth dynamics.** *BMC chemical biology*, 8(1):3, 2008.

Warringer Jonas, Zörgö Enikö, et al. **Trait variation in yeast is defined by population history.** *PLoS Genetics*, 7(6):e1002111, 2011.

Wickham Hadley. **Reshaping Data with the reshape Package [R package version 1.4.1].** *Journal of Statistical Software*, 21(12):1–20, 2007.

Wickham Hadley. ***ggplot2: Elegant Graphics for Data Analysis [R package version 2.1.0].*** Springer New York, New York, NY, 2009.

Wickham Hadley. **The Split-Apply-Combine Strategy for Data Analysis [R package version 1.8.4].** *Journal of Statistical Software*, 40(1):1–29, 2011.

Wickham Hadley. **Simple, Consistent Wrappers for Common String Operations [R package stringr version 1.1.0].** 2015.

Yang Jian, Benyamin Beben, et al. **Common SNPs explain a large proportion of the heritability for human height.** *Nature genetics*, 42(7):565–569, 2010.

Yang Jian, Lee S Hong, Goddard Michael E, and Visscher Peter M. **GCTA: a tool for genome-wide complex trait analysis.** *American journal of human genetics*, 88(1):76–82, 2011.

Young Alexander I and Durbin Richard. **Estimation of epistatic variance components and heritability in founder populations and crosses.** *Genetics*, 198(4):1405–1416, 2014.

Zackrisson Martin, Hallin Johan, et al. **Scan-o-matic: High-Resolution Microbial Phenomics at a Massive Scale.** *G3 (Bethesda, Md.)*, 6:1–12, 2016.

Zörgö Enikö, Chwialkowska Karolina, et al. **Ancient evolutionary trade-offs between yeast ploidy states.** *PLoS Genetics*, 9(3):e1003388, 2013.

Zörgö Enikö, Gjuvsland Arne, et al. **Life history shapes trait heredity by accumulation of loss-of-function alleles in yeast.** *Molecular Biology and Evolution*, 29(7):1781–1789, 2012.

Zou Fei, Gelfond Jonathan A L, et al. **Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations.** *Genetics*, 170(3):1299–1311, 2005.

Zuk Or, Hechter Eliana, Sunyaev Shamil R, and Lander Eric S. **The mystery of missing heritability: Genetic interactions create phantom heritability.** *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–1198, 2012.

Chapter 5

Predicting quantitative traits from genome and phenome with near perfect accuracy

In spite of decades of linkage and association studies and its potential impact on human health, reliable prediction of an individual's risk for heritable disease remains difficult. Large numbers of mapped loci do not explain substantial fractions of heritable variation, leaving an open question of whether accurate complex trait predictions can be achieved in practice. Here, we use a genome sequenced population of 7,000 yeast strains of high but varying relatedness, and predict growth traits from family information, effects of segregating genetic variants and growth in other environments with an average coefficient of determination R^2 of 0.91. This accuracy exceeds narrow-sense heritability, approaches limits imposed by measurement repeatability and is higher than achieved with a single assay in the laboratory. Our results prove that very accurate prediction of complex traits is possible, and suggest that additional data from families rather than reference cohorts may be more useful for this purpose.

Kaspar Märtens*, **Johan Hallin*** Jonas Warringer,
Gianni Liti & Leopold Parts

Published in Nature
Communications (2016)
doi:10.1038/ncomms11512

Introduction Disease incidence can be predicted based on the health record (Dahlem et al., 2015), the family history (Do et al., 2012) or the genetic risk due to predisposing genetic variants segregating in the population (Dudbridge, 2013). Each of these sources of information carries signal about the trait, but is not sufficient for accurate prediction (Do et al., 2012; Wray et al., 2007; Kraft and Hunter, 2009). For example, the genetic variants mapped to a trait in genome-wide association studies do not estimate disease risk well, with the vast majority of the heritable variation not accounted for (Manolio et al., 2009; So et al., 2011a). Even with very large numbers of mapped alleles (Visscher et al., 2012), purely genomic prediction accuracies still lag far behind narrow sense heritability estimates (Makowsky et al., 2011).

An important question of whether this is due to paucity of data, or perhaps more fundamental limitations, can be attacked by predicting phenotypes in model organisms (Jelier et al., 2011; Mehmood et al., 2011). In particular, crosses of founders in the yeast system have circumvented many of the technical difficulties associated with human genetic analyses, and illuminated genetic basis of variation in molecular traits (Parts et al., 2014; Albert et al., 2014; Brem and Kruglyak, 2005), cellular phenotypes (Parts et al., 2011; Ehrenreich et al., 2010; Cubillos et al., 2013), missing heritability (Bloom et al., 2013) and role of interactions (Bloom et al., 2015; Taylor and Ehrenreich, 2015; Gerke et al., 2009). Genome-based prediction has successfully explained most of the trait variation in two organism phenotypes using up to five mapped alleles (Taylor and Ehrenreich, 2015; Gerke et al., 2009), and approached narrow-sense heritability accuracy in a large-scale cross (Bloom et al., 2013). For yeast, growth in various environments is an analogue of the health record, family history is approximated by phenotypes of closely related individuals, and risk variants can be mapped as for humans. Thus, we can test whether accurate phenotype prediction for more complex traits is possible in practice, and what the constraints are.

Here, we use a recent resource of over 7,000 diploid hybrid yeast strains of high relatedness (Hallin et al., 2016) to predict their growth phenotypes. Combining genetic and phenotypic data in a linear mixed model (LMM) framework, as well as using a recently introduced mixed random forest (MRF) approach, we predict growth traits with accuracies above their narrow-sense heritability, and approaching limits set by measurement repeatability. We find that both relatedness and variant-based predictions are greatly aided by availability of very close relatives, whereas information from a large number of more distant relatives fail to improve predictive performance when closer relatives are included. Our results suggest that prediction is improved by both data from closer relatives that share much of the genome, as well as additional phenotype measurements that can capture aspects of unique environment and effects too small to be detected by mapping.

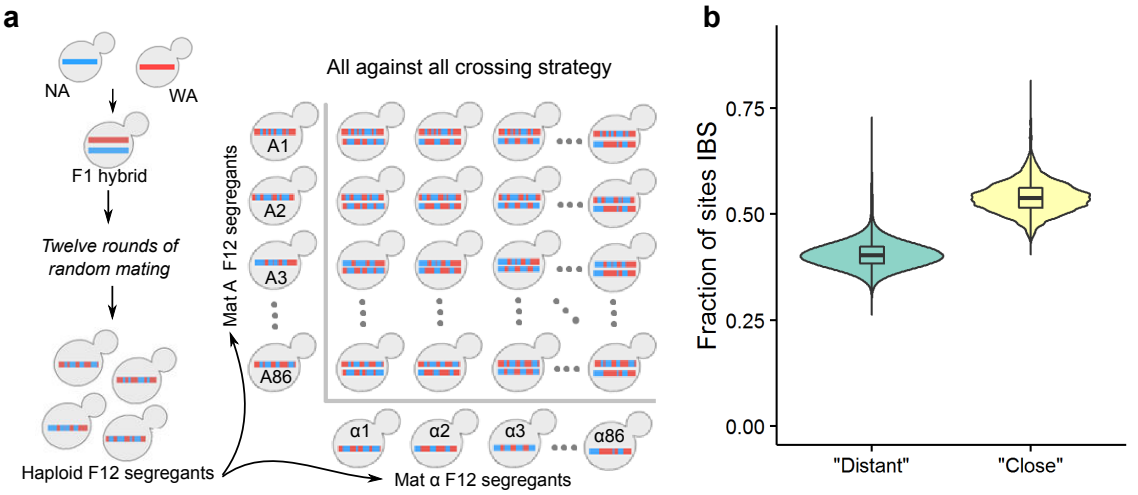


Figure 5.1. Experiment population. The 7,396 studied individuals are diploid hybrids that were constructed by systematic mating of 86 F12 *MAT***a** haploid yeast segregants to 86 *MATα* individuals, in all pairwise combinations. (a) Two-stage crossing scheme, starting from the West African (WA) and North American (NA) parents gives a large, diverse, diploid population. (b) Distribution of fraction of sites with identical genotype for pairs of hybrids is bimodal. The frequency of individual pairs that are identical by genotype state (IBS) at fraction f of the sites (y-axis) is different for pairs that share one parent ('close', right), and ones that do not ('distant', left).

Results

Study population We made use of 7,396 diploid hybrid *Saccharomyces cerevisiae* strains with phased whole-genome sequences from the collection of diploid phased outbred lines (Hallin et al., 2016). Owing to the two-stage crossing scheme (Fig. 5.1a), each of these hybrids has 170 relatives that share one chromosome in every chromosome pair (expected fraction of segregating site genotypes identical by state $f = 0.5$), and 7,225 ones for which no complete chromosome is shared, but a substantial part of linkage blocks and allele combinations are (expected $f = 0.375$, Fig. 5.1b). We refer to these levels of relatedness as 'close' and 'distant', respectively, noting that both classes correspond to close kinship. After filtering out individuals with aneuploidies and contamination, we retained 6,642 strains for analysis. Population growth of individual diploid hybrids was measured (Zackrisson et al., 2016) in nine environments in technical and biological duplicate, growth estimates were normalized against hundreds of densely spaced internal standards and the replicate average was used for analysis. The environments challenge different cellular functions, covering energy sources (for example, galactose), osmotic stress (for example, NaCl) and cancer drugs (for example, rapamycin, Supplementary Table 1). As reported before (Hallin et al., 2016), the phenotype means have large narrow-sense heritabilities (h^2) and repeatabilities (H^2, broad-sense heritability; median $h^2 = 80\%$, $H^2 = 94\%$, standard error = 0.09,

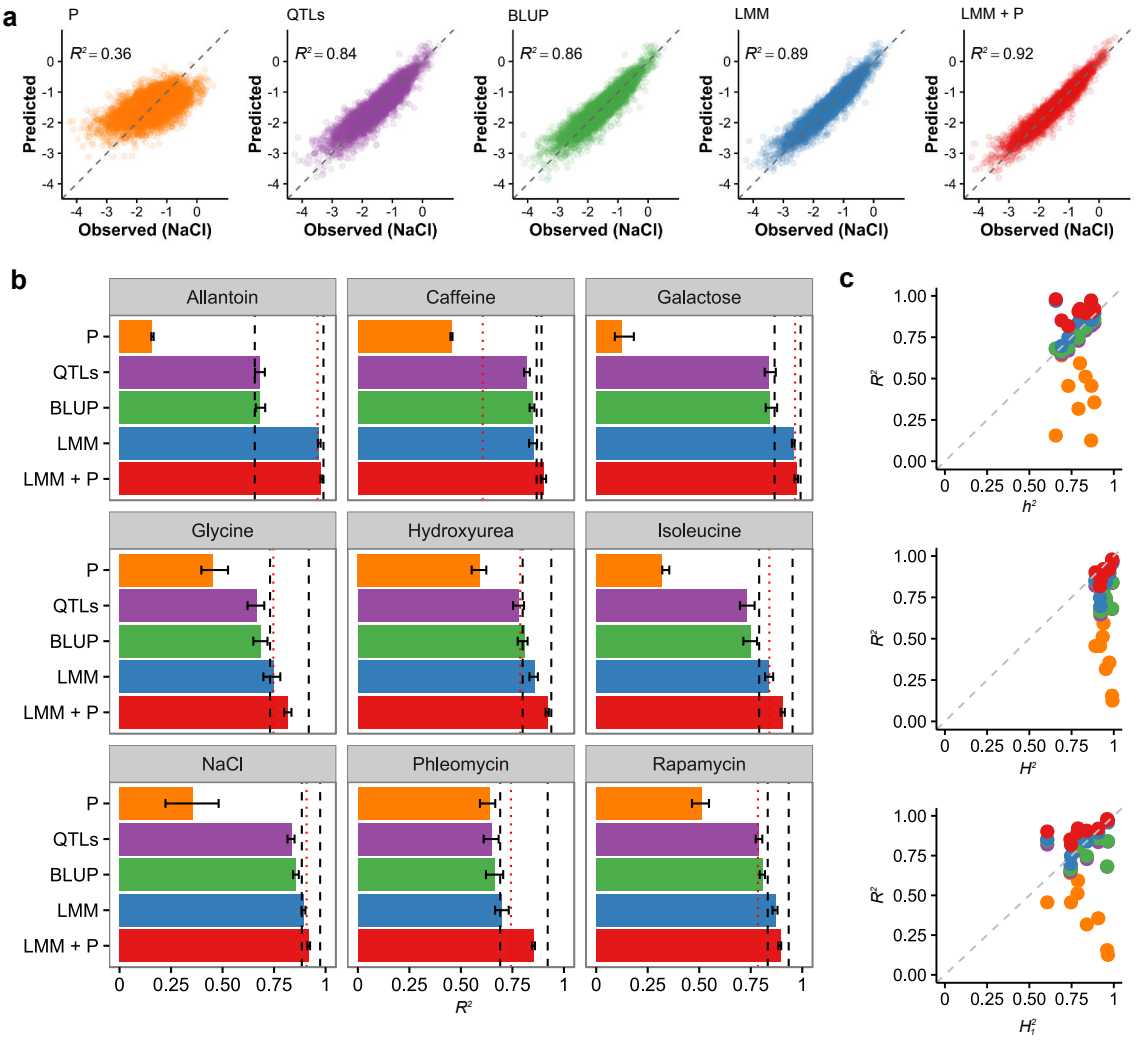


Figure 5.2. Prediction accuracy. All panels contain five model classes: linear regression on other phenotypes ('P', yellow), linear regression with additive effects determined by forward selection ('QTLs', purple), prediction based on the realized genetic relatedness ('BLUP', green), the best LMM with additive and interaction effects ('LMM', blue) and the best LMM with additive and interaction effects together with other phenotypes ('LMM+P', red). All prediction accuracies denote coefficient of determination R^2, and are determined by fourfold cross-validation. (a) Models using a single source of information predict less accurately than a combined one. Predicted (y axis) and observed (x axis) growth in NaCl for every measured hybrid strain (dots) for each model class, with coefficient of determination (R^2) of the predictions labelled. Perfect predictions would lie on the grey dashed line $y{=}x$. (b) Linear mixed models with information from other phenotypes give very accurate predictions. Predictive performance (R^2, x axis) for different models (y axis) for each of the measured phenotypes (nine boxes). Bars indicate the range of R^2 over the four cross-validation folds. The dashed lines show narrow-sense heritability h^2 (black, left) and repeatability H^2 (black, right) estimates for the mean phenotype, and the dotted line (red) shows repeatability of a single measurement H_1^2. (c) Prediction can be more accurate than one measurement. Prediction accuracy of mean phenotype (R^2, y axis) compared with different types of heritability estimates (x axis) for the four model classes: narrow-sense heritability of average phenotype (h^2, top panel), repeatability of average phenotype (H^2, middle panel) and repeatability of a single measurement (H_1^2, bottom panel). Grey dashed lines denote the identity $y{=}x$.

Supplementary Tables 2 and 3), and the traits are not independent (pairwise Pearson's r^2 = 0.01–0.49, Supplementary Fig. 1), reflecting shared genetic, epigenetic and environmental influences (Supplementary Fig. 2).

Accurate genome-aided phenotype prediction We first tested how well different genomic and phenomic data predicted growth phenotypes in our population (Fig. 5.2a) and Supplementary Fig. 3), and then combined them using LMMs (Lippert et al., 2014). We obtained predictions via fourfold cross-validation, with the training set randomly sampled from both close and distant relatives (Methods). One growth trait could be predicted from the rest with reasonable accuracy (Fig. 5.2b 'P', median R^2=0.48), and the quality of prediction depends on the strength of pairwise correlations of the phenotypes. The genomic best linear unbiased predictor (BLUP), an additive model based on realized genetic relatedness alone, captures the pedigree structure in the population, and achieves prediction accuracies very close to the narrow-sense heritability estimates (Fig. 5.2b 'BLUP', median R^2=0.77, 98% of h^2 explained). These predictions are near-identical to a simple midparent approach (Pearson's r^2>0.99, Supplementary Fig. 4). Thus, the genetic similarity between individuals explains nearly all additively heritable variation in our population.

Next, we mapped quantitative trait loci (QTLs) in each environment, and asked how well they predict growth in that environment. A small number of single nucleotide polymorphisms (SNPs) with the largest effects explain a sizeable portion of additive variance, but for all traits the prediction accuracy remains lower than BLUP's (for example, median R^2=0.58 versus 0.81 for 10 QTLs, Supplementary Fig. 5). When up to 50 SNPs are included in the model, the accuracy reaches h^2 (Fig. 5.2b, 'QTLs', median R^2=0.78, 98% of h2 explained), with predictions very similar to BLUP (r^2>0.97, Supplementary Fig. 6). Therefore, all tested methods that consider additive genetic effects reach the same, near-h^2 performance, and there is no missing narrow-sense heritability in our experiment. Extending to the LMM framework to include genetic background, dominance and interaction effects gave a modest further improvement (median increase of R^2 by 0.06), mainly due to dominance effects of strongest QTLs for allantoin and galactose (Fig. 5.2b, 'LMM', median R^2=0.86).
We then included other phenotypes measured for the same individual as covariates in the model, and achieved median prediction accuracy of 0.91 (Fig. 5.2b 'LMM+P'). To our knowledge, this is the highest for complex traits to date (de los Campos et al., 2013; Daetwyler et al., 2013), exceeding narrow-sense heritability for all nine phenotypes and approaching repeatability (Fig. 5.2c, 96% of H^2 explained). For each of the measured traits, our predictions of the mean phenotype (that is, the average of four replicate measurements) have lower error than a single growth experiment (Fig. 5.2c). The combined model improves over others especially when a large proportion of heritable non-additive variation is not

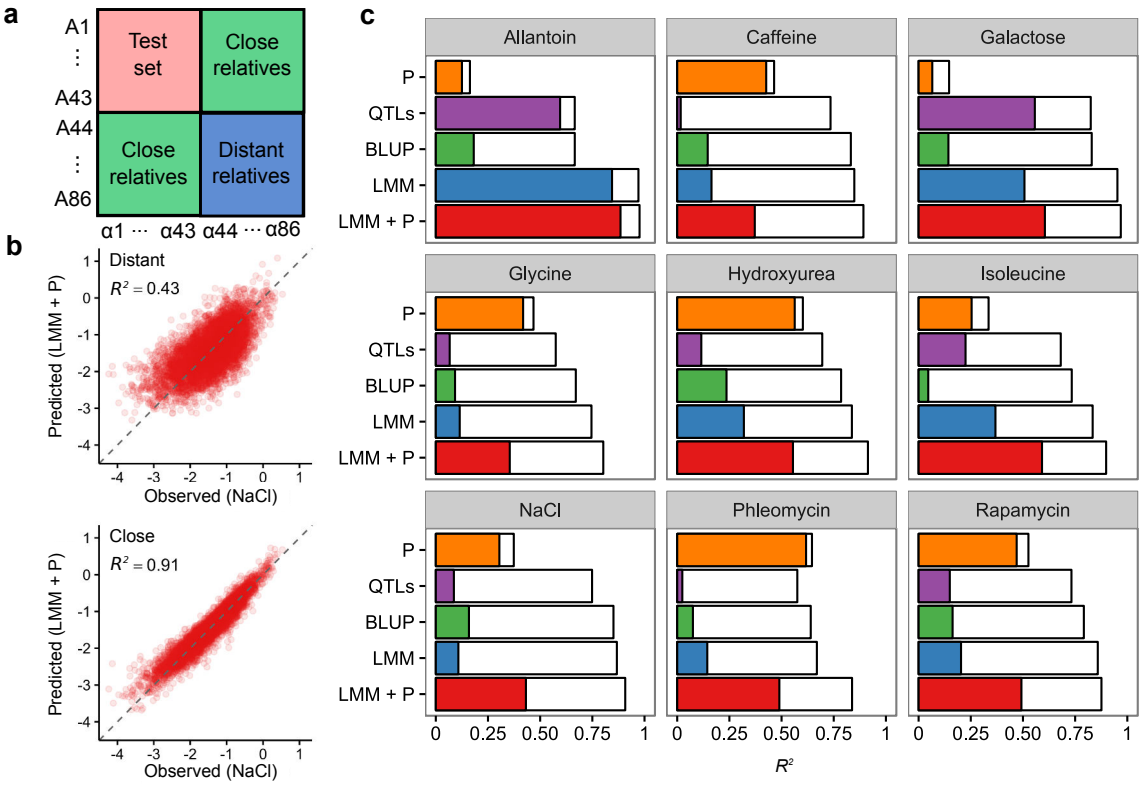


Figure 5.3. Close relatives improve predictions. (a) To cover two training scenarios, that is, fitting models on 'close' (expected fraction of sites identical in genotype f=0.5) or 'distant' (expected f=0.375) relatives, we partitioned all individuals into four equally sized groups. For a fixed test set (red box), we distinguish between training on close relatives (individuals who have a common parent with one test set individual, green box) and more distant relatives (no common parents with any test individual, blue box). As the number of close relatives is twice the number of distant relatives, we downsampled the former. Predictions are obtained by fourfold cross-validation. (b) Close relatives greatly contribute to genome-based prediction accuracy. Predicted (y axis) and observed (x axis) growth for test set individuals (red dots) in NaCl using the best LMM+P model in 'distant' (top) and 'close' (bottom) training scenarios. Grey dashed line denotes the identity y=x; coefficient of determination R^2 is labelled on the plot. (c) Distant relatives are more difficult to predict in each environment. Predictive performance (R^2, x axis) of different model classes (y axis) in two training scenarios: 'Distant' (colored bars) and 'Close' (white bars) for each of the nine environments (boxes).

captured by interaction and dominance effects (Supplementary Fig. 2).

Predictions based on closer relatives are more accurate So far, our predictions for each test individual were obtained from models that were trained with data from its close relatives that share half of the complete chromosomes. We observed that errors were larger when close relatives were not available (for example, Fig. 5.3b and Supplementary Fig. 7). Thus, we next compared two training scenarios—'close relatives', where each member of the test set has several close relatives in the training set (expected fraction of identical site genotypes f=0.5), and 'distant relatives', where test set individuals are not as closely related to anyone in the training set (expected f=0.375, Fig. 5.3a). When training on close relatives, predictions based on other traits of the same individual are slightly more accurate (median improvement=0.04, Fig. 5.3c, 'P'), whereas BLUP performs substantially better. On average, BLUP achieves R^2 of 0.14 when trained on distant relatives and 0.76 on close ones (Fig. 5.3c, 'BLUP'). This difference is explained by the larger uncertainty of the predictive distribution based on distant relatives: the observed errors are near-perfectly calibrated to their model-derived standard errors (Fig. 5.4a, r^2=0.96). Accuracy increases markedly even with a small number of close relatives included in the training data, whereas adding more distant relatives to close ones does not improve predictions (Fig. 5.4b, Supplementary Fig. 8). For example, adding on average just five close relatives per test individual rises the median R^2 from 0.15 to 0.65, but complementing the training set of close relatives by all distant relatives has a negligible effect (median R^2=0.79 versus 0.81).

Perhaps surprisingly, training on close relatives also improved QTL-based predictions. For near-monogenic traits (for example, growth in allantoin and galactose), the accuracies were similar for both training scenarios (Fig. 5.3c 'QTLs'). However, for more complex traits, the QTL model trained on distant relatives reaches high accuracy in the training data, but does not perform well out of sample, with 61% median decrease in accuracy (respective decrease for close relatives is 3%, Fig. 5.4e). In this case, the prediction uncertainties are similar (Fig. 5.4c), and most of this difference is explained by model selection. When we mapped QTLs in close relatives, but estimated their weights on distant relatives, the prediction accuracy decreased from 0.73 to 0.65 compared with carrying out both procedures on close relatives (Fig. 5.4d and Supplementary Fig. 9). Conversely, mapping QTLs in distant relatives and fitting their weights in close relatives resulted in a much lower R^2 of 0.31. Including close relatives in training gives a more faithful approximation of the phenotypic covariance structure (Supplementary Fig. 10), which explains the large gap between out-of-sample and in-sample performance for distant relatives (Fig. 5.4e). Notably, prediction accuracy drops substantially, even when just 1% of the training data changes (Fig. 5.4e, filled versus empty markers).

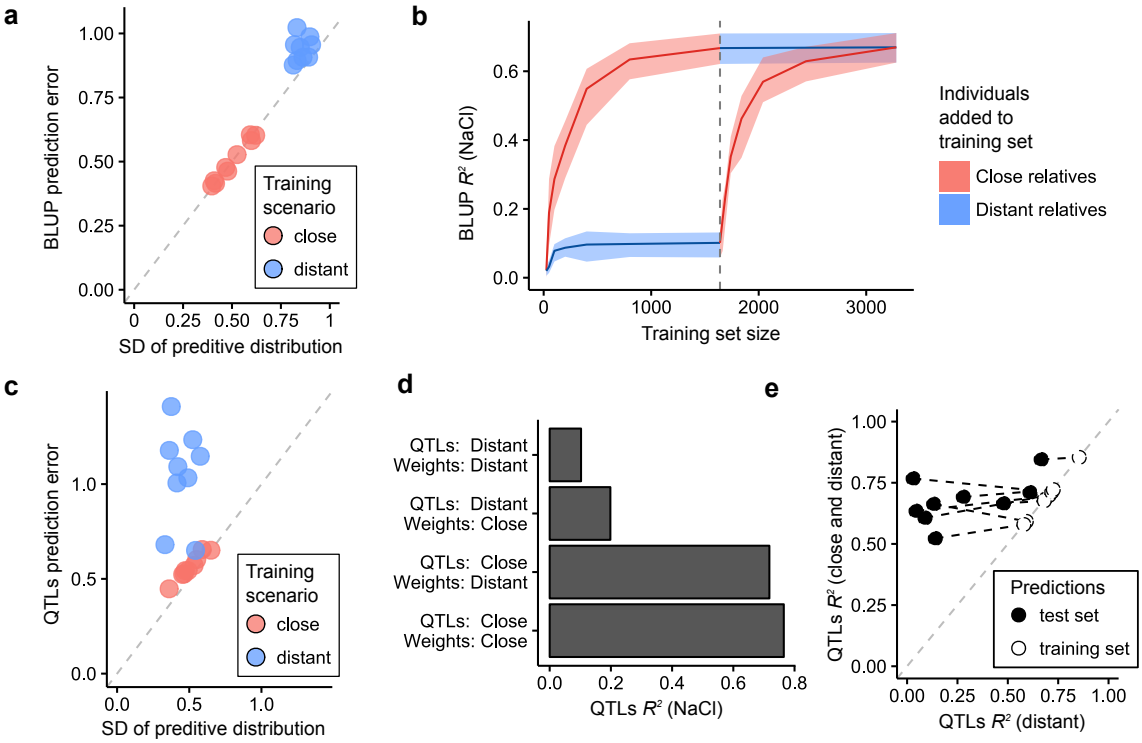


Figure 5.4. Causes of improved prediction performance for close relatives. (a) BLUP predictions from distant relatives are less accurate because of a more uncertain model-derived predictive distribution. Prediction error (y axis, standard deviation of the residuals) compared with the standard deviation of the predictive distribution (x axis) for the nine environments, when trained on distant (blue dots) or close relatives (red dots). (b) BLUP predictions are more accurate when the model is trained on a small number of close relatives compared with a large set of distant relatives. Predictive performance of BLUP (R^2, y axis) improves with expanding the training set (size on x axis) with individuals closely (red line) or distantly (blue line) related to the test set. From the dashed grey line onwards, distant relatives are added to the training set of closely related individuals, and vice versa. Shaded regions denote the range of R^2 over the four cross-validation folds. (c) Unlike for BLUP in a, the less accurate predictions from the QTL model in the 'Distant' training scenario are not in accordance with uncertainty in the model-based predictive distribution. (d) Low QTL predictive ability for out-of-sample distant relatives is mainly due to discrepancies between the sets of mapped QTLs, not their estimated effects. Predictive performance (R^2, x axis) of the QTLs model, stratified by training sets used for QTL mapping (model selection) and weight estimation (model fitting). QTL mapping and weight estimation are carried out under four training scenarios (y axis): both stages in distant relatives ('QTLs: Distant, Weights: Distant'), both in close relatives ('QTLs: Close, Weights: Close'), QTLs mapped in distant relatives and weights estimated in close relatives ('QTLs: Distant, Weights: Close'), or vice versa ('QTLs: Distant, Weights: Close'). (e) A minor change in the training set (replacing 1% of distant relatives with close ones) has a profound effect on out-of-sample QTL-based prediction accuracy. Out-of-sample (black dots) and in-sample (white dots) predictive performance (R^2) of QTLs model in two scenarios: trained on distant relatives only (x axis) or when 1% is replaced with close relatives (y axis).

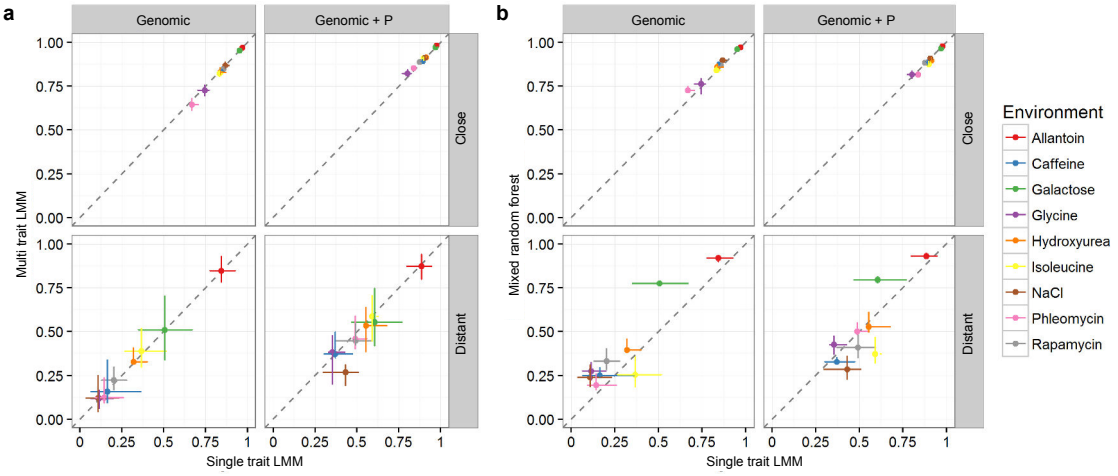


Figure 5.5. Prediction performance is similar for a range of model classes. Prediction performances of additional published methods to standard linear mixed models (LMMs), both on close and distant relatives. All results are shown for two training scenarios (close and distant relatives, panels 'close' (top) and 'distant' (bottom)) and two types of prediction: purely genomic prediction (panel 'genomic', left), and combined genomic and phenomic prediction (panel 'genomic + P', right). Both x and y axes represent the coefficient of determination R^2, and the horizontal and vertical error bars denote the range of R^2 over four cross-validation folds. (a) Multi-trait linear mixed models (MT-LMMs) perform similar to single-trait LMMs. Predictive performance (R^2) for each environment (dots with various colours) for single-trait models (x axis) and multi-trait models (y axis). (b) Mixed random forests (MRFs) perform similar to single-trait LMMs. Predictive performance (R2) for single-trait LMMs (x axis) and MRFs (y axis).

Combining genomic and phenotypic information (LMM+P) to predict from distant relatives gives accuracies similar to combining QTLs and phenotypic information. For traits where genomic prediction on distant relatives does not work well (for example, caffeine, glycine, phleomycin), this model performs similarly to using other phenotypes only or even slightly worse (median improvement 0.02, Fig. 5.3c 'LMM+P'). However, for traits with large effect QTLs (allantoin, galactose, isoleucine), genetic information helps prediction even if BLUP is not accurate.

Prediction performance is consistent for alternative models Other methods for genome-aided trait prediction have either included other phenotypes directly in the model or are compatible with doing so (Lippert et al., 2014; Stephan et al., 2015; Mrode, 2014). We confirmed that these prediction implementations give results that are concordant with ours. First, we tested the multi-trait LMM (MT-LMM) that jointly infers the effects of genotype and other phenotypes (Lippert et al., 2014). This method gave results nearly identical to the LMM+P approach on both close and distant relatives, in which we first regressed the effect of phenotypes, and then fit a genomic model on the residuals (Fig. 5.5a). Second, we applied the recently published MRF, which accounts for population structure and captures nonlinear genetic effects (Stephan et al., 2015), and can use

the other measured phenotypes as predictors. This method also performed similar to the combined LMM (median R^2 0.91 versus 0.91) for close relatives, with no consistent difference across the traits (Fig. 5.5b, top row). For distant relatives, the MRF had more accurate pure genomic predictions than a LMM for 8 of 9 traits, and when including phenotype information for both models, 4 of 9 traits (Fig. 5.5b, bottom row).

Discussion

We predicted nine heritable traits in a population of 6,642 yeast strains of varying high relatedness, and achieved accuracies over 90%, very near the repeatability limit. To our knowledge, these are the most precise out-of-sample predictions of complex traits to date. There is almost no missing narrow- or broad-sense heritability, proving that very accurate genome-aided predictions can be obtained in practice, in contrast to relatively poor genomic prediction performance for human cohorts, for example, R^2<0.16 using unrelated individuals, and <0.37 for close relatives (Makowsky et al., 2011). Our predictions outperformed the traditional mid-parent approach that is limited to narrow-sense heritability, but has been predicted to remain unsurpassed in accuracy for humans (Aulchenko et al., 2009).

The improvement in predictive ability using phenotype data is due to capturing additional signal from the non-additive genetic and environmental components, reflecting the extent to which these are shared between the traits. Their relative contribution can somewhat be gauged from the additional accuracy of the LMM+P model over the standard LMM that accounts for mapped additive, dominance and interaction effects. The improvement is largest for traits that have a large gap between narrow and broad-sense heritabilities (phleomycin, hydroxyurea, glycine, isoleucine), which is not caused by a single dominant allele (galactose, allantoin). Any remaining difference is potentially due to both weak interaction and dominance effects not included in the LMM during model selection. Standardization, distribution of replicates across multiple preculture and experimental batches, and normalization of phenotypes to very densely spaced internal controls are expected to minimize the influence of shared environmental variation across plates (Zackrisson et al., 2016). A small contribution of shared environment is consistent with the phenotypic covariance decomposition (Supplementary Fig. 2), and sizes of variance components due to the 2nd and 3rd order interactions that are difficult to map (Hallin et al., 2016; Young and Durbin, 2014). Although we cannot completely exclude that a small fraction of the phenotype covariance reflects shared environmental variation, for example, in the form of nutrient access, initial population size or exposure to stress, the residual covariance has been empirically demonstrated to be smaller than our prediction improvements for most traits (Zackrisson et al., 2016). Regardless, additional measured phenotypes from the individual can clearly inform on all these sources of variation, circumvent-

ing the need to explicitly ascertain their effects.

Genomic prediction methods have recently been extended to include more fine-grained decomposition of trait variances, both for phenotypes (for example, multi-trait models (Lippert et al., 2014)) and genotypes (partitioning sites by chromosome (Speed et al., 2012), allele frequency (Yang et al., 2015) or functional class (Finucane et al., 2015)). In latter group, the genetic covariance matrix is partitioned by allele category, and a BLUP model is fit for each. BLUP is a linear combination of training data, with uncertainties stemming from genetic relatedness only for prediction. Accordingly, we found that genomic BLUP estimates became uncertain when closer relatives were unavailable (Fig. 5.4a), and prediction error increased. This source of error is not circumvented by the partitioning methods, as the relatedness-derived uncertainty remains, and therefore these approaches are unlikely to improve our suboptimal predictions for more distant relatives.

It is important to note that our study population does not share many of the features of human cohorts. We used data from a diallel cross, in which only two alleles are present at any locus, and their frequencies are close to 50%; there is no spectrum of low frequency and rare alleles. Further, due to the controlled phenotyping design, there is little environmental variation and the heritability estimates in our populations are therefore very high. Although this is atypical for most human traits, our results concern prediction accuracies relative to the heritabilities, regardless of their numerical value. Finally, human complex traits can be influenced by hundreds if not thousands of loci. Nevertheless, their combined predictive ability has remained far below the narrow-sense heritability estimates. We capture nearly all of the broad-sense heritability with the most precise models, demonstrating that knowledge of additional phenotypes helps estimate the combined influence of small effect alleles and interactions that are difficult to map. Therefore, making use of the accumulated personal phenotype data is also expected to improve human trait prediction.

When no very close relatives were available, and no single QTL explained a large fraction of variance, the pure genomic methods were inaccurate, even in our population of 6,642 individuals with high relatedness. At the same time, when the number of very close relatives in the training sample was sufficiently large, the predictions were not improved by adding all remaining more distant relatives. Thus, observing phenotypes for parental haplotypes in at least a few cases causes BLUP to upweight their contributions, and for QTL mapping to prioritize alleles that capture their signal. In concert, these observations suggest that efforts directed towards creating genotype-based scores using common variants to predict disease risk could benefit dramatically from being complemented by systematic collection of family history and relatedness data (Aulchenko et al., 2009; So et al., 2011b; Guttmacher

et al., 2004). As information from as few as five close relatives gave large gains, we expect such an approach to be a cost-effective solution for achieving better prediction in a clinical setting with finite resources.

Methods

Panel design and phenotyping 172 haploid F12 segregants (86 Mat**a** and 86 Matα) from a cross between YPS128 and DVPBG6044 ((Illingworth et al., 2013)) were crossed in an all against all fashion to obtain 86 × 86=7,396 diploid hybrids using standard yeast protocols (Fig. 5.1). After removing strains spawning from one contaminated and eight aneuploid haploid founders, we were left with 81 × 82=6,642 crosses for analysis. The strains were grown in biological and technical duplicates (four measurements total) in 1536-position solid agar plate cultures, with all replicates on different plates and taken from two different pre-cultures to reduce systematic bias. Medium preparation, plate pouring, robotic pinning and pre-culture and experimental conditions were all extensively standardized to reduce systematic bias. Every fourth position was occupied by genetically identical internal controls in the form of the reference YPS128 strain, and the 384 controls on each plate were used to remove any remaining bias by normalization. Although complete randomization with respect to all known confounders (for example, plate position, fixture position, machine, pre-culture, temperature, humidity, neighbouring colony size, amount of light) and unknown sources of bias is not feasible, the dense grid of reference strains provides an excellent standard. We extracted the area under the growth curve relative to the starting point in each of the nine environments, converted the values to log-scale, and normalized them to a surface constructed from the surrounding internal YPS128 controls, as described earlier (Zackrisson et al., 2016). The four replicate values were then averaged to obtain the final phenotype (that is, mean growth) for each individual and environment. Panel design, genotyping, phenotyping and normalization are described in detail in Hallin et al. (2016) and Zackrisson et al. (2016).

Modelling and predictions We used a range of models to predict a trait of interest either on genomic information only, individual phenotypic information only or both.

Phenotype ('P'). Let y be the vector containing the phenotype of interest for all N individuals, and let $P_1,...,P_8$ be the remaining phenotypes. We modelled y as $y \sim N(\beta_0 + \beta_1 P_1 + \cdots + \beta_8 P_8, \sigma^2 I)$ to fit the phenotype weights β used for prediction.

Best linear unbiased predictor. Let x_j be the genotype vector for SNP $j = 1,...,M$, and let X be the genotype matrix $X = (x_1,...,x_M)$. In the genomic BLUP model, $y = \mu 1 + \sum_j b_j x_j + \epsilon$ with random coefficients $b_j \sim N(0, \sigma_g^2)$ and measurement noise $\epsilon \sim N(0, \sigma_e^2 I)$. This model implies the multivariate Gaussian distribution, $y \sim N(\mu 1, \sigma_g^2 K + \sigma_e^2 I)$ where $K = \frac{1}{c} XX^T$ is the realized genetic relatedness matrix, with the scaling constant c being the average diagonal value of XX^T.

Prediction for the test individual can be obtained by conditioning on the observed data in a standard way for multivariate normal distributions. When calculating the standard deviation of the predictive distribution (Fig. 5.4a), we averaged the variances on the predictive distributions (that is, averaged the diagonal elements of the covariance matrix of the predictive multivariate normal distribution) and reported the square root of this number.

Quantitative trait loci. To identify the strongest QTLs, we first carried out forward selection for up to 50 iterations in the linear regression model $y \sim N(\beta_0 + \sum_{j \in Q_t})$, where Q_t denotes the selected collection of QTL indexes at iteration t. The number of QTLs in the final model was determined by out-of-sample prediction accuracy, with fourfold cross-validation on the training portion of data (hence, altogether a double cross-validation scheme).

Midparent. Let y_{ij} the phenotype for individual who has parents i and j. Let P_i^1 and P_j^2 the parental phenotype values. We model y_{ij} the mid-parent value $y_{ij} = 0.5(P_i^1 + P_j^2) + \epsilon_{ij}$, where ϵ_{ij} is uncorrelated noise. We first fit the parental values from the y_{ij} observed in training data, and used them to predict phenotypes of test individuals.

LMM with dominance and interaction effects. The LMM model combines additive, dominance and interaction effects with genetic relatedness, $y \sim N(QTLs + dom + int, \sigma_g^2 K + \sigma_e^2 I)$. The fixed effects ($QTLs$+$dom$+$int$) are constructed with forward selection among additive QTLs and interaction between all such SNP pairs x_i and x_j, where x_i has previously been selected into the model. Although we miss interactions where neither locus has a significant additive effect, it has been shown that such occurrences are rare (Costanzo et al., 2010), and their contribution to explaining variance is negligible (Bloom et al., 2015). By allowing self-interactions, we also incorporated dominance effects. We selected the final model by performing cross-validation on training data after each of the feature selection steps.

LMM including phenotypes ('LMM+P'). The $LMM + P$ model combines additive, dominance and interaction effects with genetic relatedness and other traits, $y \sim N(QTLs + dom + int + P, \sigma_g^2 K + \sigma_e^2 I)$. The fixed effects contains a genetic ($QTLs + dom + int$) and non-genetic (P) part. The latter includes the linear combination of all other traits $P_1, ..., P_8$. First, we regress y on P, and then we construct the genetic component as described for the LMM model.

Multi-trait LMM. MT-LMMs model multiple phenotypes jointly. The correlation between two traits is modelled in two parts, via a genetic and non-genetic component as follows (Lippert et al., 2014). Let $Y = [y_1, ..., y_9]$ be the matrix for phenotypes $y_1, ..., y_9$, and let F denote the fixed effects for each of these phenotypes, $F = [f_1, ..., f_9]$. We used the same fixed effects f_i that we constructed in the LMM model. Let C be the genetic covariance matrix between phenotypes and $\sum$ the non-genetic one. Then $vecY \sim N(vecF, C \otimes K + \sum \otimes I)$ according to the MT-LMM. To obtain MT-LMM pre-

dictions which correspond to the LMM+P model, we condition the multivariate normal distribution.

Mixed random forest. We applied the MRF approach (Stephan et al., 2015), available via LIMIX (Lippert et al., 2014). We ran the MRF with 25 trees and otherwise default settings. For genomic predictions (corresponding to the LMM model), we included all SNPs as potential features. For genomic and phenomic prediction (corresponding to the LMM+P model), we added also other phenotypes as potential features.

Training and obtaining predictions All models were fitted with the Python package LIMIX (Lippert et al., 2014). We used four-fold cross-validation to obtain out-of-sample predictions for all 6642 individuals. We partitioned the set of all individuals into four folds analogously as shown in (Fig. 5.3a), i.e. by splitting the two sets of parents (i.e. one in rows, the other in columns) into two equally sized groups. We use each one of these four subsets of size N^2 as a test set to obtain predictions and the remaining three as a training set to fit the models. First, we did not take into account the relatedness structure and divided individuals into subsets randomly (results in Fig. 5.2). Later, we distinguished between closely and distantly related individuals (results in Fig. 5.3). The latter correspond to siblings in a traditional sense, sharing many of the haplotype blocks (expected fraction of sites identical by state 0.375), whereas the former share one complete chromosome in each pair (expected fraction of sites identical by state 0.5). The four test sets remained the same as before, but instead of training on all $3N^2$ remaining individuals, we picked the $N \times N$ individuals who do not share a parent with anyone in the test set ('distant relatives'), as well as sampled N^2 from the $2N^2$ remaining individuals who do share one parent with someone in the test set ('close relatives').

Heritability estimation Narrow-sense heritability was estimated from the genomic BLUP model as $\sigma_g^2/(\sigma_g^2 + \sigma_e^2$, when fitted to all of the data. To estimate repeatability, we fitted the following fixed effects model $r_{ij} = y_i + \epsilon_{ij}$, where r_{i1}, r_{i2}, r_{i3}, r_{i4} are the four replicate measurements for individual i, y_i is the average r_{ij} value for this individual and $\epsilon_{ij} \sim N(0, \sigma^2)$. Repeatability was estimated as $1 - \sigma^2/Var(r)$.

Data availability The data used in this study are available in the Supporting Information of Hallin et al. (2016). Analysis code is available at `https://github.com/kasparmartens/y10k-prediction`.

Acknowledgements

We thank Francisco Salinas for technical help with large-scale crossing, Martin Zackrisson for much appreciated technical assistance with extraction and analysis of growth estimates, and Oliver Stegle, Cornelis Albers and Daniel Gaffney for comments on the text. K.M. was supported by the European Regional Development Fund

through the BioMedIT project, J.H. by the Labex SIGNALIFE (ANR-11-LABX-0028-01), Swedish Research Council (grant numbers 325-2014-6547 and 621-2014-4605) and the Research Council of Norway (grant number 222364/F20), J.H. and G.L. by ATIP-Avenir (CNRS/INSERM), ARC (grant number SFI20111203947), FP7-PEOPLE-2012-CIG (grant number 322035), ANR (ANR-13-BSV6-0006-01) and Cancéropôle PACA (AAP emergence), Labex SIGNALIFE (ANR-11-LABX-0028-01), and L.P. by a Marie Curie International Outgoing Fellowship, the Wellcome Trust and Estonian Research Council (IUT34-4).

Author information

Johan Hallin & Kaspar Märtens
These authors contributed equally to this work

Affiliations
Institute of Computer Science, University of Tartu, Tartu 50409, Estonia
Kaspar Märtens & Leopold Parts

Institute for Research on Cancer and Aging, University of Sophia Antipolis, Nice 02 06107, France
Johan Hallin & Gianni Liti

Department of Chemistry and Molecular Biology, Gothenburg University, Gothenburg 40530, Sweden
Jonas Warringer

Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås N-1432, Norway
Jonas Warringer

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB101SA, UK
Leopold Parts

Contributions
K.M. analysed the data. J.H. established the resource and generated data. K.M. and L.P. conceived and designed the modelling approaches. J.W., G.L. and L.P. supervised the project. All authors wrote and approved the manuscript.

Competing interests
The authors declare no competing financial interests.

Corresponding authors
Correspondence to Jonas Warringer or Gianni Liti or Leopold Parts.

References

Albert Frank W, Treusch Sebastian, Shockley Arthur H, Bloom Joshua S, and Kruglyak Leonid. **Genetics of single-cell protein abundance variation in large yeast populations.** *Nature*, 506(7489):494–497, 2014.

Aulchenko Yurii S, Struchalin Maksim V, et al. **Predicting human height by Victorian and genomic methods.** *European journal of human genetics : EJHG*, 17(8):1070–1075, 2009.

Bloom Joshua S, Ehrenreich Ian M, Loo Wesley T, Lite Thúy-Lan Võ, and Kruglyak Leonid. **Finding the sources of missing heritability in a yeast cross.** *Nature*, 494(7436):234–237, 2013.

Bloom Joshua S, Kotenko Iulia, et al. **Genetic interactions contribute less than additive effects to quantitative trait variation in yeast.** *Nature Communications*, 6:8712–6, 2015.

Brem Rachel B and Kruglyak Leonid. **The landscape of genetic complexity across 5,700 gene expression traits in yeast.** *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1572–1577, 2005.

Costanzo Michael, Baryshnikova Anastasia, et al. **The genetic landscape of a cell.** *Science (New York, N.Y.)*, 327(5964):425–431, 2010.

Cubillos Francisco A, Parts Leopold, et al. **High-resolution mapping of complex traits with a four-parent advanced intercross yeast population.** *Genetics*, 195(3):1141–1155, 2013.

Daetwyler Hans D, Calus Mario P L, Pong-Wong Ricardo, de los Campos Gustavo, and Hickey John M. **Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking.** *Genetics*, 193(2):347–365, 2013.

Dahlem Dominik, Maniloff Diego, and Ratti Carlo. **Predictability Bounds of Electronic Health Records.** *Scientific reports*, 5(1):11865, 2015.

de los Campos Gustavo, Hickey John M, Pong-Wong Ricardo, Daetwyler Hans D, and Calus Mario P L. **Whole-genome regression and prediction methods applied to plant and animal breeding.** *Genetics*, 193(2):327–345, 2013.

Do Chuong B, Hinds David A, Francke Uta, and Eriksson Nicholas. **Comparison of family history and SNPs for predicting risk of complex disease.** *PLoS Genetics*, 8(10):e1002973, 2012.

Dudbridge Frank. **Power and predictive accuracy of polygenic risk scores.** *PLoS Genetics*, 9(3):e1003348, 2013.

Ehrenreich Ian M, Torabi Noorossadat, et al. **Dissection of genetically complex traits with extremely large pools of yeast segregants.** *Nature*, 464(7291):1039–1042, 2010.

Finucane Hilary K, Bulik-Sullivan Brendan, et al. **Partitioning heritability by functional annotation using genome-wide association summary statistics.** *Nature genetics*, 47(11):1228–1235, 2015.

Gerke Justin, Lorenz Kim, and Cohen Barak. **Genetic interactions between transcription factors cause natural variation in yeast.** *Science (New York, N.Y.)*, 323(5913):498–501, 2009.

Guttmacher Alan E, Collins Francis S, and Carmona Richard H. **The family history–more important than ever.** *The New England journal of medicine*, 351(22):2333–2336, 2004.

Hallin Johan, Märtens Kaspar, et al. **Powerful decomposition of complex traits in a diploid model.** *Nature Communications*, 7:13311, 2016.

Illingworth Christopher J R, Parts Leopold, Bergström Anders, Liti Gianni, and Mustonen Ville. **Inferring genome-wide recombination landscapes from advanced intercross lines: application to yeast crosses.** *PloS one*, 8(5):e62266, 2013.

Jelier Rob, Semple Jennifer I, Garcia-Verdugo Rosa, and Lehner Ben. **Predicting phenotypic variation in yeast from individual genome sequences.** *Nature genetics*, 43(12):1270–1274, 2011.

Kraft Peter and Hunter David J. **Genetic risk prediction–are we there yet?** *The New England journal of medicine*, 360(17):1701–1703, 2009.

Lippert Christoph, Casale Francesco Paolo, Rakitsch Barbara, and Stegle Oliver. **LIMIX: genetic analysis of multiple traits.** *bioRxiv*, page 003905, 2014.

Makowsky Robert, Pajewski Nicholas M, et al. **Beyond missing heritability: prediction of complex traits.** *PLoS Genetics*, 7(4):e1002051, 2011.

Manolio Teri A, Collins Francis S, et al. **Finding the missing heritability of complex diseases.** *Nature*, 461(7265):747–753, 2009.

Mehmood Tahir, Martens Harald, Saebø Solve, Warringer Jonas, and Snipen Lars. **Mining for genotype-phenotype relations in Saccharomyces using partial least squares.** *BMC bioinformatics*, 12(1):318, 2011.

Mrode R A. ***Linear models for the prediction of animal breeding values***. CABI, Wallingford, 3 edition, 2014.

Parts Leopold, Cubillos Francisco A, et al. **Revealing the genetic structure of a trait by sequencing a population under selection.** *Genome research*, 21(7):1131–1138, 2011.

Parts Leopold, Liu Yi-Chun, et al. **Heritability and genetic basis of protein level variation in an outbred population.** *Genome research*, 24(8):1363–1370, 2014.

So Hon-Cheong, Gui Allen H S, Cherny Stacey S, and Sham Pak C. **Evaluating the**

heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genetic epidemiology*, 35(5):310–317, 2011a.

So Hon-Cheong, Kwan Johnny S H, Cherny Stacey S, and Sham Pak C. **Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening.** *American journal of human genetics*, 88(5):548–565, 2011b.

Speed Doug, Hemani Gibran, Johnson Michael R, and Balding David J. **Improved heritability estimation from genome-wide SNPs.** *American journal of human genetics*, 91(6):1011–1021, 2012.

Stephan Johannes, Stegle Oliver, and Beyer Andreas. **A random forest approach to capture genetic effects in the presence of population structure.** *Nature Communications*, 6:7432, 2015.

Taylor Matthew B and Ehrenreich Ian M. **Transcriptional Derepression Uncovers Cryptic Higher-Order Genetic Interactions.** *PLoS Genetics*, 11(10):e1005606, 2015.

Visscher Peter M, Brown Matthew A, McCarthy Mark I, and Yang Jian. **Five years of GWAS discovery.** *American journal of human genetics*, 90(1):7–24, 2012.

Wray Naomi R, Goddard Michael E, and Visscher Peter M. **Prediction of individual genetic risk to disease from genome-wide association studies.** *Genome research*, 17(10):1520–1528, 2007.

Yang Jian, Bakshi Andrew, et al. **Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index.** *Nature genetics*, 47(10):1114–1120, 2015.

Young Alexander I and Durbin Richard. **Estimation of epistatic variance components and heritability in founder populations and crosses.** *Genetics*, 198(4):1405–1416, 2014.

Zackrisson Martin, Hallin Johan, et al. **Scano-matic: High-Resolution Microbial Phenomics at a Massive Scale**. *G3 (Bethesda, Md.)*, 6:1–12, 2016.

The Genetic Basis for Gamete Inviability

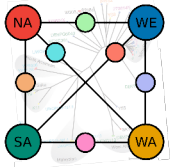
Johan Hallin[1] Jia-Xing Yue[1] Marine Poullet[1] Luca Crepaldi[2] Stephan Lorenz[2]
Jonas Warringer[3] Leopold Parts[2] Alexander Young[4] Gianni Liti[1]

SIGNALIFE · UNIVERSITÉ CÔTE D'AZUR · 1 IRCAN · 2 wellcome trust sanger institute · 3 UNIVERSITY OF GOTHENBURG · 4 UNIVERSITY OF OXFORD

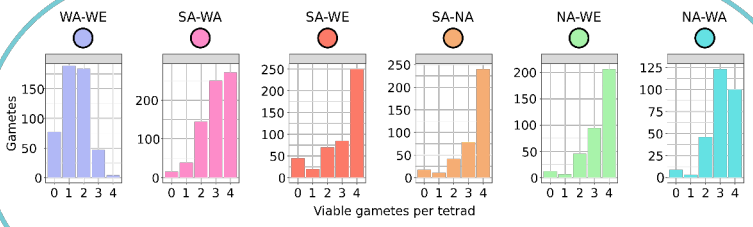
Gamete Inviability

In this project we are investigating the reasons for why gametes may be inviable or less fit. Faulty chromosome segregation and genetic interactions might render a gamete inviable but what underlies these phenomena? Does it come down to average genetic distance in SNPs, or perhaps genomic rearrangements? We do large scale, low coverage whole genome sequencing of gametes from six divergent *Saccharomyces cerevisiae* hybrids to find out. The budding yeast is the perfect model since all four resulting gametes from a meiosis can be isolated together.s

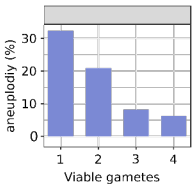
Six divergent hybrids



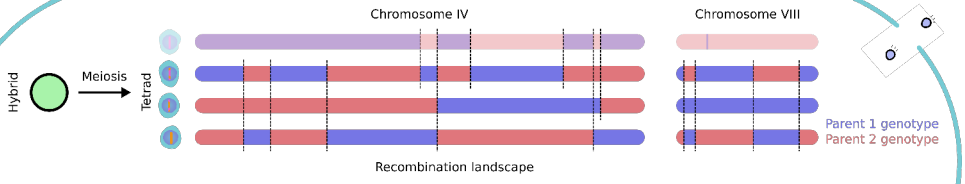
By using six hybrids from four parents with differing amounts and types of genetic variation we can correlate all our results with the underlying genetic differences between the parents to get a comprehensive view of why some gametes are inviable. Sequencing 2000 gametes from each hybrid including gametes from not fully viable tetrads ensures robust results (a tetrad being the four gametes resulting from one meiosis). In tetrads with three viable gametes we will also be able to infer the genome of the invable gamete



The gamete viability differs between the different hybrids, but what genetic differences between the parents underly this and what mechanisms underly the gamete inviability?



Few viable gametes in your tetrad? Then you are more likely to have an aneuploidy!



To give an exhaustive view of the genetics underlying inviable gametes we will investigate the recombination landscapes, QTL's and genetic interactions in the context of the genomic differences between the parents. How does an inversion affect the recombination landscape, and how does that in turn affect the viability of the gametes of the hybrids? How do the different amounts of SNPs influence the viability, and what kind of genes are incompatible when shuffled together in the gamete? How does the recombination landscape look in inviable gametes?

With our large sample size, six hybrids and full knowledge of the parental genomes, we hope to bring a complete view of the reasons for gamete inviability. This project will have implications in speciation and human birth defects as well as any other branch of biology and genetics concerned with the process of meiosis

Jia-Xing Yue *et al.* 2017 *Contrasting evolutionary genome dynamics between domesticated and wild yeasts.* Nature Genetics
Francisco Cubillos *et al.* 2011 *Assessing the complex architecture of polygenic traits in diverged yeast populations.* Molecular Ecology

Chapter 6

The genetic basis for gamete inviability – ongoing

Sexually reproducing organisms are dependent on the production of gametes for the continuation of their genetic lineage. Therefore, the ability to undergo a successful meiosis, producing viable and fully functional gametes is critical; failure to do so may result in weak or inviable offspring and the end of the lineage. Compounding on the difficulty to pass alleles on to the next generation, interactions between different alleles may also result in sub-optimal gametes. To investigate the underlying genetics behind why gametes are inviable we have constructed six hybrids spawning from crosses between highly diverged representatives of four *Saccharomyces cerevisiae* lineages. We recently published reference quality genome assemblies for the four parents and these end-to-end assemblies give us a thorough understanding of all the genetic differences in the hybrids, from single nucleotide polymorphisms to structural variation. Thanks to this, we are in a position to accurately describe how gamete viability in a hybrid is dependent on the genetic makeup of the parents. By dissecting and whole genome sequencing 2,500 gametes from each of the six hybrids, we are producing a resource of 15,000 gametes with varying viability and fitness. Using the sequence data we are exploring the impact of the recombination landscape, aneuploidies and genetic interactions on gamete inviability, and relating these phenomena to underlying genomic differences between the parents. Numbers and types of aneuploidies varied across gametes depending on parent combinations and genetic distance between parents. Aneuploidies correlate well with the gamete inviability but the majority of inviable gametes are not explained by this. We are currently exploring the effect of the recombination landscape on gamete viability and fitness, and investigating the role of allele-allele interactions.

Johan Hallin, Jia-Xing Yue, Marine Poullet, Luca Crepaldi, Stephan Lorenz, Jonas Warringer, Leopold Parts, Alexander Young, & Gianni Liti

Presented at the International Conference on Yeast Genetics and Molecular Biology (2017)

6.1 Project summary

The abstract you just read was the one I submitted to the 28th International Conference on Yeast Genetics and Molecular Biology. I was chosen to present this project as a poster and as an oral presentation during the yeast population, comparative and evolutionary genomics workshop. The number of strains to be sequenced has changed along the way and is now 1,500. As this project is still a work in progress, I will mostly share methodological aspects and summary statistics of the work that has been done so far.

Using four diverged parents (Fig. 3.2), YPS128 (NA), DBVPG6044 (WA), Y12 (SA) and DBVPG6765 (WE) we are going to investigate genetic properties of hybrid genomes. Our research team is in a great position for doing this study as we recently published end-to-end reference quality genome assemblies of these four parents (Yue et al., 2017). These complete assemblies gives a detailed view of all types of genetic variation between the parents, from SNPs to structural variation, which allows us to look at the underlying genetic factors that affect, for example, the recombination landscape.

By crossing the four parents in all possible combinations (Table 6.1) we created six hybrids with genetic divergence in both degree and kind (Liti et al., 2009; Bergström et al., 2014; Yue et al., 2017). We then pushed these six hybrid through meiosis and isolated gametes (or spores) using a micromanipulator in order to collect 2,000 viable spores from each hybrid. Using a custom made R program I documented the viability of each tetrad and the colony area (measured as number of pixels). We are underway with the sequencing of 1,500 spores from each hybrid that will be used for downstreams analysis (at the time of writing this, ~6,000 spores have been sequenced; due to time constraints 2,296 of these made it into this thesis (Table 6.4)). Sequencing is done at the Single Cell Genomics Core Facility at the Wellcome Trust Sanger Institute, we opted for a sequencing coverage that would allow us to confidently call genotypes, while keeping the cost low enough to allow for a large sample size.

All 12,000 spores that were collected were phenotyped by me at the University of Gothenburg using the Scan-o-matic methodology (Zackrisson et al., 2016) in nine different conditions.

Shortly, the main goals of this project are *i*) to characterize the recombination landscape of different hybrids and how they depend on the genetic structure of the parents. *ii*) to investigate the QTL landscape, and how it might differ from hybrid to hybrid as well as *iii*) look for genetic contributions to gamete inviability, such as aneuploidies or gene-gene interactions.

In the sections below, I will give a more detailed account of the process and progress of the project so far.

6.2 Parental strains

The four parents (NA, WA, WE, SA) were chosen for a good reason. Jia-Xing Yue recently spearheaded a project in our team

(Yue et al., 2017) (see chapter 8 for abstract) in which these strains (among others) were sequenced using PacBio (Pacific Biosciences of California, Inc.) and Illumina (Illumina, Inc) at high coverage. The long reads gained from PacBio sequencing allowed most chromosomes to be assembled into single contigs, and complex regions of the genome to be delineated.

In the context of the new project, the end-to-end high quality genome assemblies of the parents means that we have close to complete knowledge of all genetic differences between our four parents. This is important, since we can look at what genetic differences between the parents affect spore viability (for example through probability of acquiring aneuploidies during meiosis) or the recombination landscape. Additionally, we can use the parental genomes for calling genotypes rather than relying on the S288C reference genome which would inevitably miss information.

Jia-Xing Yue is the designated bioinformatician of this project and he is building and perfecting pipelines for read mapping against both parental strains to call genotypes and for classification of different types of recombination events.

Tetrad. The four gametes (or spores) resulting from one meiotic event are collectively called a tetrad.

6.3 Gamete acquisition

The six hybrids and the four diploid parents were sporulated at 23°C in 25ml liquid potassium acetate media (2%) after a 48 hour incubation in respiratory media. The amount of time spent in potassium acetate differed between the strains since they sporulate with different efficiency.

The tetrads were treated with 100µg/ml zymolyase at 37°C for between 20 and 30 minutes (depending on the strain) before being separated using a dissection microscope (Zeiss Axioskop 40), ensuring that all spores from one meiosis event (one **tetrad**) were properly isolated and could be identified as being from the same tetrad.

The tetrad dissection was performed on solid agar plates. After dissection, the plates were incubated at 30°C for three

Table 6.1. Strain genotypes. This table shows the genetic markers that are segregating in the strains used for this study, all strains are *ho::HygMX*.

	Cross	Genotype
Parents	NA-NA	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*
	SA-SA	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*
	WE-WE	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*
	WA-WA	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*
Hybrids	SA-NA	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*
	NA-WE	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*
	NA-WA	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*
	SA-WE	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*
	SA-WA	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*
	WA-WE	Mat **a**, *ura3::KanMX* – *Matα*, *ura3::KanMX*, *lys2::URA3*

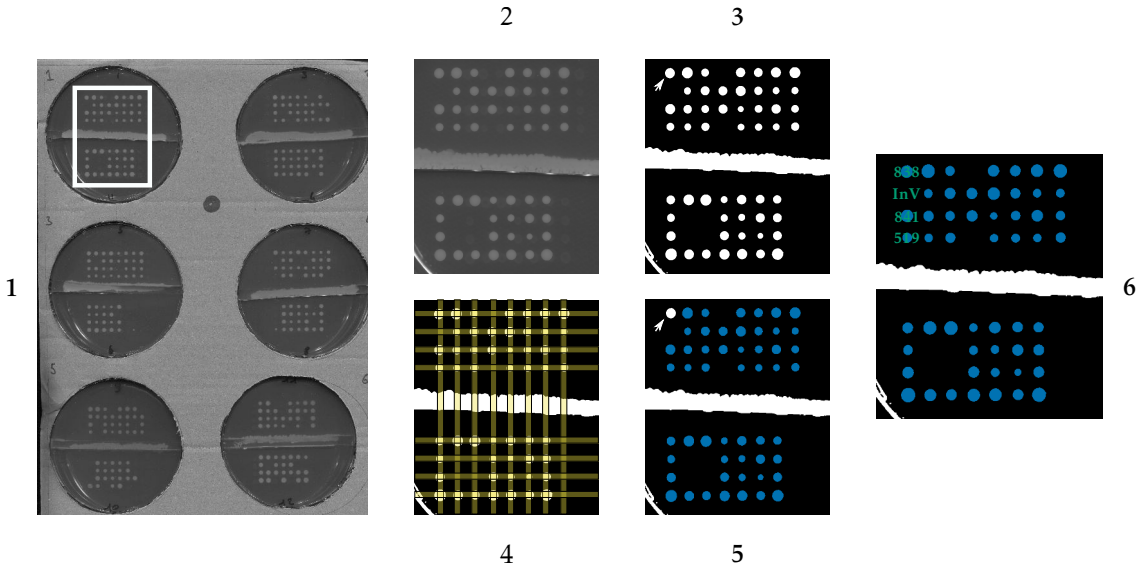


Figure 6.1. Image analysis pipeline. (1) An image is taken using a Epson flatbed scanner, the image is then processed in a program based on the EBImage package (Pau et al., 2010) for R. (2) The image is broken down so as to analyze one plate at the time, we will be following the top left plate. The cropping of the image is based on absolute values; since the plates are placed in a fixture, they will be in the same position every time an image is taken. A median filter is applied in order to smoothen the image, getting rid of any dust and noise from the plate. (3) Using an otsu threshold, the colonies are separated from the background making them possible to identify. (4) By using the courser to click on the top left colony, a grid is placed on top of the image, this grid allows me to also identify colonies that are not growing, and to know which spores are coming from the same tetrad (one column per tetrad). (5) The identified colonies are colored in blue and manually inspected. In this example, the top left colony has not been identified, if this happens you can use the courser to click on the unidentified colony and the program will find it. (6) Identified colonies are attributed with their colony size in number of pixels while inviable colonies are set to 'InV' (shown for one tetrad).

days, and subsequently scanned and the viabilities and colony sizes were documented. The tetrad dissection was a joint effort between Agnès Llored and myself.

6.4 Image analysis

All plates were scanned using an Epson Perfection V330 Photo scanner, in 8-bit greyscale at 300dpi in .tif format. The plates were placed in a custom made fixture when the image was taken, and the image analysis was performed using a custom made R program, written by me and based on the R package EBImage (Pau et al., 2010).

The analysis pipeline is detailed in figure 6.1. Once the analysis is done the colony size data is exported to a spreadsheet together with associated metadata, such as the date of dissection, and amount of days in sporulation media. Using a time-resolved phenotyping such as Scan-o-matic was unfortunately not possible due to the practical difficulties and the lack of an appropriate imaging facility at the lab in Nice. It would have been very interesting to look

at, for example, time until appearance of a colony to have a more fine-grained germination phenotype. Nevertheless, the program allowed us to easily document the viability of the tetrads, and hopefully the colony area can give us some insight into the genetics of spore germination. Due to the nature of the experiment, each colony is unique, which means that we cannot have replicates for the germination phenotype and may limit our ability to locate QTLs.

Once the colonies had been scanned, their DNA was extracted and sent to the Wellcome Trust Sanger Institute for sequencing.

6.5 Large scale DNA extraction

The ambition of this project got very tangible when the DNA extractions needed to be done; performing 9,000 DNA extractions calls for a more high throughput approach than Eppendorf tubes. During the beginning of this project I adapted the MasterPure^{TM} Yeast DNA Purification Kit from Epicentre to be used with 96-well plates, making sure that the yield and quality of the DNA was sufficient for sequencing.

Colonies from the dissection plates were scraped off and put into 2ml round bottomed 96-well plates (PP-Masterblock, Grainer bio-one), with 1ml liquid YPD. I would generally prepare four plates at a time, allowing me to perform 384 DNA extractions during one day. The plates were then incubated over night at 30°C in order to increase the number of cells. After incubation, I would take 100µl out of the plate and place in a 200µl round bottomed 96-well plate (Falcon). I would spin down the cells from the preculture (Centrifuge 5810 R, Eppendorf; 3min, 3000rpm), remove the YPD with a multi-channel pipette (Eppendorf) and then add 1ml of 25% glycerol before mixing the wells and freezing at -80°C.

The DNA extractions were then made in the 200µl 96-well plates. The main change to the original MasterPure protocol was reducing the quantities of reagents to reflect the reduced amount of cells, as well as increasing the centrifugation times due to the fact that the centrifuge taking 96-well plates does not reach the speed called for in the original protocol. The optimization of the DNA extraction protocol allowed me to scale up the amount of extractions, increase the speed with which they were done and decrease the cost, since less time and reagent was spent.

The DNA was stored in TE buffer and sent to the Single Cell Genomics Core Facility at the Wellcome Trust Sanger Institute (Hinxton, United Kingdom). Once there, it was sequenced with coverage high enough to call genotypes but low enough to allow for a large number of sequenced spores (~8x).

6.6 Growth phenotyping

Apart from the germination phenotypes we also phenotyped the growth of all 12,000 spores that were isolated using Scan-o-matic (Zackrisson et al., 2016). We chose nine different environments (Table. 6.2) based on their ability to facilitate a large spread of phenotypes, with the exception

of galactose (known from previous experiments). We also phenotyped the diploid parents and the diploid hybrids together with the spores.

Table 6.2. Phenotyping environments. All environments used for growth phenotyping with Scan-o-matic.

Environment	Note
YPD	-
Synthetic complete	-
Heat	40°C
Galactose	2%
NaCl	1.5M
$CuCl_2$	0.5mM
Caffeine	2mg/ml
Rapamycin	0.05µg/ml
Paraquat	400µg/ml

I performed all phenotyping during my stay at the University of Gothenburg between November 2017 and January 2018, with technical help from Simon Stenberg and Karl Persson in the lab of Dr. Jonas Warringer. Phenotyping was done with four replicates which were distributed over different positions in different scanners in order to minimize systematic bias. I have yet to perform the quality control and phenotype extraction, so at the moment, I only have phenotype data from my initial colony size screen to share.

6.7 Genotyping and recombination landscape

Each hybrid is the result of a cross between two different parents, which implies that all hybrids will have different segregating sites. For a given hybrid, we identified segregating sites to be used for genotyping by aligning the two parental genomes to each other, as well as aligning the Illumina reads from parent 1 to the genome assembly of parent 2 and vice versa (see table 6.3 for the amount of markers for each hybrid). Jia-Xing Yue has written this pipeline and thanks to the high quality genome assemblies from one of his projects (Yue et al., 2017) we can confidently locate SNPs in regions that are difficult when mapping sequencing reads to the S288C reference genome, such as subtelomeric regions.

By calling segregating sites using the parental genomes and by mapping the Illumina reads from the spores to both parental genomes we have greater chance to accurately call the genotypes (as opposed to basing it on the reference as is usually done (Bloom et al., 2013; Treusch et al., 2015; Hallin et al., 2016; Ziv et al., 2017)). The genotypes will be used to define the recombination landscape of these different hybrids, and again thanks to the genome assemblies of the parents we can correlate the recombination landscape with any kind of genetic variation between the parents. This will let us look at, for example, the effect of inversion on recombination.

Another interesting aspect of this project is that we are sequencing spores from tetrads that are not fully viable. I.e. from tetrads where one or more spores did not grow. Traditionally, the recombination landscape has only been looked at in tetrads where all four spores are viable (Mancera et al., 2008; Cubillos et al., 2011) which limits these studies to fully functioning meiosis.

We want to characterize successful meiosis events as well as those that do not turn out perfect. What kind of recombination landscape does a spore have where it is the only living gamete from a meiosis? This will hopefully allow us to identify patterns explaining why some meioses are dysfunctional. Furthermore, in tetrads where only one spore is inviable, we can infer its genome by matching the recombination events in the three viable gametes; giving us a window into the genome of a dead gamete. Do inviable gametes have less recombination events than viable ones?

The recombination landscape analysis is ongoing, with the pipeline being developed by Jia-Xing. Calling and classifying different types of recombination events is based on the ReCombine suite (Anderson et al., 2011) but has been re-written with slightly differing definitions and has been optimized for our two-parent approach for calling the events. As this is still ongoing, I will unfortunately not be sharing any recombination data at this point.

Faulty recombination can results in aneuploidies, we are going to use the sequence data for the spores in order to give an account of the contribution of aneuploidies to gamete inviability

6.8 Calling aneuploidies

Having more or less than the normal copy number of any given chromosome is called an aneuploidy. In gametes, they are caused by faulty chromosome segregation during meiosis and are the cause of, for example, Down's syndrome in humans (Down, 1866; Antonarakis et al., 2004).

The study by Down (1866) describing Down's syndrome is well worth a read. He wrongly attributes the hereditary origin of the syndrome to tuberculosis, but he elegantly uses the condition as an argument for the shared ancestry of all humans (seven years after Darwin's On the Origin of Species was published).

Using the coverage data from Jia-Xing's read mapping I search for possible aneuploidies. Aneuploidies should appear as an increase of reads for a specific chromosome by a factor of two and can by that logic be identified by the coverage data. I.e. the coverage would be twice as high for an aneuploid chromosome as for an euploid chromosome.

Highly repetitive regions of the genome (e.g. the rDNA on chromosome XII) were masked in the files with the coverage data

Table 6.3. Segregating sites. This table shows the amount of markers for each hybrid, the markers were called by using the genome assemblies and Illumina reads of the two parental genomes. This set of markers is not necessarily the final one and may change. "Sequenced" refers to the amount of sequenced spores that were used for the analysis in section 6.9

.

	Markers			Intermarker distance (bp)			
Cross	SNP	Indel	Total	Mean	Median	stdev	Sequenced
SA-NA	43,863	1,141	45,004	258.20	124.00	499.15	834
NA-WE	65,559	2,017	67,576	170.65	86.00	339.56	52
NA-WA	53,702	1,444	55,146	210.50	103.00	389.46	571
SA-WE	71,285	2,051	73,336	156.45	81.00	447,13	780
SA-WA	60,379	1,602	61,981	186.31	95.00	350.38	0
WA-WE	72,880	2,100	74,980	152.78	78.00	310.03	59

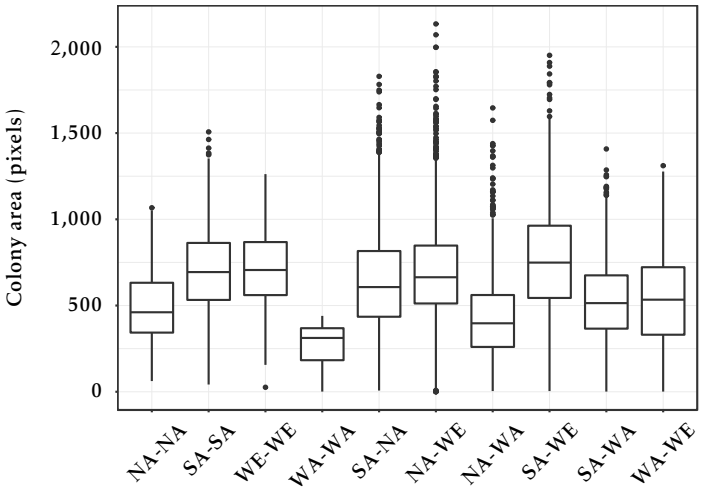


Figure 6.2. Germination phenotypes. The area of colonies, in pixels, were calculated after three days of growth using a custom made R program. Box and whisker plots show the median (horizontal line inside box), 1st and 3rd quartile (box), and the whiskers go up to the last data point within 1.5 interquartile ranges. The WA-WA parents is the weakest grower and that is reflected in that the hybrids containing this parent are generally worse growers than other hybrids. The sample size of each box corresponds to the total amount of viable spores for the parent or hybrid (i.e. Spores × Viability in table 6.4, ~2,000 for the hybrids.)

in order to not have the variation of these regions in the different strains affect the aneuploidy calling. For a given spore, the mean coverage of each chromosome was calculated as well as the mean coverage of its entire genome. Spores with an average genome coverage below 0.5 and chromosomes with an average coverage below 0.5 were excluded, to reduce false positives. Aneuploidies were called if the ratio between the chromosome and genome coverage was equal to or above 1.5. I then manually investigate each called aneuploidy by inspecting plots of the coverage of the spore.

6.9 Preliminary results

I will here share some preliminary results including the spore viability of the different hybrids, the germination phenotype and aneuploidies.

6.9.1 Spore viability and colony size

The spore viabilities are shown in table 6.4. The low viability of the WA-WE hybrid forced us to isolate a lot of its spores in order to acquire 2,000 viable spores. All parents have rather high viability as is expected, except for WA who stands out with its lower viability. This is reflected also in the lower viability of the crosses including this strain.

The weakness of the WA strain is also clear in the germination phenotype data Fig. 6.2. Furthermore, WA hybrids have on average smaller colony areas after three days of growth than hybrids not containing WA. It will be very interesting to compare these data with the growth phenotyping on YPD

to see if these data reflect a general growth defect or a defect in germination.

The amount of phenotypic variation within the homozygous diploid parents is a bit surprising. However, the fact that the haploid spores will have markers segregating that can affect growth (*ura3*, *lys2*) may explain some of this variation (Table 6.1). At the moment, we cannot exclude that this reflects noise from the germination phenotype but we intend on testing the effect of the markers to see what extent it is contributing to this variation.

6.9.2 Viability and genetic distance

Genetic divergence can effect the viability of spores by interfering with recombination during meiosis or by genetic incompatibilities between diverged genes (Greig, 2008). In figure 6.3, the viability of the different strains are plotted against the amount of markers. There may be some correlation between the viability and marker amount, however, more data points would be needed for any conclusive statements. The crosses containing the WA strain are not very efficient at producing viable gametes, this is true also for the diploid WA-WA, which means that this is not likely to be due to genetic divergence.

Interestingly, the WA-WE gametes have very low viability (figure 6.3, figure 6.4). The divergence between the two strains is not radically different from other crosses (e.g. SA-WE), which suggests that there is some particular incompatibility between these the WA and WE genomes. It is not likely to be any structural variation, since that would have come out also in other hybrids. With the end-to-end genome assemblies of the parents, and genome sequences of the gametes, we hope to explain these types of patterns.

Table 6.4. Crosses and their viabilities. This table shows the diploid parents, hybrids and their associated total amount of spores dissected and their viability. The ratio of tetrad types ranges from 0 to 4, the numbers designate the number of spores viable in a tetrad, so the values in column type 4 is the ratio of tetrads with all spores viable, while the values in column type 3 corresponds to the amount of tetrads with three viable spores.

				Ratio of tetrad types				
	Cross	Spores	Viability (%)	0	1	2	3	4
Parents	NA-NA	472	94	0	0	0.06	0.11	0.83
	SA-SA	472	88	0	0.01	0.03	0.39	0.57
	WE-WE	736	82	0.02	0.04	0.11	0.29	0.54
	WA-WA	152	60	0.00	0.13	0.37	0.26	0.18
Hybrids	SA-NA	2,420	85	0.04	0.03	0.09	0.17	0.67
	NA-WE	2,720	84	0.03	0.01	0.12	0.25	0.59
	NA-WA	2,680	81	0.03	0.02	0.12	0.35	0.48
	SA-WE	2,660	76	0.08	0.04	0.15	0.23	0.50
	SA-WA	2,776	75	0.03	0.05	0.20	0.35	0.37
	WA-WE	5,036	44	0.12	0.26	0.40	0.18	0.04

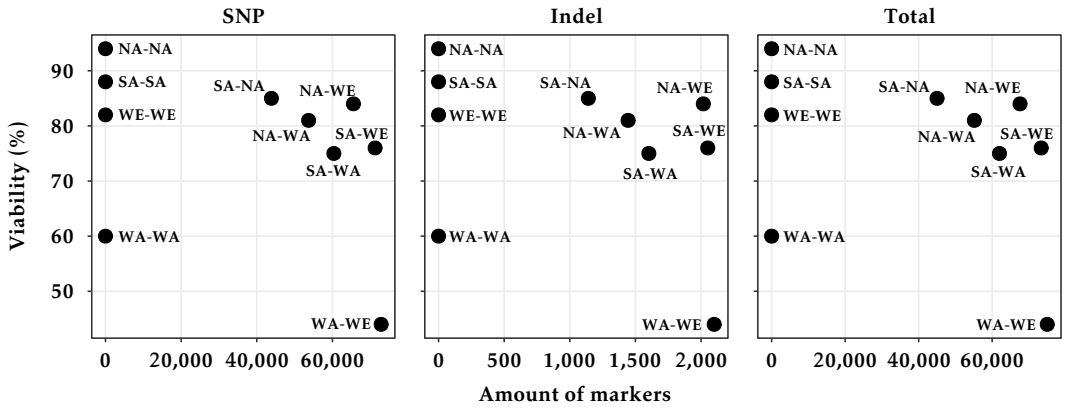


Figure 6.3. Divergence and viability. The divergence of two strains that are crossed together can have an effect on the gamete viability (Greig, 2008). The SNP, Indel, and total amount of markers are plotted on the x-axis (Table 6.3) while the average viability of the gametes of a given cross is on the y-axis (Table 6.4). The data points on $x = 0$ corresponds to the diploid parents.

6.9.3 Aneuploidies

I ran the aneuploidy analysis on the first sets of sequencing that we received from the sequencing facility, therefore, the data shown here are from 2,296 sequenced spores. Due to the, for now, limited sample size and the frequency of aneuploidies, we do not have a substantial amount of spores with aneuploidies; the amount of aneuploidies found is at the moment 103, but will increase as the sample size increases which will strengthen (or refute) any trends seen thus far.

The amount of aneuploidies as a function of chromosome size corroborates previous studies in that amount of aneuploidies increase as the size of the chromosomes decrease (Fig. 6.4a) (Mancera et al., 2008; Cubillos et al., 2011).

The amount of aneuploidies in a specific hybrid for a specific chromosome might have more to it than just differing chromosome sizes. We are investigating the possibility of parent or hybrid specific aneuploidies, i.e. aneuploidies that occur disproportionately often in a cross. In figure 6.4b, all aneuploidies found so far are sorted according to chromosome and cross. From this data, it seems like the chr XI has an increase in aneuploidies in the SA-WE cross compared to the others. This may be due to some incompatibility between the two parental strains that only manifest in that specific cross, with our end-to-end genome assemblies, we can investigate what kind of genetic or genomic incompatibility might cause this pattern.

6.10 Perspectives

With the large amount of data generated in this project we hope to give a complete view of the influence of genetic background to recombination landscapes, aneuploidies, quantitative traits, and gamete inviability.

This project is based on four parents rather than two (as in my previous project), giv-

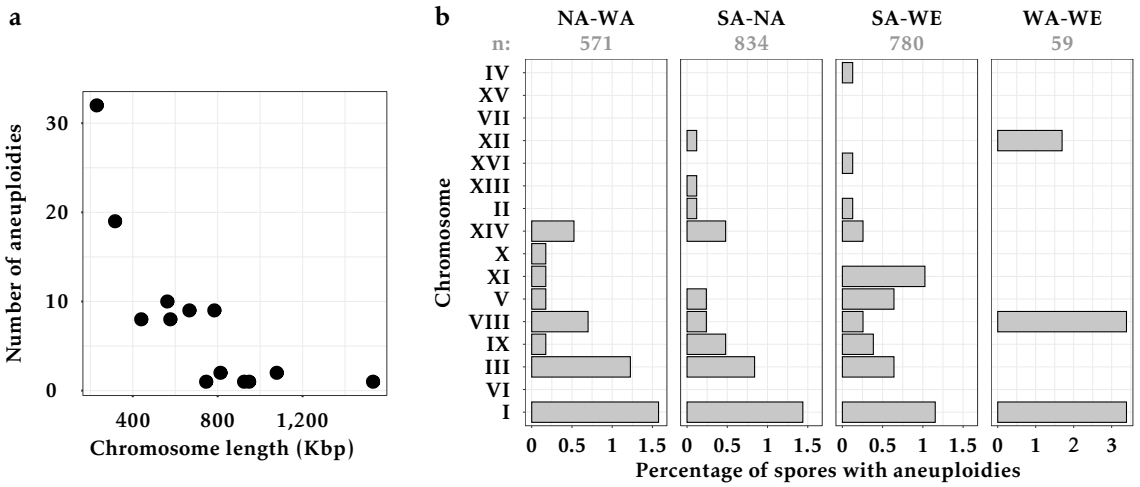


Figure 6.4. Chromosome size, hybrids and aneuploidies. (a) Corroborating previous studies, the amount of aneuploidies has an inverse relationship with the length of the chromosomes, such that smaller chromosomes tend to gain more aneuploidies than larger ones. (b) Sorting the aneuploidies according to chromosome (*y*-axis, sorted according to chromosome length) and hybrid (facets) reveals possible hybrid specific aneuploidies. Chromosome XI has a large number of aneuploidies in the SA-WE hybrid but not in the others, which may be due to some genetic or genomic incompatibility between the SA and WE strain. Values are normalized by the amount of spores sequenced for the given hybrid. Note that the WA-WE facet has a different scale on the *x*-axis, due to its smaller samples size (n=59) these values are not as robust. n refers to the amount of spores that have been sequenced, per hybrid, and included into this analysis (Table 6.3). The NA-WE hybrid is not shown since no aneuploidies were found for it (n = 52).

ing me the opportunity to investigate how the landscape of QTLs can be affected by different parental contributions. Additionally, in light of the recent article by She and Jarosz (2018), we can also use real data to look at the increase in resolution of QTLs that comes with less marker density Table 6.3. In contrast to my previous project, I would also like to investigate QTL-QTL or QTL-genome interactions, using our six different hybrids we could look at how conserved different interactions are.

In contrast to Hallin et al. (2016) and Märtens et al. (2016), the segregants in this project are the result of just one meisis, which means that they have had less recombination events and will due to this have a lower QTL resolution. However, using the logic from the round-robin approach by Treusch et al. (2015) we might be able to use the different crosses in order to narrow down the regions.

Our mapping population is haploid which limits our genetic contributions to variation to additive and epistatic components. Looking at haploid gametes does however open up a completely new set of interesting analyses that can be done. Looking at the genetic contributions to gamete inviability, for example, naturally cannot be done in diploids. Additionally, this large dataset of haploids means that a new big cross grid experiment could be done, with a much larger sample size (albeit, with bigger linkage regions).

References

Anderson Carol M, Chen Stacy Y, et al. **ReCombine: a suite of programs for detection and analysis of meiotic recombination in whole-genome datasets.** *PloS one*, 6(10):e25509, 2011.

Antonarakis Stylianos E, Lyle Robert, Dermitzakis Emmanouil T, Reymond Alexandre, and Deutsch Samuel. **Chromosome 21 and Down syndrome: from genomics to pathophysiology**. *Nature Reviews Genetics*, 5(10):725–738, 2004.

Bergström Anders, Simpson Jared T, et al. **A high-definition view of functional genetic variation from natural yeast genomes.** *Molecular Biology and Evolution*, 31(4):872–888, 2014.

Bloom Joshua S, Ehrenreich Ian M, Loo Wesley T, Lite Thúy-Lan Võ, and Kruglyak Leonid. **Finding the sources of missing heritability in a yeast cross.** *Nature*, 494(7436):234–237, 2013.

Cubillos Francisco A, Billi Eleonora, et al. **Assessing the complex architecture of polygenic traits in diverged yeast populations.** *Molecular ecology*, 20(7):1401–1413, 2011.

Down J Langdon H. **Observation on an ethnic classification of idiots**. *328*, 1866.

Greig D. **Reproductive isolation in Saccharomyces**. *Heredity*, 102(1):39–44, 2008.

Hallin Johan, Märtens Kaspar, et al. **Powerful decomposition of complex traits in a diploid model.** *Nature Communications*, 7:13311, 2016.

Liti Gianni, Carter David M, et al. **Population genomics of domestic and wild yeasts.** *Nature*, 458(7236):337–341, 2009.

Mancera Eugenio, Bourgon Richard, Brozzi Alessandro, Huber Wolfgang, and Steinmetz Lars M. **High-resolution mapping of meiotic crossovers and non-crossovers in yeast**. *Nature*, 454(7203):479–485, 2008.

Märtens Kaspar, Hallin Johan, Warringer Jonas, Liti Gianni, and Parts Leopold. **Predicting quantitative traits from genome and phenome with near perfect accuracy.** *Nature Communications*, 7:11512, 2016.

Pau G, Fuchs F, Sklyar O, Boutros M, and Huber W. **EBImage–an R package for image processing with applications to cellular phenotypes**. *Bioinformatics*, 26(7):979–981, 2010.

She Richard and Jarosz Daniel F. **Mapping Causal Variants with Single-Nucleotide Resolution Reveals Biochemical Drivers of Phenotypic Change**. *Cell*, 172(3):478–490.e15, 2018.

Treusch Sebastian, Albert Frank W, Bloom Joshua S, Kotenko Iulia E, and Kruglyak Leonid. **Genetic mapping of MAPK-mediated complex traits Across S. cerevisiae.** *PLoS Genetics*, 11(1):e1004913, 2015.

Yue Jia-Xing, Li Jing, et al. **Contrasting evolutionary genome dynamics between domesticated and wild yeasts.** *Nature genetics*, 49(6):913–924, 2017.

Zackrisson Martin, Hallin Johan, et al. **Scan-o-matic: High-Resolution Microbial Phenomics at a Massive Scale**. *G3 (Bethesda, Md.)*, 6:1–12, 2016.

Ziv Naomi, Shuster Bentley M, Siegal Mark L, and Gresham David. **Resolving the Complex Genetic Basis of Phenotypic Variation and Variability of Cellular Growth.** *Genetics*, page genetics.116.195180, 2017.

Chapter 7

Discussion and perspectives

The discussion is, arguably, the heart of any scientific publication. However, given that the scientific papers included in this thesis contain within them their own discussion, I will here simply discuss potential limitations of my work. As well as what could have been done differently and what could have been done additionally.

7.1 A QTL mapping population

The mapping population used in both papers accompanying this thesis (Märtens et al., 2016; Hallin et al., 2016) was constructed using a Singer ROTOR HDA robot (Singer Ltd, UK.). The rationale behind this mapping population was to increase the power to detect QTLs by increasing the sample size while keeping the sequencing cost down, and at the same time maximize the resolution for QTL mapping by using advanced intercrossed lines, and doing all this in diploids.

As we stated in the article, the resulting population structure of the mapping population inhibited the power with which we could call additive QTLs. A large proportion of the variation between the strains were explained by the relatedness between them, i.e. sharing half a genome made phenotypes quite similar. In order to remove the population structure for mapping additive QTLs, we calculated the inferred diploid hybrid phenotype (see paper), this effectively calculates the average effect of a haplotype by fixing the shared parental haplotype and randomizing the second parental haplotype. This removes the population structure, but unfortunately also reduces the sample size of the mapping population down to the amount of haploid parents that we sequenced. This removing of the population structure also removes any nonadditive effects that comes from interactions with the second haplotype. However, there may still be some epistatic interaction effects in this scaled down mapping population, but these effects cannot be mapped with the traditional marker regression method.

Instead, for mapping nonadditive effects we took advantage of the larger sample size as the nonadditive phenotypes were calculated to be the deviation between the diploid hybrid and the mean of the two inferred parents. The rationale behind this is that the deviations from average effect

of the average effect of the two haplotypes would constitute a purely additive inheritance, and any deviation from that would be due to within or between locus effects. This means that the sample size was again increased to the actual number of strains that we phenotyped, and explains why we might have been calling nonadditive QTLs with higher power than additive.

Althought the additive QTLs suffered a bit from the crossing scheme, the differences in genetic relatedness did aid in the phenotype predictions as the quality of predictions could be compared between the two groups. The crosses also allowed us to look at within locus contributions to heterosis at a large scale, which will be discussed later on.

Recently, a paper was published using advanced intercross lines (although they don't call it that) (She and Jarosz, 2018). They manage to map QTLs at nucleotide resolution by a very simple solution: they decreased the genetic divergence between the two parental strains. This reduces the complexity of the model, and increases the amount of space between the markers. This increase in of inter-marker distance allowed them to locate the variants that had the effect. The decrease in complexity in genome differences did not effect the diversity of phenotypes in the mapping population, and effectively only gave them the opportunity to really distinguish the causal variants down to the nucleotide.

The approach of reducing the divergence between the parental strains applied by She and Jarosz (2018) gives nothing new in terms of traditional mapping attempts where the goal of choosing parents have been to have as high variation between them as possible. However, it begs the question of why the choice of parents was to maximize the variation, rather than maximizing the probability of capturing the causal variant.

However, when using *S. cerevisiae* as a model for a specific complex trait, rather than as a model for complex traits in general, you might find yourself with less strains to choose from. This would then force you to go with whichever strains that are differing in your phenotype of interest. That being said, the amount of strains that are available now for these types of studies (e.g. The 1002 Yeast Genomes Project) might contain enough strains to easily find a good combination, hitting the sweet spot between diverging phenotypes and diverging genotypes.

In the context of my papers, it would have been interesting to have used less diverged strains and perhaps being able to narrow down the QTLs to single nucleotides. It is known that QTLs can encompass several causal variants with high linkage (Steinmetz et al., 2002; Lorenz and Cohen, 2012), and due to that the number of QTLs that we detect is likely to be an underestimation of the actual amount of causal SNPs for the traits.

7.2 Contributions to heterosis

We defined heterosis in our hybrids as deviations from the mean of the two inferred

parents. However, we never did a strict check of the assumption that the inferred parental phenotypes would correlate well with actual diploid parents. Another approach could have been to simply use the phenotype values of the haploid parents, which were phenotyped at the same time as the POLs. However, since the ploidy of cells affects phenotypes (Zörgö et al., 2013; Gerstein and Otto, 2009) and, furthermore, the parents had segregating markers that also affect the phenotype, this comparison would not have been relevant.

Instead, what could have been done is to diploidize the haploid parents. This path of action is complicated by the fact that the **HO locus** of our strains is deleted in order to repress mating type switching. However, there are ways to do this, for example introducing an inducible HO gene (Furukawa et al., 2011). This, of course, requires some genetic engineering of the parental strains, which was beyond the scope of this article. However, given the fact that the hybrids only have two variants segregating at each position, the randomization of the second haplotype that results from calculating the inferred parental phenotype, should be well represented by both variants at any given loci, and should make the inferred phenotype quite close to an actual value from a homozygous diploid strain.

In a conventional yeast heterosis study, a cross is made between two diverged parents, and then the phenotype between the completely heterozygous hybrid and the two homozygous parents are compared (Zörgö et al., 2012; Plech et al., 2014; Shapira et al., 2014; Bernardes et al., 2017). As in my study, they would then define the hybrid as heterotic depending on its relationship to the average phenotype of the two parents. However, when investigating whether dominance or overdominance has the most important effect, these studies (excluding Bernardes et al. (2017) which only looked at heterosis) average the effect of all loci in the genome. I.e. they look at the phenotype of the hybrids, and depending on the degree to which they deviate from the mid-parent expectation, dominance or overdominance is invoked. Since dominance and overdominance are, by definition, phenomena that occur within single loci, the rather blunt approach of looking at the phenotype as averaged over the entire genome may not be the best way at distinguishing the contributions.

In our study, we instead look at the individual contribution of loci to heterosis by investigating the relative frequencies of different genotypes at QTLs in hybrids that were heterotic. We chose to only look at the contribution of QTLs since they are regions that we know significantly contribute to the trait. Although our method refines the search for contributions to heterosis, it is still not at its most refined state. In order to look at the contribution to heterosis at individual loci, we used the markers that were below the apex of the QTL peaks. However, we are not 100% confident that that is the causal variant. If the causal variant is in fact the marker next to the marker we have chosen, this may decrease our power to find significant contributions. This is due to the fact that some hybrids may have recombi-

HO locus. The locus containing the gene for the HO endonuclease which creates a double-strand break at the MAT locus to facility mating type switching (Haber, 2012).

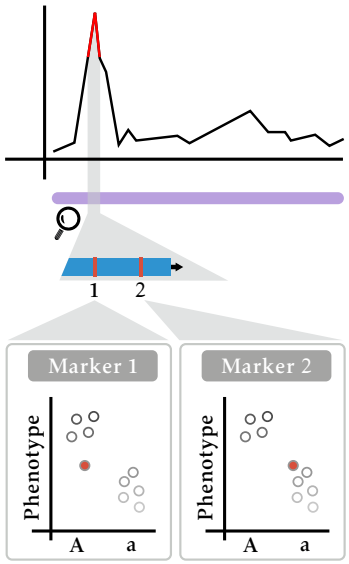


Figure 7.1. Heterosis and QTL resolution. The power to detect significant dominant or overdominant contributions can be affected by the resolution of the QTL. Using the marker below the apex of the QTL (marker 1) could result in a lower power to detect dominance or overdominance contributions, if the causal marker is not the one under the apex. Here marker 2 is the causal marker and is the marker contributing to the phenotype. The colored circle indicates an individual that has different genotypes at the two markers. This results in it having the phenotype of allele a (since it has allele a at the causal marker) but being grouped with the individuals with allele A (since it has allele A at the marker under the apex).

nation events between the causal marker and the marker that we chose to represent the causal marker, this means that the phenotype of a hybrid may not represent the genotype of the marker we are looking at Fig. 7.1. That being said, this should not have a very large effect since the amount of individuals with a recombination break point between markers with such high linkage should not be very common.

Along the lines of the QTL mapping discussed above, it would be interesting to have even higher resolution in order to distinguish between tightly linked loci that have an effect. This would, for example, remove the risk of calling pseudo-overdominant loci.

50% of the QTLs do not show any contribution to heterosis in our study. It is possible that within these 50% we have QTLs that contribute to heterosis but that we did not detect it. QTLs with contribution might be missed due to it having a too small effect size to be significant given our sample size. Alternatively, they might have an epistatic effect. However, if that were the case, they would be more difficult to detect. Nevertheless, it would be interesting to also look for QTL-QTL interactions that contribute to heterosis. QTL-QTL and QTL-genome interactions have been shown to affect the phenotypic variation (Bloom et al., 2015), and would warrant an investigation into their potential effect on heterosis.

7.3 Closing remarks

In my first project (chapter 4 and chapter 5), we devised the POL approach as a novel methodology to answer questions about quantitative genetics in a diploid model. Quite successfully, we managed to decompose the genetic components of the phenotypic variation, map QTLs, investigate the genetic contributions to heterosis and predict traits with unparalleled accuracy. However, there were things that we could not do, things that we can investigate in my new project (chapter 6). In this project, the use of four different parents will allow us to look at, for example, con-

text dependent or independent QTLs and interactions. Using gametes rather than diploid hybrids gives us the opportunity to look at the effect of aneuploidies on phenotypic variation and how the underlying genomes can induce aneuploidies. The aspect of context dependence will be very interesting to investigate. This was an aspect that we could not address using the POLs as they were originally from a two-parent cross.

During my PhD I have used innovative approaches to address long standing questions in genetics, and taken together, the work I have done during my PhD has contributed not only to the knowledge of the genetics behind complex traits, but also to the methods with which we try to understand it.

References

Bernardes J P, Stelkens R B, and Greig D. **Heterosis in hybrids within and between yeast species.** *Journal of evolutionary biology*, 30(3):538–548, 2017.

Bloom Joshua S, Kotenko Iulia, et al. **Genetic interactions contribute less than additive effects to quantitative trait variation in yeast.** *Nature Communications*, 6:8712–6, 2015.

Furukawa Kentaro, Furukawa Takako, and Hohmann Stefan. **Efficient Construction of Homozygous Diploid Strains Identifies Genes Required for the Hyper-Filamentous Phenotype in Saccharomyces cerevisiae**. *PloS one*, 6(10):e26584, 2011.

Gerstein A C and Otto S P. **Ploidy and the Causes of Genomic Evolution**. *Journal of Heredity*, 100(5):571–581, 2009.

Haber James E. **Mating-Type Genes and MAT Switching in Saccharomyces cerevisiae**. *Genetics*, 191(1):33–64, 2012.

Hallin Johan, Märtens Kaspar, et al. **Powerful decomposition of complex traits in a diploid model.** *Nature Communications*, 7:13311, 2016.

Lorenz Kim and Cohen Barak A. **Small- and large-effect quantitative trait locus interactions underlie variation in yeast sporulation efficiency.** *Genetics*, 192(3):1123–1132, 2012.

Märtens Kaspar, Hallin Johan, Warringer Jonas, Liti Gianni, and Parts Leopold. **Predicting quantitative traits from genome and phenome with near perfect accuracy.** *Nature Communications*, 7:11512, 2016.

Plech Marcin, de Visser J Arjan G M, and Korona Ryszard. **Heterosis is prevalent among domesticated but not wild strains of Saccharomyces cerevisiae.** *G3 (Bethesda, Md.)*, 4(2):315–323, 2014.

Shapira R, Levy T, Shaked S, Fridman E, and David L. **Extensive heterosis in growth of yeast hybrids is explained by a combination of genetic models.** *Heredity*, 113(4):316–326, 2014.

She Richard and Jarosz Daniel F. **Mapping Causal Variants with Single-Nucleotide Resolution Reveals Biochemical Drivers of Phenotypic Change**. *Cell*, 172(3):478–490.e15, 2018.

Steinmetz Lars M, Sinha Himanshu, et al. **Dissecting the architecture of a quantitative trait locus in yeast.** *Nature*, 416(6878):326–330, 2002.

Zörgö Enikö, Chwialkowska Karolina, et al. **Ancient evolutionary trade-offs between yeast ploidy states.** *PLoS Genetics*, 9(3):e1003388, 2013.

Zörgö Enikö, Gjuvsland Arne, et al. **Life history shapes trait heredity by accumulation of loss-of-function alleles in yeast.** *Molecular Biology and Evolution*, 29(7):1781–1789, 2012.

Chapter 8

Publications

All publications that I have been associated to during my PhD are listed below. The two main articles of my PhD can be read in their full length in chapter 4 and chapter 5, while the abstract and my personal contribution to the three articles that were not part of my main project can be seen in the following pages. Shortly, in all articles I have contributed with my expertise in large scale phenotyping and experience in handling large experiments, at the same time as these different projects have added to that experience.

Johan Hallin, Kaspar Märtens *et al*.
Powerful decomposition of complex traits in a diploid model.
Nature Communications, 2016

Kaspar Märtens, Johan Hallin *et al*.
Predicting quantitative traits from genome and phenome with near perfect accuracy.
Nature Communications, 2016

Martin Zackrisson *et al*.
Scan-o-matic: high-resolution microbial phenomics at a massive scale.
G3, 2016

Jia-Xing Yue *et al*.
Contrasting evolutionary genome dynamics between domesticated and wild yeasts.
Nature Genetics, 2017

Ignacio Vázquez-García *et al*.
Clonal heterogeneity influences the fate of new adaptive mutations.
Cell Reports, 2017

Personal contribution

Scan-o-matic: high-resolution microbial phenomics at a massive scale

This article describes a novel methodology for high through-put high quality colony growth phenotyping, showcasing its precision by implementing it to further elucidate the genetics of salt resistance in yeast. My contribution to this project was mainly performing experiments using the methodology, fine-tuning the experimental protocol, and contributing to the development of the program by giving feedback on the usability to Martin. I have taken advantage of Scan-o-matic in all the projects that I have been associated to so far. My extensive use of this methodology has made me comfortable with all aspects of large scale phenotyping; from experiment design, to execution, to data analysis and interpretation.

Scan-o-matic: high-resolution microbial phenomics at a massive scale

The capacity to map traits over large cohorts of individuals—phenomics—lags far behind the explosive development in genomics. For microbes, the estimation of growth is the key phenotype because of its link to fitness. We introduce an automated microbial phenomics framework that delivers accurate, precise, and highly resolved growth phenotypes at an unprecedented scale. Advancements were achieved through the introduction of transmissive scanning hardware and software technology, frequent acquisition of exact colony population size measurements, extraction of population growth rates from growth curves, and removal of spatial bias by reference-surface normalization. Our prototype arrangement automatically records and analyzes close to 100,000 growth curves in parallel. We demonstrate the power of the approach by extending and nuancing the known salt-defense biology in baker's yeast. The introduced framework represents a major advance in microbial phenomics by providing high-quality data for extensive cohorts of individuals and generating well-populated and standardized phenomics databases.

Martin Zackrisson, **Johan Hallin**, Lars-Göran Ottosson, Peter Dahl, Esteban Fernandez-Parada, Erik Ländström, Luciano Fernandez-Ricaud, Petra Kaferle, Andreas Skyman, Simon Stenberg, Stig Omholt, Uroš Petrovič, Jonas Warringer & Anders Blomberg

Published in G3 (2016)
doi:10.1534/g3.116.032342/-/DC1

Personal contribution

Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Successfully assembling high-quality genomes is not easy, but by PacBio and Illumina sequencing twelve strains representing major clades of *Saccharomyces cerevisiae* and its wild cousin *Saccharomyces paradoxus* Jia-Xing assembled their genomes end-to-end with reference-quality. Their genome dynamics were compared to give a view of how different selection pressures acting on these two yeast species may have shaped their genomes. The high quality end-to-end genome assemblies allowed me to test how subtelomeric gene structures affected phenotypic variation to arsenite resistance by using the POL approach that I had developed previously. My contribution is shown in figure 7d-f of the article.

Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Structural rearrangements have long been recognized as an important source of genetic variation, with implications in phenotypic diversity and disease, yet their detailed evolutionary dynamics remain elusive. Here we use long-read sequencing to generate end-to-end genome assemblies for 12 strains representing major subpopulations of the partially domesticated yeast *Saccharomyces cerevisiae* and its wild relative *Saccharomyces paradoxus*. These population-level high-quality genomes with comprehensive annotation enable precise definition of chromosomal boundaries between cores and subtelomeres and a high-resolution view of evolutionary genome dynamics. In chromosomal cores, *S. paradoxus* shows faster accumulation of balanced rearrangements (inversions, reciprocal translocations and transpositions), whereas *S. cerevisiae* accumulates unbalanced rearrangements (novel insertions, deletions and duplications) more rapidly. In subtelomeres, both species show extensive interchromosomal reshuffling, with a higher tempo in *S. cerevisiae*. Such striking contrasts between wild and domesticated yeasts are likely to reflect the influence of human activities on structural genome evolution.

Jia-Xing Yue, Jing Li, Louise Aigrain, **Johan Hallin**, Karl Persson, Karen Oliver, Anders Bergström, Paul Coupland, Jonas Warringer, Marco Cosentino Lagomarsino, Gilles Fischer, Richard Durbin & Gianni Liti

Published in Nature Genetics (2017)
doi:10.1038/ng.3847

Personal contribution

Clonal heterogeneity influences the fate of new adaptive mutations

This paper gives an account of the contribution of *de novo* and standing genetic variation to adaptive evolution in *Saccharomyces cerevisiae* populations. My main part in this project was the cross-grid experiment which was used to shuffle the *de novo* mutations and the genetic backgrounds in order to unlink their contributions, giving us a way to estimate their respective contributions. I collected and genotyped the spores to validate their mutations, and constructed and performed the crossing into diploid hybrids as well as doing the phenotyping for all strains. My contribution is represented in figure 6 of the paper.

Clonal heterogeneity influences the fate of new adaptive mutations

The joint contribution of pre-existing and de novo genetic variation to clonal adaptation is poorly understood but essential to designing successful antimicrobial or cancer therapies. To address this, we evolve genetically diverse populations of budding yeast, *S. cerevisiae*, consisting of diploid cells with unique haplotype combinations. We study the asexual evolution of these populations under selective inhibition with chemotherapeutic drugs by time-resolved whole-genome sequencing and phenotyping. All populations undergo clonal expansions driven by *de novo* mutations but remain genetically and phenotypically diverse. The clones exhibit widespread genomic instability, rendering recessive *de novo* mutations homozygous and refining pre-existing variation. Finally, we decompose the fitness contributions of pre-existing and *de novo* mutations by creating a large recombinant library of adaptive mutations in an ensemble of genetic backgrounds. Both pre-existing and *de novo* mutations substantially contribute to fitness, and the relative fitness of pre-existing variants sets a selective threshold for new adaptive mutations.

Ignacio Vázquez-García, Francisco Salinas, Jing Li, Andrej Fischer, Benjamin Barré, **Johan Hallin**, Anders Bergström, Elisa Alonso-Perez, Jonas Warringer, Ville Mustonen & Gianni Liti

Published in Cell Reports (2017)
doi:10.1016/j.celrep.2017.09.046