Disentangling the effects of mutation and selection on the evolution of gene expression

by

Brian P.H. Metzger

A dissertation submitted in partial fulfillment
Of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in the University of Michigan
2015

Doctoral Committee:
Professor Trisha Wittkopp, Chair
Professor Michael Boehnke
Associate Professor Timothy James
Professor George Zhang

To Jen, for all your love and support.

Acknowledgments

I would first like to thank Trisha Wittkopp for her support, mentorship, and scientific insights over the years, as well as for creating a laboratory environment that encourages collaboration and sharing of all that is exciting about science. I would also like to thank: Fabien Duveau for being a great collaborator and friend, as well as his significant contributions to chapter 5; Dave Yuan for creating an amazing collection of strains that were necessary for much of this work and for significant contributions to chapter 3; Jonathan Gruber for his years of hard work laying the foundation for many of the projects included here; Andrea Hodgins-Davis for her constant reminders that the environment matters; Joe Coolon and Kraig Stevenson for statistical support and joint commiseration; All past and present members of the Wittkopp lab. Without them research will be decidedly less fun; George Zhang for his guidance and insights into molecular evolution; Calum Maclean for teaching me how to work with yeast and significant contributions to Chapter 2; Jianrong Yang for computational support and significant contributions to chapter 2; My committee members Tim James and Mike Boehnke for their guidance; The Genome Science Training Program for financial and intellectual support; My family and friends for their unwavering support, even when they don't know exactly what it is that I am doing; and finally, my wife, Jen, for being a constant source of laughter and joy in my life.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

Mutation is the ultimate source of phenotypic variation. However, little is known about the effects of new mutations in the absence of natural selection and whether these effects can influence the course of evolution. This is particularly true for changes in gene expression and regulation. In this thesis I measure the effects of new *cis*- and *trans*-regulatory mutations on the expression of the *Saccharomyces cerevisiae TDH3* gene. Using these measurements, I show that *cis*- and *trans*-regulatory mutations have fundamentally different effects on gene expression. In particular, I find that *cis*-regulatory mutations are on average larger than *trans*-regulatory mutations and skewed towards decreases in *TDH3* expression, while *trans*-regulatory mutations are often, but not always, more common than *cis*-regulatory mutations and skewed towards increases in *TDH3* expression. To determine how natural selection has acted on these differences, I generate genome sequences and genetically tractable versions of over 60 diverse *S. cerevisiae* strains previously isolated from a range of environments. I use these strains to determine the effects of *cis*- and *trans*-regulatory polymorphism on *TDH3* expression. Comparing these effects to the effects of new mutations, I find that natural selection has acted on both *cis*- and *trans*-regulatory variants. Interestingly, the effects of selection varies between *cis*- and *trans*-regulatory changes due to differences in the effects of new mutations. Using the same approach, I also identify differences in the action of natural selection on *cis*- and *trans*-regulatory changes for the variability in expression amongst

genetically identical individuals, i.e. gene expression noise. Finally, I determine the evolution of regulatory changes over long evolutionary timescales in *Saccharomyces*. I find widespread evidence for compensatory changes in regulation, particularly for *trans*-regulatory changes that act in opposite directions. Consistent with this finding, I identify hundreds of *trans*-acting QTL affecting *TDH3* expression amongst four strains of *S. cerevisiae*. Together these results suggest that *trans*-regulatory changes are a common, but individually small, source of regulatory variation. In total, this thesis shows that understanding the effects of new mutations and comparing these effects to observed differences in natural populations can be a powerful approach for elucidating the underlying molecular mechanisms governing evolution.

Chapter 1

Introduction

Natural selection may explain the survival of the fittest, but it cannot explain the arrival of the fittest

— de Vries, 1905

Mutation is the ultimate source of heritable variation in evolution and provides the raw material upon which the diversity and complexity of life is built. Because natural selection and genetic drift lead to the loss of genetic and phenotypic variation over time, new mutation is regarded as essential for evolution. However, the mutational process does more than simply generate new phenotypes. It also determines the frequency and identity of the phenotypes that arise in a population. In this way, the mutational process actively shapes the course of evolution by determining the phenotypes that are available for other evolutionary forces to act on. This has direct consequences for the evolutionary process because phenotypes that appear more frequently are more likely to contribute to differences amongst populations and species, regardless of the action of genetic drift or natural selection. These differences in the rate at which phenotypes arise reflect underlying biological constraints and are therefore an integral part of the evolutionary process. Unfortunately, empirical data on the effects of new mutations is limited. This thesis focuses on determining the effects of new mutations on gene expression and how biases in the mutational process interact with natural selection.

The tangled history of mutation and selection in the study of evolution

In The Origin of Species, Darwin argued that the diversity of life on earth was the result of natural selection applied over thousands of generations and not divine creation (DARWIN 1872). The recognition that heritable differences in phenotypes affect fitness and that differences in fitness result in the preferential retention of some phenotypes over others fundamentally changed the study of biology. By offering natural selection as a specific mechanism for the evolutionary process, Darwin's theory united all of life as descendant from a common ancestor and contributed substantially to the scientific acceptance of evolution.

Like many scientists of the time, Darwin believed in a blending mode of inheritance, such that offspring phenotypes were intermediate of those of its parents (DARWIN 1868). However, it was recognized that under such a model of inheritance, the effects of natural selection in one generation are largely erased in the phenotypic blending of subsequent generations (JENKIN 1867). In addition, without a constant source of new phenotypic variation, the removal of variants by natural selection would eventually cause evolution to stall. Such logic formed the main critiques against Darwin's theory in the late 19th and early 20th centuries and lead to an 'eclipse of Darwinism' (BOWLER 1992). Thus, while Darwin's theory of evolution by natural selection greatly expanded the acceptance of evolution, it was widely recognized as incomplete, lacking adequate mechanisms describing inheritance and the origin of phenotypic variation.

The difficulties in reconciling natural selection with blending inheritance caused alternative theories of evolution to emerge. These theories focused on the inheritance of discontinuous variation, instead of continuous variation as proposed by Darwin, as the main mechanism behind evolution (BATESON 1894; DE VRIES 1905). These early mutationalists argued that new species were the result of discontinuous changes, or mutations, that had large phenotypic effects, rather than through the action of natural selection on small continuous differences. While often portrayed as antagonistic towards natural selection, these scientists were primarily concerned with the origins of phenotypic variation instead of the maintenance of phenotypic variation and were critical of natural selections role in creating variation and new species, not natural selection as a broader process (STOLTZFUS and CABLE 2014). To support their view, the mutationalists focused on identifying and cataloging hundreds of discrete differences amongst animals and plants because such differences were unlikely to arise under a blending model of inheritance (BATESON 1894; DE VRIES 1901). However, many of the documented mutations were later shown to either not result in heritable phenotypic changes or to be caused by chromosomal instabilities and were thus not thought to be important for evolution. As a consequence, a direct role for mutation in speciation was largely discarded as a major component of the evolutionary process.

Unknown to the majority of scientists at the time, the problem of inheritance had been partially solved by Mendel contemporaneously with Darwin. Mendel's insight was the conceptualization and experimental data showing that inheritance is particulate, i.e. that the genetic material was not blended between parents in their offspring, but instead was

maintained as discrete units that could be inherited in subsequent generations (MENDEL 1866). This view of inheritance centered on discontinuous variation and was in line with the views of many mutationalists.

The rediscovery of Mendels's laws in 1900 resulted in rapid progress in the understanding of genetics, ultimately resulting in the melding of discrete inheritance of phenotypic characters with the emergence of new mutations (STOLTZFUS 2012). By applying these concepts to evolution, a relatively cohesive view of evolutionary biology as a two-step process emerged. First, phenotypic variation was generated by the mutation process, with each subsequent generation inheriting mutations according to Mendels laws. This variation was then sorted based on its fitness effects by natural selection, creating the patterns of phenotypic diversity observed in nature (MORGAN 1916). However, this view of evolution was not universally accepted because the vast majority of heritable mutations identified were deleterious and thus unlikely to contribute to evolution. Essentially, without experimental data identifying *de novo* beneficial mutation, it was hard to accept new mutation as playing a large role in evolution. Even more problematically, while discontinuous variation and mutation were seen as plausible explanations for Mendelian traits, many observed phenotypes were continuous and it was unclear how they could result from Mendelian inheritance. This friction resulted in two disparate views of evolution: the first (Mendellians) placed emphasis on mutation and the inheritance of discontinuous phenotypes, while the second (biometricians) focused on natural selection and continuous variation, as the primary mechanisms by which evolution proceeds.

In addition to providing a mechanism for the inheritance of discontinuous variants, the rediscovery of Mendel's laws also initiated the mathematical and quantitative study of genetics. Under Mendel's laws, inheritance of individual phenotypes was straightforward and could be extended for thousands of generations to predict how evolution would proceed over long time scales. The melding of Darwinian selection, Mendelian inheritance, and quantitative genetics eventually resulted in what is now referred to as the Modern Synthesis (HUXLEY *et al.* 2010). One of the most critical developments from this synthesis was the mathematical and conceptual recognition that if a trait was influenced by many loci, the application of Mendel's laws would result in continuous variation whose inheritance was consistent with the biometrician's observations (FISHER 1919). This effectively ended the debate between the Mendelian's and biometricans and is often hailed as one of the crowning achievements of evolutionary biology in the 20th century.

Along with the melding of two disparate camps, the mathematical analysis of evolution also indicated that allele frequency changes in a population are dominated by natural selection and genetic drift and that pressure from new mutations plays only a minor role. As a consequence, mutation came to be seen only as a necessary generating force in evolution, while natural selection was viewed as the primarily force responsible for the diversity of life on earth. The widespread acceptance of natural selection as the dominant force in evolution also lead many biologists to view organisms as exquisitely well adapted to their environment such that it was only upon changes in the environment that evolution took place (FISHER 1930). This view had the consequence of further

diminishing the role that new mutation played in evolution: if the majority of adaptation takes place through environmental change using already existing variation, new mutations are never directly relevant. These views fed back on themselves, such that the reduction in importance given to new mutations further strengthened support for natural selection and *vice versa*. Thus, in the course of several decades, the role of mutation in evolution went from an alternative to natural selection as a mechanism for speciation and an important force in the evolution of phenotypic diversity to being regarded as a secondary force of considerably less importance or interest than natural selection.

Early studies of molecular evolution focused on the extent of similarities and differences amongst proteins from different organisms. Surprisingly, these studies revealed that the rate of evolution is often proportional to the mutation rate (ZUCKERLANDL and PAULING 1962). Such a strong correlation between rates of evolution and rates of mutation, however, was fundamentally at odds with the dominant view of evolution at the time: if organisms are well adapted to their environment and evolution proceeds by shifting allele frequencies in response to changes in the environment due to natural selection, then the rate of evolution should be proportional to the rate of environmental change, not the mutation rate (NEI 2013). In response to these discrepancies, many molecular and phenotypic differences were reimagined as neutral, drifting in frequency and largely inconsequential to the course of evolution (KIMURA 1983). While, this 'neutral' view of evolution contrasted with the more common panselectionist view of the time, the increasing acceptance of neutral evolution bolstered the argument that natural selection was more important than mutation for evolution: if mutations typically result in neutral

variation within an environment, then there is always ample variation for natural selection to act on after a change in the environment and the effects of mutations can be safely ignored.

The ubiquity of neutral molecular variation meant that molecular changes could not easily be linked to phenotypic differences among species and research shifted to identifying the specific genetic and molecular basis for adaptation. Because proteins are the molecular components directly responsible for most phenotypes, changes in protein structure and function were viewed as likely candidates for adaptation. However, it had long been known that an organism's phenotype can vary upon changes in the environment and, as a consequence, that an organism's phenotype could not be uniquely determined by the specific set of proteins encoded in its genome. In the early 1960's, Jacob and Monod identified a molecular basis for phenotypic change without genetic change through the regulation of protein levels by altering the timing or amount of a gene expression (JACOB and MONOD 1961). In addition, it soon became apparent that many phenotypically diverse organisms were too similar at the protein level for protein coding changes alone to be responsible for their phenotypic difference (ZUCKERKANDL and PAULING 1965; HUBBY and LEWONTIN 1966; LEWONTIN and HUBBY 1966). In particular, the high similarity of numerous human and chimpanzee proteins suggested that phenotypic differences between them were due to changes in the timing and expression of proteins, instead of differences in protein function (KING and WILSON 1975).

Differences in the timing and expression of genes could also account for phenotypic differences observed amongst different cell types in multi-cellular organisms (BRITTEN and DAVIDSON 1969). This had important consequences for the study of development, as the field had largely been cut out of evolutionary biology since the modern synthesis (MÜLLER 2007). While the synthesis combined population thinking with inheritance, it assumed that genetic differences would result in phenotypic differences, and simply disregarded the exact developmental mechanisms by which this happened as unimportant for evolution. However, the growing recognition that changes in expression could underlie phenotypic differences between species combined with the realization that changes in expression were the principle mechanism of development resulted in renewed interest in the study of the processes responsible for generating phenotypes and how they evolved. While developmental biology has a rich history itself, one of the key considerations to come out of the merging of developmental biology with evolution (evo-devo) was the recognition that developmental processes are often biased or constrained (VRBA and ELDREDGE 1984). In turn, these developmental constraints meant that the identity of a mutation was essential for understanding its potential to alter phenotypes, i.e. simply knowing the fitness consequence of a mutation was insufficient for understanding its evolution (HALL 2003). Instead, the molecular mechanism altered by a mutation and the frequency of the same mechanism being altered by other mutations must also be considered in order to understand the course of evolution.

In this way, evo-devo brought the role of mutation in evolution back to the forefront. Within a developmental framework, new mutations were viewed not as passive variants

acted on by natural selection, but as a direct outcome of the underlying biology. In turn, these mutations are acted on by traditional population genetic parameters, such as natural selection and genetic drift, to determine how evolution proceeds. It is thus necessary to combine these two disparate camps to fully understand the evolutionary process (NEI 2013). Despite these insights, the role of new mutation in evolution is still largely outside of the realm of population genetics and traditional molecular evolution (STOLTZFUS 2006). For example, while the role of new mutations in evolution depends on two factors, their rate of occurrence and their phenotypic effects, only the former has been studied extensively, while little empirical data exists for the later. It is thus worth considering what is known about the rate of new mutation and their phenotypic and fitness effects.

The rate of mutation

The rate at which new mutations arise is a fundamental property of the evolutionary process. However, mutation rates vary considerably across species (DRAKE *et al.* 1998). A recent review of mutation rates across all domains of life suggests that the mutation rate varies by approximately three orders of magnitude and is typically lowest in bacteria and archaea (1×10^{-10} to 1×10^{-9} bp^{-1} $generation^{-1}$), higher in single cell eukaryotes (1×10^{-9} to 1×10^{-8} bp^{-1} $generation^{-1}$), and highest in multi-cellular plants and animals (1×10^{-8} to 1×10^{-7} bp^{-1} $generation^{-1}$) (LYNCH 2010). Variation between species in the rate of mutations is further influenced by variation in generation time, as well as the number of cell divisions per generation in multi-cellular organisms (KUMAR and SUBRAMANIAN 2002; SCALLY and DURBIN 2012). Overall, the variation in mutation rate is negatively correlated with effective population size, indicating that genetic drift likely plays a major

role in determining how low a species mutation rate can be maintained by natural selection (LYNCH 2010).

Estimates of mutation rate for many species are, at least in part, based on comparisons of sequence polymorphism levels within species, focusing on the amount of polymorphism at neutral loci and assuming that this frequency is directly proportional to the mutation rate. However, background selection may operate on such sites and rates estimated using this method may be too low on average (LYNCH 2010). An alternative approach that minimizes the effects of natural selection is to use mutation accumulation (MA) lines. By propagating individuals through single organism bottlenecks, the strength of natural selection is reduced and new mutations fix in a single generation after their occurrence. The rate at which phenotypes vary within MA lines can then be used to estimate the mutation rate. As with previous approaches, this method suggests large variation in the mutation rate that spans several orders of magnitude (HALLIGAN and KEIGHTLEY 2009). However, for most phenotypes, relatively few mutations have easily detectable effects and these estimates of mutation rate are often based on small numbers of rare events. More recently, direct sequencing of MA lines has produced highly accurate estimates of mutation rate. These estimates span two orders of magnitude, ranging from approximately 2×10^{-10} bp^{-1} $generation^{-1}$ in *Saccharomyces cerevisiae* (LYNCH *et al.* 2008; ZHU *et al.* 2014) and *Escherichia coli* (LEE and POPODI 2012) to 7×10^{-9} bp^{-1} $generation^{-1}$ in *Arabidopsis thaliana* (OSSOWSKI *et al.* 2010), and encompass estimates from *Caenorhabditis elegans* (DENVER *et al.* 2009), *Mus musculus* (UCHIMURA *et al.* 2015), and *Drosophila melanogaster* (KEIGHTLEY *et al.* 2009).

While the global rate of mutation sets the rate at which new genetic variation is created, not all mutations occur with equal frequency. For example, sequencing of MA lines has revealed that transitions (G/A and C/T mutations) are typically more common than transversions (G/C, G/T, C/A, and A/T mutations), with the exact difference in rate dependent on the specific organism (DENVER *et al.* 2009; OSSOWSKI *et al.* 2010; LEE and POPODI 2012; UCHIMURA *et al.* 2015). Sequencing has also revealed differences in mutation rate based on the direction of mutation. For example, G→A and C→T transitions are often more common than A→G and T→C transitions (DENVER *et al.* 2009; KEIGHTLEY *et al.* 2009; OSSOWSKI *et al.* 2010; LEE and POPODI 2012; UCHIMURA *et al.* 2015), and overall G→A and C→T transitions are the most abundant type of spontaneous point mutations experienced by most organisms (LYNCH *et al.* 2008; ZHU *et al.* 2014).

In addition to differences in mutation rates based on the type of mutation, the rate of mutation is often dependent on local genomic context (SUNG *et al.* 2015). For example, CpG sites typically have higher mutation rates than other sites (OSSOWSKI *et al.* 2010; LEE and POPODI 2012; UCHIMURA *et al.* 2015). In addition, homopolymeric runs of the same base, often A or T, have mutation rates that can be orders of magnitude higher than the background rate (LANG *et al.* 2013) and simple repeat structures, such as microsatellites, are also known to have increased mutation rates (SUN *et al.* 2012). Likewise, the position of bases relative to replication forks (JACKSON *et al.* 2015) and the timing of a bases replication (STAMATOYANNOPOULOS *et al.* 2009; CHEN *et al.* 2010; LANG and MURRAY 2011; AGIER and FISCHER 2011) have been shown to affect local

mutation rates. As a consequence, the mutation rate varies considerably across the genome (HODGKINSON and EYRE-WALKER 2011) and genes can evolve at different rates simply due to differences in the rate of mutational input. Interestingly, recent work suggests that the presence of polymorphism can also influence the occurrence of new mutations in diploids due to a higher rate of mutation near heterozygous sites (YANG *et al.* 2015).

Substitutions are thought to be the majority of all new mutations. However, other types of mutations, such as insertions and deletions (indels), copy number variation (CNV), and aneuploidy, can occur. Sequencing of MA lines indicates that the rate of small indel occurrence (< 50bp) is typically between one and two orders of magnitude lower than for point mutations, ranging from approximately 3% in yeast (ZHU *et al.* 2014) to approximately 50% in *C. elegans* (DENVER *et al.* 2009) the rate of point mutations (OSSOWSKI *et al.* 2010; LEE and POPODI 2012; UCHIMURA *et al.* 2015). While large CNVs are more difficult to detect using short-read sequencing than indels or point mutations, available data suggests that they occur infrequently. For example, a recent study in *S. cerevisiae* identified only three CNVs compared to over 3,100 individual point mutations, suggesting a rate that is approximately three orders of magnitude lower for CNVs than for point mutations (ZHU *et al.* 2014). Interestingly, aneuploidy appears to be quite frequent on a per genome scale, occurring at a rate of 1×10^{-4} genome^{-1} generation^{-1} in a set of yeast MA lines (ZHU *et al.* 2014). However, aneuploidy often has severe fitness costs during meiosis and its role in evolution is unclear (YONA *et al.* 2012; SUNSHINE *et al.* 2015). Taken together, these estimates indicate that mutations other than

substitutions can contribute to the total mutation rate but are unlikely to be dominant sources for new mutations that contribute to evolution.

The phenotypic and fitness consequences of mutations

The probability that a mutation contributes to evolution is jointly determined by its probability of occurring and its probability of fixing in a population. This later probability is determined by the fitness effects of a mutation in conjunction with population genetic parameters such as population size and demography. As a consequence, differences in the rate of mutation and differences in fitness effects of new mutations will both influence the probability that a mutation eventually fixes in a population. However, much less is known about the fitness effects of new mutations than the rate of mutation. Theoretical work suggests that beneficial mutations should be exponentially distributed (EYRE-WALKER and KEIGHTLEY 2007), but empirical evidence suggests a more complicated picture (ROKYTA *et al.* 2008; LEVY *et al.* 2015). More generally, recent theoretical work suggests a specific form for the distribution of fitness effects that is robust to underlying parameter choice (RICE *et al.* 2015). However, empirical data suggests that the distribution of fitness effects for new mutation can vary considerably (WOODRUFF and THOMPSON 1998; ZEYL and DEVISSER 2001; WLOCH *et al.* 2001; EYRE-WALKER *et al.* 2006; MARTIN and LENORMAND 2008). These discrepancies between theoretical and empirical results may arise due to effects of the environment or epistasis (RICE *et al.* 2015) and more empirical data is currently needed.

Because the eventual fate of a mutation is directly influenced by its effect on fitness, it is indirectly influenced by the phenotypes it causes. It is therefore necessary to understand not only the rate and distribution of fitness effects for new mutations, but also the rate and distribution of phenotypic effects of new mutations (NEI 2007). Unfortunately, detailed study of the phenotypic effects of new mutations is rare in the literature (TRAVISANO and SHAW 2013). However, knowledge of the phenotypic effects of new mutations is often indirectly generated through the creation of genotype-phenotype maps (ALBERCH 1991). These maps connect genotypes to their phenotypic effects and thus contain information about the phenotypic consequences of changes in genotype.

The structure of genotype-phenotype maps can reveal much about the evolutionary process. For example, purifying selection applied to empirically determined genotype-phenotypes maps is often capable of producing a wide range of evolutionary phenomenon that typically require positive selection under traditional population genetic models (STADLER 2006; BORENSTEIN and KRAKAUER 2008; PIGLIUCCI 2010; WAGNER 2015). In particular, the structure of the genotype-phenotype map directly determines the extent of pleiotropy, or the number of phenotypes altered by a mutation. The degree of pleiotropy and its distribution have important consequences for evolution because pleiotropy is often invoked as a property that restricts evolution: highly pleiotropic mutations that alter multiple phenotypes are likely to have deleterious effects on some phenotypes and thus be detrimental and unlikely to contribute to evolution (HILL and ZHANG 2012a; b; PAABY and ROCKMAN 2013). However, recent work based on the genotype-phenotype map suggests that pleiotropy is highly modular and that this modularity limits the ability of

pleiotropy to constrain evolution (WANG *et al.* 2010; WAGNER and ZHANG 2011, 2012; ZHANG and WAGNER 2013).

Much of the research on genotype-phenotype maps has focused on a small set of phenotypes, such as RNA structure and protein enzymatic function (PIGLIUCCI 2010). Recent work has also begun to explore genotype-phenotype maps for morphology (HO and ZHANG 2014) and there have been calls to use the extensive knowledge of naturally occurring phenotypic variation to understand genotype-phenotype maps (ROCKMAN 2008). However, there currently remains large gaps in our knowledge about how new mutations alter most phenotypes. In particular, our understanding of how new mutations affect gene expression is extremely limited and many basic questions exist. For example, how often do new mutations alter expression? Are most of these effects on expression large or small? Are the effects of new mutations on expression biased in one direction and how consistent are these biases across genes? What are the evolutionary consequences of the effects of new mutations on gene expression? Creating expectations for these questions requires knowledge of how gene expression functions and what is known about its evolution.

Evolution of gene expression and regulation

Extensive variation in gene expression exists in many systems (e.g. NAGALAKSHMI *et al.* 2008; AYROLES *et al.* 2009; BUSBY *et al.* 2011; Reviewd in WHITEHEAD and CRAWFORD 2006; ALVAREZ *et al.* 2014). Much of this variation in expression is caused by mutations altering gene regulation (ZHENG *et al.* 2011), resulting in heritable differences amongst

individuals for complex traits and the propensity for disease (ALBERT and KRUGLYAK 2015). Because gene expression is important for organismal form, function, and fitness, the evolution of gene expression is expected to be constrained by purifying selection (GILAD *et al.* 2006; HODGINS-DAVIS *et al.* 2015). However, changes in gene regulation are expected to have more nuanced effects on a phenotype than protein coding changes and are considered less deleterious. As a consequence, changes in gene regulation are expected to be a dominant mechanism of adaptation (CARROLL 1995; GERHART and KIRSCHNER 1997).

Gene expression levels are determined by the interaction of *cis*-regulatory elements, such as promoters and enhancers, and *trans*-regulatory factors, such as transcription factors and small RNAs. *trans*-regulatory factors are diffusible and can interact with multiple *cis*-regulatory elements. As a consequence, *trans*-regulatory mutations are expected to alter the expression of multiple genes simultaneously, while *cis*-regulatory mutations are expected to have more restricted effects on expression (PRUD'HOMME *et al.* 2007). Because changes in gene expression are often deleterious, the more restricted nature of *cis*-regulatory changes suggests they play a disproportionately large role compared to *trans*-regulatory changes in the evolution of gene regulation (WRAY *et al.* 2003; CARROLL 2005, 2008; DAVIDSON and ERWIN 2006; WRAY 2007; ERWIN and DAVIDSON 2009).

The hypothesis that *cis*-regulatory mutations are the dominant mechanism for evolution has attracted criticism (COYNE 2006). For example, a comparison of case studies

identifying the genetic basis of adaptation suggested that *cis*-regulatory changes were no more likely to underlie adaptation than changes within protein coding regions (HOEKSTRA and COYNE 2007). However, continued effort to determine the molecular mechanism of adaptation suggest that part of the difficulty in addressing this question is due to methodology: identifying adaptive protein coding changes is typically easier than identify adaptive *cis*-regulatory changes and direct comparisons of the frequency of case studies is thus biased towards protein coding changes (STERN and ORGOGOZO 2008, 2009; STREISFELD and RAUSHER 2010; MARTIN and ORGOGOZO 2013). Unfortunately, the extent of bias in detecting different mechanisms of adaptation is difficult to quantify and it is often unclear how to interpret the growing number of case studies.

More recently, several studies have taken advantage of repeated adaptation within a system to determine the relative frequency of different mechanisms of adaptation. For example, a study focused on repeated adaptation in sticklebacks has suggested that while both coding and regulatory changes are important for adaptation, regulatory changes are the dominant mechanism for adaptation (JONES *et al.* 2012). However, studies of the molecular basis of repeated adaptation often find that the dominate mechanism of adaptation varies across systems, but can be highly repeatable within systems (GOMPEL and PRUD'HOMME 2009; CONTE *et al.* 2012; LONFAT *et al.* 2014). This high repeatability of mechanism within a system suggests that there are large biases in either the generation of phenotypes by the mutational process or in the action of natural selection. For changes in gene regulation, both mutational and selection biases reflect the structure of the underlying gene regulatory network.

Regulatory network architecture and the effects of pleiotropy

Gene expression is controlled by a complex network of interacting *cis*-regulatory elements and *trans*-regulatory factors (LEE *et al.* 2002). The organization of regulatory interactions often consists of stereographed patterns in the form of network motifs that can have specific functional consequences on expression (SHEN-ORR *et al.* 2002; LEE *et al.* 2002; MILO *et al.* 2002; Reviewed in ALON 2007). While numerous techniques exist for reconstructing network and motif structure (SCHLITT and BRAZMA 2007), including transcriptional profiling upon mutation (HU *et al.* 2007) and comparisons of transcriptomic and proteomic data (HUANG and FRAENKEL 2009), the techniques vary considerably in their accuracy, and building high quality regulatory networks remains challenging even for 'simple' organisms (MA *et al.* 2014).

Regulatory network structure has important implications for the evolution of gene expression. In particular, the degree of pleiotropy, or the number of phenotypes likely to be affected by altering a gene's expression, varies across a regulatory network. This variation in pleiotropy emerges naturally from regulatory network structure: a small proportion of genes (hubs) interact with many other genes and have high pleiotropy, while most genes have relatively few interactions with other genes and have low pleiotropy (TYLER *et al.* 2009). This structure has two main consequences for evolution. First, because new mutations are unlikely to disrupt hub genes due to their scarcity and have a low probability of effecting a phenotype when they occur in non-hub genes, regulatory networks are relatively robust to the effects of new mutations and most new

regulatory mutations are expected to affect the expression of only a few genes (BABU *et al.* 2004; LUSCOMBE *et al.* 2004; YU and GERSTEIN 2006).

Second, genes more centrally located within a regulatory network are unlikely sources of complex phenotypic and developmental change (ERWIN and DAVIDSON 2009). Because changes in gene expression are often deleterious, mutations that alter the expression of many genes, i.e. are highly pleiotropic, are expected to be deleterious and therefore make little contribution to adaptation. For example, transcription factors often show more stable expression patterns than the downstream genes that they regulate (RIFKIN *et al.* 2003). As a result of network structure, *trans*-regulatory changes affecting expression of a gene are expected to be more pleiotropic than *cis*-regulatory changes affecting expression of the same gene. One consequence of this difference in pleiotropy is that *trans*-regulatory interactions are often more conserved than specific *cis*-regulatory elements (HABIB *et al.* 2012). For example, the *trans*-regulatory network in humans and mice has remained relatively constant over time, even though there has been substantial turnover within individual *cis*-regulatory sequences (VIERSTRA *et al.* 2014; STERGACHIS *et al.* 2014). These results are consistent with *cis*-regulatory mutations being a dominant source of regulatory variation.

Evolution of *cis*-regulatory elements

There are several *cis*-regulatory elements that influence a gene's expression and evolution. One of the main determinants of *cis*-regulatory activity is the sequence of the core promoter, particularly the presence of a canonical TATA box sequence

(TATAWAWR). While nearly all genes are predicted to have a TATA like sequence in their promoter, the degree of degeneracy varies considerably (RHEE and PUGH 2012). The sequence of the TATA box has several consequences for gene expression. For example, in yeast, genes with a canonical TATA box sequences preferentially utilize SAGA over TFIID for transcriptional initiation and are more responsive to stress (HUISINGA and PUGH 2004; BASEHOAR *et al.* 2004). Similarly, genes with canonical TATA box sequences show greater variability in expression across species and environments (TIROSH *et al.* 2006; LANDRY *et al.* 2007b). However, TATA box sequence does not deterministically influence gene expression; the location of the TATA box, its orientation, and the sequence of nearby nucleotides influence TATA box activity and gene expression levels (LUBLINER *et al.* 2015). As a consequence, a wide variety of expression levels can be generated simply by altering the placement and strength of the TATA box.

cis-regulatory sequences also influence gene expression levels through the presence, positioning, and chromatin status of nucleosomes. Promoters containing many well positioned nucleosomes are typically transcriptionally silent, while those containing fewer nucleosomes, or less precisely positioned nucleosomes, are often transcriptionally active (ZAUGG and LUSCOMBE 2011). A canonical TATA box sequence often results in well positioned nucleosomes and thus lower transcription. However, nucleosomes can be moved through the action of chromatin remodelers (TIROSH and BARKAI 2008) and TATA box containing promoters are often responsive to environmental change (LANDRY *et al.* 2007b). In addition to the TATA box, positioning of nucleosomes is influenced by

the presence of long polyA:T tracts which favor the binding of nucleosomes (RAVEH-SADKA *et al.* 2012; BROGAARD *et al.* 2012). Combined, these effects suggest that *cis*-regulatory factors are the major determinant of nucleosome status within a promoter. This is consistent with experimental evidence suggesting that differences in nucleosome positioning between yeast species is largely determined by *cis*-regulatory sequence (TIROSH *et al.* 2010; HUGHES *et al.* 2012). However, *trans*-regulatory factors can have important effects on the spacing of nucleosomes (HUGHES *et al.* 2012) . In addition, specific epigenetic marks within chromatin can influence expression (XIAO *et al.* 2012) and the combination of chromatin remodelers present in a cell is predictive of expression level (LI *et al.* 2013; KEUNG *et al.* 2014).

Variation in the binding sites for transcription factors (TFs) within *cis*-regulatory elements are a major source of variation in gene expression. While TFs have similar intrinsic DNA binding affinities as nucleosomes, transcriptional activators can displace nucleosomes and increase expression (CHAROENSAWAN *et al.* 2012). This displacement is possible because TFs exhibit large variation in their affinities towards different sequences (MAERKL and QUAKE 2007; BADIS *et al.* 2009; GORDÂN *et al.* 2011; PAYNE and WAGNER 2014b). As a consequence, TFs recognize and bind specific sequences within a *cis*-regulatory element. These specific binding interactions can result in the remodeling of chromatin, are crucial for transcriptional activation, and are thus a major determinant of variation in gene expression levels.

In addition to sequence specific effects, TF binding sites also regulate expression based on position and number (LEVO and SEGAL 2014). These relationships are often idiosyncratic and dependent on the specific TF (SHARON *et al.* 2012). While the core binding sequence is the most important for TF affinity, sequences directly outside of the core binding site can also influence TF binding (LEVO *et al.* 2015). In addition, binding sites for distinct TFs show spacing that is consistent with the physical interaction of the transcription factors and specific TF-TF interactions are often important for proper regulation (YOKOYAMA *et al.* 2009). While TFs can act as barriers to nucleosome repositioning, they can also be evicted by nucleosomes (LI *et al.* 2015). Thus, the positioning of TF binding sites relative to preferred nucleosome binding sites can influence gene expression levels and dynamics.

The mechanism by which gene expression evolves is closely related to the properties of TF binding sites. In eukaryotes, individual TF binding sites are typically 10 nucleotides long. This length represent a tradeoff in specificity and robustness, as 10 nucleotide sequences are rarely generated by chance, but longer sequences are more easily degraded by new mutations (STEWART and PLOTKIN 2012). As a consequence of this relatively short sequence, many related sequences have similar binding affinities for the same TF, resulting in a neutral network of TF binding where many mutations have little to no phenotypic or fitness effect (PAYNE and WAGNER 2014a; b). This structure gives TF binding sites a high robustness to mutation, but also makes them less constrained than other sequences, resulting in relatively high turnover (BULLAUGHEY 2011). In addition, because of the high degeneracy in binding sites, selection for the most preferred sequence

is weak and the majority of TF binding sites will be closely related to the preferred sequence instead of identical to it (LYNCH and HAGNER 2014). As a consequence, identifying TF binding sites by conservation alone is often challenging, especially when the exact position of a binding site is not conserved (KALAY and WITTKOPP 2010). These results differ substantially from the binding affinity of the TFs themselves which, on average, evolve considerably slower (BOYLE *et al.* 2014).

Evolution of *trans*-regulatory factors

While regulatory network structure and pleiotropy suggest that mutations in *trans*-acting factors should be less likely to underlie regulatory changes than mutations in *cis*-regulatory elements, it does not preclude such changes (WAGNER and LYNCH 2008; LYNCH and WAGNER 2008). For example, *trans*-regulatory changes in expression can occur involving multiple *trans*-acting factors (LAVOIE *et al.* 2010) as well as the incorporation of new *trans*-acting factors into the already existing networks (BOOTH *et al.* 2010). Recent work has begun to highlight the detailed molecular mechanisms that are necessary for *trans*-regulatory changes in expression (Reviewed in LI and JOHNSON 2010; VOORDECKERS *et al.* 2015). These experiments suggest that while gene duplication appears to be important for regulatory network growth (TEICHMANN and BABU 2004) and is necessary in some examples (BAKER *et al.* 2013; POUGACH *et al.* 2014; MCKEOWN *et al.* 2014), it is not required (LYNCH *et al.* 2011; SAYOU *et al.* 2014; but see BROCKINGTON *et al.* 2015; BRUNKARD and RUNKEL 2015). Likewise, promiscuity in the ancestral *trans*-acting factor towards multiple *cis*-regulatory sequences is often crucial for new *trans*-regulatory interactions to evolve (BAKER *et al.* 2013; SAYOU *et al.* 2014; POUGACH *et al.*

2014; MCKEOWN *et al.* 2014; ANDERSON *et al.* 2015). In some cases, sub-functionalization of the regulatory network from a previously promiscuous ancestor has occurred (BAKER *et al.* 2013; POUGACH *et al.* 2014), while other cases suggest neo-functionalization through the gain of new binding sites (LYNCH *et al.* 2011; SAYOU *et al.* 2014; MCKEOWN *et al.* 2014). Taken together, these results suggest that the evolution of *trans*-regulatory factors is more flexible than originally thought (WAGNER and LYNCH 2008; LYNCH and WAGNER 2008).

While transcriptional activity is often sequence dependent, there is considerable variation in the transcripts produced and this variation is largely mediated by *trans*-acting factors. For example, while the transcription start site is influenced by nucleosome positioning, and thus *cis*-regulatory sequences (DE BOER *et al.* 2014), the exact position of the transcription start site is often dependent on the binding of specific TFs (CHENG *et al.* 2012). Similarly, the 3' end of transcription is often variable and controlled by specific TF interactions. These differences in the 5' and 3' location of transcription can result in large numbers of isoforms. For example, in yeast, which contain only a few genes with introns and little alternative splicing, the average number of isoforms per gene is predicted at over 20 based solely on 5' and 3' end variation (PELECHANO *et al.* 2013). On top of this variation, the length of the poly-A tail is often variable and is one mechanism by which mRNA decay is controlled through *trans*-regulatory factors (MOQTADERI *et al.* 2013). Many promoters also produce expression in a bidirectional manner, with one direction producing the primary transcript, while the other produces shorter transcripts at lower levels. It has been proposed that these secondary transcripts act in *trans* and have

regulatory activity in some cases (NEIL *et al.* 2009; XU *et al.* 2009). Finally, the redundancy of *cis*-regulatory sequences means that mutations can easily create new binding sites for transcription factors. As a consequence, *cis*-regulatory elements transiently binding larger numbers of transcription factors and there is widespread cross talk between TFs and *cis*-regulatory sequences (HARBISON *et al.* 2004; MACISAAC *et al.* 2006). Interestingly, this property of regulatory networks can easily result in new expression patterns in response to new environmental stimuli or in new cell types and is therefore a source for novel phenotypes (PAYNE and WAGNER 2014a). Overall, these studies suggest that *trans*-regulatory changes in expression can occur through a number of mechanisms.

The evolution of gene expression noise

It has long been recognized that heritable differences in gene expression and regulation do not fully account for observed differences in expression between individuals (KO 1991). A major source of this additional variability in expression is the stochastic fluctuations in expression amongst genetically identical cells in the same environment, or gene expression noise (BLAKE *et al.* 2003). Like other sources of expression variation, expression noise can result in phenotypic variability and it is thus important to understand the sources and consequences of expression noise on the evolution of gene expression (RASER and O'SHEA 2005; GEILER-SAMEROTTE *et al.* 2013).

Gene expression noise is caused by stochastic fluctuations in the rates of transcription, translation, mRNA decay, and protein decay (BROWN *et al.* 2013). However, both

theoretical and experimental work supports variation in transcription as the primary source of expression noise (KEPLER and ELSTON 2001; SWAIN *et al.* 2002; JONES *et al.* 2014). This variation in transcription has two sources, burst size and burst frequency. First, to produce an mRNA molecule, a gene's promoter must be in a transcriptionally active state. The switching between active and inactive states is dependent on the binding and positioning of nucleosomes and TFs within the *cis*-regulatory sequence. Because these events are stochastic, the transcriptional status of a promoter fluctuates (BROWN *et al.* 2013). In addition, because TF binding is concentration dependent, expression noise in TF levels will affect the frequency of the promoter moving in and out of the transcriptionally active state, contributing to variation in expression (ELOWITZ *et al.* 2002; SWAIN *et al.* 2002). Together, these processes influence the frequency at which a promoter region fires, or burst frequency. Second, once a promoter is in a transcriptionally active state, RNA polymerases must bind and transcription must be initiated. These processes are also stochastic, and the number of mRNA molecules created per transcription burst determines burst size. Together, burst frequency and burst size determine both the mean level of expression and the amount of expression noise.

Gene expression noise varies amongst genes (NEWMAN *et al.* 2006; RICHARD and YVERT 2014) and is particularly sensitive to a gene's promoter sequence (SANCHEZ *et al.* 2013). Within a promoter, one of the main determinants of expression noise is the presence of nucleosome disfavoring sites. These sites stabilize the promoter in a transcriptionally active state and reduce switching of promoter states and therefore reduce expression noise (KIM and MARIONI 2013; SHARON *et al.* 2014). Interestingly, unregulated

promoters have low expression noise be default (WOLF *et al.* 2015) and it is the addition of regulatory elements that results in increased expression noise. This is because each new element imparts its own stochastic fluctuations on expression towards the expression of the target gene (PEDRAZA and PAULSSON 2008; WOLF *et al.* 2015). As a consequence, increased regulation and TF binding tend to increase expression noise (DADIANI *et al.* 2013; SHARON *et al.* 2014).

Because expression noise is largely determined by promoter sequence, *cis*-regulatory mutations can alter gene expression noise. While a gene's mean level of expression and its expression noise are often correlated, with higher mean expression typically associated with lower transcriptional noise (BAR-EVEN *et al.* 2006), mutations can affect mean expression and expression noise independently (THATTAI and VAN OUDENAARDEN 2001; MUNSKY *et al.* 2012). For example, because the TATA box sequence has a large influence on nucleosome positioning, mutations within the TATA box sequence have been specifically implicated in altering expression noise independently of the mean level of expression (MURPHY *et al.* 2010; HORNUNG *et al.* 2012). Interestingly, mutations in different parts of a promoter appear to affect expression noise in different ways. For example, mutations throughout most of the promoter have been shown to primarily affect expression noise through burst frequency. By contrast, mutations in the TATA box or that alter nucleosome binding primarily affect expression noise through changes in burst size (HORNUNG *et al.* 2012). Data also suggests that expression noise at low TF concentrations is primarily due to variation in burst frequency because of poor stabilization of the active transcriptional state, while at higher TF concentrations,

promoters are almost always in an active state and expression noise is primarily determined by burst size (CAREY *et al.* 2013). These results suggest that there is a large degree of flexibility in how gene expression noise can evolve.

Because expression noise can alter phenotypes it can be acted on by natural selection. Theory suggests that increased expression noise is generally deleterious and thus under purifying selection (FRASER *et al.* 2004). Consistent with this view, essential and highly conserved genes show lower levels of expression noise than other genes (SILANDER *et al.* 2012). The total fitness cost due to expression noise at all genes is substantial, and one of the main fitness benefits to genome duplication is a reduction in expression noise (WANG and ZHANG 2011). Combined with the mechanisms that produce expression noise, this presents an interesting evolutionary dilemma; increased binding to TFs can result in specific regulation, and thus fitness benefits, but it carries the associated cost of increased noise due to additional regulation. As a consequence, there are tradeoffs in the evolution of gene regulation and precise control of both mean expression and expression noise may not be possible or beneficial (WOLF *et al.* 2015). In addition, this selection to minimize noise can have further consequences for the evolution of gene expression by limiting the evolvability of expression due to pleiotropic effects on expression noise (LEHNER 2008, 2010; WOLF *et al.* 2015).

While expression noise is typically deleterious, it can be beneficial under specific circumstances. In particular, expression noise can be beneficial as a bet-hedging strategy to ensure that at least some individuals have the appropriate level of expression upon a

change in the environment (VARDI *et al.* 2013). This strategy is particularly effective when the environment fluctuates in non-deterministic ways. In addition, increased expression noise has been linked to higher evolvability of gene expression and under circumstances in which evolvability is beneficial, increased expression noise is also expected to be beneficial (ZHANG *et al.* 2009).

Finally, because expression noise is often deleterious, strategies for alleviating expression noise have evolved (CHALANCON *et al.* 2012). For example, gene's which repress their own expression have relatively low expression noise and repressor cascades are often less noisy than activator cascades (ALON 2007). In multicellular organisms, expression noise for extracellular products is often reduced due to spatial and temporal averaging across multiple cells (LITTLE *et al.* 2013). In a related manner, averaging inputs from multiple sources can help to reduce expression noise (FRANK 2013). Finally, gene expression can be controlled by multiple *cis*-regulatory elements, such as multiple enhancers, that together drive more robust expression than either enhancer alone (FRANKEL *et al.* 2010). Taken together, there is considerable evidence that gene expression noise is a fundamental component of gene expression and that its presence has large influences on the evolution of gene expression.

The relative contributions of *cis* and *trans*-regulatory changes to evolution

Because both *cis* and *trans*-regulatory changes can result in heritable differences in gene expression, it is important to understand the relative contributions of these mechanisms to regulatory change within natural populations (WITTKOPP 2005). The contributions of *cis*

and *trans*-regulatory changes to differences in gene expression can be determined mechanistically: because *cis*-regulatory sequences in the same cell share a *trans*-regulatory environment, *cis*-regulatory changes can be directly identified by measuring differences in allele specific expression (ASE) within a hybrid between two strains or species (COWLES *et al.* 2002). In addition, because differences in expression between parental strains or species are caused by the combined effects of *cis*- and *trans*-regulatory changes, differences in ASE in hybrids and expression differences amongst the parents can be attributed to *trans*-regulatory changes (WITTKOPP *et al.* 2004).

Numerous studies have used this approach to determine the relative contributions of *cis*- and *trans*-regulatory changes to differences in expression. Within species, studies suggest that both *cis* and *trans*-regulatory changes make substantial contributions to variation in expression (WITTKOPP *et al.* 2004; WANG *et al.* 2007; GENISSEL *et al.* 2008; ZHANG and BOREVITZ 2009; GE *et al.* 2009; SUNG *et al.* 2009; MCMANUS *et al.* 2010; EMERSON *et al.* 2010; PASTINEN 2010; GONCALVES *et al.* 2012; SCHAEFKE *et al.* 2013). Interestingly, however, between species, *cis*-regulatory changes typically contribute more to regulatory divergence than *trans*-regulatory changes (LANDRY *et al.* 2005; ZHUANG and ADAMS 2007; WITTKOPP *et al.* 2008a; TIROSH *et al.* 2009; SHI *et al.* 2012; LEMMON *et al.* 2014). In particular, *cis*-regulatory changes appear to preferentially accumulate with divergence time (WITTKOPP *et al.* 2008a; BULLARD *et al.* 2010; GORDON and RUVINSKY 2012; SCHRAIBER *et al.* 2013; COOLON *et al.* 2014).

One possible explanation for the accumulation of *cis*-regulatory changes with divergence time is preferential retention of *cis*-regulatory changes over *trans*-regulatory changes by natural selection. For example, *cis*-regulatory changes tend to be additive, while *trans*-regulatory change tend to be dominant or recessive (LEMOS *et al.* 2008; EMERSON *et al.* 2010; COOLON *et al.* 2014). Because natural selection is more efficient on additive beneficial mutations than recessive or dominant beneficial mutations, difference in additivity could cause *cis*-regulatory changes to preferentially accumulate with time (GRAZE *et al.* 2012). Another, not mutually exclusive, possibility is that *trans*-regulatory mutations are preferentially selected against. This could arise due to differences in pleiotropy between *cis* and *trans*-regulatory mutations (STERN 2000; WITTKOPP 2005). However, there is little direct evidence of differences in pleiotropy between *cis* and *trans*-regulatory mutations.

In addition to differences in expression and regulation, these studies often identify genes that are mis-expressed within the hybrids. While mis-expression can be caused by heterochronic shifts in expression (RENAUT *et al.* 2009; LENZ *et al.* 2014), it is most often thought to arise from interactions amongst *cis* and *trans*-regulatory changes within two strains or species when the hybrid is created (LANDRY *et al.* 2005, 2007a). These interactions can reveal compensatory changes in expression where *cis* and *trans*-regulatory changes cancel out, resulting in regulatory divergence with little expression divergence (KUO *et al.* 2010; GORDON and RUVINSKY 2012; GONCALVES *et al.* 2012). Interestingly, if *trans*-regulatory changes in one parental strains or species interact less with the *cis*-regulatory changes in the other strain or species, the outlined approach to

classification of *cis* and *trans*-regulatory changes can be misleading. This is because any *trans*-regulatory change that preferentially interacts with its own *cis*-regulatory allele within a hybrid will be incorrectly interpreted as a *cis*-regulatory change (TAKAHASI *et al.* 2011). However, the extent to which *cis*-regulatory activity depends on the *trans*-regulatory background is currently unclear (WITTKOPP *et al.* 2008b; COOLON *et al.* 2013) and more data is needed.

While such studies can determine genome wide patterns of *cis* and *trans*-regulatory changes, knowledge of the specific genes and nucleotides causing changes in expression are needed to identify more specific molecular mechanisms. Mapping of genomic regions underlying differences in expression (eQTL mapping) can provide this information. These studies confirm that both local (*cis*-) regulatory changes and distal (*trans*-) regulatory changes contribute to differences in expression in natural populations (BREM *et al.* 2002; YVERT *et al.* 2003; KULP and JAGALUR 2006). Interestingly, these studies suggest that *cis*-regulatory changes are often larger than *trans*-regulatory changes (SCHADT *et al.* 2003) and in some cases more numerous (GILAD *et al.* 2008). However, the exact frequencies are difficult to determine (MORLEY *et al.* 2004), in part because many eQTL are expected to have small effects on expression and are therefore difficult to reliably detect (BREM and KRUGLYAK 2005).

The identification of eQTL links genotype to phenotype, and eQTL can therefore be used to determine regulatory network structure (HANSEN *et al.* 2008). However, recent studies have indicated that the genotype-phenotype link is complicated by several factors. For

example, eQTL can affect not only the level of expression, but also the dynamics of expression, making the stage in which phenotyping is performed crucial for eQTL identification (ACKERMANN *et al.* 2013). Likewise, eQTL can have different effects in different tissues and can be tissue specific (FLUTRE *et al.* 2013; CONSORTIUM 2015). In addition, an eQTL can affect expression in one environment, but not in another, and it is unclear the extent to which eQTL identified in laboratory environments influence expression in natural populations (GIBSON and WEIR 2005; SMITH and KRUGLYAK 2008). Finally, eQTL can affect expression outside of average transcription levels. For example, eQTL have been identified that alter gene expression noise (HULSE and CAI 2013) and post-transcriptional effects on expression (FAZLOLLAHI *et al.* 2014).

The presence of post-transcriptional effects on gene expression level is potentially problematic for the identification of eQTL: for methodological reasons the study of mRNA abundance is typically mush easier than the study of protein abundance. As a consequence, most eQTL studies map the genetic basis of variation in mRNA levels. This problem is particularly disconcerting in light of evidence that there is a relatively poor correlation between mRNA expression levels and protein expression levels (GYGI *et al.* 1999; GREENBAUM *et al.* 2003; GHAEMMAGHAMI *et al.* 2003; FOSS *et al.* 2011). In addition, early eQTL for protein abundance had little overlap with eQTL affecting mRNA levels (FOSS *et al.* 2007). Together, these results suggested that much of the variation in gene expression levels between individuals was due to post-transcriptional processes. However, more recent work suggests that at least 70% of protein expression levels are determined at the transcriptional stage (LU *et al.* 2007). In addition, larger

eQTL mapping studies for protein eQTL indicate that there is much greater overlap with mRNA eQTL than initially thought, with the discrepancy in part arising from differences in the effect size of protein eQTL and mRNA eQTL (PRITCHARD and GILAD 2013; ALBERT *et al.* 2014b; PARTS *et al.* 2014; PAI *et al.* 2015). Finally, allele specific measurements of protein abundance or translation are typically consistent with allele specific measurements of mRNA levels, suggesting that post-transcriptional effects are relatively small (KHAN *et al.* 2012; ARTIERI and FRASER 2013; MCMANUS *et al.* 2014; ALBERT *et al.* 2014a). Overall, a consensus is emerging that mRNA abundance is the primary determinant of gene expression levels, with post-transcriptional steps having secondary, but occasionally important, effects. Taken together, eQTL studies and allele specific expression measurements support a large role for *cis*-regulatory changes affecting mRNA abundance as the dominant mechanism affecting gene expression levels in natural populations (BABAK *et al.* 2010; LAGARRIGUE *et al.* 2013).

The effects of mutation and a test for natural selection on gene expression

Much of our knowledge about the evolution of gene expression comes from studies of naturally occurring variation. However, the evolutionary mechanisms responsible for this variation are still unclear. In particular, it remains unknown what the relative contributions of mutation and selection are towards the patterns of regulatory variation observed in natural populations. This is because both mutation and selection have acted to produce variation in gene regulation and their effects cannot be separated by studying natural variation alone. Instead, addressing this question requires knowledge of the effects on gene expression of new mutations in the absence of natural selection.

Unfortunately, our knowledge of mutational effects on gene expression is limited (CHARLESWORTH 2013; HODGINS-DAVIS *et al.* 2015).

What little data is available suggests that a wide range of expression patterns can be generated through small changes in *cis*-regulatory sequences (MAYO *et al.* 2006). These changes in expression are often due to changes in TFBS (SHULTZABERGER *et al.* 2010), but can be located throughout a promoter or enhancer (PATWARDHAN *et al.* 2009, 2012; KWASNIESKI and MOGNO 2012; MELNIKOV *et al.* 2012). In general, promoter architecture appears to play a large role in determining the effect of new *cis*-regulatory mutations (LANDRY *et al.* 2007b; ROSIN *et al.* 2012). By contrast, the effects of *trans*-regulatory changes appear more limited in effect (MAERKL and QUAKE 2009), and the regulatory network structure appears to provide robustness to many *trans*-regulatory changes (DENBY *et al.* 2012). These results suggest that new *cis*-regulatory mutations may have larger effects on expression than new *trans*-regulatory mutations.

A recent study attempted to address this question by determining the frequency and effects of new regulatory mutations on the expression of the *S. cerevisiae TDH3* gene (GRUBER *et al.* 2012). They found that *trans*-regulatory mutations were considerably more common than *cis*-regulatory mutations. Because *TDH3* is one of the most highly expressed genes in the genome (MCALISTER and HOLLAND 1985; GHAEMMAGHAMI *et al.* 2003), most new mutations were expected to decrease expression. Surprisingly, however, *trans*-regulatory mutations were biased towards increased *TDH3* expression.

Unfortunately, *cis*-regulatory mutations were too rare to perform a direct comparison of effect sizes, and the relative magnitude of effect for *cis* and *trans*-regulatory mutations remains unknown.

In addition to questions about the effect size of *cis* and *trans*-regulatory mutations, there remains questions about whether natural selection acts on *cis* and *trans*-regulatory variation in the same manner. To address this question, a test for the action of natural selection within regulatory regions is needed. However, such tests are in their infancy (FAY and WITTKOPP 2008). One common approach is to identify conserved sequences within regulatory regions and compare the frequency or type of change within these conserved regions to random expectations (MOSES *et al.* 2003; ANDOLFATTO 2005; ROMERO *et al.* 2012). However, given that such an approach first requires sequence conservation, it is largely limited to cases of strong purifying selection or recent changes in expression. Because many regulatory regions are under only weak purifying selection, such an approach is likely to miss natural selection in many regulatory elements. An alternative approach is to compare the extent of expression variation or transcription factor binding within species (as a proxy for neutral changes in expression) to the extent of variation between species (NUZHDIN *et al.* 2004; MOSES 2009). However, this approach requires that selective and mutational constraints are similar between species and it is unclear how often this is the case. Finally, while both of these tests can occasionally detect the action of natural selection, they rarely can identify the mechanistic basis for this selection.

Because of regulatory network structure, altering the expression of a single downstream gene or pathway can often be accomplished by altering the expression of multiple genes. Thus, one signature of positive selection on gene expression levels is consistent shifts in gene expression of functionally related genes (FRASER *et al.* 2010, 2012; BULLARD *et al.* 2010; MARTIN *et al.* 2012; CHANG *et al.* 2013). Extending this approach to multiple species can also identify instances of purifying selection on gene expression (SCHRAIBER *et al.* 2013). However, it is unclear how often genes in the same pathway should have altered expression in the same direction due to natural selection. For example, gene expression is often controlled by both activators and repressors and natural selection on a downstream gene's expression level is expected to alter these gene's expression in opposite directions. In addition, this approach implicitly assumes that increases and decreases in gene expression are equally likely. However, if the mutational process is biased in its direction of effect, then consistent changes in gene expression in the same direction could result from relaxation of selection and not positive selection.

Finally, in the absence of natural selection, changes in gene expression in natural populations are expected to be consistent with the effects of new mutations on expression. As a consequence, comparing the effects of new mutations and polymorphisms on expression can be used to test for the presence of natural selection. If the effects of mutation and polymorphism are similar, then there is no evidence of selection acting on expression levels. By contrast, if the effects of mutations and polymorphism differ from one another, how they differ points to the target and mechanism of natural selection (RICE and TOWNSEND 2012). Such an approach has been

used by comparing the effects of mutation accumulation lines on gene expression to natural variation within *Caenorhabditis elegans*, finding widespread evidence for purifying selection on expression (DENVER *et al.* 2005). Related approaches based on the direction of effect of mutations and polymorphism within regulatory elements for primates and rodents suggests the presence of positive selection on expression (SMITH *et al.* 2013). The logic underlying this approach can also be used to identify selection within coding regions (STOLTZFUS and YAMPOLSKY 2009). Applying this test to *cis* and *trans*-regulatory effects individually could reveal differences in the action of natural selection.

Thesis overview

In this thesis I examine the relative roles of mutation and selection on the evolution of gene expression within the *Saccharomyces* genus. I first focus on creating and characterizing a set of genetically tractable strains derived from wild isolates of *S. cerevisiae* for experimental work. I then compare the effects on gene expression of naturally occurring variants and *de novo* mutations to disentangle the contributions of mutation and selection on the evolution of gene expression, focusing separately on *cis*- and *trans*-regulatory changes. These comparisons indicate a substantial role for mutation in the evolution of gene expression and I next focus on differences in the frequency and effects of *de novo cis*- and *trans*-regulatory mutations. Finally, I compare the long term evolutionary patterns of gene expression and regulatory change across *Saccharomyces* species, incorporating the evolutionary implications of the earlier work to build more complete models of how gene expression and regulation evolves.

The data in chapter two lay the groundwork for subsequent chapters. It focuses on the extent of genetic diversity segregating within dozens of wild and domesticated *Saccharomyces cerevisiae* strains using whole genome sequencing and describe the steps needed to generate genetically tractable versions of these strains for use in the laboratory. Using these strains, I determine the extent of phenotypic variation in several environments and perform a genome wide association study (GWAS), linking specific genomic regions to naturally occurring phenotypic differences. Supplementary information for this chapter is found in Appendix A.

Chapter three compares the effects on expression of *cis*-regulatory polymorphisms and *de novo* mutations for a single yeast gene, *TDH3*. Using ancestral state reconstruction of the *TDH3* promoter, I determine the likely evolutionary history of the promoter and the effects on expression of each polymorphism as it occurred. Comparison of these effects with the effects of hundreds of *de novo* mutations within the *TDH3* promoter revealed that there is little evidence of selection acting to maintain the mean level of *TDH3* expression. Instead, selection retained polymorphisms that maintained low levels of *TDH3* expression noise. These results were caused by differences in the distribution of mutational effects for mean expression and expression noise and not the action of natural selection, highlighting one way in which the mutational process can directly influence the course of evolution. Supplementary information for this chapter is found in Appendix B.

Chapter four compares the effects on *TDH3* expression of naturally occurring and *de novo trans*-regulatory changes. I show that natural selection on *trans*-regulatory changes

has acted on both the mean level of expression and on expression noise. However, the effects of natural selection on expression noise are in the opposite direction from that identified for *cis*-regulatory changes, suggesting purifying selection. I then focus on a novel technique for mapping naturally occurring *trans*-regulatory variants. Applying this technique to *S. cerevisiae* revealed that segregating *trans*-regulatory variants underlying *TDH3* expression is highly polygenenic, caused by variation at hundreds of unique sites throughout the genome. Most of these loci are not shared between strains and the majority of the yeast genome is physically linked to at least one variant influencing expression. Appendix C contains technical improvements for high throughput mapping related to this chapter. Appendix D contains theoretical and practical considerations for using the same technique when mapping *de novo* mutations.

The analysis in chapter five centers on the effects of the mutational process and compares the frequencies and effects of *de novo cis*- and *trans*-regulatory mutations on *TDH3* expression (Appendix E contains preliminary data on differences in pleiotropy between *cis*- and *trans*-regulatory changes). I show that while *trans*-regulatory mutations are more common than *cis*-regulatory mutations, *cis*-regulatory mutations have on average larger effects on expression. These patterns are consistent with observations of the frequencies and effects of *cis*- and *trans*-regulatory effects in natural populations, suggesting that these patterns may in part be the result of mutational pressure and not natural selection. In addition, I show that the effects on *TDH3* expression of *cis* and *trans*-regulatory mutations are skewed in opposite directions, suggesting that compensatory changes in

gene expression can arise through non-adaptive processes. Supplementary information for this chapter is found in Appendix F.

Chapter six explores the long term evolutionary patterns of expression and regulatory changes in the genus *Saccharomyces*. The results suggest that gene expression evolves largely under a model of compensation, such that underlying regulation can change drastically, even though the levels of gene expression remain relatively constant. In particular, this chapter highlights the potential for widespread compensation in expression due to counteracting *trans*-regulatory changes, a mechanism of expression and regulatory evolution rarely considered. I provide a simple model of regulatory evolution that is capable of explaining the observed data.

Data in chapters two through six highlight how *cis* and *trans*-regulatory evolution changes over time, focusing specifically on the role of differences in the mutational process in driving observed patterns in natural populations. Chapter seven focuses on the implications of these results for understanding the patterns of gene expression observed in natural populations. In particular, I discuss the role of gene expression noise, the relative contributions of *cis* and *trans*-regulatory changes, and the consequences of compensatory changes in regulation on the evolution of gene expression.

References

ACKERMANN M., SIKORA-WOHLFELD W., BEYER A., 2013 Impact of natural genetic variation on gene expression dynamics. PLoS Genet. **9**: e1003514.

AGIER N., FISCHER G., 2011 The mutational profile of the yeast genome is shaped by replication. Mol. Biol. Evol. **29**: 905–13.

ALBERCH P., 1991 From genes to phenotype: dynamical systems and evolvability. Genetica **84**: 5–11.

ALBERT F. W., KRUGLYAK L., 2015 The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. **16**: 197–212.

ALBERT F. W., MUZZEY D., WEISSMAN J. S., KRUGLYAK L., 2014a Genetic influences on translation in yeast. PLoS Genet. **10**: e1004692.

ALBERT F. W., TREUSCH S., SHOCKLEY A. H., BLOOM J. S., KRUGLYAK L., 2014b Genetics of single-cell protein abundance variation in large yeast populations. Nature **506**: 494–497.

ALON U., 2007 Network motifs: theory and experimental approaches. Nat. Rev. Genet. **8**: 450–61.

ALVAREZ M., SCHREY A. W., RICHARDS C. L., 2014 Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? Mol. Ecol. **24**: 710–725.

ANDERSON D. W., MCKEOWN A. N., THORNTON J. W., 2015 Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. Elife **4**: e07864.

ANDOLFATTO P., 2005 Adaptive evolution of non-coding DNA in Drosophila. Nature **437**: 1149–52.

ARTIERI C. G., FRASER H. B., 2013 Evolution at two levels of gene expression in yeast. Genome Res.: 411–421.

AYROLES J. F., CARBONE M. A., STONE E. A., JORDAN K. W., LYMAN R. F., MAGWIRE M. M., ROLLMANN S. M., DUNCAN L. H., LAWRENCE F., ANHOLT R. R. H., MACKAY T. F. C., 2009 Systems genetics of complex traits in Drosophila melanogaster. Nat. Genet. **41**: 299–307.

BABAK T., GARRETT-ENGELE P., ARMOUR C. D., RAYMOND C. K., KELLER M. P., CHEN R., ROHL C. A., JOHNSON J. M., ATTIE A. D., FRASER H. B., SCHADT E. E.,

2010 Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation. BMC Genomics **11**.

BABU M. M., LUSCOMBE N. M., ARAVIND L., GERSTEIN M., TEICHMANN S. A., 2004 Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. **14**: 283–91.

BADIS G., BERGER M. F., PHILIPPAKIS A. A., TALUKDER S., GEHRKE A. R., JAEGER S. A., CHAN E. T., METZLER G., VEDENKO A., CHEN X., KUZNETSOV H., WANG C.-F., COBURN D., NEWBURGER D. E., MORRIS Q., HUGHES T. R., BULYK M. L., 2009 Diversity and complexity in DNA recognition by transcription factors. Science **324**: 1720–1723.

BAKER C. R., HANSON-SMITH V., JOHNSON A. D., 2013 Following gene duplication, paralog interference constrains transcriptional circuit evolution. Science **342**: 104–108.

BAR-EVEN A., PAULSSON J., MAHESHRI N., CARMI M., O'SHEA E., PILPEL Y., BARKAI N., 2006 Noise in protein expression scales with natural protein abundance. Nat. Genet. **38**: 636–43.

BASEHOAR A. D., ZANTON S. J., PUGH B. F., 2004 Identification and distinct regulation of yeast TATA box-containing genes. Cell **116**: 699–709.

BATESON W., 1894 *Materials for the study of variation*. Macmillan, London.

BLAKE W. J., KAERN M., CANTOR C. R., COLLINS J. J., 2003 Noise in eukaryotic gene expression. Nature **422**: 633–7.

BOER C. G. DE, BAKEL H. VAN, TSUI K., LI J., MORRIS Q. D., NISLOW C., GREENBLATT J. F., HUGHES T. R., 2014 A unified model for yeast transcript definition. Genome Res. **24**: 154–66.

BOOTH L. N., TUCH B. B., JOHNSON A. D., 2010 Intercalation of a new tier of transcription regulation into an ancient circuit. Nature **468**: 959–963.

BORENSTEIN E., KRAKAUER D. C., 2008 An end to endless forms: epistasis, phenotype distribution bias, and nonuniform evolution. PLoS Comput. Biol. **4**: e1000202.

BOWLER P. J., 1992 *The Eclipse of Darwinism: Anti-Darwinian Evolution Theories in the Decades Around 1900*. Johns Hopkins University Press, Baltimore.

BOYLE A. P., ARAYA C. L., BRDLIK C., CAYTING P., CHENG C., CHENG Y., GARDNER K., HILLIER L. W., JANETTE J., JIANG L., KASPER D., KAWLI T., KHERADPOUR P., KUNDAJE A., LI J. J., MA L., NIU W., REHM E. J., ROZOWSKY J., SLATTERY M., SPOKONY R., TERRELL R., VAFEADOS D., WANG D., WEISDEPP P., WU Y.-C., XIE

D., YAN K.-K., FEINGOLD E. a., GOOD P. J., PAZIN M. J., HUANG H., BICKEL P. J., BRENNER S. E., REINKE V., WATERSTON R. H., GERSTEIN M., WHITE K. P., KELLIS M., SNYDER M., 2014 Comparative analysis of regulatory information and circuits across distant species. Nature **512**: 453–456.

BREM R. B., KRUGLYAK L., 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. PNAS **102**: 1572–7.

BREM R. B., YVERT G., CLINTON R., KRUGLYAK L., 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296**: 752–5.

BRITTEN R. J. J. R., DAVIDSON E. H. H. E., 1969 Gene regulation for higher cells: a theory. Science **165**: 349–357.

BROCKINGTON S. F., MOYROUD E., SAYOU C., MONNIAUX M., NANAO M. H., THÉVENON E., CHAHTANE H., 2015 Response to Comment on "A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity." Science **347**: 621–b.

BROGAARD K., XI L., WANG J.-P., WIDOM J., 2012 A map of nucleosome positions in yeast at base-pair resolution. Nature **486**: 496–501.

BROWN C. R., MAO C., FALKOVSKAIA E., JURICA M. S., BOEGER H., 2013 Linking Stochastic Fluctuations in Chromatin Structure and Gene Expression. PLoS Biol. **11**: e1001621.

BRUNKARD J. O., RUNKEL A. M., 2015 Comment on "A promicuous intermediate underlies the evolution of LEAFY DNA binding specificity." Science **347**: 621–a.

BULLARD J. H., MOSTOVOY Y., DUDOIT S., BREM R. B., 2010 Polygenic and directional regulatory evolution across pathways in Saccharomyces. PNAS **107**: 5058–63.

BULLAUGHEY K., 2011 Changes in selective effects over time facilitate turnover of enhancer sequences. Genetics **187**: 567–82.

BUSBY M. a, GRAY J. M., COSTA A. M., STEWART C., STROMBERG M. P., BARNETT D., CHUANG J. H., SPRINGER M., MARTH G. T., 2011 Expression divergence measured by transcriptome sequencing of four yeast species. BMC Genomics **12**: 635.

CAREY L. B., DIJK D. VAN, SLOOT P. M. a., KAANDORP J. a., SEGAL E., 2013 Promoter Sequence Determines the Relationship between Expression Level and Noise. PLoS Biol. **11**: e1001528.

CARROLL S. B. S., 1995 Homeotic genes and the evolution of arthropods and chordates. Nature **376**: 479–485.

CARROLL S. B., 2005 Evolution at two levels: on genes and form. PLoS Biol. **3**: e245.

CARROLL S. B., 2008 Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell **134**: 25–36.

CHALANCON G., RAVARANI C. N. J., BALAJI S., MARTINEZ-ARIAS A., ARAVIND L., JOTHI R., BABU M. M., 2012 Interplay between gene expression noise and regulatory network architecture. Trends Genet. **28**: 221–232.

CHANG J., ZHOU Y., HU X., LAM L., HENRY C., GREEN E. M., KITA R., KOBOR M. S., FRASER H. B., 2013 The molecular mechanism of a cis-regulatory adaptation in yeast. PLoS Genet. **9**: e1003813.

CHARLESWORTH B., 2013 Stabilizing Selection, Purifying Selection and Mutational Bias in Finite Populations. Genetics **194**: 955–971.

CHAROENSAWAN V., JANGA S. C. C., BULYK M. L. L., BABU M. M. M., TEICHMANN S. A. a, 2012 DNA Sequence Preferences of Transcriptional Activators Correlate More Strongly than Repressors with Nucleosomes. Mol. Cell **47**: 183–192.

CHEN C., RAPPAILLES A., DUQUENNE L., HUVET M., GUILBAUD G., FARINELLI L., AUDIT B., AUBENTON-CARAFA Y., ARNEODO A., HYRIEN O., THERMES C., 2010 Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res. **20**: 447–457.

CHENG C., ALEXANDER R. P., MIN R., LENG J., YIP K. Y., ROZOWSKY J., YAN K.-K., DONG X., DJEBALI S., RUAN Y., DAVIS C. a., CARNINCI P., LASSMAN T., GINGERAS T. R., GUIGO R., BIRNEY E., WENG Z., SNYDER M., GERSTEIN M., 2012 Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. **22**: 1658–1667.

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

CONTE G. L., ARNEGARD M. E., PEICHEL C. L., SCHLUTER D., 2012 The probability of genetic parallelism and convergence in natural populations. Proc. Biol. Sci. **279**: 5039–5047.

COOLON J. D., MCMANUS C. J., STEVENSON K. R., GRAVELEY B. R., WITTKOPP P. J., 2014 Tempo and mode of regulatory evolution in Drosophila. Genome Res.: gr.163014.113–.

COOLON J. D., WEBB W., WITTKOPP P. J., 2013 Sex-specific effects of cis-regulatory variants in Drosophila melanogaster. Genetics **195**: 1419–22.

COWLES C. R., HIRSCHHORN J. N., ALTSHULER D., LANDER E. S., 2002 Detection of regulatory variation in mouse genes. Nat. Genet. **32**: 432–437.

COYNE J. a, 2006 Comment on “Gene regulatory networks and the evolution of animal body plans”. Science **313**: 761b.

DADIANI M., DIJK D. VAN, SEGAL B., FIELD Y., BEN-ARTZI G., RAVEH-SADKA T., LEVO M., KAPLOW I., WEINBERGER A., SEGAL E., 2013 Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise. Genome Res. **23**: 966–976.

DARWIN C., 1868 *The Variation of Animals and Plants Under Domestication Vol. 1*.

DARWIN C., 1872 *The Origin of Species*.

DAVIDSON E. H., ERWIN D. H., 2006 Gene regulatory networks and the evolution of animal body plans. Science **311**: 796–800.

DENBY C. M., IM J. H., YU R. C., PESCE C. G., BREM R. B., 2012 Negative feedback confers mutational robustness in yeast transcription factor regulation. PNAS **109**: 3874–8.

DENVER D. R., DOLAN P. C., WILHELM L. J., SUNG W., LUCAS-LLEDÓ J. I., HOWE D. K., LEWIS S. C., OKAMOTO K., THOMAS W. K., LYNCH M., BAER C. F., 2009 A genome-wide view of Caenorhabditis elegans base-substitution mutation processes. PNAS **106**: 16310–16314.

DENVER D. R., MORRIS K., STREELMAN J. T., KIM S. K., LYNCH M., THOMAS W. K., 2005 The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nat. Genet. **37**: 544–548.

DRAKE J. W., CHARLESWORTH B., CHARLESWORTH D., CROW J. F., 1998 Rates of Spontaneous Mutation. Genetics **148**: 1667–1686.

ELOWITZ M. B., LEVINE A. J., SIGGIA E. D., SWAIN P. S., 2002 Stochastic gene expression in a single cell. Science **297**: 1183–1186.

EMERSON J. J., HSIEH L.-C., SUNG H.-M., WANG T.-Y., HUANG C.-J., LU H. H.-S., LU M.-Y. J., WU S.-H., LI W.-H., 2010 Natural selection on cis and trans regulation in yeasts. Genome Res.: 826–836.

ERWIN D. H., DAVIDSON E. H., 2009 The evolution of hierarchical gene regulatory networks. Nat. Rev. Genet. **10**: 141–148.

EYRE-WALKER A., KEIGHTLEY P. D., 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. **8**: 610–618.

EYRE-WALKER A., WOOLFIT M., PHELPS T., 2006 The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics **173**: 891–900.

FAY J. C., WITTKOPP P. J., 2008 Evaluating the role of natural selection in the evolution of gene regulation. Heredity (Edinb). **100**: 191–9.

FAZLOLLAHI M., LEE E., MUROFF I., LU X.-J., GOMEZ-ALCALA P., CAUSTON H. C., BUSSEMAKER H. J., 2014 Harnessing Natural Sequence Variation to Dissect Post-Transcriptional Regulatory Networks in Yeast. G3: 1539–1553.

FISHER R. a, 1919 The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans. R. Soc. Edinburgh **52**: 399–433.

FISHER R. A., 1930 *The genetical theory of natural selection*. Oxford University Press.

FLUTRE T., WEN X., PRITCHARD J., STEPHENS M., 2013 A Statistical Framework for Joint eQTL Analysis in Multiple Tissues (G Gibson, Ed.). PLoS Genet. **9**: e1003486.

FOSS E. J., RADULOVIC D., SHAFFER S. a., GOODLETT D. R., KRUGLYAK L., BEDALOV A., 2011 Genetic Variation Shapes Protein Networks Mainly through Non-transcriptional Mechanisms (MB Eisen, Ed.). PLoS Biol. **9**: e1001144.

FOSS E. J., RADULOVIC D., SHAFFER S. a, RUDERFER D. M., BEDALOV A., GOODLETT D. R., KRUGLYAK L., 2007 Genetic basis of proteome variation in yeast. Nat. Genet. **39**: 1369–75.

FRANK S. a, 2013 Evolution of robustness and cellular stochasticity of gene expression. PLoS Biol. **11**: e1001578.

FRANKEL N., DAVIS G. K., VARGAS D., WANG S., PAYRE F., STERN D. L., 2010 Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature **466**: 1–5.

FRASER H. B., HIRSH A. E., GIAEVER G., KUMM J., EISEN M. B., 2004 Noise Minimization in Eukaryotic Gene Expression. PLos Biol. **2**: e137.

FRASER H. B., LEVY S., CHAVAN A., SHAH H. B., PEREZ J. C., ZHOU Y., SIEGAL M. L., SINHA H., 2012 Polygenic cis-regulatory adaptation in the evolution of yeast pathogenicity. Genome Res. **22**: 1930–9.

FRASER H. B. H., MOSES A. M., SCHADT E. E., 2010 Evidence for widespread adaptive evolution of gene expression in budding yeast. PNAS **107**: 2977–82.

GE B., POKHOLOK D. K., KWAN T., GRUNDBERG E., MORCOS L., VERLAAN D. J., LE J., KOKA V., LAM K. C. L., GAGNÉ V., DIAS J., HOBERMAN R., MONTPETIT A., JOLY M.-M., HARVEY E. J., SINNETT D., BEAULIEU P., HAMON R., GRAZIANI A., DEWAR K., HARMSEN E., MAJEWSKI J., GÖRING H. H. H., NAUMOVA A. K., BLANCHETTE M., GUNDERSON K. L., PASTINEN T., 2009 Global patterns of cis variation in

human cells revealed by high-density allelic expression analysis. Nat. Genet. **41**: 1216–22.

GEILER-SAMEROTTE K. A., BAUER C. R., LI S., ZIV N., GRESHAM D., SIEGAL M. L., 2013 The details in the distributions: why and how to study phenotypic variability. Curr. Opin. Biotechnol. **24**: 752–759.

GENISSEL A., MCINTYRE L. M., WAYNE M. L., NUZHDIN S. V, 2008 Cis and trans regulatory effects contribute to natural variation in transcriptome of Drosophila melanogaster. Mol. Biol. Evol. **25**: 101–10.

GERHART J., KIRSCHNER M., 1997 *Cells, Embryos and Evolution*. John Wiley & Sons, Incorporated.

GHAEMMAGHAMI S., HUH W.-K., BOWER K., HOWSON R. W., BELLE A., DEPHOURE N., O'SHEA E. K., WEISSMAN J. S., 2003 Global analysis of protein expression in yeast. Nature **425**: 737–41.

GIBSON G., WEIR B., 2005 The quantitative genetics of transcription. Trends Genet. **21**: 616–623.

GILAD Y., OSHLACK A., RIFKIN S. A., 2006 Natural selection on gene expression. Trends Genet. **22**: 456–461.

GILAD Y., RIFKIN S. a., PRITCHARD J. K., 2008 Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. **24**: 408–415.

GOMPEL N., PRUD'HOMME B., 2009 The causes of repeated genetic evolution. Dev. Biol. **332**: 36–47.

GONCALVES A., LEIGH-BROWN S., THYBERT D., STEFFLOVA K., TURRO E., FLICEK P., BRAZMA A., ODOM D. T., MARIONI J. C., 2012 Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res. **22**: 2376–2384.

GORDÂN R., MURPHY K. F., MCCORD R. P., ZHU C., VEDENKO A., BULYK M. L., GORDAN R., 2011 Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biol. **12**: R125.

GORDON K. L., RUVINSKY I., 2012 Tempo and Mode in Evolution of Transcriptional Regulation (HS Malik, Ed.). PLoS Genet. **8**: e1002432.

GRAZE R. M., NOVELO L. L., AMIN V., FEAR J. M., CASELLA G., NUZHDIN S. V, MCINTYRE L. M., 2012 Allelic imbalance in Drosophila hybrid heads: exons, isoforms, and evolution. Mol. Biol. Evol. **29**: 1521–32.

GREENBAUM D., COLANGELO C., WILLIAMS K., GERSTEIN M., 2003 Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. **4**: 117.

GRUBER J. D., VOGEL K., KALAY G., WITTKOPP P. J., 2012 Contrasting Properties of Gene-specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects and Dominance. PLoS Genet. **8**: e1002497.

GYGI S. P., ROCHON Y., FRANZA B. R., AEBERSOLD R., 1999 Correlation between Protein and mRNA Abundance in Yeast. Mol. Cell. Biol. **19**: 1720–1730.

HABIB N., WAPINSKI I., MARGALIT H., REGEV A., FRIEDMAN N., 2012 A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. Mol. Syst. Biol. **8**: 619.

HALL B. K., 2003 Evo-Devo: evolutionary developmental mechanisms. Int. J. Dev. Biol. **47**: 491–495.

HALLIGAN D. L., KEIGHTLEY P. D., 2009 Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. Annu. Rev. Ecol. Evol. Syst. **40**: 151–172.

HANSEN B. G., HALKIER B. a., KLIEBENSTEIN D. J., 2008 Identifying the molecular basis of QTLs: eQTLs add a new dimension. Trends Plant Sci. **13**: 72–77.

HARBISON C. T., GORDON D. B., LEE T. I., RINALDI N. J., MACISAAC K. D., DANFORD T. W., HANNETT N. M., TAGNE J.-B., REYNOLDS D. B., YOO J., JENNINGS E. G., ZEITLINGER J., POKHOLOK D. K., KELLIS M., ROLFE P. A., TAKUSAGAWA K. T., LANDER E. S., GIFFORD D. K., FRAENKEL E., YOUNG R. A., 2004 Transcriptional regulatory code of a eukaryotic genome. Nature **431**: 99–104.

HILL W. G., ZHANG X.-S., 2012a On the Pleiotropic Structure of the Genotype-phenotype Map and the Evolvability of Complex Organisms. Genetics **190**: 1131–1137.

HILL W. G., ZHANG X.-S., 2012b Assessing pleiotropy and its evolutionary consequences: pleiotropy is not necessarily limited, nor need it hinder the evolution of complexity. Nat. Rev. Genet. **13**: 296.

HO W.-C., ZHANG J., 2014 The genotype-phenotype map of yeast complex traits: basic parameters and the role of natural selection. Mol. Biol. Evol. **31**: 1568–1580.

HODGINS-DAVIS A., RICE D. P., TOWNSEND J. P., 2015 Gene expression evolves under a House-of-Cards model of stabilizing selection. Mol. Biol. Evol.

HODGKINSON A., EYRE-WALKER A., 2011 Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. **12**: 756–766.

HOEKSTRA H., COYNE J., 2007 The locus of evolution: evo devo and the genetics of adaptation. Evolution (N. Y.). **61**: 995–1016.

HORNUNG G., BAR-ZIV R., ROSIN D., TOKURIKI N., TAWFIK D. S., OREN M., BARKAI N., 2012 Noise-mean relationship in mutated promoters. Genome Res. **22**: 2409–2417.

HU Z., KILLION P. J., IYER V. R., 2007 Genetic reconstruction of a functional transcriptional regulatory network. Nat. Genet. **39**: 683–687.

HUANG S.-S. C., FRAENKEL E., 2009 Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci. Signal. **2**: 1–10.

HUBBY J. L., LEWONTIN R. C., 1966 A Molecular Approach To the Study of Genic Heterozygosity in Natural Populations I. The number of alleles at different Loci in Drosophila Pseudoobscura. Genetics **54**: 577–594.

HUGHES A. L. A. A. L., JIN Y., RANDO O. J. O. O. J., STRUHL K., 2012 A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. Mol. Cell **48**: 1–11.

HUISINGA K. L., PUGH B. F., 2004 A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. Mol. Cell **13**: 573–585.

HULSE A. M., CAI J. J., 2013 Genetic variants contribute to gene expression variability in humans. Genetics **193**: 95–108.

HUXLEY J., PIGLIUCCI M., MÜLLER G. B., 2010 *Evolution: The Modern Synthesis: The Definitive Edition*. Mit Press.

JACKSON A. P., TAYLOR M. S., REIJNS M. a M., KEMP H., DING J., PROCE S. M. De, 2015 Lagging-strand replication shapes the mutational landscape of the genome. Nature **518**: 502–506.

JACOB F., MONOD J., 1961 Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. **3**: 318–56.

JENKIN F., 1867 (Review of) "The origin of species." North Br. Rev. **46**: 277–318.

JONES D., BREWSTER R., PHILLIPS R., 2014 Promoter architecture dictates cell-to-cell variability in gene expression. Science: 1533–1537.

JONES F. C., GRABHERR M. G., CHAN Y. F., RUSSELL P., MAUCELI E., JOHNSON J., SWOFFORD R., PIRUN M., ZODY M. C., WHITE S., BIRNEY E., SEARLE S., SCHMUTZ J., GRIMWOOD J., DICKSON M. C., MYERS R. M., MILLER C. T., SUMMERS B. R.,

KNECHT A. K., BRADY S. D., ZHANG H., POLLEN A. a., HOWES T., AMEMIYA C., BALDWIN J., BLOOM T., JAFFE D. B., NICOL R., WILKINSON J., LANDER E. S., PALMA F. DI, LINDBLAD-TOH K., KINGSLEY D. M., 2012 The genomic basis of adaptive evolution in threespine sticklebacks. Nature **484**: 55–61.

KALAY G., WITTKOPP P. J., 2010 Nomadic enhancers: Tissue-specific cis-regulatory elements of yellow have divergent genomic positions among Drosophila species. PLoS Genet. **6**.

KEIGHTLEY P. D., TRIVEDI U., THOMSON M., OLIVER F., KUMAR S., BLAXTER M. L., 2009 Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. Genome R **19**: 1195–1201.

KEPLER T. B., ELSTON T. C., 2001 Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophys. J. **81**: 3116–3136.

KEUNG A. J., BASHOR C. J., KIRIAKOV S., COLLINS J. J., KHALIL A. S., 2014 Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. Cell **158**: 110–20.

KHAN Z., BLOOM J. S., AMINI S., SINGH M., PERLMAN D. H., CAUDY A. a, KRUGLYAK L., 2012 Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. Mol. Syst. Biol. **8**: 1–12.

KIM J. K., MARIONI J. C., 2013 Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. **14**: R7.

KIMURA M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press.

KING M. C. C., WILSON A. C. C., 1975 Evolution at two levels in humans and chimpanzees. Science **188**: 107–116.

KO M. S., 1991 A stochastic model for gene induction. J. Theor. Biol. **153**: 181–194.

KULP D. C., JAGALUR M., 2006 Causal inference of regulator-target pairs by gene mapping of expression phenotypes. BMC Genomics **7**: 125.

KUMAR S., SUBRAMANIAN S., 2002 Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. U.S.A. **99**: 803–808.

KUO D., LICON K., BANDYOPADHYAY S., CHUANG R., LUO C., CATALANA J., RAVASI T., TAN K., IDEKER T., 2010 Coevolution within a transcriptional network by compensatory trans and cis mutations. Genome Res. **20**: 1672–8.

KWASNIESKI J., MOGNO I., 2012 Complex effects of nucleotide variants in a mammalian cis-regulatory element. PNAS **109**: 19498–19503.

LAGARRIGUE S., MARTIN L., HORMOZDIARI F., ROUX P.-F., PAN C., NAS A. VAN, DEMEURE O., CANTOR R., GHAZALPOUR A., ESKIN E., LUSIS A. J., 2013 Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. Genetics **195**: 1157–66.

LANDRY C. R., HARTL D. L., RANZ J. M., 2007a Genome clashes in hybrids: insights from gene expression. Heredity (Edinb). **99**: 483–93.

LANDRY C. R., LEMOS B., RIFKIN S. A., DICKINSON W. J., HARTL D. L., 2007b Genetic properties influencing the evolvability of gene expression. Science **317**: 118–121.

LANDRY C. R., WITTKOPP P. J., TAUBES C. H., RANZ J. M., CLARK A. G., HARTL D. L., 2005 Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. Genetics **171**: 1813–22.

LANG G. I., MURRAY A. W., 2011 Mutation rates across budding yeast chromosome VI Are correlated with replication timing. Genome Biol. Evol. **3**: 799–811.

LANG G. I., PARSONS L., GAMMIE A. E., 2013 Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast. G3 **3**: 1453–65.

LAVOIE H., HOGUES H., MALLICK J., SELLAM A., NANTEL A., WHITEWAY M., 2010 Evolutionary Tinkering with Conserved Components of a Transcriptional Regulatory Network. PLoS Biol. **8**: e1000329.

LEE H., POPODI E., 2012 Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. PNAS **109**: E2774–E2783.

LEE T. I. T. I., RINALDI N. J. N. J., ROBERT F., ODOM D. T. D. T., BAR-JOSEPH Z., GERBER G. K. G. K., HANNETT N. M. N. M., HARBISON C. T. C. T., THOMPSON C. M. C. M., SIMON I., OTHERS, ZEITLINGER J., JENNINGS E. G., MURRAY H. L., GORDON D. B., REN B., WYRICK J. J., TAGNE J., VOLKERT T. L., FRAENKEL E., GIFFORD D. K., YOUNG R. a, 2002 Transcriptional regulatory networks in Saccharomyces cerevisiae. Science **298**: 799–804.

LEHNER B., 2008 Selection to minimise noise in living systems and its implications for the evolution of gene expression. Mol. Syst. Biol. **4**: 1–6.

LEHNER B., 2010 Conflict between Noise and Plasticity in Yeast (JM Akey, Ed.). PLoS Genet. **6**: e1001185.

LEMMON Z. H., BUKOWSKI R., SUN Q., DOEBLEY J. F., 2014 The Role of cis Regulatory Evolution in Maize Domestication (H Fraser, Ed.). PLoS Genet. **10**: e1004745.

LEMOS B., ARARIPE L. O., FONTANILLAS P., HARTL D. L., 2008 Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. PNAS **105**: 14471–6.

LENZ D., RILES L., FAY J., 2014 Heterochronic meiotic misexpression in an interspecific yeast hybrid. Mol. Biol. Evol.

LEVO M., SEGAL E., 2014 In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. **15**: 453–468.

LEVO M., ZALCKVAR E., SHARON E., CAROLINA A., MACHADO D., KALMA Y., LOTAM-POMPAN M., WEINBERGER A., YAKHINI Z., ROHS R., SEGAL E., 2015 Unraveling determinants of transcription factor binding outside the core binding site. Genome Res. **25**: 1018–1029.

LEVY S. F., BLUNDELL J. R., VENKATARAM S., PETROV D. a., FISHER D. S., SHERLOCK G., 2015 Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature **519**: 181–186.

LEWONTIN R. C., HUBBY J. L., 1966 A Molecular Approach To the Study of Genic Heterozygosity in Natural Populations. II. Amount of Variation and degree of heterozygosity in Natural populations of Drosophila Pseudoobscura. Genetics **54**: 595–609.

LI M., HADA A., SEN P., OLUFEMI L., HALL M. a, SMITH B. Y., FORTH S., MCKNIGHT J. N., PATEL A., BOWMAN G. D., BARTHOLOMEW B., WANG M. D., 2015 Dynamic regulation of transcription factors by nucleosome remodeling. Elife **4**: 1–16.

LI H., JOHNSON A. D., 2010 Evolution of transcription networks - lessons from yeasts. Curr. Biol. **20**: R746–R753.

LI J., LIU Y., LIU M., HAN J.-D. J., 2013 Functional Dissection of Regulatory Models Using Gene Expression Data of Deletion Mutants (R Dowell, Ed.). PLoS Genet. **9**: e1003757.

LITTLE S. C., TIKHONOV M., GREGOR T., 2013 Precise developmental gene expression arises from globally stochastic transcriptional activity. Cell **154**: 789–800.

LONFAT N., MONTAVON T., DARBELLAY F., GITTO S., DUBOULE D., 2014 Convergent evolution of complex regulatory landscapes and pleiotropy at Hox loci. Science **346**: 1004–1006.

LU P., VOGEL C., WANG R., YAO X., MARCOTTE E. M., 2007 Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. **25**: 117–24.

LUBLINER S., REGEV I., LOTAN-POMPAN M., EDELHEIT S., WEINBERGER A., SEGAL E., 2015 Core promoter sequence in yeast is a major determinant of expression level. Genome Res. **25**: 1008–1017.

LUSCOMBE N. M., BABU M. M., YU H., SNYDER M., TEICHMANN S. a, GERSTEIN M., 2004 Genomic analysis of regulatory network dynamics reveals large topological changes. Nature **431**: 308–312.

LYNCH M., 2010 Evolution of the mutation rate. Trends Genet. **26**: 345–352.

LYNCH M., HAGNER K., 2014 Evolutionary meandering of intermolecular interactions along the drift barrier. PNAS **2014**: E30–E38.

LYNCH V. J., MAY G., WAGNER G. P., 2011 Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB. Nature: 1–5.

LYNCH M., SUNG W., MORRIS K., COFFEY N., LANDRY C. R., DOPMAN E. B., DICKINSON W. J., OKAMOTO K., KULKARNI S., HARTL D. L., THOMAS W. K., 2008 A genome-wide view of the spectrum of spontaneous mutations in yeast. PNAS **105**: 9272–9277.

LYNCH V. J., WAGNER G. P., 2008 Resurrecting the role of transcription factor change in developmental evolution. Evolution (N. Y). **62**: 2131–2154.

MA S., KEMMEREN P., GRESHAM D., STATNIKOV A., 2014 De-Novo Learning of Genome-Scale Regulatory Networks in S. cerevisiae. PLoS One **9**: e106479.

MACISAAC K. D., WANG T., GORDON D. B., GIFFORD D. K., STORMO G. D., FRAENKEL E., 2006 An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics **7**: 113.

MAERKL S. J., QUAKE S. R., 2007 A systems approach to measuring the binding energy landscapes of transcription factors. Science **315**: 233–237.

MAERKL S. J., QUAKE S. R., 2009 Experimental determination of the evolvability of a transcription factor. PNAS **106**: 18650–5.

MARTIN G., LENORMAND T., 2008 The distribution of beneficial and fixed mutation fitness effects close to an optimum. Genetics **179**: 907–16.

MARTIN A., ORGOGOZO V., 2013 The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution (N. Y). **67**: 1235–1250.

MARTIN H. C., ROOP J. I., SCHRAIBER J. G., HSU T. Y., BREM R. B., 2012 Evolution of a Membrane Protein Regulon in Saccharomyces. Mol. Biol. Evol. **29**: 1747–1756.

MAYO A. E., SETTY Y., SHAVIT S., ZASLAVER A., ALON U., 2006 Plasticity of the cis-regulatory input function of a gene. PLoS Biol. **4**: e45.

MCALISTER L., HOLLAND M. J., 1985 Differential expression of the three yeast glyceraldehyde-3-phosphate dehydrogenase genes. J. Biol. Chem. **260**: 15019–15027.

MCKEOWN A. N., BRIDGHAM J. T., ANDERSON D. W., MURPHY M. N., ORTLUND E. A., THORNTON J. W., 2014 Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. Cell **159**: 58–68.

MCMANUS C. J., COOLON J. D., O'DUFF M., EIPPER-MAINS J., GRAVELEY B. R., WITTKOPP P. J., DUFF M. O., 2010 Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. **20**: 816–25.

MCMANUS J., MAY G. E., SPEALMAN P., SHTEYMAN A., 2014 Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. Genome Res.: 422–430.

MELNIKOV A., MURUGAN A., ZHANG X., TESILEANU T., WANG L., ROGOV P., FEIZI S., GNIRKE A., CALLAN C. G., KINNEY J. B., KELLIS M., LANDER E. S., MIKKELSEN T. S., 2012 Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. **30**: 271–279.

MENDEL G., 1866 Versuche Über Pflanzenhybriden. Verh Naturforsch Ver Brunn **4**: 3–47.

MILO R., SHEN-ORR S., ITZKOVITZ S., KASHTAN N., CHKLOVSKII D., ALON U., 2002 Network motifs: simple building blocks of complex networks. Science **298**: 824–7.

MOQTADERI Z., GEISBERG J. V, JIN Y., FAN X., STRUHL K., 2013 Species-specific factors mediate extensive heterogeneity of mRNA 3' ends in yeasts. PNAS **110**: 11073–11078.

MORGAN T. H., 1916 *A critique of the theory of evolution*. Princeton University Press, Princeton.

MORLEY M., MOLONY C. M., WEBER T. M., DEVLIN J. L., EWENS K. G., SPIELMAN R. S., CHEUNG V. G., 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430**: 743–7.

MOSES A. M., 2009 Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. BMC Evol. Biol. **9**: 286.

MOSES A. M., CHIANG D. Y., KELLIS M., LANDER E. S., EISEN M. B., 2003 Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol. Biol. **3**: 19.

MÜLLER G., 2007 Evo–devo: extending the evolutionary synthesis. Nat. Rev. Genet. **8**: 943–949.

MUNSKY B., NEUERT G., OUDENAARDEN A. VAN, 2012 Using gene expression noise to understand gene regulation. Science **336**: 183–187.

MURPHY K. F., ADAMS R. M., WANG X., BALÁZSI G., COLLINS J. J., 2010 Tuning and controlling gene expression noise in synthetic gene networks. Nucleic Acids Res. **38**: 2712–2726.

NAGALAKSHMI U., WANG Z., WAERN K., SHOU C., RAHA D., GERSTEIN M., SNYDER M., 2008 The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320**: 1344–1349.

NEI M., 2007 The new mutation theory of phenotypic evolution. PNAS **104**: 12235–12242.

NEI M., 2013 *Mutation-Driven Evolution*. Oxford University Press.

NEIL H., MALABAT C., D'AUBENTON-CARAFA Y., XU Z., STEINMETZ L. M., JACQUIER A., 2009 Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature **457**: 1038–1042.

NEWMAN J. R. S., GHAEMMAGHAMI S., IHMELS J., BRESLOW D. K., NOBLE M., DERISI J. L., WEISSMAN J. S., 2006 Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature **441**: 840–846.

NUZHDIN S. V, WAYNE M. L., HARMON K. L., MCINTYRE L. M., 2004 Common pattern of evolution of gene expression level and protein sequence in Drosophila. Mol. Biol. Evol. **21**: 1308–17.

OSSOWSKI S., SCHNEEBERGER K., LUCAS-LLEDÓ J. I., WARTHMANN N., CLARK R. M., SHAW R. G., WEIGEL D., LYNCH M., 2010 The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science **327**: 92–94.

PAABY A. B., ROCKMAN M. V., 2013 Pleiotropy: what do you mean? Reply to Zhang and Wagner. Trends Genet.: 9525.

PAI A. a., PRITCHARD J. K., GILAD Y., 2015 The Genetic and Mechanistic Basis for Variation in Gene Regulation (T Lappalainen, Ed.). PLoS Genet. **11**: e1004857.

PARTS L., LIU Y.-C., TEKKEDIL M. M., STEINMETZ L. M., CAUDY A. a, FRASER A. G., BOONE C., ANDREWS B. J., ROSEBROCK A. P., 2014 Heritability and genetic basis of protein level variation in an outbred population. Genome Res. **24**: 1363–70.

PASTINEN T., 2010 Genome-wide allele-specific analysis: insights into regulatory variation. Nat. Rev. Genet. **11**: 533–538.

PATWARDHAN R. P., HIATT J. B., WITTEN D. M., KIM M. J., SMITH R. P., MAY D., LEE C., ANDRIE J. M., LEE S.-I., COOPER G. M., AHITUV N., PENNACCHIO L. a, SHENDURE J., 2012 Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. **30**: 265–270.

PATWARDHAN R. P., LEE C., LITVIN O., YOUNG D. L., PE'ER D., SHENDURE J., 2009 High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. **27**: 1173–1175.

PAYNE J. L., WAGNER A., 2014a Latent phenotypes pervade gene regulatory circuits. BMC Syst. Biol. **8**: 64.

PAYNE J. L., WAGNER a., 2014b The Robustness and Evolvability of Transcription Factor Binding Sites. Science **343**: 875–877.

PEDRAZA J. M., PAULSSON J., 2008 Effects of molecular memory and bursting on fluctuations in gene expression. Science **319**: 339–343.

PELECHANO V., WEI W., STEINMETZ L. M., 2013 Extensive transcriptional heterogeneity revealed by isoform profiling. Nature **497**: 1–46.

PIGLIUCCI M., 2010 Genotype-phenotype mapping and the end of the "genes as blueprint" metaphor. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **365**: 557–566.

POUGACH K., VOET A., KONDRASHOV F. a, VOORDECKERS K., CHRISTIAENS J. F., BAYING B., BENES V., SAKAI R., AERTS J., ZHU B., DIJCK P. VAN, VERSTREPEN K. J., 2014 Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. Nat. Commun. **5**: 4868.

PRITCHARD J. K., GILAD Y., 2013 Impact of regulatory variation from RNA to protein. Science **347**: 664–667.

PRUD'HOMME B., GOMPEL N., CARROLL S. B., 2007 Emerging principles of regulatory evolution. PNAS **104**: 8605–8612.

RASER J., O'SHEA E., 2005 Noise in gene expression: origins, consequences, and control. Science **309**: 2010–2013.

RAVEH-SADKA T., LEVO M., SHABI U., SHANY B., KEREN L., LOTAN-POMPAN M., ZEEVI D., SHARON E., WEINBERGER A., SEGAL E., 2012 Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat. Genet. **44**: 743–750.

RENAUT S., NOLTE A. W., BERNATCHEZ L., 2009 Gene expression divergence and hybrid misexpression between lake whitefish species pairs (Coregonus spp. Salmonidae). Mol. Biol. Evol. **26**: 925–36.

RHEE H. S., PUGH B. F., 2012 Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature **483**: 295–301.

RICE D. P., GOOD B. H., DESAI M. M., 2015 The Evolutionarily Stable Distribution of Fitness Effects. Genetics **200**: 321–329.

RICE D. P. D., TOWNSEND J. P. J., 2012 A test for selection employing quantitative trait locus and mutation accumulation data. Genetics **190**: 1533–1545.

RICHARD M., YVERT G., 2014 How does evolution tune biological noise? Front. Genet. **5**: 1–8.

RIFKIN S. A., KIM J., WHITE K. P., 2003 Evolution of gene expression in the Drosophila melanogaster subgroup. Nat. Genet. **33**: 138–144.

ROCKMAN M. V, 2008 Reverse engineering the genotype-phenotype map with natural genetic variation. Nature **456**: 738–744.

ROKYTA D. R., BEISEL C. J., JOYCE P., FERRIS M. T., BURCH C. L., WICHMAN H. a, 2008 Beneficial fitness effects are not exponential for two viruses. J. Mol. Evol. **67**: 368–376.

ROMERO I. G., RUVINSKY I., GILAD Y., 2012 Comparative studies of gene expression and the evolution of gene regulation. Nat. Rev. Genet. **13**: 505–516.

ROSIN D., HORNUNG G., TIROSH I., GISPAN A., BARKAI N., 2012 Promoter nucleosome organization shapes the evolution of gene expression. (HD Madhani, Ed.). PLoS Genet. **8**: e1002579.

SANCHEZ A., CHOUBEY S., KONDEV J., 2013 Regulation of noise in gene expression. Annu. Rev. Biophys. **42**: 469–491.

SAYOU C., MONNIAUX M., NANAO M. H., MOYROUD E., BROCKINGTON S. F., THÉVENON E., CHAHTANE H., WARTHMANN N., MELKONIAN M., ZHANG Y., WONG G. K.-S.,

WEIGEL D., PARCY F., DUMAS R., 2014 A Promiscuous Intermediate Underlies the Evolution of LEAFY DNA Binding Specificity. Science **343**: 645–648.

SCALLY A., DURBIN R., 2012 Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. **13**: 745–753.

SCHADT E. E. E., MONKS S. S. a, DRAKE T. T. a, LUSIS A. J. A., CHE N., COLINAYO V., RUFF T. G., MILLIGAN S. B., LAMB J. R., CAVET G., LINSLEY P. S., MAO M., STOUGHTON R. B., FRIEND S. H., 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **205**: 1–6.

SCHAEFKE B., EMERSON J. J., WANG T.-Y., LU M.-Y. J., HSIEH L.-C., LI W.-H., 2013 Inheritance of gene expression level and selective constraints on trans- and cis-regulatory changes in yeast. Mol. Biol. Evol. **30**: 2121–33.

SCHLITT T., BRAZMA A., 2007 Current approaches to gene regulatory network modelling. BMC Bioinformatics **8 Suppl 6**: S9.

SCHRAIBER J. G., MOSTOVOY Y., HSU T. Y., BREM R. B., 2013 Inferring evolutionary histories of pathway regulation from transcriptional profiling data. PLoS Comput. Biol. **9**: e1003255.

SHARON E., DIJK D. VAN, KALMA Y., KEREN L., MANOR O., YAKHINI Z., SEGAL E., 2014 Probing the effect of promoters on noise in gene expression using thousands of designed sequences. Genome Res. **24**: 1698–1706.

SHARON E., KALMA Y., SHARP A., RAVEH-SADKA T., LEVO M., ZEEVI D., KEREN L., YAKHINI Z., WEINBERGER A., SEGAL E., 2012 Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. **30**: 521–530.

SHEN-ORR S. S., MILO R., MANGAN S., ALON U., 2002 Network motifs in the transcriptional regulation network of Escherichia coli. Nat. Genet. **31**: 64–68.

SHI X., NG D. W.-K., ZHANG C., COMAI L., YE W., CHEN Z. J., 2012 Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. Nat. Commun. **3**: 950.

SHULTZABERGER R. K., MALASHOCK D. S., KIRSCH J. F., EISEN M. B., 2010 The Fitness Landscapes of cis-Acting Binding Sites in Different Promoter and Environmental Contexts (DS Guttman, Ed.). PLoS Genet. **6**: e1001042.

SILANDER O. K., NIKOLIC N., ZASLAVER A., BREN A., KIKOIN I., ALON U., ACKERMANN M., 2012 A Genome-Wide Analysis of Promoter-Mediated Phenotypic Noise in Escherichia coli. PLoS Genet. **8**: e1002443.

SMITH E. N., KRUGLYAK L., 2008 Gene-environment interaction in yeast gene expression. PLoS Biol. **6**: e83.

SMITH J. D., MCMANUS K. F., FRASER H. B., 2013 A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers. Mol. Biol. Evol. **30**: 2509–2518.

STADLER P. F., 2006 Genotype-Phenotype Maps. Biol. Theory **1**: 268–279.

STAMATOYANNOPOULOS J. a, ADZHUBEI I., THURMAN R. E., KRYUKOV G. V, MIRKIN S. M., SUNYAEV S. R., 2009 Human mutation rate associated with DNA replication timing. Nat. Genet. **41**: 393–395.

STERGACHIS A. B., NEPH S., SANDSTROM R., HAUGEN E., REYNOLDS A. P., ZHANG M., BYRON R., CANFIELD T., STELHING-SUN S., LEE K., THURMAN R. E., VONG S., BATES D., NERI F., DIEGEL M., GISTE E., DUNN D., VIERSTRA J., HANSEN R. S., JOHNSON A. K., SABO P. J., WILKEN M. S., REH T. a., TREUTING P. M., KAUL R., GROUDINE M., BENDER M. a., BORENSTEIN E., STAMATOYANNOPOULOS J. a., 2014 Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature **515**: 365–370.

STERN D. L., 2000 Perspective: Evolutionary Developmental Biology and the Problem of Variation. Evolution (N. Y). **54**: 1079.

STERN D. L., ORGOGOZO V., 2008 The loci of evolution: how predictable is genetic evolution? Evolution (N. Y). **62**: 2155–2177.

STERN D. D. L., ORGOGOZO V., 2009 Is genetic evolution predictable? Science **323**: 746–751.

STEWART A. J., PLOTKIN J. B., 2012 Why Transcription Factor Binding Sites are Ten Nucleotides Long. Genetics **192**: 973–985.

STOLTZFUS A., 2006 Mutationism and the dual causation of evolutionary change. Evol. Dev. **8**: 304–317.

STOLTZFUS A., 2012 Constructive neutral evolution: exploring evolutionary theory's curious disconnect. Biol. Direct **7**: 35.

STOLTZFUS A., CABLE K., 2014 Mendelian-Mutationism: The Forgotten Evolutionary Synthesis. J. Hist. Biol. **47**: 501–546.

STOLTZFUS A., YAMPOLSKY L. Y., 2009 Climbing mount probable: mutation as a cause of nonrandomness in evolution. J. Hered. **100**: 637–647.

STREISFELD M. a., RAUSHER M. D., 2010 Population Genetics, Pleiotropy, and the Preferential Fixation of Mutations During Adaptive Evolution. Evolution (N. Y.). **65**: 1–14.

SUN J. X., HELGASON A., MASSON G., EBENESERSDÓTTIR S. S., LI H., MALLICK S., GNERRE S., PATTERSON N., KONG A., REICH D., STEFANSSON K., 2012 A direct characterization of human mutation based on microsatellites. Nat. Genet. **44**: 1161–1165.

SUNG W., ACKERMAN M. S., GOUT J.-F., MILLER S. F., WILLIAMS E., FOSTER P. L., LYNCH M., 2015 Asymmetric Context-dependent Mutation Patterns Revealed through Mutation-accumulation Experiments. Mol. Biol. Evol. **32**: 1672–1683.

SUNG H.-M., WANG T.-Y., WANG D., HUANG Y.-S., WU J.-P., TSAI H.-K., TZENG J., HUANG C.-J., LEE Y.-C., YANG P., HSU J., CHANG T., CHO C.-Y., WENG L.-C., LEE T.-C., CHANG T.-H., LI W.-H., SHIH M.-C., 2009 Roles of trans and cis variation in yeast intraspecies evolution of gene expression. Mol. Biol. Evol. **26**: 2533–8.

SUNSHINE A. B., PAYEN C., ONG G. T., LIACHKO I., TAN K. M., DUNHAM M. J., 2015 The Fitness Consequences of Aneuploidy Are Driven by Condition-Dependent Gene Effects. PLOS Biol. **13**: e1002155.

SWAIN P. S., ELOWITZ M. B., SIGGIA E. D., 2002 Intrinsic and extrinsic contributions to stochasticity in gene expression. PNAS **99**: 12795–12800.

TAKAHASI K. R., MATSUO T., TAKANO-SHIMIZU-KOUNO T., 2011 Two types of cis-trans compensation in the evolution of transcriptional regulation. PNAS **108**: 15276–81.

TEICHMANN S. a, BABU M. M., 2004 Gene regulatory network growth by duplication. Nat. Genet. **36**: 492–496.

THATTAI M., OUDENAARDEN A. VAN, 2001 Intrinsic noise in gene regulatory networks. PNAS **98**: 8614–8619.

TIROSH I., BARKAI N., 2008 Two strategies for gene regulation by promoter nucleosomes. Genome Res. **18**: 1084–1091.

TIROSH I., REIKHAV S., LEVY A. a, BARKAI N., 2009 A yeast hybrid provides insight into the evolution of gene expression regulation. Science **324**: 659–662.

TIROSH I., SIGAL N., BARKAI N., 2010 Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences. Mol. Syst. Biol. **6**: 365.

TIROSH I., WEINBERGER A., CARMI M., BARKAI N., 2006 A genetic signature of interspecies variations in gene expression. Nat. Genet. **38**: 830–834.

TRAVISANO M., SHAW R. G., 2013 Lost in the map. Evolution (N. Y). **67**: 305–314.

TYLER A. L., ASSELBERGS F. W., WILLIAMS S. M., MOORE J. H., 2009 Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. Bioessays **31**: 220–227.

UCHIMURA A., HIGUCHI M., MINAKUCHI Y., OHNO M., TOYODA A., FUJIYAMA A., MIURA I., WAKANA S., NISHINO J., YAGI T., 2015 Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res. **25**: 1–10.

VARDI N., LEVY S., ASSAF M., CARMI M., BARKAI N., 2013 Budding yeast escape commitment to the phosphate starvation program using gene expression noise. Curr. Biol. **23**: 2051–2057.

VIERSTRA J., RYNES E., SANDSTROM R., ZHANG M., CANFIELD T., HANSEN R. S., STEHLING-SUN S., SABO P. J., BYRON R., HUMBERT R., THURMAN R. E., JOHNSON A. K., VONG S., LEE K., BATES D., NERI F., DIEGEL M., GISTE E., HAUGEN E., DUNN D., WILKEN M. S., JOSEFOWICZ S., SAMSTEIN R., CHANG K.-H., EICHLER E. E., BRUIJN M. DE, REH T. a, SKOULTCHI A., RUDENSKY A., ORKIN S. H., PAPAYANNOPOULOU T., TREUTING P. M., SELLERI L., KAUL R., GROUDINE M., BENDER M. a, STAMATOYANNOPOULOS J. a, 2014 Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. **346**: 1007–1012.

VOORDECKERS K., POUGACH K., VERSTREPEN K. J., 2015 How do regulatory networks evolve and expand throughout evolution? Curr. Opin. Biotechnol. **34**: 180–188.

VRBA E. S., ELDREDGE N., 1984 Individuals, hierarchies and processes: towards a more complete evolutionary theory. Paleobiology **10**: 146–171.

VRIES H. DE, 1901 *Die mutationstheorie*. Verlag Von Veit & Comp, Leipzig,.

VRIES H. DE, 1905 *Species and Varieties, Their Origin by Mutation*. Open Court Publishing Company, Chicago.

WAGNER A., 2015 Causal Drift, Robust Signaling, and Complex Disease. PLoS One **10**: e0118413.

WAGNER G. P., LYNCH V. J., 2008 The gene regulatory logic of transcription factor evolution. Trends Ecol. Evol. **23**: 377–385.

WAGNER G. P., ZHANG J., 2011 The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nat. Rev. Genet. **12**: 204–213.

WAGNER G. P., ZHANG J., 2012 Universal pleiotropy is not a valid null hypothesis: reply to Hill and Zhang. Nat. Rev. Genet. **13**: 296.

WANG Z., LIAO B. B.-Y. B.-Y., ZHANG J., 2010 Genomic patterns of pleiotropy and the evolution of complexity. PNAS **107**: 18034–18039.

WANG D., SUNG H.-M., WANG T.-Y., HUANG C.-J., YANG P., CHANG T., WANG Y.-C., TSENG D.-L., WU J.-P., LEE T.-C., SHIH M.-C., LI W.-H., 2007 Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. Genome Res. **17**: 1161–9.

WANG Z., ZHANG J., 2011 Impact of gene expression noise on organismal fitness and the efficacy of natural selection. PNAS **108**: E67–E76.

WHITEHEAD A., CRAWFORD D. L., 2006 Variation within and among species in gene expression: raw material for evolution. Mol. Ecol. **15**: 1197–1211.

WITTKOPP P. J., 2005 Genomic sources of regulatory variation in cis and in trans. Cell. Mol. Life Sci. **62**: 1779–83.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2004 Evolutionary changes in cis and trans gene regulation. Nature **430**: 85–88.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2008a Regulatory changes underlying expression differences within and between *Drosophila* species. Nat. Genet. **40**: 346–350.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2008b Independent effects of cis- and trans-regulatory variation on gene expression in Drosophila melanogaster. Genetics **178**: 1831–5.

WLOCH D. M., SZAFRANIEC K., BORTS R. H., KORONA R., 2001 Direct estimate of the mutation rate and the distribution of fitness effects in the yeast Saccharomyces cerevisiae. Genetics **159**: 441–452.

WOLF L., SILANDER O. K., NIMWEGEN E. J. VAN, 2015 Expression noise facilitates the evolution of gene regulation. Elife **4**: e05856.

WOODRUFF R. C., THOMPSON J. N. (Eds.), 1998 *Mutation and Evolution*. Springer Netherlands, Dordrecht.

WRAY G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. **8**: 206–216.

WRAY G. a, HAHN M. W., ABOUHEIF E., BALHOFF J. P., PIZER M., ROCKMAN M. V, ROMANO L. A., 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**: 1377–1419.

XIAO S., XIE D., CAO X., YU P., XING X., CHEN C.-C., MUSSELMAN M., XIE M., WEST F. D., LEWIN H. a, WANG T., ZHONG S., 2012 Comparative epigenomic annotation of regulatory DNA. Cell **149**: 1381–1392.

XU Z., WEI W., GAGNEUR J., PEROCCHI F., CLAUDER-MÜNSTER S., CAMBLONG J., GUFFANTI E., STUTZ F., HUBER W., STEINMETZ L. M., 2009 Bidirectional promoters generate pervasive transcription in yeast. Nature **457**: 1033–1037.

YANG S., WANG L., HUANG J., ZHANG X., YUAN Y., CHEN J.-Q., HURST L. D., TIAN D., 2015 Parent–progeny sequencing indicates higher mutation rates in heterozygotes. Nature **5**: 463–467.

YOKOYAMA K. D., OHLER U., WRAY G. a, 2009 Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. Nucleic Acids Res. **37**: e92.

YONA A. H., MANOR Y. S., HERBST R. H., ROMANO G. H., MITCHELL A., KUPIEC M., PILPEL Y., DAHAN O., 2012 Chromosomal duplication is a transient evolutionary solution to stress. PNAS **109**: 21010–21015.

YU H., GERSTEIN M., 2006 Genomic analysis of the hierarchical structure of regulatory networks. PNAS **103**: 14724–14731.

YVERT G., BREM R. B., WHITTLE J., AKEY J. M., FOSS E., SMITH E. N., MACKELPRANG R., KRUGLYAK L., OTHERS, 2003 Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat. Genet. **35**: 57–64.

ZAUGG J. B., LUSCOMBE N. M., 2011 A genomic model of condition-specific nucleosome behaviour explains transcriptional activity in yeast. Genome Res. **22**: 84–94.

ZEYL C., DEVISSER J. a, 2001 Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae. Genetics **157**: 53–61.

ZHANG X., BOREVITZ J. O., 2009 Global analysis of allele-specific expression in Arabidopsis thaliana. Genetics **182**: 943–54.

ZHANG Z., QIAN W., ZHANG J., 2009 Positive selection for elevated gene expression noise in yeast. Mol. Syst. Biol. **5**: 1–12.

ZHANG J., WAGNER G. P., 2013 On the definition and measurement of pleiotropy. Trends Genet. **29**: 383–384.

ZHENG W., GIANOULIS T. a, KARCZEWSKI K. J., ZHAO H., SNYDER M., 2011 Regulatory variation within and between species. Annu. Rev. Genomics Hum. Genet. **12**: 327–346.

ZHU Y. O., SIEGAL M. L., HALL D. W., PETROV D. a, 2014 Precise estimates of mutation rate and spectrum in yeast. PNAS **111**: E2310–E2318.

ZHUANG Y., ADAMS K. L., 2007 Extensive allelic variation in gene expression in populus F1 hybrids. Genetics **177**: 1987–96.

ZUCKERKANDL E., PAULING L., 1965 *Evolutionary divergence and convergence in proteins* (V Bryson and HJ Vogel, Eds.). Academic Press, New York.

ZUCKERLANDL E., PAULING L., 1962 Molecular Disease, Evolution, and Geneic Heterogenetiy. In: Kasha M, Pullman B (Eds.), *Horizons in Biochemistry*, Academic Press, pp. 189–225.

Chapter 2

Whole genome sequencing and high-throughput phenotypic analysis of diverse *Saccharomyces cerevisiae* strains[1]

Abstract

Understanding the relationship between genetic variation and phenotypic consequences is a central goal of biology. Here we present a set of 85 high quality genomes of *Saccharomyces cerevisiae* strains that represent much of the genotypic and phenotypic diversity within the species. We find a complex pattern of phylogenetic structure that has previously masked signatures of positive selection. In addition, we find that non-reference genes, *de novo* genes, intron loss events, and copy number variants typically do not have the same phylogenetic relationships as single nucleotide polymorphisms. Finally, we construct a set of genetically tractable strains from the sequenced set of strains that will prove a valuable tool to the community. We demonstrate the utility of these strains by measuring their competitive fitness effects across a range of environmental conditions using bar-seq. Using these fitness estimates, we perform a genome-wide association study (GWAS) and detect genetic variants that are associated with altered growth rates amongst strains.

[1]This chapter will be submitted as: Calum J. Maclean*, Brian P.H. Metzger*, Jian-Rong Yang*, Wei-Chin Ho, Bryan Moyers and Jianzhi Zhang. *Equal contributions

Introduction

A central question in biology is how information encoded within a genome determines an organism's phenotypes. As the first eukaryote to have its genome fully sequenced, the budding yeast *Saccharomyces cerevisiae* has become an important model system for addressing this question. Large scale determination of phenotypes upon gene deletion (GIAEVER *et al.* 2002) and overexpression (DOUGLAS *et al.* 2012), as well as the availability of genomes from closely related species (DUJON *et al.* 2004; KELLIS *et al.* 2004; CLIFTEN *et al.* 2006; SCANNELL *et al.* 2011; LIBKIND *et al.* 2011; LITI *et al.* 2013; HITTINGER 2013) have made *S. cerevisiae* a valuable tool in understanding genome evolution. However, much of the population level genetic and phenotypic diversity within *S. cerevisiae* has been overlooked due to focus on only a few laboratory strains. This issue is compounded by the recent finding that laboratory strains, in particular the reference strain S288c, represents a phenotypic outlier (WARRINGER *et al.* 2011).

Recently, there has been increased interest in the natural population diversity and ecology of this important model organism. *S. cerevisiae* has been isolated globally from diverse natural and man-made environments, each of which present potentially unique challenges and adaptive paths, and therefore offers a rich system in which to investigate the relationship between genotype and phenotype. The hunt for genetic variants responsible for phenotypic differences has been continuing in earnest, particularly since low coverage whole capillary sequencing (LITI *et al.* 2009) and tiling array hybridization (SCHACHERER *et al.* 2009) have greatly increased the pool of known genetic variants. These studies, as well as restriction site associated DNA sequencing (RAD-seq) of a large strain set

(CROMIE *et al.* 2013), have revealed the complex population structure of the species. Recently, the addition of high quality genomes produced through next generation sequencing has improved our understanding of the relationship between genome structure and phenotype through the determination of copy-number variation and genomic rearrangements within natural populations (Hose *et al.* 2015; Strope *et al.* 2015).

Because of influences from human activity, many *S. cerevisiae* strains have complex ancestry from multiple lineages and are mosaics. In addition, strains representing pure lineages are often phenotypically distinct (WARRINGER *et al.* 2011). This strong and complex population structure has made genome-wide association studies, an important tool in detecting influential genetic variants in many species, particularly difficult in yeast and requires careful consideration when looking for meaningful associations (CONNELLY and AKEY 2012; DIAO and CHEN 2012).

Here we improve upon the resources available to the community by producing high quality genomes for 85 yeast strains isolated from a wide range of laboratory, clinical, natural, and industrial environments distributed across the globe. Using both reference genome mapping and *de novo* assembly of the short reads, the resulting genomes provide a valuable resource from which phenotypically important single base polymorphisms (SNPs) can be identified. These assemblies also allow the identification of large scale genetic variants, previously absent from the majority of genome builds. These variants are likely important in explaining the broader range of phenotypes evident across *S. cerevisiae* strains (0.8% sequence divergence) relative to its more polymorphic, but

phenotypically static, sister species *S. paradoxus* (~3.5% sequence divergence) (ZÖRGÖ *et al.* 2012; BERGSTRÖM *et al.* 2014).

We also introduce a large set of genetically barcoded strains for use in future studies developed from the sequenced strain set. To demonstrate the utility of these strains, we simultaneously phenotype all barcoded strains across several environments. In combination with our high quality genome sequence, we use these phenotypes to perform a genome-wide association study (GWAS). We believe the resources presented in this work will be of great use to both the yeast and genetics communities in future studies.

Methods

Strains and strain construction

The strains used in this work were obtained from the authors of two previous studies (LITI *et al.* 2009; SCHACHERER *et al.* 2009) and are listed in Table A-1. The geographic location and environment from which each strain was isolated is listed when known.

Most strains used in this work were initially diploid, contain no tractable genetic markers, and are homothallic. This makes tracking strains difficult and the maintenance of stable haploid strains, necessary for many studies, impossible. To produces a set of strains useful to the community, we adapted the approach used in the construction of the *S. cerevisiae* gene deletion collection to introduce drug resistance markers flanked by two unique, strain identifying, 20bp DNA barcodes at the *HO (YDL227c)* locus of each strain. This simultaneously removed the strains ability to mating type switch and introduced a

reliable means by which it can be tracked. Diploid strains were transformed using the lithium acetate method as performed by Cubillos et al (2009) with minor alterations. The ~1µg of transforming *HO*-targeting DNA contained a G418 sulfate resistance marker flanked by strain specific barcodes and was produced by two successive PCR amplifications. We first amplified the KanMX4 cassette from plasmid pFA6a-KanMX4 (WACH *et al.* 1994) using two 74bp primers each contained a unique 20bp barcode, the sequences necessary for its amplification (U1 + U2 or D1 + D2), and priming sites for the second PCR step. The second PCR used a dilution of PCR 1 as the DNA template and added sequences homologous to regions upstream and downstream of HO for targeting and replacement of the locus. The primers used in this second round differed by strain to maintain lineage specific SNPS in the region. A full list of the primers used can be found in Table A-2. To ensure that the barcodes assigned to each strain were novel and maintain their compatibility with existing gene deletions (GIAEVER *et al.* 2002), plasmid collections (HO *et al.* 2009), and existing technologies used to estimate barcode frequency, we employed unused barcode sequences already present on the widely used Tag4 array (PIERCE *et al.* 2006). We confirmed successful insertion of the KanMX4 cassette using the primers listed in Table A-6.

From these heterozygous *HO* marked diploids (a/α HO/ho::Uptag-KanMX4-Downtag), stable haploid strains were obtained by sporulation on potassium acetate media followed by ascus digestion and tetrad dissection. G418 resistant colonies were identified by replication to YPD media containing 300ug/ml G418 sulfate. Colony PCR was used to determine the mating type of individual colonies and single MAT**a** and MATα colonies

were streaked to obtain a pure strain of each mating type. Samples were grown overnight and frozen at -80C in 20% glycerol for long term storage. To allow for the easy formation of diploids between any two strains, we switched the drug resistance cassette carried by MAT**a** and MATα strains to hygromycin B and nourseothricin respectively. This was achieved using standard LiAc method using a PCR product produced by the use of primers specific to the TEF promoter and terminator common to all three drug resistance cassettes (WACH *et al.* 1994; GOLDSTEIN and MCCUSKER 1999).

Two sets of strains were treated slightly differently due to their genotypes. First, RM11a was previously made into a stable haploid strain by insertion of a *KanMX4* cassette at the *HO* locus and the region targeted resulted in the deletion of the targeting region we used in all other strains (BREM *et al.* 2002). To insert the appropriate barcodes into this background, unique homologous primers were used to target and replace the existing *KanMX4* cassette with a *HphMX4* marker amplified from plasmid pAG32 (GOLDSTEIN and MCCUSKER 1999). Unique barcodes were then added and the cassette was switched back to *KanMX4*. Second, three strains (S288c, W303 and RM11) were already heterothallic haploids. After insertion of the barcoded cassette at the *HO* locus, these strains were transformed with plasmid pCM66 to obtain strains of both mating types. pCM66 contains a galactose inducible copy of *HO* and a nourseothricin drug resistance marker. After transformation, nourseothricin resistant cells were grown with galactose as the sole carbon source at 30°C without shaking for 8 hours to induce expression of *HO*. This allowed for mating type switching and subsequent mother-daughter cell mating to produce diploids. Cells were then streaked for single colonies on YPD plates and the

ploidy of single colonies checked by colony PCR using mating-type specific primers. Diploid colonies were streaked for single colonies on fresh, non-selective, YPD plates and assayed for nourseothricin resistance. A single colony unable to grow in the presence of the drug, and therefore having lost the plasmid, was selected for each strain.

We attempted to produce genetically tractable strains for each of the 85 strains whose genomes we sequenced, but found some to be unamenable to our approach, either though natural immunity to the drugs used or an inability to successfully sporulate and produce viable offspring of both mating types. Full details of all tractable strains created and the reason for missing strains are outlined in Table A-1.

Library construction and Sequencing

Each of the 85 strains was streaked from frozen stocks onto YPD media plates (1% Yeast extract, 2% Peptone, 2% glucose, 2% agar) and following 2 days of growth a single colony was picked into 5ml of liquid YPD media and grown to saturation (36h at 30°C with shaking). Cultures were centrifuged to collect cells and DNA was extracted using standard methods. Dried DNA pellets were resuspended in 70µl of Tris-EDTA (pH8.0), the DNA quantified and purity assessed before storage at -80°C until needed.

Illumina libraries were constructed using a protocol modified from Rohland and Reich (2012). In brief, 5µg of genomic DNA were sheared using a Covaris S220 (Duty cycle 10%, Intensity 4, cycles/burst 200, time 55s) of which 2µg was used in library construction. To select DNA fragments of the desired size range (approx. 400bp), we

used DNA binding Magna beads to perform dual size selection. The fragments were blunt-end repaired, adapter ligated, and nick filled to repair the adapter overhangs. Finally, sequences necessary for multiplexing and cluster formation on an Illumina HiSeq2000 were added by PCR. Equal amounts of each library were combined and run across two paired-end 100bp lanes (43 strains in one lane and 42 in a second) of an Illumina HiSeq2000 at the University of Michigan DNA sequencing core.

Read mapping and SNP/Indel calling

Reads were first trimmed using Cutadapt (MARTIN 2011) to remove adaptor sequences (ENGEL *et al.* 2013). Bowtie2 v2.1.0 (LANGMEAD and SALZBERG 2012) was used to map reads to the S288c reference (R64-1-1) genome under the sensitive local alignment mode, allowing up to 3 mismatches/indels per read. Pertinent statistics obtained during the mapping process are listed in Table A-3. Paired reads were considered non-concordant and discarded from further analysis if apparent mapping locations were more than 1200bp apart, or if the read pair reads appeared to completely overlap one another. Paired reads were also removed from further analysis if either end was found to map ambiguously. Finally, we removed PCR duplicates by discarding all but one copy of any read pairs found to map to exactly the same genomic position.

Samtools v0.1.18 (LI *et al.* 2009) and VarScan v2.3.6 (KOBOLDT *et al.* 2012) were used to identify SNPs and indels within each genome. Only variants identified by both programs were used in our downstream analysis. To further reduce false calls due to misalignment of reads to the reference genome, variants that showed a significant strand

bias (binomial P < 0.001), or were located close to the end of the supporting read (VDB < 0.15) were also removed (DANECK *et al.* 2012). Only the most likely variant is listed for the indel and homozygous SNP lists. For the heterozygous SNP list, maximum likelihood genotype is reported. To reduce errors in estimating allele frequencies, only segregating sites with reads covering the variant in each of the 85 strains were used except when identifying pseudogenizing variants.

Phylogeny construction

We constructed a maximum composite likelihood neighbor-joining tree using MEGA 5.2 and all homozygous SNPs and substitution types (TAMURA *et al.* 2011). We allowed heterogeneous rates amongst lineages and heterogeneous rates amongst sites. Clades were identified in line with previous studies (Liti *et al.* 2009; Schacherer *et al.* 2009) and the tree was rooted using data from WANG *et al.* 2012. To assess the strength of support for the phylogeny we performed 100 bootstraps. We report all values below 100 other than in the Wine/European clade where the exact relationship amongst strains was poorly resolved and tended to contain nodes with <100 bootstrap support. Individual chromosome phylogenies were constructed using the same parameters.

Population genomics analysis

We calculated linkage disequilibrium (HARTL and CLARK 2006) between every pair of SNPs using custom code. To perform population genomic analyses, only SNP sites that were dimorphic and for which their ancestral state could be unambiguously assigned were used. To determine SNP ancestral states, we took advantage of the orthology

information and multi-species genome sequences published by Scannell *et al.* (2011). We used T-Coffee (NOTREDAME *et al.* 2000) and the default settings in BioPerl to align the coding sequences of *S. paradoxus*, *S. mikatae*, and *S. bayanus* with the orthologous coding *S. cerevisiae* reference (R64-1-1). Using these multiple-sequence alignments, we considered the states of *S. paradoxus*, *S. mikatae*, and *S. bayanus* for each SNP site and unambiguously assigned its ancestral state if at least two of these outgroup species were in agreement. Using the SNP frequency spectrum, we calculated Tajima's D (TAJIMA 1989), Fu and Li's F (FU and LI 1993), and Fay and Wu's H (Fay and Wu 2000; Zeng *et al.* 2006).

To assess selection at the protein level, we counted the number of synonymous polymorphisms (P_s), nonsynonymous polymorphisms (P_n), synonymous substitutions between *S. cerevisiae* and *S. paradoxus* (D_s) and nonsynonymous substitutions between *S. cerevisiae* and *S. paradoxus* (D_n) for each gene. We performed a two-tailed Fisher's exact test within R applying a Bonferroni multiple test correction. To calculate the proportion of amino acids under positive selection we first determined whether D_s/P_n was larger than D_n/P_s (SMITH and EYRE-WALKER 2002). For genes under positive selection, D_s/P_n is larger than D_n/P_s, while for genes subject to purifying selection, D_s/P_n is smaller than D_n/P_s. For genes showing signs of positive selection, we calculated α as 1-$D_s \cdot P_n/D_n \cdot P_s$. For genes with evidence of purifying selection, we calculated α' as 1-$D_n \cdot P_s/D_s \cdot P_n$.

Following Messer and Petrov (2013), we applied the extended version of the McDonald-Kreitman test to all genes. Starting with all SNPs, regardless of allele frequency, we sequentially removed SNPs with the lowest derived allele frequency and recalculated α and α'. Using MATLAB, we applied nonlinear least square methods to fit the results to the function y=a+b*exp(-cx), using α values for as y and different derived allele frequencies as x. We restricted $b < 0$ and $c > 0$. To ensure that specific strains did not have large influences on our estimates of α, we randomly sampled 85 strains with replacement 100 times and calculated α values for different derived allele frequencies (DAFs) within each set.

In yeasts, genes with high expression levels tend to have high codon bias (COGHLAN and WOLFE 2000) as measured by the codon-adaptation index (SHARP and LI 1987). To determine the effects of codon usage, we used Codon Adaptation Index (CAI) data from Qian *et al.* (2012). To calculate the derived allele frequency for synonymous polymorphisms in genes with high levels of codon bias we used only genes with an average CAI > 0.6.

To assess the population structure of the 85 strains, we used a model-based Markov Chain Monte Carlo (MCMC) algorithm implemented using STRUCTURE v2.3.4 (PRITCHARD *et al.* 2000). For genome-wide population structure analysis, we randomly selected 10% of the non-singleton homozygous SNPs. STRUCTURE performed 100,000 rounds of burn-in (first 50,000 with population admixture model enabled), followed by 100,000 rounds of MCMC. After three runs for each of K=2 to 8, K=7 was found to be

most appropriate (EVANNO *et al.* 2005). The population structure that exhibited the maximum mean likelihood was plotted using distruct v1.1 (ROSENBERG *et al.* 2002). Finally, the population structure of each individual chromosome was determined using all homozygous SNPs on the specific chromosome at K=7 using the same parameters as the genome-wide analysis.

De novo assembly and identification of new genes

De novo genome assembly was performed using SOAPdenovo2 v2.04 (LUO *et al.* 2012) with K=51 for all adaptor-trimmed reads from each of the 85 strains. Basic statistics of the genome assemblies are listed in Table A-4. To check the quality of the assemblies, we used BLASTn to search for the kanMX4 vector sequence present at the HO locus in the majority (81/85) of and used it as an anchor to extract the strain-specific barcodes (UPTAG and DSTAG). We successfully recovered the strain specific barcodes for all 81 marked strains. To rule out false positive *de novo* gene calls, Exonerat v2.2.0 (SLATER and BIRNEY 2005) was used to align known reference genome (S288c) genes to the assembled contigs. Known genes were localized onto contigs by prioritizing the Exonerate hits by (1) best hit with syntenic neighbor genes at either side, (2) best hit with 100% query sequence coverage, and (3) hits longer than 200 nt that do not overlap with better hits by more than 30 nt - in case a gene is split among multiple contigs. Having identified the locations of known genes on *de novo* contigs we then used GeneMarkS v4.17 (BESEMER *et al.* 2001) to perform gene predictions and compared these to the locations of known genes.

Predicted genes that showed no overlap with known gene locations were considered candidate non-reference genes. To avoid false positives caused by un-localized known genes, we used BLASTn to align the predicted genes with cDNA sequences of known genes. All predicted genes with hits covering 80% of the query, or with a > 200nt region that is > 90% identical with any reference gene, were removed. In order to classify the origin of remaining new genes, we retrieved the best hit in the NCBI "nr" database reported by BLASTn and tBLASTx for each new gene. If the best hit for a candidate new gene was the reference S288c genome, it was also removed unless there was a premature stop codon for the hit region in the S288c genome. Finally we removed non-reference genes with best hits to sequences derived from vectors, synthetic constructs, phage, bacterial genomes, or tandem elements from further analysis (Table A-7).

To access the expression level of the non-reference genes, RNA-seq data for the 23 available strains were downloaded (SKELLY *et al.* 2013). The color space RNA-seq reads from each strain were mapped to known and predicted new genes using bowtie (LANGMEAD *et al.* 2009), allowing up to 2 mismatches in color space. The best hits for each read were then used to calculate the RPKM for each gene.

Identification of intron loss event

Based on S288c reference genome annotation, we built a sequence database containing all known exon-exon junctions up to 101 bp from either side of the junction. To search for reads supporting an intron loss event, all reads were mapped to this database by Bowtie2 (LANGMEAD and SALZBERG 2012). We required at least 95% coverage of the

query and that the read mapped to least 20bp in both exons. We further filtered ambiguous mappings on the reference genome by BLASTn with e-value cutoff at 0.01. To search for intron conservation reads, the same procedure was conducted for all sequences annotated as exon-intron borders in the S288c reference genome. Finally, intron loss was declared in strains in which at least two reads spanning the exon-exon junction, but no reads spanning the two corresponding exon-intron borders, were detected.

Identifying potential copy number variants (CNV) and aneuploidies

To assess gene duplication/deletion events, read pairs that concordantly mapped to the reference genome following filtering of potential PCR duplicates were analyzed using Cufflinks v2.1.1 (TRAPNELL *et al.* 2010) to generate FPKM values for each CDS. Potential CNVs of individual genes, as well as aneueploidies and large scale duplications, were then identified by dividing each FPKM by that obtained from the same CDS in the reference strain S288c.

TATA box evolution

To investigate TATA box evolution in our sequenced strains, we collected canonical and non-canonical TATA box position annotations for *S. cerevisiae* genes from Rhee and Pugh (2012). The expression levels of yeast genes are from (NAGALAKSHMI *et al.* 2008) and the expression levels of 21 yeast strains are from (SKELLY *et al.* 2013). To determine the expected frequency of changed/unchanged SNPs in different positions of the TATA box, we determined the genome wide nucleotide composition (f(A) = f(T) = 0.309, and

f(G) = f(C) = 0.191). The expected frequencies of changed/unchanged SNPs in the highest and lowest expressed genes was calculated by the average frequencies across all eight positions.

Simultaneous phenotyping of barcoded yeast strains

To estimate the phenotypes of all strains, we used the unique barcodes inserted into each strain to calculate their relative fitness across several diverse environmental conditions by bar-seq (SMITH *et al.* 2009). Each of the barcoded genotypes were mixed equally and combined with the diploid homozygous gene deletion collection (Invitrogen). To ensure good coverage of the non-deletion yeast strains, non-deletion strains were present at approximately twice the initial population size of the gene deletion strains. The initial pool of strains was grown for approximately two generations in 25ml of YPD media (1% yeast extract, 2% peptone, 2% glucose) at 30°C before the resulting culture, termed generation 0, was used to initiate competitions in each of the experimental conditions. To reduce the potential effect of drift, large populations were maintained throughout competitions with regular transfers to fresh media (every 4-5 generations) to maintain populations in exponential growth. Populations were competed for approximately 30 generations (6 transfers). Remaining samples were stored at -80C following each transfer. Populations were competed in YPD at 30°C, YPD at 40°C, YPD + 1.25M NaCl, YPD + 8% EtOH, YPD + 4mM paraquat (superoxide), YPD + 3mM Hydrogen peroxide, and YPD + 1 mM cobalt chloride.

To determine the frequency of strains in the pooled population at a given time point, we extracted genomic DNA from samples using a Puregene Yeast/Bacteria DNA extraction kit (Qiagen). DNA barcodes were amplified by polymerase chain reaction (PCR) using Accuprime pfx (Invitrogen). The primers used for barcode amplification also result in the addition of sequence necessary for cluster formation and sequencing primer annealing on the Illumina platform. Because the downstream barcode is known to be missing in some deletion strains (DEUTSCHBAUER *et al.* 2005), only the upstream barcodes were used. Fifty base-pair single end sequence reads were obtained using one lane of an Illumina Genome Analyzer IIx at the University of Michigan DNA Sequencing Core. The Illumina Pipeline software version 1.6 was used for base calling from the image data. Because all sequences started with the same 18 base pairs of the PCR primer region and this uniformity adversely affected base calling, we removed the first 18 sequencing cycles before base calling. We used the previously published "gene-barcode map" (QIAN *et al.* 2012a) with the addition of our own strains barcode identities to assign each read to a particular strain allowing for only a single mismatch. We followed the outline of Bar-seq analysis as provided by Robinson et al (2014). We required that barcodes be represented by at least 40 total counts over each of the timepoints (across generation 0 samples as well as the end time point population data for each of the 7 conditions). A pseudo-count was added to each total for all remaining genes. Count numbers were then normalized using the TMM method implemented in the R package egdeR (ref). Because we are focused on phenotypes that are specific to a given condition, as opposed to whether a strain is generally fitter over all, we took the log2 ratio of the normalized counts obtained from a given gene relative to the mean normalized counts of the same

generation in YPD (benign condition). Overall there were 76 natural *S. cerevisiae* strains and 4498 deletions with usable data.

Genome-wide association study and causal variant confirmations

We followed a multistep GWAS approach as outlined by (LISTGARTEN *et al.* 2012, 2013). For each environment, we first scaled and centered the fitness values. We removed all sites with a minor allele frequency below 5% across the set of strains for which we had phenotype data. We then removed sites for which we were missing information in greater than 5% of strains and all mitochondrial sites. The remaining 120,314 sites were converted into a 0, 0.5, 1 format to represent homozygous non-reference, heterozygous, and homozygous reference states in each strain respectively.

Individual SNPs can be significantly associated with a phenotype either because they are true positives or because they are confounded with population structure. Because over or under correcting of population structure can lead to lack of power or false positives, we performed the genome wide association in a number of steps to attempt to find an appropriate balance. First, we performed an association between the normalized fitness values and each SNP without controlling for population structure using a simple linear regression method without covariates. Second, we used this unstructured association to rank all SNPs based on their statistical significance of association. From this list, we performed a series of associations by maximum likelihood using EMMA (KANG *et al.* 2008). To control for population structure, we estimated a kinship matrix based on a specific set of SNPs. To define this set, we started with the 1000 SNPs most significantly

associated with the phenotype in the unstructured analysis and then successively added the next 1000 most significant SNPs from the unstructured analysis until 120,000 SNPs were included. For each association, the genomic control factor, lambda, was calculated using gcontrol2 within R (DEVLIN and ROEDER 1999). We identified the minimal kinship set that controlled for population structure based on where lambda first hit 1, or if it failed to do so, was minimized (Figure A-3). Third, we ran an additional series of associations centered on the 3000 SNP region identified by lambda using kinship sets in 50 SNP windows. Again, we found the smallest kinship set where lambda hit 1, or was at its minimum. Finally, we performed an association for the 1000 most significant SNPs from the unstructured association. In each case, we used the estimated kinship set that was optimal for lambda, minus any SNPs within 10 kb of the focal SNP, to estimate population structure. Any variant in this final association with a p-value below 0.0001 was classified as significant. For these SNPs, we identified the coding region nearest to the variant as well as its immediate neighbors. These candidate SNPs generally cover a 4-10 kb region which is approximately twice the range over which LD is seen to break down. Figure A-1 outlines the association procedure used.

Results and Discussion

Whole-genome resequencing of 85 diverse yeast strains reveals a complex population structure

To improve the genomic resources available to the yeast community, we sequenced 85 *S. cerevisiae* strains with the aim of making high quality genome sequences readily available to researchers. This genetically and phenotypically diverse set of strains was

collected from numerous countries distributed across 6 continents and from a variety of human associated and wild environments (Figure 2-1A, Table A-1). We obtained an average of 3.75 million 100bp paired-end reads per genotype - approximately 97% of which were successfully mapped to the S288c reference genome. This resulted in an average coverage of 60X per genome (range 38X-99X) (Table A-3). On average, only 0.06% of the reference genome was not covered by a read in each sequenced strain. The lack of coverage at these sites is likely due to a combination of stochastic sampling of reads and strain differences in gene content. In total, >86% of sites were covered >10X and allowed us to identify 311,288 SNPs (~2.6% of possible sites) and 15,884 indels (~0.1% of possible sites), providing an unprecedented overview of yeast population genomics.

To visualize the evolutionary relationships between the sequenced strains, we constructed a neighbor-joining tree (Figure 2-1B). We determined the root of this phylogeny using outgroup sequences derived from recently identified Chinese isolates (WANG *et al.* 2012). We recovered phylogenetic clustering based on geographical origins and the environment from which each strain originates, resulting in a phylogeny that is consistent with those previously published using fewer strains (Liti *et al.* 2009; Schacherer *et al.* 2009). For example, clear clustering can be seen of strains into the West African, North American, Malaysian, Sake, Laboratory, and European/Wine groups previously identified. We also identify a newly resolved "Bakery" clade that was previously missed due to small sample sizes. To further examine the population structure of the sequenced strains, we employed a model-based clustering algorithm implemented in STRUCTURE (PRITCHARD *et al.*

2000). We again identified seven distinct sub-populations (Figure 2-2), which are in close agreement with the strain isolation sources and corroborate the clustering pattern seen in the phylogeny.

Previous analysis of *S. cerevisiae* genomes has indicated that many strains are mosaics with ancestry from several lineages (LITI *et al.* 2009). To determine the extent of mosaicism in the sequenced strains, we assessed population structure for each individual chromosome. As with the whole genome analysis, we used both neighbor joining trees (Figure A-2) and the program STRUCTURE to determine population structure (Figure 2-2). We found significant variation across strains in terms of the origin of the genetic material that makes up their individual chromosomes. For example, although the five strains constituting the laboratory clade (S288c, W303, A364A, BY4716 and CEN.PK) have a clade specific genetic make-up, chromosomes II and III appear very similar to chromosomes found in Sake and Wine/European strains, respectively. Similarly the widely used strains Y55 and SK1 show a mixed chromosomal heritage. Although they largely resemble the chromosome structure of the West African strains, several chromosomes are more similar to those of Wine/European ancestry. This variation in chromosomal population structure is indicative of differences in the evolutionary histories of related strains due to pervasive gene flow. As *S. cerevisiae* largely reproduces asexually (TSAI *et al.* 2008), unique populations can be established by rare crosses between distinct lineages allowing distinct chromosome combinations to persist without outbreeding. The observed differences in chromosomal population structure is unlikely

due to stochasticity in the structural assessment since multiple runs on the same chromosome displayed only minor variation (data not shown).

To evaluate the extent of recombination among the sequenced strains, we calculated linkage disequilibrium (LD) between each pair of SNPs (Figure 2-3A). We found that linkage disequilibrium decreased smoothly with the increase of physical distance and was similar to earlier reports by Liti *et al.* (2009) and Schacherer *et al.* (2009). However, compared with these previous finding, the rate of decrease in LD was slower and does not reach half maximum value. This suggests either a lower recombination rate than previously thought, or a more complex population structure among our set of strains than the smaller sets previously analyzed.

Natural selection within the yeast genome

To determine the effects of natural selection on the yeast genome, we took advantage of genome sequences from closely related *Saccharomyces* species to determine the likely ancestral state at each SNP and estimate the derived allele frequency (DAF) spectrum. We used this distribution to study the effects of natural selection on protein sequence and categorized SNPs into three groups: synonymous, nonsynonymous, and nonsense. For each SNP category, we calculated the population genetics statistics Tajima's D (TAJIMA 1989), Fu and Li's F (FU and LI 1993) and Fay and Wu's H (FAY and WU 2000)(Table 2-1). We found that each SNP category had a significantly negative D and F value, indicating an excess of rare variants (Figure 2-3B, Table 2-1). This excess of rare variants was strongest for nonsense sites, followed by non-synonymous sites, and then

synonymous sites. This pattern is consistent with stronger purifying selection against nonsense mutations that prematurely truncate proteins compared to nonsynonymous amino acid changes or synonymous changes that maintain the amino acid composition of the protein. We also found that synonymous and non-synonymous sites had a significantly high H value, indicating an excess of SNPs with high DAF. These H values were more negative for synonymous sites than for non-synonymous sites and indicate positive selection acting within the yeast genome. We also looked for signs of selection on synonymous polymorphisms to maintain codon usage by comparing the frequency spectrum of synonymous SNPs in highly expressed genes (mean codon adaptation index (CAI) > 0.6) with the frequency spectrum of all synonymous SNPs. We identified an excess of low frequency polymorphisms for synonymous sites with high CAI genes, suggesting there is stronger purifying selection working on such synonymous polymorphisms (Figure 2-3B).

To determine the proportion of adaptive mutations, we calculated α for each gene using *S. paradoxus* as an out-group. We found that the distribution of α is largely consistent with widespread purifying selection and relatively few instances of positive selection (Figure 2-1). Consistent with this finding, McDonald-Kreitman tests (Bonferroni corrected) of individual genes failed to detect a significant signal of positive selection for any gene. This result is consistent with previous analyses on a subset of these strains (LITI *et al.* 2009). By contrast, McDonald-Kreitman tests identified 4.6% of genes as under significant purifying selection (Bonferoni corrected p-value < 0.05).

Because slightly deleterious alleles can alter the distribution of DAFs, we further analyzed α using the approach proposed by Messer and Petrov (2013) in which α values are calculated for polymorphic sites with different levels of DAF. Surprisingly, this approach converges on an estimated value of 0.5 for α, suggesting that ~50% of sequence divergence in the population is due to positive selection (Figure 2-5A). Interstingly, the plot of α vs DAF showed a clear valley around intermediate DAFs which we believe has not been observed previously. Suspecting the reason might be heterogeneous selection across strains, we partitioned strains based on phylogenetic clustering. We found that if we seperated strains from the Wine/European cluster from all other strains and performed the same analysis on the two groups independently, the plots of α vs DAF were dramatically different from one another. The extrapolated α values were -0.5329 for the Wine/European cluster and 0.5374 for all other strains. This difference in the estimate of α was not caused by sampling in general, but the specific partitions chosen (Figure 2-5B and C). In addition, the α vs DAF plot for strains not in the Wine/European cluster showed a better fit to an exponential curve than the combined analysis (Without Wine/Europen strains adjusted $R^2 = 0.95$ vs $R^2 = 0.56$ for all strains). Overall, these results suggest that the historical action of natural selection within thse two groups has been very different and that positive selection in yeast may be more common than initially expected, especially outside of domesticated wine strains.

Identification of new genes

Because *S. cerevisiae* strains demonstrate a wide range of phenotypes and have been isolated from a diverse set of environments, optimum growth in their natural habitat may

require a unique set of genes. Furthermore, because S288c, from which the reference set of genes are defined, has long been exposed to a relatively benign and unvarying laboratory environment, the possibility arises that genes important only for survival in non-laboratory conditions may have been lost. In addition, the isolation of yeast strains from disparate environments suggests that the local ecology with which they interact is also unique and that horizontal gene transfer (HGT) from other species within their local environment may introduce new genes. Due to their potential importance in understanding a strains phenotype, identifying coding regions not present in the reference strain is of great interest. To find non-reference genes, we performed a *de novo* genome assembly of our Illumina sequencing reads (Table A-4). We identified a total of 615 non-reference genes (Figure 2-6). Importantly, our methods did not discover any non-reference genes within S288c (the reference strain) or BY4716 (a strain derived from S288c) assemblies, suggesting a low false positive rate in our non-reference gene identification pipeline.

To determine the likelihood that the identified new genes are in fact new coding genes, we estimated the expression level of both reference genes and the newly identified non-reference genes within 23 strains of *S. cerevisiae* (SKELLY *et al.* 2013). We found that on average, 54.9% of the new non-reference genes had a higher expression level than at least 5% of reference genes (Figure 2-7), suggesting that the majority of the newly identified genes are likely to be protein coding regions. We expect this analysis to underestimate the number of new genes because gene expression levels were estimated in only a single environment.

To identify the origins of the non-reference genes, we used BLASTn and tBLASTx to look for significant homology with NCBI "nr" database entries. We found that the newly identified non-reference genes fall into one of three categories. The first category includes genes not contained in the S288c reference, but for which homologous sequences were found in a previously sequenced *S. cerevisiae* strains. For example up to 36 identified non-reference genes were found to have greatest sequence similarity with genes found in the wine strain EC1118. Genes in this category are consistent with a gene loss scenario, perhaps due to adaptation to a benign environment, as these genes are often found in multiple *S. cerevisiae* strains. The second category of non-reference genes are those whose sequence is homologous to genes in other fungi. Non-reference genes in this category are usually found in only a few, often unrelated, strains and are consistent with horizontal transfer rather than widespread loss. Finally, non-reference genes that show no apparent significant similarity to any sequences held in the NCBI database may represent *de novo* gene birth. However, expression data indicates that only 13 genes in this set are expressed at levels greater than 5% of all genes, suggesting many may not be real (Figure 2-9). Even if rare, however, such genes provide important insights into evolutionary innovation and warrant further investigation to clarify their identity and function.

Population genetics of premature stop codons

Previous analyses using a small number of strains have shown that genes found to be essential in the reference S288c genome are more likely to be maintained across the population (Liti *et al.* 2009b; Schacherer *et al.* 2009). However the finding that at least 5

genes contain premature stop codons across a small number of strains as well as the finding that 57 genes show differential essentiality between S288c and the closely related (~0.1% diverged) strain Σ1278b (DOWELL *et al.* 2010) suggests that genes may change essentiality across the population (which is 0.8%-1% sequence diverged). As a consequence, several genes may have accumulated pseudogenizing genetic variants.

We initially detected 46 genes listed as essential that contain an apparent premature stop codon caused by a SNP. However closer manual inspection revealed that the majority of these genes (37/46) are likely incorrectly termed as essential because they overlap the coding region or promoter region of a true essential gene - the essential phenotype observed is actually that for disruption of the gene it overlaps. Of the 9 remaining genes with premature stop codons, only 2 introduce stop codons at a position <90% of the reference genomes gene length and therefore are likely to affect protein function. These potentially interesting premature stop codons were found for *SPP381* (YBR152W) in three (YPS1000, YJM145, UWOPS87.2421) genomes and for and *PAM16* (YJL104W) one (273614N) genome. *SPP381* is involved in mRNA splicing whilst *PAM16* is a subunit of the PAM complex – a translocase component of the mitochondrial inner membrane. Perhaps equally as interesting is that the three strains carrying the *SPP381* nonsense mutation are not closely clustered phylogenetically suggesting independence. However, a closer inspection at the individual chromosomal phylogenies does suggest only two independent loss events due to the close clustering of YPS1000 and UWOPS87.2421 for chromosome 2 (Figure A-2).

Intron loss

Although there are only a limited number of genes with introns in the *S. cerevisiae* genome, their evolutionary history is a long-standing area of interest (ROGOZIN *et al.* 2003). To this end, we identified intron loss events in the 84 strains relative to the reference genome. We found 21 intron loss events, all of which occurred in mitochondrial genes (Figure 2-9). We found frequent loss of two introns from *COX1* (18 occurrences across 2 introns), an observation consistent with loss of *COX1* introns in other species (SANCHEZ-PUERTA *et al.* 2008). AI5_BETA, a verified gene of unknown function, but fully located within intron 6 of *COX1,* was found to have lost two introns, and Q0255, an uncharacterized ORF, was found to have lost a single intron. Surprisingly, however, there is no clear phylogenetic signal for these loses, suggesting multiple independent events. However, the mitochondrial and nuclear genomes are independent entities and their transmission and fixation after strain hybridizations can take different trajectories. Unfortunately, the AT-rich nature of mitochondrial genome resulted in poor mapping and low sequence coverage and the catalogued intron loss events may not be accurate enough for the estimation of evolutionary dynamics of introns, much less the mitochondria as a whole. Nevertheless, our observations provides support for frequent changes in gene structure of mitochondrial genes, presumably a result of pervasive HGT and gene conversion in mitochondrial genome (HAO *et al.* 2010).

Population wide elucidation of TATA box sequence

The TATA box is a common component of eukaryotic promoters involved in guiding the transcriptional machinery to the transcriptional start site. Recent studies have

demonstrated that while some genes are regulated by canonical TATA box sequences (TATA(A/T)A(A/T)(A/G)), other genes contain non-canonical TATA box sequences with 1-3 mismatches relative to the canonical TATA sequence (RHEE and PUGH 2012). Canonical TATA boxes are present in ~20% of yeast genes (BASEHOAR *et al.* 2004) and are associated with genes with variable expression, such as those involved in response to stress (TIROSH *et al.* 2006; LANDRY *et al.* 2007). We found 958 SNPs across the 85 strains within defined TATA box regions. To determine if there is evidence of selection acting on the TATA box sequence, we determined whether each SNP changed the state the TATA box between a canonical and non-canonical sequences (Figure 2-10A). We then performed chi-square tests at each position to compare the observed and expected number of SNPs that result in canonical/non-canonical state conversions. We found that in the eighth position there has been an excess of SNPs that maintain the TATA box state. Conversely we found that at the fifth and seventh positions there has been an excess of SNPs that change the TATA box state. One possible explanation for the pattern of excess state switching SNPs is that they were selectively favored because of their contribution to fitness enhancing expression divergences. Supporting this possibility, we found a clear association between gene expression level in the laboratory strain S288c (NAGALAKSHMI *et al.* 2008) and the number of mismatches each genes TATA box region contains relative to the canonical sequence (Figure 2-10B). To investigate this possibility further, we compared the average gene expression level of genes with and without canonical TATA sequences for 23 strains (SKELLY *et al.* 2013). However, no difference in expression was apparent (Wilcoxon signed-rank test, $p > 0.05$ for all eight positions,

Figure 2-10C, Table 2-2). Thus, while there is evidence of selection within the TATA box sequence, the functional consequences of these changes are currently unclear.

Aneuploidies and large scale duplications

Recently, aneuploidies and other large scale duplications have risen to prominence as a mutation type commonly identified in both natural and experimental populations of yeast. They can have a profound effect on strain phenotypes and allow for rapid adaptation to novel environments due to their relatively higher frequency of occurrence compared to individual point mutations and their ability to simultaneously duplicate multiple genes (GRESHAM *et al.* 2008, 2010; PAYEN *et al.* 2013; SUNSHINE *et al.* 2015). Across our 85 sequenced genomes, we found that 23 strains contained whole chromosome aneuploidies, with several strains containing multiple chromosomal duplications and losses (Figure 2-11). A total of 26 chromosome copy number changes, both gains and losses, a rate of approximately 1.9% of all chromosomes sequenced (26/1344).

We also found several strains containing large duplications of chromosomal regions. We identified a single amplification of the left arm of chromosome 10 in YJM326. In addition, three strains (DBVPG6040, DBVPG1399 and T73) contained an amplification of the left arm of chromosome 16. Although T73 and DBVPG1399 are closely related strains, DBVPG6040 is quite diverged. The possibility exists that a recent introgression between these lineages is a possible explanation for the presence of the same duplication in the distinct strains. However, there is no evidence of a recent introgression between these strains for this chromosome from the structure data. Another possible explanation is

that because these strains are all isolated from wine making/fermentation environments, this duplication is adaptive, and additional work is needed to understand the functional significance of this duplication.

Strain phenotyping and detection of underlying genetic variants

Bar-seq offers a simple method by which the relative fitness of thousands of strains can be measured simultaneously. To demonstrate the utility of our strain collection, we exposed a population consisting of 76 barcoded strains combined with the *S. cerevisiae* gene deletion collection to seven distinct environmental challenges (YPD , YPD at 40°C, YPD + 1.25M NaCl, YPD + 7% Ethanol, YPD + 4 mM paraquat, YPD + 3 mM hydrogen peroxide and YPD + 1mM Cobalt chloride). The analysis revealed high levels of diversity within the strain panel, demonstrating their potential utility for genotype-phenotype investigations (Table A-6).

Due to the strong population structure within *S. cerevisiae*, genome-wide association studies within are expected to be difficult. However, recent studies suggest that with appropriate controls, detecting associations in yeast should be possible (CONNELLY and AKEY 2012). The increased resolution of our genome sequence data, combined with a significantly expanded number of genotypes, gave us hope that we would be able to uncover potentially interesting genetic associations. Using a multi-stage approach, we identified SNPs associated with improved performance in 5 of the 6 conditions relative to fitness in YPD after controlling for population structure. We identified between 3 and 19 SNPs significantly associated with improved growth in each condition (Table A-5).

Because of linkage, the SNP detected in the association is unlikely to be the causal SNP. Although we have no direct evidence that the SNPs, or indeed linked regions, are responsible for the phenotypes observed, a closer inspection of the regions surrounding the SNPs does suggest that the signals are real. For example we detected 13 SNPs associated with fitness at high temperatures (40°C). 5 of these 13 SNPs map to a ~16kb region on chromosome 11 which spans a region that contains the ribosomal gene *RPS21a,* a gene whose deletion is known to slow growth at high temperature. Similarly a second SNP on chromosome 11, located 51.5kb from this cluster, is within 7kb upstream and downstream of the genes *DBP7*, *RPC37* and *GCN3,* which again are known to reduce heat tolerance on deletion, as well as the genes *SET3* and *YKR023C*, genes that are known to reduce stress tolerance when deleted. It will be useful for future work to validate these associations and determine which gene or genes harbor functional variants.

Conclusions

The set of 85 genomes sequenced here significantly expands the set of high quality genomes available to the community. These sequences, combined with the genetically tractable haploids and diploids created, provides valuable tools to the community.

Using these genomes and strains, we find frequent incongruence between a phylogeny derived from whole genome SNP data and the inheritance of large scale genome features such as entire chromosome sequence, chromosome gains and losses, intron loss, the occurrence of new genes, and premature stop codons. These results suggest that continued effort is needed to understand the evolutionary history of such features. In

addition, we identify several regions associated with strain growth under environmental stress compared to benign conditions that are independent of population structure. This approach demonstrates the utility of expanding genomic and phenotypic analyses beyond standard laboratory strains and the unique biology this often entails.

References

BASEHOAR A. D., ZANTON S. J., PUGH B. F., 2004 Identification and distinct regulation of yeast TATA box-containing genes. Cell **116**: 699–709.

BERGSTRÖM A., SIMPSON J. T., SALINAS F., BARRÉ B., PARTS L., ZIA A., NGUYEN BA A. N., MOSES A. M., LOUIS E. J., MUSTONEN V., WARRINGER J., DURBIN R., LITI G., 2014 A High-Definition View of Functional Genetic Variation from Natural Yeast Genomes. Mol. Biol. Evol. **31**: 1–17.

BESEMER J., LOMSADZE A., BORODOVSKY M., 2001 GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. **29**: 2607–2618.

BREM R. B., YVERT G., CLINTON R., KRUGLYAK L., 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296**: 752–5.

CLIFTEN P. F., FULTON R. S., WILSON R. K., JOHNSTON M., 2006 After the duplication: gene loss and adaptation in Saccharomyces genomes. Genetics **172**: 863–72.

COGHLAN A., WOLFE K. H., 2000 Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. Yeast **16**: 1131–1145.

CONNELLY C. F., AKEY J. M., 2012 On the Prospects of Whole-genome Association Mapping in Saccharomyces cerevisiae. Genetics **191**: 1345–1353.

CROMIE G. a, HYMA K. E., LUDLOW C. L., GARMENDIA-TORRES C., GILBERT T. L., MAY P., HUANG A. a, DUDLEY A. M., FAY J. C., 2013 Genomic Sequence Diversity and Population Structure of Saccharomyces cerevisiae Assessed by RAD-seq. G3.

CUBILLOS F. a, LOUIS E. J., LITI G., 2009 Generation of a large set of genetically tractable haploid and diploid Saccharomyces strains. FEMS Yeast Res. **9**: 1217–25.

DANECK P., NELLAKER C., MCINTYRE R. E., BUENDIA-BUENDIA J. E., BUMPSTEAD S., PONTING C. P., FLINT J., DURBIN R., KEANE T. M., ADAMS D. J., 2012 High levels

of RNA-editing site conservation amongst 15 laboratory mouse strains. Genome Biol. **13**: R26.

DEUTSCHBAUER A. M., JARAMILLO D. F., PROCTOR M., KUMM J., HILLENMEYER M. E., DAVIS R. W., NISLOW C., GIAEVER G., 2005 Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics **169**: 1915–25.

DEVLIN B., ROEDER K., 1999 Genomic control for association studies. Biometrics **55**: 997–1004.

DIAO L., CHEN K. C., 2012 Local Ancestry Corrects for Population Structure in Saccharomyces cerevisiae Genome-Wide Association Studies. Genetics **192**: 1503–11.

DOUGLAS A. C., SMITH A. M., SHARIFPOOR S., YAN Z., DURBIC T., HEISLER L. E., LEE A. Y., RYAN O., GÖTTERT H., SURENDRA A., DYK D. VAN, GIAEVER G., BOONE C., NISLOW C., ANDREWS B. J., 2012 Functional analysis with a barcoder yeast gene overexpression system. G3 **2**: 1279–89.

DOWELL R. D., RYAN O., JANSEN A., CHEUNG D., AGARWALA S., DANFORD T., BERNSTEIN D. a, ROLFE P. A., HEISLER L. E., CHIN B., NISLOW C., GIAEVER G., PHILLIPS P. C., FINK G. R., GIFFORD D. K., BOONE C., 2010 Genotype to phenotype: a complex problem. Science **328**: 469.

DUJON B., SHERMAN D., FISCHER G., DURRENS P., CASAREGOLA S., LAFONTAINE I., MONTIGNY J. De, BLANCHIN S., BECKERICH J., BEYNE E., BLEYKASTEN C., BABOUR A., BOYER J., CATTOLICO L., CONFANIOLERI F., DARUVAR A. De, DESPONS L., FABRE E., 2004 Genome evolution in yeasts. Nature: 35–44.

ENGEL S. R., DIETRICH F. S., FISK D. G., BINKLEY G., BALAKRISHNAN R., COSTANZO M. C., DWIGHT S. S., HITZ B. C., KARRA K., NASH R. S., WENG S., WONG E. D., LLOYD P., SKRZYPEK M. S., MIYASATO S. R., SIMISON M., CHERRY J. M., 2013 The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. G3: 389–398.

EVANNO G., REGNAUT S., GOUDET J., 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. **14**: 2611–20.

FAY J. C., WU C. I., 2000 Hitchhiking under positive Darwinian selection. Genetics **155**: 1405–13.

FU Y. X., LI W. H., 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

GIAEVER G., CHU A. M., NI L., CONNELLY C., RILES L., VÉRONNEAU S., DOW S., LUCAU-DANILA A., ANDERSON K., ANDRÉ B., ARKIN A. P., ASTROMOFF A., EL-BAKKOURY M., BANGHAM R., BENITO R., BRACHAT S., CAMPANARO S., CURTISS M., DAVIS K., DEUTSCHBAUER A., ENTIAN K.-D., FLAHERTY P., FOURY F., GARFINKEL D. J., GERSTEIN M., GOTTE D., GÜLDENER U., HEGEMANN J. H., HEMPEL S., HERMAN Z., JARAMILLO D. F., KELLY D. E., KELLY S. L., KÖTTER P., LABONTE D., LAMB D. C., LAN N., LIANG H., LIAO H., LIU L., LUO C., LUSSIER M., MAO R., MENARD P., OOI S. L., REVUELTA J. L., ROBERTS C. J., ROSE M., ROSS-MACDONALD P., SCHERENS B., SCHIMMACK G., SHAFER B., SHOEMAKER D. D., SOOKHAI-MAHADEO S., STORMS R. K., STRATHERN J. N., VALLE G., VOET M., VOLCKAERT G., WANG C., WARD T. R., WILHELMY J., WINZELER E. a, YANG Y., YEN G., YOUNGMAN E., YU K., BUSSEY H., BOEKE J. D., SNYDER M., PHILIPPSEN P., DAVIS R. W., JOHNSTON M., 2002 Functional profiling of the Saccharomyces cerevisiae genome. Nature **418**: 387–91.

GOLDSTEIN a L., MCCUSKER J. H., 1999 Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. Yeast **15**: 1541–53.

GRESHAM D., DESAI M. M., TUCKER C. M., JENQ H. T., PAI D. a, WARD A., DESEVO C. G., BOTSTEIN D., DUNHAM M. J., 2008 The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet. **4**: e1000303.

GRESHAM D., USAITE R., GERMANN S. M., LISBY M., BOTSTEIN D., REGENBERG B., 2010 Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus. PNAS **107**: 18551–6.

HAO W., RICHARDSON A. O., ZHENG Y., PALMER J. D., 2010 Gorgeous mosaic of mitochondrial genes created by horizontal transfer and gene conversion. PNAS **107**: 21576–81.

HARTL D., CLARK A., 2006 *Principles of Population Genetics, Fourth Edition*. Sinauer Associates, Inc.

HITTINGER C. T., 2013 Saccharomyces diversity and evolution: a budding model genus. Trends Genet. **29**: 309–17.

HO C. H., MAGTANONG L., BARKER S. L., GRESHAM D., NISHIMURA S., NATARAJAN P., KOH J. L. Y., PORTER J., GRAY C. a, ANDERSEN R. J., GIAEVER G., NISLOW C., ANDREWS B., BOTSTEIN D., GRAHAM T. R., YOSHIDA M., BOONE C., 2009 A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. Nat. Biotechnol. **27**: 369–77.

HOSE J., YONG C. M., SARDI M., WANG Z., NEWTON M. A., GASCH A. P., 2015 Dosage compensation can buffer copy-number variation in wild yeast. Elife **4**: 1–27.

KANG H. M., ZAITLEN N. a, WADE C. M., KIRBY A., HECKERMAN D., DALY M. J., ESKIN E., 2008 Efficient control of population structure in model organism association mapping. Genetics **178**: 1709–23.

KELLIS M., BIRREN B. W., LANDER E. S., 2004 Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428**: 617–624.

KOBOLDT D. C., ZHANG Q., LARSON D. E., SHEN D., MCLELLAN M. D., LIN L., MILLER C. a, MARDIS E. R., DING L., WILSON R. K., 2012 VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. **22**: 568–76.

LANDRY C. R., LEMOS B., RIFKIN S. A., DICKINSON W. J., HARTL D. L., 2007 Genetic properties influencing the evolvability of gene expression. Science **317**: 118–121.

LANGMEAD B., SALZBERG S. L., 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**: 357–359.

LANGMEAD B., TRAPNELL C., POP M., SALZBERG S. L., 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**: R25.

LI H., HANDSAKER B., WYSOKER A., FENNELL T., RUAN J., HOMER N., MARTH G., ABECASIS G. R., DURBIN R., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

LIBKIND D., HITTINGER C. T., VALÉRIO E., GONÇALVES C., DOVER J., JOHNSTON M., GONÇALVES P., SAMPAIO J. P., VALERIO E., GONCALVES C., GONCALVES P., 2011 Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. PNAS **108**: 14539–14544.

LISTGARTEN J., LIPPERT C., HECKERMAN D., 2013 FaST-LMM-Select for addressing confounding from spatial structure and rare variants. Nat. Genet. **45**: 470–471.

LISTGARTEN J., LIPPERT C., KADIE C. M., DAVIDSON R. I., ESKIN E., HECKERMAN D., 2012 Improved linear mixed models for genome-wide association studies. Nat. Methods **9**: 525–6.

LITI G., CARTER D. M., MOSES A. M., WARRINGER J., PARTS L., JAMES S. a, DAVEY R. P., ROBERTS I. N., BURT A., KOUFOPANOU V., TSAI I. J., BERGMAN C. M., BENSASSON D., O'KELLY M. J. T., OUDENAARDEN A. VAN, BARTON D. B. H., BAILES E., NGUYEN A. N., JONES M., QUAIL M. a, GOODHEAD I., SIMS S., SMITH F., BLOMBERG A., DURBIN R., LOUIS E. J., 2009 Population genomics of domestic and wild yeasts. Nature **458**: 337–341.

LITI G., NGUYEN BA A. N., BLYTHE M., MÜLLER C. a, BERGSTRÖM A., CUBILLOS F. a, DAFHNIS-CALAS F., KHOSHRAFTAR S., MALLA S., MEHTA N., SIOW C. C., WARRINGER J., MOSES A. M., LOUIS E. J., NIEDUSZYNSKI C. a, 2013 High quality de novo sequencing and assembly of the Saccharomyces arboricolus genome. BMC Genomics **14**: 69.

LUO R., LIU B., XIE Y., LI Z., HUANG W., YUAN J., HE G., CHEN Y., PAN Q., LIU Y., TANG J., WU G., ZHANG H., SHI Y., LIU Y., YU C., WANG B., LU Y., HAN C., CHEUNG D. W., YIU S.-M., PENG S., XIAOQIAN Z., LIU G., LIAO X., LI Y., YANG H., WANG J., LAM T.-W., WANG J., 2012 SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience **1**: 18.

MARTIN M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: pp. 10–12.

MESSER P. W., PETROV D. a, 2013 Frequent adaptation and the McDonald-Kreitman test. PNAS **2013**.

NAGALAKSHMI U., WANG Z., WAERN K., SHOU C., RAHA D., GERSTEIN M., SNYDER M., 2008 The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320**: 1344–1349.

NOTREDAME C., HIGGINS D. G., HERINGA J., 2000 T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**: 205–217.

PAYEN C., RIENZI S. C. DI, ONG G. T., POGACHAR J. L., SANCHEZ J. C., SUNSHINE A. B., RAGHURAMAN M. K., BREWER B. J., DUNHAM M. J., 2013 The Dynamics of Diverse Segmental Amplifications in Populations of Saccharomyces cerevisiae Adapting to Strong Selection. G3: 399–409.

PIERCE S. E., FUNG E. L., JARAMILLO D. F., CHU A. M., DAVIS R. W., NISLOW C., GIAEVER G., 2006 A unique and universal molecular barcode array. Nat. Methods **3**: 601–3.

PRITCHARD J. K., STEPHENS M., DONNELLY P., 2000 Inference of population structure using multilocus genotype data. Genetics **155**: 945–59.

QIAN W., MA D., XIAO C., WANG Z., ZHANG J., 2012a The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. Cell Rep. **2**: 1399–410.

QIAN W., YANG J.-R., PEARSON N. M., MACLEAN C., ZHANG J., 2012b Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet. **8**: e1002603.

RHEE H. S., PUGH B. F., 2012 Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature **483**: 295–301.

ROBINSON D. G., CHEN W., STOREY J. D., GRESHAM D., 2014 Design and Analysis of Bar-seq Experiments. G3 **4**: 11–8.

ROGOZIN I. B., WOLF Y. I., SOROKIN A. V, MIRKIN B. G., KOONIN E. V, 2003 Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr. Biol. **13**: 1512–7.

ROHLAND N., REICH D., 2012 Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. **22**: 939–46.

ROSENBERG N. A., PRITCHARD J. K., WEBER J. L., CANN H. M., KIDD K. K., ZHIVOTOVSKY L. A., FELDMAN M. W., 2002 Genetic structure of human populations. Science **298**: 2381–2385.

SANCHEZ-PUERTA M. V., CHO Y., MOWER J. P., ALVERSON A. J., PALMER J. D., 2008 Frequent, phylogenetically local horizontal transfer of the cox1 group I Intron in flowering plant mitochondria. Mol. Biol. Evol. **25**: 1762–77.

SCANNELL D. R., ZILL O. A., ROKAS A., PAYEN C., DUNHAM M. J., EISEN M. B., RINE J., JOHNSTON M., HITTINGER C. T., 2011 The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 **1**: 11–25.

SCHACHERER J., SHAPIRO J. a, RUDERFER D. M., KRUGLYAK L., 2009 Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. Nature **458**: 342–5.

SHARP P. M., LI W. H., 1987 The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**: 1281–1295.

SKELLY D. a, MERRIHEW G. E., RIFFLE M., CONNELLY C. F., KERR E. O., JOHANSSON M., JASCHOB D., GRACZYK B., SHULMAN N. J., WAKEFIELD J., COOPER S. J., FIELDS S., NOBLE W. S., MULLER E. G. D., DAVIS T. N., DUNHAM M. J., MACCOSS M. J., AKEY J. M., 2013 Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Res. **23**: 1496–504.

SLATER G. S. C., BIRNEY E., 2005 Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6**: 31.

SMITH N. G. C., EYRE-WALKER A., 2002 Adaptive protein evolution in Drosophila. Nature **415**: 1022–1024.

SMITH A. M., HEISLER L. E., MELLOR J., KAPER F., THOMPSON M. J., CHEE M., ROTH F. P., GIAEVER G., NISLOW C., 2009 Quantitative phenotyping via deep barcode sequencing. Genome Res. **19**: 1836–42.

STROPE P. K., SKELLY D. a, KOZMIN S. G., MAHADEVAN G., STONE E. a, MAGWENE P. M., DIETRICH F. S., MCCUSKER J. H., CAROLINA N., SCIENCES B., CAROLINA N., 2015 The 100-genomes strains , an S . cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. : 1–13.

SUNSHINE A. B., PAYEN C., ONG G. T., LIACHKO I., TAN K. M., DUNHAM M. J., 2015 The Fitness Consequences of Aneuploidy Are Driven by Condition-Dependent Gene Effects. PLOS Biol. **13**: e1002155.

TAJIMA F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585.

TAMURA K., PETERSON D., PETERSON N., STECHER G., NEI M., KUMAR S., 2011 MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Bio. Evol. **28**: 2731–9.

TIROSH I., WEINBERGER A., CARMI M., BARKAI N., 2006 A genetic signature of interspecies variations in gene expression. Nat. Genet. **38**: 830–834.

TRAPNELL C., WILLIAMS B. A., PERTEA G., MORTAZAVI A., KWAN G., BAREN M. J. VAN, SALZBERG S. L., WOLD B. J., PACHTER L., 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**: 511–5.

TSAI I. J., BENSASSON D., BURT A., KOUFOPANOU V., 2008 Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle. PNAS **105**: 4957–62.

WACH A., BRACHAT A., PÖHLMANN R., PHILIPPSEN P., 1994 New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. Yeast **10**: 1793–808.

WANG Q.-M., LIU W.-Q., LITI G., WANG S.-A., BAI F.-Y., 2012 Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. Mol. Ecol. **21**: 5404–5417.

WARRINGER J., ZÖRGÖ E., CUBILLOS F. a F., ZIA A., GJUVSLAND A., SIMPSON J. T., FORSMARK A., DURBIN R., OMHOLT S. W., LOUIS E. J., LITI G., MOSES A., BLOMBERG A., 2011 Trait Variation in Yeast Is Defined by Population History. PLoS Genet. **7**: e1002111.

ZENG K., FU Y.-X., SHI S., WU C.-I., 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics **174**: 1431–1439.

ZÖRGÖ E., GJUVSLAND A., CUBILLOS F. F. a F., LOUIS E. J. E., ZORGO E., GJUVSLAND A., CUBILLOS F. F. a F., LOUIS E. J. E., ZÖRGÖ E., GJUVSLAND A., CUBILLOS F. F. a F., LOUIS E. J. E., LITI G., BLOMBERG A., OMHOLT S. W., WARRINGER J., 2012 Life history shapes trait heredity by promoting accumulation of loss-of-function alleles in yeast. Mol. Biol. Evol. **29**: 1781–9.

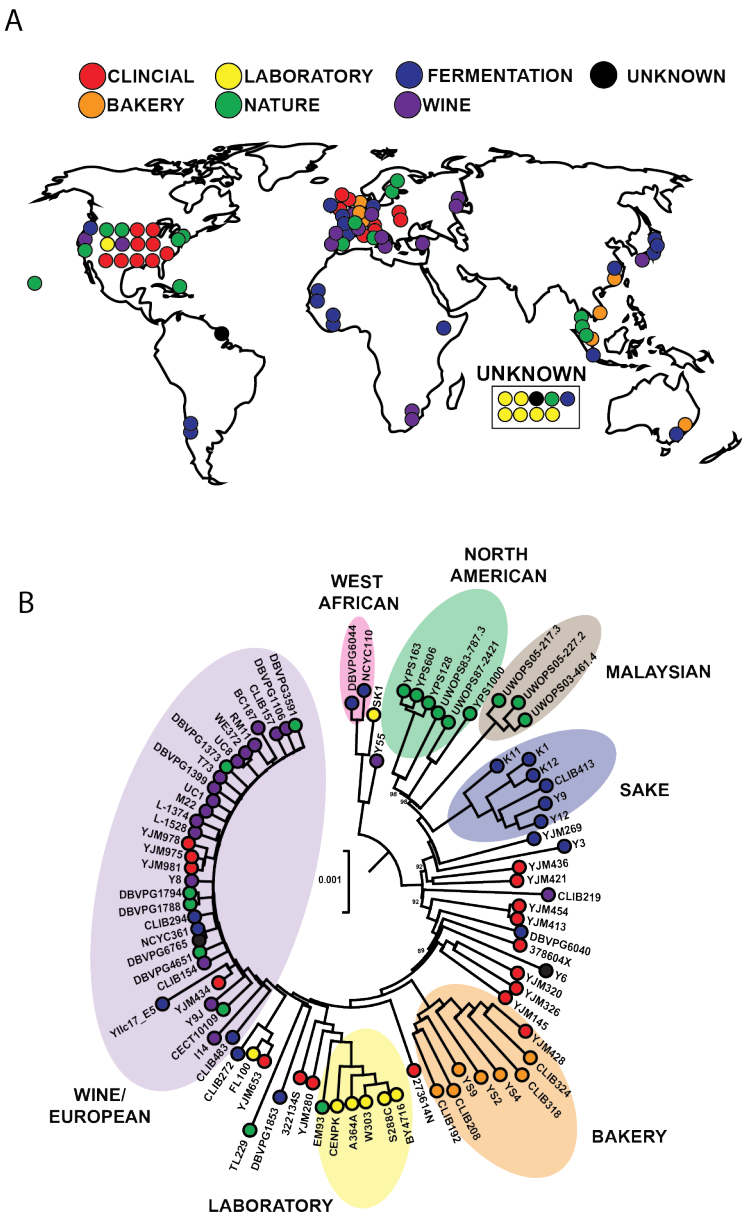


Figure 2-1 Geographical, environmental, and phylogenetic relationships

A. World map indicating geographic location of strain isolation if known. Colors represent the strains environment of isolation if known. B. Maximum composite likelihood neighbor-joining tree of all 85 sequenced genotypes. The approximate environment type from which each strain was isolated is indicated if known along with the clades into which the strains have been assigned. Bootstrap values <100 are shown except within the wine/European clade in which node support was low.

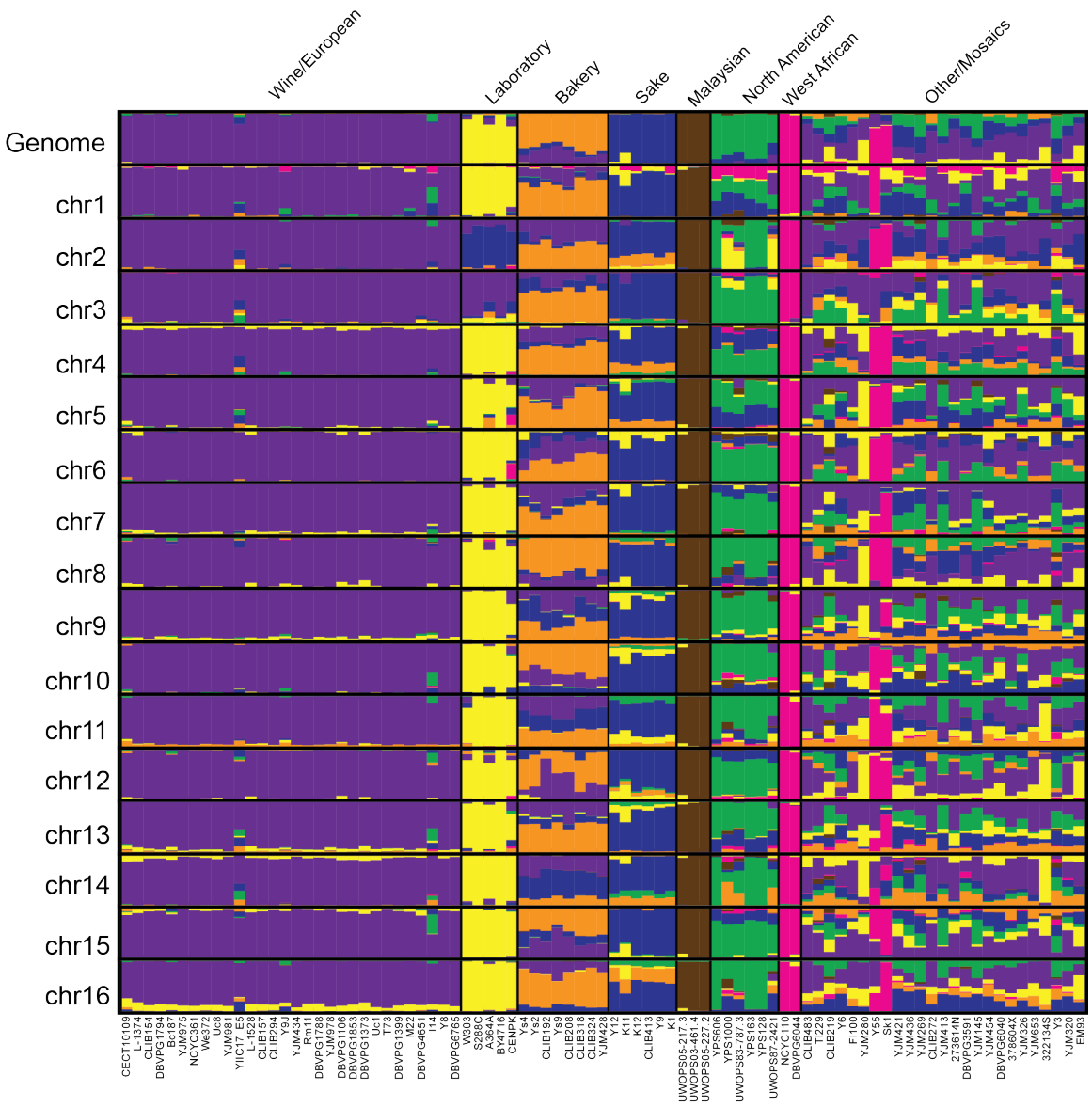


Figure 2-2 Structure of whole genome and individual chromosome relationships

Genome wide population structure of the 85 *S. cerevisiae* genomes was inferred using a randomly selected set of ~31,000 SNPs (10% of total SNPSs - see methods). Population structure assuming k = 7, found to be the best, is shown. 100,000 rounds of burn-in were followed by 100,000 of MCMC. Identified clades: Wine/European (purple); Laboratory (yellow); Bakery (orange); Sake (blue); Malaysian (brown); North American (green); West African (pink). Strains are arrayed based on clade membership derived from the phylogeny (top). Individual chromosome analyses were performed using k = 7 and all SNPs on that chromosome.

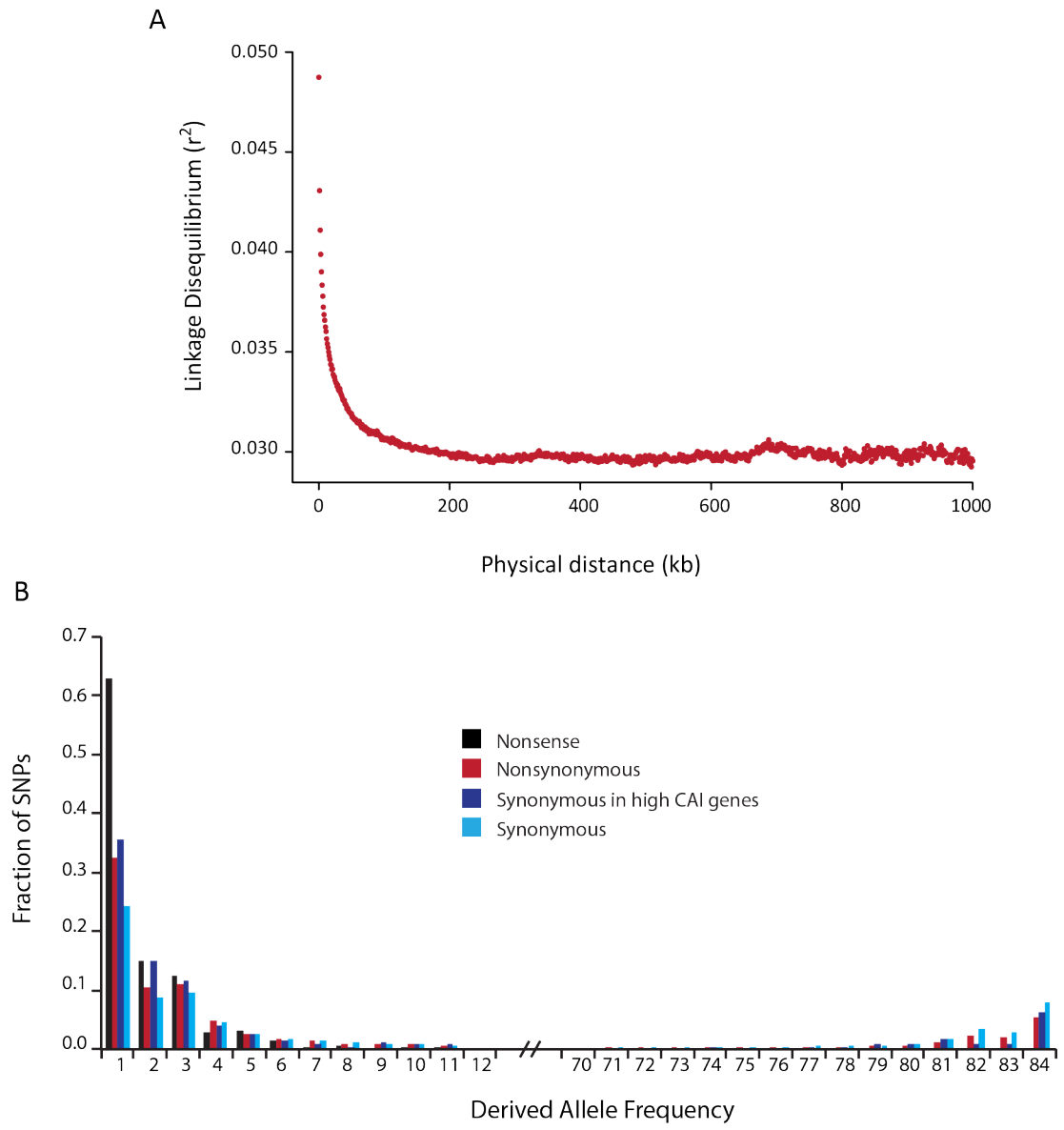


Figure 2-3 Linkage Disequilibrium and Derived Allele Frequency

A. Linkage disequilibrium (LD) decreases as physical distance increases. Each dot represents the average correlation coefficient (r^2) among pairs of SNPs within a 1kb window. B. Frequency spectrum of different SNP classes: synonymous (light blue); synonymous in high CAI genes (dark blue); nonsynonymous (red); nonsense (black). Only SNPs with a derived allele frequency (DAF) less than 12 or greater than 70 are plotted. Compared to synonymous SNPs, other SNP classes show an excess of low-frequency and deficiency of high-frequency SNPs, consistent with greater purifying selection against such variants.

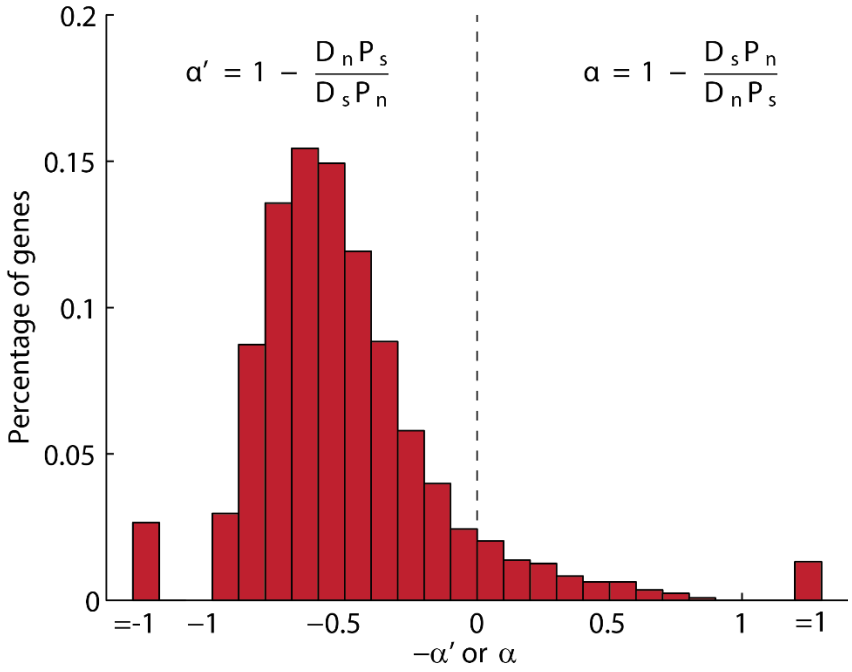


Figure 2-4 Distribution of α

Distribution of the proportion of amino-acid substitution driven by either positive selection (α) or maintained by negative selection (α') among all genes between *S. cerevisiae* and *S. paradoxus*. α' is plotted with a negative sign. P_s: numbers of synonymous polymorphisms; P_n: numbers of nonsynonymous polymorphisms; D_s: numbers of synonymous substitutions between *S. cerevisiae* and *S. paradoxus*; D_n: numbers of nonsynonymous substitutions between *S. cerevisiae* and *S. paradoxus*.

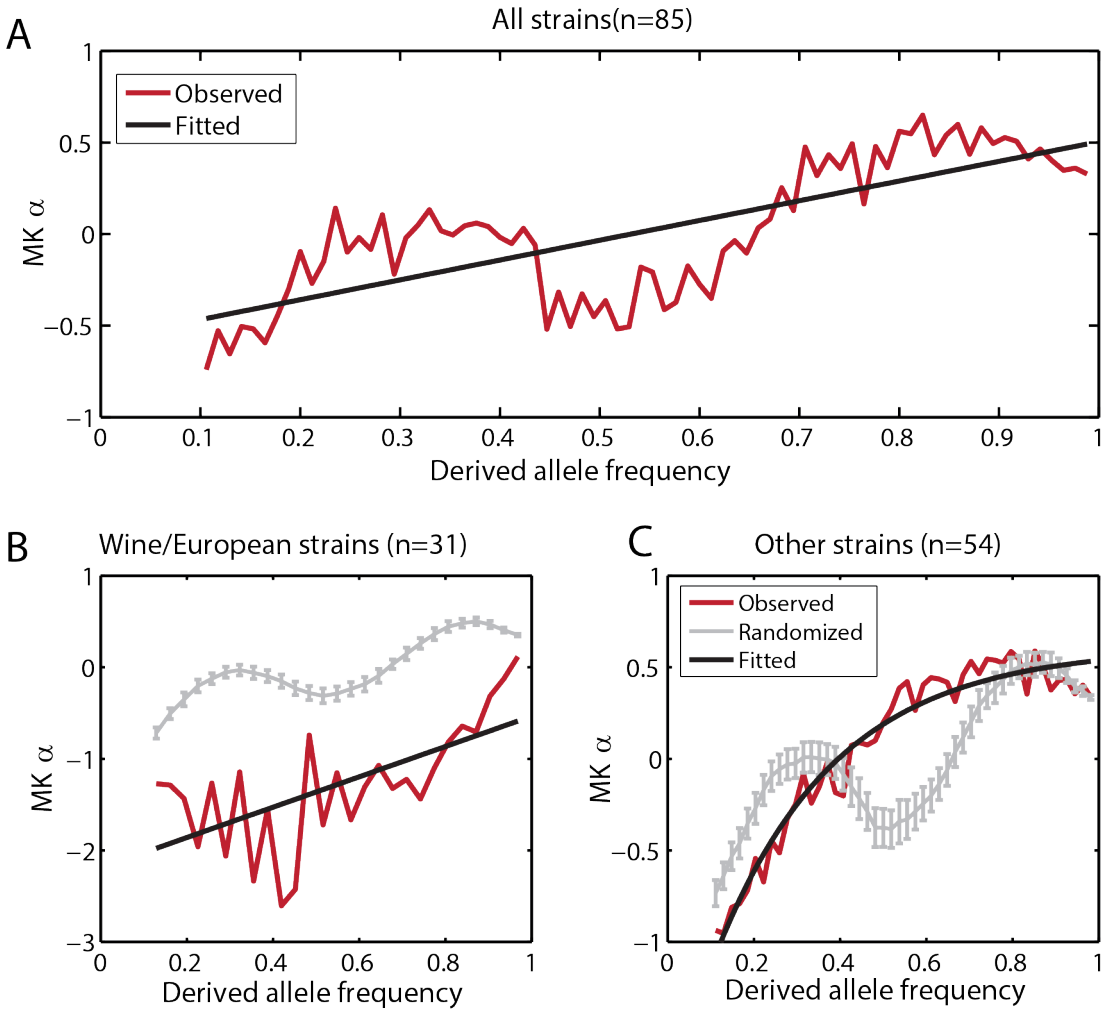


Figure 2-5 Asymptotic value of α

A. Values of α calculated using SNPs with increasing derived allele frequency for all strains. Red: observed values; Black: values fitted to exponential curve. B. Same as A using only the Wine/European strains. Gray: repeated sampling of 31 strains from all strains indicating that observed pattern is not due to sampling. C. Same as B using non-Wine/European strains. Gray: repeated sampling of 54 strains from all strains indicating the observed pattern is not due to sampling.

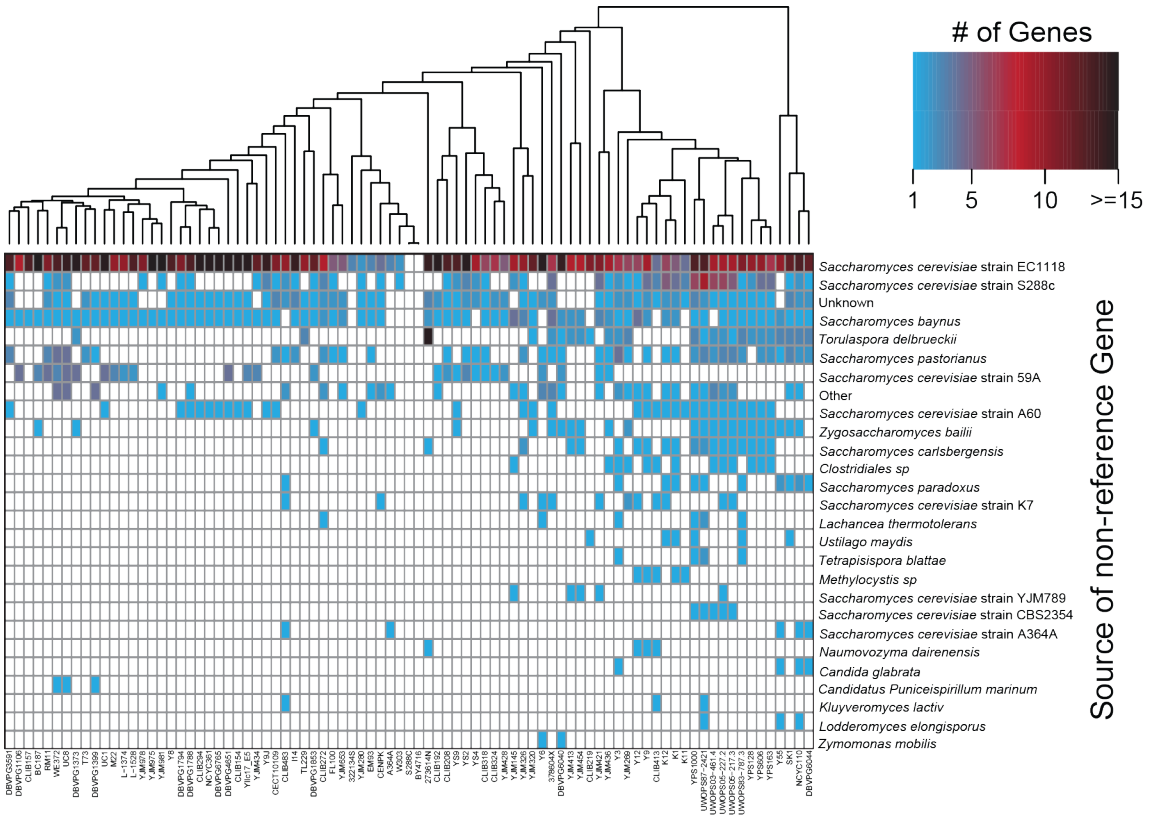


Figure 2-6 Source of non-reference genes

Distribution of non-reference genes across the 85 sequenced strains. The phylogeny is rooted as in Fig. 1B. Color indicates the number of genes detected within each genome that were identified from a particular source

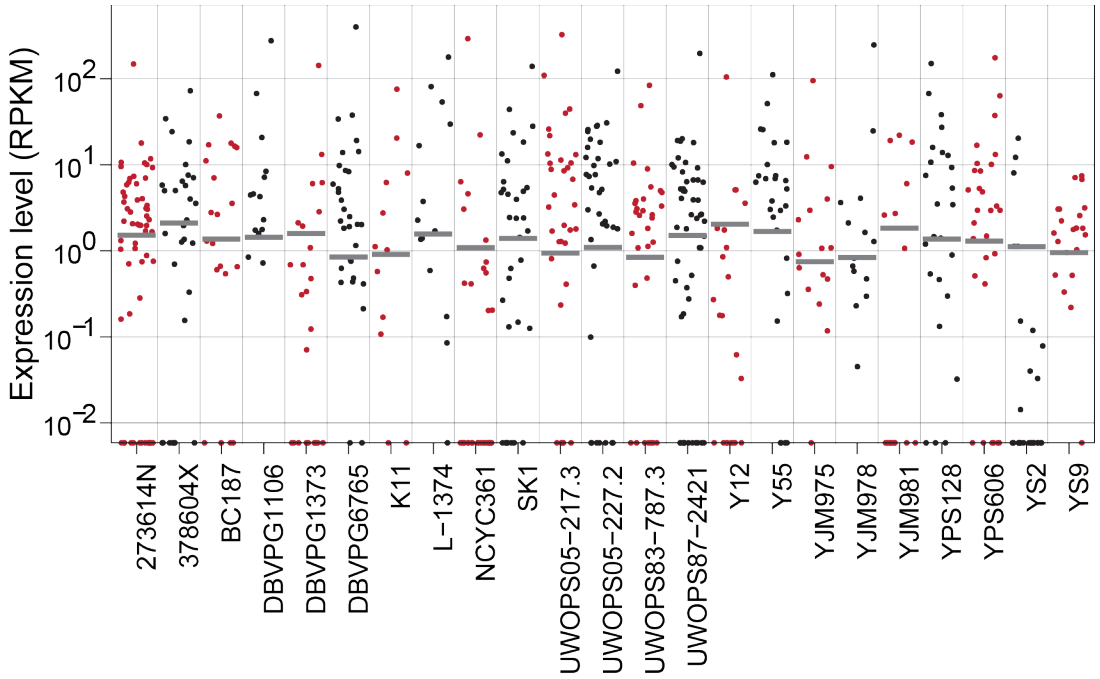


Figure 2-7 Expression levels of non-reference Genes

Gene expression levels determined by RNA-seq for 23 of the 85 sequenced strains. Reads were mapped to reference and non-reference gene sequences to determine their relative expression levels (RPKM). Gray bars represent the 5th percentile for reference genes within that strain.

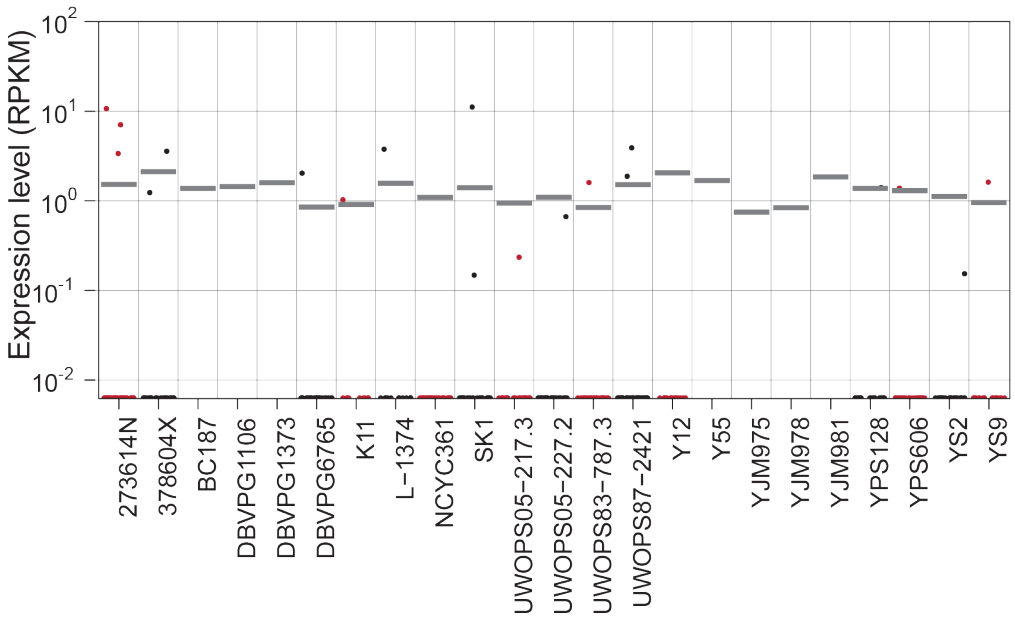


Figure 2-8 Expression levels of putative *de novo* genes

Gene expression levels determined by RNA-seq for 23 of the 85 sequenced strains. Reads were mapped to reference and putative *de novo* gene sequences to determine their relative expression levels (RPKM). Gray bars represent the 5th percentile for reference genes within that strain.

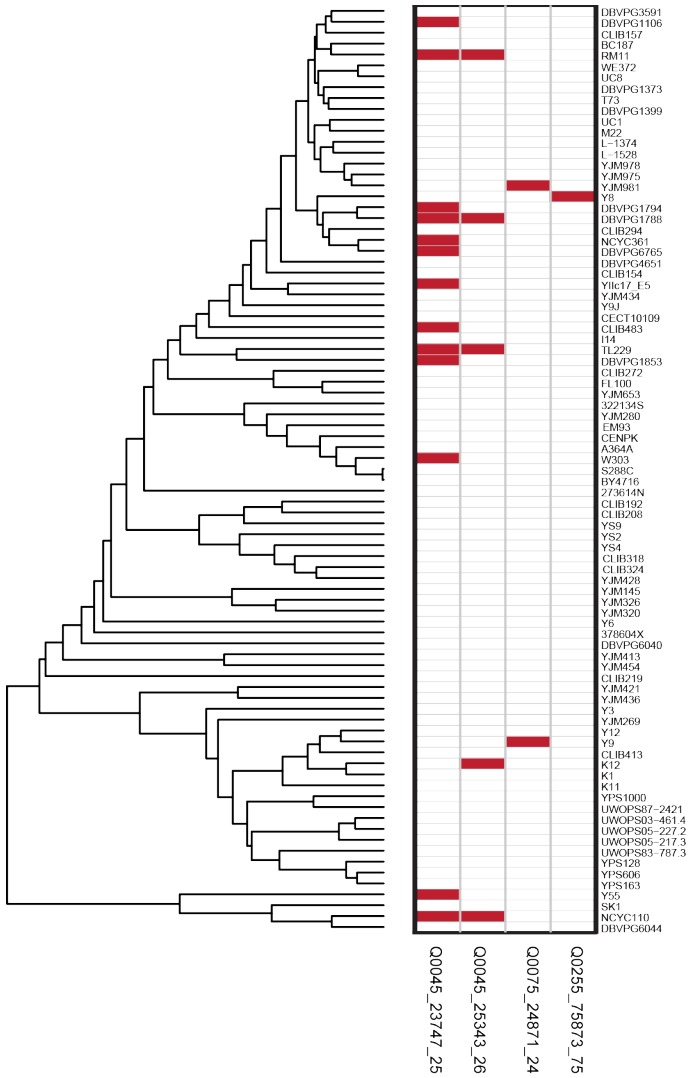


Figure 2-9 Intron loss in mitochondrial genes

Intron loss events, indicated by a red box, occurred 21 times across 4 mitochondrial genes within our population of 85 strains. There is no clear phylogenetic relationships amongst intron losses, suggesting either frequent independent intron loss or distinct evolutionary histories for the mitochondria and the nuclear genome.

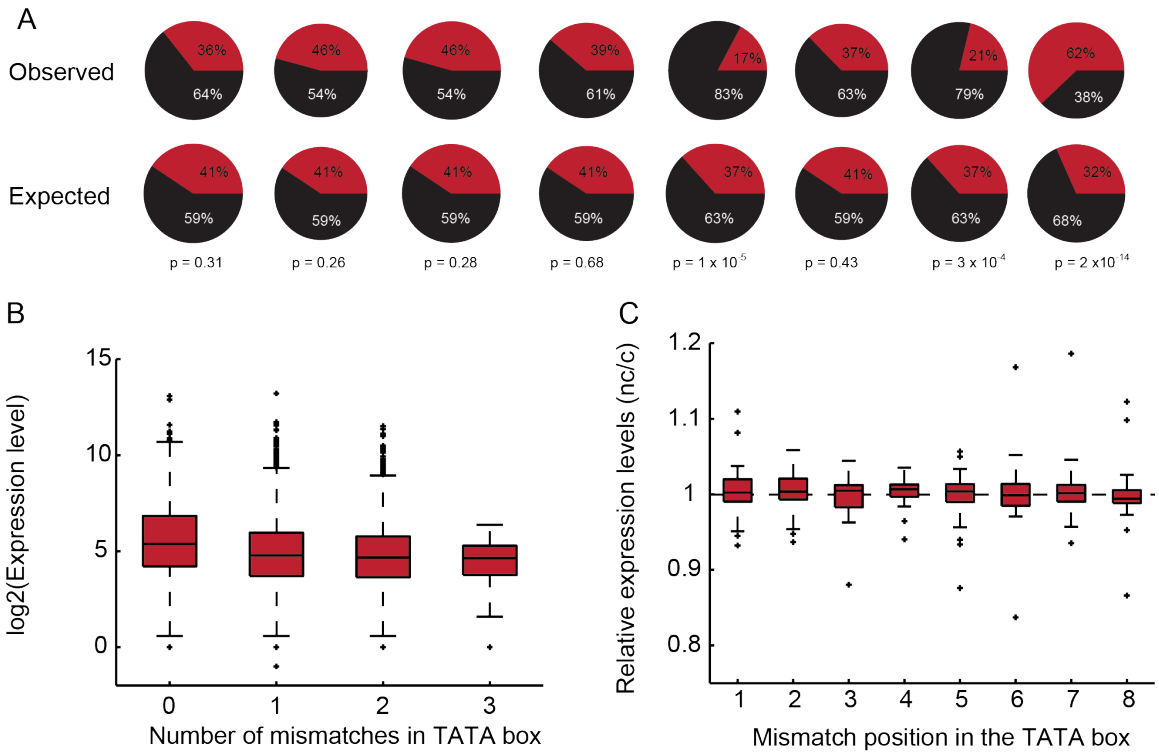


Figure 2-10 TATA box sequence conservation and expression

A. Observed and expected frequencies of SNPs changing (red) and not changing (black) the TATA box between canonical and non-canonical sequence for positions one through eight. Frequencies are listed within each pie. P-values below each paired set of pie charts result from chi-square tests comparing observed and expected base substitution frequencies. B. Expression level vs. number of mismatches in the TATA box sequence with zero mismatches representing canonical TATA box sequences. As the number of mismatches relative to the canonical TATA box sequence increases the gene expression level decreases. C. Position of TATA box mismatch from canonical sequence has no clear effect on gene expression levels. Only the effects of state changing SNPs at each of the 8 positions are assessed. If the canonical TATA box sequence increases gene expression level, then the (nc/c) expression ratio should be smaller than one. No position shows a significantly smaller or larger expression ratio (Wilcoxon signed-rank test).

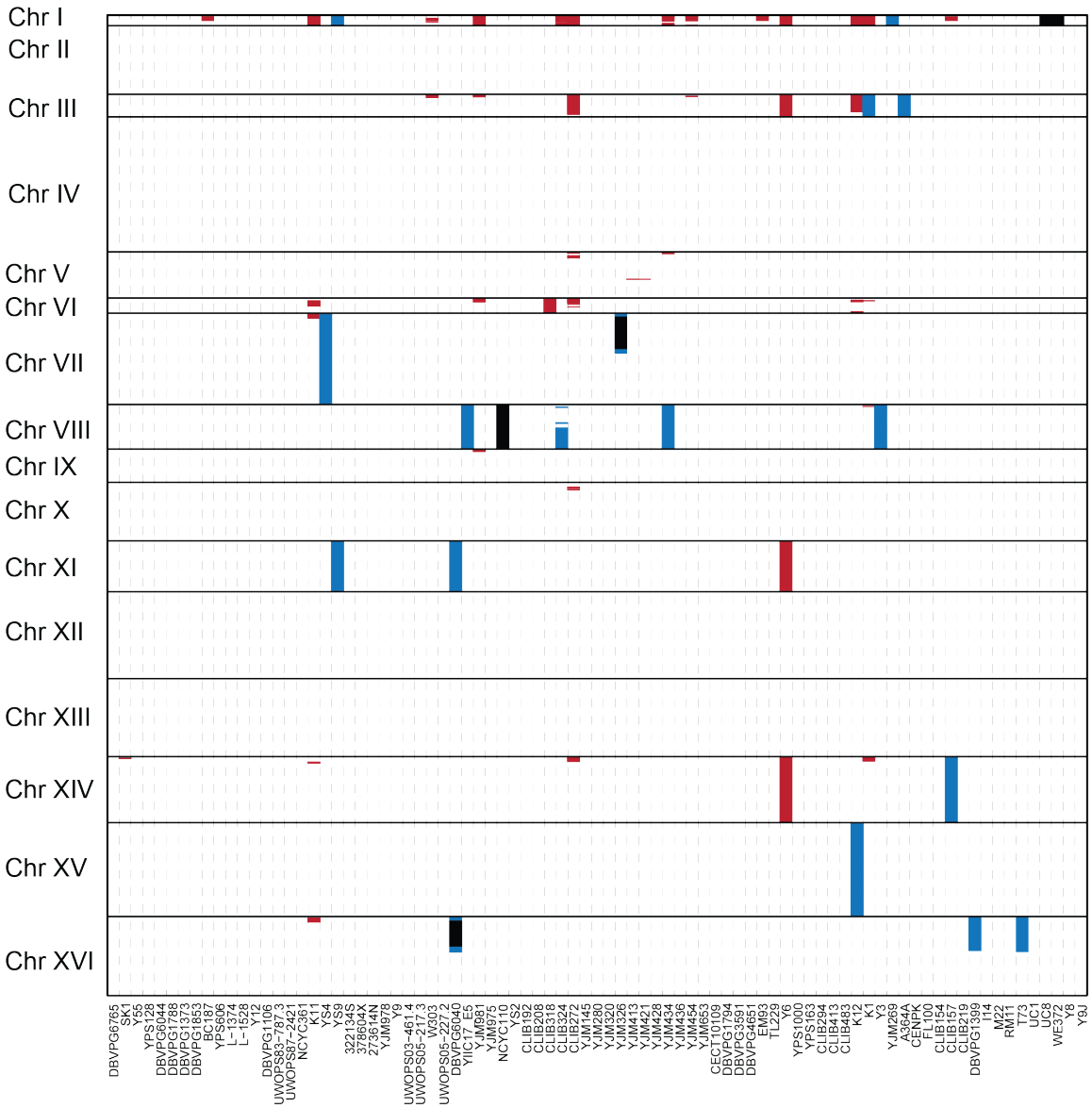


Figure 2-11 Chromosome Gains and Losses

Heatmap of sequence coverage across each chromosome for each strain. Coverage was determined by sliding windows across individual genes within each chromosome and compared to coverage of the same region in the reference strain S288c. White: ~100% of reference coverage; Red: ~50% of reference coverage; Blue: ~150% of reference coverage; Black: ~200% of reference coverage. Both whole chromosome and partial chromosome gains and losses were detected.

	n	*D*	*F*	*H*
Synonymous	99,786	-2.21	-0.68	-2.98
Nonsynonymous	76,100	-2.64	-3.53	-1.00
Nonsense	355	-3.06	-6.38	-0.00
Total	176,241	-2.39	-1.92	-2.19

Table 2-1 Population genetics statistics for different SNP classes

N: Number of SNPs in each category; D: Tajima's D; F: Fu and Li's D; H: Fay and Wu's H

	Altered	Unaltered	*p*-value *vs* expected frequency
High expression genes	201	114	0.39
Low expression genes	198	117	0.61
Expected frequency	0.614	0.386	

Table 2-2 SNPs altering canonical/non-canonical status of the TATA box

Altered: SNPs changing TATA box sequence between canonical and non-canonical states. Unaltered: SNPs not changing TATA box sequence between canonical and non-canonical states

Chapter 3

Disentangling the effects of mutation and selection on the evolution of *TDH3 cis*-regulatory variation[1]

Abstract

Genetic variation segregating within a species reflects the combined activities of mutation, selection, and genetic drift. In the absence of selection, polymorphisms are expected to be a random subset of new mutations; thus, comparing the effects of polymorphisms and new mutations provides a test for selection (DENVER *et al.* 2005; STOLTZFUS and YAMPOLSKY 2009; RICE and TOWNSEND 2012; SMITH *et al.* 2013). When evidence of selection exists, such comparisons can identify properties of mutations that are most likely to persist in natural populations (DENVER *et al.* 2005). Here, we investigate how mutation and selection have shaped variation in a *cis*-regulatory sequence controlling gene expression by empirically determining the effects of polymorphisms segregating in the *TDH3* promoter among 85 strains of *Saccharomyces cerevisiae* and comparing their effects to a distribution of mutational effects defined by 236 point mutations in the same promoter. Surprisingly, we find that selection on expression noise (i.e., variability in expression among genetically identical cells (RASER

[1]This chapter is published as: METZGER B. P. H.*, YUAN D. C.*, GRUBER J. D., DUVEAU F. D., WITTKOPP P. J., Selection on noise constrains variation in a eukaryotic promoter. Nature. **521**: 344–347. *Equal Contributions

and O'SHEA 2004)) appears to have had a greater impact on sequence variation in the *TDH3* promoter than selection on mean expression level. This is not necessarily because variation in expression noise impacts fitness more than variation in mean expression level, but rather because of differences in the distributions of mutational effects for these two phenotypes. This study shows how systematically examining the effects of new mutations can enrich our understanding of evolutionary mechanisms and provides rare empirical evidence of selection acting on expression noise.

Main Text

The *TDH3* gene encodes a highly expressed enzyme involved in central glucose metabolism (MCALISTER and HOLLAND 1985). Deletion of this gene decreases fitness (PIERCE *et al.* 2007) and its overexpression alters phenotypes (RINGEL *et al.* 2013), suggesting that the promoter controlling its expression is subject to selection in the wild. To test this hypothesis, we sequenced a 678 bp region containing the *TDH3* promoter (P_{TDH3}) as well as the 999 bp coding sequence of *TDH3* in 85 strains of *S. cerevisiae* sampled from diverse environments (Supplementary Table 1). We observed 44 polymorphisms in P_{TDH3}: 35 single nucleotide polymorphisms (SNPs) at 33 different sites and 9 insertions or deletions (indels) ranging from 1 to 32 bp (Figure B-1a). This frequency of polymorphic sites was significantly lower than the frequency of synonymous polymorphisms within the *TDH3* coding sequence (p-value = 0.03, Fisher's Exact Test) and polymorphic sites were less conserved between species than non-polymorphic sites in the promoter (p-value = 5×10^{-5}, Wilcox Rank Sum Test), consistent with purifying selection acting on P_{TDH3}. To determine whether the polymorphisms

observed in P_{TDH3} contribute to *cis*-regulatory variation, we compared relative *cis*-regulatory activity between each of 48 strains and a common reference strain. We found significant differences in *cis*-regulatory activity among strains (Figure B-1b), and 97% of the heritable *cis*-regulatory variation could be explained by sequence variation within the *TDH3* promoter (see Methods). These differences in *cis*-regulation act together with differences in *trans*-regulation to produce variation in *TDH3* mRNA abundance observed among strains (Figure B-1b).

To quantify the effect of each individual polymorphism on *cis*-regulatory activity, we used parsimony to reconstruct the evolutionary relationships among the 27 P_{TDH3} haplotypes observed in the 85 strains of *S. cerevisiae* sampled. We then inferred the most likely ancestral state for these haplotypes using P_{TDH3} sequences from an additional 15 strains of *S. cerevisiae* and all known species in the *Saccharomyces sensu stricto* genus (Supplementary Table 1, Figure B-2a). Next, we measured *cis*-regulatory activity of P_{TDH3} for the inferred ancestral state, each observed haplotype, and both possible intermediates between all pairs of observed haplotypes that differed by two mutational steps. We did this by cloning each P_{TDH3} haplotype upstream of the coding sequence for a yellow fluorescent protein (YFP), integrating these reporter genes (P_{TDH3} *–YFP*) into the *S. cerevisiae* genome, and quantifying YFP fluorescence using flow cytometry (GRUBER *et al.* 2012). For each genotype, YFP fluorescence was measured in ~10,000 single cells from each of 9 biological replicate populations (Figure 3-1a). We used these data to estimate both mean expression level (μ, Figure 3-1b) and expression noise (σ/μ, Figure 3-1) of P_{TDH3}*–YFP* for each promoter haplotype as readouts of *cis*-regulatory activity. We

then inferred the effects of individual polymorphisms by comparing the phenotypes of ancestral and descendent haplotypes that differed by only a single sequence change.

To determine how the effects of P_{TDH3} polymorphisms compare to the effects of new mutations in this *cis*-regulatory element, we estimated the distribution of mutational effects by using site-directed mutagenesis to introduce 236 of the 241 possible G:C→A:T transitions individually into P_{TDH3} *–YFP* alleles and assayed their effects on *cis*-regulatory activity using flow cytometry as described above. We used G:C→A:T transitions to estimate the distribution of mutational effects because they were the most common type of SNP observed both in the *TDH3* promoter (Figure B-1a) and genome-wide among the 85 *S. cerevisiae* strains (LITI *et al.* 2009; SCHACHERER *et al.* 2009). They were also the most frequent type of spontaneous point mutation observed in mutation accumulation lines of *S. cerevisiae* (LYNCH *et al.* 2008). To determine whether the effects of these mutations were likely to be representative of the effects of all types of point mutations, we analyzed data from previously published studies that measured the effects of single mutations on *cis*-regulatory activity (PATWARDHAN *et al.* 2009, 2012; KWASNIESKI and MOGNO 2012; MELNIKOV *et al.* 2012). We found no significant difference between the effects of G:C→A:T transitions and other types of point mutations on *cis*-regulatory activity in any of these datasets (Figure B-3 a-m). Consistent with this observation, we found no significant difference between the effects of G→A and C→T mutations on P_{TDH3} activity (mean expression level: p-value = 0.73; expression noise: p-value = 0.52, two tailed t-test, Figure B-3 n, o). We also found no significant difference

between the effects of G:C→A:T and other types of polymorphisms (mean expression level: *p*-value = 0.91; expression noise: *p*-value = 0.90, two tailed t-test, Figure B-3 p,q).

Mutations with the largest effects on mean expression level and expression noise were located within experimentally-validated transcription factor binding sites (TFBS) (BAKER *et al.* 1992; YAGI *et al.* 1994) (Figure 3-2). All of these mutations decreased mean expression level and increased expression noise. Outside of the known TFBS, 50% of the 218 mutations tested increased mean expression level and 87% increased expression noise. Despite this difference in the shape of the distributions, a negative correlation was observed between mean expression level and expression noise ($R^2 = 0.85$, Figure B-4) that was similar to previous reports for other yeast promoters (HORNUNG *et al.* 2012). The strength of this correlation was reduced to $R^2 = 0.45$ when mutations in the known TFBS were excluded.

To take the mutational process into account when testing for evidence that selection has influenced variation in the *S. cerevisiae TDH3* promoter, we compared the distributions of effects for mutations and polymorphisms on both mean expression level (Figure 3-3a) and expression noise (Figure 3-3b). We did this by randomly sampling sets of variants from the mutational distribution and comparing their effects to those observed among the naturally occurring polymorphisms. We found that the effects of observed polymorphisms on mean expression level were consistent with random samples of mutations from the distribution of mutational effects (one-sided *p*-value = 0.89, Figure B-5a,i), whereas the effects of observed polymorphisms on expression noise were not (one-

sided *p*-value = 0.0092, Figure B-5b). Specifically, polymorphisms were less likely to increase expression noise than random mutations (Figure B-5j), suggesting that selection has preferentially retained mutations that minimize expression noise from P_{TDH3} in natural populations. These results were robust to the exclusion of the large effect mutations in known TFBS from the distribution of mutational effects and the restriction of polymorphisms to G:C→A:T changes (Figure B-5c-f,k-n), the metric used to quantify expression noise (Figure B-6), and differences in genetic background that include a change in ploidy from haploid to diploid (Figure B-7).

The probability that a new mutation with a particular phenotypic effect survives within a species to be sampled as a polymorphism is related to its effect on relative fitness. The function describing relative fitness for different phenotypes can therefore be inferred by comparing the distribution of effects for new mutations to the distribution of effects for polymorphisms (Figure 3-3c,d). For mean expression level, we found that the most likely fitness function (Figure 3c) did not explain the data significantly better than a uniform fitness function representing neutral evolution (p-value = 0.87). For expression noise, we rejected a model of neutral evolution (p-value = 0.00019) and found that the most likely fitness function included higher fitness for variants that decreased gene expression noise (Figure 3-3d). Repeating this analysis using alternative metrics for expression noise produced comparable results (Figure B-6). These data suggest an evolutionary model in which purifying selection preferentially removes variants that increase expression noise, resulting in robust expression of *TDH3* among genetically identical individuals.

Consistent with this model, polymorphisms with the largest effects on expression noise

(but not mean expression level) were found at the lowest frequencies within the sampled strains of *S. cerevisiae* (mean, *p*-value = 0.43; noise *p*-value = 0.0029; permutation test, Figure B-2b-c). However, this pattern could also result from population structure among the sampled strains. To separate the effects of selection and population structure, we used the structure of the inferred haplotype network and the distribution of mutational effects to simulate neutral trajectories for *cis*-regulatory phenotypes as they diverged from the P_{TDH3} ancestral state. We then compared these trajectories to the phenotypic changes observed among naturally occurring haplotypes and their inferred intermediates for both mean expression level (Figure 3-3e) and expression noise (Figure 3-3f). We found that the observed haplotypes were consistent with neutral expectations for mean expression level (one-sided *p*-value = 0.32, Figure B-5g), but were not consistent with this neutral model for expression noise (one-sided *p*-value < 0.0001, Figure B-5h), regardless of which metric was used to measure expression noise (Figure B-6). We again saw that naturally occurring haplotypes showed smaller changes in noise relative to the common ancestor than would be expected from the mutational process alone, implying persistent selection for low noise in P_{TDH3} activity in the wild.

Taken together, our data indicate that sequence variation in the *S. cerevisiae TDH3* promoter has been affected more by selection for low levels of noise than selection for a particular level of *cis*-regulatory activity. This is not because the mean level of *cis*-regulatory activity is less important than noise for fitness, but because of differences in the distributions of mutational effects for these two phenotypes. Indeed, theoretical work shows that selection for low levels of noise is most likely to occur for phenotypes that are subject to purifying selection (LEHNER 2008). Additional evidence suggesting that

selection can act on expression noise comes from genomic analyses (FRASER *et al.* 2004; NEWMAN *et al.* 2006; BATADA and HURST 2007; LEHNER 2008; ZHANG *et al.* 2009; WANG and ZHANG 2011) and from the conservation of "shadow enhancers" that appear to maintain robust expression in multicellular organisms (FRANKEL *et al.* 2010; PERRY *et al.* 2010). By investigating not only the survival of the fittest, but also the "arrival of the fittest" (DE VRIES 1905; FONTANA and BUSS 1994), our work shows how phenotypic diversity produced by the mutational process itself has inherent biases that can influence the course of regulatory evolution. By taking empirical measurements of these mutational biases into account, we identified an unexpected target of selection that impacts how a *cis*-regulatory element evolves.

Methods

Characterizing variation segregating in the *TDH3* promoter

Variation in the *TDH3* gene was determined for 85 natural isolates of *S. cerevisiae* (LITI *et al.* 2009; SCHACHERER *et al.* 2009) (Supplementary Table 1). Sequences were obtained from each strain by PCR and Sanger sequencing using DNA extracted from diploid cells. Strains heterozygous for the *TDH3* promoter were grown on GNA plates for 12 hours (5% dextrose, 3% Difco nutrient broth, 1% Oxoid yeast extract, 2% agar) and sporulated on potassium acetate plates (1% potassium acetate, 0.1% Oxoid yeast extract, 0.05% dextrose, 2% agar). Individual spores were isolated by tetrad dissection and haploid derivatives were sequenced to empirically determine the phase of the two *TDH3* promoter haplotypes. All reagents for growth of yeast cultures were purchased from Fisher unless otherwise noted. In all, the 678 bp promoter contained SNPs at 33 sites and

the 238 synonymous sites contained 22 SNPs. 5 non-synonymous changes were also observed among these 85 strains.

Inferring the ancestral sequence and constructing the haplotype network for P_{TDH3}

Promoter haplotypes (Supplementary Table 1, Figure B-2a) were initially aligned using Pro-Coffee (TALY *et al.* 2011), followed by re-alignment with PRANK (LÖYTYNOJA and GOLDMAN 2010) and manual adjustment around repetitive elements and indels (Supplementary File 1). The *TDH3* promoter sequences from all *Saccharomyces sensu stricto* species (LITI *et al.* 2009, 2013; SCANNELL *et al.* 2011; LIBKIND *et al.* 2011), as well as an additional 15 strain of *S. cerevisiae* known to be an outgroup to the 85 focal strains (WANG *et al.* 2012), were also determined by Sanger sequencing. These sequences were used to infer the ancestral state of the *TDH3* promoter for the 85 strains with both parsimony and maximum likelihood methods implemented in MEGA 6 (TAMURA *et al.* 2013); both methods gave identical results. TCS 2.1 (CLEMENT *et al.* 2000) was used to build a haplotype network for the *TDH3* promoter, with changes polarized based on the inferred ancestral state (Figure B-2a). One haplotype (HH in Supplementary Table 1) could not be confidently placed within the network and was excluded from our analysis. Sequence conservation for individual sites was determined using sequences from all seven *Saccharomyces sensu stricto* species using ConSurf (ASHKENAZY *et al.* 2010) and the phylogeny from a prior study (HITTINGER 2013). To reduce heterogeneity in plotting, conservation was averaged over a 20bp sliding window.

Measuring variation in *TDH3* mRNA levels and *cis*-regulatory activity

Constructing reference strains

TDH3 mRNA levels and *cis*-regulatory activity were measured using pyrosequencing, with relative allelic expression in F_1 hybrids providing a readout of relative *cis*-regulatory activity (WITTKOPP *et al.* 2004). This technique requires one or more sequence differences to compare relative gDNA or cDNA abundance between two strains or two alleles within the same strain (WITTKOPP 2011). We therefore constructed reference strains of both mating types that carried a copy of the *TDH3* gene with a single, synonymous mutation (T243G). These genotypes were constructed by inserting the *URA3* gene into the native *TDH3* coding region in strains BY4741 and BY4742 and then replacing *URA3* with the modified *TDH3* coding sequence using the lithium acetate method and selection on 5-FOA (GIETZ and WOODS 2006; GRUBER *et al.* 2012). To do this, 80 bp oligonucleotides, containing a synonymous mutation and homology to each side of the target site, were transformed into these strains. Successful transformants (strains YPW342 and YPW339, respectively) were confirmed by Sanger sequencing. Resistance markers for hygromycin B (*hphMX6*) and G418 (*kanMX4*) were then inserted into the *HO* locus of these strains (producing YPW360 and YPW361, respectively) and used to construct a diploid reference strain (YPW362). A *kanMX4* resistance marker was also successfully inserted into the *HO* locus of 63 of the 85 natural strains (LITI *et al.* 2009; SCHACHERER *et al.* 2009).

Biological samples for comparing expression and cis-regulatory activity

To construct hybrids suitable for measuring *cis*-regulatory activity of natural isolates relative to a reference strain, haploid cells from each of the 63 natural isolates with a *kanMX4* resistance marker (mating type **a**) were mixed with an equal number of haploid

cells from the reference strain YPW360 (mating type α) on YPD plates (2% dextrose, 1% Oxoid yeast extract, 2% Oxoid peptone, 2% agar). After 24 hours, cultures were streaked on YPD plates to obtain single colonies and then patched to YPD plates containing G418 and Hygromycin B to select for diploids. Four replicates of each hybrid were grown in 500 µl of YPD liquid media for 20 hours at 30°C in 2 ml 96-well plates with 3 mm glass beads, shaking at 250 rpm. Cultures were diluted to an OD600 of 0.1 and then grown for an additional 4 hours. Plates were centrifuged, and the YPD liquid was removed. Cultures were then placed in a dry ice/ethanol bath until frozen and stored at -80°C. To prepare samples for measuring total *TDH3* mRNA abundance in each natural isolate relative to a common reference strain, diploids for each of the 63 natural isolates were mixed with a similar number of diploid cells from strain YPW362 based on OD600 readings after the initial growth in YPD liquid. These co-cultures were incubated and processed as described above.

Preparing genomic DNA (gDNA) and cDNA for analysis

For each hybrid and co-culture sample, gDNA and RNA were sequentially extracted from a single lysate using a modified protocol of Promega's SV Total RNA Isolation System. After thawing cultures on ice for ~30 minutes, 175 µl of SV RNA lysis buffer (with β–mercaptoethanol), 350 µl of ddH_20 and 50 µl of 400 micron RNase free beads were added to each sample. Plates were vortexed until cell pellets were completely resuspended. The plates were then centrifuged and 175 µl of supernatant was mixed with 25 µl of RNase-free 95% ethanol and loaded onto a binding plate. To extract RNA, 100 µl of RNase-free 95% ethanol was added to the flow through and loaded onto a second binding plate. These plates were then washed twice with 500 µl of SV RNA wash

solution and allowed to dry. To extract DNA, the first binding plate was washed twice with 700 µl of cold 70% ethanol and allowed to dry. For both binding plates, 100 µl of ddH_20 was added to each well, the plate was incubated at room temperature for 7.5 minutes, and the elute was collected. RNA from each sample was converted to cDNA by mixing 5 µl of extracted RNA with 2 µl RNase free water, 1 µl DNase buffer, 1 µl RNasin Plus, and 1 µl DNase 1 and incubating at 37°C for 1 hour followed by 65C for 15 minutes. 3 µl of oligo dT ($T_{19}VN$) was added and cooled to 37°C over 35 minutes. 4 µl of First Strand Buffer, 2 µl dNTPs, 0.5 µl RNasin Plus, and 0.5 µl of SuperScript II were added and incubated for 1 hour. 30 µl of ddH_20 was then added to each sample.

Pyrosequencing data collection, quality control filtering, and normalization

Pyrosequencing was performed as described previously(WITTKOPP 2011) using a PSQ 96 pyrosequencing machine and Qiagen pyroMark Gold Q96 reagents for gDNA and cDNA samples for both hybrids and co-cultured diploids. 1 µl of cDNA or gDNA was used in each PCR reaction, with primers shown in Supplementary Table 2. A single PCR and pyrosequencing reaction was performed for each gDNA and cDNA sample from each of the four biological replicate hybrid and co-culture samples for each natural haplotype, for a total of eight pyrosequencing reactions using cDNA and eight pyrosequencing reactions using gDNA for each of the 48 strains (Supplementary Table 3).

In gDNA samples from hybrids, the two *TDH3* alleles are expected to be equally abundant; however differences in PCR amplification of the two alleles (or aneuploidies altering copy number of *TDH3*) can cause unequal representation in the pyrosequencing data. Because such deviations cause estimates of relative allelic expression for these

samples to be less reliable, the 15% of samples with gDNA ratios that deviated by more than 15% from the expected 50:50 ratio were excluded. Relative abundance of the two *TDH3* alleles is expected to be more variable in the co-cultured samples because of unequal representation from differences in concentration of the two genotypes before mixing and/or after growth. Samples from co-cultured diploids with gDNA ratios in the upper or lower 10 percentile were also excluded from analysis. These quality control filters left 48 strains with at least two replicates in both the hybrid and co-cultured samples.

For each sample, relative allelic abundance in the cDNA sample was divided by relative allelic abundance for the corresponding gDNA sample to correct for remaining biases(WITTKOPP 2011). These ratios (Y_{ijk}) from strain *i*, plate *j*, and replicate *k* were fitted to the following linear model, including strain (ranging from 1-48) and plate (ranging from 1-3) as fixed effects as well as the cell density of the sample before and after growth from which the RNA and DNA were extracted (measured by OD600) as a covariate: $Y_{ijk} = \mu + Strain + Plate + Density.0 + Density.1 + \varepsilon$. An analysis of variance (ANOVA) found that strain, plate, and initial density were statistically significant for hybrids (Strain: *p*-value = $1.38\text{x}10^{-20}$; Plate: *p*-value = $1.01\text{x}10^{-10}$; Density.0: *p*-value = $5.01\text{x}10^{-3}$; ,Density.1: p-value = 0.740), and strain and plate were statistically significant for co-cultured diploids (Strain: *p*-value = $8.16\text{x}10^{-20}$; Plate: *p*-value = $2.65\text{x}10^{-3}$; Density.0: *p*-value = 0.734; ,Density.1: p-value = 0.833). Expression values for each sample were adjusted to remove the effects of plate and initial cell density. Differences in allelic abundance caused by the synonymous change introduced

for pyrosequencing were estimated by analyzing a hybrid between BY4741 and YPW360 and a co-culture of BY4741 and YPW362. The effects of this change were then subtracted from the $\log_2$-transformed expression ratio for all samples. Strains with significant *cis*-regulatory divergence from the reference were identified using *t*-tests. R code used for these analyses is provided in Supplementary File 2.

Estimating contribution variation in P_{TDH3} to cis-regulatory variation
To determine the amount of variation in *TDH3 cis*-regulatory activity explained by strain identity and the *TDH3* promoter haplotype, we fit the normalized expression values to linear models containing fixed effects of either strain identity or promoter haplotype alone. Variance among strains explained by strain identity was assumed to reflect heritable variation, with residual variance assumed to result from technical noise. Because multiple strains contained the same *TDH3* promoter haplotype, we were able to determine the proportion of this heritable variance explained by polymorphisms in the *TDH3* promoter region tested. 75% of all *cis*-regulatory variation and 97% of heritable *cis*-regulatory variation were explained by the *TDH3* promoter haplotype. To estimate the error associated with these estimates of variance explained, we analyzed 100,000 bootstrap replicates of the data with the same linear models.

Constructing strains with mutations and polymorphisms in P_{TDH3}
To efficiently assay *cis*-regulatory activity of the *TDH3* promoter, we used a P_{TDH3}-YFP reporter gene integrated near a pseudogene on chromosome 1 of strain BY4724 at position 199270(GRUBER *et al.* 2012). This P_{TDH3}-YFP transgene contains a 678bp sequence including the *TDH3* promoter that is fused to the coding sequence for YFP and the *CYC1* (cytochrome c isoform 1) terminator. The 678-bp sequence extends 5' from the

start codon of *TDH3* into the 3' untranslated (UTR) of the neighboring gene (*PDX1*), including the 5' UTR of *TDH3*. To facilitate replacing this reference haplotype with other P_{TDH3} haplotypes, we used homologous replacement to create a derivative of this starting strain in which the P_{TDH3} sequence as well as the start codon of YFP was replaced with the *URA3* gene (*URA3*-YFP; strain YPW44).

To assess *cis*-regulatory activity of naturally occurring P_{TDH3} haplotypes, we amplified the *TDH3* promoters from the 85 natural isolates using PCR and transformed these PCR products into the *URA3*-YFP intermediate. Unobserved intermediate haplotypes between all pairs of haplotypes that differ at exactly two sites were constructed by PCR-mediated site-directed mutagenesis of one of the two haplotypes in each pair and also transformed into the *URA3*-YFP strain. The 236 mutant P_{TDH3} alleles analyzed, each containing a single G:C→A:T transition, were also constructed using PCR-mediated site-directed mutagenesis, but starting with the reference P_{TDH3} haplotype. Each of these sequences was also transformed into the same *URA3-YFP* strain. All PCR primers used for amplification and site-directed mutagenesis are shown in Supplementary Table 2. In all cases, (1) transformations were performed using the lithium acetate method(GIETZ and WOODS 2006); (2) transformants were selected on 5-FOA plates, streaked for single colonies, and confirmed to not be petite (missing mitochondrial DNA) by replica plating onto YPG plates (3% (v/v) glycerol, 2% Oxoid yeast extract, 2% Oxoid peptone, 2% agar); and (3) Sanger sequencing was used to determine the sequence of potential transformants.

Quantifying fluorescence of P_{TDH3}-YFP, a proxy for *cis*-regulatory activity of P_{TDH3}

Prior work shows that fluorescence of reporter proteins such as YFP provide a reliable readout of *cis*-regulatory activity (KUDLA *et al.* 2009; GRUBER *et al.* 2012). Prior to quantifying fluorescence, all strains were revived from glycerol stocks onto YPG at the same time to control for age related effects on expression. Strains were inoculated from YPG solid media into 500 μl of YPD liquid media and grown for 20 hours at 30°C in 2 ml 96-well plates with 3 mm glass beads, shaking at 250 rpm. Immediately prior to flow cytometry, 20 μl of the overnight culture was transferred into 500 μl of SC-R (dextrose) media (GRUBER *et al.* 2012). Flow cytometry data were collected on an Accuri C6 using an intellicyt hypercyt autosampler. Flow rate was 14 μl/min and core size was 10 μm. A blue laser (λ = 488 nm) was used for excitation of YFP. Data were collected from FL1 using a 533/30 nm filter. Each culture was sampled for 2-3 seconds, resulting in approximately 20,000 recorded events.

Samples were processed using the flowClust (LO *et al.* 2009) and flowCore (HAHNE *et al.* 2009) packages within R (v 3.0.2) and custom R scripts (R CORE TEAM 2013) (Supplementary File 3). Raw data (Figure B-8a) was $\log_{10}$ transformed and artifacts were removed by excluding events with extreme FSC.H, FSC.A, SSC.H, SSC.A and width values (Figure B-8b). Samples were clustered based on FSC.A and Width to remove non-viable cells and cellular debris, and then clustered on FSC.H and FSC.A to remove doublets (Figure B-8c). Finally, samples were clustered on FL1.A and FSC.A to obtain homogeneous populations of cells in the same stage of the cell-cycle (Figure B-8d). At each filtering step, data were divided into exactly two clusters. Samples containing fewer than 1,000 events after processing were discarded. For each sample, YFP expression was

calculated as the median $\log_{10}(FL1.A)^2/\log_{10}(FSC.A)^3$. This corrects YFP expression levels for the correlation between fluorescence and cell size (measured by FSC.A) (Figure B-8e). Expression noise for each sample was calculated as σ/μ. The following alternative metrics for expression noise were also calculated and used for analysis σ, σ^2/μ^2, σ^2/μ, and residuals from a regression of σ on μ.

For each genotype, 9 independent replicate cultures were analyzed, with 3 biological replicates included on each of 3 different days. To control for variation in growth conditions, all plates contained 20 replicates of the wild-type reference strain, with at least one control sample in each row and column of the plate. For both mean expression and the standard deviation of expression, the control samples were fit to a linear model that included final cell number and average cell width as well as the day, replicate, array, read order, growth position in the incubator, array depth in incubator, measurement block, row, and column of the sample. Stepwise AIC was performed on this model to identify the most informative combination of variables to keep in the model. Plate (which incorporates effects of day, replicate, and array) and block were significant from this model. The effects of these factors were removed from measures of YFP (Figure B-8f-y) prior to the final analysis. A non-fluorescent strain containing no *TDH3* promoter was used to estimate auto fluorescence and this value was subtracted from all YFP expression values (Supplementary File 4, Supplementary Table 4).

Estimating effects of individual polymorphisms and mutations

The effect of an individual polymorphism on mean expression level and expression noise was measured as the difference in phenotype between the descendant and ancestral haplotypes that varied only for that polymorphism. The effect of an individual mutation on mean expression level and expression noise was measured as the difference in phenotype between the reference strain and the strain carrying that mutation. Statistical significance of effects for individual polymorphisms and mutations was assessed using two-sided *t*-tests.

Background effects

Although we frequently switched to fresh clones from glycerol stocks of the *URA3-YFP* strain during construction of the collection of 381 P_{TDH3}-*YFP* strains analyzed in this study, we checked for the presence of relevant second-site mutations that might have arisen spontaneously by independently reintroducing the P_{TDH3} reference allele three times. No difference in YFP fluorescence was observed among these replicate stains for either mean expression level or expression noise (mean *p*-value = 0.16, noise *p*-value = 0.069, n=1,483, ANOVA).

The reference haplotype used to determine the effect of new mutations differs from the most closely related natural haplotype (haplotype A) by a single base pair. To determine the impact of this single nucleotide difference on the distribution of mutational effects for mean expression level and expression noise, we introduced 28 of the G:C→A:T mutations into haplotype A and constructed P_{TDH3}-YFP strains that carried these alleles. The 28 mutations chosen for testing showed a range of effects on both mean expression

level and expression noise. We found that this single base difference significantly decreased mean expression level by 3.7% (p-value = 8.1×10^{-6}, ANOVA) and significantly increased expression noise by 6.8% (p-value = 1.61×10^{-4}, ANOVA), but these effects were largely consistent across genetic backgrounds, indicating little and/or weak epistasis (Figure B-9a,b). Indeed, we found that the distributions of mutational effects estimated by these 28 mutations on haplotype A and the 236 mutations on the reference haplotype were similar for both mean expression level and expression noise (Figure B-9c,d).

The reference background also contained 6 bp at the 5' end of the P_{TDH3} region derived from the 3' UTR of *PDX1* that was not included in the P_{TDH3} –YFP constructs containing natural P_{TDH3} haplotypes. To determine whether this sequence was likely to have affected our measurements of polymorphism effects, we tested for a significant change in YFP fluorescence when this 6bp were added to the P_{TDH3}-YFP alleles carrying the natural haplotypes A, D, and VV. We found no significant difference between genotypes with and without this 6 bp sequence (mean p-value = 0.64, noise p-value = 0.32, ANOVA).

Effects of cis-regulatory mutations and polymorphisms in a second trans-regulatory background

To determine the sensitivity of our conclusions to the specific genetic background used to assay *cis*-regulatory activity, we created hybrids between one of the natural *S. cerevisiae* isolates (YPS1000) and (i) 111 strains with mutations in P_{TDH3}-YFP, (ii) the strain carrying the reference P_{TDH3}-YFP allele, (iii) 39 strains with naturally occurring *TDH3* promoter haplotypes driving YFP expression, and (iv) a strain without the *TDH3* promoter in the P_{TDH3}-YFP construct and thus no YFP expression. YPS1000 was isolated

from an oak tree and is substantially diverged from BY (> 53,000 SNPs, 0.44%) (LITI *et al.* 2009; SCHACHERER *et al.* 2009; MACLEAN *et al.* 2015). We crossed all 152 of the strains described above (mating type **a**) to an isolate of YPS1000 that contained a KanMX4 drug resistance marker at the HO locus (mating type α). Hybrids were created by mixing equal cell numbers in liquid YPD and growing at 30C for 48 hours without shaking. Cultures were diluted and plated on YPG + G418 to select for hybrids and prevent petite cells from growing. Colonies were grown for 48 hours and then screened by fluorescent microscopy for YFP expression. Fluorescent colonies were streaked for single colonies and then a single colony was randomly chosen from each plate, transferred to a new plate, and confirmed to be diploid using a PCR reaction that genotyped the mating type locus. Four replicates of each strain were arrayed as in the original experiment with 20 controls per 96 well plate. Samples were grown for 20 hours in 500 ul of YPD liquid with shaking at 30C and then analyzed using the same flow cytometer machine and conditions described above. Samples were processed using the same analysis scripts described above and mean expression level and expression noise were calculated. Eight of the 111 genotypes carrying reporter genes with mutations as well as four of the 39 genotypes carrying reporter genes with polymorphisms showed phenotypes suggesting that they were aneuploidies. This rate is consistent with our previous observations of spontaneous aneuploidies produced by BY4742 (GRUBER *et al.* 2012). One additional strain (containing a mutation in the *TDH3* promoter) was also excluded for having highly inconsistent measurements among replicate populations. The R script used for this analysis is provided as Supplementary File 5 and the data are provided in Supplementary Table 5.

Tests for evidence of natural selection

Comparing the distribution of effects for single mutations and polymorphisms

In the absence of selection, the effects of polymorphisms are expected to be consistent with the effects of a random sample of new mutations. Because our data are non-normally distributed, we used non-parametric tests based on sampling to assess significance. To estimate the probability of occurrence for a mutation with a particular effect (x), we used a Gaussian kernel with a bandwidth of 0.01 to fit density curves to the distributions of mutational effects observed for both mean expression level and expression noise. We calculated the density for mean expression level values ranging from 0% to 200%, and for expression noise values ranging from 0% to 800%, ranges that extend beyond all observed effects. We set the minimum density for any effect size to 1/(number of mutations included in the mutational distribution). We expect this minimum to overestimate the true probability of most unobserved effect sizes, making this a conservative baseline for testing whether the effects of observed polymorphisms are a biased subset of all possible mutations. These density curves were then converted into probability distributions by setting the total density equal to 1 (Figure B-10a, b).

To calculate the log-likelihood of a set of n genetic variants with effects $x_1, x_2, \ldots, x_n$, we used these probability distributions to estimate the log-likelihood of a mutation with that effect, $p(x)$, and summed probabilities for all genetic variants. That is, the log-likelihood of a set of particular effects was calculated as $\sum_{i=1}^{n} \log(p(x_i))$. The log-likelihood calculated for the 45 observed polymorphisms was compared to the log-likelihoods of 100,000 samples of 45 mutations drawn randomly from the corresponding mutational distribution with replacement. To test the hypothesis that the effects of observed

polymorphisms were unlikely to result by chance from the mutational process alone, one-sided *p*-values were calculated as the proportion of random samples with log-likelihoods less than the log-likelihood value calculated for the observed polymorphisms. To determine the effects of mutations in the known TFBS on this test for selection, we excluded the effects of the mutations in the known TFBS from the distribution of mutational effects, recalculated the density curves and probability distributions, and then recalculated the log-likelihoods and *p*-values.

Inferring fitness functions from the observed effects of mutations and polymorphisms
Fitness functions relate the effect of a new mutation to its likelihood of survival within a population. We determined the most likely fitness function for mean expression level and expression noise by using a hill climbing algorithm to identify the α and β parameters of a beta distribution that maximized the likelihood of the observed polymorphism data when multiplied by the distribution of mutational effects. The beta function was started with parameters consistent with neutral evolution ($\alpha = 0$, $\beta = 0$) and new parameters were sampled randomly from a uniform distribution. The likelihood of the observed data was then calculated under the combined distribution of mutational effects and the new beta distribution. If the likelihood increased, the new parameters were kept; if not, they were discarded. This process was repeated until we observed 1,000 successive rejections. After each rejection, the width of the uniform distribution was increased in order to sample values farther away from the current parameters. A likelihood ratio test (df = 2) comparing the fitness function described by the maximum likelihood parameters for the beta distribution to a fitness function consistent with neutrality ($\alpha = 0$, $\beta = 0$) was used to test for statistically significant evidence of selection.

Comparing changes in P_{TDH3} activity observed over time to neutral expectations
If the effects of polymorphisms are determined solely by mutation, phenotypes should drift over evolutionary time in a manner dictated by the mutational process. We modeled such a neutral scenario by starting with the phenotype of the inferred common ancestor and adding to it effects randomly drawn from the mutational distribution (sampled with replacement) for each new polymorphism observed in the haplotype network, maintaining the observed relationships among haplotypes. This process was repeated 10,000 times to generate a range of potential outcomes consistent with neutral evolution of P_{TDH3} activity. We then compared the observed polymorphism data to the results of these neutral simulations to test for a statistically significant deviation from neutrality that would indicate selection. A more detailed description of this method follows.

Let x be the number of new polymorphisms added to the population to convert an observed haplotype into the most closely related descendent haplotype in each lineage that exists or must have existed in wild populations of *S. cerevisiae.* In the haplotype network for P_{TDH3}, x ranges from 0 to 5 (Figure B-2a). Pairs of haplotypes separated by 0 new polymorphisms result from recombination between existing haplotypes (e.g. haplotype RR, which is a recombinant of haplotypes W and FF).

The probability of a polymorphism with any particular effect being added to the population was assumed, in the absence of selection, to be equal to the probability of a new mutation with that effect. The log-likelihood of a single mutation ($x = 1$) with a particular effect was calculated using the probability distributions fit to density curves based on the observed mutational distributions described above. To generate equivalent probability distributions for sets of $x = 2, 3, 4,$ or 5 new mutations, we randomly drew x

mutations from the observed distribution of single mutational effects with replacement, calculated the combined effect of these mutations, and repeated this process 10,000 times. We then fit a density curve to these 10,000 combined effect values for each value of *x*, set the total density to 1 to convert this into a probability distribution, and used these curves (Figure B-10c, d) to calculate the log-likelihood of a particular set of *x* new polymorphisms with a given combined effect in the absence of selection. A likelihood of 1 was assigned to pairs of haplotypes separated only by recombination ($x = 0$), because the new genetic variant incorporated into the descendant haplotype was already known to have arisen in the population.

To calculate an overall log-likelihood for the observed set of polymorphisms, we summed the log-likelihood values for phenotypic differences observed between each pair of most closely related haplotypes seen among the natural isolates. To determine whether this overall log-likelihood for the observed polymorphisms was consistent with neutrality, we used the structure of the haplotype network to simulate 10,000 alternative sets of haplotype effects assuming that the effect of each new polymorphism was drawn randomly from the distribution of mutational effects. We calculated the log-likelihood for each node, in each set of haplotype effects, as $\log[\Pi_{x=1}^{5} (n_x! * \Pi_{i=1}^{n_x} p(x_i))]$, where x = the number of mutational steps, n_x = the number of immediately descendent haplotypes that are x mutational steps away from the focal node that exist or must have existed in *S. cerevisiae* (Figure B-2a), and $p(x_i)$ = the likelihood of the i^{th} mutation drawn from the probability distribution based on sets of x mutations. The $n_x!$ factor accounts for all possible ways that x mutations (or polymorphisms) added to the population at any given step could have been arranged among the set of descendent haplotypes observed.

To illustrate how this works for one particularly complex node in the network, consider haplotype H and its 6 immediately descendent haplotypes, L, I, VV, D, S and N (Figure B-2a). 5 of these descendent haplotypes (all except L) are all one mutational step away from H. To simulate the neutral evolution of these 5 haplotypes, we drew 5 mutational effects randomly from the probability distribution for single mutations (x = 1) with replacement, and then determined the likelihood of each of these mutational effects based on the probability distribution for x = 1. These likelihood values were multiplied together to calculate the combined probability of that particular set of 5 mutational effects occurring. This product was then multiplied by the 5! ways in which these mutations could have been arranged among the 5 descendent haplotypes. We also took into account that haplotype H has 1 additional descendent haplotype that is 5 mutational steps away from H (with none of the intermediate haplotypes known) by drawing a single value randomly from the distribution of mutational effects derived from random sets of 5 mutations (x = 5); calculated its likelihood using the probability distribution for x = 5; and multiplied it by the 1! way in which this set of 5 mutational effects could have been added to haplotype H to produce haplotype L.

The log-likelihoods for all nodes in the haplotype network were then summed to compute the log-likelihood of each set of haplotypes. To determine whether the *cis*-regulatory phenotypes observed among the natural isolates were consistent with neutral evolution, we compared the log-likelihood calculated for the observed polymorphisms to the log-likelihoods calculated for the 10,000 datasets simulated assuming neutrality. A one-sided *p*-value was calculated as the proportion of simulated neutral datasets that had a log-likelihood value less than the log-likelihood for the observed polymorphisms (Figure B-

5g,h, Table B-1).

Analysis of additional mutational data sets

To test for differences in effects among different types of point mutations, we analyzed data from previously published mutagenesis experiments in which the effects of individual mutations on *cis*-regulatory activity were determined (PATWARDHAN *et al.* 2009, 2012; KWASNIESKI and MOGNO 2012; MELNIKOV *et al.* 2012). Effects were split into each of the 12 mutation types and plotted on the same scale for all regulatory elements (Figure B-3). For each *cis*-regulatory element, we used an ANOVA to test for a significant difference among mutation types. In all cases, no significant effect was observed (p-value > 0.05). We also used a linear model including the identity of the *cis*-regulatory element and mutation type as main effects to test for a significant difference among mutational classes for sets of *cis*-regulatory elements across studies. Again, we found no significant difference among different types of mutations (p-value = 0.68, ANOVA).

Acknowledgements

We thank Calum Maclean, Jianzhi Zhang, and Chris Hittinger for strains, University of Michigan Center for Chemical Genomics for technical assistance with flow cytometry, and Joe Coolon, Rich Lusk, Kraig Stevenson, Andrea Hodgins-Davis, Jennifer Lachowiec, Calum Maclean, Jianrong Yang, Christian Landry, Jeff Townsend, and Dmitri Petrov for comments on the manuscript. Funding for this work was provided to

P.J.W. by the March of Dimes (5-FY07-181), Alfred P. Sloan Research Foundation, National Science Foundation (MCB-1021398), National Institutes of Health (1 R01 GM108826) and the University of Michigan. Additional support was provided by the University of Michigan Rackham Graduate School, Ecology and Evolutionary Biology Department and the National Institutes of Health Genome Sciences training grant (T32 HG000040) to B.P.H.M.; National Institutes of Health Genetics training grant (T32 GM007544) to D.C.Y.; National Institutes of Health NRSA postdoctoral fellowship (1 F32 GM083513-0) to J.D.G.; and EMBO postdoctoral fellowship (EMBO ALTF 1114-2012) to F.D.

Supplementary Data

Flow cytometry data was deposited to the FlowRepository (http://flowrepository.org) and assigned Repository ID FR-FCM-ZZBN. Additional data are located in Supplementary Tables 1-5 and analysis scripts are located in Supplementary Files 1-5 located online with the manuscript at:

http://www.nature.com/nature/journal/v521/n7552/full/nature14244.html

References

ASHKENAZY H., EREZ E., MARTZ E., PUPKO T., BEN-TAL N., 2010 ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. **38**: W529–33.

BAKER H. V, HUIE M. A., SCOTT E. W., DRAZINIC C. M., LOPEZ M. C., HORNSTRA I. A. N. K., YANG T. P., BAKERL H. V, 1992 Characterization of the DNA-Binding activity of GCR1 : in vivo evidence for two GCR1-binding sites in the upstream activating sequence of TPI of Saccharomyces cerevisiae. Mol. Cell. Biol. **12**: 2690–2700.

BATADA N. N., HURST L. D. L., 2007 Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat. Genet. **39**: 945–9.

CLEMENT M., POSADA D., CRANDALL K. a, 2000 TCS: a computer program to estimate gene genealogies. Mol. Ecol. **9**: 1657–1659.

DENVER D. R., MORRIS K., STREELMAN J. T., KIM S. K., LYNCH M., THOMAS W. K., 2005 The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nat. Genet. **37**: 544–548.

FONTANA W., BUSS L., 1994 "The arrival of the fittest": Toward a theory of biological organization. Bull. Math. Biol. **56**: 1–64.

FRANKEL N., DAVIS G. K., VARGAS D., WANG S., PAYRE F., STERN D. L., 2010 Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature **466**: 1–5.

FRASER H. B., HIRSH A. E., GIAEVER G., KUMM J., EISEN M. B., 2004 Noise Minimization in Eukaryotic Gene Expression. PLos Biol. **2**: 0834–0838.

GIETZ R., WOODS R., 2006 Yeast transformation by the LiAc/SS Carrier DNA/PEG method. In: *Methods in Molecular Biology, vol. 313: Yeast Protocols: Second Edition*,, pp. 107–120.

GRUBER J. D., VOGEL K., KALAY G., WITTKOPP P. J., 2012 Contrasting Properties of Gene-specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects and Dominance. PLoS Genet. **8**: e1002497.

HAHNE F., LEMEUR N., BRINKMAN R. R., ELLIS B., HAALAND P., SARKAR D., SPIDLEN J., STRAIN E., GENTLEMAN R., 2009 flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics **10**: 106.

HITTINGER C. T., 2013 Saccharomyces diversity and evolution: a budding model genus. Trends Genet. **29**: 309–17.

HORNUNG G., BAR-ZIV R., ROSIN D., TOKURIKI N., TAWFIK D. S., OREN M., BARKAI N., 2012 Noise-mean relationship in mutated promoters. Genome Res. **22**: 2409–2417.

KUDLA G., MURRAY A., TOLLERVEY D., PLOTKIN J., 2009 Coding-sequence determinants of gene expression in Escherichia coli. Science **324**: 255–258.

KWASNIESKI J., MOGNO I., 2012 Complex effects of nucleotide variants in a mammalian cis-regulatory element. PNAS **109**: 19498–19503.

LEHNER B., 2008 Selection to minimise noise in living systems and its implications for the evolution of gene expression. Mol. Syst. Biol. **4**: 1–6.

LIBKIND D., HITTINGER C. T., VALÉRIO E., GONÇALVES C., DOVER J., JOHNSTON M., GONÇALVES P., SAMPAIO J. P., VALERIO E., GONCALVES C., GONCALVES P., 2011 Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. PNAS **108**: 14539–14544.

LITI G., CARTER D. M., MOSES A. M., WARRINGER J., PARTS L., JAMES S. a, DAVEY R. P., ROBERTS I. N., BURT A., KOUFOPANOU V., TSAI I. J., BERGMAN C. M., BENSASSON D., O'KELLY M. J. T., OUDENAARDEN A. VAN, BARTON D. B. H., BAILES E., NGUYEN A. N., JONES M., QUAIL M. a, GOODHEAD I., SIMS S., SMITH F., BLOMBERG A., DURBIN R., LOUIS E. J., 2009 Population genomics of domestic and wild yeasts. Nature **458**: 337–341.

LITI G., NGUYEN BA A. N., BLYTHE M., MÜLLER C. a, BERGSTRÖM A., CUBILLOS F. a, DAFHNIS-CALAS F., KHOSHRAFTAR S., MALLA S., MEHTA N., SIOW C. C., WARRINGER J., MOSES A. M., LOUIS E. J., NIEDUSZYNSKI C. a, 2013 High quality de novo sequencing and assembly of the Saccharomyces arboricolus genome. BMC Genomics **14**: 69.

LO K., HAHNE F., BRINKMAN R. R., GOTTARDO R., 2009 flowClust: a Bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics **10**: 145.

LÖYTYNOJA A., GOLDMAN N., 2010 webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics **11**: 579.

LYNCH M., SUNG W., MORRIS K., COFFEY N., LANDRY C. R., DOPMAN E. B., DICKINSON W. J., OKAMOTO K., KULKARNI S., HARTL D. L., THOMAS W. K., 2008 A genome-wide view of the spectrum of spontaneous mutations in yeast. PNAS **105**: 9272–9277.

MACLEAN C. J., METZGER B. P. H., YANG J.-R., HO W.-C., MOYERS B., ZHANG J., 2015 Genome sequencing and high-throughput phenotypic analysis of diverse Saccharomyces cerevisiae strains. In Prep.

MCALISTER L., HOLLAND M. J., 1985 Differential expression of the three yeast glyceraldehyde-3-phosphate dehydrogenase genes. J. Biol. Chem. **260**: 15019–15027.

MELNIKOV A., MURUGAN A., ZHANG X., TESILEANU T., WANG L., ROGOV P., FEIZI S., GNIRKE A., CALLAN C. G., KINNEY J. B., KELLIS M., LANDER E. S., MIKKELSEN T. S., 2012 Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. **30**: 271–279.

NEWMAN J. R. S., GHAEMMAGHAMI S., IHMELS J., BRESLOW D. K., NOBLE M., DERISI J. L., WEISSMAN J. S., 2006 Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature **441**: 840–846.

PATWARDHAN R. P., HIATT J. B., WITTEN D. M., KIM M. J., SMITH R. P., MAY D., LEE C., ANDRIE J. M., LEE S.-I., COOPER G. M., AHITUV N., PENNACCHIO L. a, SHENDURE J.,

2012 Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. **30**: 265–270.

PATWARDHAN R. P., LEE C., LITVIN O., YOUNG D. L., PE'ER D., SHENDURE J., 2009 High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. **27**: 1173–1175.

PERRY M. W., BOETTIGER A. N., BOTHMA J. P., LEVINE M., 2010 Shadow enhancers foster robustness of Drosophila gastrulation. Curr. Biol. **20**: 1562–1567.

PIERCE S. E., DAVIS R. W., NISLOW C., GIAEVER G., 2007 Genome-wide analysis of barcoded Saccharomyces cerevisiae gene-deletion mutants in pooled cultures. Nat. Protoc. **2**: 2958–2974.

R CORE TEAM, 2013 R: A language and environment for statistical computing.

RASER J. M., O'SHEA E. K., 2004 Control of stochasticity in eukaryotic gene expression. Science **304**: 1811–1814.

RICE D. P. D., TOWNSEND J. P. J., 2012 A test for selection employing quantitative trait locus and mutation accumulation data. Genetics **190**: 1533–1545.

RINGEL A. E., RYZNAR R., PICARIELLO H., HUANG K., LAZARUS A. G., HOLMES S. G., 2013 Yeast Tdh3 (Glyceraldehyde 3-Phosphate Dehydrogenase) Is a Sir2-Interacting Factor That Regulates Transcriptional Silencing and rDNA Recombination (CS Pikaard, Ed.). PLoS Genet. **9**: e1003871.

SCANNELL D. R., ZILL O. A., ROKAS A., PAYEN C., DUNHAM M. J., EISEN M. B., RINE J., JOHNSTON M., HITTINGER C. T., 2011 The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 **1**: 11–25.

SCHACHERER J., SHAPIRO J. a, RUDERFER D. M., KRUGLYAK L., 2009 Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature **458**: 342–5.

SMITH J. D., MCMANUS K. F., FRASER H. B., 2013 A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers. Mol. Biol. Evol. **30**: 2509–2518.

STOLTZFUS A., YAMPOLSKY L. Y., 2009 Climbing mount probable: mutation as a cause of nonrandomness in evolution. J. Hered. **100**: 637–647.

TALY J.-F., MAGIS C., BUSSOTTI G., CHANG J.-M., TOMMASO P. DI, ERB I., ESPINOSA-CARRASCO J., KEMENA C., NOTREDAME C., 2011 Using the T-Coffee package to build

multiple sequence alignments of protein, RNA, DNA sequences and 3D structures. Nat. Protoc. **6**: 1669–1682.

TAMURA K., STECHER G., PETERSON D., FILIPSKI A., KUMAR S., 2013 MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. **30**: 2725–2729.

VRIES H. DE, 1905 *Species and Varieties, Their Origin by Mutation*. Open Court Publishing Company, Chicago.

WANG Q.-M., LIU W.-Q., LITI G., WANG S.-A., BAI F.-Y., 2012 Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. Mol. Ecol. **21**: 5404–5417.

WANG Z., ZHANG J., 2011 Impact of gene expression noise on organismal fitness and the efficacy of natural selection. PNAS **108**: E67–E76.

WITTKOPP P. J., 2011 Molecular Methods for Evolutionary Genetics. In: Orgogozo V, Rockman M V. (Eds.), *Molecular Methods for Evolutionary Genetics*, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 297–317.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2004 Evolutionary changes in cis and trans gene regulation. Nature **430**: 85–88.

YAGI S., YAGI K., FUKUOKA J., SUZUKI M., 1994 The UAS of the yeast GAPDH promoter consists of multiple general functional elements including RAP1 and GRF2 binding sites. J. Vet. Med. Sci. **56**: 235–244.

ZHANG Z., QIAN W., ZHANG J., 2009 Positive selection for elevated gene expression noise in yeast. Mol. Syst. Biol. **5**: 1–12.

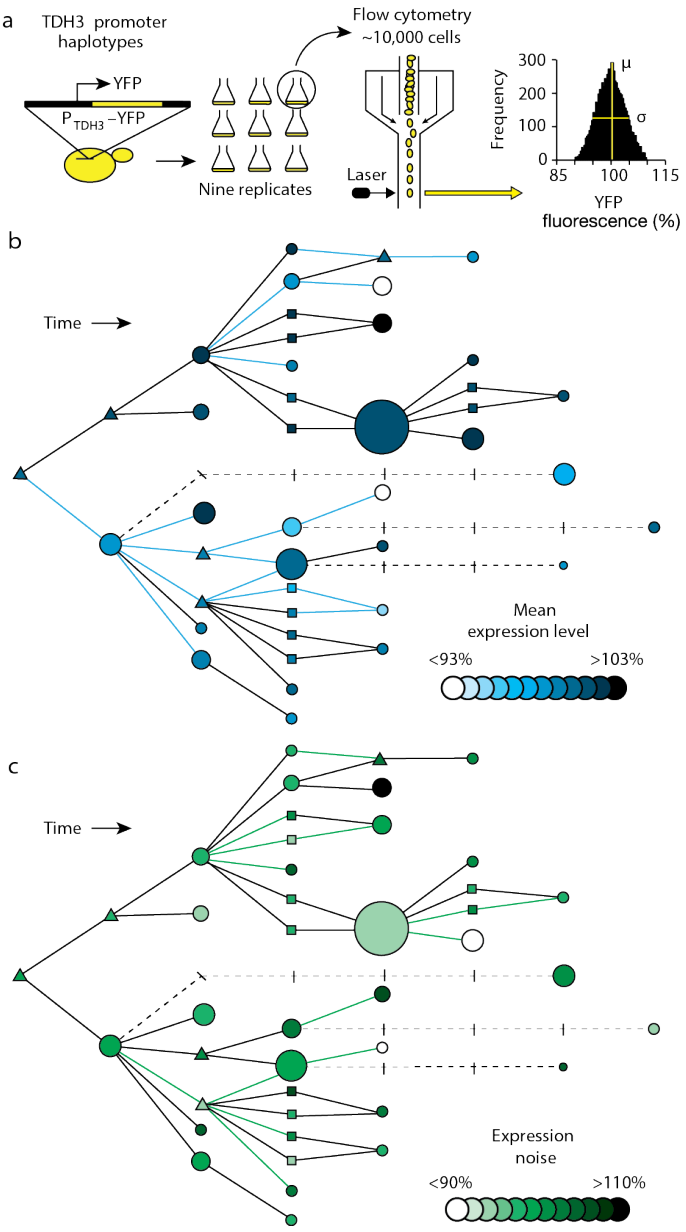


Figure 3-1 Effects of polymorphisms on P_{TDH3} activity

a, *cis*-regulatory activity was quantified as YFP fluorescence in 9 biological replicates for each P_{TDH3}-*YFP* haplotype using flow cytometry. The mean (μ) and standard deviation (σ) of single-cell fluorescence phenotypes were calculated for each sample. **b**, Mean expression level of P_{TDH3}-*YFP* for each *TDH3* promoter haplotype is shown in the haplotype network (Figure E2a), with differences in mean expression level relative to the inferred common ancestor shown with different shades. Circles are haplotypes observed among the sampled strains, with the diameter of each circle proportional to frequency of that haplotype among the 85 strains. Triangles are haplotypes that were not observed among the strains sampled, but must exist, or have existed, as intermediates between observed haplotypes. Squares are possible haplotypes that might exist, or have existed, as intermediates between observed haplotypes. Dashed lines connect haplotypes by multiple mutations. Based on *t*-tests with a Bonferroni correction, 17 of the 45 polymorphisms present in this network caused a significant change in mean expression level (blue lines). **c**, Same as **b**, but for expression noise. 18 of the 45 polymorphisms present in this network caused a significant change in expression noise (green lines, *t*-test, Bonferroni corrected)

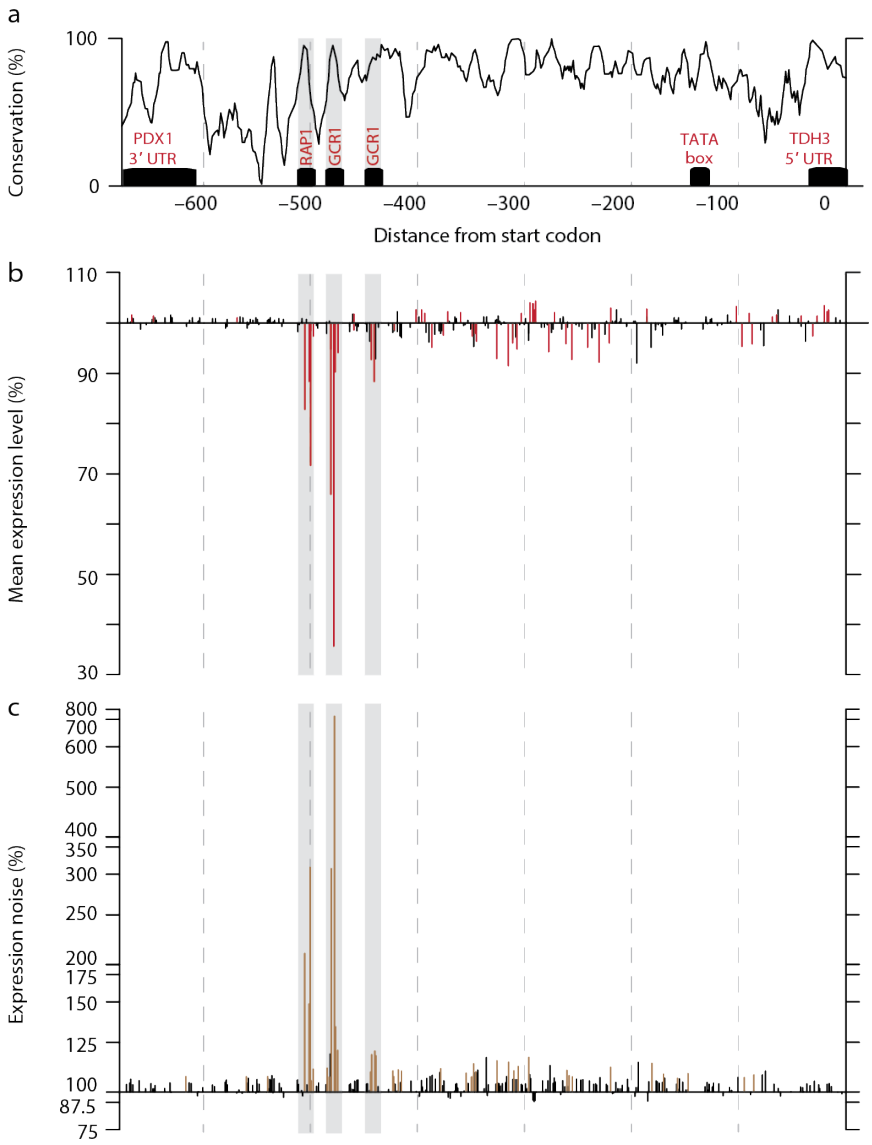


Figure 3-2 Effects of mutations on P_{TDH3} activity

a, The structure of the 678bp region analyzed, including the *TDH3* promoter with previously identified TFBS for RAP1 and GCR1, a TATA box, and UTRs for *TDH3* and *PDX1*, is shown. The black line indicates sequence conservation across the *sensu stricto* genus. **b**, Effects of individual mutations on mean expression level are shown in terms of the percentage change relative to the un-mutagenized reference allele, and are plotted according to the site mutated in the 678bp region. 59 of 236 mutations tested significantly altered mean expression levels (red lines, *t*-test, Bonferroni corrected). The shaded regions correspond to the known binding sites indicated in **a**. **c**, Same as **b,** but for expression noise. Because the effects of mutations on expression noise relative to the reference allele were much greater in magnitude than the effects of these mutations on mean expression level, they are plotted on a $\log_2$ scale. Measurements of expression noise were more variable among replicates than measurements of mean expression level, resulting in lower power to detect small changes as significant. Nonetheless, 42 of the 236 mutations tested significantly altered expression noise (brown lines, *t*-test, Bonferroni corrected).

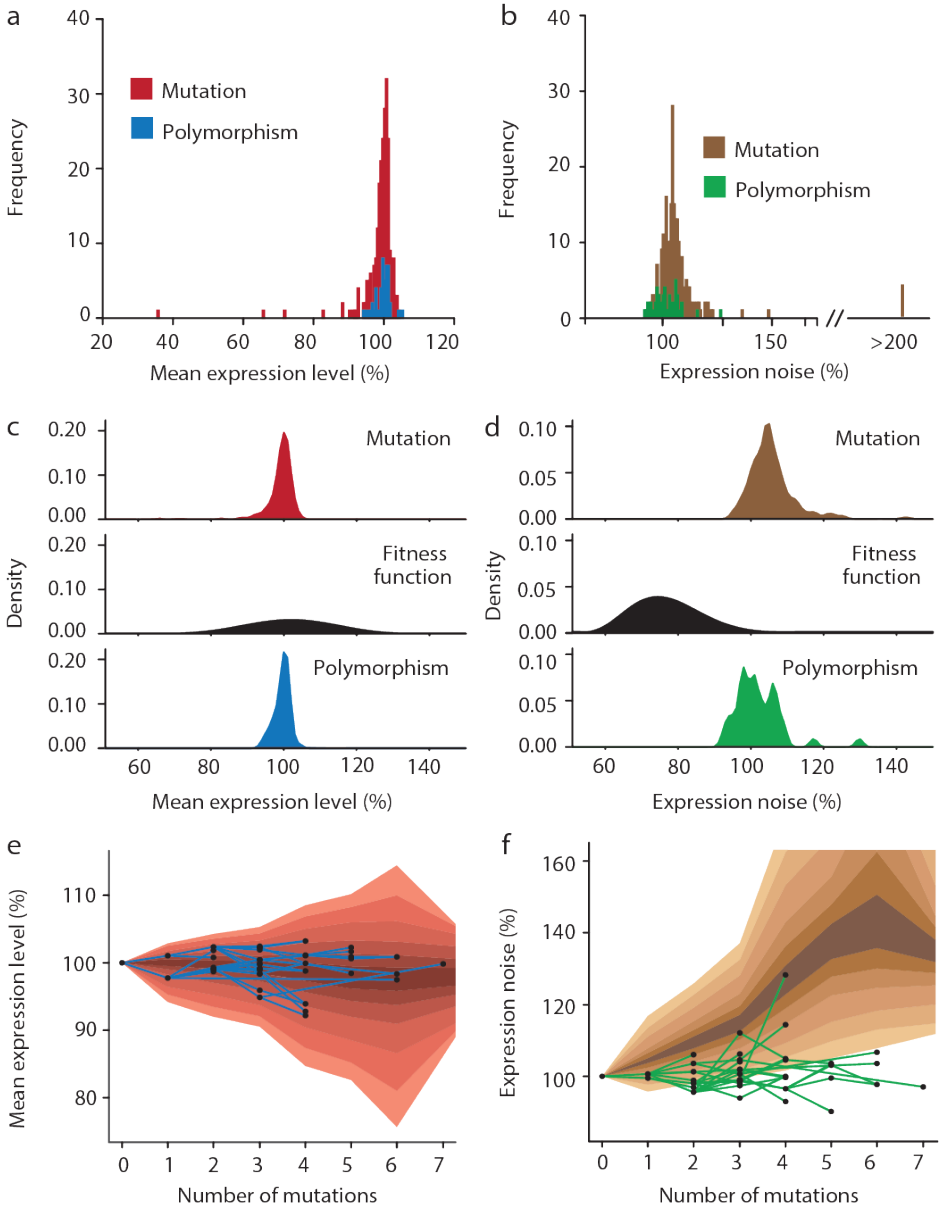


Figure 3-3 Effects of selection on P_{TDH3} activity

a, Histograms summarizing the effects of mutations (red) and polymorphisms (blue) on mean expression level are shown. **b**, Histograms summarizing the effects of mutations (brown) and polymorphisms (green) on expression noise are shown. **c**, The maximum likelihood fitness function (middle, black) relating the distribution of mutational effects (top, red) to the distribution of observed polymorphisms (bottom, blue) is shown for mean expression level. **d**, Same as **c**, but for expression noise. **e**, Changes in mean expression level observed among haplotypes over time in the inferred haplotype network (Figure E2a) are shown in blue. The red background represents the 95th, 90th, 80th, 70th, 60th and 50th percentiles, from light to dark, for mean expression level resulting from 10,000 independent simulations of phenotypic trajectories in the absence of selection. **f**, Same as **e**, but for expression noise. Effects of the mutational distribution are shown in brown. Expression noise among haplotypes is shown in green.

Chapter 4

Disentangling the effects of mutation and selection on the evolution of *TDH3 trans*-regulatory variation

Abstract

Heritable changes in gene expression are caused by *cis* and *trans*-regulatory mutations. While *trans*-regulatory mutations are expected to occur more frequently than *cis*-regulatory mutations, *cis*-regulatory changes often make substantial contributions to regulatory divergence in natural populations. Two possible explanations for this pattern include differences in the effect size of *cis* and *trans*-regulatory mutations and differences in the action of natural selection on *cis* and *trans*-regulatory changes. Here we measure the effects of new *trans*-regulatory mutations on gene expression of the *TDH3* gene in the yeast *Saccharomyces cerevisiae* and compare these effects to the effects of *trans*-regulatory polymorphisms segregating within the species to test for the action of natural selection on *TDH3* expression levels. We find evidence that selection has acted on both the mean level of expression as well as on expression noise, with the action of selection on expression noise opposite of that recently observed for *cis*-regulatory mutations. This difference in the action of natural selection arises from differences in the mutational distributions of *cis* and *trans*-regulatory changes. The effects of new *trans*-regulatory mutations also suggests that there are many small effect *trans*-regulatory segregating in

natural populations that both increase and decrease *TDH3* expression. We confirmed this prediction by mapping the genetic basis of *trans*-regulatory changes in *TDH3* expression among three diverse *S. cerevisiae* strains, finding hundreds of unique loci. Taken together, our results suggest that *trans*-regulatory changes in expression are a common source of regulatory variation in natural populations, are acted on by natural selection, and often have distinct effects on expression from *cis*-regulatory mutations.

Introduction

Heritable changes in gene expression are caused by mutations in *cis*-regulatory elements and *trans*-regulatory factors. While *cis*-regulatory elements are typically located near the gene they regulate, *trans*-regulatory factors can be located anywhere in the genome. As a consequence, the target size for *trans*-regulatory mutations is expected to be larger than the target size for *cis*-regulatory mutations and most new mutations are expected be *trans* acting (DENVER *et al.* 2005; LANDRY *et al.* 2007; GRUBER *et al.* 2012). However, *trans*-regulatory factors often interact with multiple *cis*-regulatory sequences and *trans*-regulatory changes are therefore expected to be more pleiotropic than *cis*-regulatory changes (PRUD'HOMME *et al.* 2007). As a consequence, purifying selection is expected to preferentially remove *trans*-regulatory mutations compared to *cis*-regulatory mutations. Consistent with this hypothesis, *cis*-regulatory changes often make greater contributions to regulatory divergence than *trans*-regulatory changes, despite their lower frequency of occurrence (WITTKOPP *et al.* 2008; TIROSH *et al.* 2009; EMERSON *et al.* 2010; GONCALVES *et al.* 2012; COOLON *et al.* 2014).

The relative contributions of *cis* and *trans*-regulatory changes to regulatory divergence depends on both the frequency of fixation and the effect on expression of *cis* and *trans*-regulatory mutations. For example, if *trans*-regulatory mutations are more common than *cis*-regulatory mutations, but have smaller effects on expression, then both *cis* and *trans*-regulatory changes can make substantial contributions to regulatory divergence. Studies mapping the genomic basis of expression differences in natural populations (expression Quantitative Trait Locus or, eQTL, mapping) are partly consistent with this hypothesis: likely *trans*-acting variants (distal eQTL) often have smaller effects on expression than likely *cis*-acting variants (local eQTL) (SCHADT *et al.* 2003). However, estimates of the frequency of *cis* and *trans*-regulatory changes from such studies are less clear as to which is more common (GIBSON and WEIR 2005; GILAD *et al.* 2008). Part of the difficulty in addressing this question is due to methodological limitations which bias identified eQTLs towards variants that cause large changes in expression. As a consequence, if *trans*-regulatory changes often have small effects on expression, then current eQTL mapping approaches are likely to be biased towards detecting *cis*-regulatory changes and may therefore be missing a major contributor to regulatory change, and thus evolution (ROCKMAN 2012).

In the absence of natural selection, polymorphisms are expected to be a random subset of new mutations. Thus, by comparing the phenotypic effects of polymorphisms and new mutations, the targets and extent of natural selection can be determined (DENVER *et al.* 2005; STOLTZFUS and YAMPOLSKY 2009; RICE and TOWNSEND 2012; SMITH *et al.* 2013; METZGER *et al.* 2015). While there has been considerable effort to determine the

phenotypic effects of polymorphisms, much less is known about the phenotypic effects of new mutations. In particular, it is unclear how the effects of new mutations on a phenotype differ across molecular mechanisms and if these differences have evolutionary consequences. For example, whether new *cis* and *trans*-regulatory mutations differ in their effects on expression and how such differences might interact with natural selection is unknown.

TDH3 is one of the most highly expressed genes in the *Saccharomyces cerevisiae* genome (MCALISTER and HOLLAND 1985; GHAEMMAGHAMI *et al.* 2003). It is involved in central metabolism (MCALISTER and HOLLAND 1985; VAN HEERDEN *et al.* 2014) and its deletion decreases fitness (PIERCE *et al.* 2007), suggesting that *TDH3* expression is maintained by natural selection. Surprisingly, however, previous work has shown that selection within the *TDH3* promoter has acted not to maintain high levels of *TDH3* expression, but to maintain low levels of expression noise (i.e. variability in expression between genetically identical cells) (METZGER *et al.* 2015). Previous work has also indicated substantial *trans*-regulatory variation affecting *TDH3* expression in natural populations. However, it is unclear whether natural selection has acted on *trans*-regulatory variation in the same manner as *cis*-regulatory variation.

Here we address these questions by comparing the effects on expression of *trans*-regulatory mutations and polymorphisms for the *TDH3* gene in the yeast *Saccharomyces cerevisiae*. We find that the action of natural selection varies between *cis* and *trans*-regulatory effects on expression. This difference is caused in part by differences in the

effects of new *cis* and *trans*-regulatory mutations. In addition, we find that many *trans*-regulatory mutations have small effects on expression. To determine if this is also true for *trans*-regulatory variation affecting *TDH3* expression in natural populations, we perform eQTL mapping across three diverse *S. cerevisiae* strains. We find widespread evidence of small effect *trans*-regulatory variants that are linked to nearly every region of the genome, in part due to our increased power compared to previous eQTL studies. Overall, our results suggest that *trans*-regulatory changes in expression are common in natural populations, that many of these effects are small, that they have distinct effects from *cis*-regulatory changes, and that they are therefore an important mechanism for regulatory evolution.

Results and Discussion

To determine likely sources of *trans*-regulatory effects on *TDH3* expression, we constructed a simple *TDH3* regulatory network. Using evidence of direct binding and changes in expression, we determined the set of likely direct regulators for *TDH3* expression. We expanded this network twice by determining all genes that bind to any gene already included in the network. The resulting network contains all known connections that are no more than three steps removed from the *TDH3* gene (Figure 4-1). We expect that a disproportionate number of new mutations affecting *TDH3* activity will affect genes in this network.

To determine the effects of new mutations on *TDH3* expression, we integrated a reporter gene containing the *TDH3* promoter and a yellow fluorescent protein (YFP) coding

sequence into the genome of an S288c derived strain (P_{TDH3}-YFP). We then used EMS to introduce primarily G→A and C→T mutations throughout the genome, creating approximately 32 mutations per individual. Previous work suggests that the effects of EMS on expression are similar to the effects of natural point mutations (METZGER *et al.* 2015). From this mutant population, we isolated ~1500 mutants, independent of their effects on YFP fluorescence. Given the relative target sizes of the genome and the native *TDH3* gene, we calculate that this collection contains less than one mutation in the native *TDH3* gene, and thus consists almost entirely of mutations with potential *trans*-regulatory effects on *TDH3* expression. For each isolated genotype, we determined its effects on YFP mean expression (μ) and expression noise (σ/μ) using flow cytometry of four independent biological replicates.

We found that most mutants had little effect on either mean expression level or expression noise, suggesting that only a small proportion of the yeast genome substantially alters *TDH3* activity when mutated (Figure 4-2). Surprisingly, relatively few mutants decreased *TDH3* expression. Instead, we identified several mutants with increased *TDH3* expression. This result is consistent with earlier reports of new *trans*-regulatory mutants being biased towards increases in *TDH3* expression (GRUBER *et al.* 2012). In contrast to the bias in mean expression, we found that the isolated mutants were biased towards decreased *TDH3* expression noise. These biases for mean expression and expression noise are opposite of that observed for new *cis*-regulatory mutations, indicating that the mechanism of regulation can have a large impact on the distribution of phenotypic effects expected from new mutations (METZGER *et al.* 2015). Identification of

causal mutations underlying these *trans*-regulatory effects on *TDH3* expression implicate several genes in the constructed *TDH3* regulatory network, such as *ROX1*, *TUP1*, and *TYE7* (DUVEAU *et al.* 2014, see Appendixes C and D).

To determine the effects of polymorphism on *TDH3 trans*-regulatory activity, we integrated the P_{TDH3}-YFP reporter construct into 55 phenotypically and genetically diverse strains of *S. cerevisiae* (LITI *et al.* 2009; SCHACHERER *et al.* 2009). Because the same reporter is introduced across strains, there is no *cis*-regulatory variation and any heritable difference in expression between strains must be caused by *trans*-regulatory changes. We measured YFP fluorescence in each strain using flow cytometry across 12 biological replicates, and estimated the effects of each strain's *trans*-regulatory background on both mean expression and expression noise relative to the reference strain S288c. Using the known phylogenetic relationships amongst the strains, we reconstructed the likely ancestral states for both mean expression (Figure 4-3) and expression noise (Figure 4-4) at each node and along each branch of the phylogeny.

We found that for the mean level of expression, most strains had expression levels similar to the reference strain, with *trans*-regulatory backgrounds typically increasing or decreasing *TDH3* activity by less than 10%. We also found evidence of phylogenetic structure, such that closely related strains had more similar levels of expression than more distantly related strains (Figure 4-3). These results indicate that either few difference affecting *TDH3* expression have occurred between these strains, that the majority of

effects that have occurred are small, or that there is relatively strong selection maintaining *TDH3* expression.

Likewise, we found that expression noise varied little amongst these strains, with expression noise in most strains between 80% and 120% the expression noise of the reference strain. However, we found less evidence of phylogenetic constraint for expression noise than for the mean level of expression, and we observed several large changes in expression noise typically restricted to only a single strain (Figure 4-4). This suggests the action of purifying selection to maintain *TDH3* expression noise.

To test for the action of natural selection on *TDH3 trans*-regulatory variation, we compared the effects on *TDH3* expression of new *trans*-regulatory mutations to the effects on *TDH3* expression of *trans*-regulatory polymorphisms. Surprisingly, after correcting for phylogenetic relationships amongst strain, we found that for mean expression, the effects of polymorphisms were more variable than the effects observed for new mutations (polymorphism $\sigma = 0.027$; mutation $\sigma = 0.012$; $p = 6 \times 10^{-31}$, Levene test). We observed the same pattern for expression noise (polymorphism $\sigma = 0.12$; mutation $\sigma = 0.07$; $p = 5 \times 10^{-8}$, Levene test). These results suggest that in natural populations, multiple mutations are responsible for the differences in *TDH3* activity between strains.

To account for the effects of multiple *trans*-regulatory mutations when testing for natural selection, we estimated the number of polymorphisms that occurred along each branch of

the phylogeny (MACLEAN *et al.* 2015). We then drew the appropriate number of effects from the mutational distribution and determined the combined change in expression along each branch expected from these mutations. We repeated this procedure 1000 times for each branch and used the resulting effects to determine the distribution in *TDH3* activity expected in the absence of natural selection and according to the observed phylogenetic relationships amongst strains. We did this separately for mean expression and expression noise. For mean expression, we observed that the expected distribution was centered on no change in expression, indicating that even in the absence of selection, *TDH3* mean expression is expected to remain relatively constant (Figure 4-5A). This result is a consequence of the mutational distribution and suggests that multiple *trans*-regulatory mutations will often compensate for one another. Regardless of this compensation, however, the distribution of expected mutational effects had a higher variance than the observed effects of polymorphisms, indicating that purifying selection has acted to maintain *TDH3* expression levels ($p < 1 \times 10^{-3}$, bootstrap).

We observed a related pattern for gene expression noise (Figure 4-5B). In particular, there was a large predicted skew towards decreased expression noise for new mutations. Because natural strains do not show this skew, this suggests that there has been selection to maintain the level of *TDH3* expression noise ($p < 1 \times 10^{-3}$, bootstrap). This contrasts with previous reports indicating selection to reduce *TDH3* expression noise in *cis*-regulatory sequences (METZGER *et al.* 2015). This difference in the direction of selection from *cis* and *trans*-regulatory mutations is due to differences in the distribution of mutational effects for *cis* and *trans*-regulatory mutations and not differences in the effects

observed in natural isolates. As a consequence, natural selection has maintained TDH3 expression noise by acting on two separate molecular mechanisms.

The above results suggest that there are a large number of *trans*-regulatory polymorphisms segregating within natural yeast populations affecting *TDH3* expression and noise. To test this idea, we mapped the genetic basis for *trans*-regulatory differences in *TDH3* expression for three *S. cerevisiae* strains compared to a common reference strain. These strains differ substantially from each other and the reference strain, and are thus expected to contain much of the *trans*-regulatory variation segregating within the species. To identify genomic regions altering *TDH3* expression, we used the natural isolates containing the P_{TDH3}-YFP reporter construct. For each strain, we performed three rounds of crossing and meiosis, producing populations containing millions of genetically distinct individuals. After the first and third rounds of meiosis, the 5% highest and 5% lowest fluorescence individuals from each population were isolated using fluorescence assisted cell sorting (FACS). This selection procedure was serially repeated a total of three times, each time selecting for extreme values in P_{TDH3}-YFP activity. Using Illumina sequencing, we determined allele frequencies within each sample and used differences in allele frequency between samples selected for high expression and samples selected for low expression to identify putative regions underlying differences in *TDH3* expression (ALBERT *et al.* 2014). Our procedure improves upon previous eQTL mapping approaches due to substantially higher population sizes (>100,000 at all steps), increased sporulation rates, reduced petite frequency, and better discrimination between mating types (Appendix C).

As expected, we found that increasing the number of rounds of selection on YFP fluorescence increased the power to detect differences in allele frequency, with three rounds of selection resulting in largest differences in allele frequency between samples (Figure 4-6a,b). Interestingly, this pattern was more extreme after three rounds of meiosis compared to one round of meiosis (Figure 4-6c). This effect could be caused by linkage between QTL with opposite effects on expression: because many QTL remain linked after only a single round of meiosis, selection for their individual effects is inefficient and will have low power to detect effects, regardless of the number of rounds of selection. However, upon additional rounds of meiosis, individual QTLs will be separated from one another and can be efficiently selected on. This explanation is consistent with the observed distribution of allele frequencies indicating clear blocks of linked SNPs with similar frequencies across the genome (Figure 4-7).

In total we identified 244 independent QTL affecting *TDH3* expression across these three strains (134 increases, 110 decreases) (Figure 4-6d, Figure 4-8). The number of QTL identified within each strain varied from 64 to 108, and was broadly consistent with evolutionary distance between each strain and the reference strain. Each QTL was required to be identified at least twice for a strain, suggesting that the majority are unlikely to be false positives. Interestingly, approximately an equal number of QTL increasing and decreasing expression were identified within each comparison, consistent with the relatively small net change in *TDH3* expression observed between strains. The identified QTL show little overlap between strains, suggesting independent mutations

altering *TDH3* regulation in each strain. This is consistent with the relatively short shared branch between each strain and the reference strain and suggests that overlapping QTL are likely to represent changes specific to the S288c branch.

Due to large distances and small overlap in eQTL between strains, it is currently unclear how frequently eQTL are gained or lost during evolution. It will therefore be useful for future work to consider the location and effect of QTL from additional natural isolates. In addition, while the current estimates of eQTL window size are typically on the order of 20-30 kb, and thus span numerous genes and segregating sites, the additional of multiple strains carrying the same QTL should help to narrow the identity of causative sites (SONAH *et al.* 2014). Novel statistical methods will likely need to be developed to fully capture the information presented in such data sets.

Methods

TDH3 Regulatory Network Reconstruction

To construct a putative regulatory network for *TDH3*, we used information from yeastract (TEIXEIRA *et al.* 2014). This database contains curated connections between transcription factors and the genes that they regulate. For a direct regulatory connection to *TDH3*, we required evidence of both binding of the transcription factor to *TDH3*, as well as changes in *TDH3* expression upon changes in the transcription factors expression. To expand the regulatory network, we performed the same analysis two additional times, each time including connections between any transcription factor and genes already included in the regulatory network. Network layout was produced using cytoscape (LOPES *et al.* 2010).

Yeast strains

New *trans*-regulatory mutations were created in strain YPW1139 (MATα). This strain is derived from BY4724, BY4722, BY4730 and BY4742 and contains no auxotrophies. In addition, this strain contains five mutations derived from natural yeast strains that fix two defects in the common laboratory strains: high frequency of petites and low sporulation rate. The improved alleles introduced are RME1(ins-308A), TAO3(1493Q) from Deutschbauer and Davis 2005 and SAL1, CAT5-91M and MIP1-661T from Dimitrov et al. 2009. Finally, this strain contains a copy of the *TDH3* promoter, Yellow Fluorescent Protein (YFP) coding sequence, CYC1 terminator, and KanMX4 drug resistance cassette inserted at the HO locus on chromosome IV (P_{TDH3}-YFP, see Appendix C).

Natural yeast strains were derived from LITI *et al.* 2009 and SCHACHERER *et al.* 2009. Initial construction for the strains is described in Chapter 2. Into each of 56 natural strains, as well as the reference strain S288c, we inserted the P_{TDH3}-YFP construct into the same genomic location. All strain used were MATα. Because these strains were already G4148 resistant, we used a NatMX4 resistance marker instead. For each strain, we screened for the presence of nourseothricin resistance, the loss of G418 resistance, and for the presence of YFP fluoresce. We confirmed correct insertion and the absence of mutations within the construct through Sanger sequencing of each strain. Strains are listed in Table 4-1.

Mutagenesis

EMS mutagenesis and the generation of the mutational distribution for *TDH3 trans*-regulatory effects is described in detail in Chapter 5. Briefly, a low dose of EMS was used to elevate the mutation rate, with each individual gaining ~32 mutations. Using fluorescence assisted cell sorting, individual genotypes were isolated from the EMS mutagenesis population irrespective of their effects on YFP florescence. Each isolated genotype was grown in four biological replicates in YPD and YFP fluorescence was estimated using flow cytometry relative to a non-mutant control sample. After quality filtering, 1485 individual genotypes remained.

YFP expression measurements for natural S. cerevisiae strains

To determine the effects of *trans*-regulatory polymorphism on *TDH3* expression we measured YFP activity for each natural isolate. To do this, we first revived all strains from glycerol stocks onto YPG at 30°C. After 24 hours of growth, each strain was inoculated into YPD media in a 96 well plate. In addition, control strains, the reference strain YPW1139, a strain containing no YFP coding sequence (YPW880), two *trans*-regulatory mutants, and two *cis*-regulatory mutants, were included at specific locations. This structure was replicated to solid YPD media using a pin tool and grown at 30°C. After 24 hours, the resulting colonies were pin-tool replicated into twelve 96 well plates containing 500 μl of YPD and grown at 30°C for 24 hours to reach saturation. Cultures were then diluted 1/20 into fresh 500 μl of YPD and grown for an additional four hours. Samples were then diluter 1/10 into 500 ul PBS and run on an Accuri C6 flow cytometer connected to an Intellicyt autosampler. Data was processed using the same procedure as in previous reports (METZGER *et al.* 2015). Mean expression and expression noise for

each strain was determined using custom scripts in R (R CORE TEAM 2013). To reconstruct the likely ancestral states at each node, we used the phylogeny from chapter 2 and the R package ape (PARADIS *et al.* 2004).

Test for selection

To test for the action of natural selection, we used phylogenetic contrasts to remove the effect of phylogenetic relatedness and compared the variability in expression from these contrasts to the variability in expression for new mutations. This test suggested that multiple *trans*-regulatory mutations have occurred along many branches of the phylogeny. To correct for this difference, we estimated the number of *trans*-regulatory mutations that occurred by calculating the distance between each node of the phylogeny. Because each mutant in the *trans*-regulatory distribution contained ~32 mutations, we divided the total number of mutations by 32 and used the resulting estimate to sample from the mutational distribution. We multiplied the sampled effects together and repeated this process 1000 times for each branch of the phylogeny. We did this separately for mean expression and expression noise. We then calculated the expected variability in both mean expression and expression noise for each of the 1000 replicates and compared the resulting distribution to the observed variability in the natural strains.

eQTL Mapping

To identify genomic regions responsible for differences in TDH3 expression due to *trans*-regulatory mutations, we performed eQTL mapping (Figure 4-9). We crossed versions of YPS1000 (PJW1057), SK1 (PJW1016), and M22 (PJW1072) containing the P_{TDH3}-YFP reporter to a version of S228c (PJW1240) that contains the same construct, but is MAT**a**

and G418 resistant instead of nourseothricin resistant. In addition, this strain contains an Red Fluorescent Protein (RFP) marker at its mating type locus, allowing for easy identification of mating type (Appendix C, CHIN *et al.* 2012). For each cross, we selected for diploids using a combination of nourseothricin and G418 resistance and selected a single colony to ensure homogeneity in the genetic background (P.0). We then grew each diploid on GNA media for 12-16 hours and then sporulated each diploid using KAc plates.

After greater than 50% sporulation, we isolated individual spores. Cells were first washed twice in 1 ml of H2O and then incubated with 200 ul of 0.3mg/ml 100T zymolyase for one hour with agitation. Cells were then washed with 1 ml of H2O and resuspended in 100 ul of H2O. Cells were vortexed for 2 minutes to stick spores to the tube wall. The supernatant was removed and 1 ml of H2O was added. Without agitation, this 1 ml was removed and a second 1 ml of H2O was added. This 1 ml was also removed and 1ml of triton-X (0.02%) was added. Samples were sonicated on ice for 10 seconds at medium power (3.5 on a Sonic Dismembrator Model 100, Fisher). Spores were confirmed to be separated and diploids absent by visual inspection under a microscope.

After spore isolate, the population was split into thirds. One third was added to 1 ml YPD, grown to saturation overnight, and then frozen at -80°C as a glycerol stock. The second third was used to initiate a second round of crossing by spotting onto YPD plates and grown overnight at 30°C. The final third (P.1) was sorted for the absence of the RFP marker using fluorescence assisted cell sorting (FACS) on a FACS canto II at the

University of Michigan Flow Cytometry Core. Because all MAT**a** and diploid strains should be RFP positive, this sorting capture only MATα cells. For each cross, we collected > 10^6 RFP minus individuals (F.1). These were incubated with 1ml YPD and grown for 24-28 hours at 30°C. This process (sporulation and spore isolation) was repeated two more times, each time using the spores isolated in the previous round to regenerate diploids (P.2 and P.3). After this third round of sporulation and meiosis, RFP minus cells were sorted again.

After growth to saturation, RFP minus populations (F.1 and F.3) were sorted into two distinct populations based on YFP fluorescence. Cells were transferred to 1 ml PBS and we used the middle 80% of cells based on FSC to sort the 5% highest (H.1.1 and H.3.1) or the 5% lowest (L.1.1 and L.3.1) YFP individuals after correcting for FSC. In each case, we sorted 100,000 individuals. Each sorted population was grown in YPD at 30°C for 20 hours after which one half was frozen to make a glycerol stock, and the other half used to select on YFP fluorescence an additional time. After the first round of selection, the direction of YFP selection was maintained within each sample and two more round of selection were applied.

Allele Frequency Determination

After all selection steps were completed, samples were revived from glycerol stocks and grown in 1 ml of YPD for 2 hours at 30°C. DNA was extracted from each sample using the Purgene Yeast Kit from Qiagen. DNA concentration was determined using a Qubit and Illumina Nextera XT libraries were prepared following the manufacturers guidelines. Barcodes for each sample are listed in Table 4-2. Library quality was confirmed by

bioanalzyer and all samples were pooled equally using concentration estimateds from Qubit. Samples were sequenced on a HiSeq 2000 using 125 bp paired end sequencing at the University of Michigan Sequencing Core.

QTL identification

After sequencing, samples were processed to identify individual QTL (Figure 4-10). Sickle was used to remove low quality bases from each read (JOSHI, N.A., FASS 2011) and Cutadapt was used to remove any adapter sequence from read ends (MARTIN 2011). Samples were aligned to the S228c reference genome using bowtie2 (LANGMEAD and SALZBERG 2012) and then sorted and indexed using samtools (LI *et al.* 2009). Overlapping reads were clipped using bamUtil. SNPs were jointly called within each paired set of samples selected for high and low YFP fluorescence using freebayes (GARRISON and MARTH 2012). SNP were required to reach at least 20% frequency in at least one of the two paired samples.

For each pair of samples, SNPs were filtered based on quality and depth. Each SNP required a depth of at least 20, but less than 500, a mapping quality score of greater than 30, and imbalance scores for left/right, center/end, and forward/reverse for SNP position within reads of less than 30. Finally, at each position, only the two highest likelihood SNPs were retained. For each SNP we calculated G using a likelihood ratio test of alternative and reference alleles within the high and low selected populations. For SNPs where the alternative allele was higher than the reference allele in the high selected population relative to the low selected population, we maintained the sign of G. For SNPs where the alternative allele was lower than the reference allele in the high selected

population relative to the low selected population, we flipped the sign of G. We then calculated G' by averaging these estimates over a 40 kb window centered on the SNP (MAGWENE *et al.* 2011). To identify QTL peaks, we implemented a hill climbing algorithm that identified all local maxima and minima in G'. We called peaks for values reaching above an absolute value of G' of 5. We estimated confidence intervals on the location of the peaks by drop in 2 of the absolute G' value. We then required that the same peak be identified in at least two of the six paired samples for each strain. QTL between strains whose peaks were located within each other confidence interval and in the same direction were called as the same QTL.

References

ALBERT F. W., TREUSCH S., SHOCKLEY A. H., BLOOM J. S., KRUGLYAK L., 2014 Genetics of single-cell protein abundance variation in large yeast populations. Nature **506**: 494–497.

CHIN B. L., FRIZZELL M. a., TIMBERLAKE W. E., FINK G. R., 2012 FASTER MT: Isolation of Pure Populations of a and Ascospores from Saccharomyces cerevisiae. G3 **2**: 449–452.

COOLON J. D., MCMANUS C. J., STEVENSON K. R., GRAVELEY B. R., WITTKOPP P. J., 2014 Tempo and mode of regulatory evolution in Drosophila. Genome Res.: gr.163014.113–.

DENVER D. R., MORRIS K., STREELMAN J. T., KIM S. K., LYNCH M., THOMAS W. K., 2005 The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nat. Genet. **37**: 544–548.

DEUTSCHBAUER A. M., DAVIS R. W., 2005 Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat. Genet. **37**: 1333–40.

DIMITROV L. N., BREM R. B., KRUGLYAK L., GOTTSCHLING D. E., 2009 Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. Genetics **183**: 365–83.

DUVEAU F., METZGER B. P. H., GRUBER J. D., MACK K., SOOD N., BROOKS T. E., WITTKOPP P. J., 2014 Mapping Small Effect Mutations in Saccharomyces cerevisiae: Impacts of Experimental Design and Mutational Properties. G3: 1205–1216.

EMERSON J. J., HSIEH L.-C., SUNG H.-M., WANG T.-Y., HUANG C.-J., LU H. H.-S., LU M.-Y. J., WU S.-H., LI W.-H., 2010 Natural selection on cis and trans regulation in yeasts. Genome Res.: 826–836.

GARRISON E., MARTH G., 2012 Haplotype-based variant detection from short-read sequencing. ArXiv.

GHAEMMAGHAMI S., HUH W.-K., BOWER K., HOWSON R. W., BELLE A., DEPHOURE N., O'SHEA E. K., WEISSMAN J. S., 2003 Global analysis of protein expression in yeast. Nature **425**: 737–41.

GIBSON G., WEIR B., 2005 The quantitative genetics of transcription. Trends Genet. **21**: 616–623.

GILAD Y., RIFKIN S. a., PRITCHARD J. K., 2008 Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. **24**: 408–415.

GONCALVES A., LEIGH-BROWN S., THYBERT D., STEFFLOVA K., TURRO E., FLICEK P., BRAZMA A., ODOM D. T., MARIONI J. C., 2012 Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res. **22**: 2376–2384.

GRUBER J. D., VOGEL K., KALAY G., WITTKOPP P. J., 2012 Contrasting Properties of Gene-specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects and Dominance. PLoS Genet. **8**: e1002497.

HEERDEN J. H. VAN, WORTEL M. T., BRUGGEMAN F. J., HEIJNEN J. J., BOLLEN Y. J. M., PLANQUÉ R., HULSHOF J., O'TOOLE T. G., WAHL S. A., TEUSINK B., PLANQUE R., HULSHOF J., O'TOOLE T. G., WAHL S. A., TEUSINK B., PLANQUÉ R., HULSHOF J., O'TOOLE T. G., WAHL S. A., TEUSINK B., PLANQUE R., HULSHOF J., O'TOOLE T. G., WAHL S. A., TEUSINK B., 2014 Lost in Transition: Startup of Glycolysis Yields Subpopulations of Nongrowing Cells. Science **343**: 1245114–1245114.

JOSHI, N.A., FASS J. N., 2011 Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

LANDRY C. R., LEMOS B., RIFKIN S. A., DICKINSON W. J., HARTL D. L., 2007 Genetic properties influencing the evolvability of gene expression. Science **317**: 118–121.

LANGMEAD B., SALZBERG S. L., 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**: 357–359.

LI H., HANDSAKER B., WYSOKER A., FENNELL T., RUAN J., HOMER N., MARTH G., ABECASIS G. R., DURBIN R., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

LITI G., CARTER D. M., MOSES A. M., WARRINGER J., PARTS L., JAMES S. a, DAVEY R. P., ROBERTS I. N., BURT A., KOUFOPANOU V., TSAI I. J., BERGMAN C. M., BENSASSON D., O'KELLY M. J. T., OUDENAARDEN A. VAN, BARTON D. B. H., BAILES E., NGUYEN A. N., JONES M., QUAIL M. a, GOODHEAD I., SIMS S., SMITH F., BLOMBERG A., DURBIN R., LOUIS E. J., 2009 Population genomics of domestic and wild yeasts. Nature **458**: 337–341.

LOPES C. T., FRANZ M., KAZI F., DONALDSON S. L., MORRIS Q., BADER G. D., 2010 Cytoscape web: An interactive web-based network browser. Bioinformatics **26**: 2347–2348.

MACLEAN C. J., METZGER B. P. H., YANG J.-R., HO W.-C., MOYERS B., ZHANG J., 2015 Genome sequencing and high-throughput phenotypic analysis of diverse Saccharomyces cerevisiae strains. In Prep.

MAGWENE P. M., WILLIS J. H., KELLY J. K., 2011 The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing (A Siepel, Ed.). PLoS Comput. Biol. **7**: e1002255.

MARTIN M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: pp. 10–12.

MCALISTER L., HOLLAND M. J., 1985 Differential expression of the three yeast glyceraldehyde-3-phosphate dehydrogenase genes. J. Biol. Chem. **260**: 15019–15027.

METZGER B. P. H., YUAN D. C., GRUBER J. D., DUVEAU F. D., WITTKOPP P. J., 2015 Selection on noise constrains variation in a eukaryotic promoter. Nature **521**: 344–347.

PARADIS E., CLAUDE J., STRIMMER K., 2004 APE: Analyses of phylogenetics and evolution in R language. Bioinformatics **20**: 289–290.

PIERCE S. E., DAVIS R. W., NISLOW C., GIAEVER G., 2007 Genome-wide analysis of barcoded Saccharomyces cerevisiae gene-deletion mutants in pooled cultures. Nat. Protoc. **2**: 2958–2974.

PRUD'HOMME B., GOMPEL N., CARROLL S. B., 2007 Emerging principles of regulatory evolution. PNAS **104**: 8605–8612.

R CORE TEAM, 2013 R: A language and environment for statistical computing.

RICE D. P. D., TOWNSEND J. P. J., 2012 A test for selection employing quantitative trait locus and mutation accumulation data. Genetics **190**: 1533–1545.

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

SCHACHERER J., SHAPIRO J. a, RUDERFER D. M., KRUGLYAK L., 2009 Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature **458**: 342–5.

SCHADT E. E. E., MONKS S. S. a, DRAKE T. T. a, LUSIS A. J. A., CHE N., COLINAYO V., RUFF T. G., MILLIGAN S. B., LAMB J. R., CAVET G., LINSLEY P. S., MAO M., STOUGHTON R. B., FRIEND S. H., 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **205**: 1–6.

SMITH J., MCMANUS K., FRASER H., 2013 A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers. arXiv.

SONAH H., O 'DONOUGHUE L., COBER E., RAJCAN I., BELZILE F., 2014 Combining Genome-wide Association and QTL Analysis: Opportunities and Challenges. : 3–6.

STOLTZFUS A., YAMPOLSKY L. Y., 2009 Climbing mount probable: mutation as a cause of nonrandomness in evolution. J. Hered. **100**: 637–647.

TEIXEIRA M. C., MONTEIRO P. T., GUERREIRO J. F., GONÇALVES J. P., MIRA N. P., SANTOS S. C. DOS, CABRITO T. R., PALMA M., COSTA C., FRANCISCO A. P., MADEIRA S. C., OLIVEIRA A. L., FREITAS A. T., SÁ-CORREIA I., 2014 The YEASTRACT database: An upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic Acids Res. **42**: 161–166.

TIROSH I., REIKHAV S., LEVY A. a, BARKAI N., 2009 A yeast hybrid provides insight into the evolution of gene expression regulation. Science **324**: 659–662.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2008 Regulatory changes underlying expression differences within and between *Drosophila* species. Nat. Genet. **40**: 346–350.

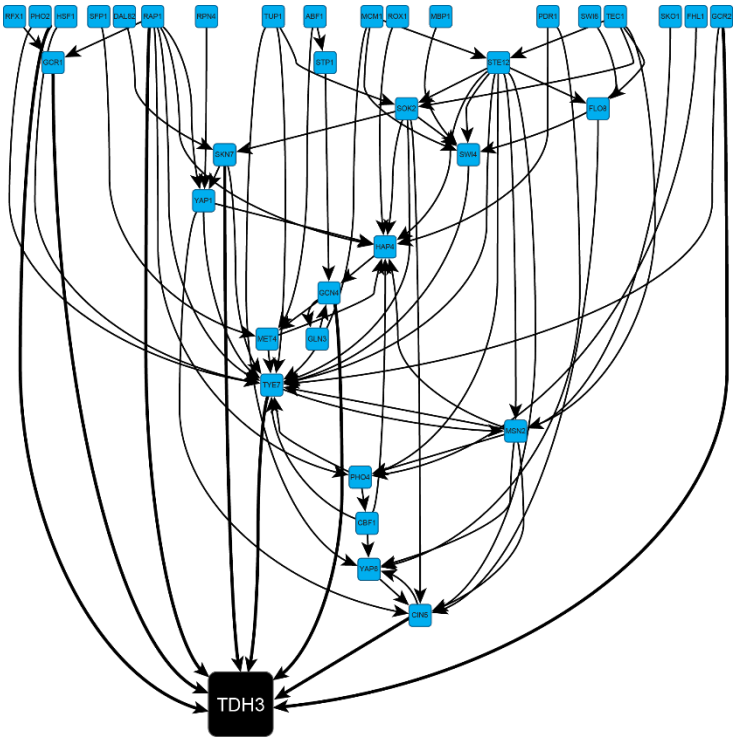


Figure 4-1 Predicted regulatory network for *TDH3* expression

Predicted regulatory network for the TDH3 gene (black). Each node (blue) represents a gene that either directly (dark arrows) or indirectly (light arrows) shows evidence of binding to the TDH3 promoter with corresponding expression changes. Network contains all connections that are three steps removed from the TDH3 gene.

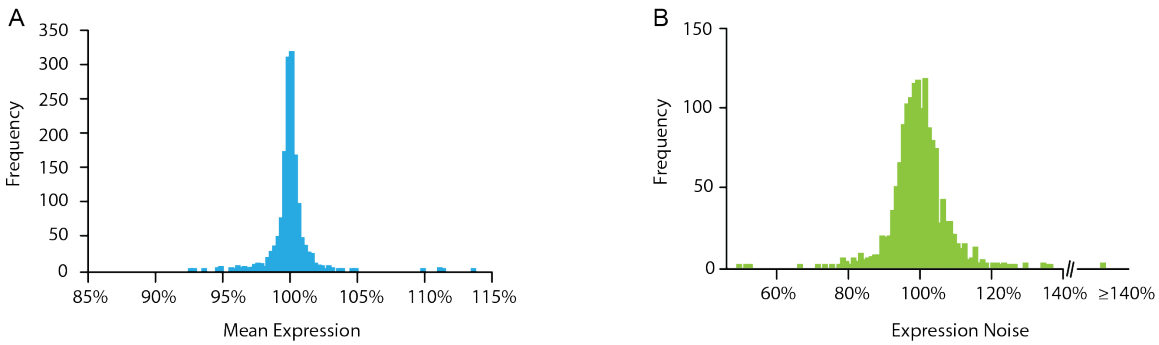


Figure 4-2 Distributions of *trans*-regulatory mutational effects for *TDH3*

Distributions of effects on TDH3 mean expression (A) and expression noise (B) for 1485 mutants. Mutants were created using EMS mutagenesis. Mean expression and expression noise were estimated using four biological replicates of each mutant.

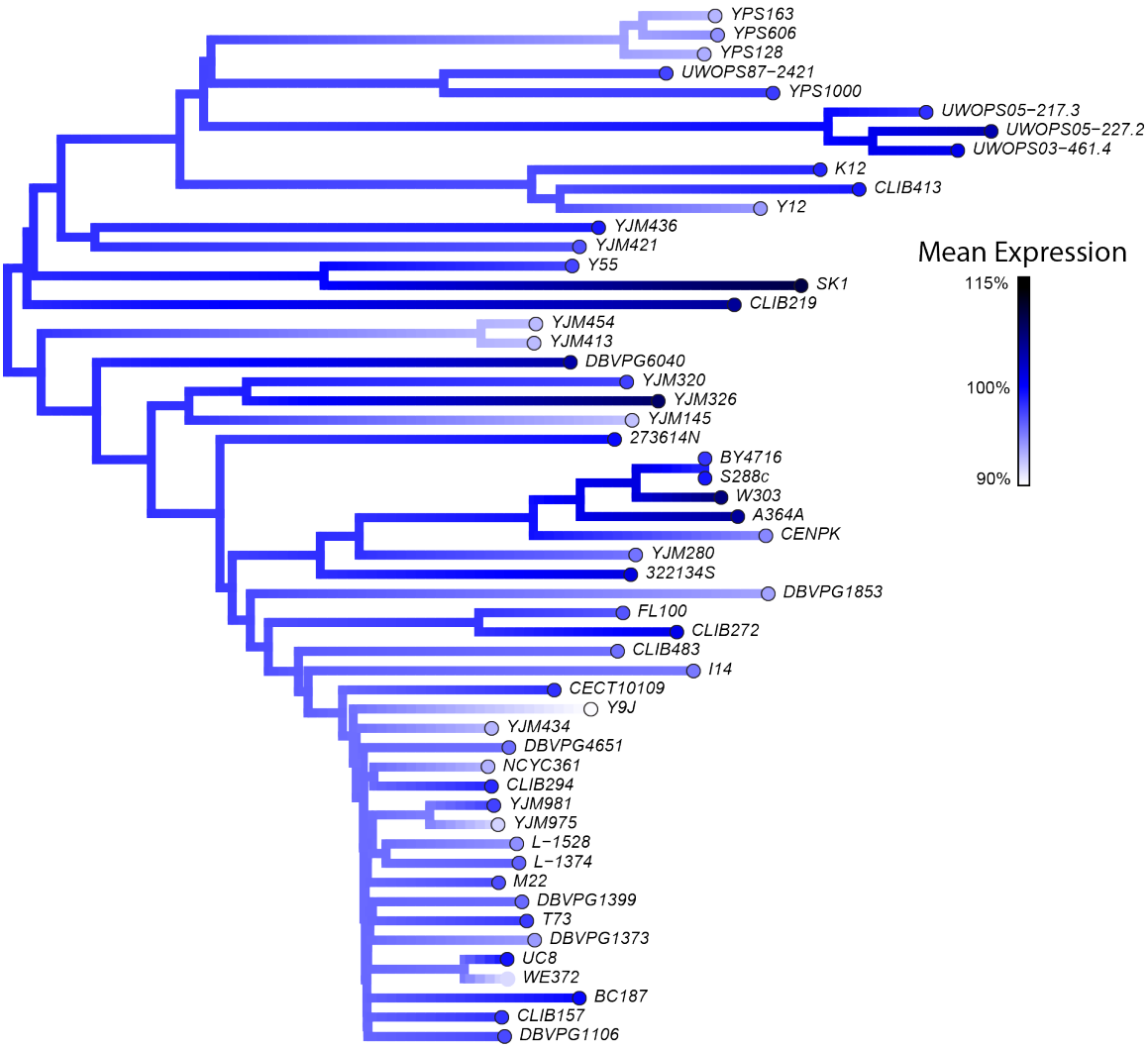


Figure 4-3 Effects of *trans*-regulatory polymorphisms on *TDH3* mean expression

Color of each circle at the ends of a branch reflects the observed effects of *trans*-regulatory polymorphisms on *TDH3* mean expression for each strain tested. Values are relative to those observed in S288c. Branch colors follow the same scale and were determined by reconstructing the most likely ancestral state at each node under a model of Brownian motion.

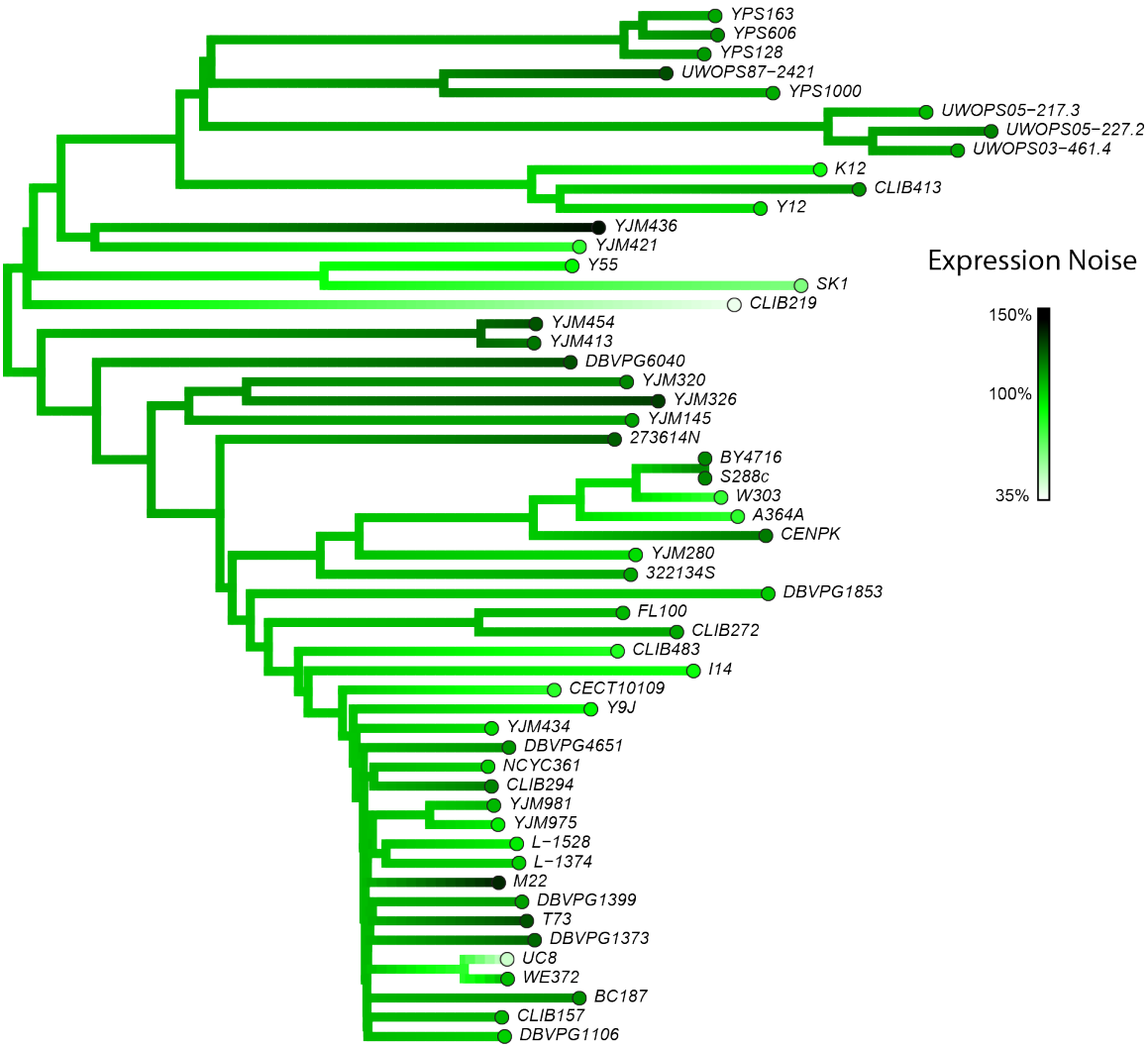


Figure 4-4 Effects of *trans*-regulatory polymorphisms on *TDH3* expression noise

Color of each circle at the ends of a branch reflects the observed effects of *trans*-regulatory polymorphisms on *TDH3* expression noise for each strain tested. Values are relative to those observed in S288c. Branch colors follow the same scale and were determined by reconstructing the most likely ancestral state at each node under a model of Brownian motion.

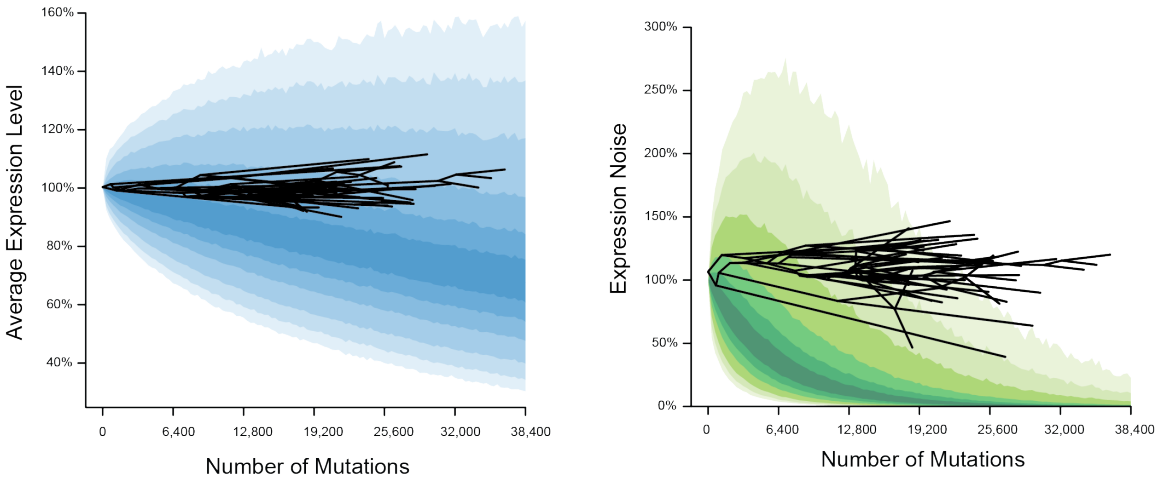


Figure 4-5 Natural selection on *TDH3* mean expression and expression noise

A. Null expectation of effects on average expression for *TDH3 trans*-regulatory polymorphisms after controlling for population structure if there was no natural selection (blue). Different colored regions indicate the 5th, 10th, 20th, 30th, 40th and 50th percentiles expected in expression in both directions from lightest to darkest. On top of this distribution is plotted the phylogenetic relationships amongst the 55 natural isolates. The effects of natural isolates fall outside of the highest probability region, yet are near the wild-type ancestral level of expression, suggesting stabilizing selection. B. Same as A, but for gene expression noise (Green). Natural isolates are higher than the null expectation suggesting selection to maintain *TDH3* expression noise.

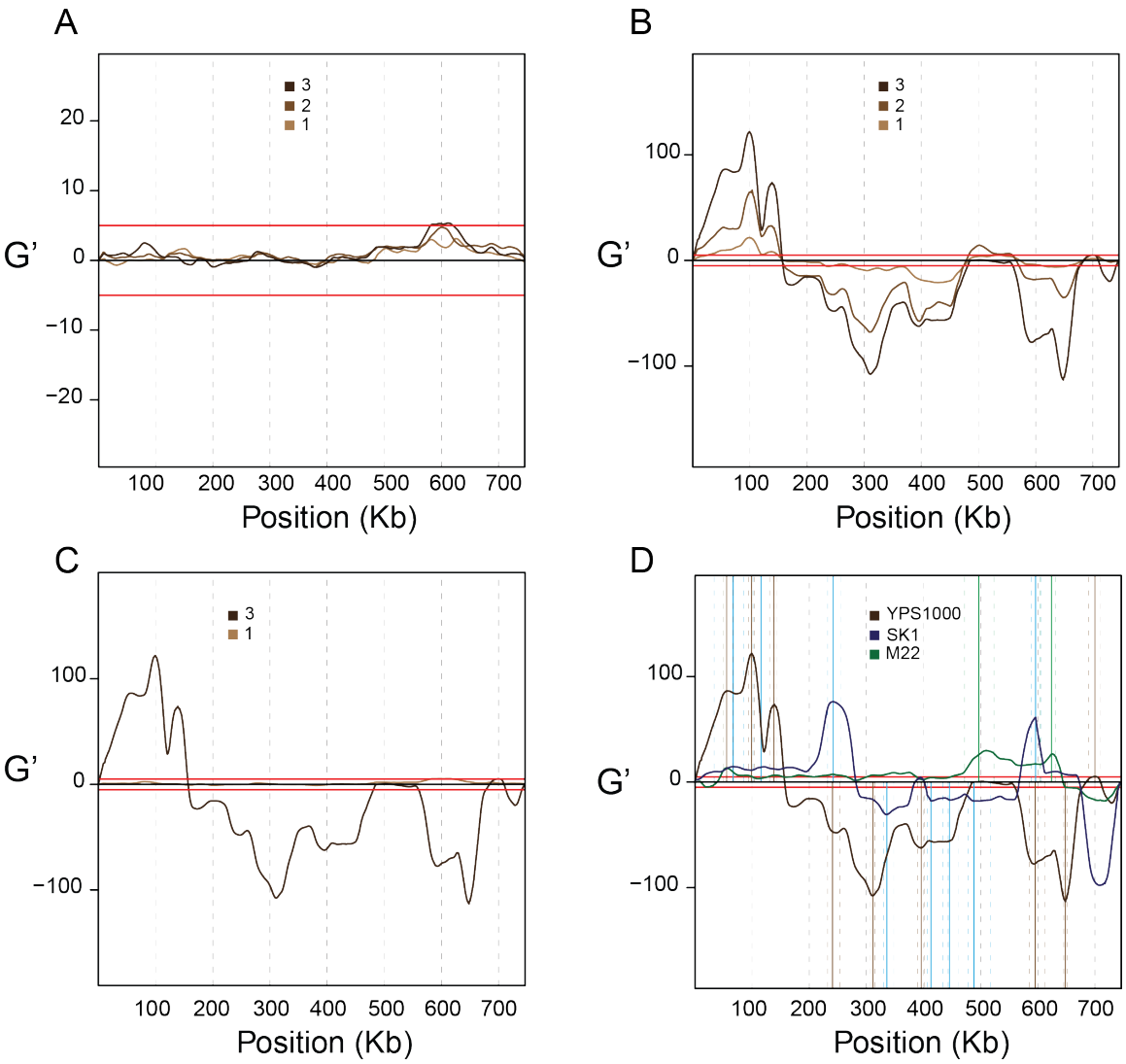


Figure 4-6 *trans*-regulatory QTL identification for *TDH3* expression

G' statistic quantifying allele frequency differences on chromosome 10 due to underlying QTL affecting *TDH3* expression. G' was calculated for each SNP using tri-cubic spline weights for all SNPs within 20 kb. Red lines show empirically derived cutoff from sample containing no QTL. A. G' values after a single round of meiosis between strains S288c and YPS1000 for either one (light brown), two (brown), or three (dark brown) rounds of selection on P_{TDH3}-YFP activity. B. G' values after three rounds of meiosis between strains S288c and YPS1000 for either one (light brown), two (brown), or three (dark brown) rounds of selection on P_{TDH3}-YFP activity. C. G' values after three rounds of selection on P_{TDH3}-YFP activity between strains S288c and YPS1000 for either one (light brown) or three (dark brown) rounds of meiosis. D. G' values and QTL locations (solid lines) with 95% confidence intervals (dashed lines) for all QTL identified on chromosome 10 between S288c and YPS1000 (brown), SK1 (blue), and M229 green).

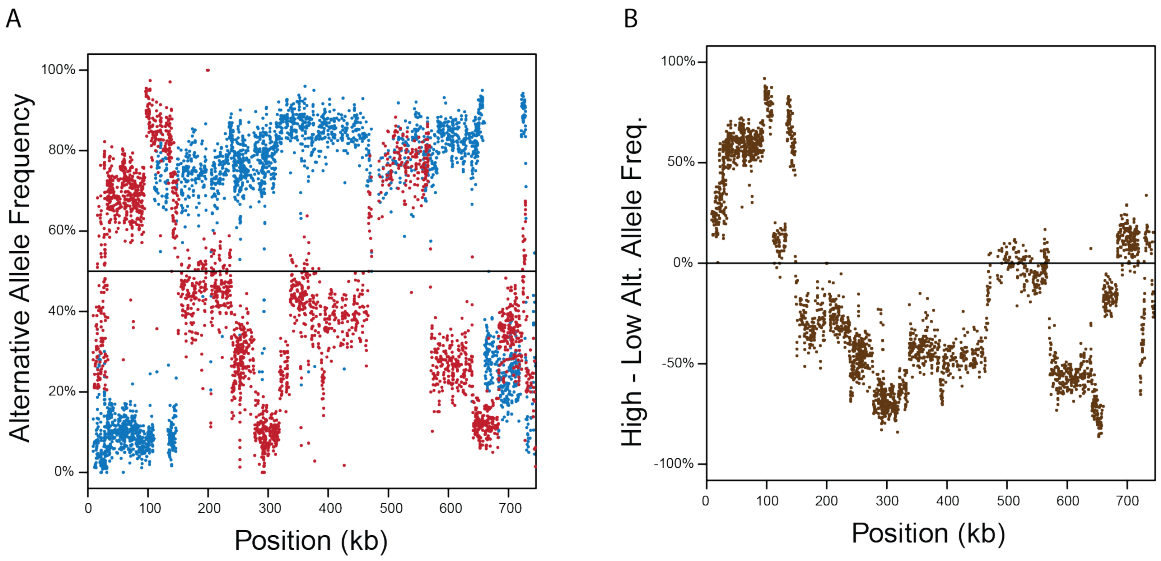


Figure 4-7 Allele frequency shifts from QTL mapping

Allele frequencies for chromosome 10. A. Each point represents the non-S288c allele frequency of a single SNP in a cross between strains S288c and YPS1000 after selecting for the 5% highest (red) or 5% lowest (blue) P_{TDH3}-YFP activity using FACS. Data shown is after three rounds of meiosis and three rounds of selection on P_{TDH3}-YFP activity for chromosome 10. B. Same as A, but showing the difference in non-S288c allele frequency between the population selected for high expression and the population selected for low expression. Discrete shifts in allele frequency suggest local linkage with many underling QTL.

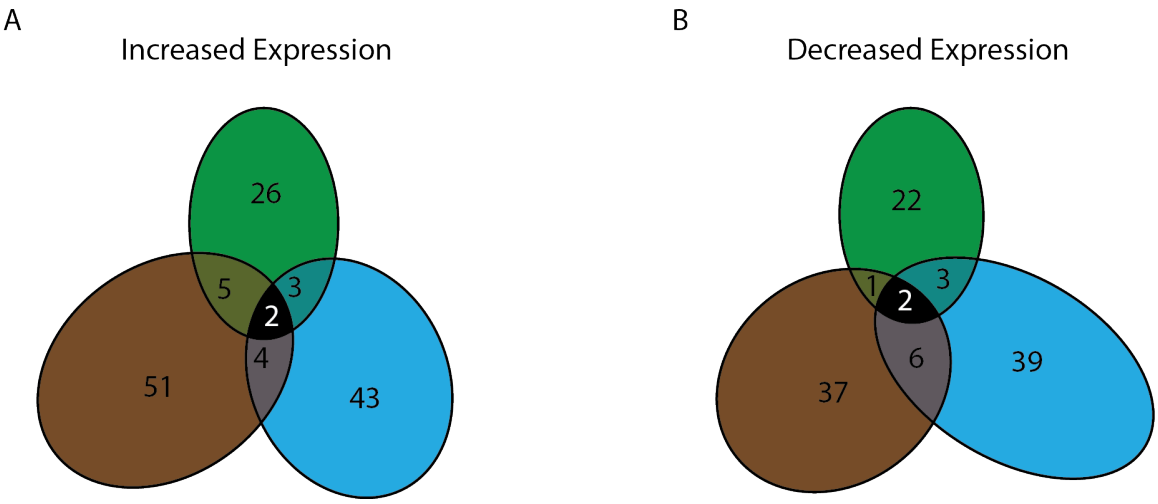


Figure 4-8 Number of *trans*-regulatory QTL identified

Number of *trans*-regulatory QTL identified between strains S288c and YPS1000 (brown), SK1 (blue), and M22 (green). Area is proportional to the number of QTL. Overlapping regions represent QTL in the same direction shared amongst strains. A. QTL where the non-S288c allele increases expression. B. QTL where the non-S288c allele decreases expression.

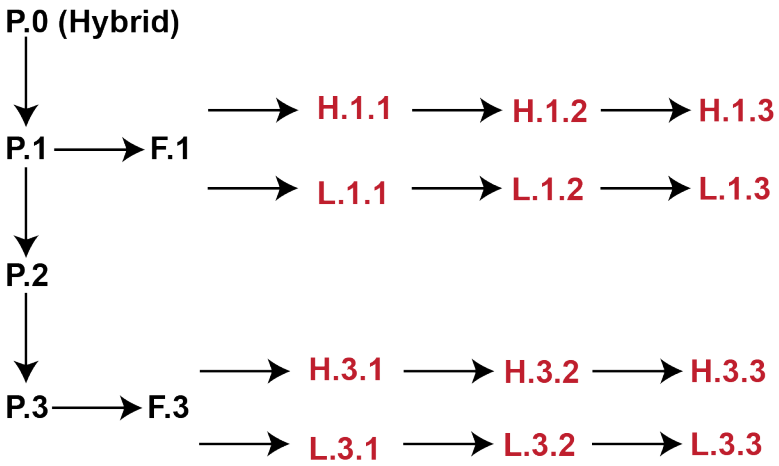


Figure 4-9 Pipeline used for mapping *TDH3 trans*-regulatory eQTL

For each cross between S228c and a natural yeast strain, the hybrid (P.0) was sporulated to generate thousands of unique genotypes (P.1). This population was mated and re-sporulated twice, producing populations P.2 and P.3. For P.1 and P.3, RFP minus MATα were selected using FACS (F.1 and F.3). Cells were then sorted using FACS based on YFP fluorescence, with 100,000 of either the 5% highest (H) or 5% lowest (L) individuals retained. This procedure was repeated a total of three times, generating a total of six comparisons for each cross. Allele frequencies for samples in red were determined using Illumina sequencing of pools. QTL were identified by differences in allele frequency between high selected and low selected pools. In addition, QTL were required to be present in at least two comparisons to be identified.

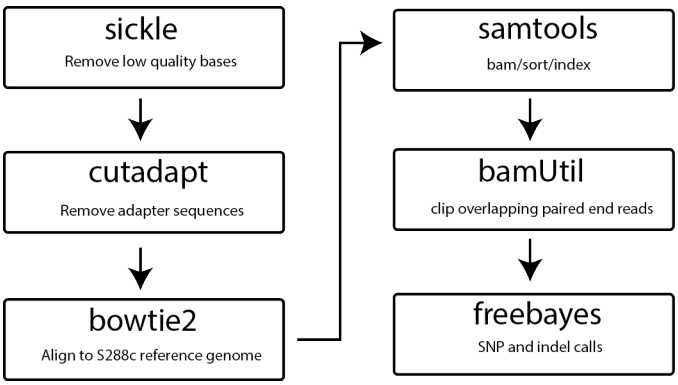


Figure 4-10 Bioinformatics pipeline used for mapping *TDH3 trans*-regulatory eQTL

Bioinformatics pipeline for identifying changes in allele frequency. All steps were performed independently on each sample except for SNP calling using freebayes, which jointly called SNPs in the high and low selected samples.

Strain Name	Collection Number
DBVPG6765	PJW1015
SK1	PJW1016
Y55	PJW1017
YPS128	PJW1018
DBVPG1373	PJW1019
DBVPG1853	PJW1020
YPS606	PJW1021
L-1374	PJW1022
L-1528	PJW1023
Y12	PJW1024
DBVPG1106	PJW1025
UWOPS83-787.3	PJW1026
UWOPS87-2421	PJW1027
NCYC361	PJW1028
322134S	PJW1029
273614N	PJW1030
YJM978	PJW1031
UWOPS03-461.4	PJW1032
UWOPS05-217.3	PJW1033
S288c	PJW1034
W303	PJW1035
UWOPS05-227.2	PJW1036
DBVPG6040	PJW1037
YIIc17_E5	PJW1038
YJM981	PJW1039
YJM975	PJW1040
NCYC110	PJW1041
CLIB272	PJW1042
YJM145	PJW1043

Strain Name	Collection Number
YJM280	PJW1044
YJM320	PJW1045
YJM326	PJW1046
YJM413	PJW1047
YJM421	PJW1048
YJM434	PJW1049
YJM436	PJW1050
YJM454	PJW1051
CECT10109	PJW1052
DBVPG3591	PJW1053
DBVPG4651	PJW1054
K12	PJW1062
YJM269	PJW1063
BY4716	PJW1064
A364A	PJW1065
CENPK	PJW1066
CLIB154	PJW1067
CLIB157	PJW1068
CLIB219	PJW1069
DBVPG1399	PJW1070
I14	PJW1071
M22	PJW1072
RM11	PJW1073
T73	PJW1074
UC8	PJW1075
WE372	PJW1076
Y9J	PJW1077
S288c Copy 2	PJW1078
BY4716 Copy 2	PJW1079

Table 4-1 Natural strains used for *trans*-regulatory polymorphism effects

Strain names and collection number for natural strain used to determine the effects of *trans*-regulatory polymorphism on *TDH3* expression.

Strain	Selection	Time point	i7 Index	i5 Index
PJW1057	L	3.1	N701	S502
PJW1057	H	3.1	N702	S502
PJW1016	L	3.1	N703	S502
PJW1016	H	3.1	N704	S502
PJW1072	L	3.1	N705	S502
PJW1072	H	3.1	N706	S502
PJW1057	L	3.2	N707	S502
PJW1057	H	3.2	N708	S502
PJW1016	L	3.2	N709	S502
PJW1016	H	3.2	N710	S502
PJW1072	L	3.2	N711	S502
PJW1072	H	3.2	N712	S502
PJW1057	L	3.3	N701	S508
PJW1057	H	3.3	N702	S508
PJW1016	L	3.3	N703	S508
PJW1016	H	3.3	N704	S508
PJW1072	L	3.3	N705	S508
PJW1072	H	3.3	N706	S508
PJW1057	L	1.1	N707	S508
PJW1057	H	1.1	N708	S508
PJW1016	L	1.1	N709	S508
PJW1016	H	1.1	N710	S508
PJW1072	L	1.1	N711	S508
PJW1072	H	1.1	N712	S508
PJW1057	L	1.2	N701	S517
PJW1057	H	1.2	N702	S517
PJW1016	L	1.2	N703	S517
PJW1016	H	1.2	N704	S517
PJW1072	L	1.2	N705	S517
PJW1072	H	1.2	N706	S517
PJW1057	L	1.3	N707	S517
PJW1057	H	1.3	N708	S517
PJW1016	L	1.3	N709	S517
PJW1016	H	1.3	N710	S517
PJW1072	L	1.3	N711	S517
PJW1072	H	1.3	N712	S517

Table 4-2 Barcode sequences used for *TDH3 trans*-regulatory eQTL mapping

Samples sequenced to determine allele frequency shifts. Selection refers to whether the sample was selected for either high (H) or low (L) expression. Time point indicates the number round of meiosis followed by the number of round of phenotypic selection. Barcodes names refer to those provided by Illumina.

Chapter 5

Contrasting frequencies and effects of *cis*- and *trans*-regulatory mutations on gene expression[1]

Abstract

Heritable differences in gene expression are caused by mutations in DNA sequences encoding *cis*-regulatory elements and *trans*-regulatory factors. These two classes of regulatory changes differ in their relative contributions to expression differences in natural populations. Here, we investigate how new mutations create the regulatory variation upon which natural selection acts by quantifying the frequencies and effects of new *cis*- and *trans*-acting mutations altering expression of a single gene. To do this, we measured the effects of 236 *cis*-regulatory mutations and >53,000 *trans*-regulatory mutations on expression of a reporter gene under control of the *Saccharomyces cerevisiae TDH3* promoter. We found that although *trans*-regulatory mutations were most common overall, *cis*- and *trans*-regulatory mutations were nearly equally abundant when only mutations with the largest effect sizes were considered. The relative frequencies of *cis*-

[1]This chapter will be submitted as: Brian P. H. Metzger*, Fabien Duveau*, David C. Yuan*, Stephen Tryban, Bing Yang, and Patricia J. Wittkopp. Contrasting frequencies and effects of *cis*- and *trans*-regulatory mutations on gene expression. *= Equal contributions

and *trans*-regulatory mutations differed significantly between mutations that increased or decreased gene expression as well, with *cis*-regulatory mutations tending to decrease expression and *trans*-regulatory mutations tending to increase expression. *cis*- and *trans*-regulatory mutations also differed in their effects on the variability in gene expression among genetically identical cells, a property of gene expression known as expression noise: *trans*-regulatory mutations were much more likely to decrease expression noise than *cis*-regulatory mutations. These differences in the frequencies and effects of *cis* and *trans*-regulatory mutations influence the raw material upon which natural selection acts and should therefore be taken into account in models of regulatory evolution.

Introduction

Variation in gene expression is a common source of phenotypic diversity within and between species (ZHENG *et al.* 2011). Much of this variation is heritable, arising from mutations in DNA sequences encoding *cis*-regulatory elements (e.g., promoters and enhancers) and *trans*-regulatory factors (e.g., transcription factors, non-coding RNAs, and signaling molecules) (STERN and ORGOGOZO 2008; CARROLL 2008). Studies investigating the genetic basis of intra- and inter-specific expression differences have shown that both *cis*- and *trans*-acting changes contribute to differences in gene expression, but the contributions of *cis*- and *trans*-acting loci are rarely equal (YVERT *et al.* 2003; GIBSON and WEIR 2005; ROCKMAN and KRUGLYAK 2006; GILAD *et al.* 2008; TIROSH *et al.* 2009; GONCALVES *et al.* 2012; SCHAEFKE *et al.* 2013; COOLON *et al.* 2014). Identifying specific genetic loci responsible for variation in gene expression has revealed

differences in both the relative frequency and average effects of *cis*- and *trans*-acting loci (GIBSON and WEIR 2005; ROCKMAN and KRUGLYAK 2006; GILAD *et al.* 2008).

These differences in the frequency and effects of *cis*- and *trans*-regulatory alleles result from the combined action of mutation, natural selection, and genetic drift, with new mutations generating the genetic variation upon which selection and drift act. Mutation accumulation experiments, in which mutations are allowed to accrue in the near absence of selection, have shown how gene expression levels change on a genomic scale in response to new mutations (DENVER *et al.* 2005; RIFKIN *et al.* 2005; LANDRY *et al.* 2007; SIMOLA *et al.* 2010; MCGUIGAN *et al.* 2014; HODGINS-DAVIS *et al.* 2015), but many questions remain about the impact of new mutations on *cis*- and *trans*-regulation. For example, what are the relative frequencies of new *cis*- and *trans*-regulatory mutations affecting expression of a particular gene? Do *cis*- and *trans*-regulatory mutations cause similar changes in a gene's expression? Are they equally likely to increase or decrease expression? Answering these questions will provide the empirical foundation needed to develop realistic neutral models of regulatory evolution that can be used to infer the impact of selection on regulatory variation observed in natural populations.

For any specific gene, *trans*-regulation is expected to have a larger mutational target size than *cis*-regulation because *cis*-regulatory mutations are typically limited to sequences close to the gene, whereas *trans*-regulatory mutations can be located anywhere in the genome. However, not all *cis* and *trans*-regulatory sites will affect the expression of a

focal gene when mutated. Instead, the subsets of *cis*- and *trans*-regulatory mutations that actually affect expression of a focal gene define realized target sizes (WITTKOPP 2005; LANG and MURRAY 2008) and it is these realized target sizes that are important for determining the relative contributions of *cis* and *trans*-regulatory changes to expression variation. While prior studies suggest that the realized target size for *trans*-regulatory mutations is larger than the realized target size for *cis*-regulatory mutations (DENVER *et al.* 2005; LANDRY *et al.* 2007; GRUBER *et al.* 2012), the relative magnitude of this difference remains uncertain.

In addition to differences in frequency, differences in the distributions of effects on gene expression for *cis* and *trans*-regulatory mutations are also important for determining the relative contributions of *cis* and *trans*-regulatory changes to expression variation. These differences include how much mutations alter gene expression level (magnitude of effects), and whether mutations increase or decrease expression (direction of effects). For example, studies of quantitative trait loci affecting gene expression (eQTL) have shown that *cis*-acting eQTLs tend to have larger magnitudes of effect on gene expression than *trans*-acting eQTL (SCHADT *et al.* 2003). One way this pattern could arise is if new *cis*-regulatory mutations have, on average, larger effects on expression than new *trans*-regulatory mutations. Differences in the direction of effects for *cis*- and *trans*-regulatory mutations are harder to predict and more likely to vary among genes, but could result from a difference in the relative frequency of activators and repressors between direct *trans*-regulatory factors (those that bind a gene's *cis*-regulatory sequences) and indirect

trans-regulatory factors (those that regulate a gene's expression by altering the abundance or activity of direct regulators).

Here, we compare the frequency and effects of *cis*- and *trans*-regulatory mutations altering expression of a focal gene using a reporter gene previously constructed for studying *cis*- (METZGER *et al.* 2015) and *trans*-regulatory (GRUBER *et al.* 2012) mutations in the baker's yeast *Saccharomyces cerevisiae*. We isolated 1485 new mutants with the potential to affect expression of this reporter gene in *trans* and compared them to a set of 235 mutants that each contained a single potential *cis*-regulatory mutation in the promoter of the reporter gene (METZGER *et al.* 2015). We also collected and analyzed 202 mutants enriched for rare *trans*-regulatory mutations with large effects on reporter gene expression. Each of these more than 1900 mutants was tested for its impact on expression level of the reporter gene as well as on the variability in reporter gene expression among genetically identical cells (expression noise). Heritable variation for gene expression noise has been described within natural populations and can also be subject to selection (ZHANG *et al.* 2009; WANG and ZHANG 2011; METZGER *et al.* 2015). We found that *cis*- and *trans*-regulatory mutations altering reporter gene expression differ in their frequency, magnitude of effect, and direction of effect for both mean expression level and expression noise. These differences in frequencies and effects indicate that the mutational input of *cis* and *trans*-regulatory changes to changes in regulation are unequal and should be considered in null models of regulatory evolution.

Results

Frequency and effects of cis- and trans-regulatory mutations on gene expression level

To quantify the effects of new *cis*- and *trans*-regulatory mutations on expression of a focal gene, we used a reporter gene (P_{TDH3}-*YFP)* containing the *S. cerevisiae TDH3* promoter (P_{TDH3}*)* and yellow fluorescent protein (YFP) coding sequence incorporated into the *S. cerevisiae* genome (GRUBER *et al.* 2012; METZGER *et al.* 2015). Effects of *cis*-regulatory mutations were determined by reanalyzing a published dataset describing the effects of 235 individual point mutations in the 678 bp *TDH3* promoter on YFP fluorescence (METZGER *et al.* 2015). Overall, *cis*-regulatory mutations were slightly more likely to decrease (130 of 235) than increase expression (105 of 235, p = 0.12, binomial test, Figure 5-1a), and *cis*-regulatory mutations that decreased expression had significantly larger effects on expression (measured as percent change in expression level) than *cis*-regulatory mutations that increased expression (p = 0.0001, Wilcoxon, Figure 5-1b). This difference resulted in a significant skew in the distribution of *cis*-regulatory effects towards decreased expression (skewness = -7.9, $p < 1 \times 10^{-15}$, D'Agostino). Mutations in known transcription factor binding sites (TFBS) for RAP1 and GCR1 (Figure F-1) contributed to this skew: 16 of 18 mutations in three previously identified TFBS decreased expression, including all mutations that decreased expression more 8% (Figure 5-1a). After excluding all mutations in known TFBS, *cis*-regulatory mutations that decreased expression still had significantly larger effects on gene expression than *cis*-regulatory mutations that increased expression (p = 0.008, Wilcoxon, Figure F-2a), indicating that the skew towards decreased expression is a general property of *cis*-regulatory mutations affecting activity of the *TDH3* promoter (skewness=-1.3,

$p=2x10^{-10}$, D'Agostino). Relative effects of *cis*-regulatory mutations were robust to changes in the reporter gene insertion site, genetic background, and even fusion of YFP to the *TDH3* coding sequence (Figure F-3).

To determine the effects of *trans*-regulatory mutations on P_{TDH3}-*YFP* expression, we generated mutations throughout the genome using a low dose of ethyl methanesulfonate (EMS) that introduced approximately 32 (21-43, 95% CI) mutations per cell. We then randomly collected 1485 mutant cells from the EMS treated population. Because the potential target size for *trans*-regulatory mutations is ~18,000-times larger than the potential target size for *cis*-regulatory mutations (12.1 Mb *S. cerevisae* genome vs 678 bp *TDH3* promoter), we expect that *cis*-regulatory mutations make a negligible contribution to changes in expression amongst these mutants and attributed all effects to *trans*-regulatory mutations. For each mutant isolated, we used flow cytometry to measure YFP fluorescence in four biological replicate populations containing ~5000 cells each and compared this measure of gene expression to YFP fluorescence of non-EMS treated controls. This approach to measurement and analysis is comparable to that used for the *cis*-regulatory mutants described above (METZGER *et al.* 2015). In contrast to the *cis*-regulatory mutants we examined, we found no significant difference in the number of *trans*-regulatory mutants that increased (n= 747) or decreased (n = 738) P_{TDH3}-*YFP* expression (p = 0.84, binomial test, Figure 5-1c), nor any significant difference in the magnitude of effects between *trans*-regulatory mutants that increased or decreased expression (p= 0.68, Wilcoxon, Figure 5-1d).

Next, we compared the relative effect sizes of the 235 *cis*-regulatory mutants and the 1485 *trans*-regulatory mutants by examining the absolute value of effects on gene expression. We found that the *cis*-regulatory mutants had significantly larger effects on gene expression than *trans*-regulatory mutants ($p = 1 \times 10^{-21}$, Wilcoxon, Figure 5-2a). This result was robust to removing the large effects of *cis*-regulatory mutants with mutations in the known TFBS ($p = 2 \times 10^{-16}$, Wilcoxon, Figure F-2b). The larger effects of *cis*-regulatory mutants than *trans*-regulatory mutants was also observed when separately considering mutations that increased expression ($p = 2 \times 10^{-6}$, Wilcoxon, Figure 5-2b) or decreased expression ($p = 2 \times 10^{-17}$, Wilcoxon, Figure 5-2c). These results were also robust to removal of the large effects of mutations in known TFBS (Figure F-2c-d). However, differences in the percentage of potential *cis*- and *trans*-regulatory mutations sampled in our collections might affect these conclusions: 34.6% of the 678 sites in the *TDH3* promoter were mutated in our collection of *cis*-regulatory mutants, but only ~0.37% of sites in the ~12.1 Mb genome were mutated in our collection of *trans*-regulatory mutants and rare, large-effect *trans*-regulatory mutations might be missing in our collection. Consistent with this possibility, 5 of the 1485 *trans*-regulatory mutants we examined showed increases in P_{TDH3}-*YFP* expression that were distinctly higher than all other mutants (expression >7.5%, Figure 5-1c,d).

To better characterize *trans*-regulatory mutations with large effects on expression, we used fluorescence-activated cell sorting (FACS) to isolate 202 mutants with the highest

and lowest YFP fluorescence levels from a new EMS-treated population. When each of these 202 mutants was grown clonally and reanalyzed in four replicate populations, the average absolute effect on P_{TDH3}-*YFP* expression was higher than for the initial *trans*-regulatory mutants isolated irrespective of their effects on expression for both mutants that increased (Figure 5-2b) and decreased expression (Figure 5-2c), indicating that FACS selection enriched for *trans*-regulatory mutants with large effects. These *trans*-regulatory mutants enriched for large effects did not have significantly different effects from *cis*-regulatory mutants for mutants that increased expression ($p = 0.14$, Wilcoxon), but had significantly smaller effects than *cis*-regulatory mutants for mutants that decreased expression ($p=0.0002$, Wilcoxon, Figure 5-2). Interestingly, 11 of the 202 isolated mutants showed changes in P_{TDH3}-*YFP* that were distinct from the bulk of mutants in our original *trans*-regulatory mutant collection (Figure 5-1e). These 11 mutants all showed increases in P_{TDH3}-*YFP* expression greater than 7.5% (Figure 5-1c-f), whereas only a single *trans*-regulatory mutant passing our quality controls with a comparable decrease in expression was isolated in either screen. Reanalysis of all strains with absolute expression changes greater than 7.5%, regardless of whether they passed our quality controls, indicates that strains with large decreases in expression are typically artifacts, whereas strains with large increases in expression are reproducible (Figure F-7). As a consequence, there is a difference in the frequency of mutants with large increases versus large decreases in P_{TDH3}-*YFP* expression and *trans*-regulatory mutants are skewed towards increased expression (skewness = 2.2, $p<1x10^{-15}$, D'Agostino). This skew towards increased expression for *trans*-regulatory mutants affecting *TDH3* expression is

consistent with earlier reports (GRUBER *et al.* 2012) and contrasts with the skew of *cis*-regulatory mutations toward decreased expression (Figure 5-1a,b).

Using *t*-tests to compare expression between each mutant and a non-mutagenized control genotype showed that 77 of 235 (33%) potential *cis*-regulatory mutants and 240 of 1485 (15%) of potential *trans*-regulatory mutants had significant changes in P_{TDH3}-*YFP* expression at a threshold of p = 0.01. For *cis*-regulatory mutations, these data suggest a realized target size of ~224 bp (0.33*678 bp in *TDH3* promoter). If we assume that genotypes with wild-type expression do not harbor mutations affecting P_{TDH3}-*YFP* expression (i.e., no compensatory mutations), the number of *trans*-regulatory mutations per cell significantly affecting P_{TDH3}-*YFP* expression follows a Poisson distribution (GRUBER *et al.* 2012), and each mutant has ~32 mutations, then these data suggest *trans*-regulatory mutations have realized target size of ~61,000 bp (45,000-93,000 bp, 95% CI), which is 0.51% (0.38-0.77%, 95% CI) of the *S. cerevisiae* genome. In other words, using statistical significance to define functional *cis*- and *trans*-regulatory mutations, *trans*-regulatory mutations that impact expression of P_{TDH3}-*YFP* should arise ~274 times more often than *cis*-regulatory mutations that significantly impact expression of the P_{TDH3}-*YFP*.

The impact of a mutation on gene expression is expected to be monotonically related to its impact on fitness (REST *et al.* 2013), thus we also examined how estimates of the realized target size for *cis*- and *trans*-regulatory mutations changed when different minimum effect sizes were used to define functional *cis* and *trans*-regulatory mutations.

To do this, we calculated the number of mutations with effects on P_{TDH3}-*YFP* expression equal to or larger than a particular value for the range of effects observed. We did this separately using the *cis*-regulatory mutants, the original collection of *trans*-regulatory mutants, and the *trans*-regulatory mutants enriched for mutations of large effect, taking into account that each *trans*-regulatory mutant has approximately 32 mutations and that the second set of *trans*-regulatory mutations was artificially enriched for mutations of large effect. We found that the realized target size for both *cis*- and *trans*-regulatory mutations dropped rapidly as the magnitude of the cutoff increased, and, unsurprisingly, the choice of the specific cutoff used had a drastic effect on the estimated target sizes. In addition, we found that *trans*-regulatory mutations remained more common than *cis*-regulatory mutations for nearly all magnitudes and directions of effect. However, the relative frequencies of *cis*- and *trans*-regulatory mutations varied considerably over the range of effect sizes examined (Figure 5-2d). For example, *trans*-regulatory mutations were inferred to be ~10,000-times more common than *cis*-regulatory mutations among mutations that alter expression less than 1%, but only 10-times more common among mutations that alter expression more than 3%. The relative frequency of *cis*- and *trans*-regulatory mutations also depended strongly on the direction of effect. For example, *cis*-regulatory mutations resulting in more than a 3% increase in P_{TDH3}-*YFP* expression appear to be either incredibly rare or nonexistent, whereas over 100 *trans*-regulatory mutations that increase expression more than 3% are predicted to exist. By contrast, decreases in expression by more than ~6% were more likely caused by *cis*-regulatory mutations than *trans*-regulatory mutations. These relationships were robust to error in the estimated number of mutations in each *trans*-regulatory mutant (Figure F-4).

Frequency and effects of cis- and trans-regulatory mutations on gene expression noise

To compare the effects of new *cis*- and *trans*-regulatory mutations on gene expression noise, we calculated the coefficient of variation (σ/μ) in YFP fluorescence for each *cis*- and *trans*-regulatory mutant. For the 235 *cis*-regulatory mutants, we found that significantly more mutants showed increased expression noise (n = 208) than decreased expression noise (n = 27, $p = 8 \times 10^{-36}$, binomial test) (Figure 5-3a), consistent with a prior analysis of these data (METZGER *et al.* 2015). This difference remained after excluding *cis*-regulatory mutations in the known TFBS (n = 190, n = 27, $p = 2 \times 10^{-31}$, binomial test). We also found that *cis*-regulatory mutants with increased expression noise had larger effects than *cis*-regulatory mutants that decreased expression noise, regardless of whether mutations in known TFBS were included ($p = 2 \times 10^{-5}$, Wilcoxon, Figure 5-3b) or not ($p = 5 \times 10^{-5}$, Wilcoxon, Figure F-5), resulting in a significant skew towards increased gene expression noise for mutations in the *TDH3* promoter (skewness=11.0, p-value < $1 \times 10^{-15,}$ D'Agostino). By contrast, the randomly collected set of 1485 *trans*-regulatory mutants contained more mutants with decreased expression noise (n = 797) than increased expression noise (n = 688, p = 0.005, binomial test, Figure 5-3c). The subsets of *trans*-regulatory mutants that increased or decreased expression noise had similar magnitudes of effects however (p = 0.71, Wilcoxon, Figure 5-3d). Similar to the effects on mean expression level, *cis*-regulatory mutants had larger effects on gene expression noise than *trans*-regulatory mutants (p =0.02, Wilcoxon, Figure 5-4a). However, this was due to increases in expression noise (p=0.001, Wilcoxon), whereas for

decreases in expression noise, *trans*-regulatory mutants had larger effects on expression noise than *cis*-regulatory mutants ($p < 0.002$, Wilcoxon, Figure 5-4b,c).

Using *t*-tests to compare expression noise between each mutant and a non-mutant control genotype showed that 40 of 235 (17%) potential *cis*-regulatory mutants and 118 of 1485 (7%) of potential *trans*-regulatory mutants had significant changes in P_{TDH3}-*YFP* expression noise at a significance threshold of $p = 0.01$. Using the same approach to compare the frequency of *cis*- and *trans*-regulatory mutations described above for the mean expression level identified a realized target size for *cis*-regulatory mutations affecting gene expression noise of ~116 bp (0.17*678 bp in *TDH3* promoter) and a realized target size for *trans*-regulatory mutations affecting gene expression noise of 30,000 bp (22,000-45,000 bp, 95% CI). For both *cis*- and *trans*-regulatory mutations, the realized target sizes for expression noise are smaller than the realized target sizes for mean expression using a statistical cutoff based on p-values; however, differences in statistical power to detect significant changes in gene expression noise and mean expression level make it difficult to compare these estimates.

To compare the frequency of *cis*- and *trans*-regulatory mutants affecting *TDH3* expression noise without relying on statistical cutoffs, we used the same approach described above for mean expression level to calculate the realized target size for *cis* and *trans*-regulatory mutations affecting gene expression noise over a range of effect sizes. We found that for all magnitudes and directions of effects, *trans*-regulatory mutations

affecting expression noise were more common than *cis*-regulatory mutations affecting expression noise (Figure 5-4d). However, as with the mean level of expression, the relative frequency of *cis* and *trans*-regulatory mutations varied considerably. Among mutations that increased expression noise, *trans*-regulatory mutations were about 10-100 times more common than *cis*-regulatory mutations for most effect sizes. By contrast, *trans*-regulatory mutations were 1,000-10,000 times more frequent than *cis*-regulatory mutations for even moderate decreases in gene expression noise because very few *cis*-regulatory mutations that decrease expression noise were observed.

Relationship between mutational effects on gene expression level and gene expression noise

New regulatory mutations can alter both the mean level of expression and expression noise simultaneously. If the effects of new mutations on mean expression level are independent of the effects of mutations on expression noise, however, natural selection can act independently on these two traits. Previous work indicates that there is often a negative correlation between mean expression and expression noise across genes (BAR-EVEN *et al.* 2006) and for individual *cis*-regulatory mutations (HORNUNG *et al.* 2012; SHARON *et al.* 2014). However, this correlation is generally insufficient to explain the observed joint variation in mean expression and expression noise for new mutations (HORNUNG *et al.* 2012; SHARON *et al.* 2014; METZGER *et al.* 2015). In addition, the expected relationship between effects on mean expression level and expression noise is less clear for individual *trans*-regulatory mutations. To compare the relationship between

effects on mean expression level and expression noise for new *cis*- and *trans*-regulatory mutations, we used Principal Components Analysis (PCA) to determine the primary axes of variation for the *cis*- and *trans*-regulatory mutants separately (Figure 5-5). As expected, we found that for *cis*-regulatory mutants the primary axis of variation had a negative slope, indicating a negative correlation between mean expression and expression noise (angle of rotation = 101°, with a 99% CI of 98°-105° based on bootstrap analysis). For *trans*-regulatory mutants, the primary axis of variation had a slightly positive slope (angle of rotation = 91° with a 99% CI of 89°-93°), but this was not significantly different from the 90° expected if *trans*-regulatory mutations had independent effects on mean expression level and expression noise (p=0.09). The angle of rotation was significantly different for *cis* and *trans*-regulatory mutations ($p < 10^{-6}$, permutation test), indicating that the *cis*- and *trans*-regulatory mutations we examined had different relationships between their effects on mean expression level and expression noise. This difference in the relationship between effects on mean expression level and expression noise for *cis*- and *trans*-regulatory mutations may contribute to the different evolutionary fates of *cis*- and *trans*-regulatory mutations.

Discussion

This study reveals many differences between *cis*- and *trans*-regulatory mutations that can impact their likelihood of contributing to variation in gene expression within and between species. For example, we find that *cis*- and *trans*-regulatory mutations differ significantly in their effects on both *TDH3* mean expression level and expression noise, with *cis*-

regulatory mutations skewed towards decreased expression and increased expression noise, while *trans*-regulatory mutations are skewed towards increased expression and often decreased expression noise. The relative frequencies of *cis*- and *trans*-regulatory mutations also differ, but the difference in frequencies depends upon the effect size used to define functional *cis*- and *trans*-regulatory mutations. For example, if only the largest changes in activity of the *TDH3* promoter are considered, then the target size for both *cis*- and *trans*-regulatory mutations is in the dozens or hundreds of bases, with *trans*-regulatory mutations more frequent for increases in expression and *cis*-regulatory mutations more frequent for decreases in expression. By contrast, if both small and large changes in P_{TDH3} activity are considered, then the *trans*-regulatory target size for altering P_{TDH3} activity is orders of magnitude larger than the *cis*-regulatory target size, and most new mutations resulting in biologically meaningful changes in *TDH3* transcription will be *trans*-regulatory. To the best of our knowledge, these data provide the first systematic comparison of *cis*- and *trans*-regulatory mutations affecting expression of a focal gene in any eukaryote.

Consequences of mutation type on mutational distributions

All of the *cis*-regulatory mutations examined in this study were G→A and C→T transitions. The majority of *trans*-regulatory mutations examined were expected to be G→A and C→T transitions as well because EMS introduces these types of changes almost exclusively (FLIBOTTE *et al.* 2010; DUVEAU *et al.* 2014). G→A and C→T transitions are also the most common spontaneous point mutations in *S. cerevisiae* (ZHU *et al.* 2014) and the most common polymorphisms segregating among natural populations

of *S. cerevisiae* (MACLEAN *et al.* 2015); however, they are still a minority of all point mutations, comprising only ~35% of all spontaneous point mutations (ZHU *et al.* 2014) and ~37% of single nucleotide polymorphisms segregating in *S. cerevisiae* (MACLEAN *et al.* 2015). So how representative are the effects of G→A and C→T transitions of all point mutations? And how might the effects of point mutations differ from other types of mutations such as insertions or deletions (indels), rearrangements, or copy number variants (CNVs)?

For *cis*-regulatory sequences, prior work suggests that there are no systematic differences in the effects of different classes of point mutations on gene expression (PATWARDHAN *et al.* 2009, 2012; KWASNIESKI and MOGNO 2012; MELNIKOV *et al.* 2012; METZGER *et al.* 2015). Nevertheless, we might have overestimated the average effect of *cis*-regulatory mutations in the *TDH3* promoter because mutations in known TFBS, which had the largest effects on expression, were overrepresented in our dataset (4% of total sequence, but 8% of mutations) due to their higher GC content (18/27 bp, 67%) relative to the rest of the promoter (~35%). *cis*-regulatory mutations had larger effects than *trans*-regulatory mutations even after excluding mutations in the TFBS, however, suggesting this overrepresentation does not alter our conclusions. Another consequence of mutating only Gs and Cs is that we failed to mutate functional elements composed of only As and Ts. The canonical TATA box contained within the *TDH3* promoter is one clear example of this. This sequence is a key determinant of both mean expression level and gene expression noise (RASER and O'SHEA 2004; HORNUNG *et al.* 2012), suggesting that

mutations within it would have had large effects that further increased the magnitude of *cis*-regulatory mutations we observed.

The impact of using EMS to introduce *trans*-regulatory mutations differs for coding and noncoding sequences of *trans*-acting factors. For mutations in noncoding *cis*-regulatory sequences of *trans*-acting factors, the consequences of using EMS are expected to be similar to the consequences of mutating only Gs and Cs in the *TDH3* promoter: potentially large effect mutations in functional elements that are AT rich such as TATA boxes might be missed, but the distribution of mutational effects should otherwise be unbiased. By contrast, in coding regions, which make up ~73% of the *S. cerevisiae* genome (ALEXANDER *et al.* 2010), using EMS to introduce mutations is expected to result in a biased sampling of amino acid changes because of the genetic code and codon usage (Figure F-6). G→A and C→T transitions are the most common type of spontaneous mutation, thus this bias should be toward the same types of amino acids changes caused most often by spontaneous mutations; however, some amino acids (Alanine, Glycine, and Proline) cannot be created by these types of mutations, whereas other amino acids (Asparagine, Isoleucine, Lysine, and Phenylalanine) and stop codons cannot be mutated. Some amino acid changes are more likely to disrupt protein function than others, resulting in differences in their magnitude of mutational effects (YAMPOLSKY and STOLTZFUS 2005). EMS-induced mutations and spontaneous mutations are also expected to differ in the relative frequency of synonymous and non-synonymous mutations: approximately 23% of naturally occurring point mutations in coding regions

are expected to be synonymous, compared to 31% for EMS-induced mutations (see Materials and Methods). Because synonymous mutations typically have smaller effects than non-synonymous mutations, the use of EMS might underestimate the effects of new *trans*-regulatory mutations in coding regions. To determine the full consequence of these biases in the distribution of mutational effects, the impact of other types of point mutations should be examined in future work.

The effects of insertions and deletions (indels), large-scale genome rearrangements, and copy number variants (CNVs) must also be considered to fully describe a mutational distribution. Point mutations are more common than these other types of mutations, but indels, rearrangements, and CNVs are known to contribute to variable gene expression in natural yeast populations (GERSTEIN *et al.* 2014) and experimentally evolved yeast populations (DUNHAM *et al.* 2002; KAO and SHERLOCK 2008; PAYEN *et al.* 2013; HOSE *et al.* 2015; SUNSHINE *et al.* 2015). For example, a prior study characterizing *trans*-regulatory mutations affecting activity of the *TDH3* promoter identified 22 spontaneous CNVs (GRUBER *et al.* 2012). In most cases, expression was increased substantially, suggesting that CNVs have distinct effects on expression from *cis*- and *trans*-regulatory changes caused by point mutations (GRUBER *et al.* 2012). The effects on gene expression for other types of CNVs (including deletions and aneuploidies of other chromosomes) are less clear, but may often be larger than for point mutations (e.g. Sunshine et al. 2015). Indels are also expected to have large effects when they occur in *trans*-acting coding regions because many will alter the reading frame and create non-functional proteins.

Together, these observations suggest that the mutational distributions we measured may be missing rare, large, effect mutations. Ultimately, the impact of indels, rearrangements, and CNVs on the distribution of mutational effects will need to be determined empirically.

Generality of the gene studied

In eukaryotes, transcription of each gene is controlled by biochemical interactions among many *trans*-acting factors that culminate in the direct binding of specific transcription factors to *cis*-regulatory sequences (WRAY *et al.* 2003). Our observation that *trans*-regulatory mutations are more common overall than *cis*-regulatory mutations is thus likely to be true for most genes, as is the observation that *cis*-acting mutations tend to have larger effects than *trans*-acting mutations (SCHADT *et al.* 2003). Other properties we report, however, such as the relative frequency of mutations that increase or decrease gene expression level or expression noise or the relative frequency of *cis*- and *trans*-regulatory mutations with particular effect sizes, are expected to vary among genes.

TDH3 encodes a glyceraldehyde-3-phosphate dehydrogenase that is involved in both glycolysis (MCALISTER and HOLLAND 1985) and chromatin remodeling (RINGEL *et al.* 2013). *TDH3* expression is regulated (at least in part) by binding sites for the RAP1 and GCR1 transcription factors in its promoter (BAKER *et al.* 1992; YAGI *et al.* 1994). Most mutations in these TFBS caused *TDH3* promoter activity to decrease, consistent with RAP1 and GCR1 activating *TDH3* expression. Other genes involved in glycolysis are

also regulated by RAP1 and GCR1 (CHAMBERS *et al.* 1995; UEMURA and FRAENKEL 2000; LIEB *et al.* 2001) and these genes potentially have similar distributions of effects for *cis*-regulatory mutations. Mutations in TFBS appear to often have the largest effects on gene expression (PATWARDHAN *et al.* 2012), suggesting that the density of TFBS within a promoter will strongly influence the distribution of effects for its *cis-regulatory* mutations. Such mutations are not expected to always decrease expression, however; loss-of-function mutations in TFBS for repressors should increase promoter activity and can thus potentially skew the effects of *cis*-regulatory mutations towards increased expression. Outside of the known TFBS, we found that *cis*-regulatory mutations tended to have small effects that were equally likely to increase or decrease expression. It remains to be seen if these mutations are disrupting unidentified binding sites for activators and repressors or simply altering chromatin structure in general ways that impact expression (VOSS and HAGER 2014).

Because *TDH3* is one of the most highly expressed proteins in the yeast genome (NEWMAN *et al.* 2006), we found the skew towards increased expression of *trans*-regulatory mutations affecting activity of the *TDH3* promoter surprising. This skew reflects the near absence of mutations decreasing reporter gene expression more than 7.5% despite the presence of multiple mutations causing expression to increase by this magnitude. Because most new *trans*-regulatory mutations are expected to disrupt activity of a *trans*-acting factor, we interpret the large frequency and magnitude of effects seen for *trans*-regulatory mutations that increase P_{TDH3} activity as an indication that repressors

play a major role in the regulation of *TDH3* expression. Because *cis*-regulatory mutations are skewed towards decreased expression, this additionally suggests that the *trans*-regulatory mutants we examined affect regulators that do not bind directly to the *TDH3* promoter.

The absence of *trans*-regulatory mutants with large decreases in P_{TDH3}-*YFP* expression (even after we selected specifically for them) could result from *trans*-regulatory mutations causing strong decreases P_{TDH3} activity being nonexistent; however, our understanding of *TDH3 cis*-regulatory sequences suggests that this is not the case: mutations in the RAP1 and GCR1 binding sites of P_{TDH3} caused large decreases in P_{TDH3} activity, suggesting that *trans*-acting mutations eliminating or significantly diminishing the function of RAP1 or GCR1 should also cause large decreases in P_{TDH3} activity. The absence of these mutants might be explained by the low fitness of *RAP1* and *GCR1* null mutants (GIAEVER *et al.* 2002), with cells carrying such mutations either dying or being out-competed by other genotypes during the ~10 generations of growth between the introduction of mutations and the isolation of individual mutant genotypes. Low fitness of such mutants could be due to their effects on *TDH3* expression, pleiotropic effects on activity of other genes, or some combination of both. In media containing glucose, null mutations in *TDH3* have much smaller fitness consequences than null mutations in *RAP1* or *GCR1* (BAKER *et al.* 1992; GIAEVER *et al.* 2002), suggesting that pleiotropy contributes to the low fitness of these mutants. The inability to recover lethal and nearly lethal mutations in studies of mutational effects such as ours is expected to have minimal

impact on the utility of these distributions for making predictions about patterns of evolutionary change since mutations causing very low fitness are also expected to be short-lived in natural populations.

Consequences of mutational properties for the evolution of gene expression

How does gene expression evolve in natural populations? And what are the forces most often responsible for shaping the patterns of regulatory divergence observed within and between species? These questions are difficult to answer, in part because we currently lack realistic null models of regulatory evolution. Generating the data needed to construct such neutral models was one of the primary goals for characterizing the frequencies and effects of new *cis*- and *trans*-regulatory mutations. Comparing regulatory evolution expected in the absence of natural selection to patterns of regulatory variation observed in natural populations can be a powerful approach to detecting natural selection and elucidating the underlying forces responsible for regulatory evolution (DENVER *et al.* 2005; RICE and TOWNSEND 2012; SMITH *et al.* 2013; METZGER *et al.* 2015). While it is not yet possible to model all aspects of regulatory evolution, the general properties we observed can be qualitatively compared to patterns of *cis* and *trans*-regulatory divergence observed in natural populations to begin disentangling the contributions of mutation and selection to the evolution of gene expression.

Using statistical significance as a cut-off for defining functional *cis*- and *trans*-regulatory mutations affecting P_{TDH3} activity, we found that new *trans*-regulatory mutations

occurred ~275 times more often than new *cis*-regulatory mutations for mean expression level and ~260 times more often for expression noise. When mutations with smaller effects (<1%) were also considered, we found that *trans*-regulatory mutations were as much as 10,000 times more common than *cis*-regulatory mutations for both properties of gene expression. These observations suggest that *trans*-regulatory mutations should be the predominant source of polymorphic expression for species where genetic variation is thought to largely reflect neutral processes. Indeed, many studies of intraspecific expression differences have found that *trans*-regulatory changes are the primary source of regulatory variation (WITTKOPP *et al.* 2008; LEMOS *et al.* 2008; EMERSON *et al.* 2010; SCHAEFKE *et al.* 2013; COOLON *et al.* 2014, 2015).

Between species, *cis*-regulatory changes appear to play a larger role (WITTKOPP *et al.* 2008; TIROSH *et al.* 2009; COOLON *et al.* 2014). The preferential fixation of *cis*-regulatory changes over evolutionary time can be caused by reduced purifying selection on *cis*-regulatory mutations compared to *trans*-regulatory mutations due to fewer pleiotropic constraints (WRAY *et al.* 2003; CARROLL 2005; STERN and ORGOGOZO 2008, 2009; CARROLL *et al.* 2013) and/or *cis*-regulatory mutations more frequently being the target of positive selection (FAY and WITTKOPP 2008; EMERSON *et al.* 2010; COOLON *et al.* 2014, 2015). The larger average effects we observed for *cis*-regulatory mutations should make them more likely to be eliminated when deleterious and fixed when advantageous, potentially explaining their greater contribution to divergence than polymorphism. The tendency of *cis*-regulatory mutations not to be recessive (LEMOS *et al.* 2008; GRUBER *et*

al. 2012) might also contribute to their more rapid elimination or fixation within a population.

Taken together, our data are consistent with a model of regulatory evolution in which the neutral process of mutation is primarily responsible for the abundant *trans*-regulatory variation observed within a species and natural selection is primarily responsible for the excess of *cis*-regulatory divergence observed between species. Such a model was also suggested and supported by comparisons of *cis*- and *trans*-regulatory polymorphisms and divergence using measures of allele-specific expression (WITTKOPP *et al.* 2008; EMERSON *et al.* 2010; COOLON *et al.* 2015). Future progress in understanding how gene expression evolves in natural populations will require building explicit models of regulatory evolution that account for differences mutational properties such as pleiotropy, effect size, and dominance between *cis* and *trans*-regulatory mutations as well as investigating how these biases are shaped by the action of natural selection. Additional information, such as estimates of pleiotropic effects of individual mutations and measures of their effects on relative fitness, is needed before such mature models can be developed.

Materials and Methods

Yeast strains

trans-regulatory mutations were created in strain YPW1139 (MATα). This strain is derived from BY4724, BY4722, BY4730 and BY4742 and contains no auxotrophies. In addition, this strain contains five mutations derived from natural yeast strains that fix two

defects in the common laboratory strains: high frequency of petites and low sporulation rate. The improved alleles introduced are RME1(ins-308A), TAO3(1493Q) from Deutschbauer and Davis 2005 and SAL1, CAT5-91M and MIP1-661T from Dimitrov et al. 2009. Finally, this strain contains a copy of the *TDH3* promoter, YFP coding sequence, CYC1 terminator, and KanMX4 drug resistance cassette inserted at the HO locus on chromosome IV (see Appendix C). *cis*-regulatory mutations were previously created in strain YPW1 (MAT**a**), as described in (METZGER *et al.* 2015). This strain is derived from BY4724 and is a *lys2- ura3-* auxotroph. YPW1 contains none of the changes made to strain YPW1139, but does carry the same P_{TDH3}-*YFP* reporter gene. This reporter gene is inserted on chromosome I near the SWH1 pseudogene instead of at the HO locus (GRUBER *et al.* 2012).

To determine whether the differences in genetic background or genomic insertion site of the reporter gene between YPW1139 and YPW1 altered the relative effects of *cis*-regulatory mutations on reporter gene expression, we compared the effects of 17 different *TDH3* promoter haplotypes on P_{TDH3}-*YFP* expression between the two strains (Table F-1). These haplotypes were chosen to capture the entire range of effects on reporter gene expression. We found that the relative effects of these haplotypes were well correlated between the two strains, with an R^2 of >0.96 for both mean expression level and expression noise (Figure F-3a,b). This strong correlation shows that differences between these two strains have a negligible effect on the relative effects on reporter expression for individual mutations. To account for absolute differences in expression between the two

strains, we used a linear model to estimate the slope of the relationship between reporter expression at the HO locus compared to reporter expression at the SWH1 locus for all 17 *TDH3* promoter haplotypes. We used this slope to correct the effects of all 235 *cis*-regulatory mutants to make their effects comparable to the effects measured for the *trans*-regulatory mutants.

To further determine whether the relative effects of *cis*-regulatory mutations measured using the P_{TDH3}-*YFP* reporter gene were representative of the relative effects of mutations in the native *TDH3* promoter, we introduced the same 17 haplotypes into the native *TDH3* promoter in a version of YPW1139 that lacked the P_{TDH3}-*YFP* reporter gene and instead contained the coding sequence for YFP added to the 3' end of the native *TDH3* coding sequence to produce a *TDH3*::YFP fusion protein at the native *TDH3* locus (YPW1452). For the effects of these *cis*-regulatory mutations on both mean expression level and expression noise, we observed a correlation coefficient of $R^2 > 0.99$ between the reporter gene in YPW1139 and the fusion protein in YPW1452 (Figure F-3c,d). Prior work has shown that fluorescence is highly correlated with protein level for fluorescent proteins ($R^2 = 0.81$, Kudla et al. 2009) and that the level of YFP fluorescence is a quantitative readout of *TDH3* promoter activity (GRUBER *et al.* 2012; DUVEAU *et al.* 2014; METZGER *et al.* 2015).

Mutagenesis

To generate *trans*-regulatory mutants, a low dose of Ethyl MethaneSulfonate (EMS) was used. The specific dose of EMS was chosen to maximize the proportion of cells with a single mutation that significantly altered reporter gene expression while maintaining a low proportion of cells with multiple mutations having significant effects on P_{TDH3}-*YFP* fluorescence (GRUBER *et al.* 2012). Genome sequencing and genetic mapping of genotypes isolated previously from populations treated similarly with EMS have confirmed that significant changes in P_{TDH3}-*YFP* expression are typically caused by only a single mutation (DUVEAU *et al.* 2014). In addition, sequencing the *TDH3* promoter in genotypes isolated from these previously EMS treated populations showed that less than 2% contain mutations in the *TDH3* promoter driving YFP fluorescence (GRUBER *et al.* 2012). These genotypes were isolated from the tails of the fluorescence distribution and were required to have statistically significant effects on YFP fluorescence before being sequenced, thus this 2% represents an upper bound on the frequency of *cis*-regulatory mutations in the EMS treated population expected in the current study.

Mutagenesis was performed on YPW1139 after reviving from a -80°C glycerol stock. All glycerol stocks were revived on YPG agar medium (10 g/L yeast extract, 20 g/L peptone, 5% v/v glycerol and 20 g/L agar) and grown for 48 hours at 30°C. Approximately 10^6 cells were then transferred to 10 mL of YPD liquid medium (10 g/L yeast extract, 20g/L peptone, 20g/L dextrose) and incubated for 24 hours at 30°C with 250 rpm shaking. After growth to a density of ~$7x10^7$ cells/mL, two aliquots of 1 mL were transferred to separate

micro-centrifuge tubes. Cells were washed twice in 1 mL H_2O and then suspended in 1 mL sodium phosphate 0.1M. 10 µL EMS 99% (Acros Organics) was added to one sample, and both samples (EMS-treated and control) were incubated for 45 minutes at room temperature. EMS mutagenesis was quenched by the addition of 1 mL sodium thiosulfate 5%, a treatment that was also applied to the control sample. Cells were pelleted and suspended twice in 1 mL sodium thiosulfate 5%, twice in 1 mL H_2O, and finally suspended in 1 mL YPD. 0.125 mL of each sample was then transferred to 3.875 mL YPD in a 15 mL culture tube that was incubated at 30°C for 24 hours. After growth to saturation, 0.125 mL of culture was diluted again for each sample to 3.875 mL YPD and grown for an additional 24 hours at 30°C to allow for ~10 generations of recovery after EMS treatment. A set of mutagen-treated and control cells were collected in each of two different experiments performed on separate days using the same protocol.

Measuring the mutation rate

After allowing cells to recover from EMS mutagenesis, the mutation rate was estimated for control and EMS-treated samples using a canavanine resistance assay (Gruber et al, 2012; Lang et al, 2008). Briefly, for each culture, 0.1 mL of a 10^{-1} dilution was plated on arginine dropout medium supplemented with 60 mg/mL L-canavanine sulfate (Sigma-Aldritch) and 0.1 mL of a 2×10^{-4} dilution of cells was plated on arginine dropout medium (6.7 g/L bacto-yeast nitrogen base, 20 g/L dextrose, 2 g/L drop-out mix minus arginine, 20 g/L agar). The number of colony forming units was counted on each medium after 48 h of growth at 30°C, and these counts were used to infer the proportion of

canavanine resistant cells in the initial cultures. Previous work has indicated that there are 88 EMS-like point mutations (G→A and C→T transitions) in the CAN1 gene that can result in canavanine resistance (LANG and MURRAY 2008). We calculated the average number of point mutations per cell as well as a 95% confidence interval for this average using the proportion of resistant colonies and this mutational target size, as described in Gruber et al. 2012.

Isolating mutants using FACS

After EMS recovery, individual EMS-treated and control cells were arrayed in a 384 well-plate layout on YPD agar plates using fluorescence activated cell sorting (BD FACS Aria III, University of Michigan Flow Cytometry Core). For each sample, 0.5 mL of saturated culture (~$7x10^7$ cells/mL) was mixed with 2 mL of PBS buffer and run on the FACS machine at a flow rate of ~15,000 cells/sec. Gating of flow cytometry events was based on width and forward scatter using FACSDiva software to avoid sorting non-yeast events or aggregates. For the first mutagenesis experiment, 1340 EMS-treated cells and 160 control cells were sorted irrespective of their fluorescence level onto five YPD agar plates.

Prior to the second mutagenesis experiment, two gates were set up corresponding to the 2^{nd} and 98^{th} percentiles of the YFP/FSC distribution (fluorescence divided by cell size) obtained from recording 10^6 non-mutant control cells. A total of 550 EMS-treated cells and 550 control cells were then sorted onto four YPD agar plates. For both EMS-treated

and control samples, 300 cells were sorted irrespective of their fluorescence level, 125 cells were sorted from the 2nd percentile gate (fluorescence lower than 98% of the control cells) and 125 cells were sorted outside of the 98th percentile gate (fluorescence higher than 98 % of the control cells). Considering the two mutagenesis experiments together, a total of 1640 EMS-treated cells and 460 control cells were collected irrespective of fluorescence level. In addition, 125 EMS-treated cells and 125 control cells were sorted from each of the 2% extreme tails of the control fluorescence distributions.

After sorting, cells were grown into colonies by incubating the plates for 48 hours at 30°C. Overall, no growth was observed at 6% of the positions, either due to the presence of a lethal mutation or because no cell was sorted. After growth, four quadrants of 96 colonies were transferred to four deep 96-well plates containing 0.5 mL YPD in each well using a V&P Scientific pin tool. Fresh YPW1139 cells that went through neither the mutagenesis procedure nor through a single cell bottleneck were revived from glycerol stocks and inoculated at 20 fixed positions on each plate. These samples were used to correct for position effects in the plate during subsequent flow cytometry experiments (note that these positions were left empty during cell sorting). After 24 hours of growth at 30°C, 100 µL of all cultures was mixed with 23 µL of glycerol 80% in sterile 96-well plates and kept frozen at -80°C. In parallel, all samples were transferred to YPG agar plates using the pin tool and grown for an additional 48 hours at 30°C. At this stage, 4.9 % of samples did not grow and were considered petites (cells lacking mitochondria). Ultimately, these procedures resulted in 1585 EMS-treated colonies sorted irrespective of

their YFP fluoresce, 202 EMS-treated colonies sorted from the 2% extreme tails of the control fluorescence distribution (99 from low fluorescence tail and 108 from high fluorescence tail), and 429 control colonies.

Quantifying fluorescence using flow cytometry

After growth on YPG agar plates, each sample was transferred to four replicate 96-well plates containing 0.5 mL of YPD for fluorescence quantification. Two replicates were inoculated 3 hours apart on two different days and grown for 22 hours at 30°C to saturation (~$7x10^7$ cells/mL). Cells were maintained in suspension during growth by the presence of a 3 mm glass bead in each well and by constant shaking at 250 rpm. Immediately prior to flow cytometry, 15 μL of each sample was diluted into 0.5 mL of PBS in a clean 96-well plate. Fluorescence was recorded for ~$2x10^4$ events per sample using a HyperCyt autosampler (Intellicyt Corp) coupled to a BD Accuri C6 instrument (488 nm laser used for excitation and 533/30 nm optical filter used for acquisition). For *cis*-regulatory mutants, flow cytometry data are publically available in the FlowRepository under Repository IDs FR-FCM-ZZBN. Flow cytometry data will be made publically available for the *trans*-regulatory mutants in the same manner.

Analysis of flow cytometry data

Flow cytometry data were analyzed with custom R scripts that are similar to those used in Metzger et al. 2015. Samples with less than 1000 events after removing budding cells and flow artifacts, or with median FSC more than three times the median average deviation,

were excluded. Strains with less than three replicates after removing poor samples were also excluded. Finally, strains with a standard deviation in mean expression greater than 0.1, a standard deviation in expression noise greater than 1, or a standard deviation in FSC (cell size) greater than 0.1, across replicates were removed. These later filters resulted in the removal of six strains from the *trans*-regulatory mutants.

Estimating mutational target sizes

Mutational target size was determined in two ways. First, *t*-tests were used to identify individual mutants with YFP fluorescence significantly different from non-mutant controls ($p < 0.01$). However, the power of this approach is dependent on sample size and does not necessarily reflect biological significance. As an alternative, we estimated mutational target size for all observed mutational effects. We first assumed that the potential target size was equal to the total number of bases available to be mutated. In addition, we assumed that the realized target size for mutations of no effect is the same as the potential target size. For *cis*-regulatory mutants, each strain carries only a single mutation and we attributed all differences in fluorescence between a mutant and control to the effect of that mutation. For each mutant we then counted the number of mutants with fluorescence equal to, or more extreme than, that specific mutants fluorescence. This was done separately for mutations that increase fluorescence and mutations that decrease fluorescence.

To estimate the target size for or *trans*-regulatory mutants, we used the same procedure with two modifications. First, we accounted for the fact that each *trans*-regulatory mutant contains multiple mutations. We assumed that all mutants that did not have effects greater than a specific cutoff did not have mutations with effects greater than that specific cutoff, i.e. we assumed large compensatory mutations were rare. We then assumed that the number of mutations with a specific effect within a mutant follows a Poisson distribution. Under this assumption, the fraction of mutants without an effect beyond a specific cutoff is proportional to the Poisson distribution rate parameter (fraction without effect = $e^{-\lambda}$). We used the estimated rate parameters to determine the number of mutants expected to have multiple mutations larger than a specific effect and correct for this bias. This bias is expected to be largest for mutants with small effects and nearly absent for mutants with large effects on YFP fluorescence.

Second, we accounted for the larger effects on expression expected when sorting of large effects. Given the distribution of effects on YFP fluorescence within a sample for non-mutant control cells and the specific cutoffs used for sorting, we calculated for each effect size the expected enrichment relative to an unsorted sample. We then used this enrichment to correct the target size estimates. We found that using this correction caused estimates of target size to be similar for trans-regulatory mutants sorted irrespective of their effects and trans-regulatory mutants sorted from the tails of the YFP florescence distribution, suggesting that it correctly accounted for the expected bias due to sorting from the extremes of the distribution. It also suggests that accurate estimates of target

size can be gained by collecting individuals with extreme phenotypes if appropriate corrections can be found.

Finally, to determine if the number of mutations expected in each *trans*-regulatory mutant altered estimates of the mutational target size, we performed the identical calculations but assuming the extremes of the 95% CI on the number of mutations within each *trans*-regulatory mutant (21 and 43).

Estimating target size and the percentage mutated

The percentage of the target size mutated was calculated as the number of sites mutated divided by the total number of possible sites that could have been mutated. For *cis*-regulatory mutants, there were 235 Gs and Cs in the 678 bp TDH3 promoter. For *trans*-regulatory mutants, we assumed that each mutant contained 32 mutations and overall there were 32*1485 = 47,520 individual mutations created. Assuming a genome size of 12,071,326, this represents less than 1% of the potential *trans*-regulatory target and the likelihood of the same mutation occurring twice is thus low.

Comparing effects of EMS and spontaneous mutations on amino acid sequences

To determine the expected frequency of amino acid changes in the absence of natural selection, we combined *S. cerevisiae* codon usage (GARDIN *et al.* 2014) with spontaneous point mutation rates determined from mutation accumulation assays (ZHU *et al.* 2014). To

determine the effects of restricting the type of mutations used to EMS like GC→AT transitions, we set the mutation rate for all other mutation types to zero and recalculated the expected frequencies of specific amino acid changes. Synonymous mutation rates were calculated as the percentage of mutations expected to result in the same amino acid assuming either natural mutation rates, or EMS like transitions only.

Statistical analyses

All statistical analyses, including calculating the angle of rotation from the principal components analysis, were performed in R (version 3.0.2, R Core Team 2013) using custom code and the following R packages: flowCore (HAHNE *et al.* 2009), flowClust (LO *et al.* 2009), mixtools (BENAGLIA *et al.* 2009), moments (KOMSTA, LUKASZ NOVOMESTKY 2015), and lawstat (HUI *et al.* 2008).

Author contributions

DCY, JG, and PJW designed the *cis*-regulatory mutational spectrum project. DCY created all initial *cis*-regulatory mutants and scored their fluorescence level. FD, BPHM, and PJW designed the *trans*-regulatory mutational spectrum project. FD collected all *trans*-regulatory mutants and scored their fluorescence level. ST and FD designed and determined the effects of *cis*-regulatory mutations at the alternative genomic location and BY and FD designed and determined the effects of *cis*-regulatory mutations at the native locus. JG determined copy number of *trans*-regulatory mutants. BPHM performed the

majority of computational analyses, with additional contributions from FD and DCY.

BPHM and PJW wrote the manuscript, with comments incorporated from all authors.

References

ALEXANDER R. P., FANG G., ROZOWSKY J., SNYDER M., GERSTEIN M. B., 2010 Annotating non-coding regions of the genome. Nat. Rev. Genet. **11**: 559–571.

BAKER H. V, HUIE M. A., SCOTT E. W., DRAZINIC C. M., LOPEZ M. C., HORNSTRA I. A. N. K., YANG T. P., BAKERL H. V, 1992 Characterization of the DNA-Binding activity of GCR1 : in vivo evidence for two GCR1-binding sites in the upstream activating sequence of TPI of Saccharomyces cerevisiae. Mol. Cell. Biol. **12**: 2690–2700.

BAR-EVEN A., PAULSSON J., MAHESHRI N., CARMI M., O'SHEA E., PILPEL Y., BARKAI N., 2006 Noise in protein expression scales with natural protein abundance. Nat. Genet. **38**: 636–43.

BENAGLIA T., CHAUVEAU D., HUNTER D. R., YOUNG D. S., 2009 mixtools: An R Package for Analyzing Finite Mixture Models. J. Stat. Softw. **32**: 1–29.

CARROLL S. B., 2005 Evolution at two levels: on genes and form. PLoS Biol. **3**: e245.

CARROLL S. B., 2008 Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell **134**: 25–36.

CARROLL S., LEE M., MARX C., 2013 Sign epistasis limits evolutionary tradeoffs at the confluence of single- and multi-carbon metabolism in Methylobacterium extorquens AM1. Evolution (N. Y). **2**: 1–40.

CHAMBERS a., PACKHAM E. a., GRAHAM I. R., 1995 Control of glycolytic gene expression in the budding yeast (Saccharomyces cerevisiae). Curr. Genet. **29**: 1–9.

COOLON J. D., MCMANUS C. J., STEVENSON K. R., GRAVELEY B. R., WITTKOPP P. J., 2014 Tempo and mode of regulatory evolution in Drosophila. Genome Res.: gr.163014.113–.

COOLON J., STEVENSON K., MCMANUS C., GRAVELEY B., PJ W., 2015 Molecular mechanisms and evolutionary processes contributing to accelerated divergence of gene expression on the Drosophila X chromosome. Mol. Biol. Evol.

DENVER D. R., MORRIS K., STREELMAN J. T., KIM S. K., LYNCH M., THOMAS W. K., 2005 The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nat. Genet. **37**: 544–548.

DEUTSCHBAUER A. M., DAVIS R. W., 2005 Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat. Genet. **37**: 1333–40.

DIMITROV L. N., BREM R. B., KRUGLYAK L., GOTTSCHLING D. E., 2009 Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. Genetics **183**: 365–83.

DUNHAM M. J., BADRANE H., FEREA T., ADAMS J., BROWN P. O., ROSENZWEIG F., BOTSTEIN D., 2002 Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. PNAS **99**: 16144–16149.

DUVEAU F., METZGER B. P. H., GRUBER J. D., MACK K., SOOD N., BROOKS T. E., WITTKOPP P. J., 2014 Mapping Small Effect Mutations in Saccharomyces cerevisiae: Impacts of Experimental Design and Mutational Properties. G3: 1205–1216.

EMERSON J. J., HSIEH L.-C., SUNG H.-M., WANG T.-Y., HUANG C.-J., LU H. H.-S., LU M.-Y. J., WU S.-H., LI W.-H., 2010 Natural selection on cis and trans regulation in yeasts. Genome Res.: 826–836.

FAY J. C., WITTKOPP P. J., 2008 Evaluating the role of natural selection in the evolution of gene regulation. Heredity (Edinb). **100**: 191–9.

FLIBOTTE S., EDGLEY M. L., CHAUDHRY I., TAYLOR J., NEIL S. E., ROGULA A., ZAPF R., HIRST M., BUTTERFIELD Y., JONES S. J., MARRA M. a, BARSTEAD R. J., MOERMAN D. G., 2010 Whole-genome profiling of mutagenesis in Caenorhabditis elegans. Genetics **185**: 431–41.

GARDIN J., YEASMIN R., YUROVSKY A., CAI Y., SKIENA S., FUTCHER B., 2014 Measurement of average decoding rates of the 61 sense codons in vivo. Elife **3**: 1–20.

GERSTEIN A. C., LO D. S., CAMPBELL M. L., KUZMIN A., OTTO S. P., ONO J., LO D. S., CAMPBELL M. L., KUZMIN A., OTTO S. P., 2014 Too Much of a Good Thing : The Unique and Repeated Paths Toward Copper Adaptation. Genetics **199**: 555–571.

GIAEVER G., CHU A. M., NI L., CONNELLY C., RILES L., VÉRONNEAU S., DOW S., LUCAU-DANILA A., ANDERSON K., ANDRÉ B., ARKIN A. P., ASTROMOFF A., EL-BAKKOURY M., BANGHAM R., BENITO R., BRACHAT S., CAMPANARO S., CURTISS M., DAVIS K., DEUTSCHBAUER A., ENTIAN K.-D., FLAHERTY P., FOURY F., GARFINKEL D. J., GERSTEIN M., GOTTE D., GÜLDENER U., HEGEMANN J. H.,

HEMPEL S., HERMAN Z., JARAMILLO D. F., KELLY D. E., KELLY S. L., KÖTTER P., LABONTE D., LAMB D. C., LAN N., LIANG H., LIAO H., LIU L., LUO C., LUSSIER M., MAO R., MENARD P., OOI S. L., REVUELTA J. L., ROBERTS C. J., ROSE M., ROSS-MACDONALD P., SCHERENS B., SCHIMMACK G., SHAFER B., SHOEMAKER D. D., SOOKHAI-MAHADEO S., STORMS R. K., STRATHERN J. N., VALLE G., VOET M., VOLCKAERT G., WANG C., WARD T. R., WILHELMY J., WINZELER E. a, YANG Y., YEN G., YOUNGMAN E., YU K., BUSSEY H., BOEKE J. D., SNYDER M., PHILIPPSEN P., DAVIS R. W., JOHNSTON M., 2002 Functional profiling of the *Saccharomyces cerevisiae* genome. Nature **418**: 387–91.

GIBSON G., WEIR B., 2005 The quantitative genetics of transcription. Trends Genet. **21**: 616–623.

GILAD Y., RIFKIN S. a., PRITCHARD J. K., 2008 Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. **24**: 408–415.

GONCALVES A., LEIGH-BROWN S., THYBERT D., STEFFLOVA K., TURRO E., FLICEK P., BRAZMA A., ODOM D. T., MARIONI J. C., 2012 Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res. **22**: 2376–2384.

GRUBER J. D., VOGEL K., KALAY G., WITTKOPP P. J., 2012 Contrasting Properties of Gene-specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects and Dominance. PLoS Genet. **8**: e1002497.

HAHNE F., LEMEUR N., BRINKMAN R. R., ELLIS B., HAALAND P., SARKAR D., SPIDLEN J., STRAIN E., GENTLEMAN R., 2009 flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics **10**: 106.

HODGINS-DAVIS A., RICE D. P., TOWNSEND J. P., 2015 Gene expression evolves under a House-of-Cards model of stabilizing selection. Mol. Biol. Evol.

HORNUNG G., BAR-ZIV R., ROSIN D., TOKURIKI N., TAWFIK D. S., OREN M., BARKAI N., 2012 Noise-mean relationship in mutated promoters. Genome Res. **22**: 2409–2417.

HOSE J., YONG C. M., SARDI M., WANG Z., NEWTON M. A., GASCH A. P., 2015 Dosage compensation can buffer copy-number variation in wild yeast. Elife **4**: 1–27.

HUI W., GEL Y., GASTWIRTH J., 2008 lawstat: an R package for law, public policy and biostatistics. J. Stat. Softw. **28**.

KAO K. C., SHERLOCK G., 2008 Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. Nat. Genet. **40**: 1499–504.

KOMSTA, LUKASZ NOVOMESTKY F., 2015 moments.

KUDLA G., MURRAY A., TOLLERVEY D., PLOTKIN J., 2009 Coding-sequence determinants of gene expression in Escherichia coli. Science **324**: 255–258.

KWASNIESKI J., MOGNO I., 2012 Complex effects of nucleotide variants in a mammalian cis-regulatory element. PNAS **109**: 19498–19503.

LANDRY C. R., LEMOS B., RIFKIN S. A., DICKINSON W. J., HARTL D. L., 2007 Genetic properties influencing the evolvability of gene expression. Science **317**: 118–121.

LANG G. I., MURRAY A. W., 2008 Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. Genetics **178**: 67–82.

LEMOS B., ARARIPE L. O., FONTANILLAS P., HARTL D. L., 2008 Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. PNAS **105**: 14471–6.

LIEB J. D., LIU X., BOTSTEIN D., BROWN P. O., 2001 Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat. Genet. **28**: 327–34.

LO K., HAHNE F., BRINKMAN R. R., GOTTARDO R., 2009 flowClust: a Bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics **10**: 145.

MACLEAN C. J., METZGER B. P. H., YANG J.-R., HO W.-C., MOYERS B., ZHANG J., 2015 Genome sequencing and high-throughput phenotypic analysis of diverse Saccharomyces cerevisiae strains. In Prep.

MCALISTER L., HOLLAND M. J., 1985 Differential expression of the three yeast glyceraldehyde-3-phosphate dehydrogenase genes. J. Biol. Chem. **260**: 15019–15027.

MCGUIGAN K., COLLET J. M., MCGRAW E. a., YE Y. H., ALLEN S. L., CHENOWETH S. F., BLOWS M. W., 2014 The Nature and Extent of Mutational Pleiotropy in Gene Expression of Male Drosophila serrata. Genetics **196**: 911–921.

MELNIKOV A., MURUGAN A., ZHANG X., TESILEANU T., WANG L., ROGOV P., FEIZI S., GNIRKE A., CALLAN C. G., KINNEY J. B., KELLIS M., LANDER E. S., MIKKELSEN T. S., 2012 Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. **30**: 271–279.

METZGER B. P. H., YUAN D. C., GRUBER J. D., DUVEAU F. D., WITTKOPP P. J., 2015 Selection on noise constrains variation in a eukaryotic promoter. Nature **521**: 344–347.

NEWMAN J. R. S., GHAEMMAGHAMI S., IHMELS J., BRESLOW D. K., NOBLE M., DERISI J. L., WEISSMAN J. S., 2006 Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature **441**: 840–846.

PATWARDHAN R. P., HIATT J. B., WITTEN D. M., KIM M. J., SMITH R. P., MAY D., LEE C., ANDRIE J. M., LEE S.-I., COOPER G. M., AHITUV N., PENNACCHIO L. a, SHENDURE J., 2012 Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. **30**: 265–270.

PATWARDHAN R. P., LEE C., LITVIN O., YOUNG D. L., PE'ER D., SHENDURE J., 2009 High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. **27**: 1173–1175.

PAYEN C., RIENZI S. C. DI, ONG G. T., POGACHAR J. L., SANCHEZ J. C., SUNSHINE A. B., RAGHURAMAN M. K., BREWER B. J., DUNHAM M. J., 2013 The Dynamics of Diverse Segmental Amplifications in Populations of Saccharomyces cerevisiae Adapting to Strong Selection. G3: 399–409.

R CORE TEAM, 2013 R: A language and environment for statistical computing.

RASER J. M., O'SHEA E. K., 2004 Control of stochasticity in eukaryotic gene expression. Science **304**: 1811–1814.

REST J. S., MORALES C. M., WALDRON J. B., OPULENTE D. a, FISHER J., MOON S., BULLAUGHEY K., CAREY L. B., DEDOUSIS D., 2013 Nonlinear fitness consequences of variation in expression level of a eukaryotic gene. Mol. Biol. Evol. **30**: 448–456.

RICE D. P. D., TOWNSEND J. P. J., 2012 A test for selection employing quantitative trait locus and mutation accumulation data. Genetics **190**: 1533–1545.

RIFKIN S. a, HOULE D., KIM J., WHITE K. P., 2005 A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438**: 220–223.

RINGEL A. E., RYZNAR R., PICARIELLO H., HUANG K., LAZARUS A. G., HOLMES S. G., 2013 Yeast Tdh3 (Glyceraldehyde 3-Phosphate Dehydrogenase) Is a Sir2-Interacting Factor That Regulates Transcriptional Silencing and rDNA Recombination (CS Pikaard, Ed.). PLoS Genet. **9**: e1003871.

ROCKMAN M. V, KRUGLYAK L., 2006 Genetics of global gene expression. Nat. Rev. Genet. **7**: 862–72.

SCHADT E. E. E., MONKS S. S. a, DRAKE T. T. a, LUSIS A. J. A., CHE N., COLINAYO V., RUFF T. G., MILLIGAN S. B., LAMB J. R., CAVET G., LINSLEY P. S., MAO M.,

STOUGHTON R. B., FRIEND S. H., 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **205**: 1–6.

SCHAEFKE B., EMERSON J. J., WANG T.-Y., LU M.-Y. J., HSIEH L.-C., LI W.-H., 2013 Inheritance of gene expression level and selective constraints on trans- and cis-regulatory changes in yeast. Mol. Biol. Evol. **30**: 2121–33.

SHARON E., DIJK D. VAN, KALMA Y., KEREN L., MANOR O., YAKHINI Z., SEGAL E., 2014 Probing the effect of promoters on noise in gene expression using thousands of designed sequences. Genome Res. **24**: 1698–1706.

SIMOLA D. F., FRANCIS C., SNIEGOWSKI P. D., KIM J., 2010 Heterochronic evolution reveals modular timing changes in budding yeast transcriptomes. Genome Biol. **11**: R105.

SMITH J., MCMANUS K., FRASER H., 2013 A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers. arXiv.

STERN D. L., ORGOGOZO V., 2008 The loci of evolution: how predictable is genetic evolution? Evolution (N. Y). **62**: 2155–2177.

STERN D. D. L., ORGOGOZO V., 2009 Is genetic evolution predictable? Science **323**: 746–751.

SUNSHINE A. B., PAYEN C., ONG G. T., LIACHKO I., TAN K. M., DUNHAM M. J., 2015 The Fitness Consequences of Aneuploidy Are Driven by Condition-Dependent Gene Effects. PLOS Biol. **13**: e1002155.

TIROSH I., REIKHAV S., LEVY A. a, BARKAI N., 2009 A yeast hybrid provides insight into the evolution of gene expression regulation. Science **324**: 659–662.

UEMURA H., FRAENKEL D. G., 2000 Glucose Metabolism in gcr Mutants of Saccharomyces cerevisiae. J. Bacteriol. **182**: 2354.

VOSS T. C., HAGER G. L., 2014 Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat. Rev. Genet. **15**: 69–81.

WANG Z., ZHANG J., 2011 Impact of gene expression noise on organismal fitness and the efficacy of natural selection. PNAS **108**: E67–E76.

WITTKOPP P. J., 2005 Genomic sources of regulatory variation in cis and in trans. Cell. Mol. Life Sci. **62**: 1779–83.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2008 Regulatory changes underlying expression differences within and between *Drosophila* species. Nat. Genet. **40**: 346–350.

WRAY G. a, HAHN M. W., ABOUHEIF E., BALHOFF J. P., PIZER M., ROCKMAN M. V, ROMANO L. A., 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**: 1377–1419.

YAGI S., YAGI K., FUKUOKA J., SUZUKI M., 1994 The UAS of the yeast GAPDH promoter consists of multiple general functional elements including RAP1 and GRF2 binding sites. J. Vet. Med. Sci. **56**: 235–244.

YAMPOLSKY L. Y., STOLTZFUS A., 2005 The exchangeability of amino acids in proteins. Genetics **170**: 1459–72.

YVERT G., BREM R. B., WHITTLE J., AKEY J. M., FOSS E., SMITH E. N., MACKELPRANG R., KRUGLYAK L., OTHERS, 2003 Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat. Genet. **35**: 57–64.

ZHANG Z., QIAN W., ZHANG J., 2009 Positive selection for elevated gene expression noise in yeast. Mol. Syst. Biol. **5**: 1–12.

ZHENG W., GIANOULIS T. a, KARCZEWSKI K. J., ZHAO H., SNYDER M., 2011 Regulatory variation within and between species. Annu. Rev. Genomics Hum. Genet. **12**: 327–346.

ZHU Y. O., SIEGAL M. L., HALL D. W., PETROV D. a, 2014 Precise estimates of mutation rate and spectrum in yeast. PNAS **111**: E2310–E2318.

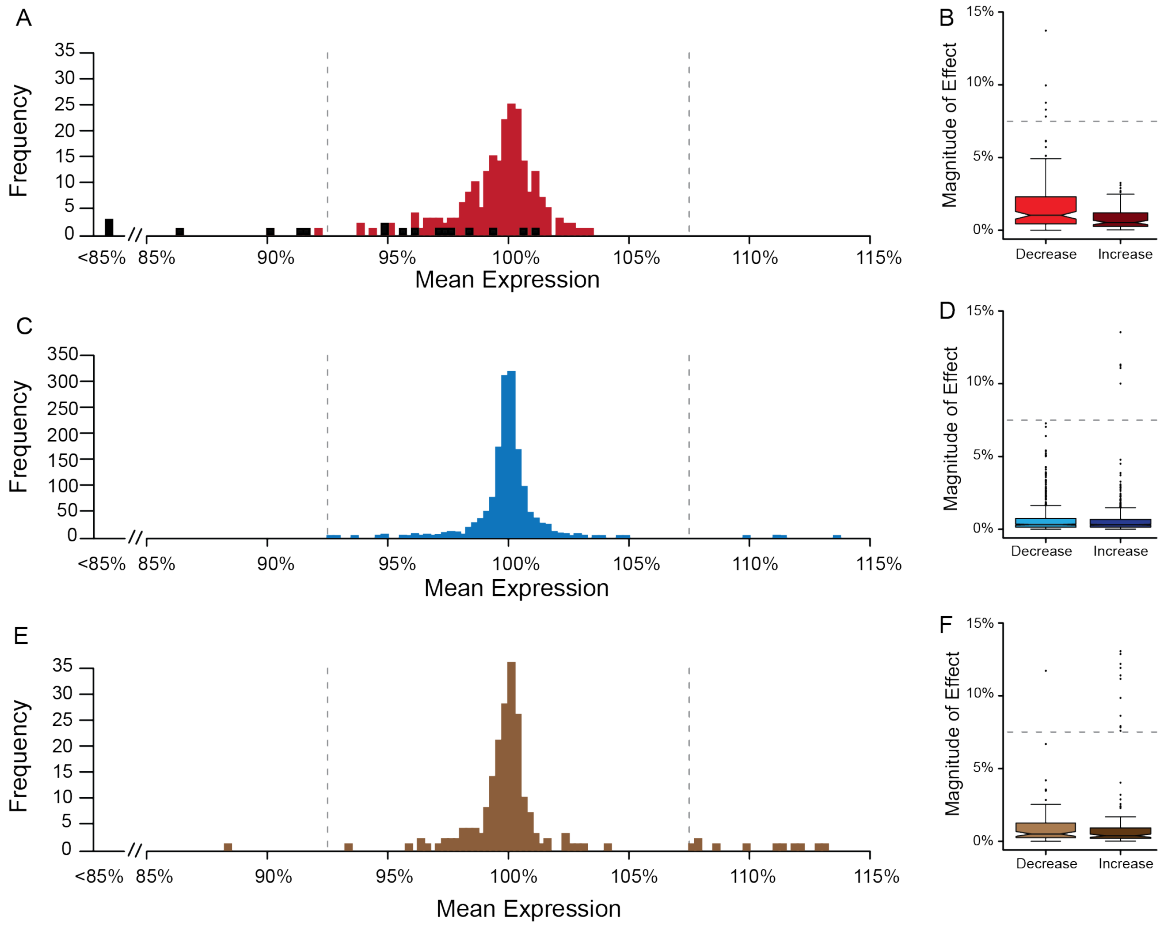


Figure 5-1 Frequency and effects of *cis*- and *trans*-regulatory mutants on mean P_{TDH3}-YFP fluorescence

(A) Frequency of *cis*-regulatory mutants with specific effects on P_{TDH3}-YFP fluorescence. Mutations outside of known TFBS (red). Mutations in known TFBS (black). (B) Effect of *cis*-regulatory mutants that increase (dark red) vs decrease (light red) P_{TDH3}-YFP fluorescence. (C) Frequency of *trans*-regulatory mutants with specific effects on P_{TDH3}-YFP fluorescence. (D) Effect of *trans*-regulatory mutants that increase (dark blue) or decrease (light blue) P_{TDH3}-YFP fluorescence. (E) Frequency of *trans*-regulatory mutants with specific effects on P_{TDH3}-YFP fluorescence for EMS induced mutants collected from the extremes of P_{TDH3}-YFP fluorescence. (F) Effect of extreme *trans*-regulatory mutants that increase (dark brown) or decrease (light brown) P_{TDH3}-YFP fluorescence.

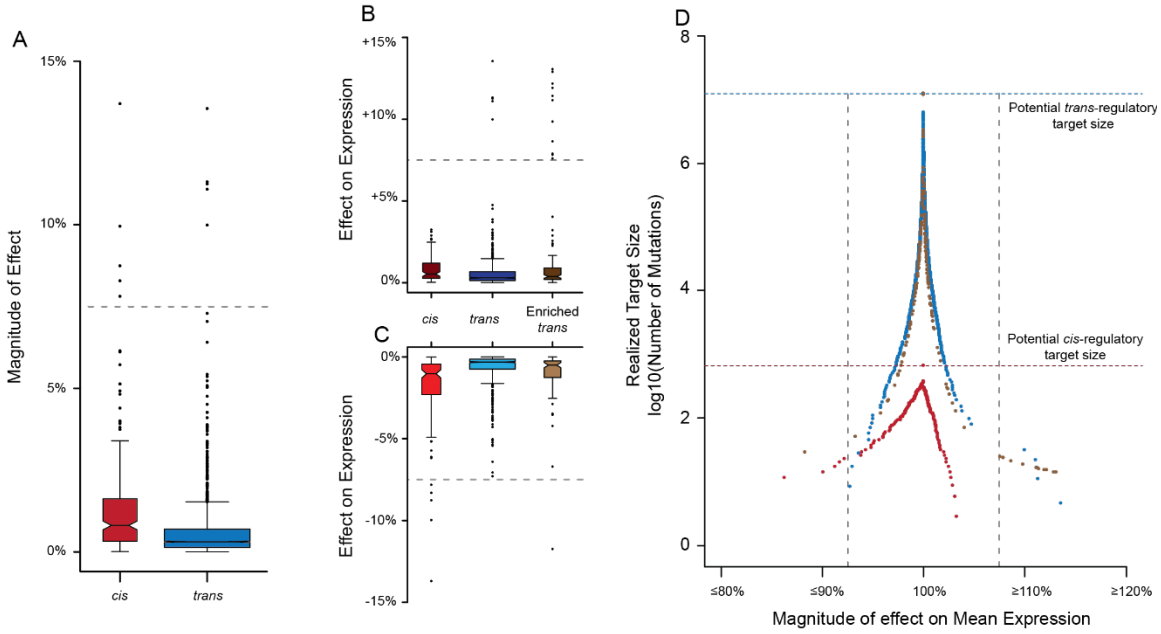


Figure 5-2 Comparison of frequency and effects for *cis*- and *trans*-regulatory mutants on mean P_{TDH3}-YFP fluorescence

(A) Absolute magnitude of effect on P_{TDH3}-YFP fluorescence for *cis*-regulatory mutants (red) and *trans*-regulatory mutants (blue). (B) Effect on P_{TDH3}-YFP fluorescence for *cis*-regulatory mutants (dark red), *trans*-regulatory mutants (dark blue), and *trans*-regulatory mutants enriched for extreme effects (dark brown) for mutants with increased P_{TDH3}-YFP fluorescence. (C) Effect on P_{TDH3}-YFP fluorescence for *cis*-regulatory mutants (light red), *trans*-regulatory mutants (light blue), and *trans*-regulatory mutants enriched for extreme effects (light brown) for mutants with decreased P_{TDH3}-YFP fluorescence. (D) Number of estimated bases in the *S. cerevisiae* genome (y-axis) that when mutated are expected to result in a change in expression equal to, or more extreme than, a specific cutoff (x-axis). *cis*-regulatory mutations (red). *trans*-regulatory mutations (blue). *trans*-regulatory mutations after enrichment (brown). Dashed lines show the maximum possible target size (potential target size) if all possible *cis*-regulatory mutations altered expression (red) or if all possible *trans*-regulatory mutations altered expression (blue).

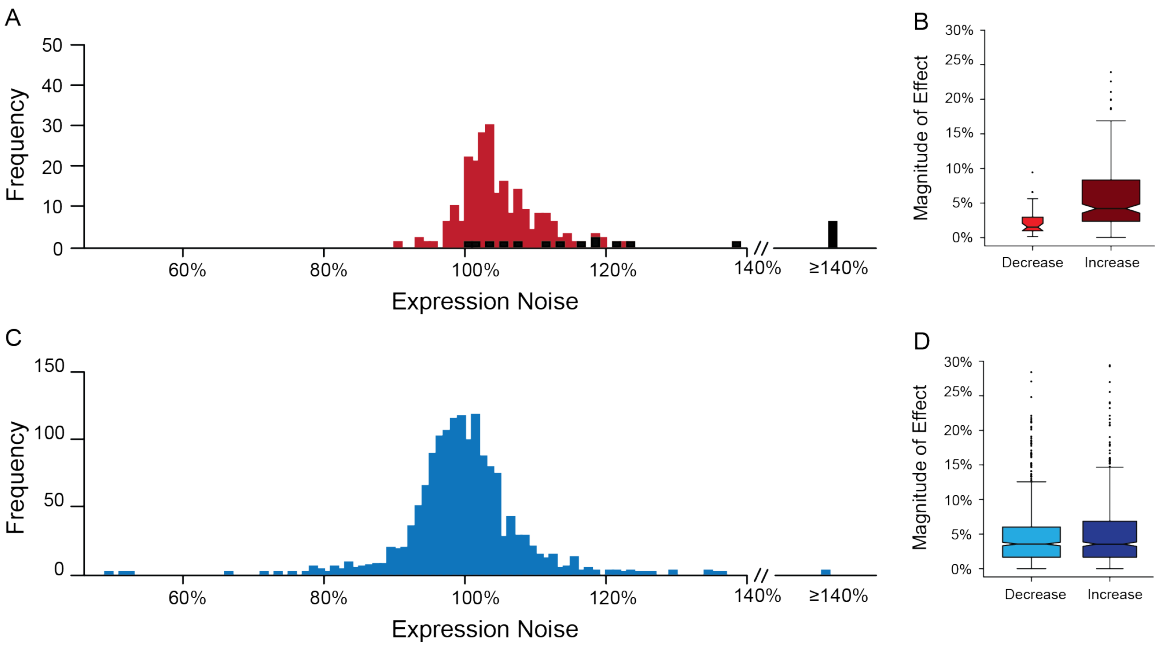


Figure 5-3 Frequency and effects of *cis*- and *trans*-regulatory mutants on P_{TDH3}-YFP fluorescence noise

(A) Frequency of *cis*-regulatory mutants with specific effects on P_{TDH3}-YFP fluorescence noise. Mutations outside of known TFBS (red). Mutations in known TFBS (black). (B) Effect of *cis*-regulatory mutants that increase (dark red) vs decrease (light red) P_{TDH3}-YFP fluorescence noise. (C) Frequency of *trans*-regulatory mutants with specific effects on P_{TDH3}-YFP fluorescence noise. (D) Effect of *trans*-regulatory mutants that increase (dark blue) or decrease (light blue) P_{TDH3}-YFP fluorescence noise. (E) Frequency of *trans*-regulatory mutants with specific effects on P_{TDH3}-YFP fluorescence noise for EMS induced mutants collected from the extremes of P_{TDH3}-YFP fluorescence. (F) Effect of extreme *trans*-regulatory mutants that increase (dark brown) or decrease (light brown) P_{TDH3}-YFP fluorescence noise.

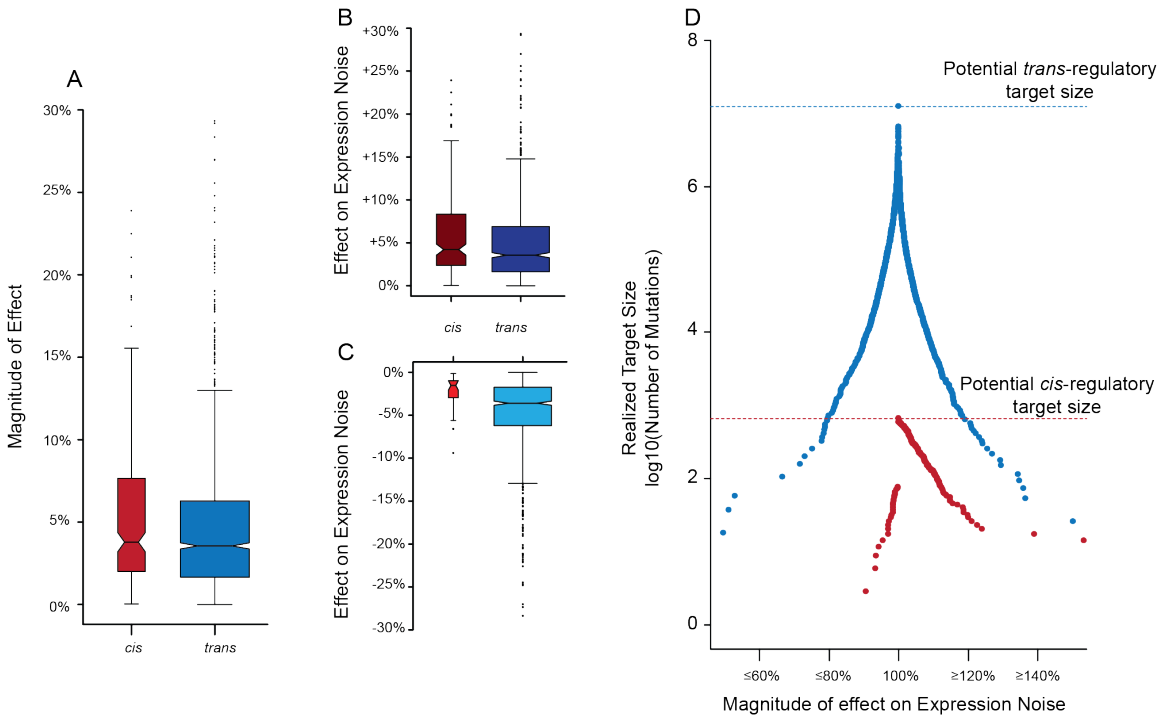


Figure 5-4 Comparison of frequency and effects for *cis*- and *trans*-regulatory mutants on P_{TDH3}-YFP fluorescence noise

(A) Absolute magnitude of effect on P_{TDH3}-YFP fluorescence noise for *cis*-regulatory mutants (red) and *trans*-regulatory mutants (blue). (B) Effect on P_{TDH3}-YFP fluorescence noise for *cis*-regulatory mutants (dark red), *trans*-regulatory mutants collected (dark blue), and *trans*-regulatory mutants enriched for extreme effects (dark brown) for mutants with increased P_{TDH3}-YFP fluorescence noise. (C) Effect on P_{TDH3}-YFP fluorescence noise for *cis*-regulatory mutants (light red), *trans*-regulatory mutants (light blue), and *trans*-regulatory mutants enriched for extreme effects (light brown) for mutants with decreased P_{TDH3}-YFP fluorescence noise. (D) Number of estimated bases in the *S. cerevisiae* genome (y-axis) that when mutated are expected to result in a change in expression noise equal to, or more extreme than, a specific cutoff (x-axis). *cis*-regulatory mutations (red). *trans*-regulatory mutations (blue). *trans*-regulatory mutations after enrichment (brown). Dashed lines show the maximum possible target size (potential target size) if all possible *cis*-regulatory mutations altered expression noise (red) or if all possible *trans*-regulatory mutations altered expression noise (blue).

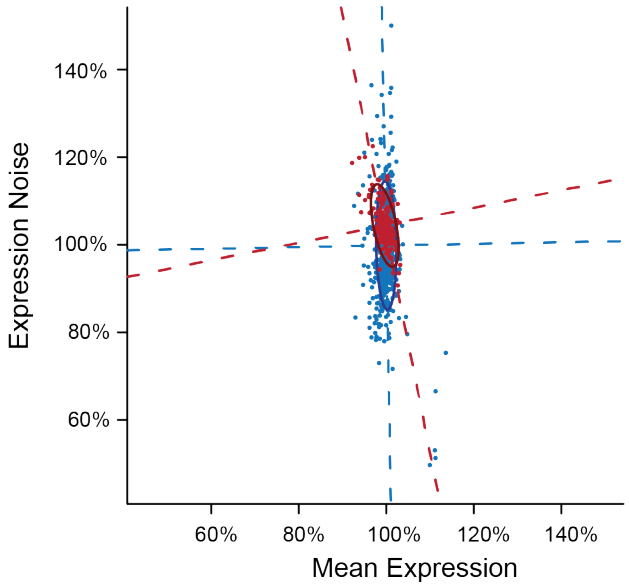


Figure 5-5 Relationship between mean expression and expression noise for *cis*- and *trans*-regulatory mutations

cis-regulatory mutations (red). *trans*-regulatory mutations (blue). Dashed line, principle components. Ovals indicate 95% of the data for *cis*- (red) and *trans*-regulatory mutations (blue). *cis*-regulatory and *trans*-regulatory mutations have a statistically significant difference in the relationship between mean expression and expression noise.

Chapter 6

Widespread compensatory changes in gene regulation maintains gene expression levels in *Saccharomyces* yeast

Abstract

Changes in gene expression are important contributors to phenotypic differences within and between species. Heritable differences in expression are caused by mutations in *cis*-regulatory elements and *trans*-regulatory factors. Previous work has shown that *trans*-regulatory changes are common within species and that *cis*-regulatory changes preferentially accumulate with time. However, biological limitations have restricted these analyses to relatively closely related species and it is unclear how gene regulation evolves on longer timescales. Here we use recently published data from distantly related *Saccharomyces* yeast to address this question. We estimate expression divergence between strains and species and compare these estimates to allele specific expression measurements from intraspecific and interspecific hybrids to estimate regulatory divergence in both *cis* and *trans*. We find that on long time scales the majority of differences in expression are due to *cis*-regulatory changes. In addition, we find that regulatory divergence and sequence divergence outpace expression divergence suggesting widespread compensatory changes in regulation. While much of this compensation is due to *cis* and *trans* changes in opposite direction, we also find evidence

for *trans*-*trans* compensation. We propose a general model of regulatory evolution that explains these results by invoking weak purifying selection on expression, differences in pleiotropy and additivity of *cis* and *trans*-regulatory mutations, and differences in the frequency of occurrence of *cis* and *trans*-regulatory mutations.

Introduction

Evolution is fundamentally a tug of war between forces that change and forces that maintain organismal form and function. The majority of new mutations are expected to be deleterious and the maintenance of phenotypes is thus thought to largely occur through the action of purifying selection. However, purifying selection does not immediately remove deleterious mutations and they can rise to high frequency in a population by genetic drift, genetic draft, or due to shifts in the environment (GILLESPIE 1994). In these instances, selection pressure exists for the mitigation of detrimental fitness effects, occurring either through the eventual extinction of lineages with deleterious alleles or by the appearance of new mutations that compensate for the deleterious effects of an initial mutation. In this later case, multiple mutations with fitness consequences will have occurred, despite little to no phenotypic change. Understanding the frequency and role such compensation plays is thus crucial to fully understanding the evolutionary process (GIBSON 1996; HARTL and TAUBES 1996; POON and OTTO 2000; MAISNIER-PATIN *et al.* 2002; WEINREICH *et al.* 2005; BADYAEV 2011; PAVLICEV and WAGNER 2012; RAJON and MASEL 2013; KHAN *et al.* 2013; SZAMECZ *et al.* 2014; ANDRIE *et al.* 2014).

Differences in gene expression are a common source of phenotypic variation within and between species (WHITEHEAD and CRAWFORD 2006; STERN and ORGOGOZO 2008; JONES *et al.* 2012; ALVAREZ *et al.* 2014). Because heritable differences in expression are caused by differences in the underlying DNA sequence, distantly related organisms are expected to diverge in expression. As expected, comparisons in flies, mammals, and yeast all indicate a clear positive relationship between sequence divergence and expression divergence (Figure 6-1) (COOLON *et al.* 2014). Surprisingly, however, the relationship between sequence and expression divergence appears similar between systems and suggests that common mechanisms may underlie patterns of gene expression divergence over long evolutionary timescales. The relationship between expression divergence and sequence divergence is non-linear and gene expression diverges more quickly during initial sequence divergence and more slowly as sequence divergence increases. This pattern could result from multiple sequence changes with opposing effects on gene expression that result in widespread compensation of expression levels (LANDRY *et al.* 2005; KUO *et al.* 2010; TAKAHASI *et al.* 2011; GONCALVES *et al.* 2012; SCHAEFKE *et al.* 2013; COOLON *et al.* 2014; ZEEVI *et al.* 2014).

Heritable differences in gene expression are caused by mutations in *cis*-regulatory elements (e.g. promoter and enhancers) and *trans*-regulatory factors (e.g. transcription factors and non-coding RNAs) (STERN and ORGOGOZO 2008; CARROLL 2008). The net contributions of *cis*- and *trans*-regulatory changes to differences in gene expression can be separated by comparing differences in expression between parental strains or species to differences in allele specific expression from F_1 hybrids made by crossing these strains

or species. In F_1 hybrids, both parental alleles are in a common *trans*-regulatory environment and allele specific differences in expression can be attributed solely to *cis*-regulatory changes (COWLES *et al.* 2002). Comparing changes in *cis* regulation with differences in expression between the parental strains or species can then be used to quantify the contribution of *trans*-regulatory changes to differences in gene expression (WITTKOPP *et al.* 2004). Numerous comparisons of the relative contribution of *cis* and *trans*-regulatory changes to expression variation within and between species have revealed the general mechanisms and patterns by which gene expression evolves (WITTKOPP *et al.* 2004, 2008; LANDRY *et al.* 2005; GENISSEL *et al.* 2008; TIROSH *et al.* 2009; MCMANUS *et al.* 2010; EMERSON *et al.* 2010; COOLON *et al.* 2014; LEMMON *et al.* 2014). These comparisons indicate that *trans*-regulatory changes are often a common source of expression differences within species (WITTKOPP *et al.* 2004; WANG *et al.* 2007; EMERSON *et al.* 2010; COOLON *et al.* 2014). This result is largely attributed to greater mutational input from *trans*-regulatory changes (WITTKOPP 2005). In addition, these studies find that the contribution of *cis*-regulatory changes to expression divergence often accumulates over time (GENISSEL *et al.* 2008; WITTKOPP *et al.* 2008; TIROSH *et al.* 2009; COOLON *et al.* 2014). This observation is consistent with either purifying selection preferentially removing highly pleiotropic *trans*-regulatory changes (PRUD'HOMME *et al.* 2007), positive selection preferentially acting on additive *cis*-regulatory changes (LEMOS *et al.* 2008; MCMANUS *et al.* 2010; SCHAEFKE *et al.* 2013), or both.

In addition to independent *cis* and *trans*-regulatory changes, previous work has shown that *cis* and *trans*-regulatory changes often jointly affect regulation of the same gene

(LANDRY *et al.* 2005; SHI *et al.* 2012; GONCALVES *et al.* 2012; SCHAEFKE *et al.* 2013; COOLON *et al.* 2014). When *cis* and *trans*-regulatory changes are in the same direction, they reinforce one another, resulting in a change in gene expression more extreme than either individual regulatory change. However, when *cis* and *trans*-regulatory changes are in opposite directions, they compensate for one another and gene expression divergence is reduced compared to the individual regulatory changes. As a consequence, sequence divergence and regulatory divergence increase more than expression divergence. In the absence of natural selection, reinforcing and compensating changes in regulation are expected to be equally abundant. However, compensation in gene regulation is more common than reinforcement, suggesting the action of natural selection (LANDRY *et al.* 2005; SCHAEFKE *et al.* 2013; COOLON *et al.* 2014).

Taken together, these observations suggest that the observed relationship between sequence divergence and expression divergence could be caused by widespread compensatory changes in regulation over long evolutionary time scales. Because *trans*-regulatory mutations are typically more common than *cis*-regulatory mutations, compensatory changes in regulation are predicted to occur primarily through initial *trans*-regulatory changes (GONCALVES *et al.* 2012). The compensating mutation can then either act in *cis* or in *trans*. When it is in *cis*, *trans*-*cis* compensation occurs. However, when both the initial mutation and the compensating mutation are in *trans*, the net *trans* change in expression is reduced. Because the partitioning of expression differences into *cis* and *trans*-regulatory changes accounts for only net contributions using the described methodology, *trans*-*trans* compensation cannot be directly detected. Instead, widespread

trans-*trans* compensation would manifest as a reduction in the amount of *trans*-regulatory divergence observed for large sequence divergences. Unfortunately, determining the extent of regulatory compensation broadly, and *trans*-*trans* compensation specifically, has been restricted by the inability to create viable hybrids among distantly related species in the majority of systems studied.

Here we utilize recently published data in the *Saccharomyces* genus to determine how *cis*, *trans*, and compensatory changes in regulation contribute to the evolution of gene expression over long evolutionary time scales. To measure gene expression and determine the underlying regulatory mechanisms involved in expression differences, we use RNA-seq data collected from both intra- and inter-specific hybrids, as well as parental strains and species, for species that span 40% sequence divergence (SCHRAIBER *et al.* 2013; SCHAEFKE *et al.* 2013). This level of sequence divergence is roughly equivalent to that observed between humans and chickens (DUJON 2006; HITTINGER 2013) and substantially increases the evolutionary distance upon which the evolution of gene regulation has been interrogated. We find that consistent with predictions, compensatory changes in gene expression are widespread over long time scales, with up to 35% of all genes showing a signature of regulatory divergence consistent with *trans*-*cis* compensation. In addition, after an initial spike in the frequency and magnitude of *trans*-regulatory changes, we find a reduction in *trans*-regulatory changes with greater sequence divergence, consistent with widespread *trans*-*trans* compensation.

Results

Expression divergence vs sequence divergence

To determine how expression divergence changes with sequence divergence in the genus *Saccharomyces*, we analyzed previously reported RNA-seq data. For short evolutionary time scales, we analyzed intraspecific levels of gene expression divergence between two *Saccharomyces cerevisiae* strains, BY4741 (BY) and RM11 (RM), which differ at approximately 34,000 sites in the ~12 Mb *S. cerevisiae* genome (0.0028 substitutions per site) (SCHAEFKE *et al.* 2013). To determine gene expression divergence at longer evolutionary time scales, we analyzed interspecific divergence in expression between *S. cerevisiae* (Sc) and three additional *Saccharomyces* species, *S. paradoxus* (Sp, 0.30 substitutions per site), *S. mikatae* (Sm, 0.48 substitutions per site), and *S. bayanus* (Sb, 0.90 substitutions per site) (SCHRAIBER *et al.* 2013). For each comparison, we correlated normalized read counts across genes and calculated gene expression divergence as the amount of variation in gene expression that was not explained by this correlation. As expected, we observed that gene expression divergence increased as sequence divergence increased, with the lowest levels of expression divergence observed between the two *S. cerevisiae* strains and the highest level of gene expression divergence observed between Sc and Sb (Figure 6-2A, black). As observed in other systems, the rate of expression divergence increased most quickly in the transition from within to between species and increased more slowly as sequence divergence between species increased, resulting in a plateau in gene expression divergence with increasing sequence divergence (Figure 6-1).

cis- and trans-regulatory divergence vs sequence divergence

Divergence in gene expression is caused by divergence in *cis*- and *trans*-regulation. To determine the relationship between *cis*-regulatory divergence and sequence divergence, we analyzed previously collected allele specific RNA-seq data from both intraspecific (BYxRM) and interspecific (Sc x Sp, Sc x Sm, and Sc x Sb) hybrids (COWLES *et al.* 2002). Analogous to the approach used to calculate gene expression divergence, we quantified *cis*-regulatory divergence as the excess variation in allele specific expression that could not be accounted for by a correlation between normalized allele specific read counts for each gene in each comparison. As with gene expression divergence, we observed that *cis*-regulatory divergence was higher for more distantly related species and lowest between *S. cerevisiae* strains (Figure 6-2A, red). The degree of *cis*-regulatory divergence tracked closely the degree of expression divergence, suggesting that a large proportion of expression divergence was due to changes in *cis*-regulation.

Differences in gene expression that cannot be attributed to allele specific expression differences in hybrids is inferred to be due to *trans*-regulatory changes (WITTKOPP *et al.* 2004). To determine the relationship between *trans*-regulatory divergence and sequence divergence, we estimated the difference in allele specific expression between strains or species and their hybrids for each gene. We then correlated these estimated differences in expression across genes and used the unexplained variation in expression divergence from the correlation as the contribution of *trans*-regulatory divergence to gene expression divergence. As with expression divergence and *cis*-regulatory divergence, we observed that *trans*-regulatory divergence was lowest within species (Figure 6-2A, blue). However,

unlike gene expression divergence and *cis*-regulatory divergence, we found that *trans*-regulatory divergence was constant between species regardless of the amount of sequence divergence. As a consequence, the relationship between *trans*-regulatory divergence and sequence divergence reached a plateau at lower levels of sequence divergence than *cis*-regulatory divergence or expression divergence. This lack of continued *trans*-regulatory divergence at high sequence divergence levels suggests the presence of *trans-trans* compensation.

Magnitude and frequency of cis and trans-regulatory changes

Correlations in gene expression and regulation take into account both the magnitude of change for each gene as well as the number of genes that have changed. To better understand how gene expression and regulation evolve at longer evolutionary time scales, we separated the effects of changes in magnitude from the effects of changes in frequency. To estimate divergence in the magnitude of change, we calculated the percent of total regulatory divergence (absolute *cis*-regulatory divergence plus absolute *trans*-regulatory divergence) that was due to *cis*-regulatory changes for each gene. We found that the median percent of regulatory divergence due to *cis*-regulatory divergence increased significantly as sequence divergence increased, ranging from approximately 40% within species, to nearly 60% between Sc and Sb ($p < 8 \times 10^{-6}$ for all comparisons, Wilcoxon rank sum test, Figure 6-2B).

Because *cis*-regulatory changes play increasingly larger roles in regulatory divergence as sequence divergence increases, *trans*-regulatory changes must by definition contribute

increasingly less to regulatory divergence. To determine how changes in the magnitude of *cis* and *trans*-regulatory changes contribute to this pattern, we calculated the absolute difference in *cis*- and *trans*-regulation for each gene for each comparison. We then took the median of these values to determine how the magnitude of regulatory divergence changed with sequence divergence. We found that median absolute *cis*-regulatory divergence increased as sequence divergence increased, consistent with genome wide patterns of *cis*-regulatory divergence and the contribution of *cis*-regulatory changes to total regulatory changes. By contrast, median absolute *trans*-regulatory divergence initially increased between species, but then decreased with increasing sequence divergence (Figure 6-2C). Thus, the increase in the contribution of *cis*-regulatory changes to regulatory divergence from within to between species is largely due to a greater increase in the average magnitude of *cis*-regulatory changes than for *trans*-regulatory changes, consistent with the preferential fixation of *cis*-regulatory changes over *trans*-regulatory changes. By contrast, the increases in the contribution of *cis*-regulatory changes to divergence between more distantly related species is caused in large part by a decrease in the average magnitude of *trans*-regulatory changes.

To determine the effect of changes in frequency of regulation on the observed patterns, we classified individual genes within each comparison into one of five regulatory divergence classes based on statistical support: *cis*-regulatory only, *trans*-regulatory only, reinforcing (*cis* and *trans* changes in the same direction), compensating (*cis* and *trans* changes in opposite directions), and conserved (none of the above). Consistent with overall expression divergence, the fraction of genes with a conserved pattern of

expression was highest within species (BY vs RM, 85%) and considerably lower between species (~23%). By contrast, the fraction of genes with only *cis*-regulatory changes increased nearly linearly with sequence divergence, going from approximately 6% within species to over 30% between Sc and Sb (Figure 6-3A). The number of genes with only *trans*-regulatory changes increased initially from within to between species, but then decreased as sequence divergence increased such that at the longest evolutionary distances considered, ~15% of genes had only a *trans*-regulatory difference in expression. This pattern for *trans*-regulatory changes is similar to the pattern observed for the magnitude of *trans*-regulatory changes. Thus, not only is the magnitude of *trans*-regulatory changes smaller with larger sequence divergence, fewer genes show evidence of *trans*-regulatory divergence as sequence divergence increases. Because this methodology detects only net *trans*-regulatory changes, these patterns suggest increasing levels of *trans*-*trans* compensation with sequence divergence.

In the absence of natural selection, *cis* and *trans*-regulatory changes in the same direction (reinforcing) and *cis* and *trans*-regulatory changes in the opposite direction (compensatory) are expected to occur with equal frequencies. As with previous reports, however, we observe that compensatory changes in expression are more common than reinforcing changes in expression, regardless of the level of sequence divergence. Interestingly, however, the proportion of genes with *cis*-*trans* compensatory changes in expression is highest at intermediate sequence divergence levels between Sc and Sp, and decreases slightly at higher sequence divergence. One possible explanation for this pattern is that there are simply fewer *trans*-regulatory changes between the most distantly

related species that can be compensated for. However, the number of genes showing a reinforcing pattern of divergence did not differ significantly between species, arguing against a simple reduction in *trans*-regulatory changes (p=0.25, chi-square test). Instead, these observations are more consistent with *trans*-*trans* compensation reducing the number of genes with net *trans*-regulatory changes, and hence the number of genes with *trans*-*cis* compensatory changes, as sequence divergence increases.

Consistent with these changes in the magnitude and frequency of *cis* and *trans*-regulatory changes, the correlation between expression differences and *cis*-regulatory differences increased with sequence divergence, while the correlation between expression differences and *trans*-regulatory differences decreases with sequence divergence (Figure 6-3B). Thus, irrespective of the exact type of compensation responsible, expression divergence on short time scales, such as within species, is dominated by changes in *trans*, while at longer evolutionary time scales, expression divergence is largely due to changes in *cis*.

Inheritance of gene expression vs regulatory divergence

The evolution of gene expression is determined not only by the effects of *cis* and *trans*-regulatory mutations, but also on the inheritance of these changes. For example, additive changes in expression are seen by natural selection as soon as they arise and can thus be efficiently fixed or removed. By contrast, mutations that are recessive must drift to moderate frequency before selection can act on them and mutations that are dominant are difficult to fix in a population by natural selection alone. In addition, combining changes in regulation can result in mis-expression, with heterozygotes having abnormal levels of

expression relative to the parental strains or species (LANDRY *et al.* 2005; RENAUT *et al.* 2009; LENZ *et al.* 2014).

To determine how inheritance of gene expression changed over long time scales, we divided all genes into four inheritance categories within each comparison based on statistical tests: conserved (no total expression difference between the hybrid and either parent), dominant (total expression difference between the hybrid and one parent only), additive (total expression difference between the hybrid and both parents, with hybrid expression intermediate that of the parents) and mis-expressed (total expression difference between the hybrid and both parents, with hybrid expression more extreme then both parents). We found that the number of genes with either a dominant or mis-expressed pattern of inheritance was relatively constant regardless of sequence divergence. By contrast, the number of conserved genes dropped with sequence divergence, while the number of additive genes increased with divergence (Figure 6-4A).

To determine how mode of inheritance and the mechanisms of regulatory divergence are related, we compared the relationship between regulatory divergence and mode of inheritance for each gene within each comparison. We found clear correspondences between regulatory divergence class and mode of inheritance class such that each regulatory divergence class was highly enriched for a single mode of inheritance. For example, genes with conserved patterns of regulatory divergence were enriched for genes with conserved inheritance and depleted for genes with additive modes of inheritance, consistent with small difference in expression between parental lines and the hybrid for

each comparison. Likewise, genes with *cis*-regulatory changes were largely additive in their effects, a result that has been observed in prior work. Similarly, genes with a reinforced pattern of divergence were enriched for additive effects, possibly because of substantial *cis*-regulatory contributions. By contrast, *trans*-regulatory changes were enriched for dominance in one strain/species over the other. Which specific allele was dominant and which allele was recessive was largely dependent on the specific cross, suggesting that *trans*-regulatory changes often result in allelic series of dominance. Finally, we observed that compensatory changes were enriched for mis-expression, consistent with underlying regulatory divergence but little expression divergence (Figure 6-4B). Overall we found highly congruent patterns across comparisons that were independent of sequence divergence, suggesting these relationships are largely constant over time. In addition, these results suggest that the mode of inheritance and mechanism of regulatory divergence are highly correlated and not fully independent.

Discussion

The percent of regulatory divergence due to *cis*-regulatory changes within species and between closely related species are consistent with previous estimates in both yeast and other systems on short evolutionary time scales. However, the percent of regulatory divergence due to *cis*-regulatory changes for the most distantly related species are substantially higher than previous reports. Why does the contribution of *cis*-regulatory changes to total regulatory divergence increase with sequence divergence? Because *trans*-regulatory mutations are expected to be more common than *cis*-regulatory mutations, regardless of sequence divergence, there must be a mechanism by which *cis*-

regulatory changes in expression are preferentially observed over *trans*-regulatory changes in expression that becomes stronger with increasing sequence divergence. There are at least three distinct possibilities; methodological artifacts, preferential selection for and against *cis* and *trans*-regulatory mutations respectively, and increasing *trans*-*trans* compensation with sequence divergence.

First, an increase in the contribution of *cis*-regulatory divergence to total regulatory divergence could be a methodological artifact. In particular, if the *trans*-regulatory elements of two strains or species do not interact equally with the two *cis*-regulatory sequences, then *trans*-regulatory changes will be incorrectly inferred as *cis*-regulatory changes (TAKAHASI *et al.* 2011). Such regulatory incompatibilities are expected to increase with sequence divergence, potentially resulting in an increase in the proportion of *cis*-regulatory divergence with increasing sequence divergence. However, it is currently unclear how often such changes in regulation occur, or how quickly they accumulate between species. In addition, such changes in regulation predict a decrease in the amount of *trans*-regulatory divergence with increasing sequence divergence. While we observe a decrease in the number of genes with only *trans*-regulatory and *cis*-*trans* compensatory changes in expression, we do not observe a decrease in the number of genes with *cis*-*trans* enhancing changes in expression, suggesting that this mechanism is not primarily responsible for the observed patterns.

Second, preferential selection against *trans*-regulatory mutations and preferential selection for *cis*-regulatory mutations have previously been proposed as contributing to

an increase in the contribution of *cis*-regulatory changes to regulatory divergence with increasing sequence divergence (WITTKOPP 2005; SCHAEFKE *et al.* 2013; COOLON *et al.* 2014). For example, *trans*-regulatory changes are expected to be more pleiotropic than *cis*-regulatory changes and because increased pleiotropy in considered deleterious, more pleiotropic mutations are under stronger purifying selection. As a consequence, a lower proportion of potential *trans*-regulatory changes than *cis*-regulatory changes are expected to contribute to regulatory divergence. If mutations with large effects on expression are more pleiotropic than mutations with small effects on expression, then pleiotropic regulatory changes should show a reduction in the magnitude of effect with increasing sequence divergence as the largest effects are removed by purifying selection. Consistent with this hypothesis, we observe that for *trans*-regulatory changes in expression, but not for *cis*-regulatory changes in expression, the magnitude of change in expression decreases with increasing sequence divergence.

Alternatively, *cis*-regulatory mutations could be preferentially selected for. One proposed mechanism for greater positive selection on *cis*-regulatory changes than *trans*-regulatory changes is a greater additivity for *cis*-regulatory changes. Because positive selection is more efficient at fixing additive beneficial mutations than recessive beneficial mutations, the proportion of regulatory changes that is additive should preferentially increase with sequence divergence. Consistent with this hypothesis, we found that *cis*-regulatory changes were enriched for an additive mode of inheritance, while *trans*-regulatory changes were enriched for a dominant/recessive mode of inheritance. This pattern has been previously observed in other systems, suggesting that it is likely a common

phenomenon (LEMOS *et al.* 2008; SCHAEFKE *et al.* 2013; COOLON *et al.* 2014). Consistent with these results, recent work has shown that new *cis*-regulatory mutations are often additive, while new *trans*-regulatory mutations are often recessive (GRUBER *et al.* 2012).

Finally, widespread *trans*-*trans* compensation can lead to an increase in the contribution of *cis*-regulatory changes over time. After an initial spike, we detected a decrease in both the magnitude and number of genes with *trans*-regulatory effects with increasing sequence divergence. This is consistent with new *trans*-regulatory changes compensating for already existing differences in gene expression. In addition, because *trans*-regulatory mutations are more common than *cis*-regulatory mutations, *trans*-*trans* compensation is expected to occur more frequently than either *trans*-*cis, cis*-*trans,* or *cis*-*cis* compensation. We found that 35% of genes showed patterns of regulatory divergence consistent with *cis*-*trans* or *trans*-*cis* compensation, suggesting that *trans*-*trans* compensation is prevalent. As a consequence, the continued increase in *cis*-regulatory divergence with sequence divergence combined with a high frequency of *trans*-*trans* compensation would lead to increasing *cis*-regulatory divergence as a proportion of net regulatory divergence.

Combined with previous analyses in other species, these data suggest a simple model of regulatory evolution containing three parts: weak, but widespread, purifying selection on expression levels, differences in pleiotropy and/or additivity between *cis* and *trans*-regulatory mutations, and differences in the frequency of *cis* and *trans*-regulatory mutations. Together, these three properties can explain the observed patterns of

expression and regulatory divergence. First, because *trans*-regulatory mutations are more common than *cis*-regulatory mutations and purifying selection is weak within species, *trans*-regulatory changes in expression are common within species. Second, the interaction of natural selection with differences in pleiotropy and additivity of *cis* and *trans*-regulatory mutations over long timescales results in an ever increasing proportion of regulatory divergence due to *cis*-regulatory changes. Finally, the combination of purifying selection on expression and the large target size for *trans*-regulatory mutations results in widespread *trans*-*trans* compensation for expression. As a consequence, net *trans*-regulatory divergence plateaus between species, further contributing to the increasing proportion of regulatory divergence due to *cis*-regulatory changes between species. It is interesting to note that such a model implicitly assumes large underlying changes in regulation with relatively static levels of expression and is therefore compatible with growing evidence that gene regulation can be fluid.

Methods

Sequence divergence (substitutions per site) estimates were taken from the literature for flies (LIN *et al.* 2008), mammals (PRASAD *et al.* 2008), and yeast (SCANNELL *et al.* 2011). Estimates of expression divergence for flies and mammals were taken from (COOLON *et al.* 2014) which used data from (BRAWAND *et al.* 2011) for mammals and from (MCMANUS *et al.* 2010; MEISEL *et al.* 2012; SUVOROV *et al.* 2013) for flies. Intraspecific estimates of expression divergence for yeast used data from (SCHAEFKE *et al.* 2013) and an interspecific estimates used data from (SCHRAIBER *et al.* 2013). For all samples, normalized read counts for each gene were correlated between samples using spearman's

rho. This represents the extent of expression conservation between strains and species and we estimated expression divergence as one minus this value.

To determine regulatory divergence within species, we used allele specific expression data between *S. cerevisiae* strains BY and RM and their F_1 hybrid from (SCHAEFKE *et al.* 2013). Relative allelic expression within the F_1 hybrids suggested the presence of aneuploidy or large copy number variants on chromosomes II and XIII and these regions were removed from further analysis in all comparisons. To determine regulatory divergence between species, we used allele specific expression data between *S. cerevisiae* (YHL068) and *S. paradoxus* (CBS 432), *S. cerevisiae* and *S. mikatae* (IFO 1815), and *S. cerevisiae* and *S. bayanus* (CBS 7001), as well as their F_1 hybrids, from (SCHRAIBER *et al.* 2013). In total we made eight comparisons, four expression comparisons between parental strains or species, and four allele specific expression comparison between these strains or species F_1 hybrids. For each comparison, we combined read counts at each gene from two independent biological replicates. To remove differences in coverage within a gene across comparisons, we down-sampled total counts to the minimum across comparisons for each gene (COOLON *et al.* 2014). As a result of this down sampling, each gene in each comparison had the same number of allele specific reads. We removed any whose counts within a comparison were less than 20 and used only the set of genes with data across all four comparisons.

For each of the four strain/species comparisons, we estimated changes in total expression at each gene as the $\log_2$ ratio of allele specific counts in the parental strains or species.

Likewise, we estimated *cis*-regulatory changes as the $\log_2$ ratio of allele specific counts within the F_1 hybrids. We then estimated *trans*-regulatory changes as the difference between total expression changes and *cis*-regulatory changes. To determine the relationship between *cis*-regulatory divergence and sequence divergence, we correlated allele specific read counts across genes for each F_1 hybrid using spearman's rho and used one minus this value as the extent of *cis*-regulatory divergence. To perform the same calculation for *trans*-regulatory divergence, we used the ratios of *trans*-regulatory differences (r) and the number of estimated reads for each specific gene (n) to estimate the number of reads expected in each comparison (x and y) due to *trans*-regulatory differences. This was done by solving the equations $x + y = n$ and $x/y = r$ for x and y. These estimates were then correlated across genes using spearman's rho to determine the extent of *trans*-regulatory similarity. One minus this value was used to estimate *trans*-regulatory divergence. Finally, we used spearman's rho to correlate the estimates of expression changes with the estimates of *cis* and *trans*-regulatory changes to determine how the dominant regulatory mechanism by which expression diverged for each comparison.

To calculate the percent of regulatory divergence due to *cis*-regulatory change at each gene within each comparison, we divided the absolute value of *cis*-regulatory changes by the sum of the absolute values of *cis*-regulatory changes and *trans*-regulatory changes. We compared the percent of regulatory divergence due to *cis*-regulatory divergence between each comparison using Wilcoxon rank sum test. To determine the change in magnitude of expression and regulatory differences with sequence divergence, we

calculated the median absolute difference in expression and *cis*- and *trans*-regulation for each comparison.

To test for significant differences in total expression between strains or species, we compared the allele specific counts between parental strains or species for each gene in each comparison using a binomial exact test and a p-value cutoff of 0.01. Likewise, to test for significant *cis*-regulatory changes, we compared allele specific counts from F_1 hybrids in each comparison for each gene using a binomial exact test with a p-value cutoff of 0.01. To detect significant *trans*-regulatory changes, we used Fisher's exact test and a p-value cutoff of 0.01 to compare the allele specific counts in the parental strains or species and the allele specific counts in the F_1 hybrids for each comparison and gene. To estimate the frequency of regulatory changes, we used the results of these statistical tests to categorize each gene in each comparison into one of five regulatory divergence categories (Table 6-1).

To test for differences in the mode of inheritance, we used the original data set before down sampling across samples within a gene. Instead, we down sampled across all comparisons globally, maintaining the same number of reads within each sample. As before, we removed any gene in which the average number of reads across samples was below 20. We then used binomial exact test to compare total expression between parental strains or species as well as between each parental strain or species and their F_1 hybrid. We used a p-value cutoff of 0.01 to determine statistical significance. Each gene was classified into one of four categories based on these statistical tests. Genes with

significant differences in expression in all three tests were classified as additive if the hybrid was intermediate the two parents, or mis-expressed if the hybrid was more extreme than both parents. Genes that were significantly different between parental strains or species as well as between the hybrid and one parental strain or species, but not the other parental strain or species, were classified as dominant. All other genes were classified as conserved. Enrichment for the overlap between regulatory and inheritance categories was calculated using chi-square tests. For each test, a 2x2 contingency tables was created by collapsing all combinations of categories but one. Each combination of regulatory and inheritance categories was tested individually.

References

ALVAREZ M., SCHREY A. W., RICHARDS C. L., 2014 Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? Mol. Ecol. **24**: 710–725.

ANDRIE J., WAKEFIELD J., AKEY J. M., 2014 Heritable variation of mRNA decay rates in yeast. Genome Res.: 2000–2010.

BADYAEV A. V, 2011 Origin of the fittest: link between emergent variation and evolutionary change as a critical question in evolutionary biology. Proc. R. Soc. B **278**: 1921–9.

BRAWAND D., SOUMILLON M., NECSULEA A., JULIEN P., CSÁRDI G., HARRIGAN P., WEIER M., LIECHTI A., AXIMU-PETRI A., KIRCHER M., ALBERT F. W., ZELLER U., KHAITOVICH P., GRÜTZNER F., BERGMANN S., NIELSEN R., PÄÄBO S., KAESSMANN H., 2011 The evolution of gene expression levels in mammalian organs. Nature **478**: 343–348.

CARROLL S. B., 2008 Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell **134**: 25–36.

COOLON J. D., MCMANUS C. J., STEVENSON K. R., GRAVELEY B. R., WITTKOPP P. J., 2014 Tempo and mode of regulatory evolution in Drosophila. Genome Res.: gr.163014.113–.

COWLES C. R., HIRSCHHORN J. N., ALTSHULER D., LANDER E. S., 2002 Detection of regulatory variation in mouse genes. Nat. Genet. **32**: 432–437.

DUJON B., 2006 Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet. **22**: 375–87.

EMERSON J. J., HSIEH L.-C., SUNG H.-M., WANG T.-Y., HUANG C.-J., LU H. H.-S., LU M.-Y. J., WU S.-H., LI W.-H., 2010 Natural selection on cis and trans regulation in yeasts. Genome Res.: 826–836.

GENISSEL A., MCINTYRE L. M., WAYNE M. L., NUZHDIN S. V, 2008 Cis and trans regulatory effects contribute to natural variation in transcriptome of Drosophila melanogaster. Mol. Biol. Evol. **25**: 101–10.

GIBSON G., 1996 Epistasis and pleiotropy as natural properties of transcriptional regulation. Theor. Popul. Biol. **49**: 58–89.

GILLESPIE J. H., 1994 *The Causes of Molecular Evolution*. Oxford University Press.

GONCALVES A., LEIGH-BROWN S., THYBERT D., STEFFLOVA K., TURRO E., FLICEK P., BRAZMA A., ODOM D. T., MARIONI J. C., 2012 Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res. **22**: 2376–2384.

GRUBER J. D., VOGEL K., KALAY G., WITTKOPP P. J., 2012 Contrasting Properties of Gene-specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects and Dominance. PLoS Genet. **8**: e1002497.

HARTL D. L., TAUBES C. H., 1996 Compensatory nearly neutral mutations: selection without adaptation. J. Theor. Biol. **182**: 303–9.

HITTINGER C. T., 2013 Saccharomyces diversity and evolution: a budding model genus. Trends Genet. **29**: 309–17.

JONES F. C., GRABHERR M. G., CHAN Y. F., RUSSELL P., MAUCELI E., JOHNSON J., SWOFFORD R., PIRUN M., ZODY M. C., WHITE S., BIRNEY E., SEARLE S., SCHMUTZ J., GRIMWOOD J., DICKSON M. C., MYERS R. M., MILLER C. T., SUMMERS B. R., KNECHT A. K., BRADY S. D., ZHANG H., POLLEN A. a., HOWES T., AMEMIYA C., BALDWIN J., BLOOM T., JAFFE D. B., NICOL R., WILKINSON J., LANDER E. S., PALMA F. DI, LINDBLAD-TOH K., KINGSLEY D. M., 2012 The genomic basis of adaptive evolution in threespine sticklebacks. Nature **484**: 55–61.

KHAN Z., FORD M. J., CUSANOVICH D. a, MITRANO A., PRITCHARD J. K., GILAD Y., 2013 Primate transcript and protein expression levels evolve under compensatory selection pressures. Science **342**: 1100–4.

KUO D., LICON K., BANDYOPADHYAY S., CHUANG R., LUO C., CATALANA J., RAVASI T., TAN K., IDEKER T., 2010 Coevolution within a transcriptional network by compensatory trans and cis mutations. Genome Res. **20**: 1672–8.

LANDRY C. R., WITTKOPP P. J., TAUBES C. H., RANZ J. M., CLARK A. G., HARTL D. L., 2005 Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. Genetics **171**: 1813–22.

LEMMON Z. H., BUKOWSKI R., SUN Q., DOEBLEY J. F., 2014 The Role of cis Regulatory Evolution in Maize Domestication (H Fraser, Ed.). PLoS Genet. **10**: e1004745.

LEMOS B., ARARIPE L. O., FONTANILLAS P., HARTL D. L., 2008 Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. PNAS **105**: 14471–6.

LENZ D., RILES L., FAY J., 2014 Heterochronic meiotic misexpression in an interspecific yeast hybrid. Mol. Biol. Evol.

LIN M. F., DEORAS A. N., RASMUSSEN M. D., KELLIS M., 2008 Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. PLoS Comput. Biol. **4**.

MAISNIER-PATIN S., BERG O. G., LILJAS L., ANDERSSON D. I., 2002 Compensatory adaptation to the deleterious effect of antibiotic resistance in Salmonella typhimurium. Mol. Microbiol. **46**: 355–66.

MCMANUS C. J., COOLON J. D., O'DUFF M., EIPPER-MAINS J., GRAVELEY B. R., WITTKOPP P. J., DUFF M. O., 2010 Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. **20**: 816–25.

MEISEL R. P., MALONE J. H., CLARK A. G., 2012 Faster-X Evolution of Gene Expression in Drosophila. PLoS Genet. **8**.

PAVLICEV M., WAGNER G. P., 2012 A model of developmental evolution: selection, pleiotropy and compensation. Trends Ecol. Evol.: 1–7.

POON A., OTTO S. P., 2000 Compensating for our load of mutations: freezing the meltdown of small populations. Evolution (N. Y). **54**: 1467–79.

PRASAD A. B., ALLARD M. W., GREEN E. D., 2008 Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol. Biol. Evol. **25**: 1795–808.

PRUD'HOMME B., GOMPEL N., CARROLL S. B., 2007 Emerging principles of regulatory evolution. PNAS **104**: 8605–8612.

RAJON E., MASEL J., 2013 Compensatory evolution and the origins of innovations. Genetics **193**: 1209–20.

RENAUT S., NOLTE A. W., BERNATCHEZ L., 2009 Gene expression divergence and hybrid misexpression between lake whitefish species pairs (Coregonus spp. Salmonidae). Mol. Biol. Evol. **26**: 925–36.

SCANNELL D. R., ZILL O. A., ROKAS A., PAYEN C., DUNHAM M. J., EISEN M. B., RINE J., JOHNSTON M., HITTINGER C. T., 2011 The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 **1**: 11–25.

SCHAEFKE B., EMERSON J. J., WANG T.-Y., LU M.-Y. J., HSIEH L.-C., LI W.-H., 2013 Inheritance of gene expression level and selective constraints on trans- and cis-regulatory changes in yeast. Mol. Biol. Evol. **30**: 2121–33.

SCHRAIBER J. G., MOSTOVOY Y., HSU T. Y., BREM R. B., 2013 Inferring evolutionary histories of pathway regulation from transcriptional profiling data. PLoS Comput. Biol. **9**: e1003255.

SHI X., NG D. W.-K., ZHANG C., COMAI L., YE W., CHEN Z. J., 2012 Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. Nat. Commun. **3**: 950.

STERN D. L., ORGOGOZO V., 2008 The loci of evolution: how predictable is genetic evolution? Evolution (N. Y). **62**: 2155–2177.

SUVOROV A., NOLTE V., PANDEY R. V., FRANSSEN S. U., FUTSCHIK A., SCHLÖTTERER C., 2013 Intra-specific regulatory variation in Drosophila pseudoobscura. PLoS One **8**.

SZAMECZ B., BOROSS G., KALAPIS D., KOVÁCS K., FEKETE G., FARKAS Z., LÁZÁR V., HRTYAN M., KEMMEREN P., GROOT KOERKAMP M. J. a, RUTKAI E., HOLSTEGE F. C. P., PAPP B., PÁL C., 2014 The Genomic Landscape of Compensatory Evolution. PLoS Biol. **12**: e1001935.

TAKAHASI K. R., MATSUO T., TAKANO-SHIMIZU-KOUNO T., 2011 Two types of cis-trans compensation in the evolution of transcriptional regulation. PNAS **108**: 15276–81.

TIROSH I., REIKHAV S., LEVY A. a, BARKAI N., 2009 A yeast hybrid provides insight into the evolution of gene expression regulation. Science **324**: 659–662.

WANG D., SUNG H.-M., WANG T.-Y., HUANG C.-J., YANG P., CHANG T., WANG Y.-C., TSENG D.-L., WU J.-P., LEE T.-C., SHIH M.-C., LI W.-H., 2007 Expression

evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. Genome Res. **17**: 1161–9.

WEINREICH D. M., WATSON R. A., CHAO L., 2005 Perspective: sign epistasis and genetic costraint on evolutionary trajectories. Evolution (N. Y). **59**: 1165–1174.

WHITEHEAD A., CRAWFORD D. L., 2006 Variation within and among species in gene expression: raw material for evolution. Mol. Ecol. **15**: 1197–1211.

WITTKOPP P. J., 2005 Genomic sources of regulatory variation in cis and in trans. Cell. Mol. Life Sci. **62**: 1779–83.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2004 Evolutionary changes in cis and trans gene regulation. Nature **430**: 85–88.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2008 Regulatory changes underlying expression differences within and between *Drosophila* species. Nat. Genet. **40**: 346–350.

ZEEVI D., LUBLINER S., LOTAN-POMPAN M., HODIS E., VESTERMAN R., WEINBERGER A., SEGAL E., 2014 Molecular dissection of the genetic mechanisms that underlie expression conservation in orthologous yeast ribosomal promoters. Genome Res.: 1991–1999.

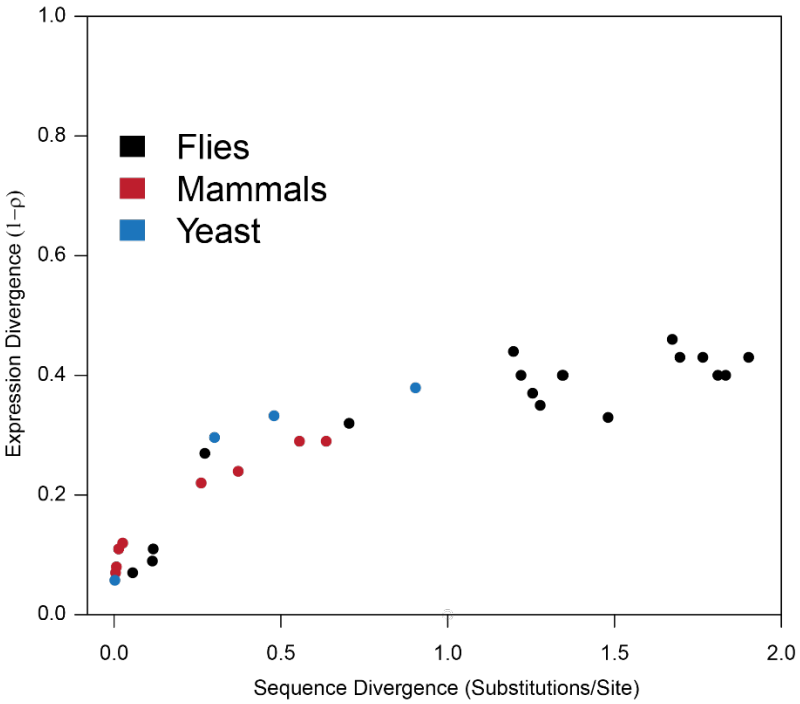


Figure 6-1 Expression divergence vs. sequence divergence

Y-axis: Divergence in gene expression between strains and species as measured by one minus spearman's rho. X-axis: Genome wide estimates of sequence divergence for flies (Black, LIN *et al.* 2008), mammals (Red, PRASAD *et al.* 2008), and yeast (Blue, SCANNELL *et al.* 2011). Expression divergence data for flies and mammals is the same used in (COOLON *et al.* 2014). Consistent pattern between diverse species suggests similar underlying mechanism maintaining gene expression levels over long evolutionary distances.

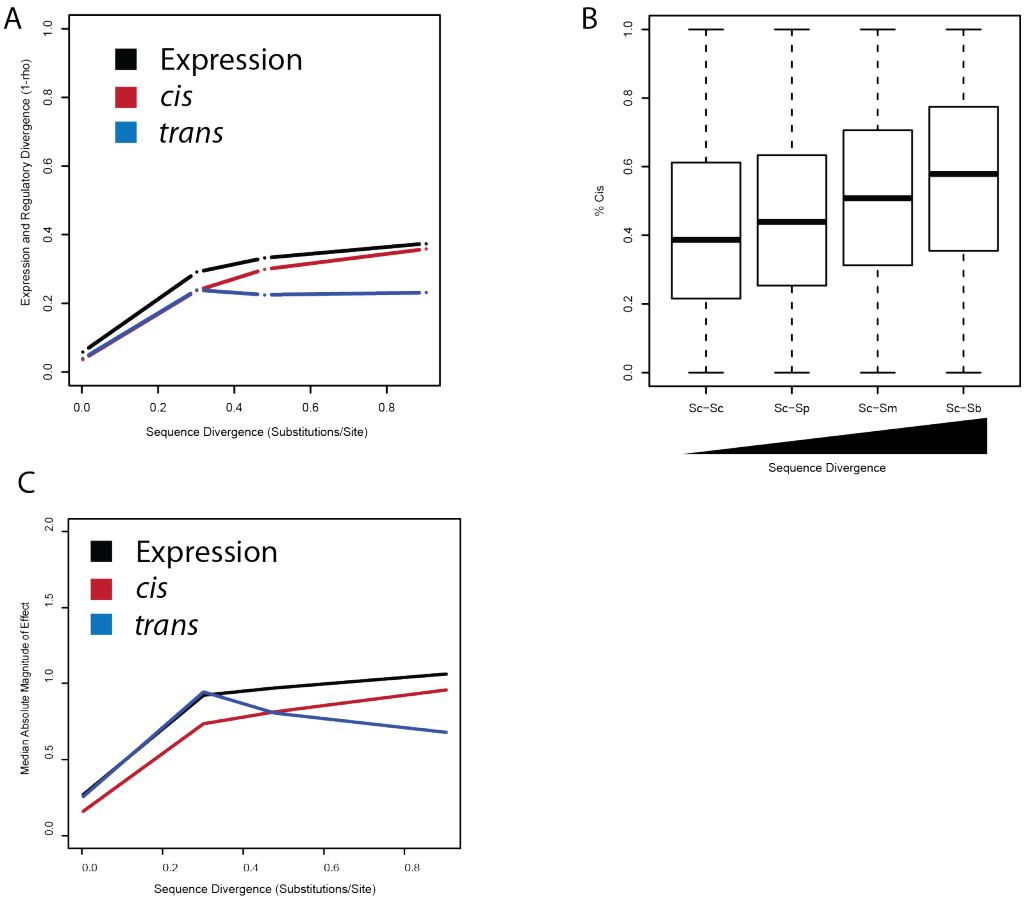


Figure 6-2 Regulatory divergence vs sequence divergence

A. Expression and regulatory divergence vs sequence divergence for *Saccharomyces* yeast. Divergence in expression and regulation was calculated as one minus spearman's rho. Sequence divergence is from four-fold degenerate sites within coding regions. Black: Total expression divergence. Red: *cis*-regulatory divergence. Blue: *trans*-regulatory divergence. B. Percent of total regulatory divergence due to *cis*-regulatory changes vs increasing sequence divergence. C. Median absolute magnitude of difference in expression and regulation vs sequence divergence. Colors are the same as in A.

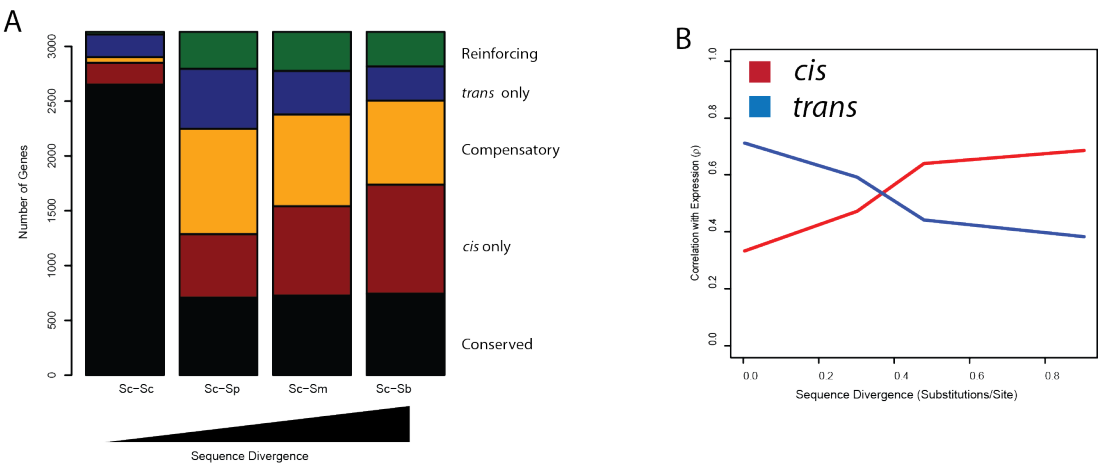


Figure 6-3 Mechanism of regulatory divergence vs. sequence divergence

A. Number of genes within each regulatory divergence category vs sequence divergence. Genes were classified into one of five categories based on statistical support: Black, conserved. Red, *cis*-regulatory change only. Orange, compensatory. Blue, *trans*-regulatory change only. Green, reinforcing. B. Correlation between *cis* (red) and *trans* (blue) regulatory divergence and total expression divergence vs sequence divergence.

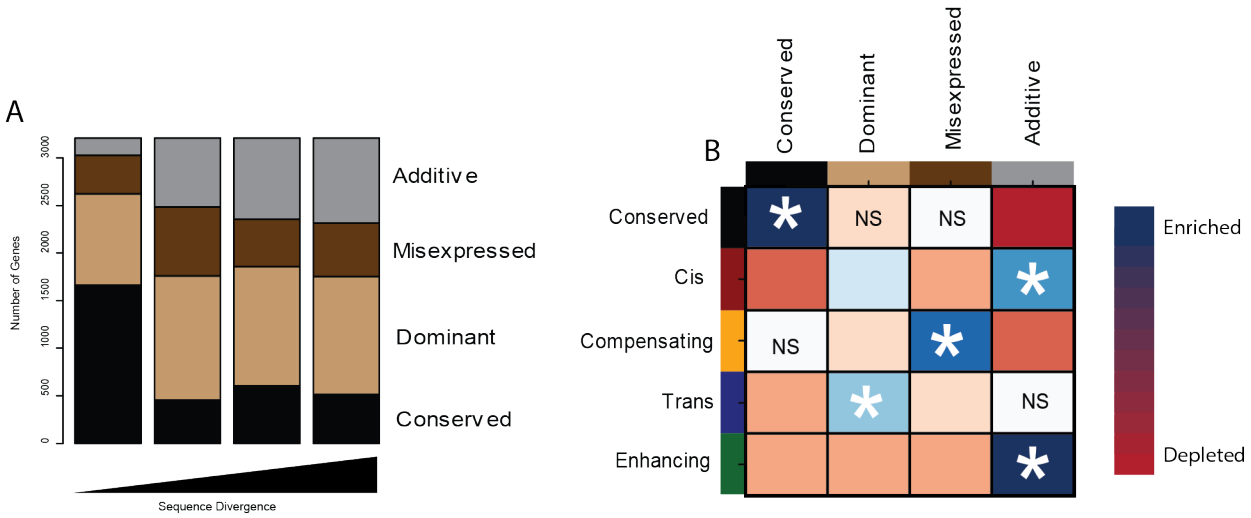


Figure 6-4 Mode of inheritance vs sequence divergence

A. Number of genes within each mode of inheritance category vs sequence divergence. Genes were classified into one of four categories based on statistical support: Black, conserved. Light brown, dominant/recessive. Dark brown, mis-expressed. Gray, additive. B. Comparison of regulatory divergence categories and modes of inheritance categories. Regulatory divergence categories are on the right with the same color as in Figure 6-3. Mode of inheritance categories are on the top with the same color as in Figure 6-4A. Intersections show enrichment (blue) and depletion (red) of genes within each combination of regulatory and inheritance categories. Asterisks mark strongest enrichment for each regulatory divergence category. All unmarked categories were significant at less than a Bonferroni corrected p-value of 0.0005.

Category	Significant expression[1]	Significant *cis*[1]	Significant *trans*[2]	*cis* and *trans*-direction?
cis only	✓	✓	×	NA
trans only	✓	×	✓	NA
Compensatory[3]	?	✓	✓	Opposite
Reinforcing	✓	✓	✓	Same
Conserved[4]	-	-	-	NA

Table 6-1 Regulatory divergence categories based on statistical tests

Genes were categorized into one of five regulatory divergence categories based on the result of three statistical tests. Checkmarks indicate a required significant test result. X indicate a required non-significant test result. [1]Binomial exact tests. [2]Fisher's exact test. [3]Compensatory changes required a significant *cis* effect and a significant *trans* effect, regardless of whether there was a significant difference in total expression. [4]Genes were categorized as conserved if they did not meet the criteria for any of the other four categories.

Chapter 7

Discussion and Future Directions

A very great deal more truth can become known than can be proven

— Richard Feynman, 1965

Mutation is the ultimate source of heritable phenotypic variation. While mutations are random in their occurrence, the phenotypic effects they produce are biased and some phenotypic changes are more likely to occur than others. This difference in the frequency at which phenotypes are produced by the mutational process reflect biological organization and structure. As a consequence, determining what these biases are and understanding their origins is crucial for a complete picture of the evolutionary process. This thesis examined the role of new mutations in the evolution of gene expression. Here, I recap some of the main conclusions and the consequences of this work for future research.

Consequences of mutational biases for the evolution of gene expression

Changes in gene expression are caused by *cis*- and *trans*-regulatory mutations. In natural populations, the relative contributions of *cis*- and *trans*-regulatory changes to differences in regulation often varies (WITTKOPP *et al.* 2008; TIROSH *et al.* 2009; EMERSON *et al.*

2010; COOLON *et al.* 2014). However, it is unclear whether the observed patterns of regulatory divergence reflect differences in the effects of new *cis*- and *trans*-regulatory mutations, differences in how natural selection acts on *cis*- and *trans*-regulatory changes, or some combination of both mutation and selection. To begin addressing this question, we first examined the effects of new *cis*- and *trans*-regulatory mutations on expression of the *S. cerevisiae TDH3* gene.

Interestingly, while *trans*-regulatory mutations were more common than *cis*-regulatory mutations, *cis*-regulatory mutations had on average larger effects on *TDH3* expression than *trans*-regulatory mutations. This result is consistent with previous work indicating larger effect of likely *cis*-regulatory (local) eQTL than likely *trans*-regulatory (distal) eQTL (SCHADT *et al.* 2003). Our work suggests that this pattern may in part reflect biases from the mutational process and is not entirely a product of natural selection. In addition, the larger average effect size for *cis*-regulatory mutations compared to *trans*-regulatory mutations means that in the absence of natural selection, *cis*-regulatory mutations will make a greater contribution to regulatory divergence than if *cis*- and *trans*-regulatory mutations were of equal size. It is therefore insufficient to rely solely on the frequency of *cis* and *trans*-regulatory mutations to determine their expected contributions to regulatory divergence in the absence of natural selection. Instead, both the frequency and effect size of new mutations must be considered. This has important implications for the evolution of gene expression because while differences in the frequency of *cis* and *trans*-regulatory mutations can be predicted, differences in the effects of *cis* and *trans*-regulatory mutations currently cannot. In addition, this thesis represents one of the first attempts to

directly compare the effects of *cis* and *trans*-regulatory mutations and is limited to only a single gene. There is therefore a lack of empirical data regarding what might be a typical difference in the effects of *cis* and *trans*-regulatory mutations.

We also found that new regulatory mutations are often biased in their direction of effect. This has important implications for tests of natural selection on gene expression. In particular, tests that assume increases and decreases in expression are equally likely may be prone to error when mutational biases are strong (FRASER *et al.* 2010, 2011, 2012; BULLARD *et al.* 2010; CHANG *et al.* 2013; SCHRAIBER *et al.* 2013). For example, if a set of related genes all show a consistent shift in expression in the same direction, this is often interpreted as positive selection for a change in gene expression. However, in the presence of biased mutational effects, this pattern could also arise due the relaxation of purifying selection. It is thus preferable to directly measure mutational biases and use these biases as the appropriate null distribution to test for the action of natural selection (SMITH *et al.* 2013).

Finally, we observed that *cis* and *trans*-regulatory mutations were biased in opposite directions, with *cis*-regulatory mutations often decreasing expression and *trans*-regulatory mutations often increasing expression. To our knowledge, this is the first indication of such differences in bias for different molecular mechanisms of regulatory change. This pattern likely arises due to differences in the proportion of activators and repressors for direct regulators of *TDH3* expression (those that bind the *TDH3* promoter) and indirect regulators of *TDH3* expression (those that bind to the direct regulators). As

such, it is likely to vary from gene to gene. However, if *cis* and *trans*-regulatory differences are often in opposite directions, this suggests that they will often compensate for one another, even in the absence of natural selection. Consistent with this, we find widespread evidence of compensatory changes in *Saccharomyces* gene regulation over long timescales (LANDRY *et al.* 2005; KUO *et al.* 2010; GONCALVES *et al.* 2012; SZAMECZ *et al.* 2014). It will be worth considering what percentage of this compensation is driven by natural selection to maintain expression levels and what percentage results largely from neutral processes. In theory, this question could be addressed using similar approaches to those taken here: by comparing the frequency of compensatory changes in natural populations to those observed from new mutations, the relative contributions of mutation and selection could be determined.

The action of natural selection depends on the effects of new mutations

In the absence of natural selection, variation in gene expression segregating within natural populations should reflect the biases of the mutational process. As a consequence, comparing the effects on expression of new mutations to the effects on expression of polymorphisms segregating in natural populations can be used to test for the action of natural selection. To perform this test, we first determined the extent of variation in *S. cerevisiae* by producing high quality genome sequences for 85 diverse strains. From this set, we created a collection of genetically tractable strains that could be experimentally manipulated in the laboratory. Using these strains, we determined the extent of both *cis* and *trans*-regulatory variation on *TDH3* expression and compared the observed effects to the empirically derived mutational distributions. We found evidence of natural selection

acting on *TDH3* expression for *trans*-regulatory mutations, but not *cis*-regulatory mutations, indicating that the action of natural selection can vary across molecular mechanisms. This result is a direct consequence of differences in the mutational effects of *cis* and *trans*-regulatory changes and therefore highlights the need to first understand how new mutation impacts a phenotype to determine the effects of natural selection.

Consequences of gene expression noise on the evolution of gene expression

Mutations can alter not only the average level of expression, but also the amount of variability in expression between genetically identical individuals, or gene expression noise. Comparing the effects of new mutations on gene expression noise revealed further differences between *cis*- and *trans*-regulatory mutations: *cis*-regulatory mutations were skewed towards increasing expression noise, while *trans*-regulatory mutations often decreased expression noise. Comparing these mutational effects to the observed effects of polymorphism on *TDH3* expression noise revealed that natural selection has acted to maintain the level of expression noise. Surprisingly, however, it has done this in two distinct ways: first, by removing *cis*-regulatory variants that increase expression noise and second, by removing *trans*-regulatory variants that decrease expression noise. As a consequence, the maintenance of *TDH3* expression noise is maintained by natural selection operating on separate regulatory mechanisms.

This observation is particularly interesting when combined with the growing evidence that increased regulation of gene expression levels is often accompanied by higher expression noise (JOTHI *et al.* 2009; DADIANI *et al.* 2013; VARDI *et al.* 2013; WOLF *et al.*

2015). While increased regulation and better matching of gene expression levels to the environment often increases fitness, increased regulation of expression levels often results in increased expression noise which can overcome this beneficial effect (WOLF *et al.* 2015). As a consequence, there is a tradeoff in increased regulation between expression levels and expression noise. Interestingly, this tradeoff is primarily restricted to *trans*-regulatory changes, such as the binding of new transcription factors. Consistent with this view, we find that new *trans*-regulatory mutations, which are likely to disrupt regulatory interactions, are skewed towards decreased noise. Because lower expression noise is often advantageous, *trans*-regulatory mutations that decrease the number of regulatory interactions may therefore occasionally be beneficial.

Compensatory changes in gene expression

Changes in gene expression are often stable across long evolutionary times. To understand the contribution of regulatory changes to this stability in gene expression levels, we compared the contribution of *cis* and *trans*-regulatory changes to differences in expression across several *Saccharomyces* species. These comparisons indicated the widespread occurrence of compensatory changes in regulation, such that changes in gene expression proceed much more slowly than changes in sequence or regulation. In particular, the data suggest widespread evidence of multiple *trans*-regulatory changes occurring in opposite directions, effectively canceling one another out. As a consequence, net *trans*-regulatory divergence as a proportion of total regulatory divergence decreases over time and *cis*-regulatory changes become the predominate mechanism for net changes in expression and regulation.

A high frequency of *trans*-regulatory mutations compensating for one another is consistent with our results in natural *S. cerevisiae* populations where we find evidence of hundreds of loci contributing to *trans*-regulatory changes in *TDH3* expression between natural strains of *S. cerevisiae*. These loci appear to affect TDH3 expression in both directions, resulting in small net changes in expression. Together, these results suggest that *trans*-acting changes in regulation are more common than often acknowledged. It is thus worth asking why this pattern was not previously detected by either eQTL methods nor estimates of regulatory divergence using allele specific expression measurements. Interesting, in both instances, the answer may largely be methodological: for eQTL studies, small changes in expression, as suggested here, are difficult to detect (YVERT *et al.* 2003), while regulatory divergence measurements are only able to capture net changes in *cis*- and *trans*-regulation, not total changes in regulation. As a consequence, small, compensating effects on expression are the most likely class of regulatory change to be missed.

In light of these observation, it would interesting to determine how quickly such *trans*-regulatory changes turn over in evolution. For example, identification of additional *trans*-acting eQTL in additional strains of *S. cerevisiae* and in different *Saccharomyces* species could shed light on whether the majority of these changes are regulatory 'noise' that result in short term differences in expression but little long term evolutionary change, or if they make substantial contributions to the evolution of gene expression in the long run. Other avenues towards pursuing this question include measuring allele specific

expression within F2 individuals to break apart *trans*-regulatory compensatory changes. Both approaches are feasible given the technology, but would both require considerable investment in developing new methodological and statistical tools for analysis.

In total, this work has shown how *cis* and *trans*-regulatory mutations vary in their effects on expression and how these differences can be acted on by natural selection to produce the patterns of expression and regulatory divergence observed in natural populations. The results of this work suggest large roles for changes in gene expression noise, compensatory changes in regulation, and large numbers of small effect *trans*-regulatory mutations in the evolution of gene expression. Collectively, these results highlight the biological insights that can be gained by considering the dual causes of molecular evolution, mutation and selection.

References

BULLARD J. H., MOSTOVOY Y., DUDOIT S., BREM R. B., 2010 Polygenic and directional regulatory evolution across pathways in Saccharomyces. PNAS **107**: 5058–63.

CHANG J., ZHOU Y., HU X., LAM L., HENRY C., GREEN E. M., KITA R., KOBOR M. S., FRASER H. B., 2013 The molecular mechanism of a cis-regulatory adaptation in yeast. PLoS Genet. **9**: e1003813.

COOLON J. D., MCMANUS C. J., STEVENSON K. R., GRAVELEY B. R., WITTKOPP P. J., 2014 Tempo and mode of regulatory evolution in Drosophila. Genome Res.: gr.163014.113–.

DADIANI M., DIJK D. VAN, SEGAL B., FIELD Y., BEN-ARTZI G., RAVEH-SADKA T., LEVO M., KAPLOW I., WEINBERGER A., SEGAL E., 2013 Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise. Genome Res. **23**: 966–976.

EMERSON J. J., HSIEH L.-C., SUNG H.-M., WANG T.-Y., HUANG C.-J., LU H. H.-S., LU M.-Y. J., WU S.-H., LI W.-H., 2010 Natural selection on cis and trans regulation in yeasts. Genome Res.: 826–836.

FRASER H. B., BABAK T., TSANG J., ZHOU Y., ZHANG B., MEHRABIAN M., SCHADT E. E., 2011 Systematic detection of polygenic cis-regulatory evolution. (MW Nachman, Ed.). PLoS Genet. **7**: e1002023.

FRASER H. B., LEVY S., CHAVAN A., SHAH H. B., PEREZ J. C., ZHOU Y., SIEGAL M. L., SINHA H., 2012 Polygenic cis-regulatory adaptation in the evolution of yeast pathogenicity. Genome Res. **22**: 1930–9.

FRASER H. B. H., MOSES A. M., SCHADT E. E., 2010 Evidence for widespread adaptive evolution of gene expression in budding yeast. PNAS **107**: 2977–82.

GONCALVES A., LEIGH-BROWN S., THYBERT D., STEFFLOVA K., TURRO E., FLICEK P., BRAZMA A., ODOM D. T., MARIONI J. C., 2012 Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res. **22**: 2376–2384.

JOTHI R., BALAJI S., WUSTER A., GROCHOW J. a, GSPONER J., PRZYTYCKA T. M., ARAVIND L., BABU M. M., 2009 Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Mol. Syst. Biol. **5**: 294.

KUO D., LICON K., BANDYOPADHYAY S., CHUANG R., LUO C., CATALANA J., RAVASI T., TAN K., IDEKER T., 2010 Coevolution within a transcriptional network by compensatory trans and cis mutations. Genome Res. **20**: 1672–8.

LANDRY C. R., WITTKOPP P. J., TAUBES C. H., RANZ J. M., CLARK A. G., HARTL D. L., 2005 Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. Genetics **171**: 1813–22.

SCHADT E. E. E., MONKS S. S. a, DRAKE T. T. a, LUSIS A. J. A., CHE N., COLINAYO V., RUFF T. G., MILLIGAN S. B., LAMB J. R., CAVET G., LINSLEY P. S., MAO M., STOUGHTON R. B., FRIEND S. H., 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **205**: 1–6.

SCHRAIBER J. G., MOSTOVOY Y., HSU T. Y., BREM R. B., 2013 Inferring evolutionary histories of pathway regulation from transcriptional profiling data. PLoS Comput. Biol. **9**: e1003255.

SMITH J. D., MCMANUS K. F., FRASER H. B., 2013 A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers. Mol. Biol. Evol. **30**: 2509–2518.

SZAMECZ B., BOROSS G., KALAPIS D., KOVÁCS K., FEKETE G., FARKAS Z., LÁZÁR V., HRTYAN M., KEMMEREN P., GROOT KOERKAMP M. J. a, RUTKAI E., HOLSTEGE F. C. P., PAPP B., PÁL C., 2014 The Genomic Landscape of Compensatory Evolution. PLoS Biol. **12**: e1001935.

TIROSH I., REIKHAV S., LEVY A. a, BARKAI N., 2009 A yeast hybrid provides insight into the evolution of gene expression regulation. Science **324**: 659–662.

VARDI N., LEVY S., ASSAF M., CARMI M., BARKAI N., 2013 Budding yeast escape commitment to the phosphate starvation program using gene expression noise. Curr. Biol. **23**: 2051–2057.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2008 Independent effects of cis- and trans-regulatory variation on gene expression in Drosophila melanogaster. Genetics **178**: 1831–5.

WOLF L., SILANDER O. K., NIMWEGEN E. J. VAN, 2015 Expression noise facilitates the evolution of gene regulation. Elife **4**: e05856.

YVERT G., BREM R. B., WHITTLE J., AKEY J. M., FOSS E., SMITH E. N., MACKELPRANG R., KRUGLYAK L., OTHERS, 2003 Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat. Genet. **35**: 57–64.

Appendix A

Supplementary Figures and Tables for Chapter 2

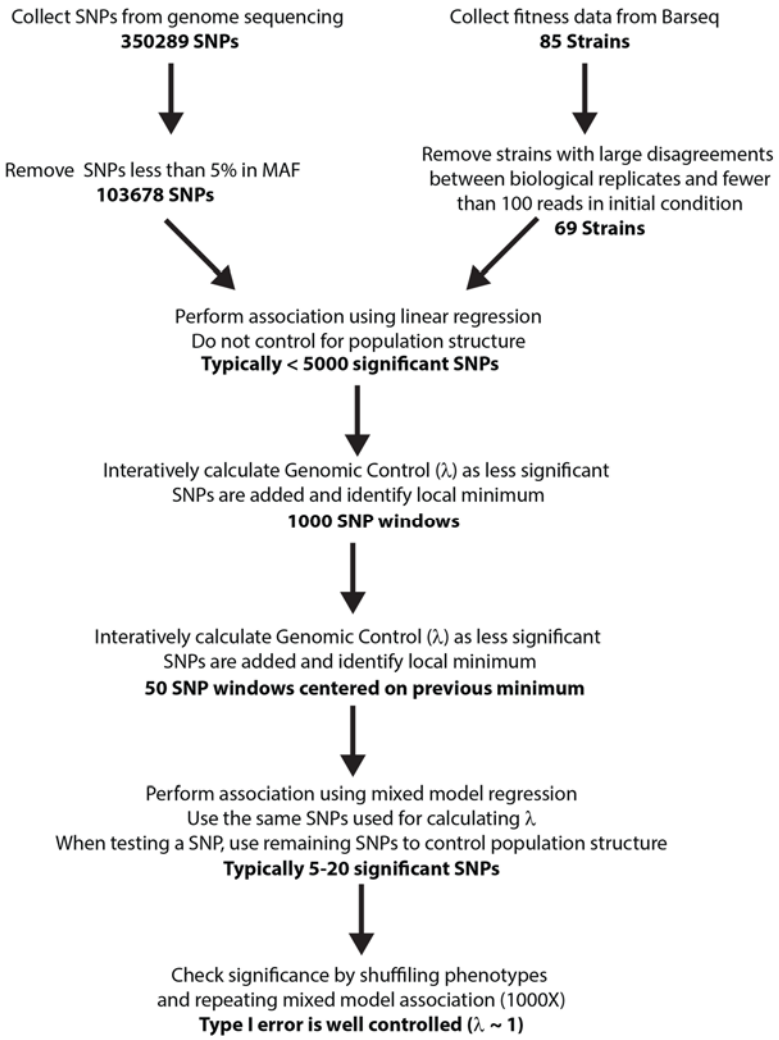


Figure A-1 Association study pipeline

Steps taken to determine SNPs significantly associated with growth rate across different environments. Each environment was run independently. Only SNPs with minor allele frequency greater than 5% in the subset of strains with quality phenotype data were used. Genomic control was used to assess whether population structure was appropriately controlled.

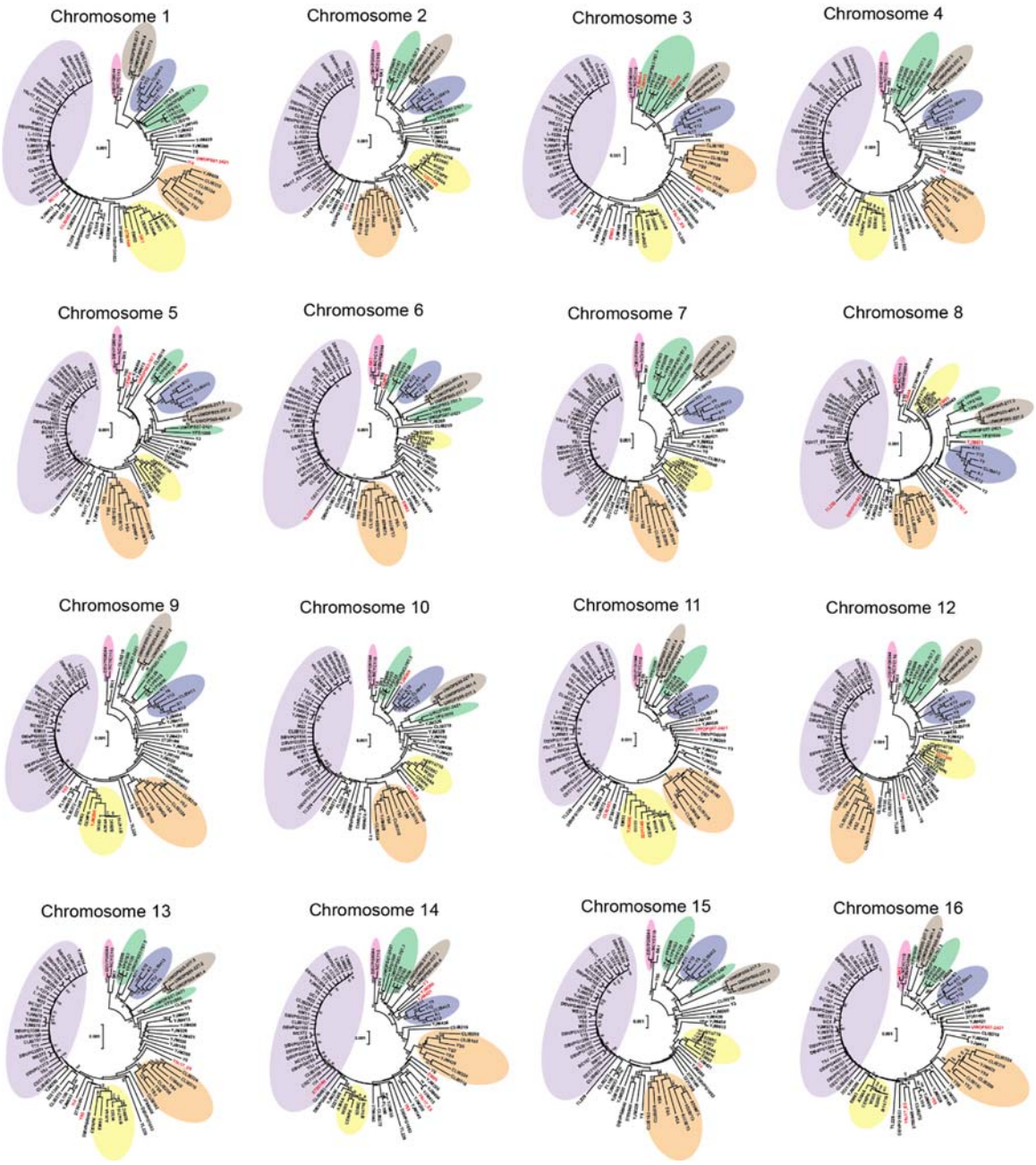


Figure A-2 Individual chromosome phylogenies

Neighbor joining phylogenies using all SNPs on a specific chromosome. Phylogenies are rooted in the same way as the entire genome phylogeny. Clades are identified based on whole genome clustering. Strains outlined in red show distinct placement from genome wide phylogeny.

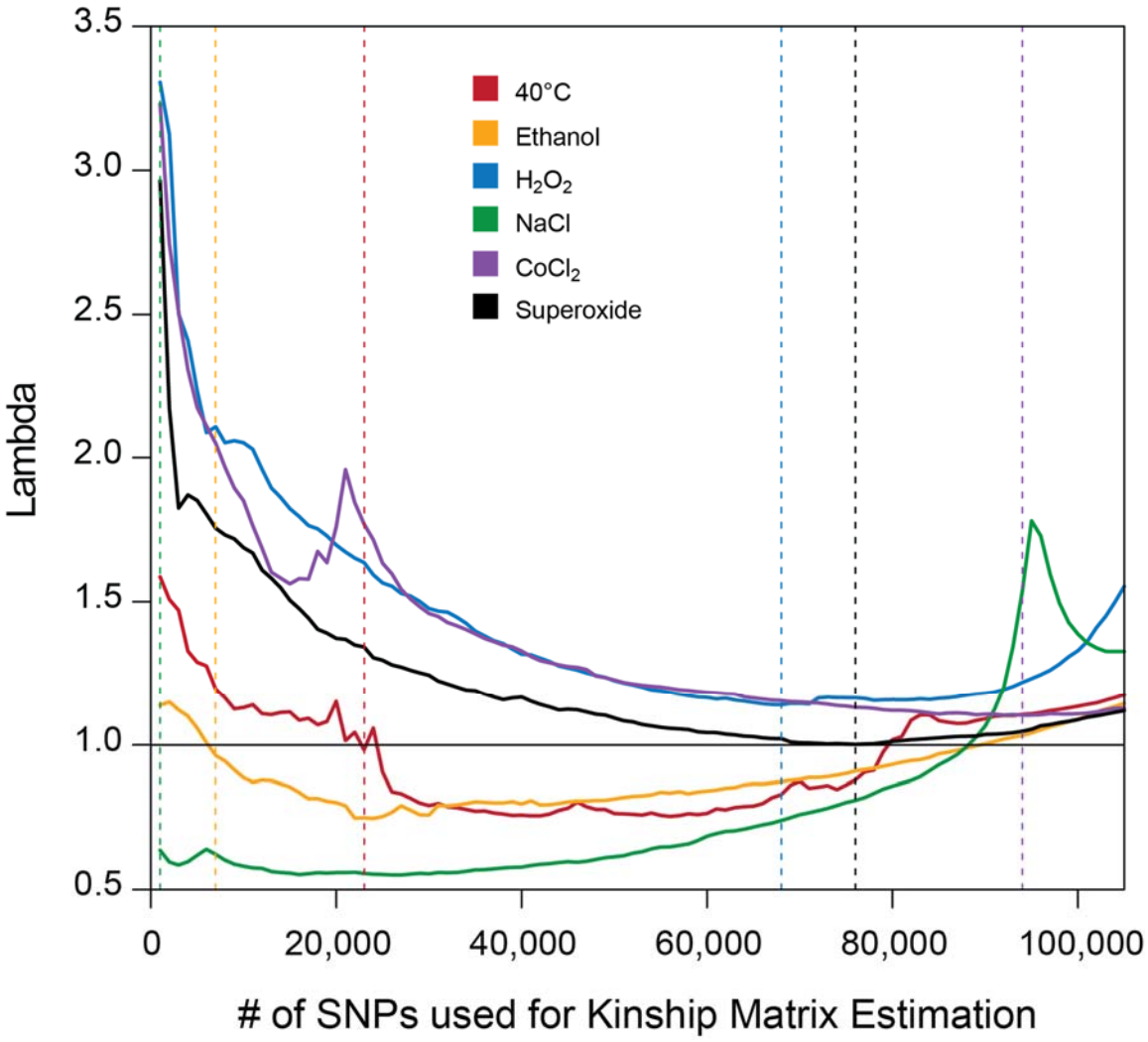


Figure A-3 Consequence of including additional SNPs when estimating a kinship matrix

Genomic control (y-axis) compared with the number of SNPs used in estimating the kinship matrix (x-axis). SNPs were ranked by p-value based on association not controlling for population structure and sequentially added in groups of 1000 to the previous sets. Analysis for each environment was performed independently. Dashed lines indicate the number of SNPs used in the final analysis to estimate the kinship matrix.

Strain	Location	Source	Strain Barcodes (5'-3')[a]		HO Homology[b]		Tractable Isolate ID (YCM)[c]				
			Uptag	Dstag	UP Group	DS Group	2n Het ho::KMX	MATa ho::KMX	MATα ho::KMX	MATa ho::HYG	MATα ho::NAT
DBVPG6765	Unknown	Unknown	CCGAGCTATTTCATGGCATT	AGCTCGACACGTATATCACG	2	1	760	761	762	1568	1900
SK1	USA	Soil	AGTACCGACTGCCACTGGAT	ATGTTAGGGCACTCGCCGAT	3	3	764	765	766	1343	1901
Y55	France	Wine	ATTGTAGGTCACGCGCCCAT	AATGATTGGTCCATCTCCCT	2	1	768	769	770	1344	1902
YPS128	Pennsylvania, USA	Oak	CGAGAGCGTTTCATATTGGT	AGATCGTTCTCATACGCCTT	3	3	772	773	774	1345	1903
DBVPG6044#	West Africa	Bili wine	TACACCACTTCATGTGAGGG	GAGTTTCTCTCAATATCCGC	2	3	776	-	-	-	-
DBVPG1788*	Finland	Soil	-	-	2	1	-	-	-	-	-
DBVPG1373	Netherland	soil	TTTCAGTCTGGCACCGAGGT	AATACTGACTTCCACCGTGT	2	1	784	785	786	1346	1904
DBVPG1853	Ethiopia	White Tecc	GACCTGGCGATTGTCACGTT	CCTTTACAGAGATTTGGAGG	2	1	788	789	790	1347	1905
BC187	Napa Valley, USA	Barrel fermentation	TAAGGAGCCATAACCTATCC	AAACTGGCAGGGTCCGAATT	2	1	792	793	794	1348	2754
YPS606	Pennsylvania, USA	Oak	CCGGGTATGTTACTAGAGTT	CCTGATGAGGCAGTTATGAT	3	3	796	797	798	1349	1906
L-1374	Chile	Wine	GTTACCTACTTCTACACGGT	CTCTATGCTAGGCCAGGATG	2	1	800	801	802	1350	1907
L-1528	Chile	Wine	GAAGCATAAGCATGTGTGAC	ATCTATGTACGTGGAGCTGT	2	1	804	805	806	1351	1908
Y12	Africa	Palm wine strain	ATCTAAGCTCGACTCAGGTG	CTTCATGCGAGATTAGTTGG	2	3	808	809	810	1352	1909
DBVPG1106	Australia	Grapes	GTATGACACTACCAGACAGC	CCGCAGAGGCAATAACCTGA	2	1	812	813	814	1353	1910
UWOPS83-787.3	Bahamas	Fruit, *Opuntia stricta*	CATTACGGGTGACCAGTGAT	TTGCATCGGGCAGTCGTTCT	2	1	816	817	818	1569	1911
UWOPS87-2421	Hawaii	Cladode, *O. megacantha*	ATATAACGCTCATGTCCCGT	TAGAGGTCTTGACACTGCGT	2	3	820	821	822	1354	1912
NCYC361	Ireland	Beer spoilage strain	AGGGCTGCTTCAGTCCTCTT	AGTGAGTCGTATTCAGCCTT	2	3	824	825	826	1570	1913
K11#	Japan	Shochu Sake strain	TAGCAGTTCCCACGTAGCTG	TTTCATGGACCGACGGGCAT	2	3	828	-	-	-	-
YS4#	Netherland	Baker strain	TTTCTATCGGGACGTGCGCT	TTATCTTCCCGAGCGGTGCT	2	2	832	-	-	-	-
YS9#	Singapore	Baker strain	TTGATCGACTCCCGCGTGTT	TCCGACCAGGAACAGTGAAC	2	2	836	-	-	-	-
322134S	RVI, Newcaslte UK	Clinical (Throat-sputum)	TAGTCTCGTACAGTCGGCTG	TACTTCAGTCCAGTCGCGTG	2	1	840	841	842	1355	1914
378604X#	RVI, Newcaslte UK	Clinical (Sputum)	GCCATGCAGCCAATCTAATA	AGCTAATGTCGAGTCACGCT	3	3	844	-	-	-	-
273614N	RVI, Newcaslte UK	Clinical (Fecal)	GCTATATGTATCCAGAGTGG	CCGCTCATTCAAGAGAACTA	2	1	848	849	450	1571	1915
YJM978	Bergamo, Italy	Clinical (Vaginal)	TTCACTGTGGGTAAGCTGAT	TGGACTGTACGTCACGCCAT	2	1	852	853	854	1357	1916

Strain	Location	Source	Strain Barcodes (5'-3')[a]		HO Homology[b]		Tractable Isolate ID (YCM)[c]				
			Uptag	Dstag	UP Group	DS Group	2n Het ho::KMX	MAT**a** ho::KMX	MATα ho::KMX	MAT**a** ho::HYG	MATα ho::NAT
Y9[#]	Japan	Ragi	TGATTCGGACGCGACTGTCT	AACTCCCTATGGATGACAGT	2	3	856	-	-	-	-
UWOPS03-461.4	Malaysia	Nectar, Bertam palm	ATATACCGCTAGGCGACTGT	ATTCACGTCGTAGCGAGCGT	4	3	860	861	862	1359	1917
UWOPS05-217.3	Malaysia	Nectar, Bertam palm	TGGACTGTCAGACTCGCGTT	ATGGATCTGGTCACCGAGTT	4	3	864	865	866	1360	1918
S288c[&]	California, USA	Rotting fig	AGTTACATCCCATGCGGTCG	TAGTCTACACGATCCGCAGG	1	1	868	869	870	1361	1919
W303[&]	Unknown	Unknown	ACTGAGCGATGCAGCGTCAT	AAGATTACGCGAAGCCAGCC	1	1	872	873	874	1362	1920
UWOPS05-227.2	Malaysia	Trigona, Bertam palm	GTCTTACAGAAAGCGCCACA	TTTATGGGCCGCACCGTCAT	4	3	876	877	878	1363	1921
DBVPG6040	Netherland	Fermenting fruit juice	TTCTTTGCAGTCGCCCAGGT	GATCTACTGTCACCCGTTGT	2	3	880	881	882	1364	1922
YIIc17_E5	Sauternes, Franxe	Wine	TGATTGAGCCTCCCGCGTTT	CCACTGAGGTTAAGTATAGG	2	1	884	885	886	1365	1923
YJM981	Bergamo, Italy	Clinical (Vaginal)	TACTACCACGAAGCAGGGAC	ATATGCGCTGAACGCTCTGC	2	1	888	889	890	1366	1924
YJM975	Bergamo, Italy	Clinical (Vaginal)	AGGGACCCGTTACCGGATTT	TCAGGAATAGCATGTGAGCC	2	1	892	893	894	1367	1925
NCYC110	West Africa	Ginger beer	CAGCAGGGCCAATCATTATA	AGTCAGGTTATGGCACATCT	3	3	896	897	898	1368	1926
YS2*	Australia	Baker strain	CCTTTGCTAGAGATGTAGTG	AATCTCAGCTAACGTCTCAG	2	1	-	-	-	-	-
CLIB192*	France	Baker	GCATATAGAGTCGCATAAGG	GTACAGCATGTCCTGAACTC	2	1	-	-	-	-	-
CLIB208 *	China	Baker	CCTACCGACAAATATGTGCA	CCGCTGGATCAACGAATATA	2	1	-	-	-	-	-
CLIB318*	Netherlands	Baker	CAGGATTATACGGATCGTCT	GGTATGTCTCCGCATCGTCT	2	1	-	-	-	-	-
CLIB324*	Vietnam	Baker	TTGATCTGCCTCACGGGTCT	TGATTCACGCTACGCGACGT	2	1	-	-	-	-	-
CLIB272	USA	Beer	GCTTCATTCTTGACAGACGT	GGATCTATGGTCACTGGATT	2	1	920	921	922	1371	1927
YJM145	Ireland	Beer	ACGACTTGATTCACTGGCTT	AGCAGGCTTGTCACTTGGTT	2	1	924	925	926	1372	1928
YJM280	USA	Clinical (Peritoneal fluid)	GGGTAGACCAGCTCACATTC	ATTCTTTGCCGACGGAGGCT	2	1	928	929	930	1373	1929
YJM320	USA	Clinical (Blood)	TACTGTCTGGCATACCGGGT	ATACGGTGGTCATTCGGCTT	2	1	932	933	934	1374	1930
YJM326	USA	Clinical (Unknown)	TAGTCTTAGCCAGTACGGTG	TACTCAGATCGGGTAATAGG	2	1	936	937	938	1375	1931
YJM413	USA	Clinical (Blood)	CTACGCCAGCGGATTTAAGG	AATTTAGGGTCACGCCGCCT	2	1	940	941	942	1376	1932
YJM421	USA	Clinical (Ascites fluid)	ATAACTCCTGGGACGACGGT	GGGTCACCGTTTCCTCAAGT	2	3	944	945	946	1377	1933
YJM428[#]	USA	Clinical (Paracenteis)	GGAGTAGCCGTAACCCTTTC	CGCTCATACGGATTATCTCT	2	1	948	-	-	-	-

Strain	Location	Source	Strain Barcodes (5'-3')[a]		HO Homology[b]		Tractable Isolate ID (YCM)[c]				
			Uptag	Dstag	UP Group	DS Group	2n Het ho::KMX	MAT**a** ho::KMX	MATα ho::KMX	MAT**a** ho::HYG	MATα ho::NAT
YJM434	Europe	Clinical (Unknown)	CGCTAGAGGTCATTCATACT	CACTATACTCGGCAGAGAGG	2	1	952	953	954	1378	1934
YJM436	Europe	Clinical (Mouth)	TGCATCAACTTCTGCTAAGG	GATCAAACAGTGTGGCAATC	2	1	956	957	958	1572	1935
YJM454	USA	Clinical (Blood)	CATCGGTTAGCAGGCCGTAT	TGAGTTCCTGCACCGTTGCT	2	1	960	961	962	1379	1936
YJM653*	USA	Clinical (Brochoalveolar)	-	-	2	1	-	-	-	-	-
CECT10109	Spain	Prickly pear	TCGCACCTTGTCCATAGATT	ACGACTCTATGACCTGTGTT	2	1	968	969	970	1380	1937
DBVPG1794*	Finland	Soil	-	-	2	1	-	-	-	-	-
DBVPG3591	Unknown	Cocoa beans	CCCGCTTTCAAATGTGCTAA	GTCGCAGGCAAACAGTCCAA	2	1	976	977	978	1381	1938
DBVPG4651	Italy	*Tuber Magnatum*	TACTTTCTGCGAGGGACGCT	TCCTTTCGCATAGAGGCCGT	2	1	980	981	982	1382	1939
EM93	California	Rotting Fig	ATTGGCTTCATGGCCCGGTT	AACCCACTAGGGAGCATAGG	1	1	984	985	986	1383	1940
TL229[#]	France	Cheese	ACCAGCGCGTTAAGTCTAGG	TTCCAGACCGAAGGAGACAC	2	1	988	-	-	-	-
Y6	French Guiana	Unknown	AAACCCGAGTGATCGACGGT	AACCTCGTATGGACGGTGTT	2	1	992	993	994	1385	1941
YPS1000	USA	Oak exudates	AGACCCGCTTAATCGGTTCG	ATGTATGTGCCATCCCGCTG	3	3	996	997	998	1386	1942
YPS163	USA	Oak exudates	ACGAGCGTCAGATCGAGTAT	CTATCGGCGCGACTGTATGT	3	3	1000	1001	1002	1387	1943
CLIB294	France	Distillery	TGACTAGACTGACCTCATTG	GCATGGACATAAGTGTAAGC	2	1	1004	1005	1006	1388	1944
CLIB413	China	Fermentation	CCGTGGATCAAATCAGTCAA	TATGTATCCCTGCACGGGCT	2	3	1008	1009	1010	1389	1945
CLIB483	France	Cider	AAGCCCTCTTCGATGATTGT	GAGGACACCCGATCTGATTC	2	1	1012	1013	1014	1390	1946
K12	Japan	Sake	GCCTGTCTATAACTAGCTTC	GGTATCACGCTACACCGGAT	2	3	1016	1017	1018	1391	1947
K1[#]	Japan	Sake	ATGATTTCGCCCGACGGTCT	TTTGCTTCCACGAGTGCCGT	2	3	1020	-	-	-	-
Y3*	Africa	Palm Wine	-	-	2	3	-	-	-	-	-
YJM269	Unknown	Apple juice	TTATGCTACCGGAGCCGAGT	TTGTTCCCGAGCCGAGCATT	2	3	1028	1029	1030	1392	1948
BY4716[&]	-	Laboratory	TATTGTCAGCCAGGTTCCCG	TATTCCCATCGGGTGCAAGG	1	1	1032	1033	1034	1393	1949
A364A	-	Laboratory	AAAGCGCCATTCTCCTAAGG	AAAGTCATTTCAACCGACGC	1	1	1036	1037	1038	1394	1950
CENPK	-	Laboratory	GCGACGACGGTATTACGACT	ACGGCTACCTTACTTCAGTT	1	1	1040	1041	1042	1395	1951
FL100	-	Laboratory	ACTTGTAGAGTAGGGCATCT	ATTTCACCGCGCTGGTAAGG	1	1	1044	1045	1046	1396	2753
CLIB154	Russia	Wine	CCCAGATATAGAGCGATAGG	GAAGTTCATGTCACGGATCT	2	1	1048	1049	1050	1397	1952

Strain	Location	Source	Strain Barcodes (5'-3')[a]		HO Homology[b]		Tractable Isolate ID (YCM)[c]				
			Uptag	Dstag	UP Group	DS Group	2n Het ho::KMX	MAT**a** ho::KMX	MATα ho::KMX	MAT**a** ho::HYG	MATα ho::NAT
CLIB157	Spain	Wine	ATTGTCATCCGACGCGGTCT	ATCACGTCGGTCCAGATCGT	2	1	1052	1053	1054	1398	1953
CLIB219	Russia	Wine	AAGCTCCCTTGCCATGTGTT	TCAGCCGGTACAGTCCTATG	3	3	1056	1057	1058	1399	1954
DBVPG1399	Netherlands	Grape	ATATGATTTACCAGCGCGGG	TTACCTATGCCAGAGCGAGG	2	1	1060	1061	1062	1400	1955
I14	Italy	Vineyard soil	TGCCATTCGGAACTAGGCTC	CCCTTTAGATAACGATGCTC	2	1	1064	1065	1066	1401	1956
M22	Italy	Wine	CCCGGCATAGTAACAGATAC	TGACAAAGAAGAGTCGGTCT	2	1	1068	1069	1070	1402	1957
RM11[&]	California	Wine	TTTACGGAGCGCGATCAGCT	CTTATCACTGTATCCGACGT	*	*	1072	1073	1074	1573	1958
T73	Spain	Wine	TTATGGTCGCCGCCAGACTT	TTATCTCGCGCCGGAGTCTT	2	1	1076	1077	1078	1403	1959
UC1[#]	France	Wine	AAAGAGCAGTTCAACCGTCC	TCAGCTTACCGAAGTTAGCC	2	1	1080	-	-	-	-
UC8	South Africa	Wine	ATGATCTGCTGCCACCAGGT	TATGTTGCACCATCACTGGC	2	1	1084	1085	1086	1404	1960
WE372	South Africa	Wine	TGATTGCAGACATATCCTCC	CAGCTCGTGTCAGTCGTATT	2	1	1088	1089	1090	1405	1961
Y8[#]	Turkey	Wine	GCGTGATGCTCAGGTATATT	GCTAGTACAGTATGAGACAG	2	1	1092	-	-	-	-
Y9J	Japan	Wine	CCGATCTTCGCACTTAGTAT	CGGAGCTTCGCCATCTTTAT	2	1	1096	1097	1098	1407	1962

Table A-1 Sequenced strain details

The geographic location (Location) and the environment/substrate from which the strains were isolated (Source) are listed where known. a = The unique strain identifying barcodes inserted into each strain. UPtag and DStag sequences, usable for strain identification and quantification by Bar-seq, are listed where successfully introduced into the HO locus of a strain. b = The identity of the homology group sequences added during PCR to target the HO regions with the barcoded drug resistance cassette in order to maintain SNPs in the region. c = The Identifier number (YCMXXXX) given to each of the mating-type/drug marker combinations for each strain. Each strain derived from the same parental genotype carries the same UPtag and DStag. Diploid strains with a single copy of HO replaced (2n Het ho::KMX) and haploid derivatives of these with opposing mating types (MATa ho::KMX and MATα ho::KMX) were made. Drug marker switches were then carried out to produce easily crossed strains that allow selection of diploids (MATa ho::HYG and MATα ho::NAT). Not all of the sequenced strains were able to produce tractable strains for various technical reasons. Strain names denoted with # were either naturally immune to drugs or produced no transformants after several attempts and therefore do not contain barcodes. Strain names marked * were heterozygously marked but found to be unable to sporulate or their fertility was very low. & denoted strains are those initially heterothallic haploids from which diploids were formed by mating type switching though plasmid borne copies of HO.

S. cerevisiae Parental Strain	Designation	oligo (5'-3')	Primer sequence (5' - 3')
DBVPG6765	UPTAG_1	CCGAGCTATTTCATGGCATT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CCGAGCTATTTCATGGCATT/CGTACGCTGCAGGTCGAC
	DSTAG_1	AGCTCGACACGTATATCACG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AGCTCGACACGTATATCACG/ATCGATGAATTCGAGCTCG
SK1	UPTAG_2	AGTACCGACTGCCACTGGAT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AGTACCGACTGCCACTGGAT/CGTACGCTGCAGGTCGAC
	DSTAG_2	ATGTTAGGGCACTCGCCGAT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATGTTAGGGCACTCGCCGAT/ATCGATGAATTCGAGCTCG
Y55	UPTAG_3	ATTGTAGGTCACGCGCCCAT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATTGTAGGTCACGCGCCCAT/CGTACGCTGCAGGTCGAC
	DSTAG_3	AATGATTGGTCCATCTCCCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AATGATTGGTCCATCTCCCT/ATCGATGAATTCGAGCTCG
YPS128	UPTAG_4	CGAGAGCGTTTCATATTGGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CGAGAGCGTTTCATATTGGT/CGTACGCTGCAGGTCGAC
	DSTAG_4	AGATCGTTCTCATACGCCTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AGATCGTTCTCATACGCCTT/ATCGATGAATTCGAGCTCG
DBVPG6044	UPTAG_5	TACACCACTTCATGTGAGGG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TACACCACTTCATGTGAGGG/CGTACGCTGCAGGTCGAC
	DSTAG_5	GAGTTTCTCTCAATATCCGC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GAGTTTCTCTCAATATCCGC/ATCGATGAATTCGAGCTCG
DBVPG1373	UPTAG_7	TTTCAGTCTGGCACCGAGGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTTCAGTCTGGCACCGAGGT/CGTACGCTGCAGGTCGAC
	DSTAG_7	AATACTGACTTCCACCGTGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AATACTGACTTCCACCGTGT/ATCGATGAATTCGAGCTCG
DBVPG1853	UPTAG_8	GACCTGGCGATTGTCACGTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GACCTGGCGATTGTCACGTT/CGTACGCTGCAGGTCGAC
	DSTAG_8	CCTTTACAGAGATTTGGAGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CCTTTACAGAGATTTGGAGG/ATCGATGAATTCGAGCTCG
BC187	UPTAG_9	TAAGGAGCCATAACCTATCC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TAAGGAGCCATAACCTATCC/CGTACGCTGCAGGTCGAC
	DSTAG_9	AAACTGGCAGGGTCCGAATT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AAACTGGCAGGGTCCGAATT/ATCGATGAATTCGAGCTCG
YPS606	UPTAG_10	CCGGGTATGTTACTAGAGTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CCGGGTATGTTACTAGAGTT/CGTACGCTGCAGGTCGAC
	DSTAG_10	CCTGATGAGGCAGTTATGAT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CCTGATGAGGCAGTTATGAT/ATCGATGAATTCGAGCTCG
L-1374	UPTAG_11	GTTACCTACTTCTACACGGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GTTACCTACTTCTACACGGT/CGTACGCTGCAGGTCGAC
	DSTAG_11	CTCTATGCTAGGCCAGGATG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CTCTATGCTAGGCCAGGATG/ATCGATGAATTCGAGCTCG
L-1528	UPTAG_12	GAAGCATAAGCATGTGTGAC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GAAGCATAAGCATGTGTGAC/CGTACGCTGCAGGTCGAC
	DSTAG_12	ATCTATGTACGTGGAGCTGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATCTATGTACGTGGAGCTGT/ATCGATGAATTCGAGCTCG
Y12	UPTAG_13	ATCTAAGCTCGACTCAGGTG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATCTAAGCTCGACTCAGGTG/CGTACGCTGCAGGTCGAC
	DSTAG_13	CTTCATGCGAGATTAGTTGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CTTCATGCGAGATTAGTTGG/ATCGATGAATTCGAGCTCG
DBVPG1106	UPTAG_14	GTATGACACTACCAGACAGC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GTATGACACTACCAGACAGC/CGTACGCTGCAGGTCGAC
	DSTAG_14	CCGCAGAGGCAATAACCTGA	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CCGCAGAGGCAATAACCTGA/ATCGATGAATTCGAGCTCG
UWOPS83-787.3	UPTAG_15	CATTACGGGTGACCAGTGAT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CATTACGGGTGACCAGTGAT/CGTACGCTGCAGGTCGAC
	DSTAG_15	TTGCATCGGGCAGTCGTTCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTGCATCGGGCAGTCGTTCT/ATCGATGAATTCGAGCTCG
UWOPS87-2421	UPTAG_16	ATATAACGCTCATGTCCCGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATATAACGCTCATGTCCCGT/CGTACGCTGCAGGTCGAC
	DSTAG_16	TAGAGGTCTTGACACTGCGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TAGAGGTCTTGACACTGCGT/ATCGATGAATTCGAGCTCG
NCYC361	UPTAG_17	AGGGCTGCTTCAGTCCTCTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AGGGCTGCTTCAGTCCTCTT/CGTACGCTGCAGGTCGAC
	DSTAG_17	AGTGAGTCGTATTCAGCCTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AGTGAGTCGTATTCAGCCTT/ATCGATGAATTCGAGCTCG
K11	UPTAG_18	TAGCAGTTCCCACGTAGCTG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TAGCAGTTCCCACGTAGCTG/CGTACGCTGCAGGTCGAC
	DSTAG_18	TTTCATGGACCGACGGGCAT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTTCATGGACCGACGGGCAT/ATCGATGAATTCGAGCTCG
YS4	UPTAG_19	TTTCTATCGGGACGTGCGCT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTTCTATCGGGACGTGCGCT/CGTACGCTGCAGGTCGAC
	DSTAG_19	TTATCTTCCCGAGCGGTGCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTATCTTCCCGAGCGGTGCT/ATCGATGAATTCGAGCTCG

S. cerevisiae Parental Strain	Designation	oligo (5'-3')	Primer sequence (5' - 3')
YS9	UPTAG_20	TTGATCGACTCCCGCGTGTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTGATCGACTCCCGCGTGTT/CGTACGCTGCAGGTCGAC
	DSTAG_20	TCCGACCAGGAACAGTGAAC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TCCGACCAGGAACAGTGAAC/ATCGATGAATTCGAGCTCG
322134S	UPTAG_54	TAGTCTCGTACAGTCGGCTG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TAGTCTCGTACAGTCGGCTG/CGTACGCTGCAGGTCGAC
	DSTAG_54	TACTTCAGTCCAGTCGCGTG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TACTTCAGTCCAGTCGCGTG/ATCGATGAATTCGAGCTCG
378604X	UPTAG_22	GCCATGCAGCCAATCTAATA	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GCCATGCAGCCAATCTAATA/CGTACGCTGCAGGTCGAC
	DSTAG_22	AGCTAATGTCGAGTCACGCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AGCTAATGTCGAGTCACGCT/ATCGATGAATTCGAGCTCG
273614N	UPTAG_23	GCTATATGTATCCAGAGTGG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GCTATATGTATCCAGAGTGG/CGTACGCTGCAGGTCGAC
	DSTAG_23	CCGCTCATTCAAGAGAACTA	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CCGCTCATTCAAGAGAACTA/ATCGATGAATTCGAGCTCG
YJM978	UPTAG_24	TTCACTGTGGGTAAGCTGAT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTCACTGTGGGTAAGCTGAT/CGTACGCTGCAGGTCGAC
	DSTAG_24	TGGACTGTACGTCACGCCAT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TGGACTGTACGTCACGCCAT/ATCGATGAATTCGAGCTCG
Y9	UPTAG_25	TGATTCGGACGCGACTGTCT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TGATTCGGACGCGACTGTCT/CGTACGCTGCAGGTCGAC
	DSTAG_25	AACTCCCTATGGATGACAGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AACTCCCTATGGATGACAGT/ATCGATGAATTCGAGCTCG
UWOPS03-461.4	UPTAG_26	ATATACCGCTAGGCGACTGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATATACCGCTAGGCGACTGT/CGTACGCTGCAGGTCGAC
	DSTAG_26	ATTCACGTCGTAGCGAGCGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATTCACGTCGTAGCGAGCGT/ATCGATGAATTCGAGCTCG
UWOPS05-217.3	UPTAG_27	TGGACTGTCAGACTCGCGTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TGGACTGTCAGACTCGCGTT/CGTACGCTGCAGGTCGAC
	DSTAG_27	ATGGATCTGGTCACCGAGTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATGGATCTGGTCACCGAGTT/ATCGATGAATTCGAGCTCG
S288c	UPTAG_28	AGTTACATCCCATGCGGTCG	TATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AGTTACATCCCATGCGGTCG/CGTACGCTGCAGGTCGAC
	DSTAG_28	TAGTCTACACGATCCGCAGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TAGTCTACACGATCCGCAGG/ATCGATGAATTCGAGCTCG
W303	UPTAG_29	ACTGAGCGATGCAGCGTCAT	TATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ACTGAGCGATGCAGCGTCAT/CGTACGCTGCAGGTCGAC
	DSTAG_29	AAGATTACGCGAAGCCAGCC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AAGATTACGCGAAGCCAGCC/ATCGATGAATTCGAGCTCG
UWOPS05-227.2	UPTAG_30	GTCTTACAGAAAGCGCCACA	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GTCTTACAGAAAGCGCCACA/CGTACGCTGCAGGTCGAC
	DSTAG_30	TTTATGGGCCGCACCGTCAT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTTATGGGCCGCACCGTCAT/ATCGATGAATTCGAGCTCG
DBVPG6040	UPTAG_31	TTCTTTGCAGTCGCCCAGGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTCTTTGCAGTCGCCCAGGT/CGTACGCTGCAGGTCGAC
	DSTAG_31	GATCTACTGTCACCCGTTGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GATCTACTGTCACCCGTTGT/ATCGATGAATTCGAGCTCG
YIIc17_E5	UPTAG_32	TGATTGAGCCTCCCGCGTTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TGATTGAGCCTCCCGCGTTT/CGTACGCTGCAGGTCGAC
	DSTAG_32	CCACTGAGGTTAAGTATAGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CCACTGAGGTTAAGTATAGG/ATCGATGAATTCGAGCTCG
YJM981	UPTAG_33	TACTACCACGAAGCAGGGAC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TACTACCACGAAGCAGGGAC/CGTACGCTGCAGGTCGAC
	DSTAG_33	ATATGCGCTGAACGCTCTGC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATATGCGCTGAACGCTCTGC/ATCGATGAATTCGAGCTCG
YJM975	UPTAG_34	AGGGACCCGTTACCGGATTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AGGGACCCGTTACCGGATTT/CGTACGCTGCAGGTCGAC
	DSTAG_34	TCAGGAATAGCATGTGAGCC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TCAGGAATAGCATGTGAGCC/ATCGATGAATTCGAGCTCG
NCYC110	UPTAG_35	CAGCAGGGCCAATCATTATA	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CAGCAGGGCCAATCATTATA/CGTACGCTGCAGGTCGAC
	DSTAG_35	AGTCAGGTTATGGCACATCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AGTCAGGTTATGGCACATCT/ATCGATGAATTCGAGCTCG
YS2	UPTAG_36	CCTTTGCTAGAGATGTAGTG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CCTTTGCTAGAGATGTAGTG/CGTACGCTGCAGGTCGAC
	DSTAG_36	AATCTCAGCTAACGTCTCAG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AATCTCAGCTAACGTCTCAG/ATCGATGAATTCGAGCTCG
CLIB192	UPTAG_37	GCATATAGAGTCGCATAAGG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GCATATAGAGTCGCATAAGG/CGTACGCTGCAGGTCGAC
	DSTAG_37	GTACAGCATGTCCTGAACTC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GTACAGCATGTCCTGAACTC/ATCGATGAATTCGAGCTCG

S. cerevisiae Parental Strain	Designation	oligo (5'-3')	Primer sequence (5' - 3')
CLIB208	UPTAG_38	CCTACCGACAAATATGTGCA	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CCTACCGACAAATATGTGCA/CGTACGCTGCAGGTCGAC
	DSTAG_38	CCGCTGGATCAACGAATATA	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CCGCTGGATCAACGAATATA/ATCGATGAATTCGAGCTCG
CLIB318	UPTAG_39	CAGGATTATACGGATCGTCT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CAGGATTATACGGATCGTCT/CGTACGCTGCAGGTCGAC
	DSTAG_39	GGTATGTCTCCGCATCGTCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GGTATGTCTCCGCATCGTCT/ATCGATGAATTCGAGCTCG
CLIB324	UPTAG_40	TTGATCTGCCTCACGGGTCT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTGATCTGCCTCACGGGTCT/CGTACGCTGCAGGTCGAC
	DSTAG_40	TGATTCACGCTACGCGACGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TGATTCACGCTACGCGACGT/ATCGATGAATTCGAGCTCG
CLIB272	UPTAG_41	GCTTCATTCTTGACAGACGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GCTTCATTCTTGACAGACGT/CGTACGCTGCAGGTCGAC
	DSTAG_41	GGATCTATGGTCACTGGATT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GGATCTATGGTCACTGGATT/ATCGATGAATTCGAGCTCG
YJM145	UPTAG_42	ACGACTTGATTCACTGGCTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ACGACTTGATTCACTGGCTT/CGTACGCTGCAGGTCGAC
	DSTAG_42	AGCAGGCTTGTCACTTGGTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AGCAGGCTTGTCACTTGGTT/ATCGATGAATTCGAGCTCG
YJM280	UPTAG_43	GGGTAGACCAGCTCACATTC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GGGTAGACCAGCTCACATTC/CGTACGCTGCAGGTCGAC
	DSTAG_43	ATTCTTTGCCGACGGAGGCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATTCTTTGCCGACGGAGGCT/ATCGATGAATTCGAGCTCG
YJM320	UPTAG_44	TACTGTCTGGCATACCGGGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TACTGTCTGGCATACCGGGT/CGTACGCTGCAGGTCGAC
	DSTAG_44	ATACGGTGGTCATTCGGCTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATACGGTGGTCATTCGGCTT/ATCGATGAATTCGAGCTCG
YJM326	UPTAG_45	TAGTCTTAGCCAGTACGGTG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TAGTCTTAGCCAGTACGGTG/CGTACGCTGCAGGTCGAC
	DSTAG_45	TACTCAGATCGGGTAATAGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TACTCAGATCGGGTAATAGG/ATCGATGAATTCGAGCTCG
YJM413	UPTAG_46	CTACGCCAGCGGATTTAAGG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CTACGCCAGCGGATTTAAGG/CGTACGCTGCAGGTCGAC
	DSTAG_46	AATTTAGGGTCACGCCGCCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AATTTAGGGTCACGCCGCCT/ATCGATGAATTCGAGCTCG
YJM421	UPTAG_47	ATAACTCCTGGGACGACGGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATAACTCCTGGGACGACGGT/CGTACGCTGCAGGTCGAC
	DSTAG_47	GGGTCACCGTTTCCTCAAGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GGGTCACCGTTTCCTCAAGT/ATCGATGAATTCGAGCTCG
YJM428	UPTAG_48	GGAGTAGCCGTAACCCTTTC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GGAGTAGCCGTAACCCTTTC/CGTACGCTGCAGGTCGAC
	DSTAG_48	CGCTCATACGGATTATCTCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CGCTCATACGGATTATCTCT/ATCGATGAATTCGAGCTCG
YJM434	UPTAG_49	CGCTAGAGGTCATTCATACT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CGCTAGAGGTCATTCATACT/CGTACGCTGCAGGTCGAC
	DSTAG_49	CACTATACTCGGCAGAGAGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CACTATACTCGGCAGAGAGG/ATCGATGAATTCGAGCTCG
YJM436	UPTAG_50	TGCATCAACTTCTGCTAAGG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TGCATCAACTTCTGCTAAGG/CGTACGCTGCAGGTCGAC
	DSTAG_50	GATCAAACAGTGTGGCAATC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GATCAAACAGTGTGGCAATC/ATCGATGAATTCGAGCTCG
YJM454	UPTAG_51	CATCGGTTAGCAGGCCGTAT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CATCGGTTAGCAGGCCGTAT/CGTACGCTGCAGGTCGAC
	DSTAG_51	TGAGTTCCTGCACCGTTGCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TGAGTTCCTGCACCGTTGCT/ATCGATGAATTCGAGCTCG
CECT10109	UPTAG_53	TCGCACCTTGTCCATAGATT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TCGCACCTTGTCCATAGATT/CGTACGCTGCAGGTCGAC
	DSTAG_53	ACGACTCTATGACCTGTGTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ACGACTCTATGACCTGTGTT/ATCGATGAATTCGAGCTCG
DBVPG3591	UPTAG_6	CCCGCTTTCAAATGTGCTAA	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CCCGCTTTCAAATGTGCTAA/CGTACGCTGCAGGTCGAC
	DSTAG_6	GTCGCAGGCAAACAGTCCAA	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GTCGCAGGCAAACAGTCCAA/ATCGATGAATTCGAGCTCG
DBVPG4651	UPTAG_56	TACTTTCTGCGAGGGACGCT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TACTTTCTGCGAGGGACGCT/CGTACGCTGCAGGTCGAC
	DSTAG_56	TCCTTTCGCATAGAGGCCGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TCCTTTCGCATAGAGGCCGT/ATCGATGAATTCGAGCTCG
EM93	UPTAG_57	ATTGGCTTCATGGCCCGGTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATTGGCTTCATGGCCCGGTT/CGTACGCTGCAGGTCGAC
	DSTAG_57	AACCCACTAGGGAGCATAGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AACCCACTAGGGAGCATAGG/ATCGATGAATTCGAGCTCG

S. cerevisiae Parental Strain	Designation	oligo (5'-3')	Primer sequence (5' - 3')
TL229	UPTAG_58	ACCAGCGCGTTAAGTCTAGG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ACCAGCGCGTTAAGTCTAGG/CGTACGCTGCAGGTCGAC
	DSTAG_58	TTCCAGACCGAAGGAGACAC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTCCAGACCGAAGGAGACAC/ATCGATGAATTCGAGCTCG
Y6	UPTAG_59	AAACCCGAGTGATCGACGGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AAACCCGAGTGATCGACGGT/CGTACGCTGCAGGTCGAC
	DSTAG_59	AACCTCGTATGGACGGTGTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AACCTCGTATGGACGGTGTT/ATCGATGAATTCGAGCTCG
YPS1000	UPTAG_60	AGACCCGCTTAATCGGTTCG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AGACCCGCTTAATCGGTTCG/CGTACGCTGCAGGTCGAC
	DSTAG_60	ATGTATGTGCCATCCCGCTG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATGTATGTGCCATCCCGCTG/ATCGATGAATTCGAGCTCG
YPS163	UPTAG_61	ACGAGCGTCAGATCGAGTAT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ACGAGCGTCAGATCGAGTAT/CGTACGCTGCAGGTCGAC
	DSTAG_61	CTATCGGCGCGACTGTATGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CTATCGGCGCGACTGTATGT/ATCGATGAATTCGAGCTCG
CLIB294	UPTAG_62	TGACTAGACTGACCTCATTG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TGACTAGACTGACCTCATTG/CGTACGCTGCAGGTCGAC
	DSTAG_62	GCATGGACATAAGTGTAAGC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GCATGGACATAAGTGTAAGC/ATCGATGAATTCGAGCTCG
CLIB413	UPTAG_63	CCGTGGATCAAATCAGTCAA	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CCGTGGATCAAATCAGTCAA/CGTACGCTGCAGGTCGAC
	DSTAG_63	TATGTATCCCTGCACGGGCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TATGTATCCCTGCACGGGCT/ATCGATGAATTCGAGCTCG
CLIB483	UPTAG_64	AAGCCCTCTTCGATGATTGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AAGCCCTCTTCGATGATTGT/CGTACGCTGCAGGTCGAC
	DSTAG_64	GAGGACACCCGATCTGATTC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GAGGACACCCGATCTGATTC/ATCGATGAATTCGAGCTCG
K12	UPTAG_65	GCCTGTCTATAACTAGCTTC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GCCTGTCTATAACTAGCTTC/CGTACGCTGCAGGTCGAC
	DSTAG_65	GGTATCACGCTACACCGGAT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GGTATCACGCTACACCGGAT/ATCGATGAATTCGAGCTCG
K1	UPTAG_66	ATGATTTCGCCCGACGGTCT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATGATTTCGCCCGACGGTCT/CGTACGCTGCAGGTCGAC
	DSTAG_66	TTTGCTTCCACGAGTGCCGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTTGCTTCCACGAGTGCCGT/ATCGATGAATTCGAGCTCG
Y3	UPTAG_67	TTTGCGTACCGCCTAGCTGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTTGCGTACCGCCTAGCTGT/CGTACGCTGCAGGTCGAC
	DSTAG_67	TTAGGTCGGCGCACTAGCTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTAGGTCGGCGCACTAGCTT/ATCGATGAATTCGAGCTCG
YJM269	UPTAG_79	TTATGCTACCGGAGCCGAGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTATGCTACCGGAGCCGAGT/CGTACGCTGCAGGTCGAC
	DSTAG_79	TTGTTCCCGAGCCGAGCATT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTGTTCCCGAGCCGAGCATT/ATCGATGAATTCGAGCTCG
BY4716	UPTAG_69	TATTGTCAGCCAGGTTCCCG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TATTGTCAGCCAGGTTCCCG/CGTACGCTGCAGGTCGAC
	DSTAG_69	TATTCCCATCGGGTGCAAGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TATTCCCATCGGGTGCAAGG/ATCGATGAATTCGAGCTCG
A364A	UPTAG_70	AAAGCGCCATTCTCCTAAGG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AAAGCGCCATTCTCCTAAGG/CGTACGCTGCAGGTCGAC
	DSTAG_70	AAAGTCATTTCAACCGACGC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/AAAGTCATTTCAACCGACGC/ATCGATGAATTCGAGCTCG
CENPK	UPTAG_71	GCGACGACGGTATTACGACT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GCGACGACGGTATTACGACT/CGTACGCTGCAGGTCGAC
	DSTAG_71	ACGGCTACCTTACTTCAGTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ACGGCTACCTTACTTCAGTT/ATCGATGAATTCGAGCTCG
Fl100	UPTAG_86	ACTTGTAGAGTAGGGCATCT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ACTTGTAGAGTAGGGCATCT/CGTACGCTGCAGGTCGAC
	DSTAG_86	ATTTCACCGCGCTGGTAAGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATTTCACCGCGCTGGTAAGG/ATCGATGAATTCGAGCTCG
CLIB154	UPTAG_73	CCCAGATATAGAGCGATAGG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CCCAGATATAGAGCGATAGG/CGTACGCTGCAGGTCGAC
	DSTAG_73	GAAGTTCATGTCACGGATCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GAAGTTCATGTCACGGATCT/ATCGATGAATTCGAGCTCG
CLIB157	UPTAG_74	ATTGTCATCCGACGCGGTCT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATTGTCATCCGACGCGGTCT/CGTACGCTGCAGGTCGAC
	DSTAG_74	ATCACGTCGGTCCAGATCGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/ATCACGTCGGTCCAGATCGT/ATCGATGAATTCGAGCTCG
CLIB219	UPTAG_75	AAGCTCCCTTGCCATGTGTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AAGCTCCCTTGCCATGTGTT/CGTACGCTGCAGGTCGAC
	DSTAG_75	TCAGCCGGTACAGTCCTATG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TCAGCCGGTACAGTCCTATG/ATCGATGAATTCGAGCTCG

S. cerevisiae Parental Strain	Designation	oligo (5'-3')	Primer sequence (5' - 3')
DBVPG1399	UPTAG_76	ATATGATTTACCAGCGCGGG	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATATGATTTACCAGCGCGGG/CGTACGCTGCAGGTCGAC
	DSTAG_76	TTACCTATGCCAGAGCGAGG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTACCTATGCCAGAGCGAGG/ATCGATGAATTCGAGCTCG
I14	UPTAG_77	TGCCATTCGGAACTAGGCTC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TGCCATTCGGAACTAGGCTC/CGTACGCTGCAGGTCGAC
	DSTAG_77	CCCTTTAGATAACGATGCTC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CCCTTTAGATAACGATGCTC/ATCGATGAATTCGAGCTCG
M22	UPTAG_78	CCCGGCATAGTAACAGATAC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CCCGGCATAGTAACAGATAC/CGTACGCTGCAGGTCGAC
	DSTAG_78	TGACAAAGAAGAGTCGGTCT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TGACAAAGAAGAGTCGGTCT/ATCGATGAATTCGAGCTCG
RM11	UPTAG_52	TTTACGGAGCGCGATCAGCT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTTACGGAGCGCGATCAGCT/CGTACGCTGCAGGTCGAC
	DSTAG_52	CTTATCACTGTATCCGACGT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CTTATCACTGTATCCGACGT/ATCGATGAATTCGAGCTCG
T73	UPTAG_80	TTATGGTCGCCGCCAGACTT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TTATGGTCGCCGCCAGACTT/CGTACGCTGCAGGTCGAC
	DSTAG_80	TTATCTCGCGCCGGAGTCTT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TTATCTCGCGCCGGAGTCTT/ATCGATGAATTCGAGCTCG
UC1	UPTAG_81	AAAGAGCAGTTCAACCGTCC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/AAAGAGCAGTTCAACCGTCC/CGTACGCTGCAGGTCGAC
	DSTAG_81	TCAGCTTACCGAAGTTAGCC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TCAGCTTACCGAAGTTAGCC/ATCGATGAATTCGAGCTCG
UC8	UPTAG_82	ATGATCTGCTGCCACCAGGT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/ATGATCTGCTGCCACCAGGT/CGTACGCTGCAGGTCGAC
	DSTAG_82	TATGTTGCACCATCACTGGC	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/TATGTTGCACCATCACTGGC/ATCGATGAATTCGAGCTCG
WE372	UPTAG_83	TGATTGCAGACATATCCTCC	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/TGATTGCAGACATATCCTCC/CGTACGCTGCAGGTCGAC
	DSTAG_83	CAGCTCGTGTCAGTCGTATT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CAGCTCGTGTCAGTCGTATT/ATCGATGAATTCGAGCTCG
Y8	UPTAG_84	GCGTGATGCTCAGGTATATT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/GCGTGATGCTCAGGTATATT/CGTACGCTGCAGGTCGAC
	DSTAG_84	GCTAGTACAGTATGAGACAG	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/GCTAGTACAGTATGAGACAG/ATCGATGAATTCGAGCTCG
Y9J	UPTAG_85	CCGATCTTCGCACTTAGTAT	CATCTATACTTTAAAATG/GATGTCCACGAGGTCTCT/CCGATCTTCGCACTTAGTAT/CGTACGCTGCAGGTCGAC
	DSTAG_85	CGGAGCTTCGCCATCTTTAT	ACTAATATACACATTTTA/CGGTGTCGGTCTCGTAG/CGGAGCTTCGCCATCTTTAT/ATCGATGAATTCGAGCTCG

Table A-2 Barcode and primer sequences

Uptag and Dstag barcodes assigned to each strain. Not all strains were successfully barcoded. Primer sequence gives the sequence of each barcode in context of the upstream and downstream sequences needed for amplification and targeting to the HO gene.

STRAIN	# RAW READS	# MAPPABLE	% MAPPABLE	# UNIQUE	% UNIQUE	# CONCORDANT UNIQUE	% CONCORDANT UNIQUE
DBVPG6765	5284304	5106683	0.9664	4136496	0.7828	3921718	0.7421
SK1	10540908	10031591	0.9517	7414727	0.7034	7154660	0.6788
Y55	11581982	11248211	0.9712	8956892	0.7733	7982302	0.6892
YPS128	9229216	9075458	0.9833	7522372	0.8151	6329116	0.6858
DBVPG6044	9142580	8933400	0.9771	6451661	0.7057	5619086	0.6146
DBVPG1788	7862564	7540263	0.959	6314620	0.8031	4961662	0.631
DBVPG1373	12440674	12032561	0.9672	9652667	0.7759	8117372	0.6525
DBVPG1853	8465550	8238381	0.9732	6407810	0.7569	5062458	0.598
BC187	7534756	7306321	0.9697	6159734	0.8175	4896614	0.6499
YPS606	7087944	6982170	0.9851	5819305	0.821	4805494	0.678
L-1374	6696198	6487796	0.9689	5256232	0.785	4165192	0.622
L-1528	8305190	8052646	0.9696	6584023	0.7928	5388218	0.6488
Y12	7145816	7061202	0.9882	5387987	0.754	4583406	0.6414
DBVPG1106	7302038	7080852	0.9697	5739361	0.786	4688844	0.6421
UWOPS83-787.3	8208416	8054253	0.9812	6806301	0.8292	5350658	0.6519
UWOPS87-2421	7549142	7381746	0.9778	6046842	0.801	4739368	0.6278
NCYC361	6940174	6643354	0.9572	5514631	0.7946	4302854	0.62
K11	8049548	7959248	0.9888	4066145	0.5051	3330790	0.4138
YS4	6749264	6535464	0.9683	5151995	0.7633	4261500	0.6314
YS9	7927482	7339364	0.9258	5896883	0.7439	5018192	0.633
322134S	6979932	6809146	0.9755	5309081	0.7606	4641194	0.6649
378604X	6384816	6215410	0.9735	4885249	0.7651	3871218	0.6063
273614N	7542110	7200338	0.9547	5847268	0.7753	4592882	0.609
YJM978	5907012	5732364	0.9704	4327857	0.7327	3599958	0.6094
Y9	8091960	7942982	0.9816	6118855	0.7562	4726592	0.5841
UWOPS03-461.4	8393732	8202423	0.9772	6555225	0.781	5130896	0.6113
UWOPS05-217.3	6523428	6391367	0.9798	5143629	0.7885	3448808	0.5287
S288C	10692668	10378610	0.9706	8102188	0.7577	6114742	0.5719
W303	6741958	6601574	0.9792	4830260	0.7164	3790168	0.5622
UWOPS05-227.2	7680810	7527610	0.9801	6420810	0.836	4552828	0.5928
DBVPG6040	7712090	7565503	0.981	5668962	0.7351	3839530	0.4979
YIIc17_E5	7207044	6907681	0.9585	5412899	0.7511	4358202	0.6047
YJM981	6938228	6686394	0.9637	4770593	0.6876	3422484	0.4933
YJM975	6971528	6799562	0.9753	5695281	0.8169	4537874	0.6509
NCYC110	6015720	5874647	0.9765	4448198	0.7394	3855086	0.6408
YS2	7453946	7216083	0.9681	5769810	0.7741	4373548	0.5867
CLIB192	6451646	6214063	0.9632	4993320	0.774	4004562	0.6207
CLIB208	8246170	7999657	0.9701	6481016	0.7859	5266968	0.6387
CLIB318	7912266	7745178	0.9789	6277986	0.7934	5328104	0.6734
CLIB324	7222384	7045908	0.9756	5862341	0.8117	4753410	0.6581
CLIB272	7259564	6992377	0.9632	5691203	0.784	4140966	0.5704
YJM145	7460498	7220732	0.9679	6054929	0.8116	4956570	0.6644
YJM280	6504562	6352620	0.9766	5090330	0.7826	4185252	0.6434
YJM320	8236790	7964151	0.9669	6169155	0.749	4968400	0.6032

STRAIN	# RAW READS	# MAPPABLE	% MAPPABLE	# UNIQUE	% UNIQUE	# CONCORDANT UNIQUE	% CONCORDANT UNIQUE
YJM326	8778880	8550041	0.9739	6981690	0.7953	5829194	0.664
YJM413	11024888	10711297	0.9716	9198454	0.8343	7119194	0.6457
YJM421	7883796	7686713	0.975	6367818	0.8077	4691324	0.5951
YJM428	7481486	7252880	0.9694	6091729	0.8142	4915452	0.657
YJM434	6867872	6358548	0.9258	5293065	0.7707	3985338	0.5803
YJM436	7426230	7185908	0.9676	5836370	0.7859	4610254	0.6208
YJM454	6249534	6051539	0.9683	4905091	0.7849	3866652	0.6187
YJM653	7187228	7015354	0.9761	5594135	0.7783	3770312	0.5246
CECT10109	8336134	8028499	0.9631	6728627	0.8072	5063486	0.6074
DBVPG1794	7825764	7649247	0.9774	6347909	0.8112	4457058	0.5695
DBVPG3591	7461794	7241459	0.9705	6059518	0.8121	4754638	0.6372
DBVPG4651	7689052	7363336	0.9576	6276467	0.8163	4350490	0.5658
EM93	7314248	7130340	0.9749	5624843	0.769	4511016	0.6167
TL229	7267760	7149656	0.9837	6139762	0.8448	3573086	0.4916
Y6	7435264	7187873	0.9667	5896607	0.7931	4192122	0.5638
YPS1000	6995148	6868688	0.9819	5776753	0.8258	4246830	0.6071
YPS163	8176392	8060384	0.9858	6769622	0.8279	4785098	0.5852
CLIB294	6420702	6177769	0.9622	5102395	0.7947	3859650	0.6011
CLIB413	6947436	6834561	0.9838	5449677	0.7844	3975356	0.5722
CLIB483	5462750	5230945	0.9576	4255055	0.7789	3165640	0.5795
K12	6748080	6640559	0.9841	4984688	0.7387	3952062	0.5857
K1	6223622	6125505	0.9842	4892367	0.7861	4249202	0.6828
Y3	8418316	8229424	0.9776	6727795	0.7992	5634360	0.6693
YJM269	9295036	8989681	0.9671	6751156	0.7263	6321864	0.6801
BY4716	6461220	6306368	0.976	5065535	0.784	4188002	0.6482
A364A	6792258	6264629	0.9223	5121458	0.754	4416780	0.6503
CENPK	6196422	6010243	0.97	4361002	0.7038	3595788	0.5803
FL100	5171790	5006896	0.9681	4090791	0.791	3295094	0.6371
CLIB154	6321356	6113115	0.9671	5162789	0.8167	3760344	0.5949
CLIB157	6233988	6047975	0.9702	5054564	0.8108	3957838	0.6349
CLIB219	4880852	4751897	0.9736	4024776	0.8246	2209384	0.4527
DBVPG1399	7799648	7512318	0.9632	6314710	0.8096	4256724	0.5458
I14	6512032	6309461	0.9689	5363899	0.8237	3544490	0.5443
M22	7678400	7459555	0.9715	6073943	0.791	4940570	0.6434
RM11	7782508	7667302	0.9852	6352186	0.8162	4208808	0.5408
T73	7060734	6818077	0.9656	5586992	0.7913	4368218	0.6187
UC1	6768116	6469654	0.9559	5525056	0.8163	3501676	0.5174
UC8	8240264	8009180	0.972	6833607	0.8293	5342598	0.6484
WE372	8661582	8414823	0.9715	7230800	0.8348	5723326	0.6608
Y8	7791856	7582299	0.9731	6384654	0.8194	4257828	0.5464
Y9J	6091740	5869791	0.9636	4909005	0.8058	4105464	0.6739

Table A-3 Reference genome mapping statistics

Raw number of reads obtained for each strain and the total number and percentage of reads remaining after various filtering steps. Reads were mapped to the S288c reference genome (R64-1-1). Unique reads required unique mapping positions and concordant reads required correct orientation of paired end reads.

Strain	Total Size	Longest	>10K	>1K	GC %	N50 Length	# N50
DBVPG6765	11899095	30783	95	2955	38.68%	4004	891
SK1	13016976	59712	331	1667	38.82%	8368	459
Y55	12403243	16357	20	3575	39.35%	2431	1500
YPS128	12139858	19773	98	2926	38.85%	3953	920
DBVPG6044	12425342	13880	20	3497	39.39%	2540	1424
DBVPG1788	12090444	46788	323	1761	38.60%	8345	436
DBVPG1373	12671376	29834	194	2378	38.99%	5187	711
DBVPG1853	12167788	17920	51	3249	39.12%	3246	1138
BC187	12035976	82793	391	1144	38.48%	14656	245
YPS606	12007677	24215	138	2698	38.68%	4662	774
L-1374	11784124	16971	16	3571	39.05%	2570	1391
L-1528	11897717	14810	20	3588	39.16%	2549	1395
Y12	11884110	13596	7	3667	39.32%	2104	1662
DBVPG1106	11997255	19916	89	2958	38.87%	3944	934
UWOPS83-787.3	11955284	27325	150	2598	38.66%	4947	733
UWOPS87-2421	12060037	30535	155	2568	38.67%	5003	729
NCYC361	12095277	33723	275	2012	38.62%	7168	516
K11	12381228	10514	1	3526	39.76%	1054	3282
YS4	13419563	11900	5	3205	39.27%	930	3526
YS9	14040022	14262	8	3259	39.09%	1024	3183
322134S	12811229	14814	9	3439	39.20%	1494	2220
378604X	11949961	17629	27	3596	39.04%	2575	1394
273614N	11895008	20263	14	3750	39.20%	2243	1575
YJM978	11742018	12352	4	3762	39.31%	2174	1647
Y9	12442732	79261	393	970	38.63%	17814	206
UWOPS03-461.4	12086158	24069	96	2901	38.86%	4101	898
UWOPS05-217.3	11727913	12276	6	3835	39.08%	2278	1552
S288C	12765331	42741	307	1875	38.91%	7294	512
W303	12524292	54700	397	1322	38.72%	12029	310
UWOPS05-227.2	11879355	65208	395	1355	38.43%	12141	297
DBVPG6040	12525846	8237	0	3619	39.61%	1173	2970
YIIc17_E5	13581119	24556	90	2980	38.81%	2165	1410
YJM981	12638639	59799	408	1260	38.80%	12611	308
YJM975	11916267	33113	175	2472	38.69%	5300	682
NCYC110	12176061	40091	216	2320	38.74%	5704	641
YS2	13897378	31384	96	2924	38.72%	1783	1659
CLIB192	13835905	26334	67	3096	38.66%	1646	1877
CLIB208	14551807	17074	19	3241	38.86%	984	3302
CLIB318	14776732	11030	3	3061	39.07%	651	5156
CLIB324	14743090	19400	14	3289	38.63%	912	3638
CLIB272	12575255	101743	323	1588	38.40%	9782	332
YJM145	11961334	27993	100	2831	38.76%	4231	872
YJM280	12017919	21758	88	2974	38.81%	3870	937
YJM320	12293734	13849	40	3369	39.16%	2879	1269
YJM326	12128009	19484	65	3141	38.90%	3400	1054
YJM413	12191449	81608	382	872	38.42%	20733	183
YJM421	12126830	83107	400	984	38.45%	17782	207
YJM428	14938761	14115	5	3274	38.73%	755	4424
YJM434	12222934	46795	379	1524	38.51%	10707	346
YJM436	12601033	52386	325	1725	38.57%	8355	401
YJM454	12188043	81608	393	1073	38.48%	15934	229
YJM653	13986214	20588	23	3283	38.64%	1235	2591
CECT10109	12134240	60328	337	1675	38.61%	9074	397
DBVPG1794	11972513	25146	147	2592	38.66%	4959	727
DBVPG3591	12024502	57819	402	1398	38.53%	11570	323

Strain	Total Size	Longest	>10K	>1K	GC %	N50 Length	# N50
DBVPG4651	11945996	42965	374	1615	38.48%	10017	374
EM93	14142650	31666	143	2602	38.66%	1959	1365
TL229	12466994	11531	1	3425	39.03%	986	3504
Y6	13919002	35463	58	3076	38.78%	1447	2138
YPS1000	11995832	30104	210	2313	38.63%	5922	609
YPS163	11996426	22527	88	2990	38.81%	4011	923
CLIB294	11914326	22010	78	2968	38.80%	4037	924
CLIB413	12331122	24589	63	3138	38.85%	3245	1102
CLIB483	11881269	23689	107	2819	38.67%	4251	841
K12	12378717	47201	351	1550	38.73%	9679	371
K1	12215001	71688	357	1564	38.53%	10227	349
Y3	14535592	24222	73	2907	38.66%	1139	2513
YJM269	12631158	25602	112	2729	39.06%	4275	892
BY4716	12210304	51855	366	1510	38.56%	10525	347
A364A	12115349	23315	120	2737	38.74%	4437	822
CENPK	12246267	19465	35	3336	39.13%	3124	1183
FL100	14046864	13914	14	3370	38.74%	1095	3036
CLIB154	11864123	51623	379	1340	38.44%	12378	297
CLIB157	11860328	45786	332	1691	38.51%	8892	400
CLIB219	11243613	6901	0	3522	39.02%	1011	3461
DBVPG1399	12103773	26547	181	2602	38.68%	4832	724
I14	11819917	17981	59	3273	38.74%	3420	1053
M22	12089908	29468	133	2728	38.87%	4613	792
RM11	11985968	29729	158	2631	38.75%	4889	741
T73	12197558	49398	222	2264	38.65%	5772	598
UC1	11819237	32113	230	2280	38.55%	6240	584
UC8	12021674	44529	341	1696	38.56%	8969	407
WE372	12001800	36521	333	1750	38.56%	8720	424
Y8	11942056	34281	305	1947	38.57%	7659	470
Y9J	11946302	50902	370	1604	38.47%	9880	373

Table A-4 *de novo* genome statistics

Total size of the de novo genome builds for each strain. The longest contig for each strain is given, as are the number of contigs greater than 10 kb (>10K0 and 1 kb (>1K). #N50 gives number of contigs needed to hit N50 length.

Condition	SNP Rank	SNP position	p-value	SNP Type	Gene	Systematic	Start site	Stop Site
40°C	1	chr11_490822	1.08E-06	Intergenic		YKR023W	483419	485011
					DBP7	YKR024C	487372	485144
					RPC37	YKR025W	487770	488618
					GCN3 <	YKR026C	489655	488738
					BCH2 <	YKR027W	491364	493661
					SAP190	YKR028W	494257	49735
					SET3	YKR029C	499833	497578
40°C	2 7 8 10 11	chr11_542388 chr11_544599 chr11_546863 chr11_550377 chr11_558621	1.15E-06 5.67E-05 5.77E-05 5.98E-05 6.24E-05	Coding Coding Coding Coding Coding	MRS4	YKR052C	533464	532550
					YSR3	YKR053C	535281	534067
					DYN1 ^^^	YKR054C	547925	535647
					RHO4	KR055W	548216	549091
					TRM2 ^	YKR056W	549448	551367
					RPS21A	YKR057W	551657	552242
					GLG1	YKR058W	552770	554620
					TIF18	YKR059W	554987	556174
					UTP30	YKR060W	556518	557342
					KTR2 ^	YKR061W	557677	558954
					TFA2	YKR062W	559666	560652
40°C	3	chr13_305627	1.83E-06	Intergenic	SEC59	YMR013C	296738	295179
					BUD22	YMR014W	298868	300427
					ERG5	YMR015C	302485	300869
					SOK2 <	YMR016C	305593	303236
					SPO20 <	YMR017W	307489	308682
						YMR018W	310208	311752
					STB4	YMR019W	312156	315005
					FMS1	YMR020W	315377	316903
40°C	4	chr02_170152	4.09E-06	Intergenic	HEK2	YBL032W	160184	161329
					SHE1	YBL031W	161699	162715
					PET9	YBL030C	163997	163041
						YBL029C-A	164772	164488
						YBL029W	166134	167264
						YBL028C	167838	167518
					RPL19B >	YBL027W	168423	169376
					LSM2 <	YBL026W	170623	171038
					RRN10	YBL025W	171481	171918
					NCL1	YBL024W	172534	174588
					MCM2	YBL023C	177526	174920
					PIM1	YBL022C	181275	177874
40°C	5	chr12_1004315	1.53E-05	Coding	SEN1	YLR430W	993434	1000129
					ATG23	YLR431C	1001703	1000342
					IMD3	YLR432W	1002557	1004128
					CNA1 ^	YLR433C	1006008	1004347
						YLR434C	1006408	1006025
					TSR2	YLR435W	1006378	100699
					ECM30	YLR436C	1011245	1007421
					DIF1	YLR437C	1012023	1011622
					CAR2	YLR438W	1012501	1013775
					LSM3	YLR438C-A	1014178	1013909
40°C	6	chr07_16686	3.31E-05	Intergenic		YGL260W	6860	7090
					YPS5	YGL259W	8470	8967
						YGL258W-A	9162	9395
					VEL1	YGL258W	11110	11730
					MNT2	YGL257C	14157	12481
					ADH4 >	YGL256W	15159	16307
					ZRT1 <	YGL255W	20978	22108
					FZF1	YGL254W	22304	23203
					HXK2	YGL253W	23935	25395
					RTG2	YGL252C	27484	25718
40°C	9	chr05_197502	5.94E-05	Coding	BIM1	YER016W	188277	189311
					AFG3	YER017C	191788	189503
					SPC25	YER018C	192624	191959
					ISC1	YER019W	192797	194230
					SBH2	YER019C-A	194539	194273
					GPA2	YER020W	195168	196517
					RPN3 ^	YER021W	196948	198519
					SRB4	YER022W	198812	200875
					PRO3	YER023W	201076	201936
					YAT2	YER024W	202192	204963
					GCD11	YER025W	205251	206834
40°C	12	chr12_627274	6.36E-05	Coding	VPS34	YLR240W	617533	620160
						YLR241W	620473	622821
					ARV1	YLR242C	623883	622918
					GPN3	LR243W	624203	625021
					MAP1	YLR244C	626331	625168
					CDD1	YLR245C	626930	626502
					ERF2 ^	YLR246W	627118	628197
					IRC20	YLR247C	633354	628684

					RCK2	*YLR248W*	634252	636084
					YEF3	*YLR249W*	636780	639914
40°C	13	chr04_375335	8.8E-05	Intergenic	*KNH1*	*YDL049C*	365874	365068
					STP4	*YDL048C*	368211	366739
					SIT4	*YDL047W*	369771	370706
					NPC2	*YDL046W*	371240	371761
					MRP10	*YDL045W-A*	372248	372535
					FAD1	*YDL045C*	373608	372688
					MTF2 <	*YDL044C*	375289	373967
					PRP11 >	*YDL043C*	376480	375680
					SIR2	*YDL042C*	378445	376757
					NAT1	*YDL040C*	381438	378874
					PRM7	*YDL039C*	384081	381985
					BSC1	*YDL037C*	385587	38460
Condition	**SNP Rank**	**SNP position**	**p-value**	**SNP Type**	**Gene**	**Systematic**	**Start site**	**Stop Site**
3mM H2O2	1	chr14_676834	5.35E-06	Coding	*MPP6*	*YNR024W*	672409	672969
						YNR025C	673061	672702
					SEC12	*YNR026C*	674689	673274
					BUD17	*YNR027W*	674923	675876
					CPR8 ^	*YNR028W*	676177	677103
						YNR029C	678488	677199
					ALG12	*YNR030W*	678799	680454
					SSK2	*YNR031C*	685433	680694
3mM H2O2	2	chr04_101326	7.33E-06	Coding	*ACK1*	*YDL203C*	97953	96082
					MRPL11	*YDL202W*	98475	99224
					TRM8	*YDL201W*	99561	100421
					MGT1	*YDL200C*	101067	100501
					^	*YDL199C*	103353	101290
					GGC1	*YDL198C*	104551	103649
					ASF2	*YDL197C*	106494	104917
						YDL196W	106741	107070
3mM H2O2	3	chr10_303316	1.53E-05	Coding	*ICS3*	*YJL077C*	295061	294666
					NET1	*YJL076W*	295245	298814
					APQ13	*YJL075C*	298876	298460
					SMC3	*YJL074C*	302849	299157
					JEM1 ^	*YJL073W*	303181	305118
					PSF2	*YJL072C*	305862	305221
					ARG2	*YJL071W*	306132	307856
						YJL070C	310637	307971
3mM H2O2	4	chr07_870646	3.46E-05	Coding	*UBR1*	*YGR184C*	865753	859901
					TYS1	*YGR185C*	867520	866336
					TFG1	*YGR186W*	867774	869981
					HGH1 ^	*YGR187C*	871416	870232
					BUB1	*YGR188C*	875109	872044
					CRH1	*YGR189C*	878192	876669
						YGR190C	880661	880296
					HIP1	*YGR191W*	880420	882231
3mM H2O2	5	chr08_123678	3.48E-05	Intergenic	*STP2*	*YHR006W*	117814	119439
					ERG11	*YHR007C*	121683	120091
						YHR007C-A	122765	122550
					SOD2 <	*YHR008C*	123590	122889
					TDA3 >	*YHR009C*	125680	124109
					RPL27A	*YHR010W*	126521	127492
					DIA4	*YHR011W*	127780	129120
					VPS29	*YHR012W*	129481	130448
3mM H2O2	6	chr04_275205	3.49E-05	Coding	*PHO2*	*YDL106C*	271901	270222
					NSE4	*YDL105W*	272389	273597
					QRI7	*YDL104C*	274876	273653
					QRI1 ^	*YDL103C*	276581	275148
					POL3	*YDL102W*	276872	280165
					DUN1	*YDL101C*	281848	280307
					GET3	*YDL100C*	283176	282112
					BUG1	*YDL099W*	283419	284444
3mM H2O2	7	chr14_748773	3.81E-05	Intergenic	*MNT4*	*YNR059W*	738545	736803
					FRE4	*YNR060W*	739951	742110
						YNR062C	745343	744360
					>	*YNR063W*	746943	748766
					>	*YNR064C*	750008	749136
						YNR065C	753700	750350
						YNR066C	755035	753725
					DSE4	*YNR067C*	759099	755746
3mM H2O2	8	chr06_254049	4.22E-05	Coding	*RET2*	*YFR051C*	251803	250163
					RPN12	*YFR052W*	252505	253329
						YFR052C-A	253734	253429
					HXK1 ^	*YFR053C*	255049	253592
						YFR054C	259433	258855
						YFR056C	264325	263957
					IRC7	*YFR055W*	264204	265226

3mM H2O2	9	chr15_223663	4.65E-05	Coding	*GPD2*	*YOL059W*	217126	218448
					ARG1	*YOL058W*	219211	220473
						YOL057W	220767	222902
					GPM3 ^	*YOL056W*	223268	224179
					THI20	*YOL055C*	226075	224420
					PSH1	*YOL054W*	228614	229834
					AIM39	*YOL053W*	230085	231272
3mM H2O2	10	chr10_134512	4.78E-05	Coding	*SSY5*	*YJL156C*	128952	126853
					FBP26	*YJL155C*	130643	129285
					VPS35	*YJL154C*	133935	131101
					INO1 ^	*YJL153C*	135933	134332
						YJL152W	136174	136533
					SNA3	*YJL151C*	136773	136372
						YJL150W	137123	137425
					DAS1	*YJL149W*	137379	139370
3mM H2O2	11	chr10_405572	6.60E-05	Intergenic	*APS3*	*YJL024C*	396592	395931
					PET130	*YJL023C*	398398	397355
					BBC1	*YJL020C*	402410	398937
					MPS3 >	*YJL019W*	402897	404945
					<	*YJL016W*	405588	407273
					CCT3	*YJL014W*	407558	409162
					MAD3	*YJL013C*	411040	409493
					VTC4	*YJL012C*	413399	411234
					RPC17	*YJL011C*	414770	414285
3mM H2O2	12	chr13_715643	6.66E-05	Coding	*ESC1*	*YMR219W*	707133	712109
					ERG8	*YMR220W*	712316	713671
						YMR221C	715445	713931
					FSH2 ^	*YMR222C*	716309	715638
					UBP8	*YMR223W*	716715	718130
					MRE11	*YMR224C*	720653	718575
					MRPL44	*YMR225C*	721403	720960
						YMR226C	722396	721593
					TAF7	*YMR227C*	724385	722613
3mM H2O2	13	chr07_177632	7.71E-05	Coding	*MPT5*	*YGL178W*	167352	170571
						YGL176C	173079	171415
					SAE2	*YGL175C*	174322	173285
					BUD13	*YGL174W*	174545	175345
					XRN1 ^	*YGL173C*	180113	175527
					NUP49	*YGL172W*	180700	182118
					ROK1	*YGL171W*	182390	184084
					SPO74	*YGL170C*	185394	184153
					SUA5	*YGL169W*	186059	187339
Condition	**SNP Rank**	**SNP position**	**p-value**	**SNP Type**	**Gene**	**Systematic**	**Start site**	**Stop Site**
1M NaCl	1	chr15_69652	0.000221	Coding	*ARG8*	*YOL140W*	58759	60030
	2	chr15_69778	0.000221	Coding	*CDC33*	*YOL139C*	61024	60383
					RTC1	*YOL138C*	65350	61325
					BSC6	*YOL137W*	65621	67114
					PFK27	*YOL136C*	68754	67561
					MED7 ^^	*YOL135C*	70044	69376
						YOL134C	70545	70156
					HRT1	*YOL133W*	70325	70690
					GAS4	*YOL132W*	71300	72715
						YOL131W	73031	73357
1M NaCl	3	chr08_29694	0.000992	Coding	*EFM1*	*YHL039W*	21783	23540
					CBP2	*YHL038C*	25509	23617
						YHL037C	26179	25778
					MUP3	*YHL036W*	26241	27881
					VMR1 ^	*YHL035C*	32756	27978
					SBP1	*YHL034C*	34077	33193
					RPL8A	*YHL033C*	36025	35255
					GUT1	*YHL032C*	38508	36379
					GOS1	*YHL031C*	39486	38815
						YHL030W-A	39074	39535
Condition	**SNP Rank**	**SNP position**	**p-value**	**SNP Type**	***Gene***	***Systematic***	**Start site**	**Stop Site**
1mM CoCL2	1	chr07_136453	2.38E-06	Coding	*EMP24*	*YGL200C*	123305	122694
					YIP4	*YGL198W*	123591	124298
					MDS3	*YGL197W*	124698	129161
					DSD1	*YGL196W*	129883	131169
					GCN1 ^	*YGL195W*	131525	139543
					HOS2	*YGL194C*	141726	140368
					IME4	*YGL192W*	142246	144048
					COX13	*YGL191W*	144808	145197
					CDC55	*YGL190C*	147389	145809
1mM CoCL2	2	chr14_591136	2.82E-06	Coding	*SSN8*	*YNL025C*	585291	584320
					KSH1	*YNL024C-A*	586820	586602
					EFM6	*YNL024C*	587847	587107
					FAP1 ^	*YNL023C*	591160	588263
					RCM1	*YNL022C*	592899	591427

					HDA1	*YNL021W*	593227	595347
					ARK1	*YNL020C*	597539	595623
						YNL019C	599230	598376
						YNL018C	601774	599936
1mM CoCL2	3	chr07_209438	1.76E-05	Coding	*YIP5*	*YGL161C*	200142	199210
					AIM14	*YGL160W*	200561	202273
						YGL159W	202721	203833
					RCK1	*YGL158W*	207033	208571
					ARI1 ^	*YGL157W*	209006	210049
					AMS1	*YGL156W*	210416	213667
					CDC43	*YGL155W*	214081	215211
					LYS5	*YGL154C*	216096	215278
					PEX14	*YGL153W*	216273	217298
1mM CoCL2	4	chr05_48112	1.81E-05	Coding	*SOM1*	*YEL059C-A*	42624	42400
					HHY1	*YEL059W*	42652	42960
					PCM1	*YEL058W*	43252	44925
					SDD1	*YEL057C*	45020	45721
					HAT2 ^	*YEL056W*	47168	48373
					POL5	*YEL055C*	51539	48471
					RPL12A	*YEL054C*	53218	52721
					MAK10	*YEL053C*	56102	53901
					AFG1	*YEL052W*	56571	58100
1mM CoCL2	5	chr13_12620	2.35E-05	Coding		*YML133C*	4684	461
					COS3	*YML132W*	7244	8383
						YML131W	11295	10198
					ERO1 ^	*YML130C*	13174	11483
					COX14	*YML129C*	14753	14541
					MSC1	*YML128C*	16676	15135
					RSC9	*YML127W*	17064	18809
					ERG13	*YML126C*	20535	19060
					PGA3	*YML125C*	21699	20761
1mM CoCL2	6	chr10_472011	3.59E-05	Intergenic	*REC107*	*YJR021C*	469579	468555
					LSM8	*YJR022W*	469784	470113
						YJR023C	470205	469804
					MDE1	*YJR024C*	470964	470230
					BNA1 <	*YJR025C*	471671	471138
					YJRWTy1-1		472463	478384
1mM CoCL2	7	chr16_509846	4.26E-05	Coding		*YPL025C*	503030	502473
					RMI1	*YPL024W*	503517	504242
					MET12	*YPL023C*	506312	504339
					RAD1 ^	*YPL022W*	506697	509999
					ECM23	*YPL021W*	511101	511664
					ULP1	*YPL020C*	514178	512313
					VTC3	*YPL019C*	517018	514511
					CTF19	*YPL018W*	517651	518760
					IRC15	*YPL017C*	520233	518734
1mM CoCL2	8	chr13_260820	4.91E-05	Coding	*ERG6*	*YML008C*	252990	251839
						YML007C-A	253272	253162
					YAP1	*YML007W*	253848	255800
					GIS4 ^	*YML006C*	258416	256092
					TRM12	*YML005W*	260221	261609
					GLO1	*YML004C*	262685	261705
						YML003W	263483	264355
						YML002W	264541	266754
					YPT7	*YML001W*	267174	267800
1mM CoCL2	9	chr10_401648	5.05E-05	Coding	*RNR2*	*YJL026W*	392404	393603
					RRN7	*YJL025W*	393967	395511
					APS3	*YJL024C*	396592	395931
					PET130	*YJL023C*	398398	397355
					BBC1 ^	*YJL020C*	402410	398937
					MPS3	*YJL019W*	402897	404945
						YJL016W	405588	407273
					CCT3	*YJL014W*	407558	409162
					MAD3	*YJL013C*	411040	409493
1mM CoCL2	10	chr05_342178	5.06E-05	Intergenic	*DOT6*	*YER088C*	335188	333176
					PTC2<	*YER089C*	337340	335946
					TRP2 <	*YER090W*	337949	339472
					MET6	*YER091C*	342167	339864
					IES5	*YER092W*	342855	343232
					TSC11	*YER093C*	343320	347612
					AIM11	*YER093C-A*	348400	347912
					PUP3	*YER094C*	349346	348729
					RAD51	*YER095W*	349980	351182
1mM CoCL2	11	chr07_370360	5.7E-05	Coding	*DBP3*	*YGL078C*	361859	360288
					HNM1	*YGL077C*	363916	362225
					RPL7A ^	*YGL076C*	364335	365996
					MPS2	*YGL075C*	368088	366925
					HSF1	*YGL073W*	368753	371254

					AFT1	*YGL071W*	372012	374084
					RPB9	*YGL070C*	374827	374459
					MNP1	*YGL068W*	375087	375671
					NPY1	*YGL067W*	376101	377255
					RPC37	*YKR025W*	487770	488618
					GCN3	*YKR026C*	489655	488738
					BCH2	*YKR027W*	491364	493661
					SAP190 ^	*YKR028W*	494257	497358
1mM CoCL2	12	chr11_496826	6.7E-05	Coding	*SET3*	*YKR029C*	499833	497578
					GMH1	*YKR030W*	500282	501103
					SPO14	*YKR031C*	506395	501344
					DAL80	*YKR034W*	506898	507707
					DID2	*YKR035W-A*	507939	508553
					GMC2	*YLR445W*	1024189	1024837
						YLR446W	1025214	1026515
					VMA6	*YLR447C*	1027893	1026856
					RPL6B ^	*YLR448W*	1028854	1029768
1mM CoCL2	13	chr12_1031337	6.72E-05	Coding	*FPR4*	*YLR449W*	1030834	1032012
					HMG2	*YLR450W*	1032627	1035764
					LEU3	*YLR451W*	1036093	1038753
					SST2	*YLR452C*	1041366	1039270
					RIF2	*YLR453C*	1042986	1041799
					FMP27	*YLR454W*	1043998	1051884
						YNL046W	542304	542822
					LAP2	*YNL045W*	542963	544978
					YIP3 >	*YNL044W*	545268	545877
					<	*YNL042W-B*	547113	547370
1mM CoCL2	14	chr14_546186	7.27E-05	Intergenic	*BOP3*	*YNL042W*	548100	549290
					COG6	*YNL041C*	551987	549468
						YNL040W	553380	554750
					BDP1	*YNL039W*	555048	556832
					GPI15	*YNL038W*	557020	557783
					GAP1	*YKR039W*	515063	516871
						YKR041W	518198	518950
					UTH1	*YKR042W*	519527	520624
					SHB17	*YKR043C*	521712	520897
1mM CoCL2	15	chr11_524511	9.19E-05	Coding	*UIP5*	*YKR044W*	522015	523346
						YKR045C	523969	523418
					PET10 ^	*YKR046C*	525074	524223
						YKR047W	525257	525562
					NAP1	*YKR048C*	526640	525387
					FMP46	*YKR049C*	527231	526830
					SPO14	*YKR031C*	506395	501344
						YKR032W	506517	506831
						YKR033C	507304	506879
					DAL80 ^	*YKR034W*	506898	507707
1mM CoCL2	16	chr11_509718	9.24E-05	Coding	*DID2*	*YKR035W-A*	507939	508553
	18	chr11_507344	9.68E-05	Coding	*OPI8*	*YKR035C*	508561	507920
					CAF4 ^	*YKR036C*	510633	508702
					SPC34	*YKR037C*	511797	510910
					KAE1	*YKR038C*	513159	511999
					GAP1	*YKR039W*	515063	516871
					KEX2	*YNL238W*	202428	204872
					YTP1	*YNL237W*	205188	206567
					SIN4	*YNL236W*	206930	209854
						YNL234W	210233	211513
1mM CoCL2	17	chr14_213493	9.65E-05	Coding	*BNI4* ^	*YNL233W*	211922	214600
					CSL4	*YNL232W*	214923	215801
					PDR16	*YNL231C*	217042	215987
					ELA1	*YNL230C*	218662	217523
					URE2	*YNL229C*	220201	219137
					JJJ1	*YNL227C*	222431	220659
					SMC3	*YJL074C*	302849	299157
					JEM1	*YJL073W*	303181	305118
					PSF2	*YJL072C*	305862	305221
					ARG2	*YJL071W*	306132	307856
1mM CoCL2	19	chr10_309358	9.93E-05	Coding	^	*YJL070C*	310637	307971
					UTP18	*YJL069C*	312706	310922
						YJL068C	313915	313016
					MPM1	*YJL066C*	314872	314114
					DLS1	*YJL065C*	315557	315054
Condition	**SNP Rank**	**SNP position**	**p-value**	**SNP Type**	***Gene***	***Systematic***	**Start site**	**Stop Site**
					CRD1	*YDL142C*	202570	201719
					BPL1	*YDL141W*	203039	205111
					RPO21	*YDL140C*	210561	205360
4mM Paraquat	1	chr04_212222	2.64E-05	Intergenic	*SCM3* <	*YDL139C*	212046	211375
					RGT2 <	*YDL138W*	213351	215642
					ARF2	*YDL137W*	216529	217074

					RPL35B	*YDL136W*	217600	218367
					RDI1	*YDL135C*	219288	218680
4mM Paraquat	2	chr02_170152	3.53E-05	Intergenic		*YBL029W*	166134	167264
						YBL028C	167838	167518
					RPL19B >	*YBL027W*	168423	169376
					LSM2 <	*YBL026W*	170623	171038
					RRN10	*YBL025W*	171481	171918
					NCL1	*YBL024W*	172534	174588
					MCM2	*YBL023C*	177526	174920
					PIM1	*YBL022C*	181275	177874
4mM Paraquat	3	chr13_305627	6.38E-05	Intergenic	*SEC59*	*YMR013C*	296738	296738
					BUD22	*YMR014W*	298868	300427
					ERG5	*YMR015C*	302485	300869
					SOK2 <	*YMR016C*	305593	303236
					SPO20 <	*YMR017W*	307489	308682
						YMR018W	310208	311752
					STB4	*YMR019W*	312156	315005
					FMS1	*YMR020W*	315377	316903

Table A-5 Significant GWAS SNPs and Region Features

Table includes the SNPs significantly associated with growth rate in each test condition. Genes within the region are listed. Gene names marked with ^, < and > identify whether the significant SNP is located within the coding region of that gene, directly upstream or directly downstream of the gene respectively. Dark green denotes genes that list a deletion phenotype in the yeast genome database (CHERRY *et al.* 2012) that directly relates to the specific test condition (i.e. 'Heat tolerance: Decreased' for high temperature SNPs). Light green denotes a deletion phenotype curation that lists a 'stress response: decreased' or 'stress response: increased'.

Primer	Sequence
HO-A	TATTAGGTGTGAAACCACGAAAAGT
HO-D	CATGTCTTCTCGTTAAGACTGCAT
kB	CTGCAGCCGAGGAGCCGTAAT
kC3	CCTCGACATCATCTGCCCAGAT
HO-B	ACTGTCATTGGGAATGTCTTATGAT
HO-C	GAGTGGTAAAAATCGAGTATGTGCT

Table A-6 Primers used for confirmation of HO deletion and KanMX4 insertion

G418 resistant colonies were checked by PCR to confirm deletion of HO and correct insertion of the KanMX4 marker. PCRs using primer pairs HO-A/HO-B, HO-A/kB, HO-D/HO-C, and HO-D/kC3 were required to produce the correct band sizes to confirm each diploid strain as a heterozygote deletion.

Hit ID	E-Value	Query Coverage Percentage	Percent Identity	Strains	New Gene Group
emb|FN393072.1|	9.00E-106	0.99	0.971	DBVPG1853	2
emb|FN393081.1|	9.00E-163	1	0.994	CLIB157	3
emb|FN393081.1|	1.00E-155	1	0.997	DBVPG6765, L-1528, NCYC361, DBVPG6040	3
emb|FN393081.1|	8.00E-164	1	0.997	DBVPG1788, DBVPG1794, DBVPG4651, Y8	3
emb|FN393081.1|	2.00E-157	1	1	DBVPG1373, L-1374, UWOPS87-2421, 378604X, YJM978, YJM981, YJM975, YS2, YJM320, YJM421, TL229, UC1	3
emb|FN393081.1|	1.00E-154	1	0.994	BC187, T73	3
emb|FN393081.1|	5.00E-153	1	0.99	273614N	3
emb|FN393081.1|	1.00E-125	0.68	0.932	EM93	3
emb|FN393081.1|	1.00E-147	1	0.978	UWOPS83-787.3, I14	3
emb|FN393081.1|	3.00E-120	0.68	0.919	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	3
emb|FN393081.1|	4.00E-142	1	0.965	SK1, DBVPG6044, NCYC110, CENPK	3
emb|FN393081.1|	1.00E-147	1	0.978	CLIB219	3
emb|FN393081.1|	5.00E-141	1	0.962	UWOPS87-2421, YPS1000	3
emb|FN393081.1|	7.00E-145	1	0.971	UWOPS83-787.3, YIIc17_E5	3
emb|FN393081.1|	2.00E-159	1	0.932	SK1, NCYC110	3
emb|FN393074.1|	0	0.98	0.958	YJM413, YJM454	4
emb|FN393074.1|	0	0.98	0.962	DBVPG6044, NCYC110	4
emb|FN393074.1|	0	0.98	0.959	Y12, Y9	4
emb|FN393074.1|	0	0.98	0.958	UWOPS83-787.3	4
emb|FN393074.1|	0	0.98	0.958	YPS128, YPS606, YPS163	4
emb|FN393074.1|	0	1	0.996	L-1528, CLIB157	4
emb|FN393074.1|	0	1	0.979	K11, K1	4
emb|FN393074.1|	0	1	0.979	273614N	4
emb|FN393074.1|	0	1	1	YJM975	4
emb|FN393074.1|	0	1	0.998	YS2	4
emb|FN393074.1|	0	1	0.997	Y3	4
emb|FN393074.1|	0	1	0.999	SK1, DBVPG1853, YS4, YS9, 322134S, 378604X, YJM981, YJM145, YJM320, YJM326, DBVPG3591, CLIB483	4
emb|FN393074.1|	0	1	0.998	CLIB318, TL229, Y9J	4
emb|FN393074.1|	0	1	0.998	UWOPS87-2421, CLIB272, YJM653, FL100	4
emb|FN393074.1|	0	1	0.994	CLIB324	4
emb|FN393074.1|	0	1	0.98	K12	4
emb|FN393074.1|	0	1	0.981	Y6	4
emb|FN393074.1|	0	1	0.997	UC1	4
emb|FN393074.1|	0	1	0.997	CLIB208	4
emb|FN393074.1|	0	1	0.998	Y55, DBVPG1788, YIIc17_E5, YJM434, DBVPG1794, CLIB294	4
emb|FN393074.1|	0	1	0.997	CECT10109	4
emb|FN393074.1|	0	1	0.996	T73	4
emb|FN393074.1|	0	1	0.998	BC187	4

emb|FN393074.1|	0	1	0.961	YJM421	4
emb|FN393081.1|	1.00E-149	1	0.975	SK1, DBVPG6044, NCYC110, CENPK	5
emb|FN393081.1|	1.00E-149	1	0.975	273614N, YPS1000	5
emb|FN393081.1|	2.00E-152	0.99	0.984	YIIc17_E5	5
emb|FN393081.1|	2.00E-153	1	0.984	CLIB219	5
emb|FN393081.1|	4.00E-154	0.99	0.987	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	5
emb|FN393081.1|	4.00E-155	1	0.987	I14	5
emb|FN393081.1|	4.00E-155	1	0.987	YJM269	5
emb|FN393081.1|	1.00E-160	1	1	DBVPG6765, DBVPG1788, DBVPG1373, DBVPG1853, BC187, L-1374, L-1528, NCYC361, 378604X, YJM978, DBVPG6040, YJM981, YJM975, YS2, YJM421, DBVPG1794, DBVPG4651, TL229, CLIB157, T73, UC1, Y8	5
emb|FN393081.1|	6.00E-159	1	0.997	YJM320	5
emb|FN393068.1|	0	0.99	0.798	Y9J	6
gb|U19263.1|SPU19263	1.00E-53	0.46	0.913	CLIB483	7
gb|EU864229.1|	1.00E-84	0.84	0.988	CLIB154	8
gb|EU864229.1|	0	1	0.988	DBVPG6765, NCYC361, Y9J	8
gb|EU864229.1|	1.00E-155	1	0.983	DBVPG1788, 378604X, YIIc17_E5, YJM320, CECT10109, DBVPG1794, DBVPG3591, DBVPG4651, CLIB294, UC1	8
gb|EU864229.1|	4.00E-84	0.83	0.988	YS9	8
gb|EU864229.1|	7.00E-152	1	0.975	K11, K12, K1	8
gb|EU864229.1|	6.00E-153	1	0.978	Y12, UWOPS83-787.3, UWOPS87-2421, Y9, YJM280, YJM326, YPS1000, CLIB413	8
gb|EU864229.1|	7.00E-152	1	0.975	YPS128, YPS606, YPS163	8
gb|EU864229.1|	7.00E-152	1	0.975	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	8
gb|EU864228.1|	1.00E-173	1	1	CLIB272	8
emb|FN677930.2|	0	1	0.954	SK1, Y55	9
emb|FN677930.2|	0	1	0.996	UWOPS05-217.3, UWOPS05-227.2	9
emb|FN677930.2|	0	1	0.995	K12, K1	9
emb|FN677930.2|	0	1	0.993	YJM326	9
emb|FN677930.2|	0	1	0.996	YJM145	9
emb|FN677930.2|	0	1	0.992	DBVPG6040, CLIB272, YJM436, YJM653, FL100	9
emb|FN677930.2|	0	1	0.995	YJM421	9
emb|FN677930.2|	0	1	0.996	Y12	9
emb|FN677930.2|	0	1	0.993	YJM413, YJM454	9
emb|FN677930.2|	0	1	0.996	378604X	9
emb|FR847062.1|	1.00E-142	1	0.892	Y3	10
emb|FR847062.1|	0	1	0.937	SK1, Y55, NCYC110	10
emb|FR847062.1|	0	1	0.999	YJM436, CLIB483	10
emb|FR847062.1|	0	1	1	DBVPG6040, CLIB272	10
emb|FR847061.1|	0	1	0.997	K1	10
emb|FR851879.1|	0	1	0.995	378604X, YJM421	10
emb|FR847061.1|	0	1	0.998	273614N, YJM326, YJM413, YJM454, K12	10
emb|FR847062.1|	0	1	0.994	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	10
emb|FR851879.1|	0	1	0.991	Y12	10

emb|FR851879.1|	0	1	0.992	YJM145	10
emb|FN393074.1|	2.00E-157	1	1	DBVPG6765, DBVPG1788, DBVPG1373, BC187, L-1528, NCYC361, YS9, 273614N, YJM978, YJM981, YJM975, CLIB192, CLIB318, YJM145, YJM320, YJM326, YJM413, YJM434, YJM454, CECT10109, DBVPG1794, DBVPG3591, DBVPG4651, CLIB154, DBVPG1399, M22, T73, UC1, UC8, WE372, Y8, Y9J	11
emb|FN393074.1|	6.00E-165	1	1	DBVPG1106	11
emb|FN393074.1|	1.00E-155	1	0.997	DBVPG1853, CLIB294, CLIB157, I14	11
emb|FN393074.1|	1.00E-155	1	0.997	RM11	11
emb|FN393074.1|	1.00E-154	1	0.994	TL229	11
emb|FN393074.1|	5.00E-153	1	0.99	L-1374	11
dbj|AB188515.1|	0	1	1	Y55, DBVPG6044, NCYC110, A364A	12
dbj|AB188515.1|	0	0.99	0.924	CLIB483	12
dbj|AB188681.1|	0	1	0.902	378604X	12
dbj|AB188681.1|	0	1	0.996	YJM326, Y6	12
dbj|AB188681.1|	0	1	0.998	UWOPS05-217.3, UWOPS05-227.2	12
dbj|AB188681.1|	0	1	0.995	YJM421	12
dbj|AB200246.1|	0	1	0.998	YJM269	12
dbj|AB188681.1|	0	1	0.999	CLIB413	12
dbj|AB188681.1|	0	1	1	Y12, K12, CENPK	12
ref|XM_003680075.1|	0	1	0.957	273614N	13
ref|XM_003680067.1|	0	0.93	0.74	DBVPG1373	14
emb|FN393067.1|	1.00E-148	1	0.987	CLIB272	15
emb|FN393067.1|	2.00E-151	1	0.993	Y12, K11, Y9	15
emb|FN393067.1|	1.00E-148	1	0.987	YJM145, YJM320, YJM326, YPS1000	15
emb|FN393067.1|	5.00E-147	1	0.984	YJM269	15
ref|XM_002556240.1|	6.00E-06	0.15	0.705	CLIB272	16
ref|XM_002556240.1|	6.00E-06	0.15	0.705	Y6	16
ref|XM_001526834.1|	7.00E-05	0.08	0.791	UWOPS87-2421	16
ref|XM_001526834.1|	2.00E-06	0.2	0.685	Y55, NCYC110	16
emb|FN394216.1|	4.00E-148	1	0.993	CLIB157	17
dbj|AB188681.1|	0	1	1	YJM269	18
dbj|AB188681.1|	0	1	0.886	CLIB483	18
emb|FN394216.1|	0	1	0.999	NCYC361	19
emb|FN394216.1|	0	1	0.999	YS2	19
emb|FN394216.1|	0	1	1	YIIc17_E5, CLIB192, CLIB208, Y6	19
emb|FN394216.1|	0	1	0.999	DBVPG6765	19
emb|FN394216.1|	0	1	1	YS9, DBVPG6040, CLIB294, UC1	19
ref|XM_003680078.1|	5.00E-180	1	0.966	273614N	20
dbj|AB302221.1|	0	0.97	0.995	UWOPS87-2421	21
dbj|AB302221.1|	0	0.97	0.999	DBVPG3591, RM11	21
dbj|AB302221.1|	0	0.97	1	UC8, WE372	21
dbj|AB302221.1|	0	0.97	0.995	Y3	21
dbj|AB302221.1|	0	1	0.993	Y3	21

dbj|AB302221.1|	0	0.98	0.995	SK1, DBVPG6044	21
dbj|AB302221.1|	0	0.98	0.996	YJM269	21
dbj|AB302221.1|	0	0.98	0.996	UWOPS05-217.3, UWOPS05-227.2	21
dbj|AB302221.1|	0	1	0.995	UWOPS83-787.3	21
dbj|AB302221.1|	0	0.98	0.995	CECT10109	21
dbj|AB302221.1|	0	0.98	0.996	YPS1000, I14	21
dbj|AB302221.1|	0	0.98	0.996	YPS606, YPS163	21
dbj|AB302221.1|	2.00E-139	1	0.997	UWOPS87-2421, YJM326, DBVPG3591, RM11, T73, UC8, WE372	21
emb|FN394216.1|	0	1	1	DBVPG6040	22
tpg|BK006936.2|	4.00E-175	1	0.997	YPS606, UWOPS87-2421, UWOPS03-461.4, W303, UWOPS05-227.2, YPS1000, YPS163	23
tpg|BK006936.2|	8.00E-177	1	0.992	CLIB413, K12, K1	23
gb|CP002850.1|	9.00E-10	0.15	0.767	Y6	24
gb|CP002850.1|	9.00E-10	0.15	0.767	DBVPG6040	24
emb|HE616743.1|	1.00E-130	1	0.888	273614N	26
emb|FN394216.1|	1.00E-122	0.99	0.909	L-1528, YJM978, YIIc17_E5, YJM280, YJM434, YJM653, DBVPG4651, CLIB294, CLIB483, FL100, CLIB154, DBVPG1399, I14, Y9J	27
emb|FN394216.1|	6.00E-127	0.99	0.918	SK1, YS4, CLIB318, CLIB272	27
emb|FN393086.1|	1.00E-142	1	0.994	YS4	28
emb|FN393086.1|	1.00E-143	1	0.997	Y55, DBVPG1788, DBVPG1373, BC187, NCYC361, 378604X, W303, YIIc17_E5, YJM145, YJM320, Y6, CLIB294, CLIB154, CLIB157, DBVPG1399, RM11, T73, UC1, UC8, WE372	28
emb|FN393086.1|	1.00E-142	1	0.994	YJM981, YJM975	28
emb|FN394217.1|	0	1	0.995	TL229	29
emb|FN394217.1|	0	1	0.996	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	29
emb|FN394217.1|	0	1	0.997	UWOPS83-787.3	29
emb|FN394217.1|	0	1	0.999	SK1, DBVPG6044, NCYC110, Y3	29
emb|FN394217.1|	0	1	0.996	YIIc17_E5, YJM326, Y6, Y9J	29
emb|FN394217.1|	0	1	0.992	CLIB219	29
emb|FN394217.1|	0	1	1	YJM978, YJM981, YJM975, CECT10109	29
emb|FN394217.1|	0	1	0.997	YPS128, YPS606, YPS163	29
emb|FN394217.1|	0	1	0.995	YJM145, YJM320, YPS1000	29
emb|FN394217.1|	0	1	0.996	DBVPG1853, YS4, DBVPG6040, YS2, CLIB192, CLIB318, CLIB324, YJM428	29
emb|FN394217.1|	0	1	0.997	CLIB272, YJM653, FL100	29
tpg|BK006941.2|	1.00E-160	1	0.982	YS2	30
tpg|BK006941.2|	3.00E-162	1	0.985	K11, CLIB413, CLIB483, K12, K1	30
tpg|BK006941.2|	2.00E-159	1	0.98	Y12, Y9	30
emb|FN394216.1|	5.00E-29	0.42	0.667	UWOPS83-787.3	31
emb|FN394216.1|	5.00E-29	0.41	0.667	UWOPS87-2421, YPS1000	31
emb|FN393063.1|	3.00E-176	1	0.995	YJM413, YJM454	32
emb|FN393063.1|	3.00E-176	1	0.995	TL229	32
emb|FN393063.1|	2.00E-177	1	0.997	DBVPG1788, DBVPG1853, L-1528, 273614N, YJM978, YJM981, YJM975, CLIB272, CECT10109, DBVPG1794, DBVPG3591, DBVPG1399, UC8, WE372, Y8, Y9J	32
dbj|AB195821.1|	0	1	0.913	UC8, WE372	33
dbj|AB195821.1|	0	1	0.913	DBVPG1399	33

dbj|AB195821.1|	0	0.97	0.994	Y3	33
dbj|AB195821.1|	0	0.97	0.999	DBVPG3591, RM11	33
dbj|AB195821.1|	0	0.97	0.999	YJM326, CECT10109, T73, UC8, WE372	33
dbj|AB195821.1|	0	0.97	0.993	SK1, DBVPG6044, NCYC110	33
dbj|AB195821.1|	0	0.97	0.996	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	33
dbj|AB195821.1|	0	0.97	0.996	UWOPS87-2421, YPS1000, I14	33
dbj|AB195821.1|	0	0.97	0.994	YPS606, 378604X, YS2, CLIB318, YJM421, YJM653, EM93, Y6, YPS163, K12, K1, FL100	33
dbj|AB195821.1|	0	0.97	0.994	Y9	33
dbj|AB195821.1|	0	0.97	0.994	CLIB272	33
dbj|AB195821.1|	0	1	0.994	YS4, CLIB208	33
tpg|BK006942.2|	0	0.88	0.763	FL100	34
tpg|BK006942.2|	0	0.88	0.764	YS2, CLIB272	34
tpg|BK006942.2|	0	0.88	0.763	378604X, Y9, YJM421, K12, K1	34
tpg|BK006942.2|	0	0.88	0.763	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	34
tpg|BK006942.2|	0	0.88	0.762	UWOPS87-2421, YPS1000, I14	34
tpg|BK006942.2|	0	0.88	0.763	YPS128, YPS163	34
tpg|BK006942.2|	0	0.88	0.763	YPS606	34
tpg|BK006942.2|	0	0.88	0.763	SK1, NCYC110	34
tpg|BK006942.2|	0	0.88	0.764	UWOPS83-787.3	34
tpg|BK006942.2|	0	0.88	0.764	YJM326, CECT10109	34
tpg|BK006942.2|	0	0.88	0.763	UC8, WE372	34
tpg|BK006942.2|	0	0.88	0.762	RM11	34
tpg|BK006942.2|	0	0.88	0.762	DBVPG3591	34
emb|FN393058.1|	3.00E-175	1	0.978	K12	35
emb|FN393058.1|	8.00E-177	1	0.981	K1	35
emb|FN393058.1|	7.00E-178	1	0.984	YJM436	35
emb|FN393058.1|	0	1	1	T73	36
emb|FN393058.1|	0	1	0.997	DBVPG6765, DBVPG1373, BC187, DBVPG1106, NCYC361, 322134S, YJM975, YJM320, YJM413, YJM454, DBVPG3591, DBVPG4651, CLIB294, CLIB154, CLIB157, DBVPG1399, M22, RM11, UC1, Y9J	36
emb|FN393058.1|	0	1	0.995	Y55, YJM434, Y8	36
emb|FN393058.1|	0	1	0.995	DBVPG1853	36
emb|FN393058.1|	0	1	0.987	YPS128, YPS606, Y12, UWOPS83-787.3, 273614N, Y9, W303, YJM326, YJM421, YPS1000, YPS163, CLIB413, CLIB483, K12, K1, YJM269, A364A, CENPK, I14	36
tpg|BK006935.2|	8.00E-177	1	0.981	DBVPG1788, YS9, YJM978, YJM981, DBVPG1794, Y3, UC8, WE372	36
emb|FN393058.1|	0	1	0.997	DBVPG6040	36
emb|FN393058.1|	0	1	0.979	TL229	36
tpg|BK006935.2|	3.00E-175	1	0.978	CLIB208	36
emb|FN393058.1|	0	1	0.995	Y6	36
tpg|BK006935.2|	0	1	0.995	378604X	36
tpg|BK006935.2|	0	1	0.997	YJM145	36
tpg|BK006935.2|	0	1	0.995	UWOPS87-2421	36
emb|FN393075.2|	6.00E-159	1	0.997	YJM436	37

emb|FN393075.2|	4.00E-174	1	0.991	CENPK	37
emb|FN393075.2|	9.00E-176	1	0.994	CLIB272, CECT10109	37
emb|FN393075.2|	8.00E-177	1	0.997	YJM269	37
emb|FN393075.2|	8.00E-177	1	0.997	378604X, 273614N, YJM326, CLIB483	37
emb|FN393075.2|	5.00E-160	1	1	UWOPS83-787.3, UWOPS87-2421, YJM413, YJM454, YPS1000, I14	37
emb|FN393075.2|	3.00E-169	0.99	1	DBVPG3591, M22, T73	37
emb|FN393075.2|	0	1	0.997	YJM421	37
emb|FN393075.2|	2.00E-178	1	1	DBVPG6765, Y55, YPS128, DBVPG6044, DBVPG1788, DBVPG1373, DBVPG1853, BC187, YPS606, L-1374, L-1528, DBVPG1106, NCYC361, YJM978, YIIc17_E5, YJM981, YJM975, NCYC110, CLIB192, YJM434, DBVPG1794, DBVPG4651, YPS163, CLIB294, A364A, CLIB154, CLIB157, CLIB219, DBVPG1399, RM11, UC1, UC8, WE372, Y8, Y9J	37
NA	NA	NA	NA	Y3	38
NA	NA	NA	NA	UWOPS87-2421, YPS1000, I14	38
NA	NA	NA	NA	YPS128, YPS606, YPS163	38
NA	NA	NA	NA	SK1, DBVPG6044, NCYC110	38
NA	NA	NA	NA	YS4, 378604X, Y9, YS2, CLIB192, CLIB324, CLIB272, YJM421, YJM428, YJM436, YJM653, EM93, Y6, K12, K1, FL100	38
NA	NA	NA	NA	UWOPS83-787.3	38
NA	NA	NA	NA	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2, YJM326, CECT10109, CLIB483	38
NA	NA	NA	NA	DBVPG3591, RM11	38
NA	NA	NA	NA	T73, UC8, WE372	38
emb|HE616749.1|	1.00E-28	0.88	0.775	273614N, TL229	39
gb|CP001751.1|	1.00E-33	0.58	0.67	DBVPG1399	40
gb|CP001751.1|	1.00E-34	0.58	0.672	UC8, WE372	40
ref|XM_754100.1|	1.00E-08	0.48	0.286	UWOPS83-787.3	42
ref|XM_754100.1|	2.00E-07	0.48	0.286	SK1	42
ref|XM_754100.1|	3.00E-09	0.48	0.294	K12, K1	42
ref|XM_754100.1|	1.00E-08	0.48	0.286	YPS1000	42
ref|XM_754100.1|	1.00E-08	0.48	0.286	CLIB219	42
ref|XM_003678764.1|	6.00E-24	0.8	0.304	UWOPS87-2421	43
ref|XM_003678764.1|	3.00E-26	0.83	0.299	YPS1000	43
ref|XM_003678764.1|	4.00E-26	0.83	0.299	Y3	43
emb|FN393082.1|	0	1	0.998	L-1374, L-1528, YJM981, YJM326, DBVPG3591, DBVPG4651, Y3, RM11	44
emb|FN393082.1|	0	1	0.998	DBVPG6765, DBVPG1788, DBVPG1373, BC187, DBVPG1106, NCYC361, 273614N, YJM434, DBVPG1794, CLIB294, DBVPG1399, T73, UC8, WE372, Y8	44
emb|FN393082.1|	0	1	0.991	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	44
emb|FN393082.1|	0	1	0.989	CLIB192	44
emb|FN393082.1|	0	1	0.991	Y12, K11, Y9, CLIB483, K12, K1, YJM269	44
emb|FN393082.1|	0	1	1	YJM978, YJM975, CLIB154, CLIB157, I14	44
emb|FN393082.1|	0	1	0.998	CLIB272	44
emb|FN393082.1|	0	1	0.998	YPS128, YPS606, UWOPS83-787.3, YJM413, YJM421, YJM436, YJM454, YPS163	44
emb|FN393082.1|	0	1	0.996	322134S, YIIc17_E5	44
emb|FN393082.1|	0	1	0.989	SK1, Y55, DBVPG6044, NCYC110	44
emb|FN393082.1|	0	1	0.991	UWOPS87-2421, YJM145, YJM320, YPS1000	44

emb|FN393082.1|	0	1	0.996	YJM280, CECT10109, Y9J	44
emb|X95505.1|	6.00E-114	1	0.897	Y55, DBVPG6044, NCYC110	45
emb|X95505.1|	5.00E-116	1	0.875	Y55, NCYC110	45
emb|X95505.1|	0	1	0.933	UWOPS87-2421	45
emb|Z37511.1|	0	1	0.998	DBVPG1399, UC8, WE372	45
emb|FR750555.1|	0	1	0.999	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	45
emb|FR750555.1|	0	1	0.999	UWOPS87-2421, YPS1000	45
NA	NA	NA	NA	CLIB483	46
gb|EF567080.1|	0	1	0.986	Y12, Y9, K12, K1, Y3, YJM269, CENPK	47
gb|EF567080.1|	0	1	0.986	YJM326, YJM421, Y6	47
gb|EF567080.1|	0	1	1	A364A	47
gb|EF567080.1|	0	1	0.984	UWOPS03-461.4	47
gb|EF567080.1|	0	0.99	0.875	CLIB483	47
gb|EF567080.1|	0	0.99	0.874	378604X	47
tpg|BK006948.2|	3.00E-150	1	0.997	CLIB272	49
emb|FN393074.1|	2.00E-157	1	0.997	273614N, YJM975, CLIB294, CLIB157, RM11	51
emb|FN393074.1|	6.00E-159	1	1	DBVPG6765, DBVPG1788, DBVPG1373, DBVPG1853, BC187, L-1374, L-1528, DBVPG1106, NCYC361, YS9, YJM978, YJM981, CLIB192, YJM145, YJM320, YJM326, YJM413, YJM434, YJM454, CECT10109, DBVPG1794, DBVPG3591, DBVPG4651, TL229, CLIB154, DBVPG1399, I14, M22, T73, UC1, UC8, WE372, Y8, Y9J	51
emb|FN393074.1|	6.00E-152	1	0.984	Y55, DBVPG6044, NCYC110	51
emb|FP929060.1|	1.00E-05	0.15	0.803	CLIB413, K1	52
emb|FP929060.1|	1.00E-05	0.15	0.803	YPS128, YPS606, YPS163	52
emb|FP929060.1|	1.00E-05	0.15	0.803	Y9	52
emb|FP929060.1|	1.00E-05	0.15	0.803	YJM269	52
emb|FP929060.1|	1.00E-05	0.15	0.803	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2, YJM145, YJM436	52
emb|FP929060.1|	1.00E-05	0.15	0.803	Y3	52
NA	NA	NA	NA	YIIc17_E5	53
NA	NA	NA	NA	L-1528, YS9, CLIB192, DBVPG1399, I14, M22, RM11, T73, Y8	53
NA	NA	NA	NA	DBVPG6765, DBVPG1788, DBVPG1853, L-1374, NCYC361, YS4, 378604X, 273614N, YS2, CLIB208, CLIB318, CLIB324, CLIB272, YJM280, YJM320, YJM428, YJM653, CECT10109, DBVPG1794, DBVPG3591, DBVPG4651, Y6, CLIB294, Y3, FL100, CLIB154, UC1, Y9J	53
NA	NA	NA	NA	UWOPS87-2421, K11, YPS1000, K12, K1	53
NA	NA	NA	NA	CLIB413	53
NA	NA	NA	NA	Y12, Y9	53
emb|FN394216.1|	5.00E-180	1	0.981	Y12, Y9	54
emb|FN394216.1|	0	1	0.984	UWOPS87-2421, K11, YJM320, YPS1000, K12, K1	54
emb|FN394216.1|	0	1	0.984	CLIB219	54
emb|FN394216.1|	0	1	0.984	YPS128, YPS606, YPS163	54
emb|FN394216.1|	5.00E-180	1	0.981	YJM269	54
emb|FN394216.1|	5.00E-180	1	0.981	YJM326	54
emb|FN394216.1|	0	1	0.989	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2, Y6	54
emb|FN394216.1|	0	1	0.987	SK1, DBVPG6044, NCYC110	54

emb|FN394216.1|	5.00E-180	1	0.981	YS4	54
ref|XM_003683475.1|	2.00E-134	1	0.896	TL229	55
emb|Z37510.1|	0	1	0.998	DBVPG1399, UC8, WE372	56
emb|FN677930.2|	6.00E-159	1	0.997	DBVPG1853, YJM428	57
emb|FN677930.2|	6.00E-159	1	0.997	TL229	57
emb|FN677930.2|	1.00E-160	1	1	DBVPG6765, Y55, DBVPG1788, DBVPG1373, BC187, L-1374, L-1528, DBVPG1106, NCYC361, YS9, 378604X, 273614N, YJM978, YIIc17_E5, YJM981, YJM975, YS2, CLIB192, CLIB272, YJM320, YJM434, YJM653, CECT10109, DBVPG1794, DBVPG3591, DBVPG4651, CLIB294, YJM269, CENPK, CLIB154, CLIB157, DBVPG1399, I14, M22, RM11, T73, UC1, UC8, WE372, Y8, Y9J	57
emb|FN677930.2|	6.00E-159	1	0.997	CLIB208, CLIB318, CLIB324	57
emb|FN677930.2|	1.00E-142	1	0.949	UWOPS83-787.3, YJM326	57
emb|FN677930.2|	7.00E-152	1	0.985	Y12, Y9, K12, K1	57
emb|FN677930.2|	2.00E-126	1	0.958	DBVPG6040, YJM280, YJM436	57
emb|FN677930.2|	1.00E-143	1	0.952	YPS128, YPS606, UWOPS87-2421, YJM145, YJM413, YJM454, EM93, YPS1000, YPS163, Y3	57
emb|FN677930.2|	1.00E-142	1	0.949	DBVPG6044, NCYC110	57
tpg|BK006945.2|	7.00E-146	1	0.984	YS4, CLIB318	58
tpg|BK006945.2|	7.00E-146	1	0.984	378604X, YJM436, CLIB483	58
tpg|BK006945.2|	1.00E-148	1	0.989	YPS128, YPS606, YPS163	58
tpg|BK006945.2|	7.00E-146	1	0.984	CENPK	58
tpg|BK006945.2|	2.00E-147	1	0.986	UWOPS87-2421, YS2, YJM145	58
tpg|BK006945.2|	6.00E-128	0.83	0.982	YJM421	58
tpg|BK006945.2|	2.00E-129	0.83	0.985	YPS1000	58
emb|HE616745.1|	1.00E-44	0.82	0.692	YJM269	59
emb|HE616745.1|	5.00E-124	0.91	0.693	CLIB219	59
ref|XM_003681408.1|	1.00E-51	0.78	0.72	SK1, Y55, DBVPG6044, NCYC110	59
emb|HE616745.1|	1.00E-131	0.91	0.693	DBVPG6040	59
emb|HE616745.1|	1.00E-131	0.91	0.693	YPS128, YPS606, YPS163	59
emb|HE616745.1|	7.00E-129	0.91	0.692	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	59
ref|XM_003681408.1|	1.00E-132	0.9	0.696	YJM436	59
ref|XM_003681408.1|	5.00E-130	0.9	0.694	UWOPS83-787.3, YJM413, YJM454	59
emb|HE616745.1|	2.00E-115	0.91	0.691	378604X	59
emb|HE616745.1|	4.00E-132	0.9	0.694	YJM320, YPS1000	59
ref|XM_003681408.1|	2.00E-49	0.76	0.717	YJM269	59
emb|HE616745.1|	6.00E-35	0.65	0.713	SK1, Y55, DBVPG6044, NCYC110	59
ref|XM_453073.1|	4.00E-96	0.9	0.708	CLIB413	60
emb|FN394216.1|	0	1	1	DBVPG6765, NCYC361, YIIc17_E5, CLIB294, UC1	61
tpg|BK006948.2|	2.00E-134	1	0.984	YJM269	62
tpg|BK006948.2|	8.00E-132	1	0.991	K11, YS9, Y9, CLIB272, CLIB413, K12, K1	62
tpg|BK006948.2|	2.00E-134	1	0.997	YS4	62
gb|HQ615872.1|	0	1	1	DBVPG1373, L-1374, DBVPG1106, YIIc17_E5, YJM434, DBVPG4651, M22, UC1, UC8	63
gb|HQ615872.1|	0	1	0.999	RM11	63
gb|HQ615872.1|	0	1	0.999	YS2	63

gb|HQ615872.1|	0	1	0.999	BC187	63
emb|FN393081.1|	1.00E-175	0.97	0.961	UWOPS83-787.3, YIIc17_E5, I14	64
emb|FN393081.1|	0	1	0.958	CLIB219	64
emb|FN393081.1|	2.00E-178	0.97	0.966	YPS128, YPS606, UWOPS87-2421, YPS1000	64
emb|FN393081.1|	7.00E-172	0.97	0.953	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	64
emb|FN393081.1|	0	0.97	0.987	RM11	64
emb|FN393081.1|	0	1	0.998	DBVPG1788, DBVPG1794, UC1, Y8	64
emb|FN393081.1|	0	1	0.998	L-1374, 378604X, DBVPG6040, DBVPG4651, T73	64
emb|FN393081.1|	0	1	0.997	CLIB192, CLIB324	64
emb|FN393081.1|	0	1	0.998	YJM421	64
emb|FN393081.1|	0	1	0.994	TL229	64
emb|FN393081.1|	0	1	0.995	YS4, YS2	64
emb|FN393081.1|	0	1	0.997	BC187, UWOPS87-2421, 273614N	64
emb|FN393081.1|	0	0.97	0.974	DBVPG6765, DBVPG1373, NCYC361	64
emb|FN393081.1|	0	0.97	0.974	Y55	64
emb|HG316466.1|	5.00E-153	0.99	0.736	YJM269	65
emb|HG316466.1|	5.00E-107	0.99	0.734	DBVPG6040	65
emb|HG316466.1|	0	0.91	0.728	378604X	65
emb|HG316466.1|	0	0.91	0.727	YPS128, YPS606, YPS163	65
emb|HG316466.1|	0	0.91	0.728	UWOPS83-787.3, YJM413, YJM454	65
emb|HG316466.1|	0	0.91	0.727	SK1, Y55, NCYC110	65
emb|HG316466.1|	0	0.9	0.727	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	65
emb|HG316466.1|	0	0.91	0.728	YJM436	65
emb|HG316466.1|	0	0.91	0.728	YJM320, YPS1000	65
emb|HG316466.1|	3.00E-91	0.75	0.722	YJM269	65
emb|HG316466.1|	0	1	1	DBVPG1373	66
gb|HQ615872.1|	0	1	0.999	YS2	66
gb|HQ615872.1|	0	0.99	1	DBVPG1373	66
gb|HQ615872.1|	0	1	1	BC187, L-1528, DBVPG1106, YIIc17_E5, CLIB192, CLIB208, CLIB324, YJM421, YJM428, YJM434, DBVPG4651, Y6, M22, RM11, UC1, UC8, WE372	66
gb|HQ615872.1|	0	1	0.999	YJM436	66
gb|HQ615872.1|	0	1	0.999	DBVPG6040	66
emb|HG316466.1|	0	1	0.999	DBVPG1853	66
emb|FN393070.1|	4.00E-148	1	0.993	Y55, DBVPG6044, NCYC110	67
emb|FN393070.1|	7.00E-151	1	1	TL229	67
emb|FN393070.1|	7.00E-145	1	0.987	DBVPG1853	67
tpg|BK006937.2|	1.00E-26	0.8	0.42	UWOPS87-2421, UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2, YPS1000	68
emb|FN393060.2|	2.00E-15	0.18	0.812	Y3	69
tpg|BK006937.2|	2.00E-05	0.4	0.488	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	70
tpg|BK006937.2|	6.00E-06	0.4	0.512	UWOPS87-2421, YPS1000	70
emb|FN393075.2|	0	0.88	0.68	DBVPG1399, UC8, WE372	70

emb|FN394217.1|	4.00E-154	1	0.997	DBVPG6765, DBVPG1373, NCYC361, YJM978, YJM981, YJM975, YJM413, YJM454, CECT10109, DBVPG3591, Y8	72
emb|FN394217.1|	5.00E-153	1	0.994	SK1, DBVPG6044, NCYC110	72
emb|FN394217.1|	1.00E-155	1	1	DBVPG1788, L-1374, L-1528, DBVPG1106, UWOPS87-2421, 273614N, CLIB272, YJM320, YJM434, DBVPG1794, DBVPG4651, CLIB294, CLIB483, Y3, A364A, CLIB154, CLIB157, DBVPG1399, I14, M22, RM11, T73, UC1, UC8, WE372, Y9J	72
emb|FN394217.1|	2.00E-151	1	0.987	Y12, Y9	72
emb|FN394217.1|	5.00E-153	1	0.994	BC187	72
gb|DQ443739.1|	0	1	0.989	Y3	73
gb|DQ443739.1|	0	1	0.986	YPS1000	73
gb|DQ443739.1|	0	1	0.987	SK1, K12	73
gb|DQ443739.1|	0	1	0.986	K1	73
gb|DQ443739.1|	0	1	0.988	UWOPS83-787.3	73
ref|XM_003681409.1|	5.00E-55	0.53	0.682	YPS1000	74
ref|XM_003681409.1|	6.00E-54	0.53	0.681	YJM269	74
ref|XM_003681409.1|	5.00E-55	0.53	0.682	YPS128, YPS606, YPS163	74
ref|XM_003681409.1|	6.00E-54	0.53	0.681	YJM436	74
ref|XM_003681409.1|	1.00E-55	0.53	0.686	UWOPS83-787.3, YJM413, YJM454	74
ref|XM_003681409.1|	5.00E-55	0.53	0.682	SK1, Y55, DBVPG6044, NCYC110	74
ref|XM_003681409.1|	2.00E-52	0.53	0.681	UWOPS03-461.4, UWOPS05-227.2	74
ref|XM_003681409.1|	7.00E-59	0.53	0.686	378604X	74
emb|FN393087.1|	0	1	0.993	UWOPS83-787.3, UWOPS87-2421, YPS1000	75
emb|FN393087.1|	0	1	0.991	CLIB219	75
emb|FN393087.1|	0	1	1	M22	75
emb|FN393087.1|	0	1	1	DBVPG6765, Y55, DBVPG1788, DBVPG1853, BC187, L-1528, DBVPG1106, NCYC361, YJM978, YIIc17_E5, YJM981, YJM975, CLIB208, YJM421, YJM434, CECT10109, DBVPG1794, DBVPG3591, DBVPG4651, EM93, CLIB483, CLIB157, I14, RM11, UC8, WE372, Y8	75
emb|FN393087.1|	0	1	0.999	YPS128, YPS606, YS4, YS2, CLIB318, CLIB324, YJM428, YJM436, Y6	75
dbj|AB031349.1|	0	1	1	SK1, 378604X, DBVPG6040, CLIB192	75
emb|FN393087.1|	0	1	0.997	YJM413, YJM454	75
emb|HE616744.1|	0	1	0.949	273614N	76
ref|XM_452193.1|	2.00E-157	0.89	0.5	UWOPS87-2421	77
emb|CR382122.1|	2.00E-26	0.67	0.487	CLIB483	77
emb|FN393063.1|	4.00E-155	1	0.992	CLIB483	78
emb|FN394216.1|	0	1	1	DBVPG6765, YS2, CLIB192, CLIB294	79
emb|FN394216.1|	0	1	1	DBVPG6765	79
emb|FN393070.1|	2.00E-120	1	0.961	Y9	80
emb|HE580276.1|	3.00E-24	0.32	0.568	273614N	81
emb|HE580276.1|	7.00E-23	0.38	0.554	Y12, Y9	81
emb|HE580276.1|	7.00E-23	0.38	0.554	CLIB413	81
ref|XM_004181469.1|	6.00E-05	0.34	0.59	Y3	82
ref|XM_004181469.1|	6.00E-05	0.34	0.59	UWOPS83-787.3, UWOPS87-2421, YPS1000	82
ref|XM_003042257.1|	6.00E-117	0.86	0.485	DBVPG1399	83
ref|XM_002143158.1|	2.00E-105	0.83	0.484	UC8, WE372	83

emb|FN393062.1|	4.00E-174	1	0.991	DBVPG1373, Y6, Y8	85
emb|FN393062.1|	2.00E-178	1	1	273614N, DBVPG3591, CLIB294, I14, Y9J	85
emb|FN393062.1|	4.00E-174	1	0.991	L-1528, YJM434	85
emb|FN393062.1|	9.00E-176	1	0.994	DBVPG1106, YJM981, YJM975, CLIB192, YJM421, CLIB154, CLIB157, DBVPG1399, UC8, WE372	85
emb|FN393062.1|	8.00E-177	1	0.997	DBVPG6765, Y55, L-1374, NCYC361, CECT10109, M22, UC1	85
emb|FN393062.1|	4.00E-174	1	0.991	YJM436	85
emb|FN393062.1|	4.00E-174	1	0.991	YS9, CLIB483	85
tpg|BK006942.2|	1.00E-141	0.92	0.978	YJM421	86
tpg|BK006942.2|	1.00E-135	1	0.99	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	86
tpg|BK006942.2|	2.00E-138	1	0.997	UWOPS87-2421, YPS1000	86
tpg|BK006948.2|	6.00E-159	1	0.997	UWOPS87-2421	87
emb|HE616746.1|	2.00E-165	1	0.936	273614N	88
emb|FN393070.1|	9.00E-150	1	1	BC187, DBVPG3591, Y8	89
emb|FN393070.1|	1.00E-155	1	0.997	YJM981, YJM434	89
emb|FN393070.1|	4.00E-148	1	0.997	YJM421, CLIB154	89
emb|Z37510.1|	3.00E-113	0.97	0.872	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	90
emb|Z37510.1|	2.00E-114	0.97	0.875	UWOPS87-2421, YPS1000	90
ref|XM_003680073.1|	0	1	0.956	273614N	91
gb|AC158404.2|	2.00E-08	0.32	0.412	378604X	92
emb|FN394216.1|	0	1	1	DBVPG6765, NCYC361, 273614N, YIIc17_E5, CLIB192, DBVPG4651, CLIB294, CLIB154, UC1	93
emb|FN394216.1|	0	1	0.963	M22	94
emb|FN394216.1|	0	1	0.96	DBVPG1373	94
emb|FN394216.1|	0	1	0.99	CLIB413	94
emb|FN394216.1|	0	1	0.997	CLIB154	94
emb|FN394216.1|	0	1	0.983	YS2	94
emb|FN394216.1|	1.00E-97	0.68	0.746	Y3	95
emb|FN394216.1|	4.00E-97	0.66	0.749	UWOPS87-2421, YPS1000, I14	95
emb|FN394216.1|	2.00E-99	0.68	0.748	DBVPG3591, RM11, T73, UC8, WE372	95
emb|FN394216.1|	5.00E-102	0.68	0.751	YJM326	95
emb|FN394216.1|	1.00E-96	0.68	0.744	CLIB483	95
emb|FN394216.1|	2.00E-100	0.68	0.75	CECT10109	95
emb|FN394216.1|	5.00E-102	0.68	0.751	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	95
emb|FN394216.1|	5.00E-102	0.68	0.751	YPS128, YPS606, YPS163	95
emb|FN394216.1|	4.00E-103	0.68	0.753	SK1, DBVPG6044, NCYC110	95
emb|FN394216.1|	2.00E-100	0.68	0.75	YS4, 378604X, Y9, YS2, CLIB192, CLIB208, CLIB318, CLIB324, CLIB272, YJM421, YJM428, YJM436, YJM653, EM93, Y6, K12, FL100	95
emb|FN394216.1|	4.00E-103	0.68	0.753	UWOPS83-787.3	95
emb|FN394216.1|	5.00E-102	0.68	0.751	CLIB219	95
emb|FN394216.1|	2.00E-87	0.66	0.735	UC8, WE372	95
emb|FN394216.1|	2.00E-87	0.66	0.735	DBVPG1399	95
emb|FN394216.1|	0	1	1	DBVPG6765, NCYC361, YS9, DBVPG6040, YIIc17_E5, YS2, CLIB192, CLIB208, Y6, CLIB294, UC1	95
ref|XM_002553465.1|	2.00E-72	0.8	0.659	UWOPS87-2421	96

NA	NA	NA	NA	273614N	97
ref|XM_003680067.1|	0	1	0.965	273614N	98
emb|FN394216.1|	0	1	1	DBVPG6765, NCYC361, YS9, YIIc17_E5, CLIB192, Y6, CLIB294, UC1	99
emb|FN394216.1|	0	1	0.999	DBVPG6040	99
ref|XM_003680079.1|	0	0.98	0.96	273614N	100
gb|HQ615872.1|	0	0.99	1	DBVPG1373, L-1374, L-1528, DBVPG1106, DBVPG6040, YIIc17_E5, CLIB318, YJM428, DBVPG4651, RM11, UC1, UC8, WE372	101
gb|HQ615872.1|	0	1	0.951	DBVPG1853	101
gb|HQ615872.1|	0	0.99	0.978	YS4, YS2, CLIB208	101
tpg|BK006934.2|	5.00E-109	1	0.871	378604X	102
emb|FN393063.1|	0	1	0.995	CLIB219	103
emb|FN393063.1|	0	1	0.998	YPS1000	103
emb|FN393063.1|	0	1	0.995	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	103
emb|FN393063.1|	0	1	0.995	YPS128, YPS606	103
ref|XM_003683022.1|	6.00E-36	0.2	0.689	273614N	104
ref|XM_003681397.1|	1.00E-175	0.88	0.707	DBVPG1373	105
emb|HE956757.1|	1.00E-17	0.91	0.552	Y12, K11, Y9, CLIB413, K1	106
gb|EU004203.1|	6.00E-52	0.58	1	UWOPS05-227.2	107
gb|EU004203.1|	2.00E-58	0.61	1	YJM145	107
emb|FN394216.1|	0	1	1	DBVPG6765, DBVPG1373, NCYC361, DBVPG6040, DBVPG4651, Y6, CLIB294, CLIB413, CLIB154, M22, UC1	108
emb|FN393074.1|	0	1	0.869	YIIc17_E5, YJM653, FL100, CLIB154, I14	109
emb|FN393067.1|	8.00E-132	1	0.953	DBVPG1788, DBVPG1373, BC187, NCYC361, YIIc17_E5, YJM981, CLIB324, YJM428, YJM434, DBVPG3591, Y3, UC1, Y9J	110
emb|CR382139.2|	5.00E-32	0.92	0.443	CLIB483	111
ref|XM_003680070.1|	0	1	0.97	273614N	112
emb|FN677930.2|	3.00E-152	0.99	0.915	Y12, 378604X, 273614N, Y9, YJM145, YJM326, YJM421	113
emb|FN394216.1|	0	1	0.999	CLIB154	114
emb|FN394216.1|	0	1	1	DBVPG1373, NCYC361, DBVPG4651, CLIB413, UC1	114
emb|FN394216.1|	0	1	0.999	CLIB294	114
emb|FN393074.1|	1.00E-173	1	0.989	SK1, DBVPG6044, NCYC110	115
tpg|BK006937.2|	1.00E-167	1	0.965	K12	116
tpg|BK006937.2|	6.00E-166	1	0.962	K11	116
gb|EU004203.1|	9.00E-97	0.51	1	YJM413, YJM421, YJM454	117
NA	NA	NA	NA	DBVPG3591	118
NA	NA	NA	NA	CECT10109	118
NA	NA	NA	NA	YPS1000, I14	118
gb|M81158.1|	0	1	0.986	UWOPS83-787.3, YJM326	119
gb|M81158.1|	0	1	0.986	UWOPS87-2421, YPS1000	119
gb|M81158.1|	0	1	0.982	Y3	119
gb|M81158.1|	0	1	0.985	Y12, Y9	119
gb|M81158.1|	0	1	0.984	YPS128, YPS606, YPS163	119
gb|GQ995455.1|	0	0.96	0.814	Y3	120
tpg|BK006940.2|	1.00E-60	0.95	0.799	Y9J	122

gb|HQ615872.1|	0	0.93	1	DBVPG1373, DBVPG1853, BC187, DBVPG1106, YJM434, DBVPG4651, Y6, M22, RM11, UC1, UC8, WE372	123
gb|HQ615872.1|	0	0.93	0.979	YS9	123
emb|FN393060.2|	3.00E-175	1	0.992	YJM145	124
emb|FN393060.2|	8.00E-177	1	0.995	DBVPG1106	124
emb|FN393060.2|	7.00E-178	1	0.997	UC8, WE372	124
ref|XM_003681432.1|	4.00E-123	1	0.713	273614N	125
emb|HG316456.1|	1.00E-100	0.84	0.673	UWOPS87-2421	126
emb|CR380947.2|	1.00E-08	0.13	0.74	Y3	127
emb|CR380947.2|	2.00E-10	0.29	0.74	Y55, DBVPG6044, NCYC110	127
gb|GU268671.1|	6.00E-68	1	0.925	YJM145	128
ref|XM_003681430.1|	2.00E-65	0.95	0.777	273614N, TL229	129
NA	NA	NA	NA	SK1, DBVPG6044, NCYC110, CENPK	130
NA	NA	NA	NA	YJM269	130
emb|AJ585533.2|	9.00E-62	0.95	0.818	EM93	131
ref|XM_002554171.1|	2.00E-45	0.94	0.417	UWOPS83-787.3	132
ref|XM_002554171.1|	7.00E-45	0.94	0.414	Y3	132
emb|CU928170.1|	1.00E-20	0.91	0.395	UWOPS87-2421, YPS1000	132
ref|XM_003680306.1|	0	1	0.973	273614N	133
emb|FN394216.1|	1.00E-60	1	0.793	UWOPS87-2421	134
emb|FN394216.1|	0	1	1	DBVPG6765, NCYC361, YS9, YIIc17_E5, YS2, CLIB208, CLIB294, UC1	135
emb|FN393080.1|	4.00E-80	0.75	0.852	CLIB219	136
tpg|BK006938.2|	1.00E-154	1	0.994	DBVPG1853	137
ref|XM_003683023.1|	4.00E-41	0.55	0.698	273614N	138
ref|XM_003680076.1|	0	1	0.944	273614N	140
tpg|BK006940.2|	3.00E-40	0.39	0.757	Y9J	141
emb|AJ585563.1|	0	1	0.851	CLIB272, YJM653	142
emb|FN394216.1|	0	1	1	CLIB208	143
emb|FN393084.1|	0	1	1	DBVPG1373, DBVPG1853, BC187, L-1528, YS9, CLIB192, CLIB324, YJM428, DBVPG4651, M22, RM11, UC1, UC8	144
emb|FN393084.1|	0	1	0.999	L-1374	144
tpg|BK006935.2|	1.00E-124	1	0.991	UWOPS05-217.3	145
emb|HE616744.1|	0	1	0.949	273614N	146
tpe|HG323650.1|	0	1	0.995	UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2	147
ref|XM_004177543.1|	3.00E-10	0.27	0.664	UWOPS87-2421	148
tpg|BK006941.2|	8.00E-139	0.98	0.914	Y9	149
tpg|BK006948.2|	6.00E-159	1	0.965	YJM421	150
ref|XM_003680074.1|	0	1	0.965	273614N	151
emb|FN393063.1|	5.00E-160	1	1	YJM436, Y6	153
emb|FN394216.1|	0	1	1	DBVPG6765, NCYC361, YS9, DBVPG6040, YIIc17_E5, YS2, CLIB192, CLIB208, Y6, CLIB294, UC1	154
tpg|BK006940.2|	1.00E-133	0.85	0.722	Y9J	155
NA	NA	NA	NA	YJM269	156
NA	NA	NA	NA	SK1, NCYC110, CENPK	156

tpg|BK006941.2|	0	1	0.998	YJM434	157
emb|FN394216.1|	0	1	1	DBVPG6765, NCYC361, YS9, 273614N, YIIc17_E5, CLIB192, CLIB208, DBVPG4651, Y6, CLIB294, CLIB154, UC1	158
ref|XM_003680068.1|	0	1	0.905	273614N	159
tpg|BK006934.2|	5.00E-160	1	0.968	UWOPS03-461.4, UWOPS05-227.2	160
emb|FN394216.1|	4.00E-167	1	1	DBVPG6765, NCYC361, DBVPG6040, YIIc17_E5, CLIB192, Y6, CLIB294, UC1	161
emb|FN393070.1|	0	1	0.981	SK1	162
ref|XM_003680064.1|	0	1	0.944	273614N	163
emb|FN393078.1|	7.00E-96	0.8	0.994	YJM436	164
ref|XM_001482837.1|	4.00E-11	0.85	0.341	Y6	165
emb|FN394216.1|	0	1	1	DBVPG6765, NCYC361, YIIc17_E5, YS2, CLIB192, CLIB208, Y6, UC1	167
NA	NA	NA	NA	273614N, TL229	168
tpg|BK006937.2|	0	1	0.81	YPS606, YPS163	169
emb|Z37510.1|	0	1	0.995	DBVPG1399, UC8, WE372	171
ref|XM_003680071.1|	0	0.99	0.911	273614N	172
ref|XM_003680065.1|	0	1	0.96	273614N	173
gb|M10604.1|YSCMEL1A	0	1	0.996	UWOPS87-2421, UWOPS03-461.4, UWOPS05-217.3, UWOPS05-227.2, YPS1000	174
emb|FN393060.2|	1.00E-148	1	0.993	CLIB219, Y9J	175
emb|HG316466.1|	1.00E-149	1	0.991	BC187, YS9	176
ref|XM_003678764.1|	2.00E-25	0.97	0.467	UWOPS83-787.3, Y3	177
tpg|BK006945.2|	0	1	0.988	YJM280	178
tpg|BK006937.2|	5.00E-147	1	0.975	DBVPG6044, NCYC110	179
emb|FN394216.1|	1.00E-154	1	1	DBVPG6765, NCYC361, YS9, DBVPG6040, YIIc17_E5, YS2, CLIB208, Y6, CLIB294, UC1	180
emb|FN393060.2|	8.00E-157	1	1	YJM975	181
emb|FN394216.1|	0	1	0.999	DBVPG6765, NCYC361, YS9, DBVPG6040, YIIc17_E5, CLIB192, CLIB208, Y6, CLIB294, UC1	183
tpg|BK006936.2|	8.00E-137	0.9	0.757	Y3	184
ref|XM_003680077.1|	0	1	0.967	273614N	185
emb|HE616744.1|	0	1	0.925	273614N	186
emb|HE616748.1|	1.00E-162	1	0.969	273614N	187
emb|HE616743.1|	0	0.76	0.907	273614N	188
emb|FN393070.1|	4.00E-168	1	1	CLIB294	189
ref|XM_001482837.1|	1.00E-36	0.81	0.503	Y6	190
emb|Z86109.1|	0	0.99	0.775	CLIB272	191
gb|M19944.1|YSCRSDS	2.00E-144	1	1	DBVPG1788, YJM981	192
emb|FN394216.1|	0	1	1	DBVPG6765, NCYC361, YIIc17_E5, YS2, CLIB192, CLIB208, CLIB294, UC1	193
dbj|AB195821.1|	0	0.97	0.994	UWOPS83-787.3	NA

Table A-7 List of non-reference genes identified

Best hit BLAST ID for *de novo* assembled genes that do not align to the reference genome. E-value, percent of query covered by hit, and the percent identity of the hit are listed. Strains with the same best BLAST hit are grouped. New gene group combines BLAST hits with an e-value smaller than $1x10^{-10}$.

Appendix B

Supplementary Figures and Tables for Chapter 3

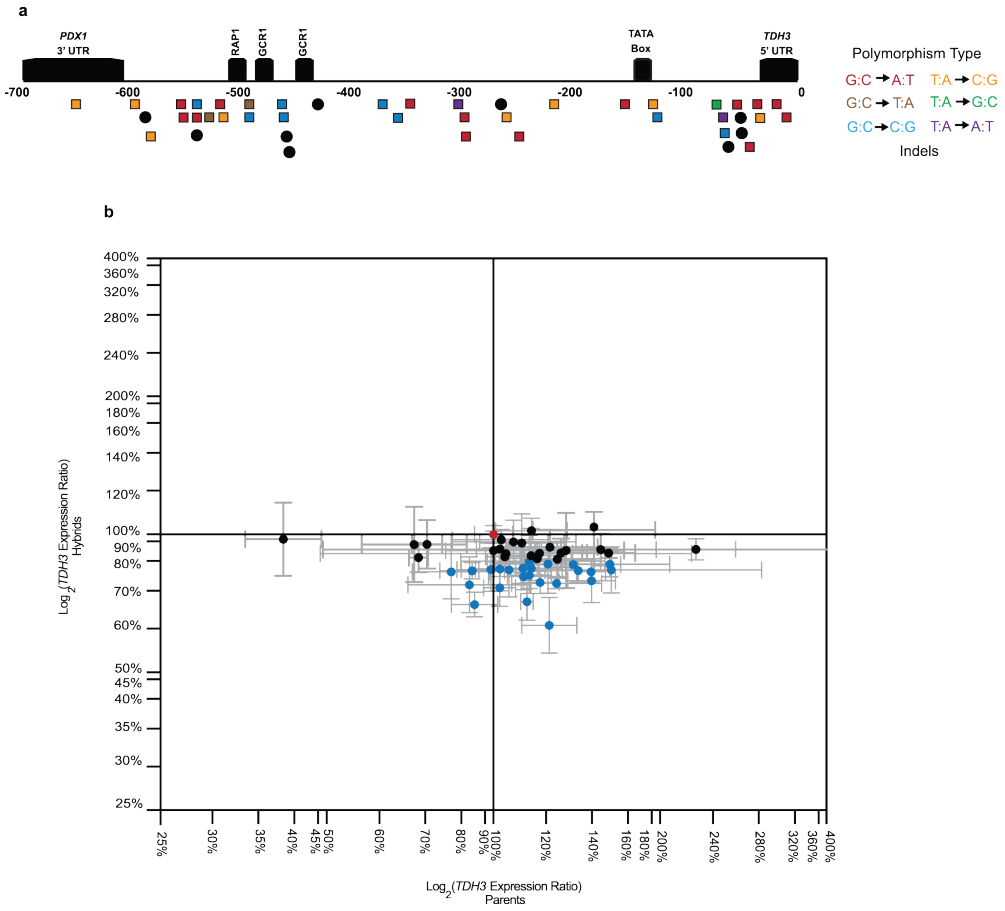


Figure B-1 *TDH3* promoter polymorphisms influence *TDH3* mRNA levels

a, Locations of polymorphisms within the *TDH3* promoter relative to known functional elements, including RAP1 and GCR1 transcription factor binding sites, are shown. Squares are point mutations, circles are indels. red, G:C→A:T; yellow, G:C→T:A; blue, G:C→C:G; orange T:A→C:G; green, T:A→G:C; purple, T:A→A:T. **b,** The $\log_2$ ratio of total expression divergence between natural isolates and a reference strain (x-axis) versus the $\log_2$ ratio of total *cis*-regulatory expression divergence between natural isolates and the reference strain (y-axis) is shown. Error bars are 95% CI. The 25 of 48 strains with significant *cis*-regulatory differences from the reference strain are shown in blue. Reference strain is shown in red. These data show differences in *cis*- and *trans*- regulation among strains, but do not reveal the evolutionary changes that give rise to these differences.

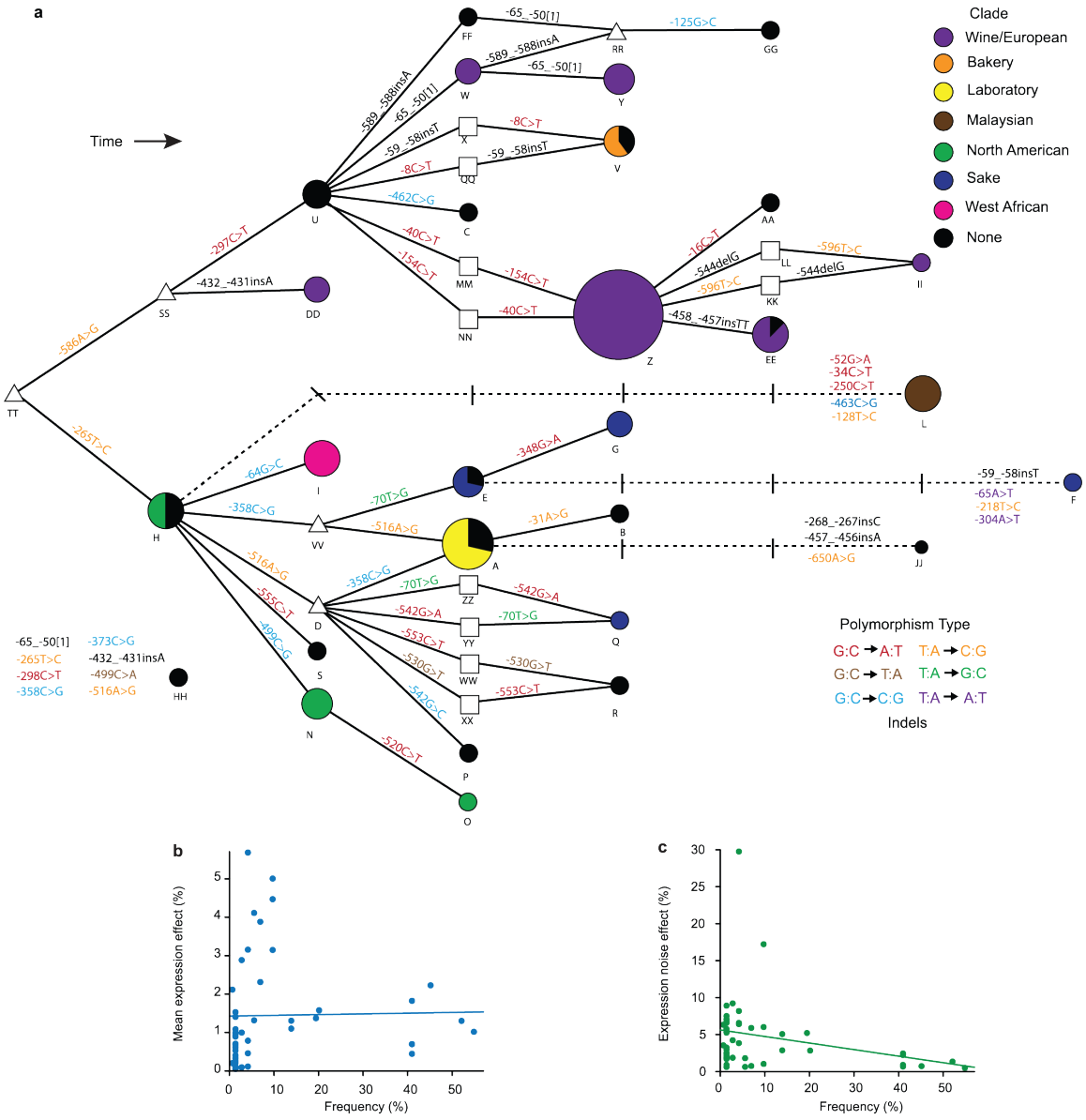


Figure B-2 Ancestral state reconstruction of the TDH3 promoter

a, The *TDH3* promoter haplotype network is shown with the inferred ancestral strain at the left. Circles represent haplotypes observed among the 85 strains with their diameters proportional to haplotype frequency. The haplotypes are colored according to clade (Supplementary Table 1). Triangles are haplotypes that were not observed among the strains sampled, but must exist or have existed as intermediates between observed haplotypes. Squares are possible intermediates connecting two observed haplotypes, but it is unknown which of these actually exists or existed in *S. cerevisiae*. Solid lines connect haplotypes that differ by a single mutation; dashed lines connect haplotypes that differ by multiple mutations. Mutations on each branch are colored by the mutation type as in Extended Figure 1a. b, Relationship between the effect of a polymorphism on mean expression level and the frequency of that polymorphism among the strains sampled (p-value = 0.43). c, Relationship between the effect of a polymorphism on expression noise and the frequency of that polymorphism among the strains sampled (p-value = 0.0028).

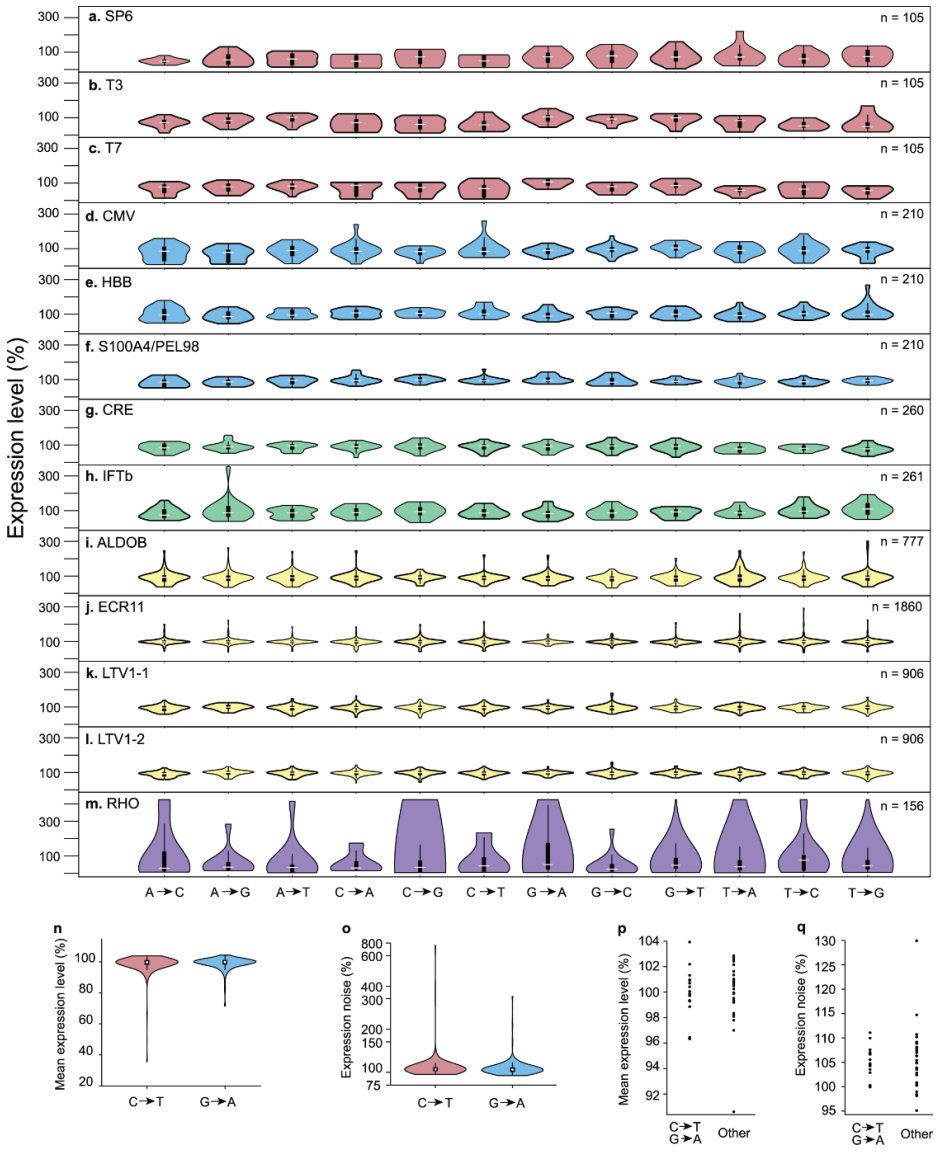


Figure B-3 No significant difference between mutation types

Distributions of effects on mean expression level from previous random mutagenesis experiments are shown partitioned by mutation type. For each mutation type, the distribution (inside) and density (outside, colored) of the effects on mean expression level are shown. The number of mutations tested for each promoter is shown in the upper right corner of each panel. **a**, bacteriophage SP6 promoter. **b**, bacteriophage T3 promoter. **c**, bacteriophage T7 promoter. **d**, human CMV promoter. **e**, human HBB promoter. **f**, human S100A4/PEL98 promoter. **g**, synthetic cAMP-regulated enhancer. **h**, interferon-B enhancer. **i**, ALDOB enhancer. **j**, ECR11 enhancer. **k**, LTV1 enhancer replicate 1. **l**, LTV1 enhancer replicate 2. **m**, rhodopsin promoter. Red: Patwardhan *et al.* 2009 bacteriophage promoters(PATWARDHAN *et al.* 2009). Blue: Patwardhan *et al*. 2009 mammalian promoters(PATWARDHAN *et al.* 2009). Green: Melnikov *et al*. 2012 mammalian enhancers(MELNIKOV *et al.* 2012). Yellow: Patwardhan *et al.* 2012 mammalian promoters(PATWARDHAN *et al.* 2012). Purple: Kwasnieski *et al.* 2012 promoter(KWASNIESKI and MOGNO 2012). **n**, Distribution of effects for C→T (red) and G→A (blue) mutations for mean expression level in this study. **o**. Same as **n**, but for expression noise. **p**, Distribution of effects for C→T/G→A polymorphisms compared to other polymorphism types for mean expression level in this study. **q**, same as **p**, but for gene expression noise.

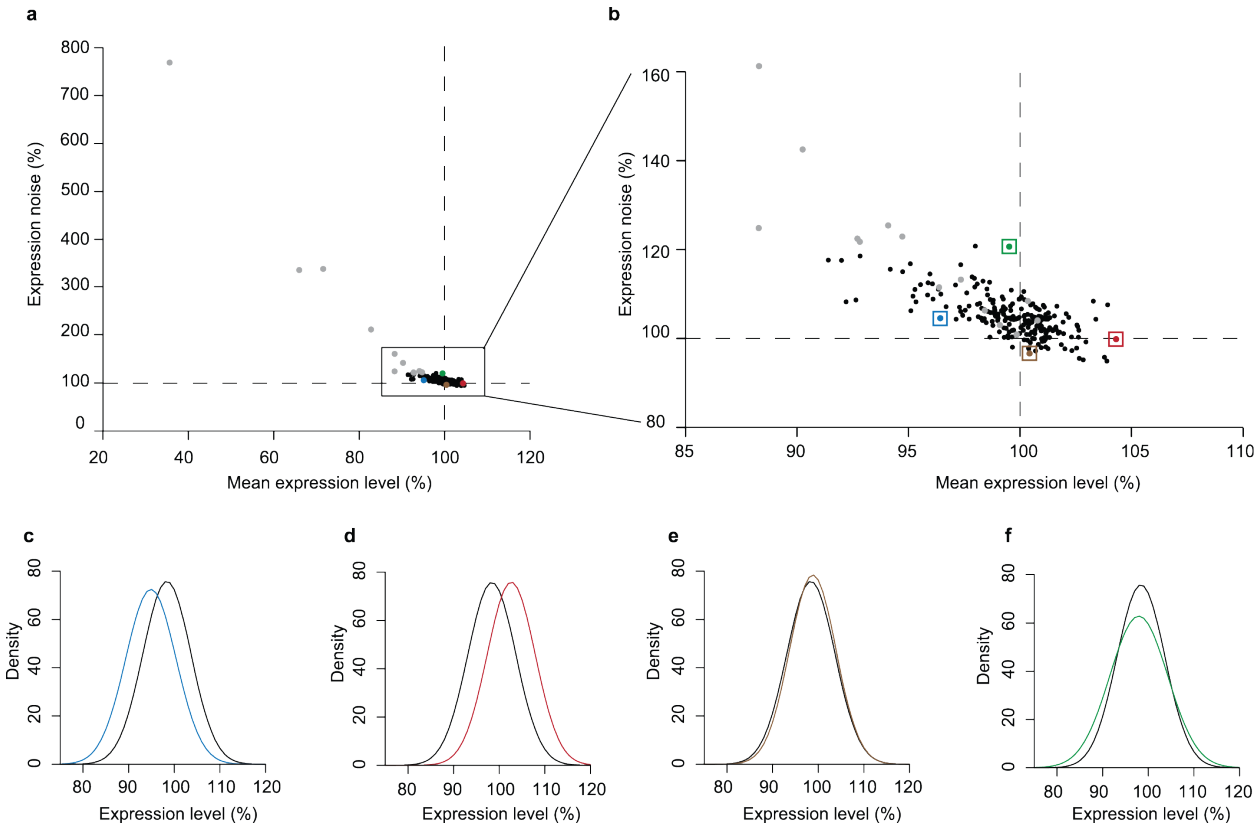


Figure B-4 Correlation between mean expression level and expression noise.

a, Correlation between mean expression level (x-axis) and expression noise (y-axis) for the 236 point mutations in the *TDH3* promoter (R^2=0.85) is shown. Gray points correspond to mutations in known transcription factor binding sites. Colored points correspond to individual mutations highlighted in **c-f. b,** Alternative plot showing the majority of data from **a** more clearly**,** gray and colored points are the same as in **a. c,** Distribution of gene expression phenotypes from a mutant (blue) with decreased mean expression level but similar expression noise as the reference strain (black). Outside of the known TFBS, 50% of mutations decreased mean expression. **d,** Distribution of gene expression phenotypes from a mutant (red) with increased mean expression level but similar gene expression noise as the reference strain (black). Outside of the known TFBS, 50% of mutations increased mean expression. **e,** Distribution of gene expression phenotypes from a mutant (brown) with decreased gene expression noise but similar mean expression level as the reference strain (black). Outside of the known TFBS, 13% of mutations decreased expression noise. **f,** Distribution of gene expression phenotypes from a mutant (green) with increased gene expression noise but similar mean expression level as the reference strain (black). Outside of the known TFBS, 87% of mutations increased expression noise.

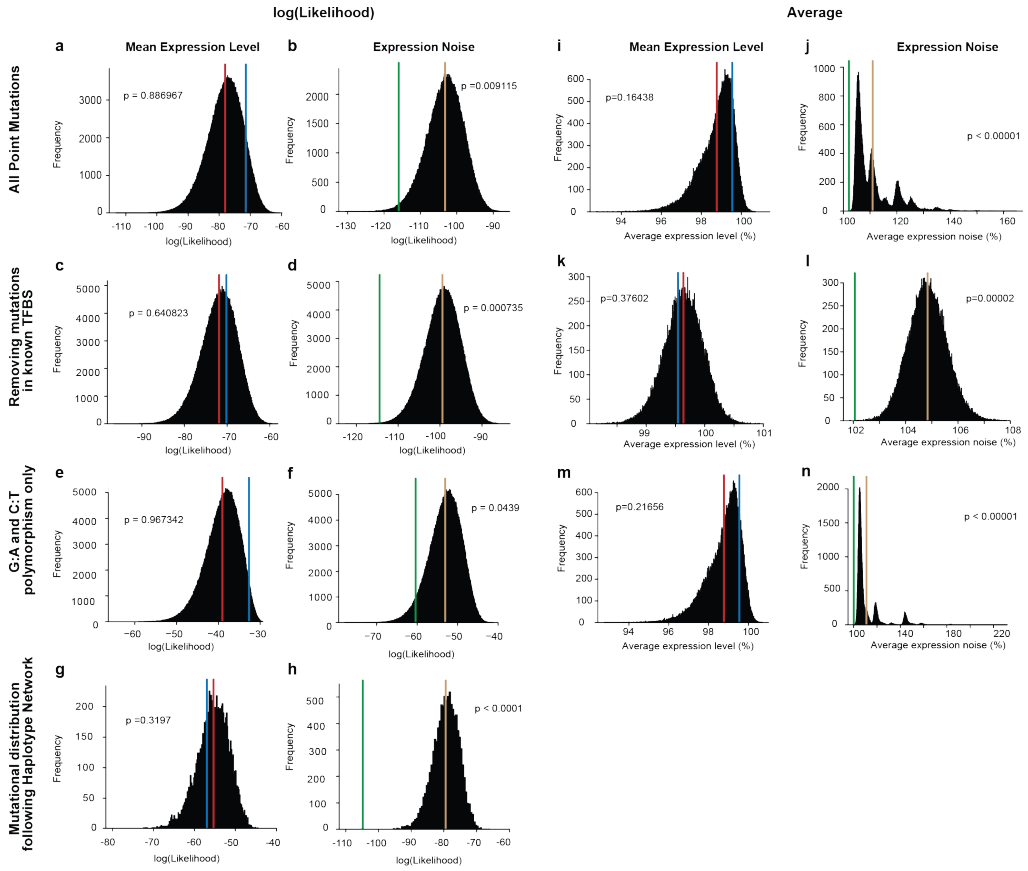


Figure B-5 Tests for selection

a-h, Tests for selection using likelihood. **a,** The distribution of likelihood values for 100,000 randomly sampled sets of 45 mutations drawn from the mutational effect distribution is shown for mean expression level. The average likelihood for all samples of mutations tested (red) as well as the likelihood of the observed polymorphisms (blue) are also shown. **b**, Same as **a**, but for expression noise. The average likelihood for all mutation samples tested is shown in brown and the likelihood of the observed polymorphisms is shown in green. **c**, Same as **a**, but with the large effect mutations in the TFBS removed from the mutational effect distribution used for sampling. **d**, Same as **b**, but after removing the mutations in the TFBS from the mutational effect distribution. **e**, Same as **a**, but using only G→A and C→T polymorphisms. **f**, same as **b**, but using only G→A and C→T polymorphisms. **g**, Distribution of likelihoods for 10,000 random walks along the *TDH3* promoter haplotype network using the effects from the mutational distribution is shown. **h**, Same as **e**, but for expression noise. **i-n,** Tests for selection using average effects. **i**, The distribution of average effects for 100,000 randomly sampled sets of 45 mutations drawn from the mutational effect distribution is shown for mean expression level (black). Polymorphisms do not have a significantly different average mean expression (blue, 99.5%) than sets of mutations (red, 98.8%; *p*-value = 0.16438). This figure is comparable to Extended Data figure 5a, but uses average effects instead of the likelihoods to test for differences in distribution between random mutations and polymorphisms. **j**, Same as **i**, but for expression noise. Polymorphisms have significantly lower average expression noise (green, 102.1%) than sets of random mutations (brown, 110.9%; *p*-value < 0.00001). **k**, Same as **i**, but with the large effect mutations in the TFBS removed from the mutational effect distribution used for sampling (polymorphisms, 99.5%; mutations, 99.6%; *p*-value = 0.37602). **l**, Same as **j**, but after removing the mutations in the TFBS from the mutational effect distribution (polymorphisms, 102.1%; mutations, 104.8%; *p*-value = 0.00002). **m**, Same as **i**, but using only G→A and C→T polymorphisms (polymorphisms, 99.7%; mutations, 98.8%; *p*-value = 0.21656). **n**, same as **j**, but using only G→A and C→T polymorphisms (polymorphisms, 100.0%; mutations, 110.9%; *p*-value < 0.00001).

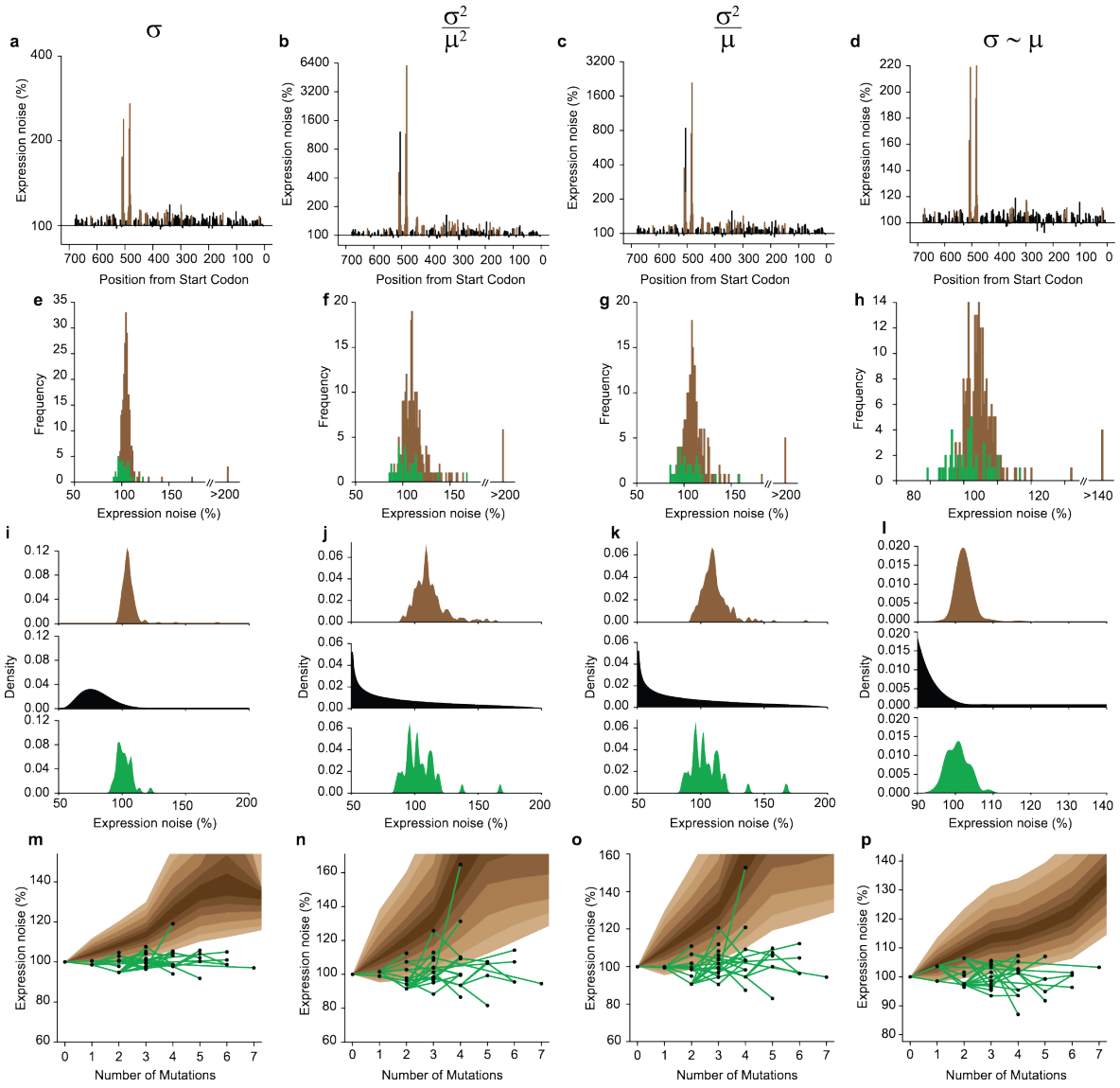


Figure B-6 Alternative Metrics for Quantifying Expression Noise

a-d, Distributions of effects for mutations on gene expression noise across the *TDH3* promoter with expression noise quantified as σ (**a**), σ^2/μ^2 (**b**), σ^2/μ (**c**), and residuals from the regression of σ on μ (**d**), **e-h**, Distributions of effects for mutations on gene expression noise (brown) compared to polymorphisms (green) with noise quantified as σ (**e**), σ^2/μ^2 (**f**), σ^2/μ (**g**), and residuals from the regression of σ on μ (**h**). **i-l**, The maximum likelihood fitness function (middle, black) relating the distribution of mutational effects (top, brown) to the distribution of observed polymorphisms (bottom, green) for expression noise quantified as σ (**i**), σ^2/μ^2 (**j**), σ^2/μ (**k**), and residuals from the regression of σ on μ (**l**). **m-p**, Changes in expression noise observed among haplotypes over time in the inferred haplotype network (Figure E2a) are shown in green. The brown background represents the 95th, 90th, 80th, 70th, 60th and 50th percentiles, from light to dark, for expression noise resulting from 10,000 independent simulations of phenotypic trajectories in the absence of selection where noise is quantified as σ (**m**), σ^2/μ^2 (**n**), σ^2/μ (**o**), and residuals from the regression of σ on μ (**p**). p-values for all test are located in Table AI - 1.

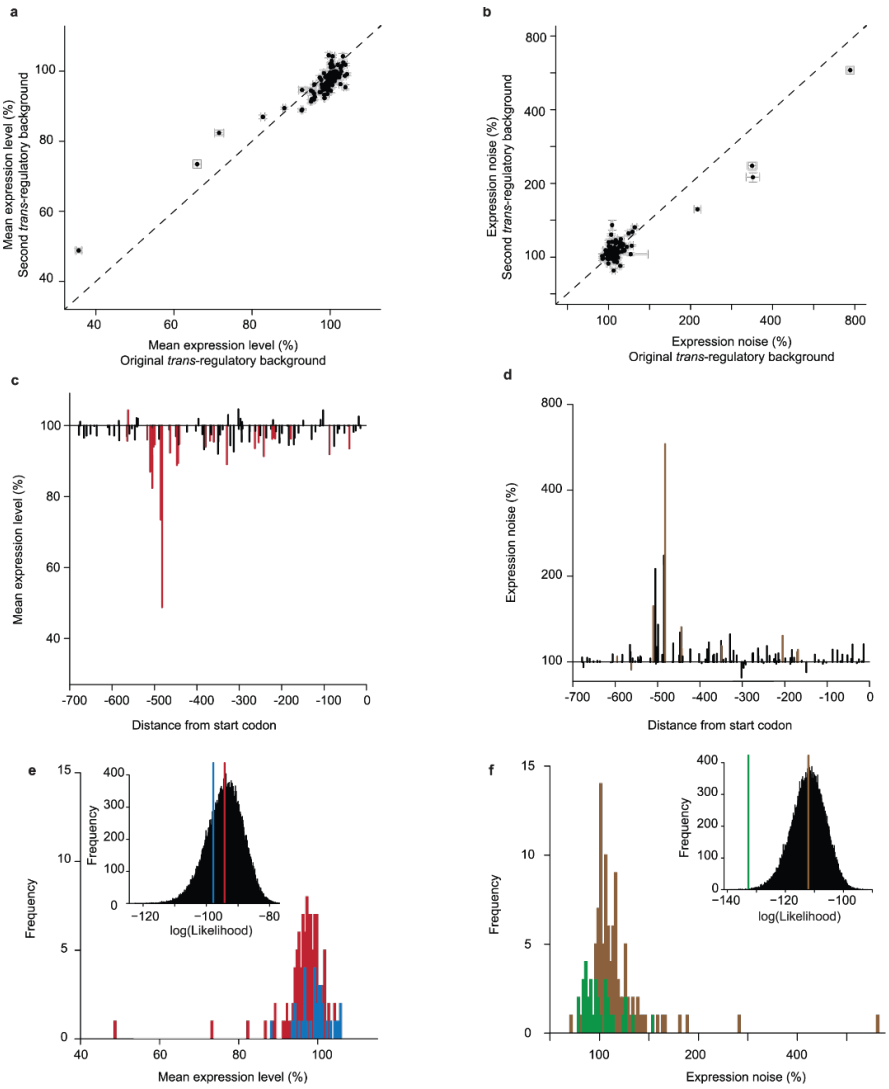


Figure B-7 Effects in a second *trans*-regulatory background

a, A comparison between effects of mutations on mean expression in the original *trans*-regulatory background (x-axis) and a hybrid *trans*-regulatory background between BY4741 and YPS1000 (y-axis) is shown. Error bars are 95% confidence intervals. **b**, Same as **a**, but for gene expression noise. **c,** Effects of individual mutations on mean expression level in the hybrid trans-regulatory background are shown in terms of the percentage change relative to the un-mutagenized reference allele, and are plotted according to the site mutated in the 678bp region (significant mutations: red lines, *t*-test, Bonferroni corrected). Note that most mutations decrease expression, unlike in the original genetic background. **d**, Same as **c**., but for gene expression noise (significant mutations: brown lines, *t*-test, Bonferroni corrected). **e**, Distribution of *de novo* mutation effects in the second *trans*-regulatory background (red) compared with the effects of naturally occurring haplotypes in this *trans*-regulatory background (blue). Inset: the distribution of likelihood values for 100,000 randomly sampled sets of 27 mutations drawn from the mutational effect distribution is shown for mean expression level. The average likelihood for all samples of mutations tested (red) as well as the likelihood of the observed polymorphisms (blue) are also shown (p-value = 0.2584). Removing mutations in the known TFBS resulted in a significant difference between mutations and polymorphisms (p-value = 0.00781). **f**, Same as **e**, but for gene expression noise. Mutations, brown. Polymorphisms, green (p-value = 0.00037). Removing mutations in the known TFBS did not change this result (p-value < 0.00001).

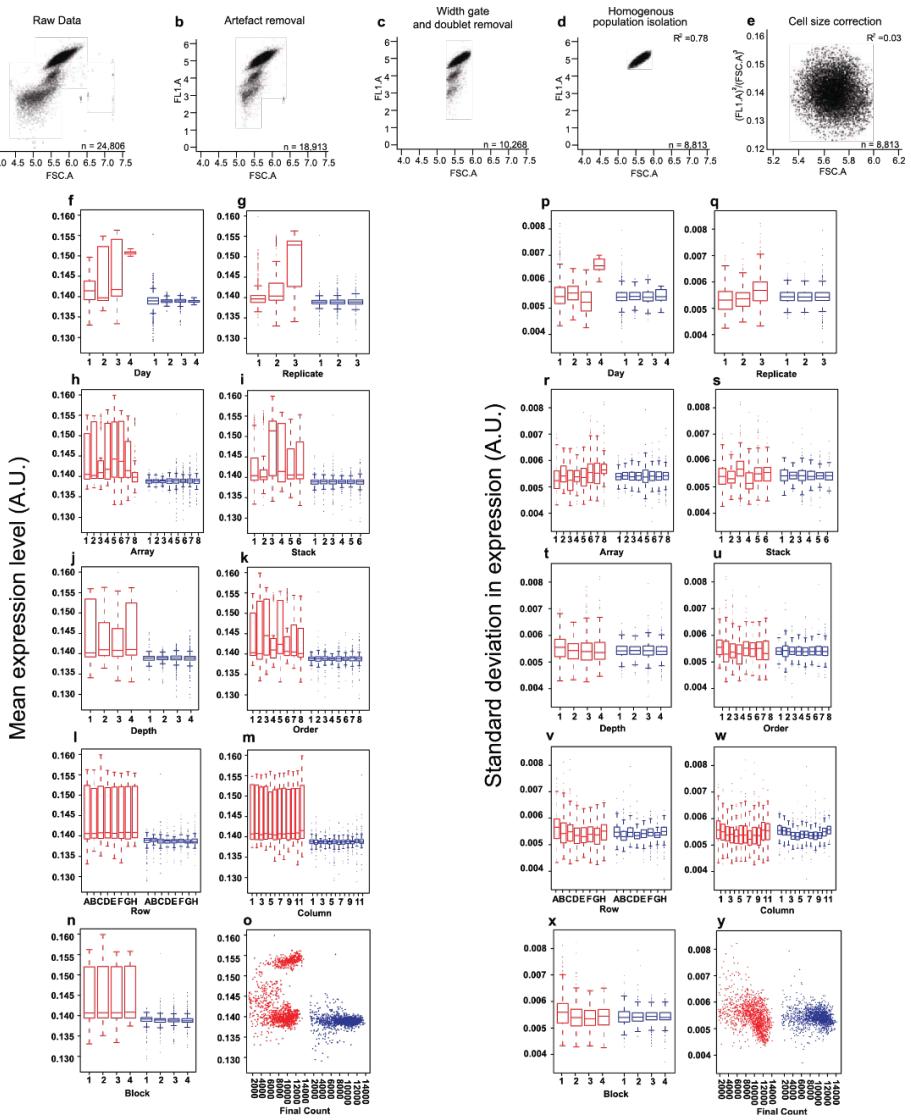


Figure B-8 Methodology for the analysis of flow cytometry data

a, Raw data from the flow cytometer is shown for the first control sample collected. Each point is an individual event scored by the flow cytometer, the vast majority of which are expected to be cells. FSC.A is a proxy for cell size, and FL1.A is a measure of YFP fluorescence. Log_{10} values are plotted for both FSC.A and FL1.A. **b**, The same sample is shown after events found in the negative control sample (using hard gates on FSC.A and FL1.A) were excluded. **c**, The same sample is shown after flowClust was used to remove events likely to be from multiple cells entering the detector simultaneously. **d**, The same sample is shown after flowClust was used to isolate the densest homogenous population within the sample. The R^2 value shown is the correlation between YFP fluorescence and cell size. **e**, After correcting for differences in cell size, the correlation between YFP fluorescence and cell size was nearly 0 and not significant. In all panels, the number of events analyzed (i.e., sample size) is shown in the bottom right corner. Box plots of mean expression of control samples before (red) and after (blue) correcting for the effects of individual plates for each day on which samples were run (**f**), for replicates nested within day (**g**), for array nested within day and replicate (**h**), for stack nested within day (**i**), for depth nested within day (**j**), for order nested within day and replicate (**k**), for row nested within array (**l**), for column nested within array (**m**), for block nested within array (**n**), and for the final cell count (**o**). The y-axis is in arbitrary units. **p-x**, same as **f-o**, but for gene expression noise.

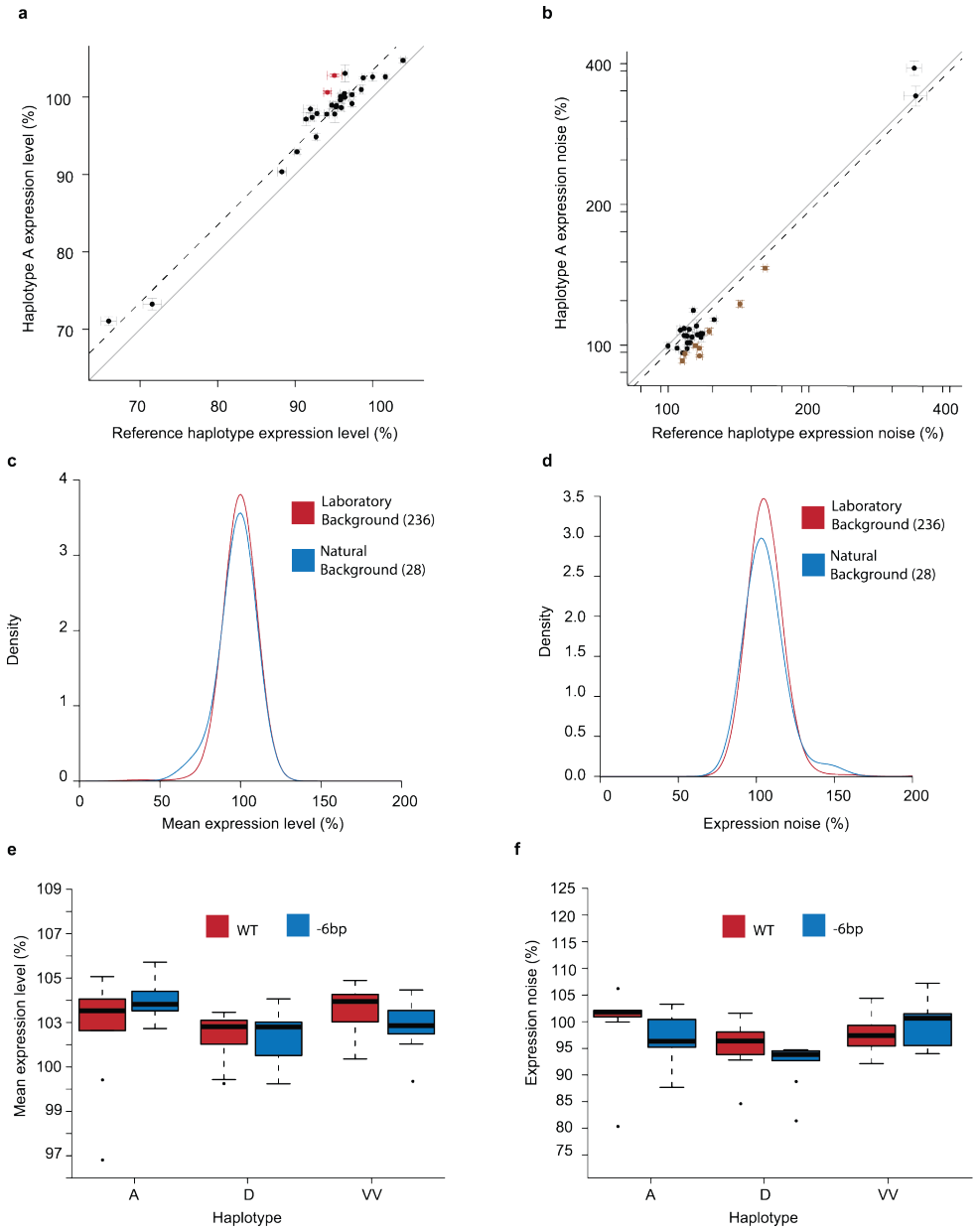


Figure B-9 Consistency of mutational effects on different genetic backgrounds

a, The effects on mean expression level for each of the 28 mutations tested on both the reference haplotype (x-axis) and natural haplotype A observed in wild strains (y-axis) are shown. These two haplotypes differ by a single point mutation. Solid lines show expression from the P_{TDH3} haplotypes on which the two sets of mutations were created, both of which were defined as 100% activity. The gray line shows y = x. The dashed line shows the consistent increase in mean expression level when these mutations were tested on haplotype A. Error bars show 95% CI. Colored points have significantly different effects on the two backgrounds (p-value < 0.05, ANOVA, Bonferroni corrected), indicating weak epistasis. **b**, Same as **a**, but for gene expression noise. **c**, Distributions of mutational effects for mean expression levels are shown based on the 236 point mutations on tested on the reference haplotype (red) as well as for the 28 mutations tested on haplotype A (blue). **d**, Same as **c**, but for gene expression noise. **e**, The effect on mean expression of the full *TDH3* promoter (red) compared to promoters containing 6 fewer bp at the 5' end (blue). Each box plot summarizes data from 9 replicates. **f**, Same as **e**, but for expression noise.

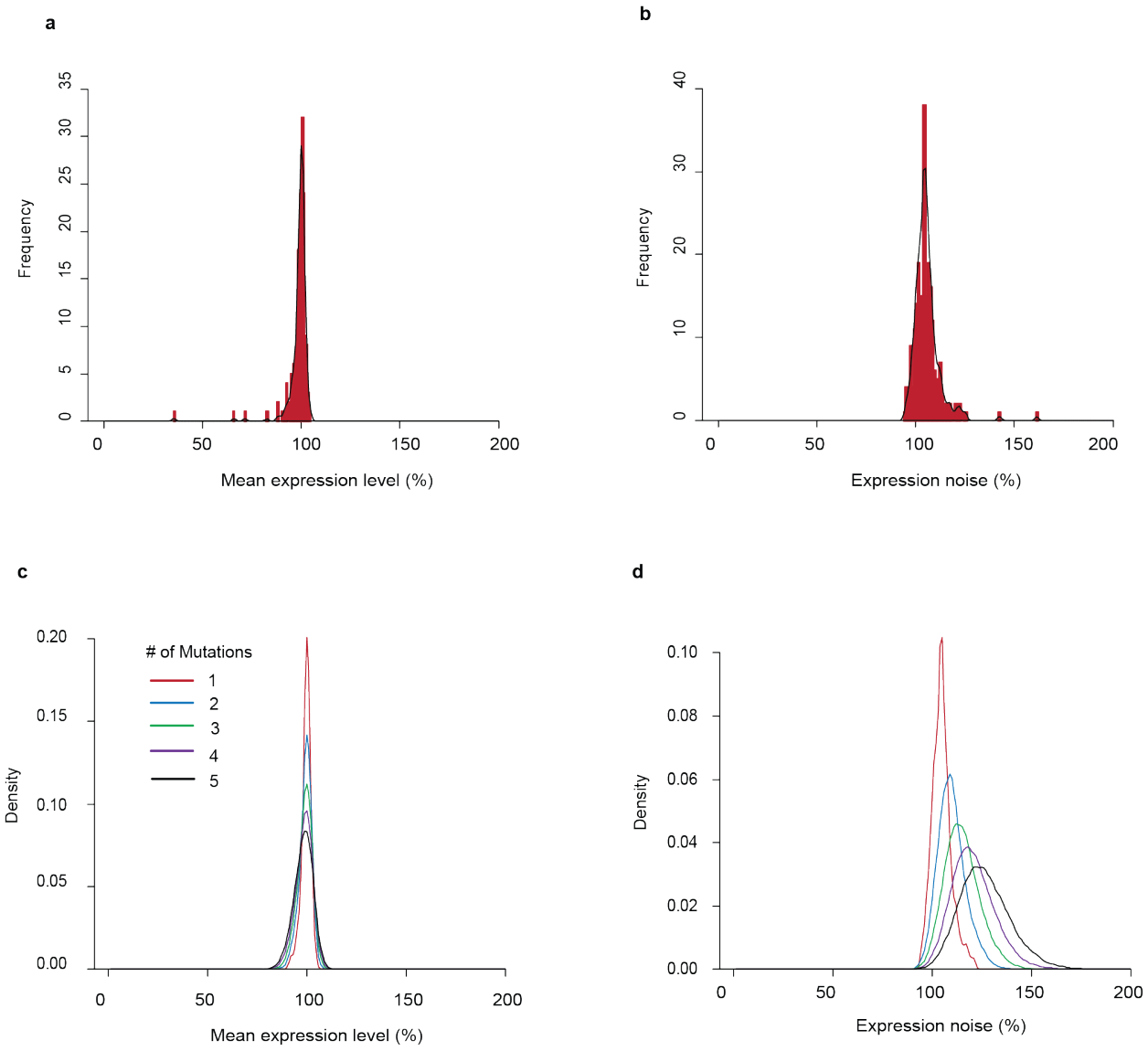


Figure B-10 Probability distributions for mutational effects

a, A histogram summarizing the mutational effects on mean expression level is shown (red), overlaid with the density curve (black line) used to calculate the likelihood of an effect on mean expression level. **b**, Same as **a**, but for expression noise. **c**. Density curves for the effects of one (red), two (blue), three (green), four (purple) or five (black) mutations randomly drawn from the distribution of mutational effects observed for mean expression level. **d**, Same as **c**, but for expression noise.

Metric*†	Distribution§	No TFBS Mutations§	G:A/C:T Polymorphism§	Fitness Function¶	Network‖
μ	0.88697	0.64082	0.96734	0.87	0.3197
σ	<0.00001	<0.00001	<0.00001	9×10^{-6}	<0.0001
σ/μ	0.00912	0.00076	0.04390	0.00019	<0.0001
σ^2/μ^2	<0.00001	<0.00001	0.00760	0.00002	<0.0001
σ^2/μ	<0.00001	<0.00001	<0.00001	0.00015	<0.0001
Regression‡	<0.00001	<0.00001	<0.00001	4×10^{-7}	<0.0001

*mean expression (μ) †standard deviation in expression (σ) ‡Residuals from linear model: $\sigma \sim \mu$
§100,000 ermutations ‖10,000 ermutations ¶Likelihood Ratio

Table B-1 Alternative Metrics for Quantifying Expression Noise

p-values for tests of selection using mean expression (μ) and five metrics of expression noise, including σ/μ which is used throughout the main text.

Appendix C

Creation of an improved strain for mapping complex traits in *Saccharomyces cerevisiae*

Introduction

Determining the genetic and molecular mechanisms underlying complex phenotypes often requires identifying the causative genetic loci and nucleotides contributing to these traits (RAUSHER and DELPH). However, in the yeast *Saccharomyces cerevisiae*, the current laboratory strains, S288c and its descendants, have several phenotypes that limit their usefulness in high throughput mapping approaches.

For example, *S. cerevisiae* isolates from the wild readily undergo meiosis under nutrient starvation and the majority of individual diploids sporulate. By contrast, S288c enters meiosis slowly and only a small proportion of individuals successfully complete meiosis, even under ideal conditions (DEUTSCHBAUER and DAVIS 2005; GERKE *et al.* 2006). Because genetic mapping requires recombination, and thus, meiosis, the limited meiotic abilities of S288c reduces the number and speed at which mapping populations can be created.

In addition to poor sporulation, S288c and its descendants generate petite cells lacking mitochondria with high frequency. As a consequence, these individuals cannot perform aerobic respiration and often have altered phenotypes compared to wild-type individuals (CHEN and CLARK-WALKER 1999). Because linking phenotypes to their genomic location requires high quality phenotyping, additional variation introduced by petite individuals can reduce the accuracy and power of genetic mapping.

Finally, many complex traits are affected by alleles that are recessive and it preferable to map the genetic basis of a trait within haploids when possible. However, upon meiosis, yeast generate both **a** and α haploids that will readily reform diploids. Current techniques for limiting the recreation of diploids suffer from a lack of throughput and poor specificity (TONG *et al.* 2001). As a consequence, the power to map the genetic basis of recessive traits in yeast is reduced.

Results

To overcome these deficiencies, we modified S288c to increase its sporulation rate and density, reduce the frequency at which it generated petites, and to express a fluorescent marker that allowed easy identification of mating type. To accomplish these goal, we obtained several strains derived from S288c, specifically variants of BY4722, BY4724, BY4730, BY4724, and BY4742. These strains vary in their mating type and auxotrophies, facilitating crossing. In addition, these strains differ at a set of alleles derived from natural *S. cerevisiae* strains that either improve sporulation rate or lower petite frequency. These include versions of *TAO3* and *RME1* that increase sporulation

rate (DEUTSCHBAUER and DAVIS 2005) and versions of *SAL1*, *CAT5*, and *MIP1* that decrease petite frequency (DIMITROV *et al.* 2009). An allelic variant at *MKT1* has also been identified that affects both sporulation and petite frequency. However, while the wild-type S288c allele decreases sporulation rate, it also substantially reduces petite frequency compared to the alternative allele and we kept the S288c version (DEUTSCHBAUER and DAVIS 2005; DIMITROV *et al.* 2009).

Through a series of crosses, transformations, and sporulations, we isolated a single individual that contained the desired set of alleles and was free of all auxotrophies except for ura3Δ0 (Figure C-1). We retained the *URA3* auxotrophy to facilitate future genetic manipulation by the *delitto perfetto* method, which requires *5-FOA* counter-selection and therefore a starting strain that is ura- (STORICI and RESNICK 2006). To facilitate the creation of the correct strain, we tracked the allelic identity of each segregating locus using pyrosequencing (Table C-1,Table C-2). After identification, the isolated individual was turned into a diploid and sporulated to generate isogenic **a** and α haploids. To the **a** haploid, we introduced a red fluorescent protein into the *MAT* locus (CHIN *et al.* 2012). This marker allows identification of individuals based on their mating type using fluorescence assisted cell sorting (FACS). Using this marker, populations containing millions of individuals of the same mating type can be collected in minutes. Finally, both **a** and α strains contain a *TDH3* promoter driving Yellow Fluorescent Protein expression located at the *HO* gene to facilitate mapping of mutations and polymorphisms influencing *TDH3* expression.

Methods

When permitted, growth was performed using YPD (20g Glucose, 20g peptone, 10g yeast extract per 1 L water; 20g agar for solid plates). For crosses involving auxotrophies, synthetic complete media was used, minus the appropriate amino acids (1.7g Yeast nitrogen base, 5 g Ammonium Sulfate, 20g glucose per 1 L water; 20g agar for solid plates). Sporulation was induced by growth on YPD plates for 24 hours at room temperature, followed by plating on KAc plates at room temperature (10g Potassium acetate, 0.5g glucose per 1 L water; 20g agar for solid plates). Ascus walls were dissolved prior to tetrad dissections by incubating spores in 200 μl zymolyase (1 mg/ml 20T) for 1 hour without shaking.

To create homozygous diploids from haploids, strains were transformed with plasmid pCM66. pCM66 contains a galactose inducible copy of *HO* and a selective nourseothricin resistance marker. After transformation, nourseothricin resistant cells were grown with galactose as the sole carbon source at 30°C without shaking for 8 hours to induce expression of *HO*. This allowed for mating type switching and subsequent mother-daughter cell mating to produce diploids. Cells were streaked for single colonies on YPD plates and the ploidy of single colonies checked by colony PCR using mating-type specific primers. Diploid colonies were streaked onto fresh, non-selective, YPD plates and assayed for loss of nourseothricin resistance, and thus pCM66.

Pyrosequencing was used to follow sporulation and petite QTN. Methods are as described in (WITTKOPP 2012). PCR primers used are in Table C-1. Dispensation order for pyrosequencing is in Table C-2.

References

CHEN X. J., CLARK-WALKER G. D., 1999 The Petite Mutation in Yeasts: 50 Years On. Int. Rev. Cytol. **194**: 197–238.

CHIN B. L., FRIZZELL M. a., TIMBERLAKE W. E., FINK G. R., 2012 FASTER MT: Isolation of Pure Populations of a and Ascospores from Saccharomyces cerevisiae. G3 **2**: 449–452.

DEUTSCHBAUER A. M., DAVIS R. W., 2005 Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat. Genet. **37**: 1333–40.

DIMITROV L. N., BREM R. B., KRUGLYAK L., GOTTSCHLING D. E., 2009 Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. Genetics **183**: 365–83.

GERKE J. P., CHEN C. T. L., COHEN B. a, 2006 Natural isolates of Saccharomyces cerevisiae display complex genetic variation in sporulation efficiency. Genetics **174**: 985–97.

RAUSHER M. D., DELPH L. F., Commentary : When does understanding phenotypic evolution require identification of the underlying genes ? : 1–33.

STORICI F., RESNICK M. a M., 2006 The delitto perfetto approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. Methods Enzymol. **409**: 329–45.

TONG a H., EVANGELISTA M., PARSONS A. B., XU H., BADER G. D., PAGÉ N., ROBINSON M., RAGHIBIZADEH S., HOGUE C. W., BUSSEY H., ANDREWS B. J., TYERS M., BOONE C., 2001 Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science **294**: 2364–8.

WITTKOPP P. J., 2012 Using Pyrosequencing to Measure Allele-Specific mRNA Abundance and Infer the Effects of Cis- and Trans-regulatory Differences. In: Orgogozo V, Rockman M V. (Eds.), *Molecular Methods for Evolutionary Genetics*, Methods in Molecular Biology. Humana Press, Totowa, NJ.

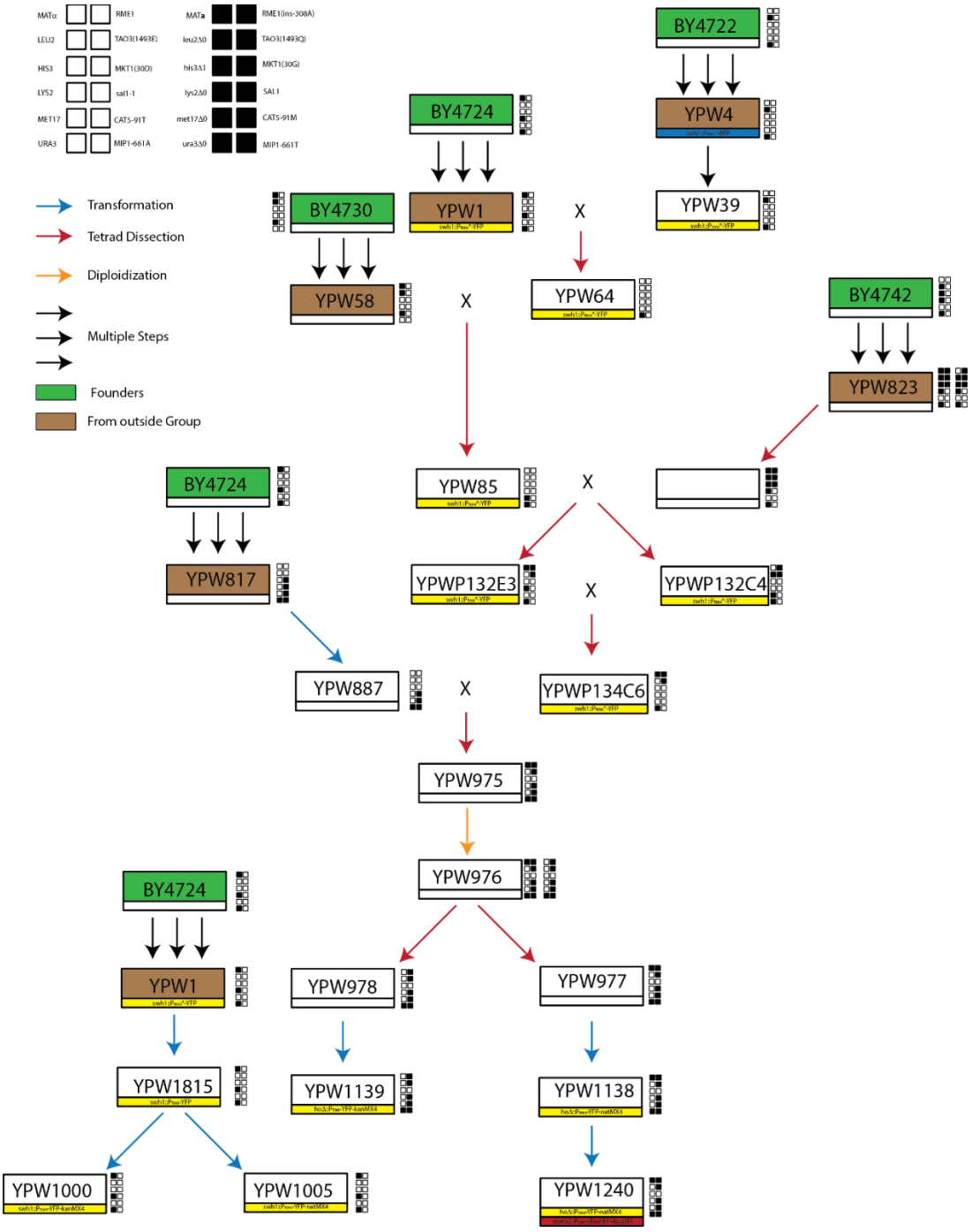


Figure C-1 Crossing, transformation, and sporulation scheme for strain creation

Each box represents a unique strain. Strains in green are the original founding strains and strains in brown are those originally obtained by the lab. Genotypes at segregating markers are shown to the right of each strain. Auxotrophies and mating type on the left, sporulation and petite alleles on the right. The shading of the box indicates the allele at a particular locus for each strain as given in the upper left corner. All crosses are designated by an X, tetrad dissections by a red arrow, transformations by a blue arrow, and diploidizations by an orange arrow. Black arrows represent multiple steps taken prior to receiving a strain. Below each strain, the identity of fluorescent markers is labeled.

Primer Number	Primer Name[1]	Sequence[2]	Gene
1530	RME1_psq_R_bio	/5Biosg/GCACTCTGGCCTTTGTTCTC	RME1
1529	RME1_psq_F	TGCTTCGTCACGTAAAATGG	RME1
1538	RME1_pyro_F	AAGTGGCCGGGCATGTA	RME1
1532	TAO3_psq_F_bio	/5Biosg/TCTCCTTGGGTTTGTATGGTTT	TAO3
1533	TAO3_psq_R	GAGAGAGCAGTTCGGCAAAT	TAO3
1534	TAO3_pyro_R	AGAATTATGTAATTTCGTTT	TAO3
1536	MKT1_psq_R_bio	/5Biosg/TGGCTCTTGGGGTTGAATAG	MKT1
1535	MKT1_psq_F	TCCTATGCCATTGAGGCTCT	MKT1
1537	MKT1_pyro_F	TCTGAATAATTGTACCCTGG	MKT1
1588	SAL1_psq_F_bio	/5Biosg/CATTGCTGGTGGTTTAGCTG	SAL1
1589	SAL1_psq_R	TGTGGTAGGTTCAGGGTCTTTG	SAL1
1590	SAL1_pyro_R	CACCTCTGTAAAATAATCTG	SAL1
1591	CAT5_psq_F_bio	/5Biosg/AGTACTTCGTGTTGGCTCATAGGT	CAT5
1592	CAT5_psq_R	AGTGCCCTCCGATTACTGTC	CAT5
1593	CAT5_pyro_R	TGATGTATCTCCTGGTCC	CAT5
1594	MIP1_psq_F_bio	/5Biosg/CCAATTTGTAGTCCCCAGTTGTAA	MIP1
1595	MIP1_psq_R	CCTGGAGGAGCTTTGACTTGAGT	MIP1
1596	MIP1_pyro_R	GGATGCGGTTAACCA	MIP1

Table C-1 Primer sequences for tracking sporulation and petite QTN by pyrosequencing

[1]psq R and F primers used for PCR; pyro primer used for sequencing. [2] /5Biosg/ designates biotinylated primer

Gene	Sequence to analyze[1]	Dispensation order	S228c Allele	Alt Allele
RME1	AT/CAAATATATCACCGTATTTTC	GATCGATATATCA	-	A
TAO3	G/CAACAGCTGAAA	TGCTACAG	C	G
MKT1	A/GTATAGACG	TAGCTATAG	A	G
SAL1	AC/GCCCC/ACCCTCA	GATCGCCACCGTCA	T	C
CAT5	CAC/TATGTGC/TTTT	GCAGTCGATGT	C	G
MIP1	CGC/TATTTTCCA	GCGACTGATTCA	C	T

Table C-2 Sequences and dispensation order for pyrosequencing

[1]Bases on either side of / designates alternative alleles

Appendix D

Mapping small effect mutations in *Saccharomyces cerevisiae*: impacts of experimental design and mutational properties[1]

Abstract

Genetic variants identified by mapping are biased toward large phenotypic effects because of methodological challenges for detecting genetic variants with small phenotypic effects. Recently, bulk segregant analysis combined with next generation sequencing (BSA-seq) was shown to be a powerful and cost-effective way to map small effect variants in natural populations. Here, we examine the power of BSA-seq for efficiently mapping small effect mutations isolated from a mutagenesis screen. Specifically, we determined the impact of segregant population size, intensity of phenotypic selection to collect segregants, number of mitotic generations between meiosis and sequencing, and average sequencing depth on power for mapping mutations with a range of effects on the phenotypic mean and standard deviation as well as relative fitness. We then used BSA-seq to map the mutations responsible for three EMS-induced mutant phenotypes in Saccharomyces cerevisiae. These mutants display small

[1] This appendix is published as: Duveau F., Metzger B. P. H., Gruber J. D., Mack K., Sood N., Brooks T. E., Wittkopp P. J., 2014 Mapping Small Effect Mutations in *Saccharomyces cerevisiae*: Impacts of Experimental Design and Mutational Properties. G3: 1205–1216.

quantitative variation in the mean expression of a fluorescent reporter gene (-3%, +7% and +10%). Using a genetic background with increased meiosis rate, a reliable mating type marker, and Fluorescence-Activated Cell Sorting (FACS) to efficiently score large segregating populations and isolate cells with extreme phenotypes, we successfully mapped and functionally confirmed a single point mutation responsible for the mutant phenotype in all three cases. Our simulations and experimental data show that the effects of a causative site not only on the mean phenotype, but also on its standard deviation and relative fitness should be considered when mapping genetic variants in microorganisms such as yeast that require population growth steps for BSA-seq.

Introduction

Characterizing the causal relationships between genotypes and phenotypes is a major goal of modern genetics. Bulk segregant analysis (BSA), in which two phenotypically distinct sub-populations (bulks) of recombinant progeny (segregants) are isolated from a genetic cross and genotyped, is one way to achieve this goal (MICHELMORE *et al.* 1991). With this method, regions of the genome contributing to the phenotypic difference between the two pools of segregants are identified because causative alleles (and linked loci) occur at different frequencies in the two bulks. BSA is a cost-effective approach to mapping because genotypes are determined only for the two bulk samples rather than each of the individual recombinants. The recent development of high throughput sequencing, which can be used to determine allele frequencies for nearly all sites in the genome in each phenotypic pool simultaneously, has made BSA particularly effective for mapping polymorphisms in organisms with small genomes such as yeast (EHRENREICH *et*

al. 2010; POMRANING *et al.* 2011; LITI and LOUIS 2012; WILKENING *et al.* 2013). Even small differences in allele frequency between bulks can be detected with this genotyping-by-sequencing approach (Parts et al. 2011), allowing detection of small effect variants. Because BSA requires sorting large numbers of individuals based on their phenotype, it is particularly well suited to the analysis of traits that can easily be selected or scored in the laboratory, such as growth in different environments (WENGER *et al.* 2010; SWINNEN *et al.* 2012; EHRENREICH *et al.* 2012; YANG *et al.* 2013) or expression of a fluorescent reporter gene (ALBERT *et al.* 2014).

BSA can be used to identify sites contributing to natural variation (Leeuwen et al. 2012; Granek et al. 2012; Bastide et al. 2013; Parts et al. 2011) or mutant phenotypes isolated from genetic screens (WICKS *et al.* 2001; BRAUER *et al.* 2006; XIA *et al.* 2010). Experimental design and statistical properties of BSA coupled with high-throughput sequencing (BSA-seq) for mapping quantitative trait loci (QTL) have been examined in detail (MAGWENE *et al.* 2011; EDWARDS and GIFFORD 2012); however, methods for mapping mutations using BSA-seq after a mutagenesis screen have received less theoretical attention (but see Birkeland et al. 2010). Compared to natural variation, the density of polymorphic sites is usually much lower after a mutagenesis screen and the mutations are more likely to have effects on fitness. As a result, optimal experimental design and statistical power are expected to be different for BSA-seq when analyzing natural variation and mutant genotypes created by random mutagenesis in the lab. For example, sequencing information from linked segregating sites can be combined when mapping natural variation to increase the power of detection (MAGWENE *et al.* 2011;

EDWARDS and GIFFORD 2012), but this is usually not possible with the lower genetic diversity present after mutagenesis. In such cases, sequencing coverage sufficient for statistical analysis must be recovered from the causative site itself.

Here, we examine the influence of experimental design and mutational properties on the mapping success of BSA-seq when the density of segregating sites is low, with the goal of providing a general framework for large-scale mapping of small effect mutations after a mutagenesis screen. We describe the effect on mapping sensitivity of four experimental parameters (population size, intensity of phenotypic selection, number of mitotic generations between meiosis and sequencing, and sequencing depth) as well as three mutation properties (effect on mean phenotype, effect on standard deviation of the phenotype, and effect on fitness) that can potentially bias genotype frequencies in the segregant bulks. Previous studies modeling BSA-seq for QTL mapping primarily considered the effects of a genetic variant on the mean phenotype for the trait of interest (PARTS *et al.* 2011; MAGWENE *et al.* 2011).

We used the results from this computational modeling to design a bulk segregant mapping experiment suitable for identifying mutations in yeast causing small changes in expression of a YFP fluorescent reporter gene controlled by the S. cerevisiae TDH3 promoter. These mutations were previously isolated from a low-dose mutagenesis screen in which each haploid mutant recovered was predicted to have, on average, 47 new mutations with only one affecting fluorescence of the reporter gene (GRUBER *et al.* 2012). Our simulations indicated that isolating very large pools of haploid segregants (>105

cells) with extreme fluorescence phenotypes was essential for mapping success given the biological properties of the mutant strains. To achieve this, we developed an experimental system for efficiently collecting phenotypically divergent cells from a population of haploid segregants that uses (i) a genetic background with a higher meiosis rate than the typical laboratory strain (DEUTSCHBAUER and DAVIS 2005), (ii) a robust and tractable mating type marker to efficiently isolate stable haploid bulks (CHIN *et al.* 2012), and (iii) fluorescence-activated cell sorting (FACS) for high-throughput phenotyping and selection of individuals with extreme fluorescence levels. Genetic variants responsible for changes in mean YFP expression as small as 3% relative to the wild type genotype were then successfully mapped despite their significant impact on fitness and confirmed using allele replacement, showing that BSA-seq is a powerful method for identifying small effect mutations after a mutagenesis screen.

Materials and Methods

Power analyses

To identify parameters that influenced and maximized power for BSA-seq, we modeled the effects of sequencing depth, phenotypic selection cutoff for choosing bulks, total population size, and generations of growth after meiosis as functions of a causal mutation's effect on mean expression, standard deviation of expression, and fitness. Because of the low density of mutations expected in mutants isolated from mutagenesis screens, we assumed only one causal mutation influenced the phenotype of interest in each mutant. We also assumed that non-causal mutations were in linkage equilibrium with the causal site. Finally, for simplicity, we assumed that non-causal mutations did not

affect fitness. Violating this final assumption should not affect allele frequencies for the causal site as long as it is not linked to these non-causal mutations.

Power analyses were performed in two steps. First, a deterministic model was used to calculate the expected mutant and reference allele frequencies for the causal site in both phenotypically high and phenotypically low bulks prior to sequencing (see Figure D - S1). Then, using these expected frequencies, sampling was used to account for variation introduced by library preparation, sequencing depth, and allele frequency from sequencing. For each set of parameters, we simulated 1000 sets of reference and mutant allele read counts for both the high and low bulks. These modeling and simulation steps were all performed in R (v 2.14.1, R Development Core Team 2013) and are described fully in File S1 with R code provided in File S2.

Strains used for mapping

Haploid mutant strains from Gruber et al. 2012 with trans-regulatory effects on expression of a fluorescent reporter gene in S. cerevisiae were used in this study. These mutants were isolated from a low-dose EMS mutagenesis of a BY4724 (MATa lys2Δ0 ura3Δ0) derivative called YPW1 with a PTDH3-YFP reporter gene inserted on chromosome I at position 199,270 (GRUBER *et al.* 2012). Based on Canavanine resistance assays, each strain was estimated to contain 47 ± 17 (99% confidence interval) EMS-induced point mutations, with exactly 1 mutation expected to affect YFP expression in 98.7% of the strains (GRUBER *et al.* 2012). An RFP marker was inserted at the MATa locus in each of these mutant strains before crossing to the mapping strain described

below to avoid diploid contamination when sorting haploid segregant progeny (FASTER MT approach, Chin et al. 2012). The genetic basis of altered fluorescence was mapped for the YPW89, YPW94 and YPW102 mutants from Gruber et al. (2012), which showed +10%, +7% and -3% changes in mean fluorescence relative to the non-mutagenized reference strain, respectively (Table D-1 and see Figure D - S2A). These mutants also reduced the standard deviation of fluorescence phenotypes for each strain (Table D-1 and see Figure D - S2B). The mapping strain (MATα met17Δ0 ura3Δ0 PTDH3-YFP RME1(ins-308A)) that each of these mutants was crossed to was obtained from a series of crosses involving YPW1 (MATa lys2Δ0 ura3Δ0 PTDH3-YFP), BY4722 (MATα leu2Δ0 ura3Δ0), BY4730 (MATa leu2Δ0 met17Δ0 ura3Δ0) and a YAD373 derivative (MATa leu2Δ0 ura3Δ0 RME1(ins-308A) TAO3(E1493Q) MKT1(D30G)) from Deutschbauer and Davis (2005). The dominant RME1(ins-308A) allele increased sporulation frequency of heterozygous diploids relative to the starting strain, which facilitated isolating large numbers of spores.

Obtaining bulk segregant populations

For each mutant, the segregating pools were obtained as follows. First, the mutant strain was crossed to the mapping strain on YPD plate and diploid colonies were isolated on SC-Lys-Met medium. A single diploid colony was then inoculated to 2 ml GNA (5% D-glucose, 3% Difco nutrient broth, 1% Yeast Extract) and grown to saturation at 30°C. 0.2 ml of this culture was diluted into 1.8 ml of GNA and grown 4 hours to log-phase. Next, cells were washed twice in 1 ml H2O, resuspended in 30 µl H2O and spotted on KAc plate to induce sporulation. Sporulation plates were incubated at room temperature

without Parafilm sealing to allow oxygenation. Sufficient sporulation (> 5%) was usually observed after 4 days, at which point random spores were isolated.

For each strain, the whole yeast spot (about 5x107 cells and tetrads) was resuspended in 1 ml H2O in a microcentrifuge tube, washed once in 1 ml H2O and incubated in 200 µl zymolyase 20T (1 mg/ml) for 50 min on a rotor at room temperature. Once ascus walls were digested, samples were washed in 1 ml H2O and resuspended in 100 µl H2O. In order to enrich for spores relative to vegetative cells, each tube was vortexed vigorously for 2 minutes, which resulted in spores visibly sticking to the tube wall; diploid cells remained in suspension. The supernatant was then carefully aspirated off and the spores were washed once more with 1 ml H2O. To release spores from tube walls, 1 ml Triton-X (0.02%) was added to the empty tubes on ice and samples were briefly sonicated at low power (10 seconds at power 3.5 using a Sonic Dismembrator Model 100 from Fisher Scientific). Spore suspensions with non-aggregated cells were observed under a microscope after this step, with less than 5% vegetative contaminants (diploid carryover). Spores suspension needed to be grown to log-phase to express the fluorescence phenotype. To avoid mating during growth, approximately 3x105 MATα spores were sorted using a BD FACSAria II based on the absence of RFP expression. These cells were then centrifuged, resuspended in 2 ml of YPD (to clear off traces of Triton-X) and incubated at 30°C for 14 h.

Next, log-phase cultures (≈2x107 cells/ml) were washed in water and resuspended in 3 ml SC-Arg (media that has lower autofluorescence than YPD). Cells were acclimated to

their medium for 3 h at 30°C before sorting through the FACSAria II instrument. Cytometric gating was set up using FACSDiva software in order to sort 2x105 cells from both tails (<2-5 and >95-98 percentiles) of the YFP distribution in two separate tubes. Only events of intermediate size were sorted (10% tails of FSC.A distribution were discarded) and special care was made to keep the FSC.A median (a proxy for median cell size) the same in the two bulks of cells. Finally, low fluorescent and high fluorescent bulks were resuspended in 3 ml of YPD, grown to saturation at 30°C and genomic DNA was extracted using a Gentra Puregene Yeast/Bact. Kit from QIAGEN. From this, genomic DNA libraries were prepared using a modified version of a previously described approach (ROHLAND and REICH 2012) as explained in File S1. 100 bp paired-end sequencing was performed on Illumina HiSeq2000 platform at the University of Michigan Sequencing Core Facility.

Analysis of Illumina sequencing data

For each sample, FASTQ files containing all paired-end reads data were generated with CASAVA v1.8.2 software. Prior to alignment, low quality ends were trimmed from reads using sickle v1.2 (https://github.com/najoshi/sickle) with default settings (-q 20 –l 20). Trimmed reads were then aligned to the S288c reference genome (http://www.yeastgenome.org, R64 release from 03-Feb-2011) with PTDH3-YFP inserted on chromosome I using Bowtie2 v2.1.0 (bowtie2 -p 2 -x ref.fasta -1 SampleX.R1.fastq -2 SampleX.R2.fastq -I 0 -X 900 -S SampleX.sam –t; [Langmead and Salzberg 2012]). Next, bamUtil v1.0.9 was used to clip overlaps between mate reads that could bias our estimation of allele frequencies (bam clipOverlap --in SampleX.sam --out

SampleX.clipped.sam --stats --readName). MPILEUP files containing base calls from overlapping reads at each genomic position were generated with SAMtools v0.1.19 (samtools view –q 10 –bS; samtools sort; samtools mpileup –BD –f; Li et al. 2009).

For each mutant, two different MPILEUP files were generated: one was used to call a set of high confidence single nucleotide polymorphisms (SNPs) using VarScan v2.3.6 (Koboldt et al. 2009, 2012, http://varscan.sourceforge.net) and the other was used to estimate allele frequencies for a broader set of genomic positions using Popoolation2 v1.201 (KOFLER *et al.* 2011). The first MPILEUP file was obtained from a BAM file containing sequencing data for the mapping strain and another BAM file merging reads for the low and high fluorescence bulks. SNPs were called using VarScan somatic and somaticFilter commands with the mapping strain considered as "normal" and the merged F1 segregant bulks as "tumor" (somatic --min-coverage 5 --min-var-freq 0.01; somaticFilter --min-coverage 5 --min-reads2 3 --min-strands2 2 --min-var-freq 0.05). Filtering out sites with strong strand bias was critical to remove false positive variants. In parallel, the second MPILEUP file was generated from two separate BAM files containing reads data for the low and high bulks. Allele frequencies at variable sites were computed using Popoolation2 (mpileup2sync.jar --min-qual 20 --threads 2; snp-frequency-diff.pl --min-count 10 --min-coverage 10 --max-coverage 500).

Finally, G-tests were computed for each variable sites using likelihood.test() function from the R package Deducer (FELLOWS 2012). A large fraction of variable sites identified through Popoolation2 were absent from the set of high confidence SNPs obtained with

VarScan and were removed before plotting P-values of G-tests. Mutant alleles at these sites were considered as mapping errors as they usually occured at low frequency, at the end of reads or only in one strand direction.

Single site mutagenesis

Targeted mutagenesis was performed using the delitto perfetto approach (STORICI and RESNICK 2006) to introduce each candidate mutation into the genetic background of the mutant ancestor (BY4724 with PTDH3-YFP inserted on chromosome I). The CORE-UK cassette (COunterselectable REporter KlURA3 and kanMX4) was first inserted at the candidate mutation position in each target gene (SSN2, TUP1 or ROX1) by homologous recombination. Then, coding sequence harboring the candidate mutation was PCR-amplified from the mutant strain (YPW89, YPW94 or YPW102) and introduced in place of the CORE-UK cassette. Sanger sequencing of the target gene confirmed allele replacement.

Expression level of fluorescent reporter for PTDH3 activity: Expression level of the YFP fluorescent reporter protein was quantified in the wild type, EMS-induced mutants and single-site mutants using flow cytometry. Eight replicates of each strain and the non-fluorescent BY4724 were arrayed at random positions in a 96-well format on YPG agar rectangular plates (OmniTrays). In addition, 20 replicates of the wild-type fluorescent strain were arrayed at specific positions to control for plate position effects. After growth on YPG, arrayed strains were transferred into 0.5 ml of YPD in a 96 deep-well plate using a V&P Scientific Pin Tool and grown for 20 h at 30°C to saturation. Cells were

maintained in suspension by the addition of a 3 mm glass bead in each well and constant shaking at 220 rpm. Immediately before flow cytometry, 20 µl of each culture was diluted into 0.5 ml of SC-Arg in another 96-well plate. Fluorescence was then quantified for an average of 104 events per sample using a HyperCyt Autosampler (IntelliCyt Corp.) coupled to a BD Accuri C6 Flow Cytometer (533/30nm optical filter used for YFP acquisition).

Flow cytometry data were analyzed with custom R scripts. First, a set of cytometric events considered as single fluorescing cells was filtered for each sample using Bioconductor flowCore and flowClust packages. An average of 5x103 events per sample were retained after this step. Next, a fluorescence phenotype was calculated for each single event corresponding to log(FL1.A)2/log(FSC.A)3, which corrected for the correlation between fluorescence level and cell size. FL1.A and FSC.A are the area of the YFP fluorescence signal and forward scatter signal (proxy for cell size), respectively. The phenotype of a given sample corresponds to the median phenotype of all filtered events. Finally, we tested for plate position effects by fitting a linear model to the fluorescence data obtained from the 20 control samples. We included in this model the effects of each plate, row, column and half-plate (plates were scored one half at a time). A stepwise approach based on Akaike information criterion for model selection was conducted using the step() function in R and showed that a simple model including only the half-plate effect explained 74.9% of the fluorescence variation across the 20 control samples. Therefore, the effect of each half-plate was extracted from the linear model and subtracted from all samples occurring on the same half-plate. To calculate mean

expression relative to wild type, the fluorescence phenotype of the reference strain YPW1 was subtracted from the fluorescence phenotype of each tested strain and this was divided by the difference in fluorescence phenotype between YPW1 and the non-fluorescent control BY4724.

Fitness assay

Competition experiments were performed to estimate fitness of the parental strain YPW1, the EMS mutants YPW89, YPW94 and YPW102 and the 3 single site mutants with mutation in SSN2, TUP1 and ROX1. Fitness was also measured for five mutants randomly chosen from the 179 trans-regulatory mutants described in Gruber et al. (2012) to estimate the range of selection coefficients for this set of mutants (see Figure D - S2). In all cases, fitness was measured relative to a common reference strain expressing GFP (MATa lys2Δ0 ura3Δ0 PTDH3-GFP) as follows. Each strain was grown for 24 h to saturation in 5 ml of YPD medium, and then cultures were diluted to 6x106 cells/ml based on optical density measurement. 500 µl of each YFP yeast culture was mixed thoroughly with 500 µl of GFP culture and 9 ml of YPD. Next, 500 µl samples of each mix were randomly arrayed in 8 wells of a first 96 deep-well plate and 10 µl samples were diluted to 104 cells/ml into the same 8 random wells of a second 96 deep-well plate containing 490µl of YPD per well. The first plate was used to estimate the proportion of YFP and GFP cells at the beginning of the competition assay (T0) using flow cytometry. The second plate was grown for 24 h at 30°C with constant shaking (220 rpm with glass beads to keep cells in suspension). 20 µl of each culture was diluted in 500 µl of YPD in a clean 96-well plate and the proportion of YFP and GFP cells was estimated at the end

of the competition assay (T1) using flow cytometry. Fluorescence was recorded for at least 2x104 events per sample using a HyperCyt Autosampler (IntelliCyt Corp.) coupled to a BD Accuri C6 Flow Cytometer (585/40 nm optical filter used for YFP acquisition and 533/30 nm optical filter used for GFP acquisition). Despite considerable overlap of the YFP and GFP signal detected through the 533/30 nm filter, control experiments showed that cells expressing YFP or GFP could be distinguished using this filter combination. Custom R scripts were used to filter out spurious events from flow data and to compute the proportion of YFP and GFP cells for each sample. A selection coefficient was then calculated for each replicate using the following formula:

$$s = \frac{\left(\ln\left(\frac{YFP_1}{GFP_1}\right) - \ln\left(\frac{YFP_0}{GFP_0}\right)\right)}{g},$$

where YFP_0 and GFP_0 are the observed number of cells expressing YFP and GFP at time T0 and YFP_1 and GFP_1 are analogous numbers of cells at time T1 for each sample. The experiment was started from an average density of 104 cells/ml and stopped at 6x107 cells/ml, yielding an approximate number of generations g = 12.55. The selection coefficient obtained by competing mutants expressing YFP against a GFP reference strain can be explained by the mutant genetic background or by YFP expression itself. The fitness effect of the YFP marker was quantified by competing the parental strain YPW1 to the GFP reference strain. From this, the selection coefficient associated with the mutant background was calculated as follows, assuming the effect of EMS-induced mutations and YFP on fitness were additive:

$$s_{mut} = \frac{s+1}{s_{YFP}+1} - 1,$$

where s_{mut} is the selection coefficient for the mutant background, s_{YFP} is the selection coefficient for the YPW1 reference strain expressing YFP measured in competition to the GFP strain and s is the selection coefficient for the mutant strain expressing YFP measured in competition with the same GFP reference strain.

Results

Optimizing experimental design for mapping mutations of small effect

To determine how power for detecting different types of mutations varies with mutational properties and experimental parameters, we developed a flexible simulation that models mapping via BSA-seq computationally. This power analysis was parameterized for mapping mutations affecting fluorescence in S. cerevisiae that can be efficiently scored in large populations of recombinant cells using FACS, but can be adapted to other biological systems with different genome sizes, mutation effects or attainable population sizes. Variant discovery and allele-frequency estimation were modeled assuming whole genome sequencing of DNA extracted from bulks of recombinants with high- and low-fluorescence phenotypes. We included two phases of cell growth during which competition among genotypes can affect allele frequencies. The first growth phase between meiosis and bulk selection was required to express the fluorescence phenotype and the second between bulk selection and DNA extraction to increase the amount of genomic DNA for sequencing (Figure D-1A). Experimentally controllable parameters included in our simulation were population size, intensity of phenotypic selection, sequencing depth, and number of generations between spore isolation and DNA extraction. Innate biological properties of a mutation included were

the effects on the mean and standard deviation of the fluorescence phenotype as well as the effect on fitness, which changes the frequency of the mutation during growth. We used the range of biological properties observed for trans-regulatory mutants affecting fluorescence of a reporter gene isolated in Gruber et al. 2012 (see Figure D - S2) to identify experimental parameters that provide a high probability of detecting a causal mutation for a minimal cost.

We found that the size of the population from which the bulks were selected did not have a large impact on power as long as the population size was at least an order of magnitude higher than sequencing depth (Figure D-1). When segregant pools were smaller than sequencing depth, a high rate of false positives was observed (see Figure D - S3). For the range of biological parameters observed among mutants isolated in Gruber et al. (2012) (shaded areas in Figure D-1), we found that 20 generations or less of mitotic growth provided sufficient power to detect most causal mutations (Figure D-1E-G). Increasing this number to 50 generations allowed competition among genotypes to strongly bias allele frequencies, causing a loss of power to detect mutations with selection coefficients greater than 0.05 (Figure D-1G). Our analyses also suggested that selecting cells from the 10% or smaller tails of the fluorescence distribution is sufficient to achieve high detection power under most conditions. Generally, sampling cells from more extreme tails by decreasing the fluorescence cutoff for selection increased power (Figure D-1H-J); however, mutations causing very large increase or decrease in the standard deviation of the fluorescence phenotype were found to require a particular range of cutoff percentages to maximize power (Figure D-1I).

Unlike in our simulations, population size, generations of growth and the fluorescence cutoff for bulk selection are inter-dependent in a real experiment. Starting with a larger population size, for instance, decreases the number of generations of growth required to obtain sufficient cells for analysis, which in turn decreases the impact of the selection coefficient of the mutation on mapping success. However, increasing the population size requires more time and money spent phenotyping individuals prior to bulk selection. Considering the ease of creating large populations of yeast and the high-throughput phenotyping possible when using FACS to measure fluorescence, we decided to fix the population size at 107 cells and the cutoff for bulk selection at 2% for the rest of this study, resulting in 2x105 cells in each bulk. This allowed us to keep the total number of generations to 20, with 10 generations of growth between spore isolation and bulk selection, and 10 generations between bulk selection and DNA extraction.

Using these conditions, we investigated the impact of sequencing depth on mapping power. As expected, increasing the average genome coverage always improved the power to detect causal variants (Figure D-1K-M), but this requires increased cost. Selecting the ideal sequencing coverage therefore depends on the properties of the mutation(s) that the researcher seeks to identify. For mutations similar to those isolated in Gruber et al. (2012), we deemed 100x genome coverage the best compromise between power and cost. We note, however, that mutations with a mean effect of 5% or larger can be reliably detected with sequencing coverage as low as 25x (Figure D-1K) as long as the mutation does not have a large impact on the phenotypic standard deviation (Figure D-1L) or

relative fitness (Figure D-1M). Mutations with mean phenotypic effects smaller than 2% or with selection coefficients larger than 0.15 would likely require >100x sequencing coverage in each bulk to be mapped.

With the population size, generations of growth, intensity of phenotypic selection, and sequencing coverage fixed as described above, we used our simulation to further investigate the relationship between power to detect a causative mutation and that mutation's effects on mean expression (Figure D-2A), standard deviation of expression (Figure D-2E) and relative fitness (Figure D-2I). We found that mutations that change the mean phenotype at least 3%, cause a phenotypic standard deviation ranging from 75% to 150% of the wild type, and have a selection coefficient less than 0.1 should be detected with a power higher than 90% with this experimental design (Figure D-2). This combination of effects on mean and standard deviation includes >90% of the 179 trans-acting mutants isolated by Gruber et al. 2012 (see Figure D - S2A,B). It also includes 5 of the 8 mutants for which we measured relative fitness (see Figure D - S2C). Fitness measurements for mutants isolated from a mutagenesis screen are expected to overestimate the fitness effects of a causative mutation, however, because mutations that do not affect the phenotype of interest can also affect fitness. Therefore, using the relative fitness of a mutant strain to determine the best experimental design is expected to underestimate the true power for mapping the causative mutation in that strain. Increasing the mean effect of a mutation always improved detection power (Figure D-2B,C,D,G), but a more complex relationship was observed between the effects of a mutation on the phenotypic standard deviation and selection coefficient. Specifically, when a mutation

had a large effect on the standard deviation, more deleterious mutations could have greater detection power than less deleterious mutations under some conditions (Figure D-2F,H). This is because these parameters bias mutation frequency in the two bulks in opposite directions: increasing the selection coefficient lowers the mutation frequency in both bulks, while increasing the standard deviation raises it.

Identifying single candidate mutations in trans-acting mutants

To empirically evaluate the BSA-seq approach using parameters selected based on the simulations described above, we attempted to map mutations responsible for altered fluorescence in three trans-regulatory mutants carrying a YFP reporter protein under the control of the S. cerevisiae TDH3 promoter. Assuming that a single causative mutation explains the phenotypic effects observed in each mutant (Table 1), a mapping power >97% is expected for bulks consisting of 2x105 cells sorted from the 2% tails of the fluorescence distribution, with 20 generations of growth and an average sequencing coverage of at least 75x (see Figure D - S4).

To efficiently obtain such large and stringently selected bulks of pure haploid segregants, we followed the protocol shown in Figure D-3. Each haploid mutant strain was mated to a common mapping strain and sporulation (meiosis) was induced in the resulting diploids. Including the RME1(ins-308A) allele (DEUTSCHBAUER and DAVIS 2005) in the mapping strain increased sporulation frequency from 2% to 20%, making it easier to isolate a large population of F1 haploid segregants. To prevent mating between MATa and MATα haploids in the population of segregants, we sorted ~3x105 cells lacking

expression of a red fluorescent protein (RFP) reporter gene that we had inserted at the mata-2 locus in each MATa parent. More than 99.6% of cells lacking expression of this FASTER MT marker (CHIN *et al.* 2012) were confirmed to be MATα haploids. After 10 generations of growth to allow these MATα cells to robustly express their YFP fluorescence phenotype, high- and low-fluorescence bulks of ~2x105 cells each were isolated via FACS. Attention was paid during cell sorting to avoid introducing other phenotypic variation between the two bulks. For instance, no more than 1% variation in median cell size (FSC) was allowed between bulks (see Figure D - S5). After an additional 10 generations of growth, genomic DNA from the low- and high-fluorescing bulks was sequenced to at least 75x coverage (Table 2) using 100 bp paired-end Illumina sequencing.

Sequencing data were analyzed using a pipeline with two main parallel steps (Figure D-4). First, a set of high confidence variants was called for each mutant with VarScan (KOBOLDT *et al.* 2009, 2012) based on genome sequencing data from the mapping strain and genomes of the segregant bulks, with reads from both bulks merged. The number of mutations identified in each mutant ranged from 33 to 77 (Table 2), which closely matches the number of mutations (30-64) predicted using Canavanine-resistance mutation rates in Gruber et al. (2012). Most of these mutations (85-94%) were G:C to A:T transitions, as expected for EMS-induced mutations (Table 2). Allele frequencies were then estimated at every variable site for each bulk using Popoolation2 (KOFLER *et al.* 2011). These allele frequencies were strongly correlated (r = 0.983) with independent estimates determined by pyrosequencing (see Figure D - S6 and File S3). The number of

reads containing reference and mutant alleles at high confidence polymorphic sites were compared between low and high fluorescence bulks using a two-sided G-test.

For each mutant, we observed a single, highly significant ($P<0.001$) association with YFP fluorescence level (Figure D-5); physically linked sites also showed significant associations with comparatively lower P-values (Figure D-5). To determine the likelihood that similar associations would have been detected for other mutants, we examined the distribution of aligned sequence reads in more detail (see File S4). We found that ~3% of the genome had little-to-no sequencing coverage in each mutant (see Figure D - S9) and that this was due to difficulty obtaining and/or aligning sequence reads from these regions rather than stochastic fluctuations in coverage due to sampling (see Figure D - S10A,B). We also found that reducing sequencing depth would have caused us to miss the significantly associated site in YPW89 -- the mutant with the largest effect on mean fluorescence of the PTDH3-YFP reporter gene -- because of its strong deleterious effect on fitness (see Figure D - S10C,D).

The sites with the strongest statistical associations in YPW89 (Figure 5A), YPW94 (Figure D-5B), and YPW102 (Figure D-5C) were non-synonymous substitutions affecting the SSN2, TUP1 and ROX1 genes located on chromosomes IV, III and XVI, respectively (Table 3). These three mutations are all coding substitutions, which is not surprising given that open reading frames constitute 73% of S. cerevisiae genome (Saccharomyces Genome Database). TUP1(G696D) and ROX1(R12K) are both missense mutations that change one amino acid, while SSN2(Q971*) introduces an early stop

codon truncating 450 amino acids of the protein. All three mutations affect amino acids that are highly conserved across Saccharomyces species, with TUP1(G696D) and ROX1(R12K) substitutions predicted to be deleterious using SIFT (SIFT score = 0 for both; Kumar et al. 2009). The fitness estimates of each substitution are similar to the fitness consequences reported for deletion alleles of these genes (W = 0.859, 0.954 and 0.983 for single site substitutions in SSN2, TUP1 and ROX1; W = 0.896, 0.921 and 0.971 for deletions of SSN2, TUP1 and ROX1 [Breslow et al. 2008; Deutschbauer et al. 2005]), suggesting that the mutations observed severely impair the function of the corresponding proteins. Interestingly, TUP1 appears to be a direct regulator of TDH3 expression: Tup1 protein is a general transcriptional repressor that was shown to directly bind TDH3 promoter in ChIP-chip experiments (HANLON *et al.* 2011). Rox1 is a repressor of hypoxic genes that might indirectly affect TDH3 expression through the regulation of Pdr1 transcription factor (HARBISON *et al.* 2004; LAROCHELLE *et al.* 2006). The Tup1-Ssn6 complex is also a well-established regulator of ROX1 expression (MENNELLA *et al.* 2003), suggesting that Tup1 acts at multiple levels of the regulatory network controlling TDH3 expression. Finally, SSN2 encodes a facultative subunit of the RNA polymerase II holoenzyme (SONG *et al.* 1996), which could potentially act on several components of the TDH3 regulatory network.

Validating bulk segregant mapping results

In parallel to the method described above, we used a more traditional mapping approach involving tetrad dissection (BIRKELAND *et al.* 2010) to analyze the YPW89, YPW94 and YPW102 mutants. Each mutant was crossed to a common mapping strain (see File S1);

the resulting diploids were sporulated; and a dozen tetrads were dissected. For each tetrad, the fluorescence phenotype of each spore was determined by measuring mean fluorescence of their haploid mitotic progeny using flow cytometry, and the two colonies most likely to carry the causative mutation were identified manually (see Figure D - S7). Segregant progeny deemed to show the mutant phenotype were pooled together, genomic DNA from this pool was extracted and subjected to Illumina sequencing, and allele frequencies were estimated for each variable site. G-tests were used to compare the observed mutation frequency at variable sites to a null model with a frequency of 0.5.

For YPW89, YPW94, and YPW102, the best candidate mutations identified with the BSA-seq approach were also highly significant with the tetrad dissection method (see Figure D - S8). Compared to the mass sporulation and BSA, however, tetrad dissection was tedious and less amenable to the analysis of a large number of mutants. The tetrad approach was also very sensitive to errors in phenotype assignment, which can be caused by environmental variation or stochastic noise. Indeed, several mutants failed to yield a significant candidate site when using this approach (data not shown). For such mutants, some tetrads showed a clear 2:2 segregation of the fluorescence phenotype (see Figure D - S7A), while others were harder to characterize (see Figure D - S7B). This might be explained by the fact that each of the four sister spores was grown in a separate vial prior to phenotyping with the tetrad dissection approach, but all spores were grown together in the same vial for BSA-seq, minimizing the influence of micro-environmental factors.

We also tested directly whether the variants identified by BSA-seq in YPW89, YPW94, and YPW102 were responsible for their mutant phenotypes by using site-directed mutagenesis to introduce each candidate mutation individually into the genetic background carrying the PTDH3-YFP transgene that was originally used for the EMS mutagenesis screen. In all three cases, the single site mutation completely recapitulated the fluorescence phenotype of the EMS-mutant from which it was identified (Figure D-6). This result, combined with the absence of any other significant mutation unlinked to these causative sites (Figure D-5), shows that each original mutant carried exactly one causative mutation and that this mutation could be unambiguously identified using BSA-seq despite its small phenotypic effect.

Discussion

Using both simulated and empirical data, we describe the impact of innate properties of genetic variants and controllable experimental factors on the success of mapping single nucleotide variants using a bulk-segregant analysis with high-throughput sequencing after a mutagenesis screen. We show how mapping success is affected by a mutation's effect on the mean phenotype as well as its effects on phenotypic variance and fitness. By using simulations to determine optimal experimental conditions and new genetic tools to efficiently isolate large pools of informative segregants, we demonstrated the efficiency of the approach by identifying mutations in SSN2, TUP1 and ROX1 with small effects on the expression of the PTDH3-YFP reporter gene in S. cerevisiae. Below, we discuss (1) how the effects of a mutation on fitness affect mapping using BSA-seq; (2) how our

findings can be applied to mapping other traits in other organisms; and (3) how our conclusions and methods can be used to study QTLs underlying natural variation.

Impact of fitness on mapping success

Prior statistical models of BSA-seq focused on the effect of segregating sites on the mean phenotype (MAGWENE *et al.* 2011; EDWARDS and GIFFORD 2012), but their effects on the standard deviation and relative fitness can also impact mapping success when using pooled segregant approaches in S. cerevisiae (WILKENING *et al.* 2013). Fitness effects could be an especially important source of discovery bias in BSA-seq data given the high proportion of random mutations showing detrimental effects on growth (WLOCH *et al.* 2001; EYRE-WALKER and KEIGHTLEY 2007). We examined this issue computationally and found that starting with a large population of spores and selecting large pools of segregants (2x105) was essential for achieving high mapping power when a mutation affected both the phenotype of interest and relative fitness. For example, our model predicted that selection of only 103 segregants would be sufficient to map a mutation with a 3% effect on the mean and no effect on fitness, yet reducing the size of the segregant bulk in this study from 2x105 to 103 would have caused the SSN2(Q971*) mutation with a >10% effect on the mean to remain undetected because of its deleterious fitness effects. Because deleterious alleles tend to be purged from both bulks during growth, the power to detect a significant difference in allele frequencies between the two bulks is decreased. One solution to reduce the impact of fitness on mapping success would be to decrease the generations of growth after meiosis, but this is not always possible. For example, in our case, growth was needed for the cells to express the

fluorescence phenotype as well as to increase the amount of genomic DNA available for sequencing. This latter growth phase could be shortened by using a protocol for preparing DNA libraries that requires less genomic DNA, but this usually increases noise and cost.

In the absence of fitness effects and competitive growth, alternative alleles for a site not affecting the trait of interest should be found in 50% of segregants in each bulk. Comparing the allele frequency in a single segregant bulk after phenotypic selection to a null frequency of 0.5 to detect causative mutations should be avoided, however, because any effects of a genetic variant on fitness can cause allele frequencies to deviate from this null model, increasing false positives. Rather, allele frequencies should be compared between bulks from the extremities of the phenotypic distribution. If only one tail of the distribution is amenable to phenotypic selection, for instance when selecting for drug resistance, cells that have not been subjected to phenotypic sorting but have otherwise undergone the same experimental steps as the segregant bulks should be used to define the null model (Ehrenreich et al. 2010; Parts et al. 2011).

Applications for other traits and organisms

BSA is a powerful approach to mapping for species and traits where large numbers of recombinant offspring can be analyzed and individuals with extreme phenotypes can be efficiently isolated. Selected bulks can now be genotyped en masse by high-throughput sequencing whenever a reference genome is available or can be obtained. When these conditions are met, BSA-seq can quickly identify mutations causing a mutant phenotype, even when the phenotypic effect of a mutation is very small. Our data show that the

experimental design needed to most reliably and cost-effectively identify such mutations is different in each case. We encourage researchers to use the simulations and statistical models described in this study to identify experimental parameters that will maximize their own mapping success by tuning the parameters in the model to their specific system. These parameters include not only a mutation's effect on the mean phenotype and fitness, but also its effect on the standard deviation of the phenotype. For example, we found that when the mutant phenotype has a standard deviation much larger or much smaller than the wild-type phenotype, mapping power decreases quickly with a wide range of BSA-seq experimental designs. Under these conditions, sequencing individuals from the two symmetric tails of the phenotypic distribution is not recommended and an alternative approach should be considered, such as selection of asymmetric bulks. Analysis of larger bulks should also help increase power in these cases.

When extrapolating our findings to mapping other species and traits, it is important to consider that we modeled a BSA-seq experiment in yeast including population growth between meiosis and phenotyping as well as between phenotyping and DNA sequencing. This competitive growth is not necessary when using BSA-seq in multicellular organisms such as fruit flies or nematodes. Therefore, mapping power should be much less affected by mutations that impact reproductive fitness, allowing smaller population sizes to be used. Still, the minimum effect size that can be mapped in these types of organisms will usually be larger than the minimum effect size that can be mapped in yeast, both because of the increased genome size and because of the smaller attainable bulk size. For example, if a single causative mutation is segregating in an F2 population, our model

predicts that the power of BSA-seq to detect a mutation changing the phenotype by 5% relative to wild type is greater than 0.9 for a total population size of 104, a 5% cutoff for phenotypic selection, and an average sequencing coverage of 25x. Although this is not a simple task, these parameter values can be achieved in D. melanogaster and C. elegans, respectively, using tools such as fly cages to raise large populations of flies or worm sorters to automate phenotypic scoring and selection.

Mapping phenotypic variation in natural populations

BSA-seq has been shown to be a powerful approach for mapping small effect QTLs underlying natural variation in S. cerevisiae (ALBERT *et al.* 2014). Compared to the mutants characterized in our study, strains used for QTL mapping typically have more segregating sites and more causative loci. The large number of sites segregating in these strains leads to many linked polymorphisms, which reduces mapping resolution, but can improve the power to detect small effect QTLs (MAGWENE *et al.* 2011; EDWARDS and GIFFORD 2012). However, the presence of multiple QTLs acting in the same direction can decrease the power to detect a polymorphism of small effect compared to the case where it segregates alone (YANG *et al.* 2013). Our work suggests that the effects of QTLs on phenotypic noise and/or fitness should also be considered in future statistical models of QTL mapping via BSA-seq to avoid discovery biases.

The three genetic tools we used to increase the sensitivity of BSA-seq for finding novel mutations can also be used to study natural variation in yeast. Specifically, the dominant RME1(ins-308A) allele that we inserted into our mapping strain to increase meiosis rate (DEUTSCHBAUER and DAVIS 2005) can also be incorporated into other strains, allowing

for the efficient recovery of large numbers of segregants. This is important because many strains of S. cerevisiae, including the commonly used S288c lab strain and its derivatives as well as some wild isolates, have low sporulation rates that limit the efficiency of BSA-seq (GERKE *et al.* 2009). The FASTER MT cassette (CHIN *et al.* 2012) can also be inserted into other genotypes to allow for robust and efficient recovery of MATα spores, preventing mating among F1 segregants. Compared to the Yeast Magic Marker (TONG *et al.* 2001; PAN *et al.* 2004) used for a similar purpose in previous BSA studies (EHRENREICH *et al.* 2010; WILKENING *et al.* 2013; ALBERT *et al.* 2014), FASTER MT requires less genetic manipulation of the mapping strain(s), reduces biases caused by diploids in the sorted haploid cultures (GERSTEIN and OTTO 2011; WILKENING *et al.* 2013), and limits the impact of competitive growth on mapping power by allowing MATα cells to be sorted immediately after spore isolation. Finally, if the phenotype of interest can be coupled to a fluorescent reporter gene, FACS can be used for high throughput phenotyping and selection. Other easy-to-score phenotypes should also be well suited for genetic mapping using BSA-seq.

In conclusion, this study provides a methodological framework for efficiently mapping genetic variants with small effects, illustrates the importance of considering the fitness effects of causative variants when using BSA-seq in microorganisms such as yeast, and describes the use of experimental tools that can reduce the bias against detection of variants with small effects by allowing very large populations of phenotypically divergent individuals to be collected and analyzed.

Extended Materials and Methods

Power analyses

Modeling exact allele frequencies in bulks: The goal of the deterministic step of the model was to calculate the frequency of mutant and reference alleles expected in each phenotypically divergent bulk of cells depending on total population size used for sorting (n), phenotypic selection cutoff used for isolating bulks (c), generations of growth after meiosis (g), mutation effect on mean expression (μ), mutation effect on standard deviation in expression (σ), and selection coefficient for the mutation (s). We modeled the total population distribution with respect to expression, X_T, as a mixture distribution of two populations, X_R and X_M, where X_R is the population carrying the reference allele at the causative locus and X_M is the population carrying the mutant allele at the causative locus, and tracked each population separately. Each population was assumed to follow a normal distribution with respect to expression:

(1) $X_R \sim N(0,1)$

(2) $X_M \sim N(\mu, \sigma^2)$

We represent the mean effect of a causal mutation, μ, relative to the standard deviation of the reference strain such that an increase of μ by 1 is equivalent to a shift in mean expression by one standard deviation (an approximately 7.5% change in expression in our data). Mutations were assumed not to influence sporulation efficiency or spore survival and X_R and X_M were started at equal frequencies. Populations were allowed to grow deterministically assuming a selection coefficient for the mutant causative allele of s. The reference allele was assumed to have fitness of 1 and after g generations the frequency of the mutant population in the population was:

(3) $f_{M_W}^0 = \frac{(1-s)^g}{(1-s)^g+1}$

where W indicates the whole mutant or reference population prior to selection of phenotypic bulks (see Figure D - S1 for diagram). The reference allele frequency was then the difference:

(4) $f_{R_W}^0 = 1 - f_{M_W}^0$

After determining the frequencies of the mutant and reference populations, phenotypic selection using flow cytometry was modeled on the total population, X_T, at a predetermined population cutoff, c. The goal was to quantify the frequency of mutant and reference genotypes in each phenotypic bulk. Because X_T is a mixture distribution, the fractions of individuals with mutant and reference alleles present in each bulk were determined from the reference and mutant phenotypic distributions X_R and X_M. For the high bulk, this required determining the quantiles q_{R_H} and q_{M_H} on X_R and X_M such that q_{R_H} and q_{M_H} equaled the same expression value $e_{T_H}$and c percent of the total population had higher expression than e_{T_H}.

(5) $f_{R_W}^0 * q_{R_H} + f_{M_W}^0 * q_{M_H} = 1 - c$

(6) $e_{T_H} = \Phi^{-1}(q_{R_H}) = \mu + \sigma * \Phi^{-1}(q_{M_H})$

Likewise, for the low bulk this required determining quantiles q_{R_L} and q_{M_L} on X_R and X_M such that q_{R_L} and q_{M_L} equaled the same expression value e_{T_L} and c percent of the total population had lower expression than e_{T_L}.

(7) $f_{R_W}^0 * q_{R_L} + f_{M_W}^0 * q_{M_L} = c$

(8) $e_{T_L} = \Phi^{-1}(q_{R_L}) = \mu + \sigma * \Phi^{-1}(q_{M_L})$

In both instances, $\Phi^{-1}(q)$ is the standard normal quantile function, H and L index the high and low bulks respectively, and e_{T_H} and e_{T_L} are the expression values for the high and low bulks relative to the entire population X_T. We solved the above equations numerically for q_{M_H} and q_{M_L} using *solnp* within Rsolnp (Ghalanos & Theussl 2006) by minimizing the following functions for the high and low bulks respectively:

(9) $$[f^0_{R_W} * \Phi\left(\mu + \sigma * \Phi^{-1}(q_{M_H})\right) + f^0_{M_W} * q_{M_H} + c - 1]^2$$

(10) $$[f^0_{R_W} * \Phi\left(\mu + \sigma * \Phi^{-1}(q_{M_L})\right) + f^0_{M_W} * q_{M_L} - c]^2$$

where $\Phi(y)$ is the cumulative distribution function for the standard normal distribution. From these quantiles, the frequencies of the mutant and reference alleles in the high and low bulks were calculated as the weighted proportion of mutant and reference alleles more extreme than the phenotypic cutoff:

(11) $$f^0_{M_H} = \frac{f^0_{M_W} * \Phi\left(\frac{e_{T_H} - \mu}{\sigma}\right)}{f^0_{R_W} * \Phi(e_{T_H}) + f^0_{M_W} * \Phi\left(\frac{e_{T_H} - \mu}{\sigma}\right)} = \frac{f^0_{M_W} * q_{M_H}}{f^0_{R_W} * q_{R_H} + f^0_{M_W} * q_{M_H}} = \frac{f^0_{M_W} * q_{M_H}}{1-c}$$

(12) $$f^0_{R_H} = 1 - f^0_{M_H}$$

(13) $$f^0_{M_L} = \frac{f^0_{M_W} * \Phi\left(\frac{e_{T_L} - \mu}{\sigma}\right)}{f^0_{R_W} * \Phi(e_{T_L}) + f^0_{M_W} * \Phi\left(\frac{e_{T_L} - \mu}{\sigma}\right)} = \frac{f^0_{M_W} * q_{M_L}}{f^0_{R_W} * q_{R_L} + f^0_{M_W} * q_{M_L}} = \frac{f^0_{M_W} * q_{M_L}}{c}$$

(14) $$f^0_{R_L} = 1 - f^0_{M_L}$$

To model the additional growth necessary to create libraries from the sorted bulks, each bulk was allowed to undergo another g generations of growth, assuming that the relative fitness (1-*s*) between genotypes with the mutant and reference alleles of the site affecting fluorescence was the same before and after bulk selection:

(15) $f_{M_H}^1 = \frac{f_{M_H}^0(1-s)^g}{f_{M_H}^0(1-s)^g + f_{R_H}^0 * 1}$

(16) $f_{M_L}^1 = \frac{f_{M_l}^0(1-s)^g}{f_{M_L}^0(1-s)^g + f_{R_L}^0 * 1}$

Simulation of allele frequency estimates from sequencing data: Using the deterministic allele frequencies described above, we simulated the library creation and sequencing processes by drawing the proportion of 'reads' containing the mutant allele from a binomial distribution in each bulk independently:

(17) $T_{M_H} \sim B(V_H, F_{M_H})$

(18) $T_{M_L} \sim B(V_L,\ F_{M_L})$

where V is the distribution of sequencing coverage and F the mutant allele frequency distribution. The sequencing coverage distribution was simulated as a negative binomial distribution (ROBINSON and SMYTH 2007, 2008):

(19) $V \sim NB(\alpha, \beta)$ with mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha(\beta+1)}{\beta^2}$

To adjust coverage, we varied β (inverse scale) because our data suggested α (shape) was approximately 80 regardless of sequencing depth. Average coverage was set to reflect coverage after mapping and we did not explicitly model sequencing error. To account for sampling during library creation, the mutant allele frequencies were simulated from the deterministic frequencies assuming a binomial distribution:

(21) $F_{M_H} \sim \frac{B\left(n, f_{M_H}^1\right)}{n}$

(22) $F_{M_L} \sim \frac{B\left(n, f_{M_L}^1\right)}{n}$

Reference 'reads' were then assumed to make up the difference between the coverage and the number of mutant 'reads'

(23) $t_{R_H} \sim v_H - t_{M_H}$

(24) $t_{R_L} \sim v_L - t_{M_L}$

A G-test was performed on the counts t_{M_H}, t_{M_L}, t_{R_L} and $t_{R_H}$to determine significance. Power was calculated as the frequency of simulations where the *P*-value was below 0.001, representing a Bonferonni correction assuming 50 possible mutations.

Comparison between G-test and Fisher's exact test: The Fisher's exact test commonly used in the analysis of next generation sequencing data (KOFLER *et al.* 2011) assumes that the row and column totals of the two-by-two contingency table are fixed. This assumption is violated by sequencing data, however, because coverage for each allele results from sampling reads from an underlying distribution. When marginal totals are free to vary, the G-test is more appropriate than the Fisher's exact test. We analyzed our data using both tests and found that their results were very similar (although not identical) except when sequencing coverage was low (Figure D - S11).

DNA library preparation

Genomic DNA libraries were produced in parallel by modifying a low cost method developed for Illumina sequencing (ROHLAND and REICH 2012). Briefly, DNA was sheared, Illumina adapters were attached by blunt-end ligation and indexed using PCR. Between enzymatic reactions, DNA was cleaned using custom MagNA beads (Carboxyl-modified Sera-Mag Magnetic Speed-beads in a PEG/NaCl buffer) as a lower-cost

substitute for AMPure XP kit. For each sample, 2 µg of genomic DNA (120 µl) was sheared to an average fragment size of 400 bp with a Covaris S220 instrument (Duty cycle: 10%, intensity: 4, cycles/burst: 200, time: 55 s). 1 µg (60 µl) of sheared DNA was purified in 96 µl (1.6x) of MagNA bead solution and resuspended in 20 µl of water. Blunt-end repair was performed using a NEB Quick Blunting Kit by mixing 19 µl of DNA with 2.5 µl of blunting buffer, 2.5 µl of 1 mM dNTP mix and 1 µl of blunt enzyme mix. This mix was incubated for 20 min at 12°C followed by 15 min at 37°C. DNA was then cleaned up in 2x MagNA beads and eluted in 25 µl of water. Next, adapters were ligated using a NEB Quick Ligation Kit. 23.8 µl of blunt DNA was mixed with 30 µl of ligation buffer, 4 µl of P5 + P7 adapter mix (100 µM each) and 1.2 µl of Quick T4 DNA ligase and incubated at 25°C for 20 min. DNA was then cleaned in 1.6x beads, eluted in 40 µl and nick-fill in was done using Bst DNA Polymerase Large Fragment from NEB. 39 µl of DNA sample was mixed with 5 µl of ThermoPol buffer, 4 µl of 25 mM dNTP mix and 2 µl of Bst DNA polymerase (2 U/µl). After 20 min at 37°C, samples were mixed with 1.6x MagNA beads and eluted in 30 µl water. KAPA HiFi PCR Kit was used for indexing PCR: 10 µl of template DNA was mixed with 5 µl of HiFi buffer, 0.75 µl of 10 mM dNTP mix, 0.75 µl primer IS4 (10 µM), 0.75 µl indexing primer (10 µM), 7.25 µl sterile H_2O and 0.5 µl KAPA HiFi polymerase (1 U/µl). PCRs were incubated at 95°C for 4 min followed by 12 cycles at 98°C for 20 s, 64°C for 15 s and 72°C for 20 s with a final extension at 72°C for 5 min. PCR products were then cleaned up in 1.6x MagNA beads and eluted in 40 µl of water. Samples were then processed at the UM Sequencing Core Facility. For each sample, DNA concentration was quantified through qPCR with primers targeting P5 and P7 adapters and using an Agilent 2100 Bioanalyzer. Equimolar amounts

of each sample were pooled together for multiplexed sequencing before gel electrophoresis size selection of DNA fragments ranging from 350 bp to 850 bp on a 1% agarose gel. The 8 libraries produced for this project (high- and low-fluorescing bulks for each of the three mutants plus the original non-mutagenized strain and the mapping strain, all of which were haploid) were combined with 16 libraries constructed for other projects and subjected to 100 bp paired-end sequencing in one lane on Illumina HiSeq2000 platform. Oligonucleotide sequences used for library preparation are listed in Table S1 and barcode sequences used for multiplexing in Table S2. Because average sequencing depth was lower than 75x for two of the samples (YPW89 low bulk and YPW102 low bulk), we decided to re-sequence the corresponding genomic libraries in an independent sequencing lane using the same procedure. All data from the two runs of sequencing were combined for analyses presented in this study.

Tetrad dissection-based approach for mapping

In addition to the high-sensitivity method described above, we mapped the causative mutation altering YFP expression in several mutants including YPW89, YPW94 and YPW102 using a tetrad dissection-based approach (BIRKELAND *et al.* 2010). First, mutants YPW89 and YPW94 were crossed to Y39 (*MATα leu2Δ0 ura3Δ0* P_{TDH3}*-YFP*) and YPW102 was crossed to Y85 (*MATα met17Δ0 ura3Δ0* P_{TDH3}*-YFP*). Resulting diploids were sporulated in KAc medium, several tetrads were dissected and individual spores were grown on YPD (11 tetrads for YPW89xY39, 8 tetrads for YPW94xY39 and 9 tetrads for YPW102xY85). The fluorescence level of the resulting colonies was quantified through flow cytometry. Each spore was grown in YPD to saturation, then

diluted in SC-Arg medium and grown to log-phase at 30°C. Fluorescence (FL1-A) and forward scatter (FSC-A) of thousands of cells were recorded using a HyperCyt Autosampler (IntelliCyt Corp.) coupled to a BD Accuri C6 Flow Cytometer (533/30 nm optical filter used for YFP acquisition). Based on these data, a mutant phenotype was assigned for 2 of the 4 spore progeny from each tetrad. For tetrads derived from YPW89 and YPW94 (increased YFP expression), the two progeny with highest median of FL1-A/FSC-A were considered as mutants. For tetrads derived from YPW102 (decreased YFP expression), the two progeny with lowest median of FL1-A/FSC-A were considered as mutants. These mutant progeny were then cultured separately to saturation in YPD and mixed evenly to a final volume of 2.5 ml. 22 progeny were mixed together for YPW89, 16 for YPW94 and 18 for YPW102. For each pool, genomic DNA was extracted using a Gentra Puregene Yeast/Bacteria Kit from QIAGEN. Next, 2 μg of DNA was sheared with a Covaris S220 instrument and genomic libraries were prepared using NEBNext E6040 kit. An in-line barcoding strategy was adopted for multiplexing. Briefly, 3' A overhang was added to end-repaired DNA fragments. Then, barcoded adapters were ligated to dA-tailed DNA, creating Y-shaped products whose extremities are single-stranded. PCR using standard Illumina primers allowed the addition of adapter sequences attaching to Illumina flow cells. PCR products ranging from 400bp to 800bp were size selected on an agarose gel. Barcodes, adapters and PCR primer sequences are listed in Table S3 and Table S4. 22 libraries were pooled together and 100 bp paired-end reads were sequenced on a single lane of HiSeq2000 flow cell at the University of Michigan Sequencing Core. Sequencing data were analyzed through the same pipeline as described above, except that

only mutant segregant pools were sequenced in this case. G-tests were performed by comparing observed mutation frequency in the mutant pool to a null expectation of 0.5.

Quantification of allele frequencies through pyrosequencing

To assess the accuracy of allele frequency estimates obtained through Illumina sequencing, quantitative genotyping of the low and high fluorescence bulks was performed for three variable sites in each mutant using pyrosequencing. These included the site with strongest allele frequency difference between bulks as well as two sites showing no significant difference in allele frequency. Pyrosequencing assays (see File S3) were designed following manufacturer instructions (PyroMark Assay Design software from QIAGEN), except that a universal biotinylated primer was used to reduce the cost. For each variant assessed, PCR reactions were performed as previously described (AYDIN *et al.* 2006) on 5 different genomic DNA templates from the original haploid mutant, the haploid mapping strain, the F1 diploid hybrid and the low and high fluorescence haploid segregants. Quantitative genotyping was performed on a PyroMark ID instrument following the protocol described in Wittkopp (2011). Data from parental strains and the hybrid were used to correct for potential PCR or sequencing biases. Knowing that true allele frequencies are 1, 0 and 0.5 in the mutant, mapping strain, and hybrid, a 2nd degree polynomial regression model was fitted to the observed data and used to correct allele frequencies in the segregant bulks.

Robustness of the BSA-seq approach

Impact of genomic position on mapping success: To determine the limits of our bulk segregant mapping protocol, we tested whether the three causal mutations we identified would have been successfully mapped if they had been located somewhere else in the genome. This might not be the case if the power to map a mutation of a given effect size was uneven across the genome, either because of random fluctuation in sequencing depth or because of reads failing to align uniquely to the genome. To examine this possibility, we first computed for each bulk sample the sequencing depth at every genomic position using *genomecov* tool in BEDTools v2.17.0 (QUINLAN and HALL 2010). We then inferred, for each genomic position in each segregant bulk, the number of mutant and wild type alleles we would have observed if the site was causative given the coverage of the position and the mutation frequency at the actual causative site. We then calculated the fraction of genomic positions for which a mutation with the same effect as the actual causative mutation would have been detected and called significant using the analysis pipeline described in Figure D-4.

Depending on the mutant considered, we found that 2.9% to 3.4% of genomic positions were not covered by any sequencing reads in at least one sample (Figure D - S10A, left bars), making it impossible to test for a significant association. Additionally, 4.1% to 5.1% of genomic positions failed to meet the minimum cutoff of 10 reads in the merged bulks that we required for the site to be called as a high confidence SNP (Figure D - S10A, middle bars). These sites were thus not tested for a significant association with the fluorescence phenotype and the causative mutation would have remained undetected if

located at one of these positions. Finally, we found that 4.2% to 4.9% of sites had insufficient sequencing coverage to yield a significant phenotypic association in a G-test (Figure D - S10A, right bars), most of which also failed to meet the 10 read minimum criterion to be called a SNP. Low sequence read coverage at these sites could be caused by random fluctuations in sequencing depth or problems aligning sequence reads that contain these sites.

To determine how often sites with low coverage resulted from poor alignment of sequence reads, we assessed mappability for each position in the reference genome using software from the GEM library (DERRIEN *et al.* 2012). A genomic site was considered to have perfect mappability if and only if every possible read overlapping that site aligned uniquely to the correct genomic position (STEVENSON *et al.* 2013). Aligning 100 bp sequences to the reference genome while allowing up to five mismatches showed imperfect mappability for 6.8% of the *S. cerevisiae* genome (Figure D - S10B). More than 97% of these sites were included in at least one of the three groups of problematic sites described above (Figure D - S10B), indicating that the inability to uniquely map sequence reads, rather than random variation in sequencing depth, was responsible for the vast majority of sites with low coverage in our dataset. This interpretation is further supported by the genome-wide distributions of sequencing coverage showing two peaks -- one centered at the mean coverage for each sample and the other at 0 (see Figure D - S9). If a causative mutation occurs in a low mappability region, it would remain undetected, but linked mutations could still yield a significant association of the phenotype to a broader genomic region. However, such mapping by linkage is likely to

occur only if the average distance between mutations is smaller than the extent of genetic linkage. Linkage extends approximately 50 kb after a single generation of meiosis in *S. cerevisae* (MORTIMER *et al.* 1991). Given the number of mutations in each mutant isolated in Gruber *et al.* (2012), an average of one mutation is expected every 255 kb, making linkage unlikely for most pairs of sites. Assuming all of these mutations are indeed unlinked, we conclude that a small portion of the genome (~4% on Figure D - S10A, middle bars) is unsuitable to mapping in these mutants using short-read data regardless of sequencing depth.

Impact of decreased sequencing depth on mapping success: To determine how variant calling might have affected our results, we assessed the total number of mutations called for each mutant using the bulk sequencing data when reads from the SAM files were randomly subsampled to a genome coverage ranging from 10x to 110x using the Picard (v1.97) command-line tool *DownsampleSam* (http://picard.sourceforge.net). For all three mutants, a steep drop was observed in the total number of mutations called at 10x coverage relative to 25x coverage (Figure D - S10C). As expected, sites with the lowest read counts for mutant alleles were the first to be missed when sequencing depth was decreased. Interestingly, the only mutation missed in YPW89 mutant when sequencing coverage was reduced to 75x was the causative mutation. This was because this mutation also strongly reduced fitness (Table 3), causing the number of mutant alleles in both bulks to be very low. With decreased coverage, the number of sequencing reads overlapping this site quickly fell below the minimum required for detection as a high confidence SNP.

Finally, we determined how the significance of G-tests used to identify associated sites varied with sequencing depth. The read number for reference and mutant alleles at the causative site were divided by the same values, so that the average sequencing depth between low and high bulks at the site was 80, 70, 60, 50, 40, 30, 20 and 10. We found that the statistical significance of associations between causal sites and YFP fluorescence decreased linearly with sequencing depth, but at different rates for different mutants (Figure D - S10D). For YPW102, as few as 10 reads overlapping the causal site were required to detect a significant association, whereas 15 and 41 reads were required in YPW94 and YPW89, respectively. YPW89 was again found to be the most sensitive to a decrease in sequencing depth despite having the strongest effects on mean fluorescence because its effects on fitness decreased its frequency in both bulks (Figure D - S10C,D).

Supplementary Information

Additional tables and data sets concerning this appendix can be found online with the manuscript at: http://www.g3journal.org/content/4/7/1205.long

Acknowledgments

We thank the University of Michigan Flow Cytometry Core, as well as the Center for Chemical Genomics and High Throughput Screening for the access to flow cytometry and cell sorting platforms, and the University of Michigan Sequencing Core for Next Generation Sequencing. We also thank Brian Chin and William Timberlake for sharing the FASTER MT plasmid, Adam Deutschbauer and Barry Williams for sharing yeast strains, and Joseph Coolon, Alisha John, Rich Lusk, Calum Maclean and Kraig Stevenson for helpful discussions and comments on the manuscript. Financial support for this work was provided to PJW by grants from the University of Michigan, National Science Foundation (MCB-1021398), National Institutes of Health (1 R01 GM108826) and March of Dimes (Basil O'Connor Starter Scholar Research Award 5- FY07-181). FD was funded by a Long Term Fellowship from the European Molecular Biology Organization (EMBO ALTF 1114-2012); BPHM was supported by the University of Michigan Genome Sciences Training Program (T32 HG000040) and by a University of Michigan Rackham Merit Fellowship; JDG was supported by a National Institutes of Health NRSA (1F32GM083513-0); NS was supported by the University of Michigan Undergraduate Research Opportunities Program; and TEB was supported by the National Science Foundation Research Experiences for Undergraduate Program at University of Michigan (ED-QUE2ST).

References

ALBERT F. W., TREUSCH S., SHOCKLEY A. H., BLOOM J. S., KRUGLYAK L., 2014 Genetics of single-cell protein abundance variation in large yeast populations. Nature **506**: 494–497.

AYDIN A., TOLIAT M. R., BÄHRING S., BECKER C., NÜRNBERG P., 2006 New universal primers facilitate Pyrosequencing. Electrophoresis **27**: 394–397.

BASTIDE H., BETANCOURT A., NOLTE V., TOBLER R., STÖBE P., FUTSCHIK A., SCHLÖTTERER C., 2013 A genome-wide, fine-scale map of natural pigmentation variation in Drosophila melanogaster. PLoS Genet. **9**: e1003534.

BIRKELAND S. R., JIN N., OZDEMIR a. C., LYONS R. H., WEISMAN L. S., WILSON T. E., 2010 Discovery of Mutations in Saccharomyces cerevisiae by Pooled Linkage Analysis and Whole-Genome Sequencing. Genetics **186**: 1127–1137.

BRAUER M. J., CHRISTIANSON C. M., PAI D. a, DUNHAM M. J., 2006 Mapping novel traits by array-assisted bulk segregant analysis in Saccharomyces cerevisiae. Genetics **173**: 1813–6.

BRESLOW D. K., CAMERON D. M., COLLINS S. R., SCHULDINER M., STEWART-ORNSTEIN J., NEWMAN H. W., BRAUN S., MADHANI H. D., KROGAN N. J., WEISSMAN J. S., 2008 A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. Nat. Methods **5**: 711–718.

CHIN B. L., FRIZZELL M. a., TIMBERLAKE W. E., FINK G. R., 2012 FASTER MT: Isolation of Pure Populations of a and Ascospores from Saccharomyces cerevisiae. G3 **2**: 449–452.

DERRIEN T., ESTELLÉ J., MARCO SOLA S., KNOWLES D. G., RAINERI E., GUIGÓ R., RIBECA P., 2012 Fast Computation and Applications of Genome Mappability. PLoS One **7**: e30377.

DEUTSCHBAUER A. M., DAVIS R. W., 2005 Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat. Genet. **37**: 1333–40.

DEUTSCHBAUER A. M., JARAMILLO D. F., PROCTOR M., KUMM J., HILLENMEYER M. E., DAVIS R. W., NISLOW C., GIAEVER G., 2005 Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics **169**: 1915–25.

EDWARDS M. D., GIFFORD D. K., 2012 High-resolution genetic mapping with pooled sequencing. BMC Bioinformatics **13 Suppl 6**: S8.

EHRENREICH I. M., BLOOM J., TORABI N., WANG X., JIA Y., KRUGLYAK L., 2012 Genetic Architecture of Highly Complex Chemical Resistance Traits across Four Yeast Strains (AP Gasch, Ed.). PLoS Genet. **8**: e1002570.

EHRENREICH I. M., TORABI N., JIA Y., KENT J., MARTIS S., SHAPIRO J. a, GRESHAM D., CAUDY A. a, KRUGLYAK L., 2010 Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature **464**: 1039–42.

EYRE-WALKER A., KEIGHTLEY P. D., 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. **8**: 610–8.

FELLOWS I., 2012 Deducer : A Data Analysis GUI for R. J. Stat. Softw. **49**: 1–15.

GERKE J., LORENZ K., COHEN B., 2009 Genetic interactions between transcription factors cause natural variation in yeast. Science **323**: 2007–2010.

GERSTEIN A. C., OTTO S. P., 2011 Cryptic fitness advantage: diploids invade haploid populations despite lacking any apparent advantage as measured by standard fitness assays. PLoS One **6**: e26599.

GRANEK J. a, MURRAY D., KAYIKÇI O., MAGWENE P. M., 2012 The Genetic Architecture of Biofilm Formation in a Clinical Isolate of Saccharomyces cerevisiae. Genetics: 1–53.

GRUBER J. D., VOGEL K., KALAY G., WITTKOPP P. J., 2012 Contrasting Properties of Gene-specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects and Dominance. PLoS Genet. **8**: e1002497.

HANLON S. E., RIZZO J. M., TATOMER D. C., LIEB J. D., BUCK M. J., 2011 The stress response factors Yap6, Cin5, Phd1, and Skn7 direct targeting of the conserved co-repressor Tup1-Ssn6 in S. cerevisiae. PLoS One **6**: e19060.

HARBISON C. T., GORDON D. B., LEE T. I., RINALDI N. J., MACISAAC K. D., DANFORD T. W., HANNETT N. M., TAGNE J.-B., REYNOLDS D. B., YOO J., JENNINGS E. G., ZEITLINGER J., POKHOLOK D. K., KELLIS M., ROLFE P. A., TAKUSAGAWA K. T., LANDER E. S., GIFFORD D. K., FRAENKEL E., YOUNG R. A., 2004 Transcriptional regulatory code of a eukaryotic genome. Nature **431**: 99–104.

KOBOLDT D. C., CHEN K., WYLIE T., LARSON D. E., MCLELLAN M. D., MARDIS E. R., WEINSTOCK G. M., WILSON R. K., DING L., 2009 VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics **25**: 2283–2285.

KOBOLDT D. C., ZHANG Q., LARSON D. E., SHEN D., MCLELLAN M. D., LIN L., MILLER C. a, MARDIS E. R., DING L., WILSON R. K., 2012 VarScan 2: Somatic mutation

and copy number alteration discovery in cancer by exome sequencing. Genome Res. **22**: 568–76.

KOFLER R., PANDEY R. V., SCHLÖTTERER C., 2011 PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics **27**: 3435–6.

KUMAR P., HENIKOFF S., NG P. C., 2009 Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. **4**: 1073–1081.

LANGMEAD B., SALZBERG S. L., 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**: 357–359.

LAROCHELLE M., DROUIN S., ROBERT F., TURCOTTE B., 2006 Oxidative stress-activated zinc cluster protein Stb5 has dual activator/repressor functions required for pentose phosphate pathway regulation and NADPH production. Mol. Cell. Biol. **26**: 6690–6701.

LEEUWEN T. Van, DEMAEGHT P., OSBORNE E. J., DERMAUW W., GOHLKE S., NAUEN R., 2012 Population bulk segregant mapping uncovers resistance mutations and the mode of action of a chitin synthesis inhibitor in arthropods. : 2–7.

LI H., HANDSAKER B., WYSOKER A., FENNELL T., RUAN J., HOMER N., MARTH G., ABECASIS G., DURBIN R., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

LITI G., LOUIS E. J., 2012 Advances in quantitative trait analysis in yeast. PLoS Genet. **8**: e1002912.

MAGWENE P. M., WILLIS J. H., KELLY J. K., 2011 The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing (A Siepel, Ed.). PLoS Comput. Biol. **7**: e1002255.

MENNELLA T. A., KLINKENBERG L. G., ZITOMER R. S., 2003 Recruitment of Tup1-Ssn6 by yeast hypoxic genes and chromatin-independent exclusion of TATA binding protein. Eukaryot. Cell **2**: 1288–1303.

MICHELMORE R. W., PARAN I., KESSELI R. V, 1991 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. PNAS **88**: 9828–32.

MORTIMER R. K., SCHILD D., CONTOPOULOU C. R., KANS J. A., 1991 [57] Genetic and physical maps of Saccharomyces cerevisiae. Methods Enzymol. **194**: 827–863.

PAN X., YUAN D. S., XIANG D., WANG X., SOOKHAI-MAHADEO S., BADER J. S., HIETER P., SPENCER F., BOEKE J. D., 2004 A robust toolkit for functional profiling of the yeast genome. Mol. Cell **16**: 487–496.

PARTS L., CUBILLOS F. a., WARRINGER J., JAIN K., SALINAS F., BUMPSTEAD S. J., MOLIN M., ZIA A., SIMPSON J. T., QUAIL M. a., MOSES A. M., LOUIS E. J., DURBIN R., LITI G., 2011 Revealing the genetic structure of a trait by sequencing a population under selection. Genome Res. **21**: 1131–8.

POMRANING K. R., SMITH K. M., FREITAG M., 2011 Bulk segregant analysis followed by high-throughput sequencing reveals the Neurospora cell cycle gene, ndc-1, To be allelic with the gene for ornithine decarboxylase, spe-1. Eukaryot. Cell **10**: 724–733.

QUINLAN A. R., HALL I. M., 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**: 841–2.

R CORE TEAM, 2013 R: A language and environment for statistical computing.

ROBINSON M. D., SMYTH G. K., 2007 Moderated statistical tests for assessing differences in tag abundance. Bioinformatics **23**: 2881–7.

ROBINSON M. D., SMYTH G. K., 2008 Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics **9**: 321–32.

ROHLAND N., REICH D., 2012 Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. **22**: 939–46.

SONG W., TREICH I., QIAN N., KUCHIN S., CARLSON M., 1996 SSN genes that affect transcriptional repression in Saccharomyces cerevisiae encode SIN4, ROX3, and SRB proteins associated with RNA polymerase II. Mol. Cell. Biol. **16**: 115–120.

STEVENSON K. R., COOLON J. D., WITTKOPP P. J., 2013 Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. BMC Genomics **14**: 536.

STORICI F., RESNICK M. a M., 2006 The delitto perfetto approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. Methods Enzymol. **409**: 329–45.

SWINNEN S., SCHAERLAEKENS K., PAIS T., CLAESEN J. J., HUBMANN G., YANG Y., DEMEKE M., FOULQUIE-MORENO M. R., GOOVAERTS A., SOUVEREYNS K., CLEMENT L., DUMORTIER F. F., THEVELEIN J. M., FOULQUIÉ-MORENO M. R., 2012 Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis. Genome Res. **22**: 975–84.

TONG a H., EVANGELISTA M., PARSONS A. B., XU H., BADER G. D., PAGÉ N., ROBINSON M., RAGHIBIZADEH S., HOGUE C. W., BUSSEY H., ANDREWS B. J., TYERS M., BOONE C., 2001 Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science **294**: 2364–8.

WENGER J. W., SCHWARTZ K., SHERLOCK G., 2010 Bulk Segregant Analysis by High-Throughput Sequencing Reveals a Novel Xylose Utilization Gene from Saccharomyces cerevisiae (Z Gu, Ed.). PLoS Genet. **6**: e1000942.

WICKS S. R., YEH R. T., GISH W. R., WATERSTON R. H., PLASTERK R. H., 2001 Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. Nat. Genet. **28**: 160–164.

WILKENING S., LIN G., FRITSCH E. S., TEKKEDIL M. M., ANDERS S., KUEHN R., NGUYEN M., AIYAR R. S., PROCTOR M., SAKHANENKO N. a, GALAS D. J., GAGNEUR J., DEUTSCHBAUER A., STEINMETZ L. M., 2013 An Evaluation of High-Throughput Approaches to QTL Mapping in Saccharomyces cerevisiae. Genetics: 853–865.

WITTKOPP P. J., 2012 Using Pyrosequencing to Measure Allele-Specific mRNA Abundance and Infer the Effects of Cis- and Trans-regulatory Differences. In: Orgogozo V, Rockman M V. (Eds.), *Molecular Methods for Evolutionary Genetics*, Methods in Molecular Biology. Humana Press, Totowa, NJ.

WLOCH D. M., SZAFRANIEC K., BORTS R. H., KORONA R., 2001 Direct estimate of the mutation rate and the distribution of fitness effects in the yeast Saccharomyces cerevisiae. Genetics **159**: 441–52.

XIA Y., WON S., DU X., LIN P., ROSS C., VINE D. LA, WILTSHIRE S., LEIVA G., VIDAL S. M., WHITTLE B., GOODNOW C. C., KOZIOL J., Y MORESCO E. M., BEUTLER B., 2010 Bulk segregation mapping of mutations in closely related strains of mice. Genetics **186**: 1139–1146.

YANG Y., FOULQUIÉ-MORENO M. R., CLEMENT L., ERDEI E., TANGHE A., SCHAERLAEKENS K., DUMORTIER F., THEVELEIN J. M., 2013 QTL analysis of high thermotolerance with superior and downgraded parental yeast strains reveals new minor QTLs and converges on novel causative alleles involved in RNA processing. PLoS Genet. **9**: e1003693.

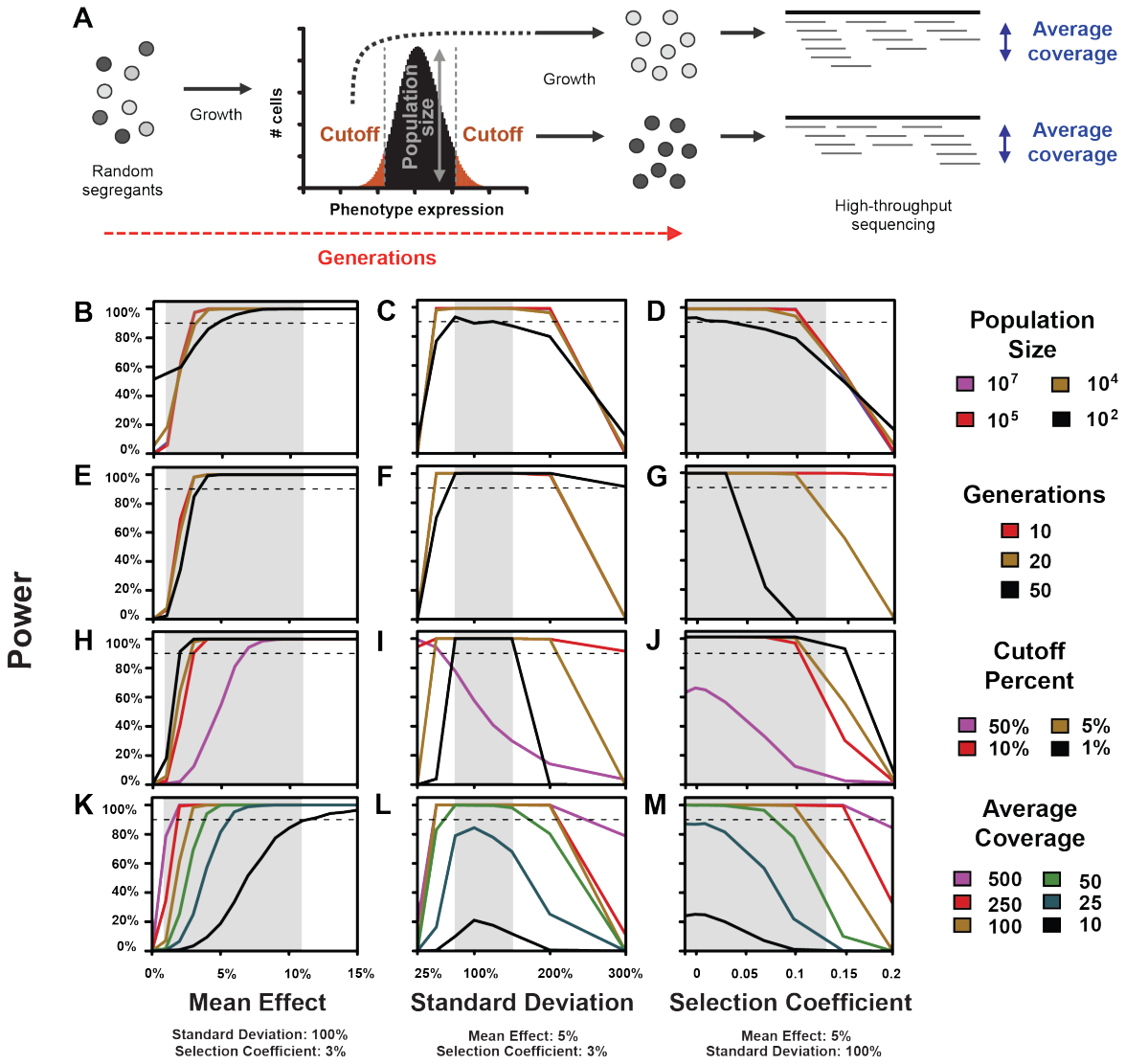


Figure D-1 Experimentally controllable parameters affecting statistical power

(A) An overview of the modeled BSA-seq experiment is shown, with the four experimental parameters we allowed to vary (population size, generations of growth, cutoff for bulk selection, and average coverage of sequencing) indicated. Power is shown for various population sizes (B-D), generations of growth (E-G), bulk selection cutoffs (H-J), and average sequencing coverages (K-M), for a range of effects of the causal mutation on mean expression (B, E, H, K), standard deviation of expression (C, F, I, L), and fitness (measured in terms of the selection coefficient) (D, G, J, M). In all plots, the dashed line indicates 90% power. Gray shaded regions represent 90% confidence intervals of the mean effect and standard deviation of the fluorescence phenotypes observed in a recent set of trans-regulatory mutants (Gruber et al. 2012, see Figure D - S2A,B). The 90% confidence interval for selection coefficients was inferred from fitness assays performed on 8 mutants (see Materials and Methods and Figure D - S2C). In all analyses, only the indicated parameters were allowed to vary; all other experimentally controllable parameters were fixed at values ultimately used in our mapping experiment (Sequencing Depth = 100, Population Size = 107, Cutoff Percent = 5%, Generations = 20), and mutational parameters were fixed at values representative of the mutants used for mapping (Mean effect = 5%, Standard Deviation = 100%, Selection Coefficient = 0.03).

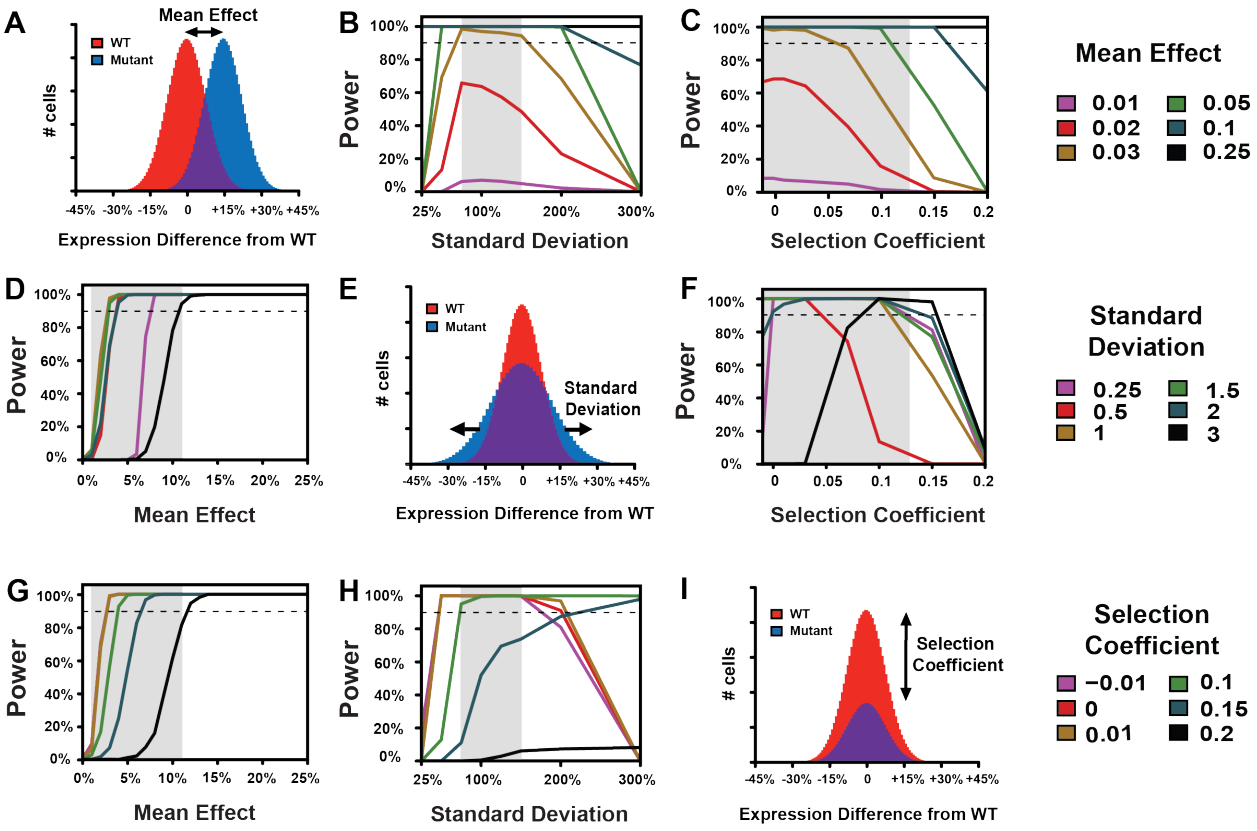


Figure D-2 Inherent properties of mutations affecting statistical power

Power is shown for various mutation effects on mean (B,C), standard deviation (D,F), and relative fitness (G,H). Comparisons of hypothetical wild-type (red) and mutant (blue) populations with effects of a mutation on mean expression (A), standard deviation of expression (E), and relative fitness (I) are also shown. In all plots, the dashed line indicates 90% power. Gray shaded regions represent values of the mean effect, standard deviation, or selection coefficient of causal mutations observed in a recent set of expression mutants (see Figure D - S2). In all analyses, only the indicated parameters were allowed to vary; all others were fixed. These fixed values were: Mean effect = 5%, Standard Deviation = 100%, Selection Coefficient = 0.03, Sequencing Depth = 100, Population Size = 107, Cutoff Percent = 5%, Generations = 20.

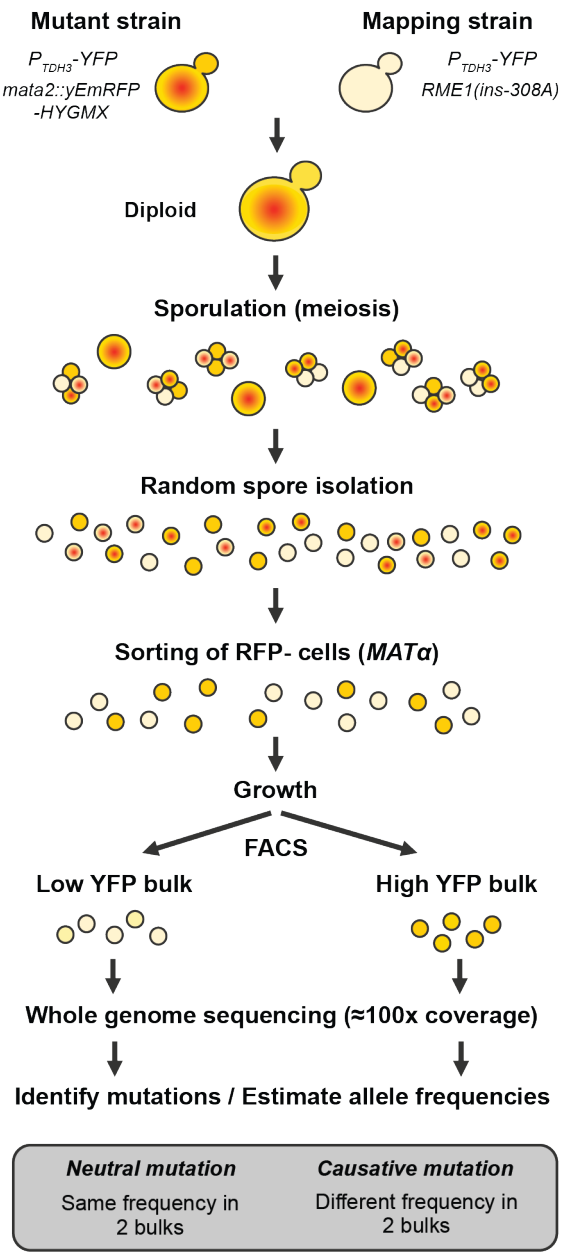


Figure D-3 Overview of experimental design for mapping small effect mutations

This approach is based on the isolation of a large number of random F1 segregant haploid cells, followed by high-throughput phenotypic selection using FACS, and estimation of allele frequencies genome-wide using next generation sequencing. Note the selection of haploid MATα cells using expression of the RFP reporter linked to MATα locus that is indicated with a red dot. Quantitative differences in the level of YFP expression are indicated by differences in the intensity of yellow background.

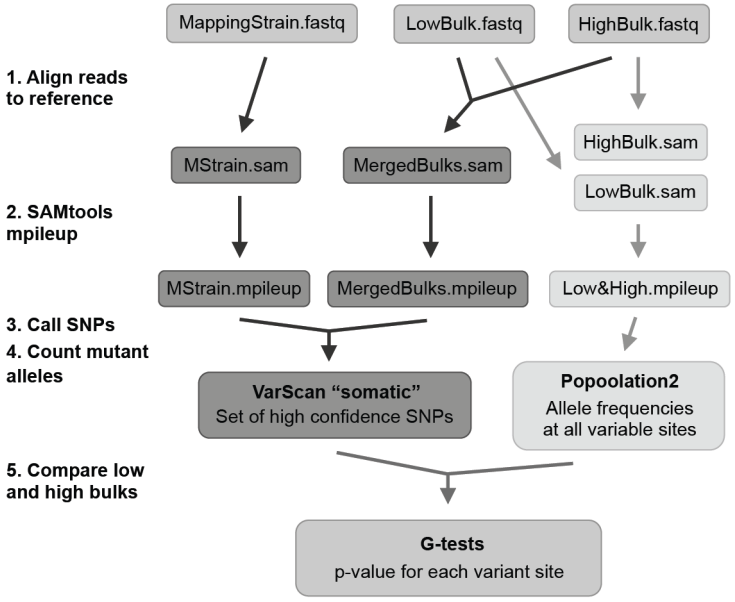


Figure D-4 Analysis of Illumina sequencing data

A set of high confidence variants was called using the somatic command in VarScan (dark gray), with reads from the mapping strain treated as "normal" data and reads from merged bulks treated as "tumor" data. Allele frequencies were then estimated for these sites in the low fluorescence and high fluorescence bulks with Popoolation2 (light gray). Differences between these two bulks were assessed using G-tests.

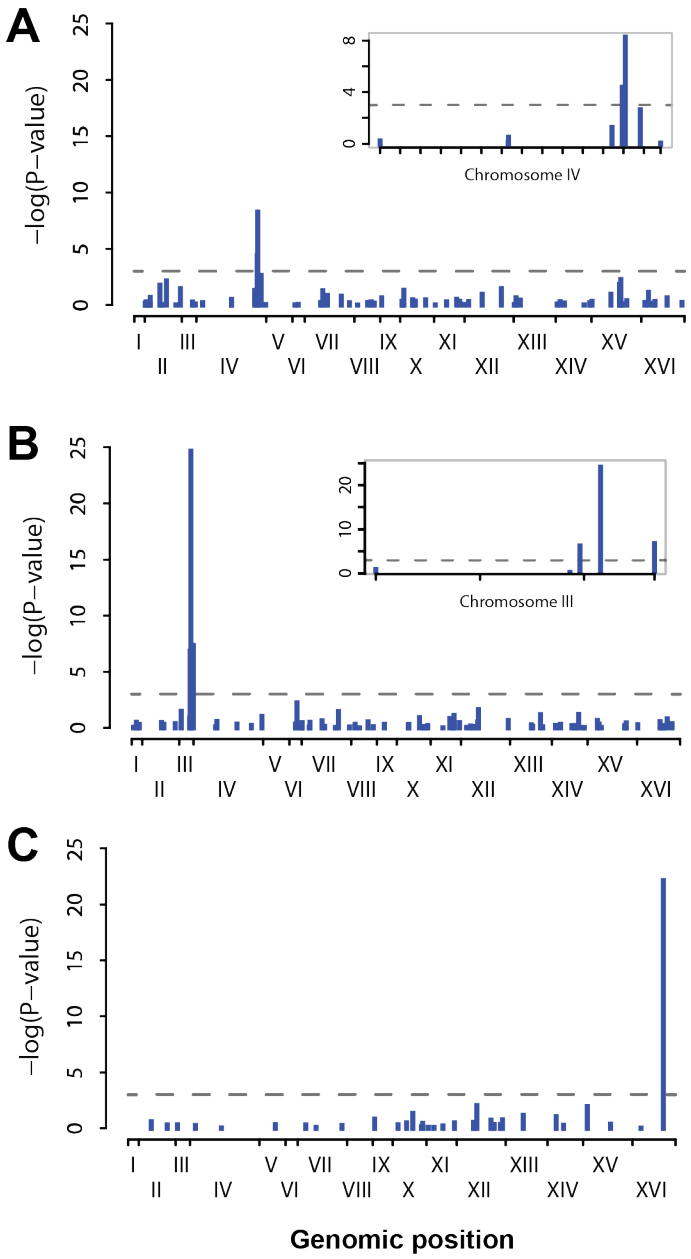


Figure D-5 BSA-seq identifies a single causative site for each mutant

Significance of the difference in allele frequency between low fluorescence and high fluorescence bulks is shown as the negative of logarithm of P-value from G-test for mutants YPW89 (A), YPW94 (B) and YPW102 (C). Each bar shows significance for an individual EMS-induced mutation with its genomic position represented on x-axis. Roman numerals indicate each of the 16 S. cerevisiae chromosomes. Insets in (A) and (B) are magnifications of chromosomes harboring causative sites and show linked mutations with significant effects. Horizontal dotted lines represent a significance threshold of $\alpha = 0.001$.

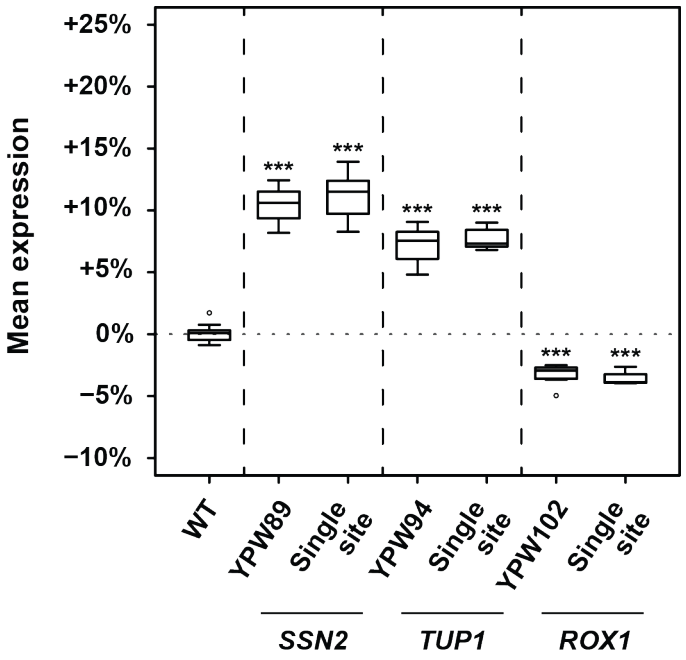


Figure D-6 Single mutations identified completely explain mutant phenotypes

Mean expression level (measured as YFP fluorescence) for 8 replicates each of the wild-type genotype (WT), the YPW89, YPW94, and YPW102 mutants, and the three allele-replacement strains ("Single site") for each mapped mutation (SSN2(Q971Stop), TUP1(G696D) and ROX1(R12K)) are shown. For each replicate, the median level of fluorescence of at least 5000 cells was quantified and expressed relative to mean fluorescence in the wild-type (WT) reference strain. Mutant genotypes and allele-replacement strains were compared to the WT strain using t-tests (*** $P<0.001$). In all three cases, the single site mutant was found to phenocopy the EMS mutant strain ($P=0.58$ for SSN2, $P=0.23$ for TUP1, and $P=0.44$ for ROX1).

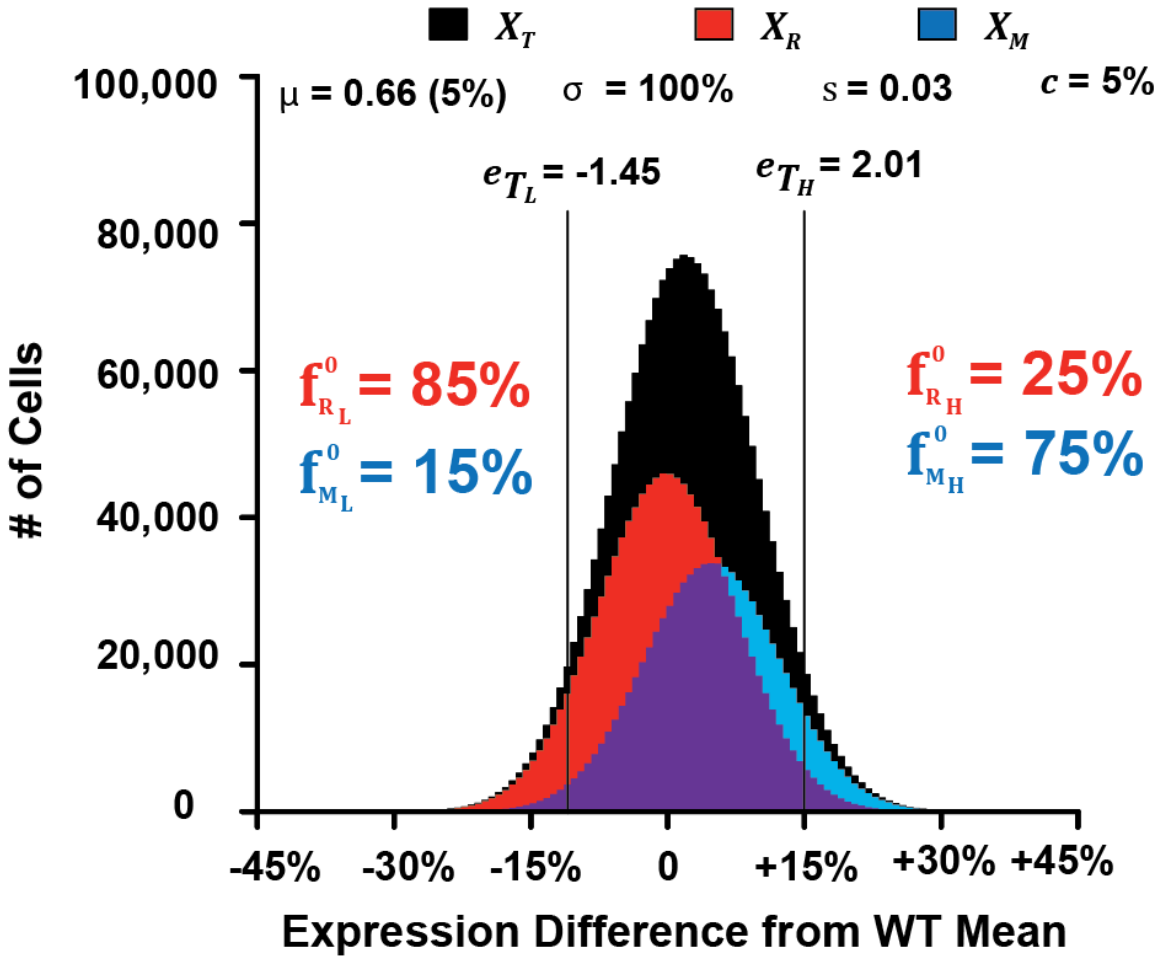


Figure D - S1 Example of phenotypic distribution after the deterministic phase of simulation

For a full description of the model used to generate these distributions, see File S1. The phenotypic distribution for all cells in the population (X_T) is shown in black, whereas the phenotypic distributions for cells carrying the reference (X_R) and mutant (X_M) alleles of the causative site are shown in red and blue, respectively. Black lines show the 5th and 95th percentiles of the phenotypic distribution for all cells, which correspond to the thresholds used for sorting with a 5% cutoff for the high and low bulks. The frequency of the reference allele (f_R) and the frequency of the mutant allele (f_M) are shown for both the low (L) and high (H) bulks. Results are shown for a causative mutation that changes the mean (μ) by 5%, has no effect on the phenotypic standard deviation (σ), and has a selection coefficient (s) of 0.03, when the selected bulks are obtained using a 5% cutoff (c).

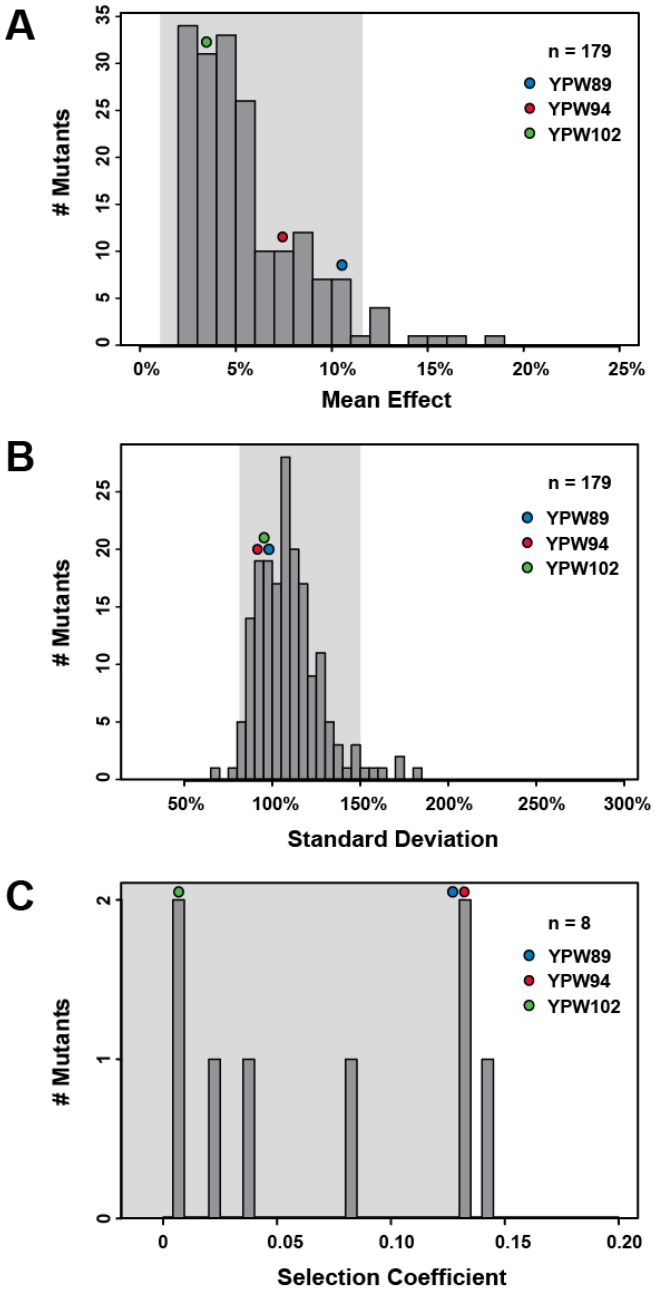


Figure D - S2 Phenotypic effects of the *trans*-regulatory mutants described

Absolute values of effects on mean expression level (A) and standard deviation of fluorescence (B) relative to wild type are shown for the full set of 179 *trans*-regulatory mutants. (C) Selection coefficients for 8 randomly selected mutants, including the three mutants used for mapping in this study (YPW89, YPW94 and YPW102), are shown. Shaded regions show confidence intervals excluding the 10% most extreme mutants and correspond to the shaded regions in Figure 1 and Figure 2. Colored dots indicate the parameter values for the three mutants analyzed in this study.

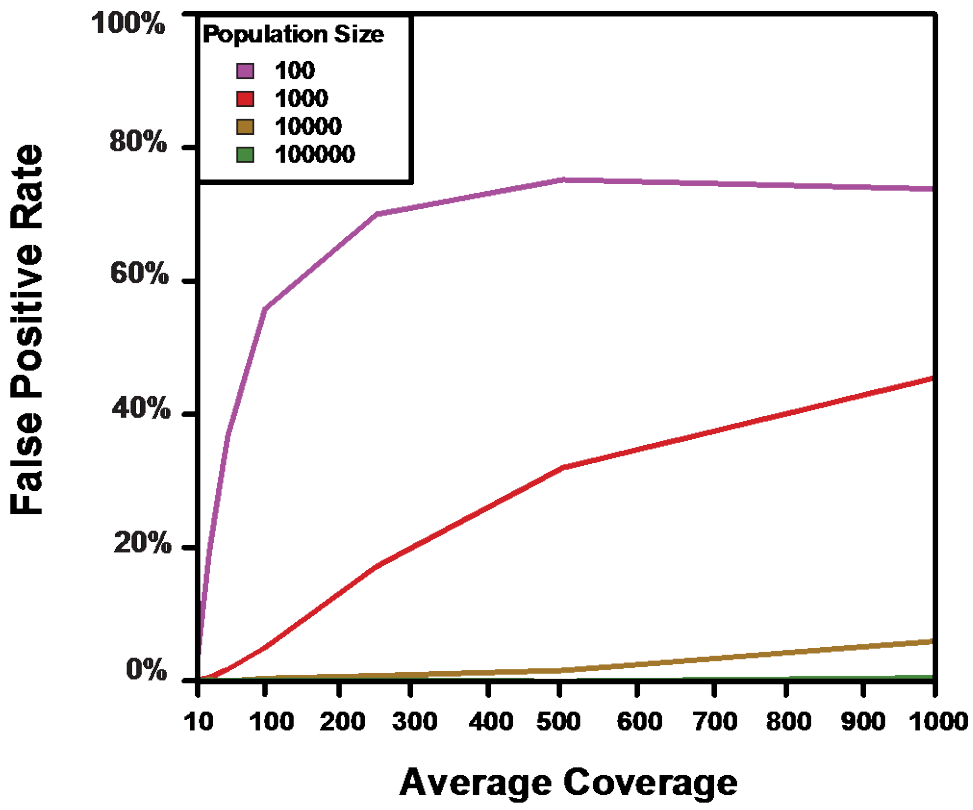


Figure D - S3 Statistical power to detect a difference in the frequency of a neutral mutation

Power for mean effect = 0% between bulks depending on average depth of coverage and population size. This power corresponds to the false discovery rate.

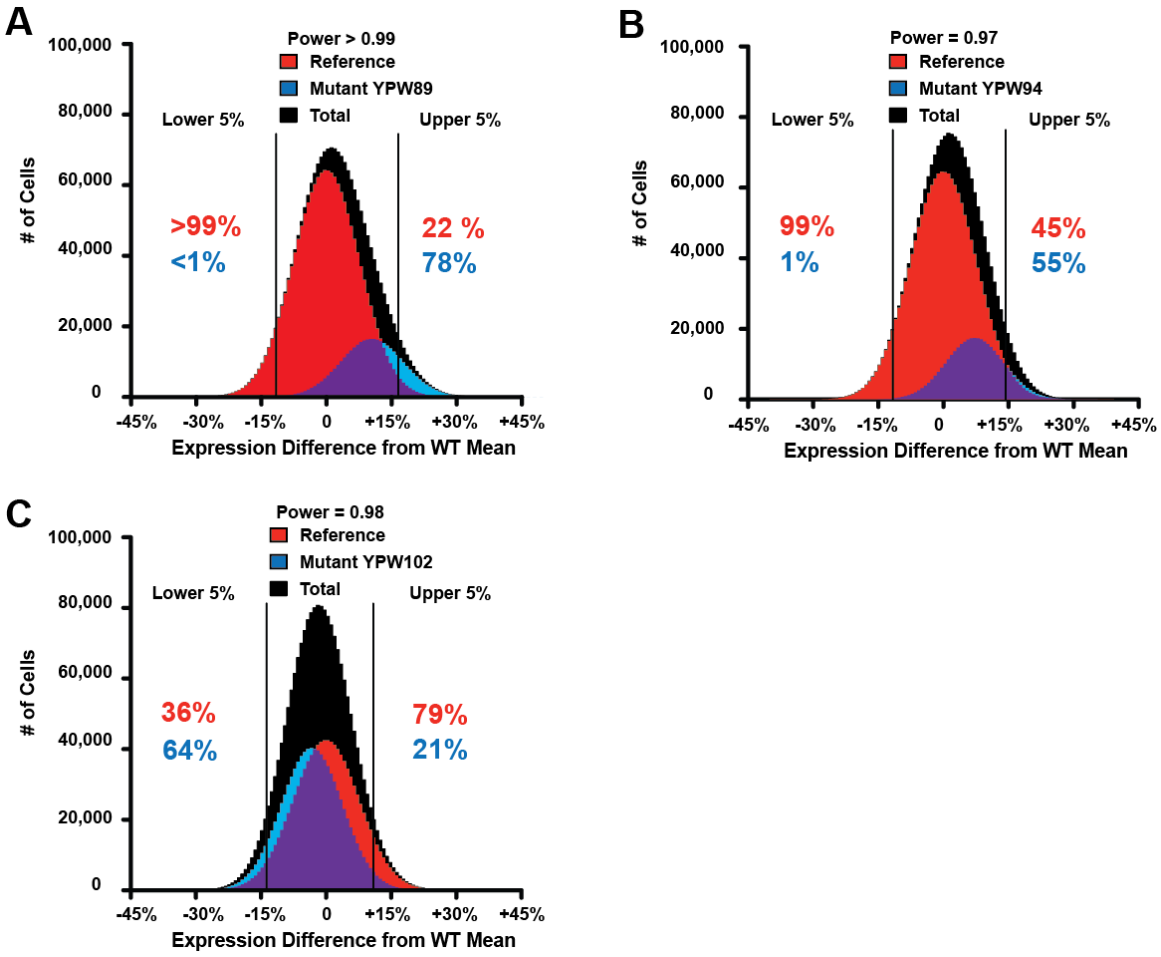
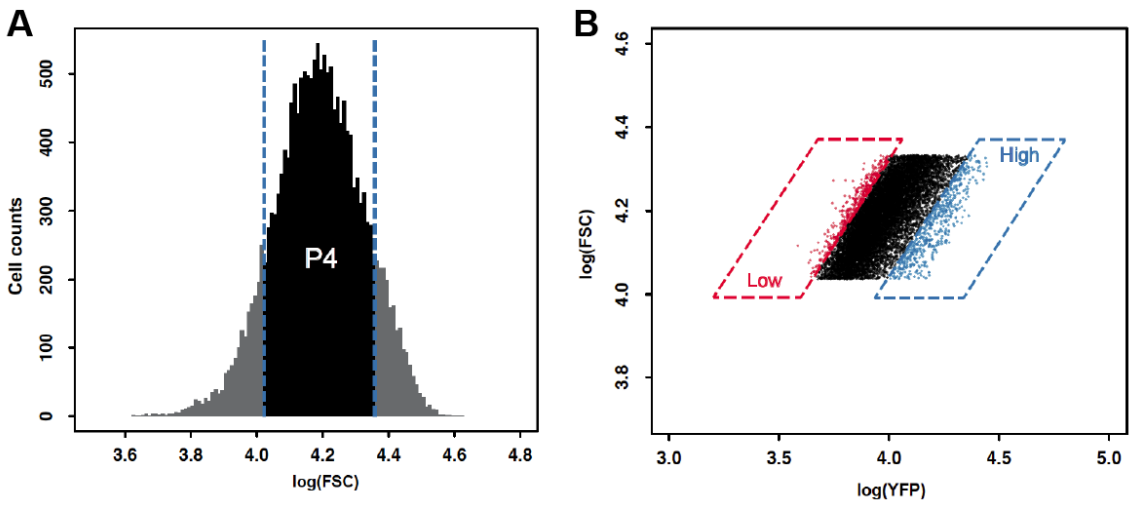


Figure D - S4 Phenotypic distributions after the deterministic phase of the simulations

Results are shown for mutant YPW89 (A), YPW94 (B), and YPW102 (C). Parameters used for the mean, standard deviation, and selection coefficient of the mutant causal allele were estimated from fluorescence and fitness phenotypes of the mutant strains (Table 1). Black: Total population distribution; Red: Reference allele containing population distribution; Blue: Mutant allele containing population distribution. Black lines show the 5% and 95% cutoffs on the total (black) distribution. Numbers in red indicate the frequency of the reference allele in the two bulks while numbers in blue indicate the frequency of the mutant allele in the two bulks. The power to detect a significant difference ($P < 0.001$) in mutation frequency between lower and higher tails in a G-test given an average sequencing coverage of 75 is shown above each plot.



C

Population	#Events	%Parent	%Total	FSC Median	YFP Median
All Events	50,000	----	100.0	25,293	13,873
P4	12,155	86.5	24.3	14,916	9,006
Low Bulk	391	3.3	0.8	14,266	6,000
High Bulk	385	3.2	0.8	14,249	17,909

Figure D - S5 FACS gating used to collect low and high fluorescence pools

Data shown is from the analysis of mutant YPW102. Gating based on the relationship between FSC-A and FSC-H was used to remove cell doublets (not shown). (A) Gating based on FSC-A, which is a proxy for cell size, was then used to exclude the smallest ~8% and the largest ~8% of events. (B) Finally, low and high bulks were selected based on fluorescence level (log(YFP)) and cell size (log(FSC-A)). Careful attention was paid to select bulks with different fluorescence levels, but similar cell sizes. (C) Changes in event number (cells) resulting from the gates shown in panels (A) and (B).

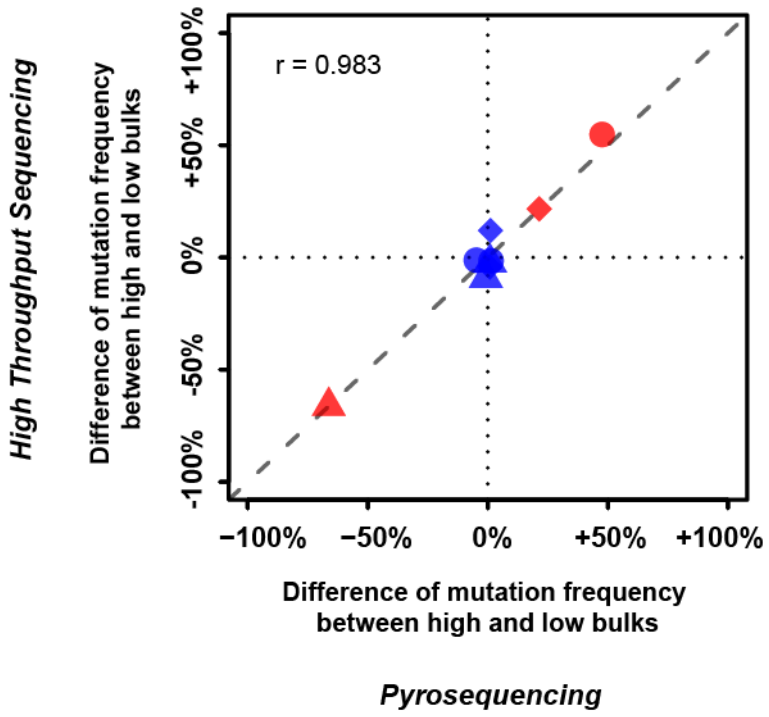


Figure D - S6 Allele frequency correlation between Illumina and pyro-sequencing

For each mutant, pyrosequencing assays were developed for quantitative genotyping of two phenotypically neutral sites (blue) as well as for the site with the highest significance of association with the fluorescence phenotype (red). The plot shows the difference in mutant allele frequency between the high fluorescence and low fluorescence bulks for each site as determined by pyrosequencing (x-axis) or whole genome sequencing (y-axis). Different shapes represent sites analyzed in different mutants (diamond: YPW89, circle: YPW94, triangle: YPW102).

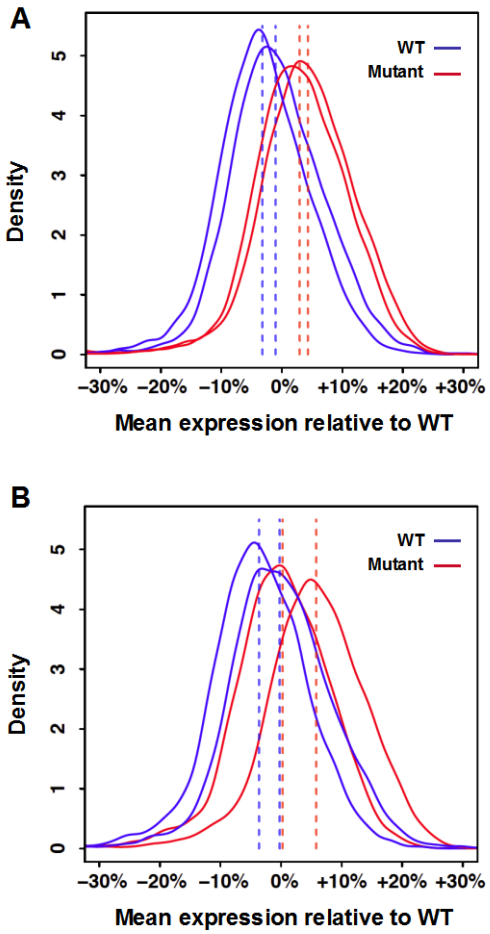


Figure D - S7 Spore phenotypes assayed after tetrad dissection

Segregation of the YFP phenotype in two tetrads derived from mutant YPW54 are shown. (A) Tetrad showing a clear 2:2 segregation of fluorescence level. (B) Tetrad for which mutant and wild-type progeny are hard to distinguish based on fluorescence, potentially leading to incorrect assignment to a phenotypic pool when assembling mutant and reference pools for mapping. Blue and red solid lines show distributions of fluorescence for populations derived from spores assumed to harbor wild type and mutant alleles of the causative site, respectively. Dotted lines indicate the median fluorescence level for each of the wild-type (blue) and mutant (red) populations.

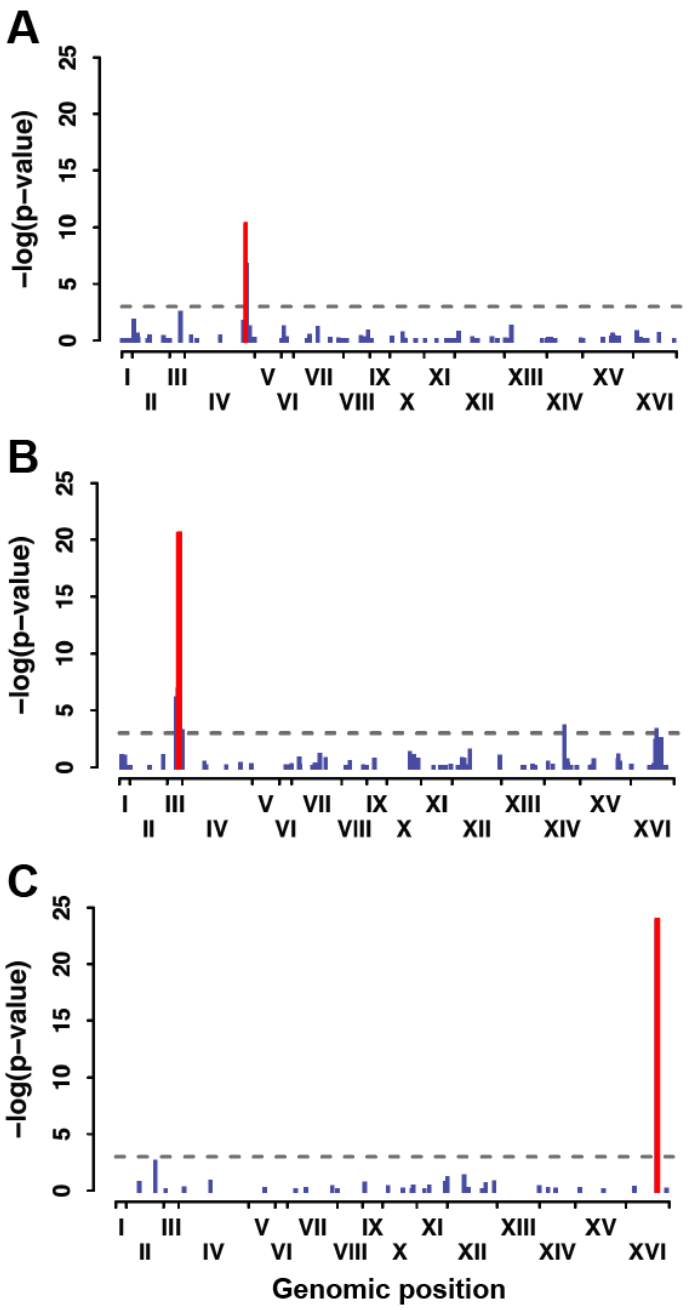


Figure D - S8 Tetrad-based mapping identifies the same candidate sites as BSA-seq.

Mapping results from a more traditional mapping approach based on tetrad dissection are shown for mutant YPW89 (A), YPW94 (B) and YPW102 (C). Colored bars represent individual EMS-induced mutations with their genomic position represented on x-axis and significance of the difference in allele frequency between low fluorescence and high fluorescence bulks represented on y-axis (negative logarithm of *P*-value from G-test). For each mutant, the most significant site identified by BSA-seq is shown in red. Horizontal dotted lines represent a significance threshold of α=0.001.

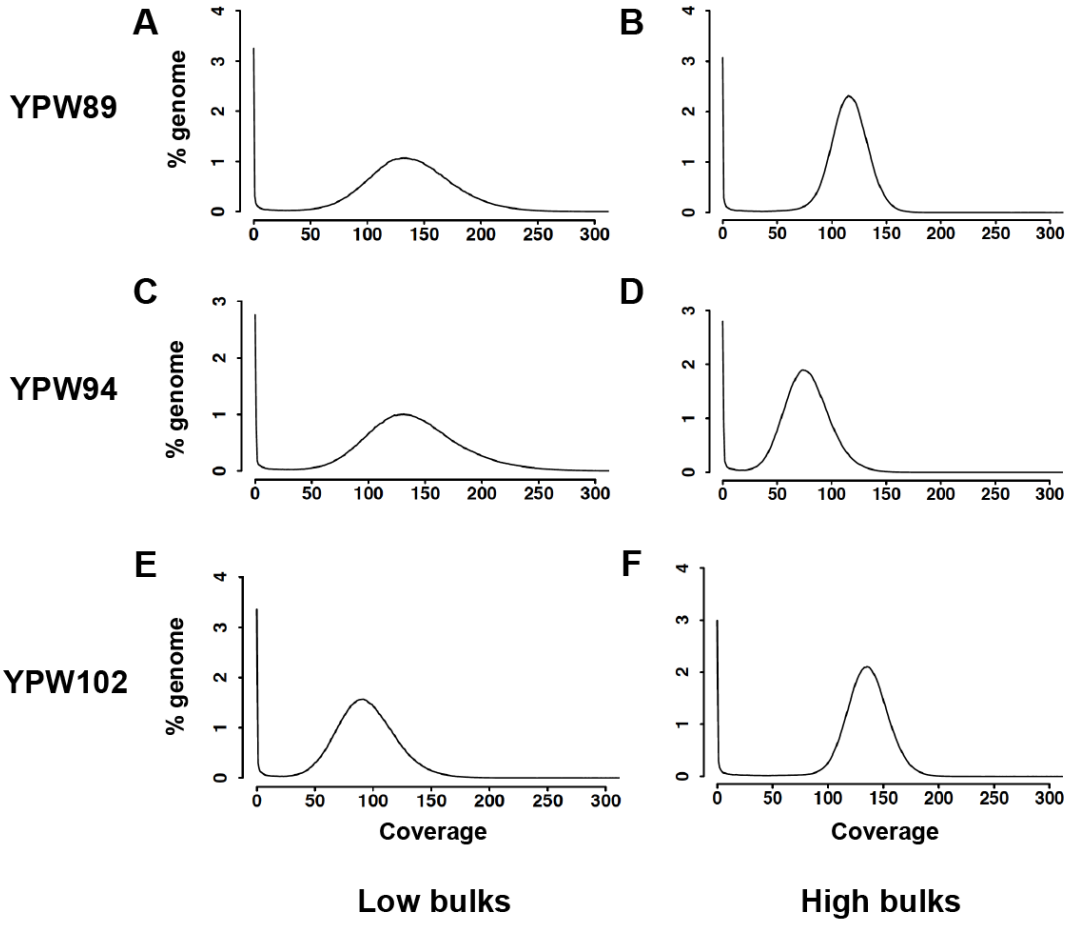


Figure D - S9 Sequencing coverage for each bulk shows two peaks

Distributions of sequencing coverage across reference genome are shown for low (A,C,E) and high (B,D,F) bulks obtained from mutants YPW89 (A,B), YPW94 (C,D) and YPW102 (E,F). Note the peak at 0, which indicates sites with no overlapping sequencing reads, in addition to the peak near the average coverage for each sample.

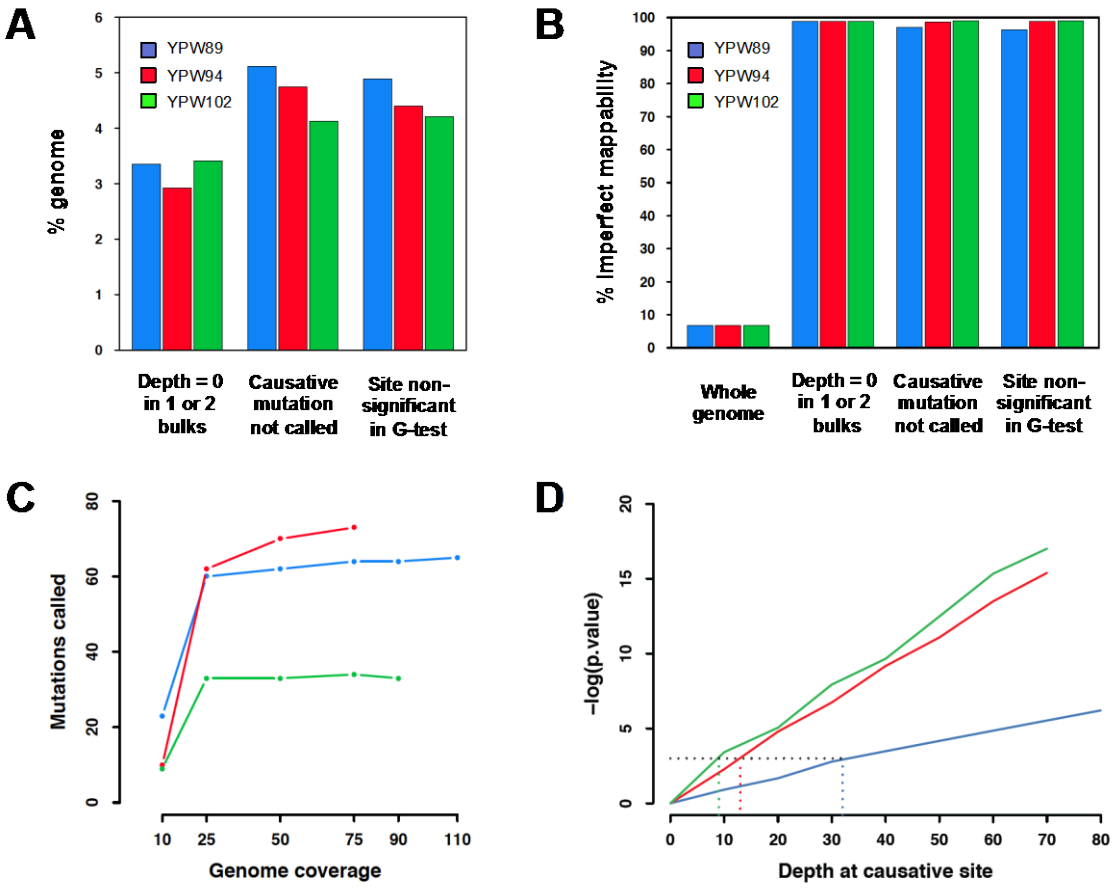


Figure D - S10 Poor mapping not low coverage responsible for most mapping blind spots.

(A) For each mutant, the fraction of the genome with insufficient coverage to detect a putative causative mutation because zero reads were observed in one or both bulks (left), the causative mutation was not called as a sequence variant (middle), or the number of reads mapped was insufficient to generate a significant G-test (right) is shown. (B) Proportion of sites with imperfect mappability across the whole genome and for each genomic class considered in (A) are shown. The vast majority of sites for which a putative mutation could not be detected also showed poor mappability. (C) Robustness of the total number of mutations called to variation in sequencing depth is shown. For each mutant, SNPs were called after subsampling mapped reads to a sequencing depth of 90x, 75x, 50x, 25x and 10x in low and high bulks. Mutants are color coded as in panel (A). (D) Significance of the causative site depending on its coverage is shown. For a constant mutation frequency at the causative site, the total number of alleles was decreased from 90 to 0 (x-axis) and the *P*-value of the G-test was computed (y-axis). Mutants are color coded as in panel (A). Dotted lines highlight the threshold of coverage below which *P*-values were considered non-significant ($P>0.001$).

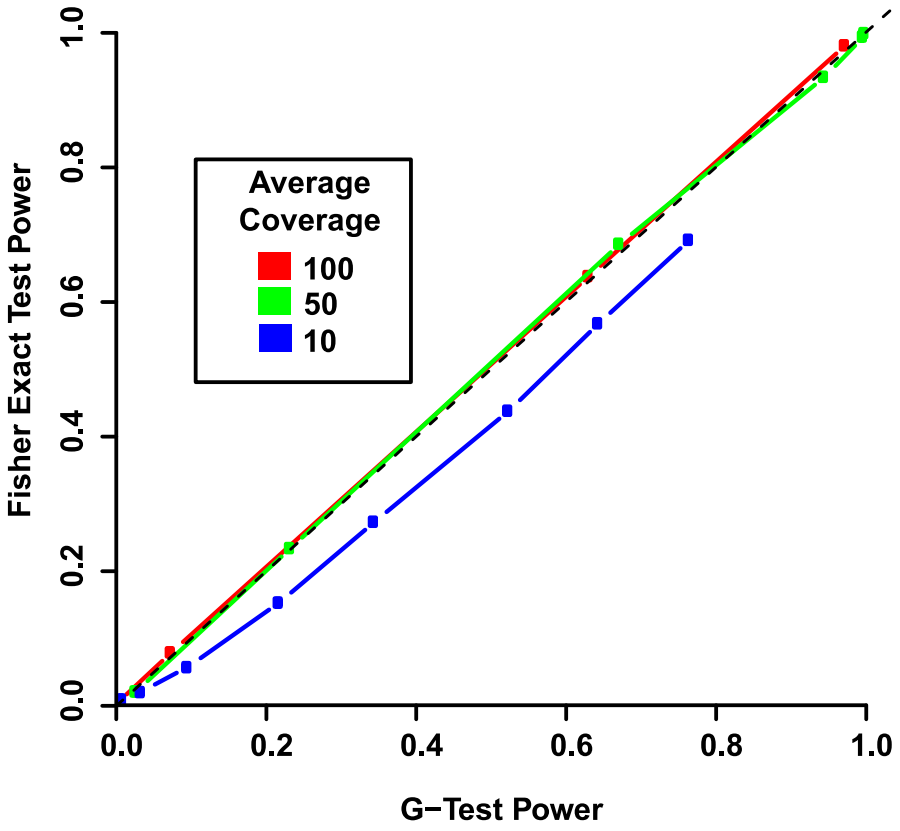


Figure D - S11 Comparison of statistical power using Fisher's exact test and G-test.

Power to detect a significant difference in allele frequency between bulks for different mutation effect sizes and sequencing depths is shown. Dots on each line represent different mutation effects ranging from 0% to +25% (bottom left to top right) relative to WT mean expression. Fixed parameter values were: Standard Deviation = 100%, Selection Coefficient = 0.03, Population Size = 10^7, Cutoff Percent = 5%, Generations = 20.

Mutant	Mean effect		Std Dev[c]	Selection Coef.[d]
	%[a]	sd[b]		
YPW89	+10.45%	+1.38	94.77%	0.127
YPW94	+7.21%	+0.95	86.92%	0.130
YPW102	-3.25%	-0.43	92.51%	0.009

Table D-1 Properties for the three mutants analyzed

[a]Mean expression of mutant relative to wild type expressed as a percentage of change in fluorescence phenotype relative to wild type. [b]Mean expression of mutant relative to wild type expressed as a number of wild type standard deviation from wild type mean. [c]Standard deviation of expression phenotype of the mutant strain relative to the reference strain. [d]Selection coefficient was measured using competitive growth of each mutant against the control population, as described in Materials and Methods.

Mutant	Low Bulk		High Bulk		# Mutations	G:C → A:T
	Mean[a]	IQR[b]	Mean[a]	IQR[b]		
YPW89	134.1	109-160	111.7	103-126	65	91.0%
YPW94	134.0	107-163	75.3	61-90	73	84.4%
YPW102	91.9	74-109	132.2	121-147	33	84.9%

Table D-2 Summary statistics for sequencing

[a]Mean coverage obtained after read alignment using genomecov tool from BEDTools. [b]Interquartile range of genome coverage.

Mutant	Mutation position		Mutation type			Phenotypic effect			Sequencing depth[d]		Mutation frequency	
	Chr.	Position	Gene	DNA	Protein	Mean[a]	Std Dev[b]	Sel Coef[c]	Low bulk	High bulk	Low bulk	High bulk
YPW89	IV	1347028	SSN2	C->T	Q971Stop	10.16%	-6.66%	0.140	121	83	0	0.22
YPW94	III	260366	TUP1	G->A	G696D	6.93%	-14.39%	0.045	152	73	0	0.55
YPW102	XVI	679727	ROX1	G->A	R12K	-4.05%	-8.89%	0.015	102	77	0.96	0.30

Table D-3 Properties of the three confirmed mutants

[a]Mean expression of single site mutant relative to wild type expressed as a percentage of change in fluorescence phenotype relative to wild type. [b]Standard deviation of expression phenotype of the single site mutant strain relative to the reference strain. [c]Selection coefficient was measured using competitive growth of each single site mutant against the control population, as described in Materials and Methods. [d]Number of sequencing reads overlapping the variable site in each bulk.

Oligo ID	Oligo Sequence 5'-3' (* indicates Phosphorothioate bound)
IS1_adapter.P5	A*C*A*CTCTTTCCCTACACGACGCTCTTCCGA*T*C*T
IS2_adapter.P7	G*T*G*ACTGGAGTTCAGACGTGTGCTCTTCCGA*T*C*T
IS3_adapter.P5+P7	A*G*A*TCGGAAG*A*G*C
IS4_indPCR.P5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT

Table D - S1 FACS based mapping oligonucleotide adapter sequences

Oligo ID	Oligo Sequence 5'-3' (Lowercase: Index barcode)	Barcode	Sample
indexing4	CAAGCAGAAGACGGCATACGAGATttgatccGTGACTGGAGTTCAGACGTGT	GGATCAA	YPW89.low
indexing5	CAAGCAGAAGACGGCATACGAGATatcttgcGTGACTGGAGTTCAGACGTGT	GCAAGAT	YPW94.low
indexing6	CAAGCAGAAGACGGCATACGAGATtctccatGTGACTGGAGTTCAGACGTGT	ATGGAGA	YPW102.low
indexing12	CAAGCAGAAGACGGCATACGAGATacttcaaGTGACTGGAGTTCAGACGTGT	TTGAAGT	YPW89.high
indexing13	CAAGCAGAAGACGGCATACGAGATtgatagtGTGACTGGAGTTCAGACGTGT	ACTATCA	YPW94.high
indexing14	CAAGCAGAAGACGGCATACGAGATgatccaaGTGACTGGAGTTCAGACGTGT	TTGGATC	YPW102.high
indexing19	CAAGCAGAAGACGGCATACGAGATgagattcGTGACTGGAGTTCAGACGTGT	GAATCTC	WT
indexing20	CAAGCAGAAGACGGCATACGAGATgagcatgGTGACTGGAGTTCAGACGTGT	CATGCTC	Mapping.Strain

Table D - S2 FACS based mapping indexing oligos and barcodes

Only the eight samples used in this study are shown. These eight samples were multiplexed with 16 other samples using the following barcodes: 1-TCGCAGG, 2-CTCTGCA, 3-CCTAGGT, 4-GGATCAA, 5-GCAAGAT, 6-ATGGAGA, 7-CTCGATG, 8-GCTCGAA, 9-ACCAACT, 10-CCGGTAC, 11-AACTCCG, 12-TTGAAGT, 13-ACTATCA, 14-TTGGATC, 15-CGACCTG, 16-TAATGCG, 17-AGGTACC, 18-TGCGTCC, 19-GAATCTC, 20-CATGCTC, 21-ACGCAAC, 22-GCATTGG, 23-GATCTCG, 24-CAATATG.

Oligo ID	Oligo Sequence 5'-3'
Indexed adapter 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNT
Indexed adapter 2	NNNNNNAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PCR primer 1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
PCR primer 2	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT

Underlined: barcode. Color: Same color shows complementary regions where annealing occurs during PCR.

Table D - S3 Tetrad-based mapping oligonucleotide adapter sequences

Barcode	Sample
ACCAGG	Y1
AAGGCC	Y39
TATTCG	Y54 x Y85
CGGAAC	Y85
ATACCT	Y89 x Y39
ACACGA	Y94 x Y39
CACATA	Y102 x Y85

Table D - S4 Tetrad-based mapping barcodes

Only the seven samples used in this study are presented. These seven samples were multiplexed with 14 other samples using the following barcodes: 1-ACCAGG, 2-AAGGCC, 3-TCTGAT, 4-CAAGTG, 5-TACGTT, 6-TATTCG, 7-CGGAAC, 8-ATACCT, 9-GTGCTG, 10-GGCGTA, 11-TGCACG, 12-CTACGC, 13-ACACGA, 14-CCGTAG, 15-GTAACA, 16-GTGTAT, 17-AGGTTC, 18-CACATA, 19-AGTTGG, 20-GCTCAA, 21-TTGACT, 22-TCTCGG.

Appendix E

Relationship between pleiotropy and fitness

Abstract

Pleiotropy is viewed as a restrictive phenomenon in evolution, preventing the fixation of mutations that have positive fitness effects on one phenotype due to negative fitness effects on other phenotypes. The degree of pleiotropy is expected to vary based on the molecular mechanism by which a mutation alters a phenotype. As a consequence, different molecular mechanisms are predicted to have different probabilities of underlying adaptation. In particular, mutations in protein coding regions are predicted to be more pleiotropic than regulatory regions and *trans*-regulatory mutations are predicted to be more pleiotropic than *cis*-regulatory mutations. Here we begin to test these hypotheses by comparing the degree of pleiotropy, as determined by the number of differentially expressed genes upon mutation, to the fitness effects of the same mutations. We find that more pleiotropic mutations are on average more deleterious. However, this relationship is weak, accounting for at most ~15% of the difference in fitness amongst mutants. In addition, initial estimates of pleiotropy and fitness for *cis* and *trans*-regulatory are presented.

Introduction

Determining the molecular mechanisms likely to contribute to adaptation under different evolutionary scenarios is a long standing goal of evolutionary biology. Numerous examples in which the molecular basis of adaptation is known, or at least strongly suspected, exist and there have been several attempts to identify larger evolutionary patterns amongst these cases (HOEKSTRA and COYNE 2007; WRAY 2007; STERN and ORGOGOZO 2008, 2009; STREISFELD and RAUSHER 2010; STERN 2013). From such studies, there have been two main conclusions. First, the molecular basis of adaptation varies considerably between systems with no clear dominant mechanism. Second, within a system, the molecular basis for adaptation often proceeds through similar mechanisms, and in many instances uses the same gene or mutation.

One contributor to these seemingly contradictory conclusions is the existence of pleiotropy, or the ability of mutations to simultaneously alter more than one phenotype: because most changes to a phenotype are deleterious, the existence of pleiotropy means that only a few mutations are expected to alter a phenotype in a beneficial way without also altering additional phenotypes in a negative way. As a consequence, the number of mutations that are strictly beneficial is limited and it is therefore likely that the same molecular mechanism will be used in multiple instances of adaptation (STERN 2013). Importantly, the structure of pleiotropy is expected to vary across systems, and thus the constraints and molecular mechanisms of adaptation are also expected to vary across systems.

To date, most studies have focused on measuring pleiotropy and fitness within natural systems. As a consequence, the independent effects of pleiotropy and fitness are confounded by natural selection. Instead, estimates of the effects of pleiotropy and fitness for new mutations in the absence of natural selection are needed. Initial estimates of the relationship between pleiotropy and fitness for new mutations suggested a relatively weak effect, with variation in pleiotropy amongst mutations altering yeast morphology accounting for only ~18% of the variation in fitness (COOPER *et al.* 2007). However, a more recent study suggests that selection against pleiotropic effects on gene expression is often quite strong (MCGUIGAN *et al.* 2014). The relationship between pleiotropy and fitness for new mutations is thus unclear.

Mutations that alter protein sequence and function are expected to be highly pleiotropic because they alter function regardless of context. By contrast, altering expression or regulation of the same protein is expected to be less pleiotropic because the activity of the protein is altered under only a specific set of conditions. As a consequence, protein coding changes are expected to contribute less often to adaptation than changes in gene expression and regulation (STERN 2000; CARROLL 2005). In addition to affecting the probability of adaptation through protein coding or regulatory changes, the specific molecular mechanisms by which expression and regulation are altered is also expected to be influenced by pleiotropy. Gene expression is controlled by interactions between *cis*-regulatory elements, such as promoters and enhancers, and diffusible *trans*-regulatory factors, such as transcription factors and small RNAs. Because *trans*-regulatory factors often interact with multiple *cis*-regulatory sequences, altering a gene's expression

through a *trans*-regulatory changes is more likely to be pleiotropic than altering a gene's expression through a *cis*-regulatory change (CARROLL 2005). However, direct evidence of differences in pleiotropy amongst different molecular mechanisms is missing.

Here we estimate the relationship between fitness and pleiotropy in the context of gene expression evolution. We estimate pleiotropy as the number of genes whose expression varies upon mutation and compare these estimates to the fitness costs of the same mutation. We first compare the pleiotropic and fitness consequences of gene deletions in the yeast *Saccharomyces cerevisiae*, finding a weak, but negative relationship. We next present initial finding on the pleiotropic and fitness consequences of *cis* and *trans*-regulatory mutations.

Results and Discussion

To estimate the relationship between pleiotropy and fitness for new mutations, we analyzed previously collected expression data from homozygous gene deletions in the yeast *S. cerevisiae*. For each deletion, we estimated pleiotropy as the number of differentially expressed genes relative to a control strain and then compared these estimates to competitive fitness assays of the deletion strain against a control strain in the same environment used for measuring expression. To estimate effects of each deletion on gene expression, we used two independent data sets, one from CHUA *et al.* 2006 and the second from HU *et al.* 2007, using a recent reanalysis of differential expression by REIMAND *et al.* 2010 for the latter. To estimate the fitness effects of each deletion strain we used data from STEINMETZ *et al.* 2002.

We found a negative, but not significant, correlation between pleiotropy and fitness for the Chua *et al*. data (R^2= 0.054, p = 0.12, Figure E-1A). Using robust linear regression, we estimated the effect of pleiotropy on fitness, finding that each differentially expressed gene reduced fitness by at most about ~0.0001. For the Reimand *et al.* data we observed a significant negative correlation between pleiotropy and fitness (R^2=0.16, p = 7 x 10^{-11}, Figure E-1B). Differences between the two data sets likely explain the discrepancy: the second data set included more genes than the first (238 vs 47) and these genes were on average more pleiotropic (p = 2 x10^{-8}, wilcoxon test). Regardless of these differences, robust regression of fitness on pleiotropy in the Reimand *et al.* data also suggested a fitness decrease of ~0.0001 per differentially expressed gene. While this effect of pleiotropy on fitness is small per gene, many deletions affected expression of hundreds of genes, suggesting the total fitness effect of pleiotropy is not negligible. Interestingly, the negative relationship between pleiotropy and fitness appears to be largely driven by a lack of highly pleiotropic mutants with high fitness and not a lack of lowly pleiotropic mutants with no fitness effects. This suggests that increased pleiotropy doesn't directly decrease fitness, but instead may impose a declining fitness ceiling with increasing pleiotropy.

To test if the observed negative relationship between pleiotropy and fitness for gene deletions also applied to point mutations, we used RNA-seq to measure expression from seven *S. cerevisiae* strains containing point mutations. Each mutant was known to significantly affect the expression of a yellow fluorescent protein (YFP) reporter under

control of the *TDH3* promoter. For each mutant, we estimated pleiotropy as the number of differentially expressed genes relative to a non-mutant reference strain. We then compared these estimates of pleiotropy to competitive fitness of each strain against a common reference strain. As with the previous data set, we found a weak, but significant, negative relationship between pleiotropy and fitness ($p=0.038$). Robust regression indicated a fitness cost of increasing the number of differentially expressed genes by one of 0.0001, which is similar in magnitude to the previous estimates (Figure E-2).

To determine whether there was a difference in pleiotropy between *cis* and *trans*-regulatory mutations, we classified each mutant based on the mechanism by which it influenced YFP expression (Table E-1, GRUBER *et al.* 2012). In total, four mutants contained *cis*-regulatory mutations. Of these mutants, one contained only a *cis*-regulatory mutation, while the remaining three strains contained additional *trans*-regulatory mutations. Two of these strains with both *cis* and *trans*-regulatory mutations contained the same *cis*-regulatory mutation as each other, but varied in their *trans*-regulatory mutations. Finally, three mutants contained only *trans*-regulatory mutations. We found no significant difference in the degree of pleiotropy between *cis* and *trans*-regulatory mutants ($p=0.63$, Wilcoxon). However, there is little power to detect a significant difference given the low number of strains tested.

Interestingly, we observed that isolating a single *cis*-regulatory mutation from additional *trans*-regulatory mutations substantially reduced pleiotropy. This could be because the single *cis*-regulatory mutation tested is located in the promoter of a reporter gene, and

thus unlikely to have effects on the expression of native yeast genes, while the *trans*-regulatory mutants can affect native yeast genes. In addition, the *trans*-regulatory mutants contained multiple mutations, while the *cis*-regulatory mutant contained only a single mutation. Consistent with the idea, the mutant with the lowest estimated pleiotropy was previously confirmed to affect YFP expression through a single *cis*-regulatory mutation in the reporter's promoter. Thus, additional work is needed to test the pleiotropic effects of *cis*-regulatory mutations at the native *TDH3* gene, test the pleiotropic effects of single *trans*-regulatory mutations, and to increase the number of mutants analyzed so that statistical measures of significance can be more appropriately applied.

Methods

For data from Chua *et al*. 2006, differentially expressed genes were called as those with an absolute z-score greater than 2 after Bonferroni multiple testing correction. For data from Reimand *et al*. 2010, differentially expressed genes were called using the provided sets of significantly up and down regulated genes. Fitness measurements from Steinmetz *et al*. 2002 were the average of two independent time course competitive fitness assays for each deletion mutant. For Chua *et al*. 2006, four gene deletions were excluded due to lack of fitness data. For Reimand *et al*. 2010, 31 gene deletion were excluded due to lack of fitness data.

cis and *trans*-regulatory mutants were taken from Gruber et al. 2012 and express yellow fluorescent protein from a *TDH3* driven reporter construct (GRUBER *et al.* 2012). Each strain was grown in YPD (10g yeast extract, 20g peptone, 20g dextrose per 1 L of water)

overnight at 30°C with shaking in duplicate and then diluted 1/10 and allowed to grow for four hours in the same conditions. RNA was extracted using the RiboPure Yeast Kit from Ambion. Illumina sequencing libraries were created using the TruSeq RNA preparation guide. Barcodes and samples are listed in Table E-2. 50 bp single end sequencing was performed by the University of Michigan sequencing center.

To detect differentially expressed genes, all reads were aligned to the S288c reference genome using bowtie (LANGMEAD *et al.* 2009) and processed using SAMtools (LI *et al.* 2009). Counts per open reading frame were produced using Bedtools (QUINLAN and HALL 2010). A custom script was then used to enumerate all possible non-redundant groupings of the eight strains into up to three groups. These three groups represent distinct levels of gene expression and the complete set of enumerated models represent all potential differential expression patterns amongst the strains. This approach assumes that most genes are not differentially expressed amongst all strains. As a consequence, information about the variability in expression for a gene is shared amongst all strains, effectively increasing the number of replicates and power to detect when a gene for one or a few strains have differential expression. For each model, we estimated posterior probabilities using bayseq (HARDCASTLE and KELLY 2010) and a gene was called as differentially expressed when the Bayes factor for the set of models with expression different from the control strain for that gene was greater than 1 in a particular mutant. The total number of genes estimated as differentially expressed was used to estimate the degree of pleiotropy for each mutant.

Fitness was measured by competing each strain, including the control strain with no mutations, against a common reference strain marked with green fluorescent protein (GFP, BY4741 SWH1::p*TDH3*-GFP). All strains were independently grown for more than 24 hours at 30°C with shaking in YPD and then combined in equal numbers with the GFP reference strain. 10 replicates of 2 ul of each mix was inoculated into 0.5 ml of YPD. For each competition, the number of GFP and YFP positive cells was counted using an Accuri C6 flow cytometer. Competitions were conducted for 18 hours at 30°C in YPD with shaking. The number of GFP and YFP cells was counted again using flow cytometry, and fitness calculated based on the difference in ratio of YFP/GFP cells observed from the initial time point to the later time point. In each instance, approximately 50,000 events were recorded. To detect positive GFP and YFP cells, pure cultures positive for each fluorescent protein were independently measured using the same procedure and used to determine counting gates. To improve separation of GFP and YFP fluorescence signals, a 540/20 band pass filter was placed in FL2 instead of the normal 585/40 filter.

References

CARROLL S. B., 2005 Evolution at two levels: on genes and form. PLoS Biol. **3**: e245.

CHUA G., MORRIS Q. D., SOPKO R., ROBINSON M. D., RYAN O., CHAN E. T., FREY B. J., ANDREWS B. J., BOONE C., HUGHES T. R., 2006 Identifying transcription factor functions and targets by phenotypic activation. PNAS **103**: 12045–50.

COOPER T. F., OSTROWSKI E. A., TRAVISANO M., 2007 A negative relationship between mutation pleiotropy and fitness effect in yeast. Evolution (N. Y). **61**: 1495–9.

GRUBER J. D., VOGEL K., KALAY G., WITTKOPP P. J., 2012 Contrasting Properties of Gene-specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects and Dominance. PLoS Genet. **8**: e1002497.

HARDCASTLE T. J., KELLY K. a, 2010 baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics **11**: 422.

HOEKSTRA H., COYNE J., 2007 The locus of evolution: evo devo and the genetics of adaptation. Evolution (N. Y). **61**: 995–1016.

HU Z., KILLION P. J., IYER V. R., 2007 Genetic reconstruction of a functional transcriptional regulatory network. Nat. Genet. **39**: 683–687.

LANGMEAD B., TRAPNELL C., POP M., SALZBERG S. L., 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**: R25.

LI H., HANDSAKER B., WYSOKER A., FENNELL T., RUAN J., HOMER N., MARTH G., ABECASIS G. R., DURBIN R., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

MCGUIGAN K., COLLET J. M., ALLEN S. L., CHENOWETH S. F., BLOWS M. W., 2014 Pleiotropic Mutations Are Subject to Strong Stabilizing Selection. Genetics: 1051–1062.

QUINLAN A. R., HALL I. M., 2010 BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics **26**: 841–842.

REIMAND J., VAQUERIZAS J. M., TODD A. E., VILO J., LUSCOMBE N. M., 2010 Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. Nucleic Acids Res. **38**: 4768–4777.

STEINMETZ L. M., SCHARFE C., DEUTSCHBAUER A. M., MOKRANJAC D., HERMAN Z. S., JONES T., CHU A. M., GIAEVER G., PROKISCH H., OEFNER P. J., DAVIS R. W., 2002 Systematic screen for human disease genes in yeast. Nat. Genet. **31**: 400–404.

STERN D. L., 2000 Perspective: Evolutionary Developmental Biology and the Problem of Variation. Evolution (N. Y). **54**: 1079.

STERN D. L., 2013 The genetic causes of convergent evolution. Nat. Rev. Genet. **14**: 751–764.

STERN D. L., ORGOGOZO V., 2008 The loci of evolution: how predictable is genetic evolution? Evolution (N. Y). **62**: 2155–2177.

STERN D. D. L., ORGOGOZO V., 2009 Is genetic evolution predictable? Science **323**: 746–751.

STREISFELD M. a., RAUSHER M. D., 2010 Population Genetics, Pleiotropy, and the Preferential Fixation of Mutations During Adaptive Evolution. Evolution (N. Y). **65**: 1–14.

WRAY G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. **8**: 206–216.

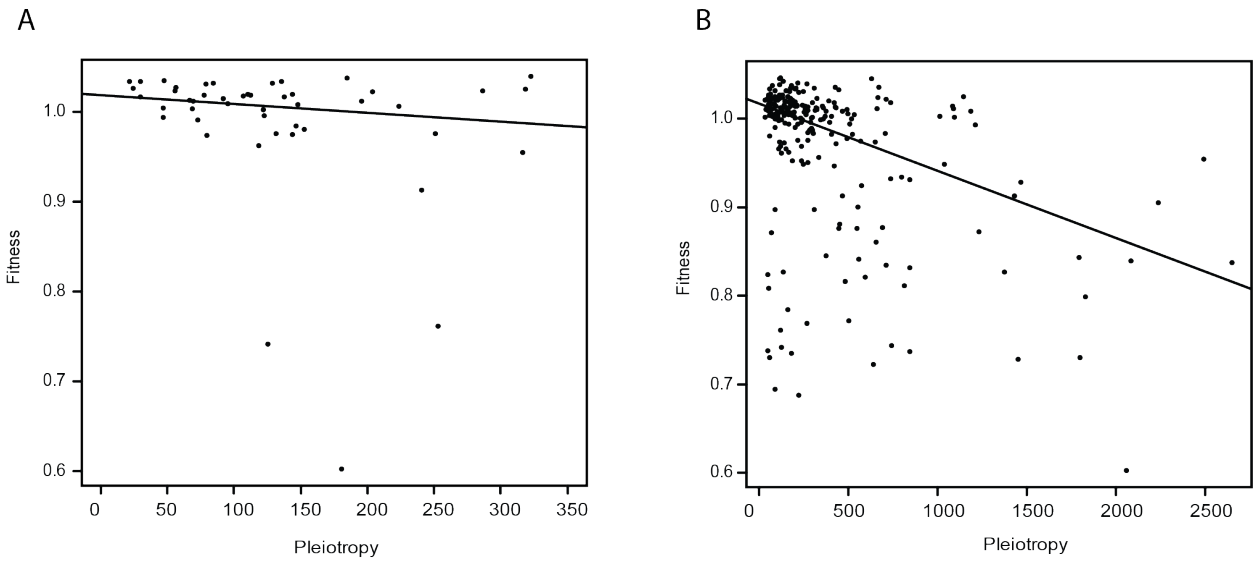


Figure E-1 Relationship between pleiotropy and fitness for gene deletions

Pleiotropy was measured as the number of differentially expressed genes upon single gene deletion in S. cerevisiae. Fitness was measured by competitive fitness assays between a gene deletion strain and a control strain. Solid line shows fit from robust linear regression.

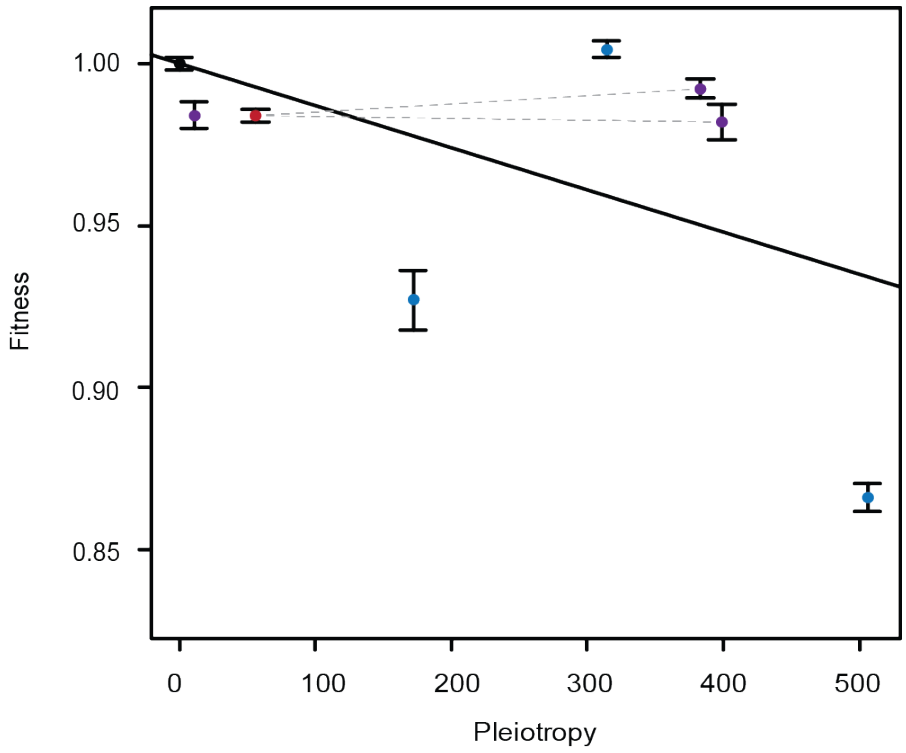


Figure E-2 Relationship between pleiotropy and fitness for regulatory mutations

Pleiotropy was measured as the number of differentially expressed genes in mutants of *S. cerevisiae* relative to a reference strain. Fitness was measured by competitive fitness assays between each strain and a common reference strain. Red: *cis*-regulatory mutant. Blue: *trans*-regulatory mutants. Purple: *cis* and *trans*-regulatory mutant. Solid line shows fit from robust linear regression. Error bars show 95% CI on fitness estimate. Gray dashed lines connect the *cis*-regulatory mutant to two mutants that contain the same *cis*-regulatory mutation plus additional *trans*-regulatory mutations.

Strain	Class	*cis*-mutation position	Pleiotropy	Average Fitness
YPW1	Reference	NA	0	1.000
YPW60	*cis*	-255	56	.984
YPW50	*cis* + *trans*	-255	399	.982
YPW51	*cis* + *trans*	-240	11	.984
YPW196	*cis* + *trans*	-255	383	.992
YPW258	*trans*	NA	506	.866
YPW255	*trans*	NA	314	1.00
YPW198	*trans*	NA	172	.927

Table E-1 Properties of the *S. cerevisiae* strains containing point mutations

Strain	Barcode Sequence
YPW1-1	TTGACT
YPW1-2	TATTCG
YPW60-1	TCTGAT
YPW60-2	TGCACG
YPW50-1	ACACGA
YPW50-2	ACCAGG
YPW51-1	AGTTGG
YPW51-2	AGGTTC
YPW196-1	CTACGC
YPW196-2	CAAGTG
YPW258-1	CACATA
YPW258-2	CCGTAG
YPW255-1	GTAACA
YPW255-2	GTGTAT
YPW198-1	GCTCAA
YPW198-2	GGCGTA

Table E-2 Illumina barcode sequences used for each sample

Appendix F

Supplementary Figures and Tables for Chapter 5

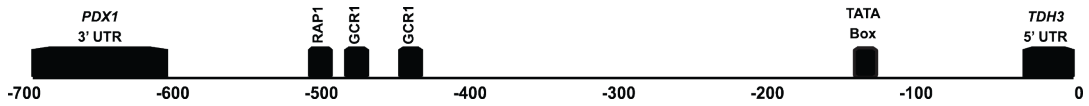


Figure F-1 Position of regulatory elements within the TDH3 promoter.

Regulatory element names are listed above each box. Bo width is proportional to regulatory element size . Numbers indicate bases upstream of start codon.

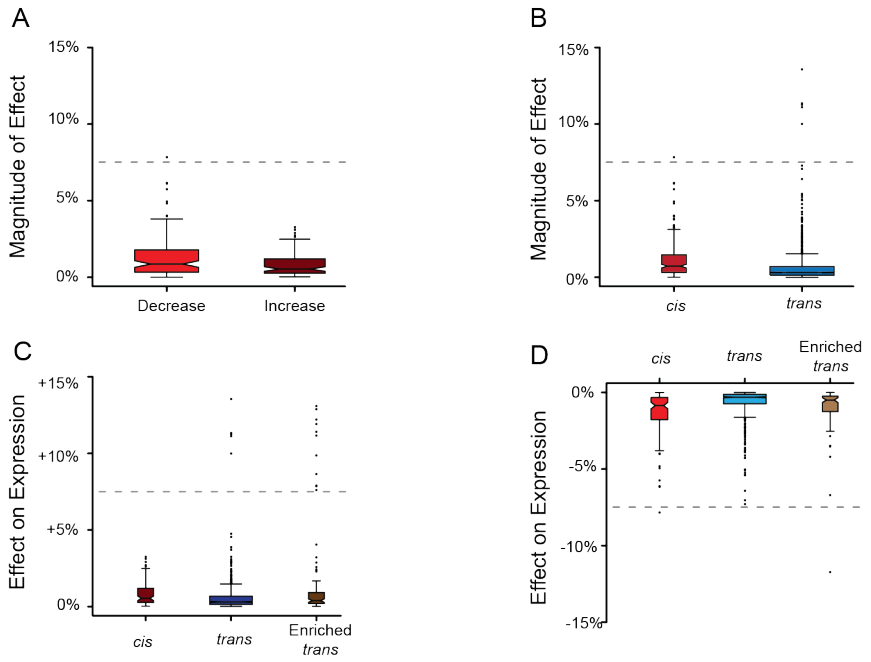


Figure F-2 Effects on expression for *cis*-regulatory mutations outside of known TFBS

(A) Magnitude of effect for *cis*-regulatory mutations outside of TFBS that increase (dark red) vs decrease (light red) P_{TDH3}-YFP fluorescence. (B). Magnitude of effect on P_{TDH3}-YFP fluorescence for *cis*-regulatory mutations outside of TFBS (red) vs *trans*-regulatory mutations (blue). (C). Effect on P_{TDH3}-YFP fluorescence for *cis*-regulatory mutants outside of known TFBS (dark red), *trans*-regulatory mutants (dark blue), and *trans*-regulatory mutants enriched for extreme effects (dark brown) for mutants with increased P_{TDH3}-YFP fluorescence. (D) Effect on P_{TDH3}-YFP fluorescence for *cis*-regulatory mutants outside of TFBS (light red), *trans*-regulatory mutants (light blue), and *trans*-regulatory mutants enriched for extreme effects (light brown) for mutants with decreased P_{TDH3}-YFP fluorescence.

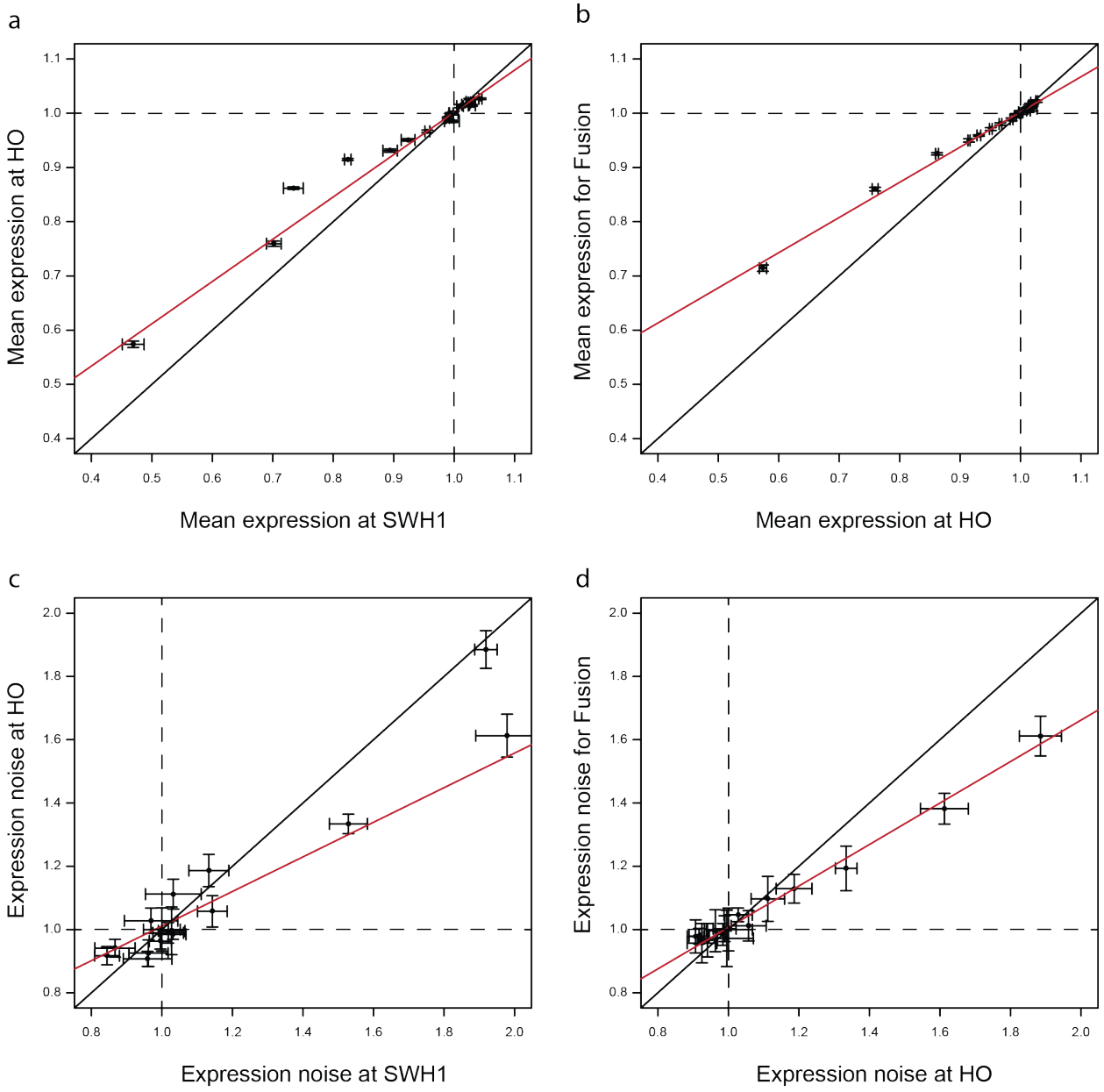


Figure F-3 Effects on reporter expression due to genomic location and background

Black solid lines are the y=x line. Dashed lines are the non-mutant control expression for each reporter. Red solid line is the slope from a linear regression. Error bars are 95% CI. (A) Effect of *cis*-regulatory mutations on P_{TDH3}-YFP fluorescence for a reporter near the SWH1 pseudogene (x-axis) vs the effect of the same *cis*-regulatory mutations on P_{TDH3}-YFP fluorescence for a reporter at the HO locus and in a second genetic background. (B) Effect of *cis*-regulatory mutations on P_{TDH3}-YFP fluorescence for a reporter at the HO locus (x-axis) vs the effect of the same *cis*-regulatory mutations on YFP fluorescence for a fusion protein of native *TDH3* to YFP in the same genetic background. (C) Effect of *cis*-regulatory mutations on P_{TDH3}-YFP fluorescence noise for a reporter near the SWH1 pseudogene (x-axis) vs the effect of the same *cis*-regulatory mutations on P_{TDH3}-YFP fluorescence noise for a reporter at the HO locus and in a second genetic background. (D) Effect of *cis*-regulatory mutations on P_{TDH3}-YFP fluorescence noise for a reporter at the HO locus (x-axis) vs the effect of the same *cis*-regulatory mutations on YFP fluorescence noise for a fusion protein of native *TDH3* to YFP in the same genetic background.

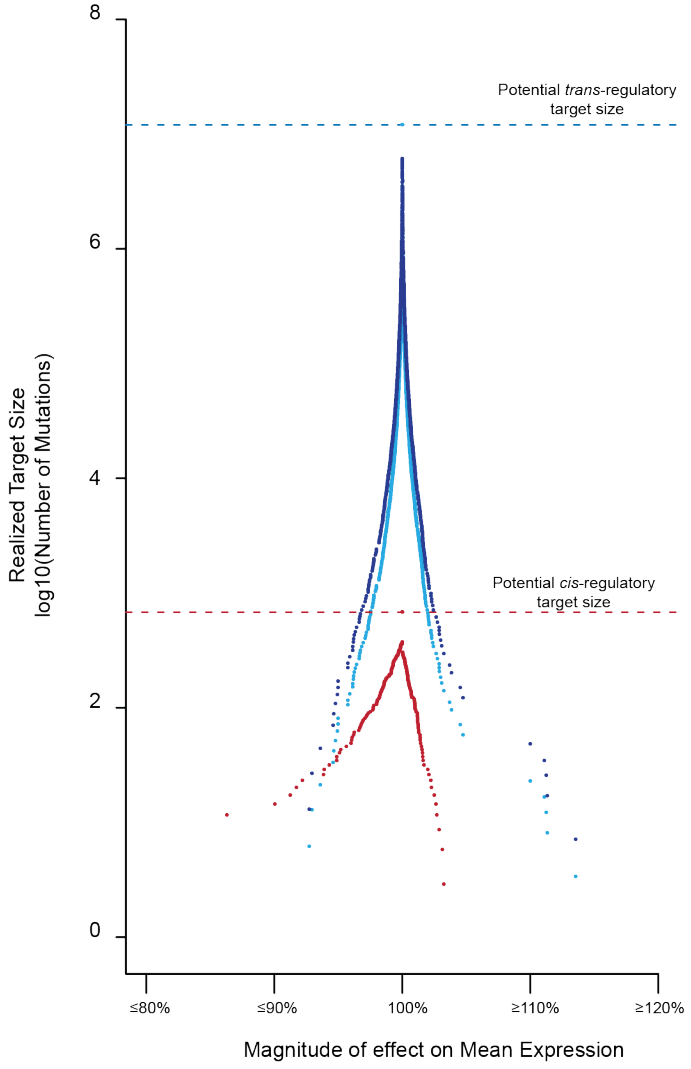


Figure F-4 Effect of estimated number of mutations in *trans*-regulatory mutants on frequency

Number of estimated bases in the *S. cerevisiae* genome (y-axis) that when mutated are expected to result in a change in expression equal to, or more extreme than, a specific cutoff (x-axis). *cis*-regulatory mutations (red). *trans*-regulatory mutations with N=21 mutations per *trans*-regulatory mutant (light blue). This corresponds to the lower-bound of the 95% confidence interval on the number of *trans*-regulatory mutations per mutant. *trans*-regulatory mutations with N=43 mutations per *trans*-regulatory mutant (dark blue). This corresponds to the upper-bound of the 95% confidence interval on the number of *trans*-regulatory mutations per mutant.

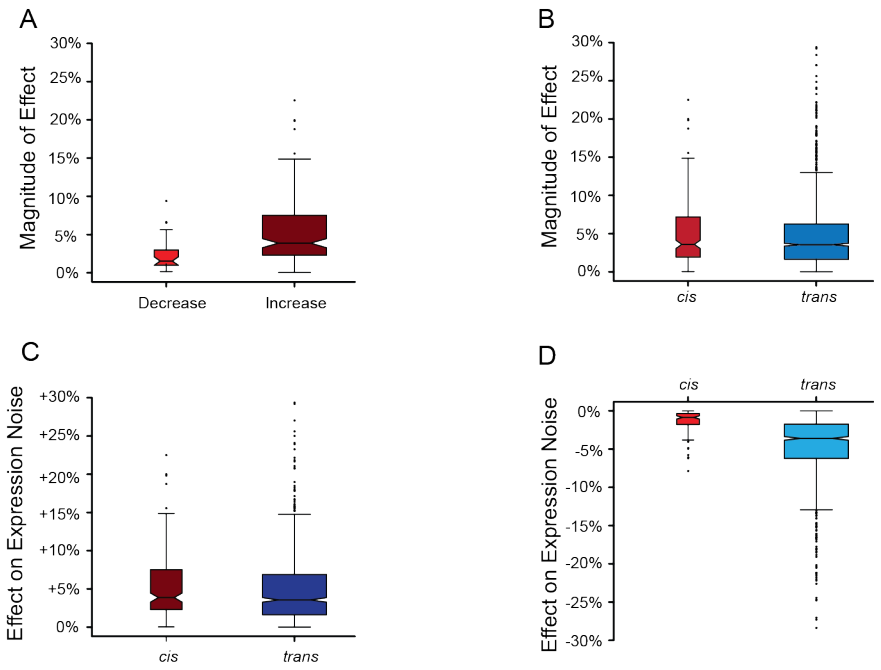


Figure F-5 Effects on expression noise for *cis*-regulatory mutations outside of known TFBS

(A) Magnitude of effect for *cis*-regulatory mutations outside of TFBS that increase (dark red) vs decrease (light red) P_{TDH3}-YFP fluorescence noise. (B). Magnitude of effect on P_{TDH3}-YFP fluorescence noise for *cis*-regulatory mutations outside of TFBS (red) vs *trans*-regulatory mutations (blue). (C). Effect on P_{TDH3}-YFP fluorescence noise for *cis*-regulatory mutants outside of known TFBS (dark red), *trans*-regulatory mutants (dark blue), and *trans*-regulatory mutants enriched for extreme effects (dark brown) for mutants with increased P_{TDH3}-YFP fluorescence noise. (D) Effect on P_{TDH3}-YFP fluorescence noise for *cis*-regulatory mutants outside of TFBS (light red), *trans*-regulatory mutants (light blue), and *trans*-regulatory mutants enriched for extreme effects (light brown) for mutants with decreased P_{TDH3}-YFP fluorescence noise.

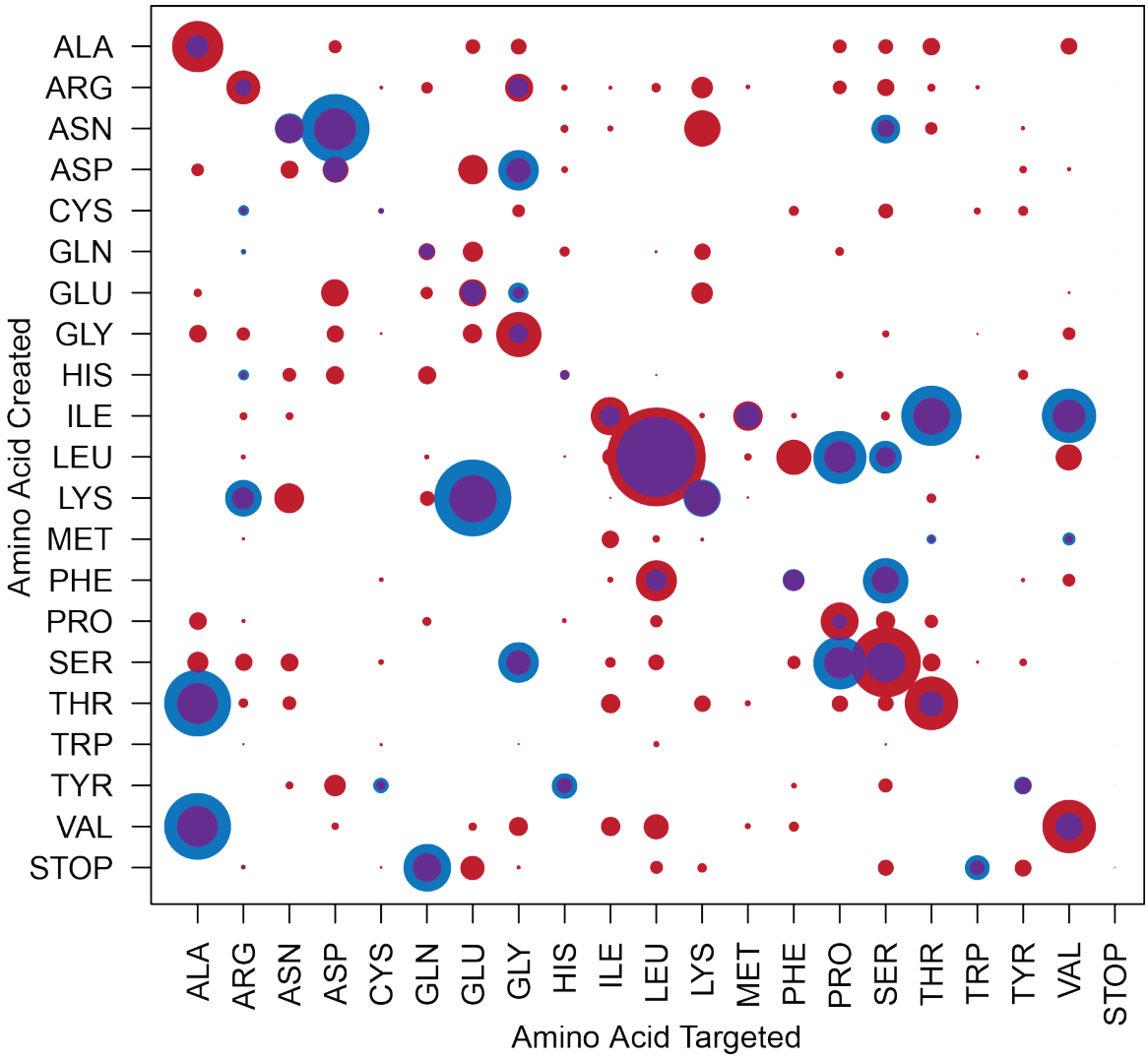


Figure F-6 Expected frequency of amino acid changes due to EMS mutagenesis

Expected frequency of EMS mutations vs natural mutations on amino acid changes. Frequency of changes from one amino acid (x-axis) to a second amino acid (y-axis) for EMS like GC→AT transitions (blue) vs all point mutations (red). Circles are proportional to frequency.

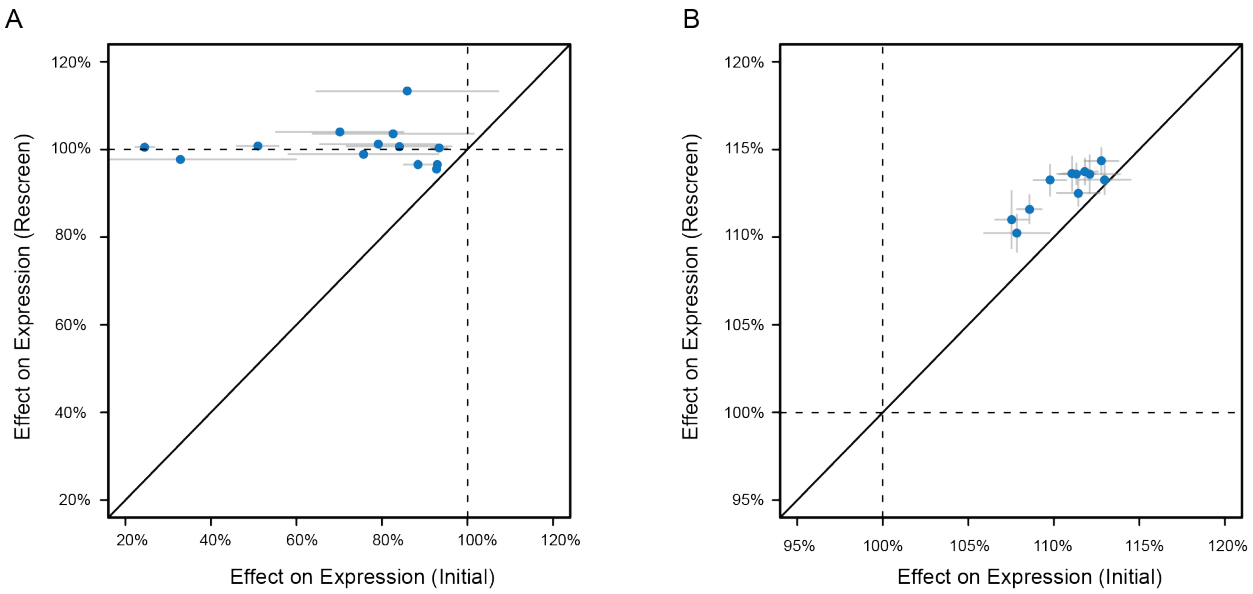


Figure F-7 Consistency of estimated effects differs for decreases and increases in expression

Estimated effect of *trans*-regulatory mutants that either decrease (A), or increase (B) expression by less than or more than 7.5% in the initial screen respectively. Gray bars are 95% confidence intervals on the mean (blue dots). A. Mutants with large decreases in expression had large variability amongst replicates in the initial screen and were removed from analysis (x-axis). Consistent with this decision, these mutants did not confirm in a follow-up analysis (y-axis). B. By contrast, mutants with large increases in expression had consistent measurements between replicates and were retained in the analysis. Consistent with this decision, these mutants had reproducible effects on expression.

	Mean Expression Relative to WT (%)						Expression Noise Relative to WT (%)					
	SWH1		HO		Fusion		SWH1		HO		Fusion	
Strain	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
m75	68.45	1.22	75.94	0.50	97.72	0.38	209.24	3.55	188.51	6.08	139.83	5.57
m221	98.51	0.54	101.32	0.31	114.18	0.37	108.59	3.48	96.32	2.56	86.41	5.88
m123	96.52	0.30	99.12	0.18	113.03	0.26	111.00	4.85	98.95	3.31	86.99	3.62
m89	90.13	1.14	95.08	0.23	110.32	0.29	112.57	8.81	111.18	4.83	95.18	6.29
m90	87.23	1.19	93.10	0.31	108.97	0.16	123.55	6.29	118.63	5.19	97.96	4.02
m153	100.11	0.65	102.11	0.41	115.80	0.28	104.88	6.17	92.52	4.30	83.02	5.52
m154	100.43	0.58	101.43	0.14	115.14	0.18	104.61	7.68	90.72	2.43	84.87	4.64
m63	80.39	0.55	91.47	0.18	107.91	0.35	166.72	6.01	133.40	3.14	103.52	6.25
m66	71.63	1.61	86.18	0.25	105.11	0.26	215.78	9.92	161.27	6.91	119.90	4.30
m156	99.42	0.65	101.56	0.21	115.27	0.43	94.53	6.37	94.09	2.87	83.84	4.67
CB	97.51	0.73	100.02	0.35	113.78	0.42	109.05	5.82	100.00	6.96	86.77	5.98
m142	97.06	0.35	99.97	0.16	113.67	0.27	112.46	4.11	98.45	1.60	85.43	3.04
m184	100.01	0.18	101.97	0.74	114.57	0.50	112.04	4.21	99.57	7.69	84.33	7.87
m160	97.20	1.23	98.51	0.23	112.35	0.28	105.66	8.43	102.76	4.08	90.79	1.91
m91	93.23	0.41	96.66	0.23	111.36	0.25	124.65	4.68	105.74	5.05	87.79	4.25
V	101.80	0.27	102.67	0.16	116.14	0.26	92.07	3.91	91.72	2.93	84.94	2.18
m76	45.78	1.79	57.38	0.61	81.15	0.68	554.18	181.66	320.04	64.24	212.84	8.59

Table F-1 Effects of mutations on reporter expression at alternative genomic positions

Mean (μ) and standard deviation (σ) of YFP florescence for each of 17 *TDH3* promoter haplotypes for their effect on both mean expression and expression noise. YFP fluorescence for all haplotypes was measured from three separate locations in different genetic backgrounds. All effects are expressed as percentages relative to a strain with the same construct at the same genomic location, but containing no mutations.