Pleiotropic Effects of *Cis*- and *Trans*-Regulatory Mutations

by

Pétra L Vande Zande

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Molecular, Cellular, and Developmental Biology)
in the University of Michigan
2021

Doctoral Committee:

Professor Patricia J. Wittkopp, Chair
Professor Anuj Kumar
Assistant Professor Cora MacAlister
Professor Jianzhi Zhang

Petra L Vande Zande

pvzande@umich.edu

ORCID iD: 0000-0002-4789-9899

Dedication

To Jesus Christ, the Author of Life

"He is before all things, and in Him all things hold together."

Colossians 1:17

Acknowledgements

I would first like to thank Trisha Wittkopp for caring deeply for her trainees, for devoting a significant amount of time to my training, and for calling me to a higher standard of research and writing. I would also like to thank: Fabien Duveau for 'teaching me the ropes' of research in the Wittkopp lab and whose excitement and enthusiasm for research is truly contagious; Brian Metzger and other members of the Wittkopp lab who laid the foundation, both intellectually and experimentally, for the research presented here; all members of the Wittkopp lab, past and present, for letting me talk your ears off, giving excellent and frequent feedback, and creating an atmosphere of encouragement and support; my committee members Anuj Kumar, Cora MacAlister, and George Zhang for your straightforward and very helpful discussions of the work in this thesis; members of the MB2 neighborhood for creating a congenial atmosphere in which to do research and for helpful conversations; the Genetics Training Program for financial support and the opportunity to interact with a wide variety of excellent scientists; the PIBS program and MCDB department for financial and intellectual support; the MCDB PhD cohort of 2015 for support and camaraderie while figuring out this whole grad school thing together; Dr. Eric Michael Bratsolias Brown for telling me I could get paid in grad school and showing me how to apply; and finally, to my family and friends for their wholehearted support and encouragement through all the highs and lows of a PhD program.

Table of Contents

List of Tables

List of Figures

Abstract

Changes in gene expression are an important source of phenotypic diversity both within and between species. Mutations generating variation in gene expression can be *cis*-regulatory to a particular gene, which typically occur in promoters or enhancers and cause allele-specific changes in expression, or *trans*-regulatory to the gene, which are mediated by diffusible factors and thus do not result in allele specific changes in expression. *Cis*-regulatory mutations are hypothesized to be less pleiotropic, or impact fewer traits, than *trans*-regulatory mutations to the same gene. Also, mutations that are more pleiotropic are hypothesized to more frequently be deleterious than those that are less pleiotropic. Thus, lower pleiotropy could contribute to a preferential fixation of *cis*-regulatory mutations relative to *trans*-regulatory mutations over time. Here I test these hypotheses by examining the fitness effects and genome-wide effects on gene expression of *cis*- and *trans*-regulatory mutations in the baker's yeast *Saccharomyces cerevisiae*. I first use RNA-sequencing data and fitness data for strains of bearing *cis*- and *trans*-regulatory mutations to the gene *TDH3* to define a distribution of the pleiotropic fitness effects of *trans*-regulatory mutations relative to *cis*-regulatory mutations and show that most pleiotropic fitness effects are indeed detrimental. I then compare the extents of the mutations' impacts on genome-wide gene expression and show that *trans*-regulatory mutations have a more widespread impact on gene expression than *cis*-regulatory mutations of similar effect size on the focal gene *TDH3*. In addition, I show that *trans*-regulatory mutations have pleiotropic effects on expression of genes affected by changes in the expression of the focal gene itself. I next use gene expression

data for a set of of ~1400 gene deletion strains of *Saccharomyces cerevisiae* to compare the genome-wide impacts on gene expression of *cis*- and *trans*-acting deletions for all genes in the dataset, and find that for the vast majority of genes, *trans*-acting deletions have more widespread effects on gene expression, or are more pleiotropic, than *cis*-acting deletions. Furthermore, this pattern can be explained by the degree distribution of the regulatory network resulting in highly pleiotropic *trans*-regulatory factors serving as *trans*-regulatory deletions to many genes. Finally, I return to the RNA-sequencing data to explore a mechanism of active compensation for reduction in *TDH3* expression by its paralog *TDH2* that is dependent on the *trans*-regulators Gcr1p and Rap1p. This compensation occurs when *TDH3* expression is lowered via *cis*-regulatory mutations, but not when it is lowered via mutations in these *trans*-regulators, resulting in the different downstream effects of changing *TDH3* expression in *cis* and in *trans*. Together these analyses provide the first empirical description of the pleiotropic effects of *cis*- and *trans*-regulatory mutations on both fitness and gene expression.

Chapter 1 Introduction[1]

Gene expression variation in evolution and disease

The regulation of gene expression is a critical step in translating genotypes into phenotypes. Variation in this regulation is common within and between species (Zheng et al. 2011) and contributes to trait diversity. For example, changes in the regulation of gene expression have been shown to contribute to divergent pigmentation in plants and animals (Kronforst et al. 2012; Wessinger and Rausher 2012), polymorphic body size in mice (Oliver et al. 2005), the sporulation rate in domesticated yeast (Deutschbauer and Davis 2005), and many other morphological, physiological and behavioral traits (Martin and Orgogozo 2013; Courtier-Orgogozo et al. 2020), including disease states in humans (Albert and Kruglyak 2015). Understanding how regulatory variation arises and evolves is thus critical for understanding many aspects of biology.

Genetic variation that affects the activity of regulatory networks underlies variation in gene expression. These networks include interactions among proteins, RNAs and DNA sequences. Transcription factor proteins and DNA sequences such as enhancers and promoters are most often considered to define the structure of gene regulatory networks (Babu et al. 2004; Yu and Gerstein 2006), but protein–protein interactions, signaling pathways and even metabolic states can also have an impact on their activity (Flint and Ideker 2019). Mutations that alter any of these elements can give rise to variation in gene expression. Such mutations can be classified

[1] Sections of this chapter are published as: Hill MS*, Vande Zande P*, Wittkopp PJ. 2021. Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* 22:203–215. *Equal contributions

as either *cis*- or *trans*- acting (Rockman and Kruglyak 2006): *cis*-acting mutations alter expression of a gene located on the same chromosome and tend to be located close to the affected gene, whereas *trans*-regulatory mutations have effects on gene expression that are mediated by diffusible molecules (such as RNAs and proteins) and can be located anywhere in the genome. Both types of mutation contribute to variation in gene expression, but differences in their molecular mechanisms suggest that they might contribute unequally to regulatory variation over evolutionary time.

Genomic studies describing variation in gene expression and the relative contributions of *cis*- and *trans*- acting variants have now been performed for diverse plant, animal and microbial species (Signor and Nuzhdin 2018). As with all traits, this variation reflects the introduction of new genetic variants by mutation, the filtering of these variants by natural selection and the chance survival of variants mediated by genetic drift. The extent to which each of these processes shapes the variation we see in wild populations, however, remains difficult to discern. For example, if one gene shows more variation in its expression than another, this might be because expression of the first gene is under less selective constraint or because a greater fraction of new mutations alters its expression (among other possibilities). Likewise, the patterns observed in *cis*- and *trans*-regulatory expression divergence within and between species could be the result of differences in the frequencies at which *cis*- and *trans*-regulatory mutations influence expression of a gene and/or differences in the selective constraints that exist for each category.

Pleiotropy in adaptive evolution

A mutation's degree of pleiotropy, or the number of traits it impacts, is thought to increase the probability of that mutation being deleterious, potentially imposing a strong selective constraint on adaptative evolution (Kimura and Ohta 1974). Fisher's geometric model

of adaptive evolution provides an intuitive explanation for why this may be. If a trait is conceptualized as a single dimension in phenotypic space, a mutation of a particular effect size on that trait can either move the phenotype toward or away from an optimum with equal probability. In contrast, a mutation that affects two traits, or moves the phenotype along two dimensions, has less than a 50% chance of moving closer to an optimum (Fig. 1-1). The more traits, or dimensions, that are simultaneously influenced by a mutation, the smaller the probability that the resulting phenotype will be closer to the optimum (Fisher 1930; Pavlicev and Wagner 2012). This led to the hypothesized 'cost of complexity,' or the idea that the more complex an organism is, the slower adaptation becomes, as smaller mutational steps have a higher proportion of potentially advantageous phenotypes in a multidimensional trait space (Orr 2000).

This theoretical work generated much interest in empirically quantifying how pleiotropic mutations typically are in order to understand how much of a constraint on adaptive evolution pleiotropy may actually impose (Wagner and Zhang 2011; Kinsler et al. 2020). However, quantifying pleiotropy by some absolute metric has proven challenging. Distinct organismal traits upon which a mutation acts are difficult to define, as traits can be correlated for genetic, mechanical, or morphological reasons (Stearns 2010). Furthermore, a quantification of the number of traits influenced by a mutation will always be limited by which traits are measured in any particular study (Paaby and Rockman 2013). Measuring the relative degrees of pleiotropy among different mutations for which the relevant traits have been carefully defined is less problematic, and in some cases a relationship between a specific description of pleiotropy and fitness has been empirically demonstrated (Featherstone and Broadie 2002; He and Zhang 2006; Cooper et al. 2007).

The pleiotropy of mutations influencing gene expression can also be described in multiple ways. In the case of a multicellular organism, mutations that influence expression of a gene in a subset of tissues, organs, or developmental stages can be considered less pleiotropic than those that influence expression in all tissues where the gene is typically expressed (Carroll 2005; Wray 2007). For a unicellular organism, the number of environmental conditions under which a mutation influences expression of a gene could similarly be considered less pleiotropic than mutations that influence expression in all environments (Dudley et al. 2005). In both of these cases, mutations that occur in non-coding elements that are specific to a developmental stage, tissue, or environment may be less pleiotropic than those in coding regions that affect the gene's function wherever it is expressed, potentially making noncoding mutations important sources of morphological variation (Stern and Orgogozo 2008). However, new studies suggest that many noncoding regulatory elements may be more pleiotropic than previously appreciated (Preger-Ben Noon et al. 2018; Sabarís et al. 2019) and systematic investigations of how this type of pleiotropy may relate to fitness are lacking.

The pleiotropy of regulatory mutations can also be defined as the number of genes for which expression is affected by a mutation. A negative relationship between this type of pleiotropy and fitness has been observed (Featherstone and Broadie 2002). In this case, mutations that occur in a gene that is less connected to other genes in various cellular networks such as metabolic, transcriptional, or protein-protein interaction networks might be considered less pleiotropic than a gene that is more connected if the level of connectivity directly translates to the number of genes that will be differentially expressed. For example, a *trans*-regulatory mutation influencing expression of a gene via a highly connected transcription factor might influence expression of more genes than a *cis*-regulatory mutation at the gene itself. This

difference in pleiotropy could in turn result in a more severe constraint on mutations changing the expression level or sequence of highly connected genes. Studies have failed, however, to find a relationship between a gene's connectivity in transcriptional regulatory networks and expression divergence (Kopp and McIntyre 2012; Yang and Wittkopp 2017), or connectivity in protein-protein interaction networks and divergence (Siegal et al. 2007). The incongruence of these findings with the observed relationship between the fitness and pleiotropy described above may result from a lack of a direct relationship between connectivity in these networks and the number of genes that are differentially expressed upon mutation (Flint and Ideker 2019), or a lack of a direct relationship between the immediate fitness impact of a mutation and its ultimate expression or sequence divergence between species. Empirical data directly assessing trends in pleiotropy and fitness of *cis*- and *trans*-regulatory mutations, in conjunction with known patterns in their contributions to expression divergence is needed to form a comprehensive picture of the role pleiotropy might play in expression divergence.

Patterns of expression divergence and regulatory variation

Distinguishing between *cis*- and *trans*- regulatory variation reveals patterns in their contributions to expression divergence. Two general strategies have primarily been used to disentangle the effects of *cis*- and *trans*- regulatory variants on a genomic scale. The first approach uses allele- specific expression (ASE) in F1 hybrids to compare the activity of *cis*-regulatory alleles in a common *trans*-regulatory background with expression in the parents of the F1 hybrid (Wittkopp et al. 2004). The second strategy uses statistical associations between genetic variants and gene expression to identify quantitative trait loci affecting gene expression (eQTLs) (Brem et al. 2002; Schadt et al. 2003). These two approaches provide complementary information about *cis*- and *trans*-regulatory variation, with the first capturing the net effect of all

cis- and *trans*-regulatory variants and the second providing information about the effects of individual loci. Studies using ASE to estimate the relative contributions of *cis*- and *trans*-regulatory variants to variation in gene expression have been conducted in various taxa, including plants (Springer and Stupar 2007; Zhang and Borevitz 2009; Shi et al. 2012; Bell et al. 2013), yeast (Wang et al. 2007; Sung et al. 2009; Emerson et al. 2010; Metzger et al. 2017), mice (Goncalves et al. 2012; Mack et al. 2016), birds (Davidson and Balakrishnan 2016; Wang et al. 2017), wasps (Wang et al. 2016) and flies (Wittkopp et al. 2008; McManus et al. 2010; Suvorov et al. 2013; Coolon et al. 2014). These studies include analysis of gene expression among individuals from outbred populations, between more isolated strains of the same species and between species. Each of these comparisons captures the evolution of gene expression at a different stage in the evolutionary process. Within species, *trans*-regulatory variants seem to contribute more to variation in gene expression than *cis*-regulatory variants (Emerson et al. 2010; Schaefke et al. 2013; Chen et al. 2015; Metzger et al. 2017; Signor and Nuzhdin 2018). This pattern has been suggested to be due to a larger mutational target size for *trans*-regulatory variants (Wittkopp 2005): that is, there are more places in the genome where a mutation can affect a gene's expression in *trans* than in *cis*. *trans*- acting variants are also often assumed to affect expression of more genes, on average, than *cis*- acting variants. However, *cis*-regulatory variants often make similar (McManus et al. 2010; Shi et al. 2012; Coolon et al. 2014; Guerrero et al. 2016) or greater (Wittkopp et al. 2008; Shi et al. 2012; Mack et al. 2016) contributions to gene expression divergence between species. Studies directly comparing the relative contributions of *cis*- and *trans*- regulatory variants with expression divergence suggest that the relative contribution of *cis*- regulatory variants increases with divergence time (Fig. 1-2A,B; Coolon et al. 2014; Metzger et al. 2017). This increasing *cis*-regulatory contribution can be

explained by *cis*-regulatory variants being more beneficial (Emerson et al. 2010; Coolon etc) and/or less deleterious (Schaefke et al. 2013) than *trans*-regulatory variants, which might result from differences in their average pleiotropy, as discussed in the section above.

Studies identifying eQTLs contributing to variation in gene expression have been conducted in a similarly diverse array of taxa (Gibson and Weir 2005; Rockman and Kruglyak 2006; Gilad et al. 2008; Nica and Dermitzakis 2013). eQTLs located close to the affected gene (that is, proximal) are often considered *cis*- acting whereas eQTLs located further from the affected gene (that is, distal) are often considered *trans*- acting (Brem et al. 2002). Consistent with this assumption, proximal eQTLs often have allele-specific effects on gene expression (Mohammadi et al. 2017). Indeed, the largest study of eQTLs to date, which was conducted by the Genotype-Tissue Expression (GTEx) Consortium and surveyed gene expression in cells derived from 49 tissues from up to 838 humans, has shown a strong correlation between the estimated effect of eQTLs designated as *cis*- acting and allele-specific measures of expression in heterozygous individuals (GTEx Consortium 2020). Several eQTL studies have reported that the majority of heritable expression variation is explained by *trans*- acting eQTLs (Grundberg et al. 2012; Wright et al. 2014; Liu et al. 2019), some of which affect the expression of many genes and are known as 'hot spots' (Yvert et al. 2003; Kliebenstein 2009; Albert et al. 2018; Lutz et al. 2019). The GTEx study detected at least one *cis*- acting eQTL for nearly 95% of protein- coding genes, whereas *trans*- acting eQTLs were detected for only 121 protein-coding genes. The number of individuals surveyed for each tissue was a strong predictor of the number of *trans*-acting eQTLs detected, however, underscoring the importance of taking statistical power into account when comparing the number of *trans*- acting eQTLs reported among studies (GTEx Consortium 2020). The unequal power for detecting *cis*- and *trans*- regulatory variants must also

be considered when comparing eQTLs: systematically testing for *trans*-regulatory variants requires many more statistical tests, and thus a greater multiple testing burden, than *cis*-regulatory variants. For this reason, some eQTL studies have focused solely on identifying *cis*-eQTLs (Lappalainen et al. 2013; Kita et al. 2017). Relative effect sizes of putatively *cis*- and *trans*- eQTLs can be more fairly compared. Such comparisons tend to show that *cis*- eQTLs have larger effects on gene expression than *trans*- eQTLs (Gilad et al. 2008; Kliebenstein 2009). For example, in the GTEx study, an average of 22% of *cis*- eQTLs caused a two fold or greater change in gene expression compared with 19% of *trans*-eQTLs (GTEx Consortium 2020). Similarly, in a recent, highly powered eQTL mapping study between two strains of *S. cerevisiae*, the average *cis*-eQTL also explained 2.8- fold more of the expression variation than the average *trans*-eQTL (Albert et al. 2018). But genes are often regulated by multiple *trans*-regulatory variants, and sets of *trans*- eQTLs affecting expression of the same gene tend to explain more of that gene's expression variation than its *cis*- eQTLs (Grundberg et al. 2012; Wright et al. 2014; Albert et al. 2018). This observation is consistent with the greater combined contribution of *trans*-regulatory variation to polymorphic gene expression inferred using ASE. Although ASE and eQTL studies reveal the relative contributions of *cis*- and *trans*-regulatory variation, they provide little insight into the specific genetic changes and molecular mechanisms altered by this variation. Only when such studies reach single-variant resolution can they provide this type of insight, which is necessary for a complete understanding of why the patterns of regulatory variation we see today exist (Bernardo Lemos, Christian R. Landry, Pierre Fontanillas, Susan C. P. Renn, Rob Kulathinal, Kyle M. Brown, and Daniel L. Hartl 2008). Considering the molecular mechanisms by which *cis*- and *trans*-regulatory mutations can have an effect give some insight

into why they may differ in their pleiotropic effects and, therefore, why natural selection may act on them differently.

The mechanistic effects of *cis*-regulatory mutations

cis-regulatory variation arises from genetic changes affecting sequences controlling the expression of a particular allele of a gene. These changes can affect the core promoter and enhancers of the gene, which both contain binding sites for transcription factors. They can also affect chromatin structure influencing the accessibility of DNA to transcription factors and sequences in the RNA transcript that affect its structure, stability or translation (Fig. 1-3), although more research is needed in these areas (Schaefke et al. 2018). At the most proximal level, a gene's expression is controlled by its core promoter sequence, which contains binding sites for the general transcription factors necessary for transcription. Despite the potential for core promoters to contribute to expression divergence, key elements of their sequence (Carninci et al. 2006), histone marks (Villar et al. 2015) and function (Lubliner et al. 2015) are often highly conserved among species. This conservation is presumably driven by the requirement for a functional promoter to express a gene as well as the strong functional constraints on proteins that bind to these sequences because they regulate so many different genes, making them highly pleiotropic. Indeed, sequences within promoters that serve as binding sites for general transcription factors, such as TATA boxes, are the most highly conserved portions of mammalian core promoters (Carninci et al. 2006). However, a comparison of core promoter sequences between human and rhesus macaque suggested that core promoters for a small number of genes might be diverging due to positive selection (Liang et al. 2008), and other work has shown that the gain and loss of core promoters contributes to expression divergence between mouse and human (Young et al. 2015). Furthermore, even if variation in the core promoter itself

is not the source of expression divergence, the structure of the core promoter can still influence expression divergence. For example, the presence of a TATA box (Tirosh et al. 2006; Landry et al. 2007), nucleosome positioning in the core promoter (Hornung et al.) and tandem repeats in the core promoter sequence have all been shown to correlate with expression divergence in yeast (Tirosh et al. 2009).

Compared with core promoters, enhancers are typically located further from the transcription start site in either upstream (5′), downstream (3′) or intronic regions (Andersson and Sandelin 2020) and seem to more often be the source of *cis*- regulatory variation affecting gene expression (Wray 2007; Wittkopp and Kalay 2011; Long et al. 2016). Because enhancers regulate gene expression in a more time-specific, tissue-specific or environment-specific manner than core promoters, they are expected to be subject to less functional constraint due to pleiotropy (Paaby and Rockman 2013) and thus more evolvable (Wray et al. 2003). Indeed, histone marks commonly associated with enhancers show greater divergence among mammalian species than histone marks associated with core promoters (Villar et al. 2015). Although single-cell organisms such as *S. cerevisiae* lack enhancers, they have upstream activating and repressing sequences that often work in a similarly context- dependent manner (Hahn and Young 2011).

The primary functional units within all of these *cis*-regulatory DNA sequences are binding sites for transcription factors, which can activate or repress transcription (Spitz and M Furlong 2012). These sequences are short, degenerate and able to evolve relatively quickly, even from random sequences (Rockman and Wray 2002; de Boer et al. 2019). Mutations that change the identity, affinity, orientation, number and/or spacing of transcription factor binding sites (TFBSs) can alter *cis*-regulatory activity (Swanson et al. 2011; Sharon et al. 2012; Long et al. 2016). Large-scale mutagenesis studies of enhancers and other similar *cis*-regulatory elements

have shown that, although many mutations in these sequences can alter gene expression, mutations in TFBSs tend to have the largest effects (Kwasnieski et al. 2012; Melnikov et al. 2012; Patwardhan et al. 2012; Metzger et al. 2015). Although TFBSs are often among the most highly conserved sequences within an enhancer (Zhang and Gerstein 2003; Cooper et al. 2006; Burgess and Freeling 2014; Glenwinkel et al. 2014), they can also harbor genetic changes responsible for variation in gene expression within species (Lewinsky et al. 2005; Claussnitzer et al. 2014; Corradin and Scacheri 2014) and between species (Arnoult et al. 2013; Chang et al. 2013). However, in most cases where functional changes have been mapped to enhancers or similar *cis*-regulatory sequences, the specific genetic changes responsible for altering their function have not yet been identified (Stern and Orgogozo 2008; Martin and Orgogozo 2013; Rebeiz and Williams 2017; Klein et al. 2018).

The mechanistic effects of *trans*-regulatory mutations

Whereas *cis*-regulatory variants tend to lie near the affected gene, *trans*-regulatory variants affecting a gene's expression can be located virtually anywhere in the genome (Fig. 1-4A). These potential sites of *trans*-regulatory variants include both coding and non-coding sequences that affect expression or activity of gene products that regulate the focal gene's expression either directly (by binding to its *cis*-acting sequences) or indirectly (by influencing the activity of direct regulators) (Fig. 1-4B; Lutz et al. 2019). The diversity of mechanisms by which trans-regulatory variants can influence a gene's expression leads to considerable variability in their effects on the expression of the focal gene, as well as their effects on other genes throughout the genome.

Coding and non-coding sequences.

Although the effects of *trans*-regulatory variants are mediated by diffusible molecules such as RNAs or proteins, studies of regulatory variation segregating in humans suggest that most *trans*- acting variants are not located within the sequences encoding these molecules (GTEx Consortium et al. 2017). Instead, in large-scale genome-wide association studies, the majority of *trans*-regulatory variants have been found in non-coding, putatively *cis*-regulatory sequences controlling the gene's expression (Battle et al. 2014; GTEx Consortium et al. 2017; Yao et al. 2017; GTEx Consortium 2020). By changing expression of the gene they affect in *cis*, such variants can affect the expression of other genes in *trans* (Yvert et al. 2003; Grundberg et al. 2012; Brynedal et al. 2017; GTEx Consortium et al. 2017). For example, a *cis*- acting eQTL located near the gene encoding lysozyme (an enzyme that breaks down bacterial cell walls) has been shown to also act as a *trans*-acting eQTL for expression of other genes in monocytes (Fairfax et al. 2012). Similarly, a *cis*-acting eQTL near the transcription factor KLF14, which regulates expression of genes in adipose tissue, explains *trans*- acting effects observed on expression of other genes (Small et al. 2011). However, studies of *S. cerevisiae* suggest that this species might have a different distribution of *trans*- regulatory variants in coding and non-coding sequences. As in humans, hotspot genes with *trans*- regulatory eQTLs affecting expression of many genes are more likely to have local, putatively *cis*- acting eQTLs than expected by chance (Albert et al. 2018), but the functional *trans*- regulatory variants mapped and validated in S. cerevisiae so far have primarily, although not exclusively, been in coding regions (Brem et al. 2002; Yvert et al. 2003; Ronald et al. 2005; Sudarsanam and Cohen 2014; Albert et al. 2018; Lutz et al. 2019). *S. cerevisiae* might have a higher proportion of *trans*-regulatory variants in coding sequences than humans because so much less of their genome is non-coding (27% in S. cerevisiae versus 97% in humans (Alexander et al. 2010); however, the higher proportion of

coding variants might also be a consequence of often using a laboratory-adapted strain that carries many variants absent from wild populations (Doniger et al. 2008). Determining the true relative contributions of coding and non-coding variants to *trans*- regulatory variation in yeast (and other species) will require much more extensive mapping and functional testing of variants from natural populations. If *trans*-regulatory variants generally do map to non-coding sequences more often than coding sequences, this might be because mutations in non-coding sequences tend to be less pleiotropic. For example, non-coding mutations that affect activity of a tissue-specific enhancer are expected to have an impact on fewer traits than coding mutations that alter the same gene's protein sequence everywhere it is expressed (Wray et al. 2003; Wray 2007; Carroll 2008; Stern and Orgogozo 2009). Indeed, most *trans*- acting eQTLs in human non-coding sequences seem more likely to affect enhancers than core promoters (GTEx Consortium et al. 2017), and often have tissue-specific effects (GTEx Consortium et al. 2017; Liu et al. 2019). Because mutations that are more pleiotropic are expected to typically be more deleterious than less pleiotropic mutations (Kimura and Ohta 1974), coding mutations might be selected against more strongly than non-coding mutations, reducing their frequency in natural populations. However, this paradigm is challenged by data showing that *cis*-regulatory sequences are more pleiotropic (Sabarís et al. 2019), and protein sequences more modular (Lynch and Wagner 2008; Wagner and Lynch 2008), than generally appreciated. Indeed, a recent study has shown how modularity in the yeast MATα2 transcription factor protein facilitated its divergence, which was then followed by changes in *cis*-regulatory, non-coding sequences of the genes it regulates (Britton et al. 2020).

These findings highlight the fact that coding and noncoding mutations are not synonymous with *cis*- and *trans*-regulatory mutations (Wittkopp 2005). The differences in

pleiotropy between coding and noncoding sequences based on the modularity of regulatory elements described above, therefore, may or may not contribute to differences in pleiotropy, and potentially fitness, between *cis*- and *trans*-regulatory mutations. Furthermore, because many *trans*-regulatory mutations occur in noncoding sequences controlling expression of the *trans*-regulator rather than in the coding sequences, they are simultaneously *cis*-acting mutations to the *trans*-regulator, and *trans*-acting mutations to the genes regulated. Since the designation of mutations as *cis*- or *trans*-regulatory is relative to the specific gene being discussed, it is likely that any difference in pleiotropy between *cis*- and *trans*-regulatory mutations is also relative, and not dependent on whether the sequence being mutated is in coding or noncoding DNA. Rather, the differences in pleiotropy between *cis*- and *trans*-regulatory mutations may be related to the number of genes they affect based on their positions relative to the focal gene in the regulatory network.

Transcription factors.

Transcription factors are often considered the most likely source of *trans*-regulatory variation, especially for hotspot eQTLs, because most transcription factors regulate expression of many target genes (Gerstein et al. 2010; modENCODE Consortium et al. 2010; ENCODE Project Consortium 2012; Kemmeren et al. 2014; Yue et al. 2014). Indeed, transcription factors do often seem to be responsible for hotspot eQTLs in both humans (Bryois et al. 2014; Yao et al. 2017; Cesar et al. 2018) and *S. cerevisiae* (Lee and Bussemaker 2010; Albert et al. 2018). However, the ability of transcription factors to affect expression of multiple downstream target genes also results in functional constraint on their variation. Indeed, their protein-coding sequences, DNA-binding specificities and general physiological roles are often conserved over long evolutionary timescales (Lambert et al. 2018). Despite these general trends of conservation,

transcription factors can and do diverge in function, as changes in protein sequences, including those that affect their DNA binding specificity, have been reported for transcription factors controlling the mating type in yeast (Gerke et al. 2009; Baker et al. 2011), flower development and cell division in plants (Sayou et al. 2014) and body patterning in insects (Galant and Carroll 2002; Ronshaugen et al. 2002), among others.

Sources of trans- regulatory variation other than transcription factors.

Variants affecting genes not encoding transcription factors are also important sources of *trans*-regulatory variation. For example, chromatin regulators can have widespread effects on gene expression (Choi and Kim 2008), and an eQTL study in *S. cerevisiae* suggests that genes encoding these types of protein harbor *trans*- acting eQTLs affecting expression of many genes (Lee et al. 2006). Functional studies in *S. cerevisiae* have also demonstrated *trans*-regulatory effects of variants in cofactors that modulate the activity of transcription factors (Fazlollahi et al. 2016) as well as genes that influence metabolism, such as the glucose receptor *RGT2*58, and a membrane protein, SSY1, that senses the concentration of extracellular amino acids (Brown et al. 2008). In humans, *trans*-eQTLs have also been shown to map to genes that do not encode transcription factors, such as the SLCO1A6 gene, in which a genetic variant was shown to alter expression of many genes by altering the transport of bile acids in pancreatic islets (Tian et al. 2015). The diverse sources of *trans*-regulatory variation illustrated by these and other studies result from the interconnectedness of transcriptional, structural, signaling and metabolic networks, and underscore the challenge of predicting and identifying *trans*-regulatory variants with our current understanding of systems biology (Flint and Ideker 2019). They are also consistent with the proposed 'omnigenic' model of heritability, in which every gene expressed has the potential to influence every trait (Boyle et al. 2017).

Comparing the pleiotropic effects of *cis*- and *trans*-regulatory mutations

While the interconnected nature of cellular networks hints at the potential for *trans*-regulatory mutations to indeed be very pleiotropic, it also suggests that *cis*-regulatory mutations may trigger many downstream effects and therefore also be very pleiotropic. It is necessary, therefore, to collect empirical data directly addressing the questions of how the pleiotropic effects of *cis*- and *trans*-regulatory mutations compare to each other. Mutants bearing individual *cis*- and *trans*-regulatory mutations to the same gene, *TDH3*, now exist in the baker's yeast *Saccharomyces cerevisiae* (Metzger et al. 2015; Duveau et al. 2021), making such a direct comparison of pleiotropic effects at a detailed molecular level possible for the first time. Analysis of these strains, in conjunction with broader analyses to address the generalizability of the findings, provide direct empirical evidence of differences in pleiotropy between *cis*- and *trans*-regulatory mutations and their relationship to fitness.

Thesis overview

In this thesis I examine the pleiotropic effects of *cis*- and *trans*-regulatory mutations on both gene expression and fitness. In addition, I explore active compensation as a molecular mechanism that may explain some of the differences in pleiotropic effects of cis- and trans-regulatory mutations to the gene *TDH3* in *Saccharomyces cerevisiae*.

In the second chapter, I investigate the hypothesis that *trans*-regulatory mutations have more extensive pleiotropic effects than *cis*-regulatory mutations, leading to differences in fitness that are an important factor in the evolution of gene expression. While this hypothesis has frequently been stated, the difficulty of quantifying pleiotropy has resulted in a dearth of empirical data to support this hypothesis. I fill this gap by designing and performing fitness assays and an RNA-sequencing strategy to collect genome-wide gene expression data for 45

yeast strains containing single point mutations affecting the expression of the same gene in *cis* or in *trans*. My analysis of these data demonstrates for the first time that the effects of *trans*-regulatory mutations are indeed more pleiotropic than the effects of *cis*-regulatory mutations, but only for mutations with a similar effect size on expression of the focal gene. In addition, *trans*-regulatory mutations frequently have negative pleiotropic fitness effects. Importantly, this work also revealed that the downstream consequences of changing a gene's expression via a *cis*-regulatory mutation are not reproduced when its expression is changed via *trans*-regulatory mutations. This unexpected finding highlights the necessity of understanding the context in which a mutation occurs in order to predict the consequences of that mutation for the organism as a whole.

In the third chapter I explore whether the pattern of more pleiotropic *trans*-regulatory mutations than *cis*-regulatory holds for other genes in the genome, and whether this is dependent on *cis* and *trans* mutations being in noncoding or coding sequences. I develop a framework to test these questions using existing gene expression data for a large number of yeast gene deletion strains. By building a perturbation network from these data and considering each node in turn as the focal gene, I find that the topology of that network is able to generate a pattern in which *trans*-acting mutations have more widespread effects than *cis*-acting mutations for the vast majority of genes analyzed. This finding demonstrates broad generalizability of *cis*-acting mutations being less pleiotropic than *trans*-acting mutations to the same focal gene in the yeast *Saccharomyces cerevisiae*.

In the fourth chapter I provide one potential explanation for why *cis*- and *trans*-regulatory mutations to the same gene do not have comparable effects on genes influenced by the change in expression of the focal gene. I do this by demonstrating that the reduction in expression of the

yeast gene *TDH3* by mutations in the *cis*-regulatory element can be compensated for by upregulation of its paralogs *TDH1* and *TDH2*. This compensation appears to happen via a feedback mechanism that involves at least two of the direct regulators of all three paralogs, *RAP1* and *GCR1*, as it increases the expression of the paralogs and a reporter gene driven by an intact *TDH3* promoter. This upregulation does not occur in yeast strains with mutated versions of *RAP1* and *GCR1*. This work provides a look into the mechanism by which active compensation for reduction in an enzyme may take place in the cell, and provides one explanation for why the genome-wide effects of *cis*- and *trans*-regulatory mutations differ as they do.

References

Albert FW, Bloom JS, Siegel J, Day L, Kruglyak L. 2018. Genetics of trans-regulatory variation in gene expression. *Elife* [Internet] 7. Available from: http://dx.doi.org/10.7554/eLife.35471

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

Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. 2010. Annotating non-coding regions of the genome. *Nat. Rev. Genet.* 11:559–571.

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

Arnoult L, Su KFY, Manoel D, Minervino C, Magriña J, Gompel N, Prud'homme B. 2013. Emergence and diversification of fly pigmentation through evolution of a gene regulatory module. *Science* 339:1423–1426.

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. 2004. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* 14:283–291.

Baker CR, Tuch BB, Johnson AD. 2011. Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc. Natl. Acad. Sci. U. S. A.* 108:7493–7498.

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, et al. 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24:14–24.

Bell GDM, Kane NC, Rieseberg LH, Adams KL. 2013. RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biol. Evol.* 5:1309–1323.

Bernardo Lemos, Christian R. Landry, Pierre Fontanillas, Susan C. P. Renn, Rob Kulathinal, Kyle M. Brown, and Daniel L. Hartl. 2008. Evolution of Genomic Expression. In: Pagel M, Pomiankowski A, editors. Evolutionary Genomics and Proteomics. Sinauer Associates. p. 81–118.

de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. 2019. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* [Internet]. Available from: http://www.nature.com/articles/s41587-019-0315-8

Boyle EA, Li YI, Pritchard JK. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 169:1177–1186.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752–755.

Britton CS, Sorrells TR, Johnson AD. 2020. Protein-coding changes preceded cis-regulatory gains in a newly evolved transcription circuit. *Science* 367:96–100.

Brown KM, Landry CR, Hartl DL, Cavalieri D. 2008. Cascading transcriptional effects of a naturally occurring frameshift mutation in Saccharomyces cerevisiae. *Mol. Ecol.* 17:2985–2997.

Brynedal B, Choi J, Raj T, Bjornson R, Stranger BE, Neale BM, Voight BF, Cotsapas C. 2017. Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *Am. J. Hum. Genet.* 100:581–591.

Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, Ho KM, Ring S, Hurles M, Deloukas P, et al. 2014. Cis and trans effects of human genomic variants on gene expression. *PLoS Genet.* 10:e1004461.

Burgess D, Freeling M. 2014. The most deeply conserved noncoding sequences in plants serve similar functions to those in vertebrates despite large differences in evolutionary rates. *Plant Cell* 26:946–961.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38:626–635.

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol.* 3:e245.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134:25–36.

Cesar ASM, Regitano LCA, Reecy JM, Poleti MD, Oliveira PSN, de Oliveira GB, Moreira GCM, Mudadu MA, Tizioto PC, Koltes JE, et al. 2018. Identification of putative regulatory regions and transcription factors associated with intramuscular fat content traits. *BMC Genomics* 19:499.

Chang J, Zhou Y, Hu X, Lam L, Henry C, Green EM, Kita R, Kobor MS, Fraser HB. 2013. The Molecular Mechanism of a Cis-Regulatory Adaptation in Yeast.Fay JC, editor. *PLoS Genet.* 9:e1003813.

Chen J, Nolte V, Schlötterer C. 2015. Temperature stress mediates decanalization and dominance of gene expression in Drosophila melanogaster. *PLoS Genet.* 11:e1004883.

Choi JK, Kim YJ. 2008. Epigenetic regulation and the variability of gene expression. *Nat. Genet.* 40:141–147.

Claussnitzer M, Dankel SN, Klocke B, Grallert H, Glunk V, Berulava T, Lee H, Oskolkov N, Fadista J, Ehlers K, et al. 2014. Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell* 156:343–358.

Coolon etc W. cis- and trans-regulation contribute to accelerated divergence of gene expression on the Drosophila X chromosome.

Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. 2014. Tempo and mode of regulatory evolution in Drosophila. *Genome Res.* 24:797–808.

Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* 16:1–10.

Cooper TF, Ostrowski EA, Travisano M. 2007. A NEGATIVE RELATIONSHIP BETWEEN MUTATION PLEIOTROPY AND FITNESS EFFECT IN YEAST. *Evolution* 61:1495–1499.

Corradin O, Scacheri PC. 2014. Enhancer variants: evaluating functions in common disease. *Genome Med.* 6:85.

Courtier-Orgogozo V, Arnoult L, Prigent SR, Wiltgen S, Martin A. 2020. Gephebase, a database of genotype-phenotype relationships for natural and domesticated variation in Eukaryotes. *Nucleic Acids Res.* 48:D696–D703.

Davidson JH, Balakrishnan CN. 2016. Gene Regulatory Evolution During Speciation in a Songbird. *G3* 6:1357–1364.

Deutschbauer AM, Davis RW. 2005. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat. Genet.* 37:1333–1340.

Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang S-P, Fay JC. 2008. A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet.* 4:e1000183.

Dudley AM, Janse DM, Tanay A, Shamir R, Church GM. 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol. Syst. Biol.* 1:2005.0001.

Duveau F, Vande Zande P, Metzger BP, Diaz CJ, Walker EA, Tryban S, Siddiq MA, Yang B, Wittkopp PJ. 2021. Mutational sources of trans-regulatory variation affecting gene expression in Saccharomyces cerevisiae. Elife [Internet] 10. Available from: http://dx.doi.org/10.7554/eLife.67806

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

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, Ellis P, Langford C, Vannberg FO, Knight JC. 2012. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* 44:502–510.

Fazlollahi M, Muroff I, Lee E, Causton HC, Bussemaker HJ. 2016. Identifying genetic modulators of the connectivity between transcription factors and their transcriptional targets. *Proc. Natl. Acad. Sci. U. S. A.* 113:E1835–E1843.

Featherstone DE, Broadie K. 2002. Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. *Bioessays* 24:267–274.

Fisher RA. 1930. The genetical theory of natural selection. 272. Available from: https://psycnet.apa.org/fulltext/1930-04698-000.pdf

Flint J, Ideker T. 2019. The great hairball gambit.Copenhaver GP, editor. *PLoS Genet.* 15:e1008519.

Galant R, Carroll SB. 2002. Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* 415:910–913.

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

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* 330:1775–1787.

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

Gilad Y, Rifkin SA, Pritchard JK. 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* 24:408–415.

Glenwinkel L, Wu D, Minevich G, Hobert O. 2014. TargetOrtho: a phylogenetic footprinting tool to identify transcription factor targets. *Genetics* 197:61–76.

Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, Flicek P, Brazma A, Odom DT, Marioni JC. 2012. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res.* 22:2376–2384.

Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, Bell JT, Yang T-P, Meduri E, Barrett A, et al. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* 44:1084–1089.

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

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. 2017. Genetic effects on gene expression across human tissues. *Nature* 550:204–213.

Guerrero RF, Posto AL, Moyle LC, Hahn MW. 2016. Genome-wide patterns of regulatory divergence revealed by introgression lines. *Evolution* 70:696–706.

Hahn S, Young ET. 2011. Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* 189:705–736.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. *Genetics* 173:1885–1891.

Hornung G, Oren M, Barkai N. Nucleosome Organization Affects the Sensitivity of Gene Expression to Promoter Mutations. Available from: https://doc-0s-0s-apps-viewer.googleusercontent.com/viewer/secure/pdf/manuqqjquh28234to5vs0s5vd6e1cv61/kfpqkl6q82uv5838kmofi9sfn8fa4ae9/1495806375000/gmail/13015333458294748387/ACFrOgCERDYP_22KXEa0hGDqVcQFsbtMjFqqVkbQLKORgkvbk-bWS3TKP5B-jgg0t-JFqNF4r2NDgW

Kemmeren P, Sameith K, van de Pasch LAL, Benschop JJ, Lenstra TL, Margaritis T, O'Duibhir E, Apweiler E, van Wageningen S, Ko CW, et al. 2014. Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. *Cell* 157:740–752.

Kimura M, Ohta T. 1974. On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 71:2848–2852.

Kinsler G, Geiler-Samerotte K, Petrov DA. 2020. Fitness variation across subtle environmental perturbations reveals local modularity and global pleiotropy of adaptation. *Elife* [Internet] 9. Available from: http://dx.doi.org/10.7554/eLife.61271

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

Klein JC, Keith A, Agarwal V, Durham T, Shendure J. 2018. Functional characterization of enhancer evolution in the primate lineage. *Genome Biol.* 19:99.

Kliebenstein D. 2009. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annu. Rev. Plant Biol.* 60:93–114.

Kopp A, McIntyre LM. 2012. Transcriptional network structure has little effect on the rate of regulatory evolution in yeast. *Mol. Biol. Evol.* 29:1899–1905.

Kronforst MR, Barsh GS, Kopp A, Mallet J, Monteiro A, Mullen SP, Protas M, Rosenblum EB, Schneider CJ, Hoekstra HE. 2012. Unraveling the thread of nature's tapestry: the genetics of diversity and convergence in animal pigmentation. *Pigment Cell Melanoma Res.* 25:411–433.

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U. S. A.* 109:19498–19503.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* [Internet] 172:650–665. Available from: http://dx.doi.org/10.1016/j.cell.2018.01.029

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic properties influencing the evolvability of gene expression. *Science* 317:118–121.

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506–511.

Lee E, Bussemaker HJ. 2010. Identifying the genetic determinants of transcription factor activity. *Mol. Syst. Biol.* 6:412.

Lee S-I, Pe'er D, Dudley AM, Church GM, Koller D. 2006. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc. Natl. Acad. Sci. U. S. A.* 103:14062–14067.

Lewinsky RH, Jensen TGK, Møller J, Stensballe A, Olsen J, Troelsen JT. 2005. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. *Hum. Mol. Genet.* 14:3945–3953.

Liang H, Lin Y-S, Li W-H. 2008. Fast evolution of core promoters in primate genomes. *Mol. Biol. Evol.* 25:1239–1244.

Liu X, Li YI, Pritchard JK. 2019. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* 177:1022-1034.e6.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* 167:1170–1187.

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.

Lutz S, Brion C, Kliebhan M, Albert FW. 2019. DNA variants affecting the expression of numerous genes in trans have diverse mechanisms of action and evolutionary histories.Fay JC, editor. *PLoS Genet.* 15:e1008375.

Lynch VJ, Wagner GP. 2008. Resurrecting the role of transcription factor change in developmental evolution. *Evolution* 62:2131–2154.

Mack KL, Campbell P, Nachman MW. 2016. Gene regulation and speciation in house mice. :451–461.

Martin A, Orgogozo V. 2013. THE LOCI OF REPEATED EVOLUTION : A CATALOG OF GENETIC HOTSPOTS OF PHENOTYPIC VARIATION. :1235–1250.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res.* 20:816–825.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30:271–277.

Metzger BPH, Wittkopp PJ, Coolon JD. 2017. Evolutionary dynamics of regulatory changes underlying gene expression divergence among Saccharomyces species. *Genome Biol. Evol.* 177:1987–1996.

Metzger BPH, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ. 2015. Selection on noise constrains variation in a eukaryotic promoter. *Nature* 521:344–347.

modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* 330:1787–1797.

Mohammadi P, Castel SE, Brown AA, Lappalainen T. 2017. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* 27:1872–1884.

Nica AC, Dermitzakis ET. 2013. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368:20120362.

Oliver F, Christians JK, Liu X, Rhind S, Verma V, Davison C, Brown SDM, Denny P, Keightley PD. 2005. Regulatory variation at glypican-3 underlies a major growth QTL in mice. *PLoS Biol.* 3:e135.

Orr HA. 2000. Adaptation and the cost of complexity. *Evolution* 54:13–20.

Paaby AB, Rockman MV. 2013. The many faces of pleiotropy. *Trends Genet.* 29:66–73.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30:265–270.

Pavlicev M, Wagner GP. 2012. Coming to Grips with Evolvability. *Evolution: Education and Outreach* 5:231–244.

Preger-Ben Noon E, Sabarís G, Ortiz DM, Sager J, Liebowitz A, Stern DL, Frankel N. 2018. Comprehensive Analysis of a cis -Regulatory Region Reveals Pleiotropy in Enhancer Function. *Cell Rep.* 22:3021–3031.

Rebeiz M, Williams TM. 2017. Using Drosophila pigmentation traits to study the mechanisms of cis-regulatory evolution. *Current Opinion in Insect Science* [Internet] 19:1–7. Available from: http://dx.doi.org/10.1016/j.cois.2016.10.002

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. *Nat. Rev. Genet.* 7:862–872.

Rockman MV, Wray GA. 2002. Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* 19:1991–2004.

Ronald J, Brem RB, Whittle J, Kruglyak L. 2005. Local regulatory variation in Saccharomyces cerevisiae. *PLoS Genet.* 1:e25.

Ronshaugen M, McGinnis N, McGinnis W. 2002. Hox protein mutation and macroevolution of the insect body plan. *Nature* 415:914–917.

Sabarís G, Laiker I, Preger-Ben Noon E, Frankel N. 2019. Actors with Multiple Roles: Pleiotropic Enhancers and the Paradigm of Enhancer Modularity. *Trends Genet.* [Internet] 0. Available from: http://dx.doi.org/10.1016/j.tig.2019.03.006

Sayou C, Monniaux M, Nanao MH, Moyroud E, Brockington SF, Thévenon E, Chahtane H, Warthmann N, Melkonian M, Zhang Y, et al. 2014. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science* 343:645–648.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422:297–302.

Schaefke B, Emerson JJ, Wang TY, Lu MYJ, Hsieh LC, Li WH. 2013. Inheritance of gene expression level and selective constraints on trans-and cis-regulatory changes in yeast. *Mol. Biol. Evol.* 30:2121–2133.

Schaefke B, Sun W, Li Y-S, Fang L, Chen W. 2018. The evolution of posttranscriptional regulation. *Wiley Interdiscip. Rev. RNA*:e1485.

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.

Shi X, Ng DW-K, Zhang C, Comai L, Ye W, Chen ZJ. 2012. Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. *Nat. Commun.* 3:950.

Siegal ML, Promislow DEL, Bergman A. 2007. Functional and evolutionary inference in gene networks: does topology matter? *Genetica* 129:83–103.

Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. *Trends Genet.* 34:532–544.

Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, Kong A, Thorsteindottir U, Shin S-Y, Richards HB, GIANT Consortium, et al. 2011. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat. Genet.* 43:561–564.

Spitz F, M Furlong EE. 2012. Transcription factors: from enhancer binding to developmental control. *Nature Publishing Group* [Internet] 13. Available from: http://dx.doi.org/10.1038/nrg3207

Springer NM, Stupar RM. 2007. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *Plant Cell* 19:2391–2402.

Stearns FW. 2010. One Hundred Years of Pleiotropy: A Retrospective. *Genetics* 186.

Stern DL, Orgogozo V. 2008. THE LOCI OF EVOLUTION : HOW PREDICTABLE IS GENETIC EVOLUTION ? :2155–2177.

Stern DL, Orgogozo V. 2009. Is genetic evolution predictable? *Science* [Internet] 323:746–751. Available from: http://dx.doi.org/10.1126/science.1158997

Sudarsanam P, Cohen BA. 2014. Single nucleotide variants in transcription factors associate more tightly with phenotype than with gene expression. *PLoS Genet.* 10:e1004325.

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, et al. 2009. Roles of trans and cis variation in yeast intraspecies evolution of gene expression. *Mol. Biol. Evol.* 26:2533–2538.

Suvorov A, Nolte V, Pandey RV, Franssen SU, Futschik A, Schlötterer C. 2013. Intra-specific regulatory variation in Drosophila pseudoobscura. *PLoS One* 8:e83547.

Swanson CI, Schwimmer DB, Barolo S. 2011. Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Curr. Biol.* 21:1186–1196.

Tian J, Keller MP, Oler AT, Rabaglia ME, Schueler KL, Stapleton DS, Broman AT, Zhao W, Kendziorski C, Yandell BS, et al. 2015. Identification of the Bile Acid Transporter Slco1a6 as a Candidate Gene That Broadly Affects Gene Expression in Mouse Pancreatic Islets. *Genetics* 201:1253–1262.

Tirosh I, Barkai N, Verstrepen KJ. 2009. Promoter architecture and the evolvability of gene expression. :1–6.

Tirosh I, Weinberger A, Carmi M, Barkai N. 2006. A genetic signature of interspecies variations in gene expression. *Nat. Genet.* 38:830–834.

Villar D, Berthelot C, Flicek P, Odom Correspondence DT. 2015. Enhancer Evolution across 20 Mammalian Species. *Cell* 160:554–566.

Wagner GP, Lynch VJ. 2008. The gene regulatory logic of transcription factor evolution. *Trends Ecol. Evol.* 23:377–385.

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* 12:204–213.

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, et al. 2007. Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. *Genome Res.* 17:1161–1169.

Wang M, Uebbing S, Ellegren H. 2017. Bayesian Inference of Allele-Specific Gene Expression Indicates Abundant Cis-Regulatory Variation in Natural Flycatcher Populations. *Genome Biol. Evol.* 9:1266–1279.

Wang X, Werren JH, Clark AG. 2016. Allele-Specific Transcriptome and Methylome Analysis Reveals Stable Inheritance and Cis-Regulation of DNA Methylation in Nasonia. *PLoS Biol.* 14:e1002500.

Wessinger CA, Rausher MD. 2012. Lessons from flower colour evolution on targets of selection. *J. Exp. Bot.* 63:5741–5749.

Wittkopp PJ. 2005. Genomic sources of regulatory variation in cis and in trans. *Cell. Mol. Life Sci.* 62:1779–1783.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* 430:85–88.

Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between Drosophila species. *Nat. Genet.* 40:346–350.

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

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8:206–216.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20:1377–1419.

Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, Madar V, Jansen R, Chung W, Zhou Y-H, et al. 2014. Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* 46:430–437.

Yang B, Wittkopp PJ. 2017. Structure of the Transcriptional Regulatory Network Correlates with Regulatory Divergence in Drosophila. *Mol. Biol. Evol.* 34:1352–1362.

Yao C, Joehanes R, Johnson AD, Huan T, Liu C, Freedman JE, Munson PJ, Hill DE, Vidal M, Levy D. 2017. Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *Am. J. Hum. Genet.* 100:571–580.

Young RS, Hayashizaki Y, Andersson R, Sandelin A, Kawaji H, Itoh M, Lassmann T, Carninci P, FANTOM Consortium, Bickmore WA, et al. 2015. The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res.* 25:1546–1557.

Yu H, Gerstein M. 2006. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.* 103:14724–14731.

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515:355–364.

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

Zhang X, Borevitz JO. 2009. Global analysis of allele-specific expression in Arabidopsis thaliana. *Genetics* 182:943–954.

Zhang Z, Gerstein M. 2003. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J. Biol.* 2:11.

Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M. 2011. Regulatory variation within and between species. *Annu. Rev. Genomics Hum. Genet.* 12:327–346.

Figures

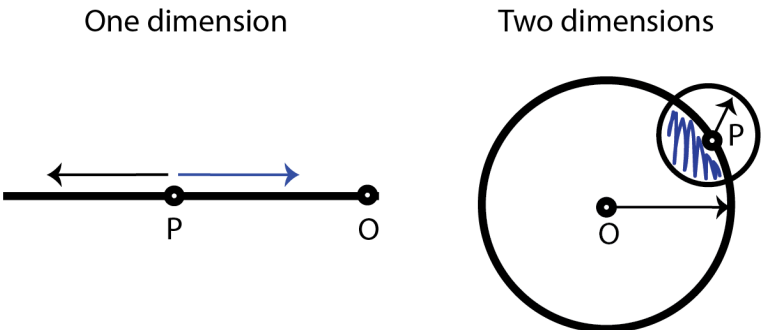

Figure 1-1: Fisher's geometric model and the fitness effects of pleiotropy

In a one-dimensional trait space, 50% of the mutations of a particular effect size will move the phenotype toward an optimum (blue arrow) and therefore be advantageous. In a two-dimensional trait space, the overlap between the movement of the phenotype and those phenotypes that are closer to the optimum is less than 50% (blue colored overlap). (Adapted from Pavlicev & Wagner, 2012)

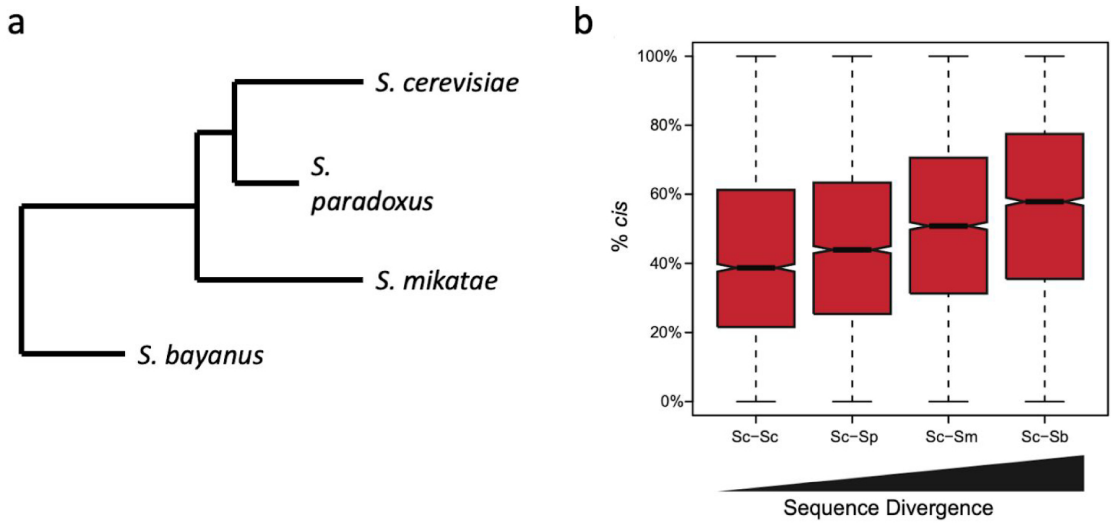

Figure 1-2: cis- and trans-regulatory contributions to expression differences between and within species

An analysis of allele-specific expression in hybrid yeast (*Saccharomyces*) species with a range of divergence times (a, branch lengths reflect relative divergence times) showed increasing contributions of *cis*-regulatory variation to expression differences with increasing divergence time (b, notches in the boxplot indicate 95% CI of the median).

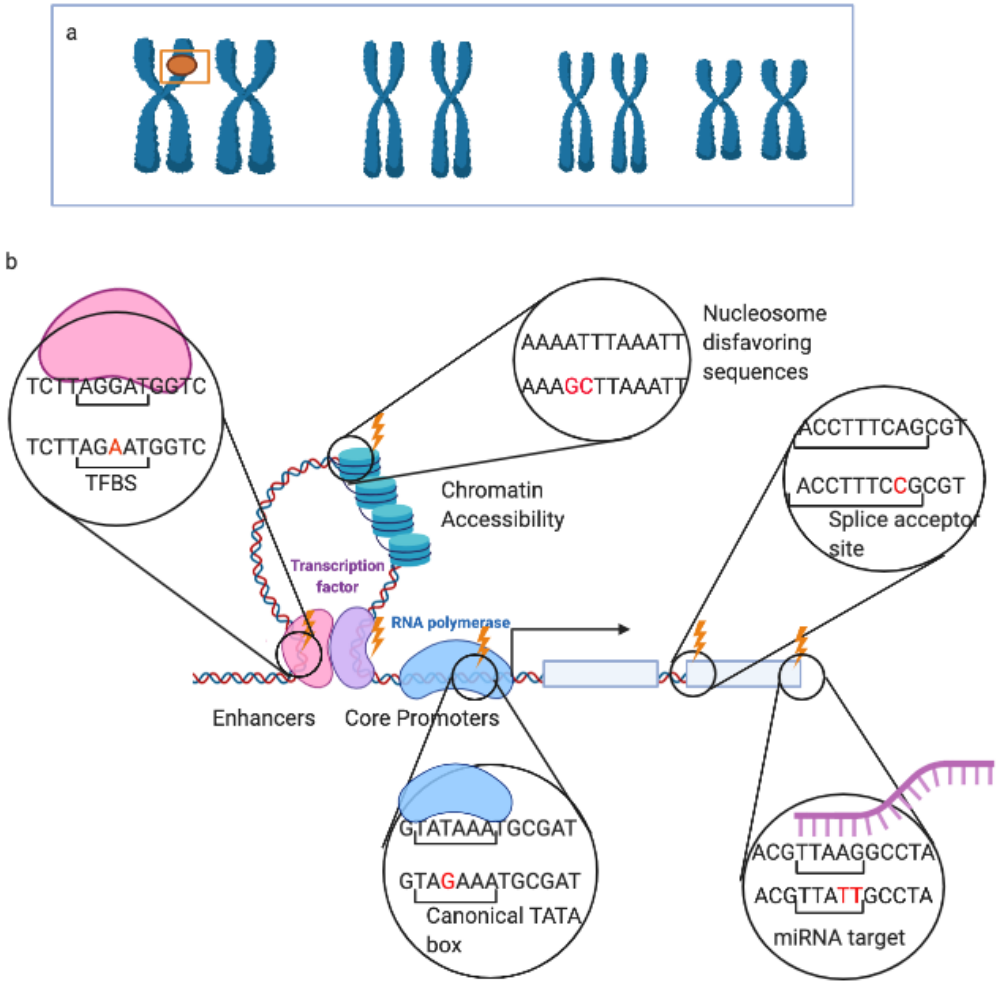

Figure 1-3: Sources of cis-regulatory variation in eukaryotes

(a) Mutations that are sources of *cis*-regulatory variation occur in close proximity to the gene of interest. (b) Mutations (indicated with lightning bolts) affecting the core promoter (including in motifs such as the TATA box used to assemble the transcription machinery activating RNA polymerase), enhancers (whose functional units are transcription factor binding sites (TFBS)), chromatin accessibility (altered by nucleosome placement and stability), and post-transcriptional regulation such as splicing sites or microRNA targets in the 3' UTR, can have *cis*-regulatory effects on gene expression.

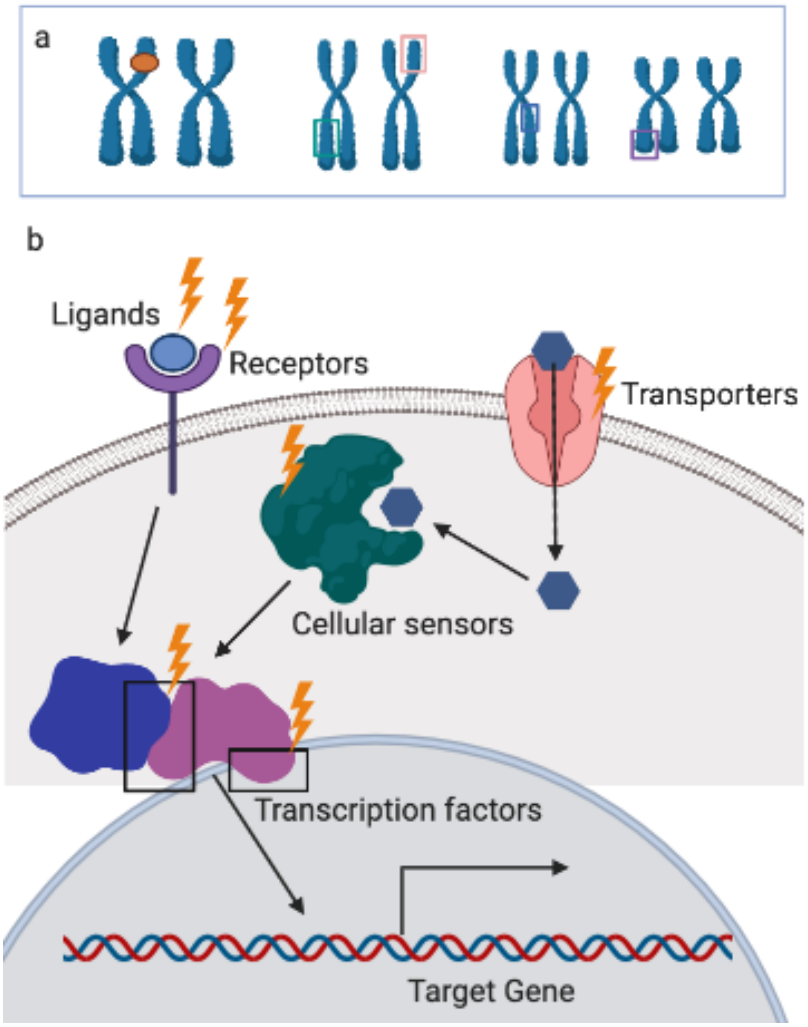

Figure 1-4: Sources of trans-regulatory variation

(a) Mutations that are sources of *trans*-regulatory variation can occur anywhere throughout the genome, in coding or non-coding sequences, and affect diffusible molecules that then influence the expression of the gene of interest. (b) These mutations (indicated by lightning bolts) can occur in non-coding or coding sequences of transcription factors, cellular sensors, transporters, signaling receptors and ligands, and other molecules that influence transcription of many genes via effects on the many interconnected cellular networks.

Chapter 2 Pleiotropic Effects of *Trans*-Regulatory Mutations on Fitness and Gene Expression

Abstract

Variation in gene expression arises from *cis*- and *trans*-regulatory mutations, which contribute differentially to expression divergence. Here, we compare the impacts on gene expression and fitness for *cis*- and *trans*-regulatory mutations affecting expression of the *TDH3* gene in *Saccharomyces cerevisiae*. We use the effects of *cis*-regulatory mutations to isolate effects of *trans*-regulatory mutations caused by impacts on *TDH3* from impacts on other genes, providing a rare distribution of pleiotropic effects. *Cis*- and *trans*-regulatory mutations had different effects on expression of genes downstream of *TDH3*, showing that the pleiotropic effects of *trans*-regulatory mutations do not only act in parallel to *cis*-regulatory mutations. The more widespread and deleterious effects of *trans*-regulatory mutations we observed are consistent with their decreasing relative contribution to expression differences over evolutionary time.

Main Text

Heritable variation in gene expression is widespread within and between species and often contributes to phenotypic diversity (Stern and Orgogozo 2008). This variation arises from mutations that alter activity of the regulatory networks that control when, where, and how much of a gene product is produced. Each regulatory mutation can act in *cis* or in *trans* with respect to a specific gene. *cis*-regulatory mutations tend to be located close to the focal gene and often impact functional elements in non-coding sequences that regulate the focal gene's expression, such as promoters or enhancers. By contrast, *trans*-regulatory mutations can be located anywhere in the genome and can impact either coding or non-coding sequences of genes that influence the focal gene's expression through activity of a diffusible molecule such as a protein or RNA. *cis*- and *trans*-regulatory variants contribute differently to the evolution of gene expression (Coolon et al. 2014; Metzger et al. 2017; Signor and Nuzhdin 2018; Gokhman et al. 2021; Hill et al. 2021): *trans*-regulatory variants appear to be the primary source of mRNA expression differences within a species but the relative contribution of *cis*-regulatory variants often increases with evolutionary time. Understanding how and why these classes of regulatory mutations contribute differently to variation in gene expression is important for understanding how gene expression evolves.

Differences in the way *cis*- and *trans*-regulatory mutations affect gene expression might contribute to a preferential fixation of *cis*-regulatory variants relative to *trans* (Wray 2007; Wittkopp et al. 2008; Schaefke et al. 2013; Hill et al. 2021). A *cis*-regulatory mutation alters expression of a focal gene, which can in turn also have effects on expression of downstream genes (orange box in Figure 2-1). By contrast, a *trans*-regulatory mutation affecting expression

of the same focal gene might have effects comparable to the *cis*-regulatory mutation plus independent effects on expression of other genes, each with its own potential downstream consequences (blue box in Figure 2-1). Effects of *cis*-regulatory mutations might thus be a subset of the effects of *trans*-regulatory mutations, and mutations altering expression of the focal gene in *trans* might have more wide-spread effects on gene expression than mutations altering expression of this focal gene in *cis.* Consequently, *trans*-regulatory mutations might be more pleiotropic (i.e., affect more traits) and fixed less often than *cis*-regulatory mutations because mutations that are more pleiotropic are predicted to be more deleterious (Kimura and Ohta 1974). However, regulatory networks are often more complex than shown in Figure 2-1 (Kemmeren et al. 2014), potentially complicating these expectations and making it important to test these ideas empirically (Paaby and Rockman 2013; Zhang and Wagner 2013, Paaby and Rockman 2013b).

Here, we examine the pleiotropic effects of *trans*-regulatory mutations by using *cis*-regulatory mutations to separate the effects of a *trans*-regulatory mutation caused by its impact on a focal gene from its effects caused by impacts on other genes. We separate these mutational effects for fitness and gene expression by measuring relative growth rate and expression profiles for 40 strains of *S. cerevisiae*: 5 with mutations that titrate expression of a focal gene in *cis* and 35 with mutations that alter expression of the same gene in *trans*. The *TDH3* gene in the baker's yeast *Saccharomyces cerevisiae,* which encodes a glyceraldehyde-3-phosphate dehydrogenase (GAPDH), was used as the focal gene for this work because prior studies have systematically identified and isolated *cis*- and *trans*-regulatory mutations that affect its expression (Metzger et al. 2015; Metzger et al. 2016; Duveau et al. 2021). The 5 *cis*-regulatory mutants examined caused expression of *TDH3* to vary from 0% to ~135% of wild-type levels and had mutations in

the *TDH3* promoter, disrupting well-characterized binding sites for the transcription factors RAP1p and GCR1p (Huie et al. 1992; Yagi et al. 1994) (Fig. 2-1B, see Methods). The 35 *trans*-regulatory mutants examined caused *TDH3* expression to vary from ~6% to ~130% of wild-type levels and had mutations in the coding sequences of direct regulators *RAP1* (4 mutants) or *GCR1* (5 mutants) or indirect regulators involved in purine biosynthesis (4 mutants), iron transport (4 mutants), transcriptional regulation (8 mutants), or other processes (10 mutants) (Fig. 2-1C, Table 1).

To separate the fitness effects of *trans*-regulatory mutations attributable to changes in *TDH3* expression from the fitness effects attributable to the pleiotropic impacts of these mutations on other genes, we first defined the relationship between *TDH3* expression and fitness using only the *cis*-regulatory mutants. Relative fitness was estimated for each mutant based on measures of clonal population growth rate under the same conditions used to grow cells for expression profiling (RNA-seq) (see Methods). To predict the fitness effects of any change in *TDH3* expression between 0 and 135% of wild-type expression, we fit a local polynomial regression (LOESS) curve to these data. We found that both increases and decreases in *TDH3* expression decreased fitness (Fig. 2-2A), consistent with prior work using competitive growth to estimate the fitness effects of changing *TDH3* expression (Duveau et al. 2017; Duveau et al. 2018).

Using these inferred effects of changes in *TDH3* expression on fitness, we estimated the pleiotropic fitness effects of each *trans*-regulatory mutant by comparing its measured fitness to the fitness predicted for a *cis*-regulatory mutant with the same change in *TDH3* expression. More specifically, we calculated the pleiotropic fitness effects of *trans*-regulatory mutants as the deviation from the *TDH3* expression -- fitness curve (Fig. 2-2B). Excluding 2 flocculant *trans*-

regulatory mutants for unreliable estimates of growth rate, 52% (17/33) of mutants had significant deleterious pleiotropic effects based on the LOESS regression curve falling above their 95% confidence intervals for fitness. By contrast, 9% (3/33) of mutants had significant beneficial pleiotropic effects based on the LOESS regression curve falling below their 95% confidence intervals. The remaining 13 *trans*-regulatory mutants (39%) showed fitness effects comparable to *cis*-regulatory mutants with similar impacts on *TDH3*. Overall, the empirical distribution of pleiotropic fitness effects was bimodal, with smaller pleiotropic effects skewed toward deleterious effects and larger pleiotropic effects entirely deleterious (Fig. 2-2C). These data provide a rare distribution of pleiotropic effects and direct empirical support for the hypothesis that *trans*-regulatory mutations tend to be more deleterious than *cis*-regulatory mutations, causing them to be more likely to be removed from a population by natural selection.

These differences in the fitness effects of *cis*- and *trans*-regulatory mutations presumably arise from differences in how they impact expression of other genes in the genome; mutants affecting expression of more genes tend to be more deleterious (Featherstone and Broadie 2002) (Fig. 2-6). As described above, *trans*-regulatory mutations might have more widespread effects on the transcriptome than *cis*-regulatory mutations because they affect expression of the focal gene as well as other genes in parallel. To determine whether the *trans*-regulatory mutants generally tended to have more widespread effects on gene expression than the *cis*-regulatory mutants, we compared the number of genes considered significantly differentially expressed in the *cis*- and *trans*-regulatory mutants at a false discovery rate (FDR) of 10%. We found no statistically significant difference in the median number of differentially expressed genes between *cis*- and *trans*-regulatory mutants (Fig. 2-3A, permutation test p-value: 0.11, Fig. 2-7A), but the *trans*-regulatory mutants showed significantly more variable effects (permutation

test p-value = 0.01, Fig. 2-7B). We also compared gene expression using the Euclidean distance among $\log_2$ fold change estimates between each mutant and wild type for all genes other than *TDH3*, which captures the magnitude of expression changes estimated for all genes regardless of statistical significance. Again, we found no significant difference in the median impact on gene expression for *cis*- and *trans*-regulatory mutants (Fig. 2-3B, permutation test p-value: 0.25, Fig. 2-7C) but a greater variance in the effects of *trans*-regulatory mutants (permutation test p-value = 0.02, Fig. 2-7D).

The absence of a larger median effect of *trans*-regulatory mutants relative to *cis*-regulatory mutants might be due to differences in the severity of mutational effects between the sets of *cis*- and *trans*-regulatory mutants examined. To test this possibility, we examined the effects of *cis*- and *trans*-regulatory mutants on the number of significantly differentially expressed genes while taking their impact on *TDH3* expression into account. We found that 83% of *trans*-regulatory mutants showed a greater number of differentially expressed genes than predicted for a *cis*-regulatory mutant with the same effect on *TDH3* expression (Fig. 2-3C). Using Euclidean distances among $\log_2$ fold changes in gene expression rather than the number of significantly differentially expressed genes showed the same pattern (Fig. 2-3D). Prior work has shown that new *trans*-regulatory mutations tend to have smaller effects on *TDH3* expression than new *cis*-regulatory mutations (Metzger et al. 2016), suggesting that greater pleiotropy of *trans*-regulatory mutations is often offset by their smaller effect sizes. Consequently, relative pleiotropy between *cis*- and *trans*-regulatory mutations is not as simple as *trans*-regulatory mutations always tending to be more pleiotropic than *cis*-regulatory mutations. Rather, *trans*-regulatory mutants should only be assumed to be more pleiotropic than *cis*-regulatory mutants when they have comparable effects on expression of the focal gene.

As shown in Figure 2-1, *trans*-regulatory mutants are hypothesized to have more widespread effects than *cis*-regulatory mutants because they are expected to impact expression of genes downstream of the focal gene similarly to *cis*-regulatory mutations but also have additional pleiotropic effects on expression of other genes. This model predicts that the effects of *cis*-regulatory mutations are a subset of the effects of *trans*-regulatory mutations. To determine whether the effects of *cis*-regulatory mutations were indeed a subset of the effects of *trans*-regulatory mutations impacting expression of the same gene, we focused on genes downstream of *TDH3* whose expression was significantly altered when *TDH3* expression was eliminated. Using an FDR of 10%, we identified 154 such downstream genes in the *TDH3* null mutant (Table 2). 55 (36%) of these 154 genes were under-expressed in the null mutant relative to the wild-type strain, and 99 (64%) were over-expressed (Table 2). This gene set was significantly enriched for genes encoding proteins involved in glycolytic processes (Fig. 2-8), suggesting that many expression changes observed in *TDH3* null mutants might be due to a homeostatic response of the cells to maintain metabolism in the absence of the TDH3p enzymatic activity involved in glycolysis and gluconeogenesis (McAlister and Holland 1985). These downstream genes were also enriched for genes associated with the gene ontology terms DNA biosynthesis, integration, and transposition (Fig. 2-8), and expression changes in these genes might be related to non-metabolic functions of TDH3p, such as its interaction with SIR2p to regulate transcriptional silencing and rDNA recombination (Ringel et al. 2013).

The median absolute $\log_2$ fold expression changes observed for this set of 154 genes downstream of *TDH3* decreased monotonically as *TDH3* expression approached wild type, with the smallest median expression change seen in the *cis*-regulatory mutant overexpressing *TDH3* (Fig. 2-4A). To determine how expression of each of these genes scaled with changes in *TDH3*

expression, we used linear regressions to test for significant correlations between *TDH3* expression and expression of each downstream gene in the 5 *cis*-regulatory mutants. 132 (86%) of these 154 genes showed a significant linear relationship with *TDH3* expression at a 10% FDR (Fig. 2-9A), with 49 genes showing a significant positive correlation and 83 genes showing a significant negative correlation (Fig. 2-4B, Fig. 2-9B). For example, the *GPD2* gene, which encodes an NAD-dependent enzyme two steps away from *TDH3* in the metabolic network, showed a strong negative correlation with *TDH3* expression (Fig. 2-4C), indicating that when *TDH3* expression was decreased by *cis*-regulatory mutations, *GPD2* expression increased in response.

To determine whether similar impacts on expression of these 132 downstream genes were observed when *TDH3* expression was altered in *trans*, we used the linear models fitted to the *cis*-regulatory mutant data to predict the change in expression expected for each downstream gene due to the impact of the *trans*-regulatory mutant on *TDH3* expression alone. Deviations from these expectations indicate pleiotropic effects of the *trans*-regulatory mutant on expression of the genes downstream of *TDH3*. For example, 13 of the 35 *trans*-regulatory mutants showed evidence of a pleiotropic effect on expression of *GPD2*, as indicated by a change in *GPD2* expression outside of the 95% confidence interval for the expression change predicted by *cis*-regulatory mutants (Fig. 2-4D). Such pleiotropic effects were observed for every downstream gene in multiple *trans*-regulatory mutants, with the magnitude of the pleiotropic effects (measured as residuals from the gene-specific regression models based on the *cis*-regulatory mutants) varying among *trans*-regulatory mutants and genes (Fig. 2-4E). These data indicate that the *trans*-regulatory mutants examined in this study often changed the relationship between *TDH3* expression and expression of its downstream genes. Consequently, the effects of *cis*-

regulatory mutations altering *TDH3* expression were not a simple subset of the effects of *trans*-regulatory mutants also altering *TDH3* expression.

The deviation in effects between *cis*- and *trans*-regulatory mutations on expression of genes downstream of *TDH3* differed among *trans*-regulatory mutants with mutations in different genes as well as among mutants with different mutations in the same gene (Fig. 2-5A). Hierarchical clustering of the residuals for genes downstream of *TDH3* also showed these different impacts (Fig. 2-5B). For example, four *trans*-regulatory mutants all caused large decreases in *TDH3* expression, but the two mutants with mutations in *GCR1* (GCR1339 and GCR1162) had different impacts on expression of genes downstream of *TDH3* than the two mutants with mutations in *RAP1* (RAP154 and RAP1238) (Fig. 2-5B). Each pair of mutants with mutations in the same gene had similar effects on expression of these downstream genes. By contrast, two other mutant alleles of *RAP1* (RAP1484 and RAP1357) both increased expression of *TDH3* but had distinct impacts on expression of genes downstream of *TDH3* (Fig. 2-5B). These different impacts of the four *RAP1* mutant alleles underscore that different coding mutations in the same gene can have different pleiotropic effects (Lynch and Wagner 2008). In other cases, pleiotropic effects for mutants with mutations in different genes that function in the same pathway (e.g., ADE4, ADE5, ADE6) were similar (Fig. 2-5B). Taken together, these data illustrate that *trans*-regulatory mutants have impacts on expression of genes downstream of *TDH3* that deviate from the effects of *cis*-regulatory mutants in a variety of ways, implying additional connections between *trans*-regulators of *TDH3* and its downstream genes not shown in Figure 2-1.

By quantifying and comparing the effects of *cis*- and *trans*-regulatory mutations on gene expression and fitness, this work provides an important complement to studies describing the

relative contributions of *cis*- and *trans*-regulatory variants to expression differences within and between species because it can help explain why we see the variation we see. For example, the tendency of *trans*-regulatory mutations to be more pleiotropic than *cis*-regulatory mutants suggests that the increasing contribution of *cis*-regulatory variation to expression differences over evolutionary time (Signor and Nuzhdin 2018; Hill et al. 2021) is due to natural selection preferentially removing *trans*-regulatory mutations. However, our data suggests that statements about the relative fitness of *cis*-and *trans*-regulatory mutations should be more nuanced, considering the relative effect sizes of *cis*- and *trans*-regulatory mutations when predicting their evolutionary fates. Differences in the rate at which new *cis*- and *trans*-regulatory mutations arise as well as their relative dominance in diploid species are also expected to contribute to the patterns of regulatory variation seen in natural populations. Ultimately, understanding both the properties of new regulatory mutations and the sources of regulatory variation seen in the wild are needed to understand the evolution of gene expression.

Materials and Methods

Yeast Genotypes

Strains of *S. cerevisiae* bearing *cis*-regulatory mutations used in this study are a subset of the strains used to assay the fitness effects of changing *TDH3* expression constructed and described in Duveau et al (2017). They are haploid, mating type **a** strains of *S. cerevisiae* derived from S288C and constructed from the progenitor strain YPW1001, which contains a wild type P_{TDH3}-*YFP* construct and a *NatMX4* drug resistance marker at the *HO* locus and alleles of *MKT1*, *SAL1*, *CAT5* and *MIP1* decreasing petite frequency and the alleles of *RME1* and *TAO3* increasing sporulation efficiency, as previously described (Duveau et al. 2017). The strains

specifically used in this study include: (1) YPW1177: a strain with a deletion of the entire native *TDH3* promoter and coding sequence (0% of wild-type expression), (2) YPW1156: a strain with a C->T point mutation in the native *TDH3* promoter 482 bp upstream of the *TDH3* start codon in a binding site for the GCR1 transcription factor (~20% of wild-type expression), (3) YPW1200: a strain with a C->T point mutation in the native *TDH3* promoter 485 bp upstream of the start codon in the same binding site for GCR1 (~50% of wild-type expression), (4) YPW1188: a strain with a G->A point mutation in the native *TDH3* promoter 510 bp upstream of the start codon in a binding site for the RAP1 transcription factor (~85% of wild-type expression), (5) YPW1189: a strain with the wild type *TDH3* promoter and coding sequence used as a reference strain for the previous four mutant strains, (6) YPW3059: a strain with two copies of the *TDH3* gene separated by a *URA3* selectable marker in which each copy of *TDH3* contained a G->A point mutation in its promoter 505 bp upstream of the start codon in a RAP1 binding site (resulting in a total ~135% of wild-type expression), and (7) YPW2682: a strain with the same *URA3* selectable marker as the strain with the *TDH3* gene duplication inserted after the native *TDH3* locus used as a reference strain for the overexpression mutant strain.

The 35 *trans*-regulatory mutants analyzed in this study (Table 1) include a subset of those described in Duveau et al (*2021*). Briefly, mutants analyzed with mutations in *GCR1* or *RAP1* were constructed by using mutagenic PCR to randomly introduce mutations within each gene and then using CRISPR-mediated allele-replacement to substitute the native locus with a mutant allele (Duveau et al. 2021). The 9 *trans*-regulatory mutants with mutations in one of these two genes each contained 1 to 6 mutations (Table 1). The remaining 26 *trans*-regulatory mutants analyzed each contained a single nucleotide change introduced into the genome by site-directed mutagenesis and either *delitto perfetto* or CRISPR, with the specific mutation introduced

identified by genetic mapping of mutant genotypes isolated from an EMS mutagenesis screen for altered expression of *TDH3* (Duveau et al. 2021). *Trans*-regulatory mutant strains and the corresponding reference strain are haploid, mating type α strains of *S. cerevisiae* derived from S288C and constructed from the progenitor strain YPW1139, which also contains a wild type P_{TDH3}-*YFP* construct and a *KanMX* drug resistance marker at the *HO* locus and alleles of *MKT1*, *SAL1*, *CAT5* and *MIP1* decreasing petite frequency and the alleles of *RME1* and *TAO3* increasing sporulation efficiency, as previously described (Duveau et al. 2017). This same progenitor strain YPW1139, re-stocked and renamed as YPW3016, was used as the reference strain for these *trans*-regulatory mutant strains.

These 43 strains (5 *cis*-regulatory mutants, 35 *trans*-regulatory mutants, and 3 reference strains), as well as 3 deletion mutants not used in this study, were randomly arrayed into a 96-well plate containing YPD media (2% dextrose monohydrate, 2% peptone, 1% yeast extract, weight to volume in milliQ purified water and sterilized by autoclave), including 3 replicates of the reference for *trans*-regulatory mutants. The outer rows and columns of this 96-well plate were filled only with sterile media because slight differences in yeast growth at the outer wells of the 96-well plate had been previously observed. 4 unique random plate arrays, each containing all mutants and 3 replicates of the *trans*-regulatory reference strain but in different plate positions were designed and assembled. Each plate was then grown to saturation in YPD media with glass beads while shaking at 250 rpm to maintain suspension. 100 uL of each culture in these four plates was mixed with 23 uL of 80% glycerol and stored at -80 C.

Estimating relative fitness

Relative fitness was estimated based on quadruplicate measures of growth rate for each genotype. A pin tool was used to transfer cells from the four replicate 96-well plates containing glycerol stocks to a solid YPG (5% glycerol by volume, 2% peptone weight to volume, 2% agar weight to volume, 1% yeast extract weight to volume in milliQ purified water and sterilized by autoclave) agar plates to prevent the formation of the petite phenotype which can be common upon thawing. These plates were then incubated at 30 C for ~3 days to allow colony growth. Cells from each genotype were then transferred using a pin tool from the agar plate into 500 uL of liquid YPD media in a 1 mL plate with glass beads and grown with shaking at 30 C for ~3.5 days. 5 uL of these saturated cultures were transferred into 100 uL of fresh YPD liquid media in a Costar 96 well plate with lid, which was then inserted into a BioTek Synergy (Agilent). Cells were grown for 24 hrs at 30 C while being continually shaken to maintain suspension, pausing to take optical density measurements of each well, including blank control wells, every hour. Two strains, containing a mutation in either *CYC8* or *SSN2* showed visible evidence of flocculation (a known phenotype for deletions of both of these genes (Cherry et al. 2012), and were excluded from growth rate analysis and subsequent analyses using growth rate measurements. Optical density curves from 0 to 18 hrs (when the diauxic shift occurred), were then plotted and fit to a sigmoidal growth curve using the R program 'growthcurver' (Sprouffske and Wagner 2016). Maximal growth rates, calculated as the maximal slope of each fitted curve, were used to calculate growth rate relative to the appropriate control strain to yield average relative growth rates and standard errors (scripts used for analysis and raw data available at Github).

RNA extraction

Cells used for RNA-seq were sampled from the 4 replicate plate glycerol stocks described above. A pin tool was used to transfer cells from each glycerol stock plate to solid YPG agar plates and cells were grown for ~3 days at 30 C to prevent the formation of the petite phenotype which can be common upon thawing. Cells were then transferred by pin tool again into 500 uL liquid YPD media in 1 mL 96-well plates with glass beads and grown at 30 C while shaking at 250 rpm for 2 days, until all strains were once again at saturation. 100 uL of the saturated culture was then transferred to a Costar 96 well plate and OD measured using a Tecan Sunrise (Tecan). The OD was then used to calculate cell density. A separate 1mL 96 well plate was then filled with 500 uL of YPD in each well, and each well inoculated with a volume of the saturated culture calculated to grow to a cell density of $5x10^6$ cell per mL (an OD of about 0.4) after 12.5 hr. These cultures were then grown rotating on a wheel at 30 C for 12.5 hrs, after which 100uL was removed and used to measure OD to ensure all cultures were between an OD of 0.26 to 0.48. Plates were then centrifuged for 5 min at 3000 g, and liquid media pipetted off. Remaining cell pellets were frozen by plunging the entire plate containing cell pellets into liquid nitrogen. Plates were sealed with foil and stored at -80 C until RNA extractions were performed. For any very slow growing strains that did not reach an OD of between 0.26 and 0.48, this process was repeated and multiple cell pellets pooled at the RNA extraction stage to achieve uniform cell numbers across all strains and replicates.

Frozen cell pellets were resuspended in 700 uL of lysis buffer (100mM Tris-HCl, pH 7.5, 500 mM LiCl, 10mMEDTA, pH 8, 1% LiDS, 5mM DTT) containing beta-mercaptoethanol and transferred to a plate containing ~250 uL of acid-washed 425-600 um beads. These plates were vortexed 10x for 1min each with 1min on ice in between. Plates were centrifuged at 3000 rpm at 4 C for 4 min, and 400 uL of lysis supernatant were removed and transferred into a new 96 well

plate containing 50 uL oligodT magnetic beads resuspended in lysis buffer (Dynabeads mRNA DIRECT Kit, Ambion, cat# 61011). Beads and lysate were incubated at room temperature with agitation for 5 minutes, and then placed on a magnetic stand. Supernatant was pipetted off and beads were washed 2x with Wash Buffer A (10mM Tris-HCl, pH 7.5, 0.15 M LiCl, 1 mM EDTA, 0.1% LiDS) and 2x with Wash Buffer B (10mM Tris-HCl, pH 7.5, 0.15 M LiCl, and 1mM EDTA) and then eluted in 10 uL of elution buffer (10mM Tris-HCl, pH 7.5) for 2 min at 72 C. After incubation, plate was placed immediately back on magnetic stand and supernatant containing eluted RNA transferred to an RNAse-free plate. After extraction, several random samples were run on Agilent Bioanalyzer to check RNA quality and concentration before moving on to library preparation. In addition, selected wells were tested for the presence of *NatMX* or *KanMX* resistance markers present in distinct strains and controls by RT-PCR and visualized on a gel to determine whether well-cross contamination had taken place at any step previous to library prep; we found no evidence of such cross contamination.

RNA-seq pipeline and DESeq2 analysis

RNA-seq libraries were prepared using ⅓ volume reactions from TruSeq RNA Sample Preparation v2 Kit, and using multiplexable adapters from the TruSeq RNA CD Index Plate (cat# 20019792). Ten of the 198 library preps used in this study failed, resulting in those strains being analyzed in triplicate rather than quadruplicate. All samples from the same replicate plate were pooled and run on one lane on the Illumina HiSeq 4000 by the University of Michigan Advanced Genomics Core, for a total of 4 replicate plates run on 4 sequencing lanes. Raw reads were run through the FastQC (version 0.11.5) read quality software and passed the program benchmarks for read quality. Reads were then trimmed using Cutadapt ((Martin 2011), version 1.10) and

pseudo-mapped to the a transcriptome index generated using Salmon ((Patro et al. 2017), version 0.9.1) on *S. cerevisiae* cDNA (Ensemble, release 38, retrieved from ftp://ftp.ensemblgenomes.org/pub/release-38/fungi/fasta/saccharomyces_cerevisiae/cdna/). Read counts from Salmon were imported into R using TxDb (version 3.2.2). Read counts for the entire dataset were supplied to DESeq2 (Love et al. 2014) to model gene expression levels, with different strain backgrounds being analyzed separately. The DESeq2 'contrast' wrapper was then used to estimate $\log_2$ fold changes for each strain relative to its appropriate reference and identify differentially expressed genes and estimate $\log_2$ fold changes in expression relative to the reference. We used PCA analysis to identify any outliers amongst the replicates sequenced and identified one *trans*-regulatory mutant control replicate as well as one of each of the *TDH3 cis*-regulatory mutants with expression of 0%, 20%, 50%, and 85%, all coming from the fourth replicate plate, as outliers (Fig. 2-10). These samples were excluded from the analyses.

To assess the reliability of DESeq2 estimates of expression levels from our RNA-seq data, we compared the RNA-seq expression measures for the *cis*-regulatory mutants to previously published expression driven by the same set of *cis*-regulatory mutant alleles controlling expression of a fluorescent reporter gene (Duveau et al. 2017). However, because the reporter gene was at the *HO* locus, the fluorescence measures were first compared to fluorescence measures in set of strains with *cis*-regulatory mutations at the native *TDH3* locus driving expression of a TDH3-YFP fusion protein (Duveau et al. 2018). This relationship was used to predict expression values at the native locus for strains that contained *cis*-regulatory mutations that were not present in a fusion protein strain themselves. These predictions were then compared with the expression values estimated directly from the RNA-seq data. RNA-seq and fluorescence estimates were strongly correlated with an r^2 value of 0.97 (Fig. 2-11).

Comparing the RNA-seq data for reference strains with a single copy of *TDH3* that were mating type α (YPW3016) and mating type **a** (YPW1189) showed that 35 genes were significantly differentially expressed between these two strains, 28 of which were annotated as mating type genes, dubious, or uncharacterized ORFs, and 7 of which were annotated with different functions (Fig. 2-12, Table 3). These genes were not significantly differentially expressed between the reference strain and mutant strains of the same mating type, indicating that all differentially expressed genes in mutant strains are attributable to the *cis*- or *trans*-regulatory mutations affecting *TDH3* expression. Expression of URA3 was discovered in 5 of the *trans*-regulatory mutant lines, suggesting that they had not lost the URA3 containing plasmid used during strain construction (Duveau et al. 2021). To ensure that these strains did not influence the patterns of analysis presented in the main text, all analyses were carried out with and without these strains, without affecting the conclusions. Analyses with these 5 mutants excluded are included in the scripts provided at Github.

Statistical analysis

Permutation tests were conducted to assess differences in the number of differentially expressed genes or the Euclidean distance between $\log_2$-fold changes for *cis*- and *trans*-regulatory mutants. These permutation tests were used because they take into account the differences in sample size for *cis*- and *trans*-regulatory mutants. Because there are 5 *cis*-regulatory mutants, 5 *trans*-regulatory mutants were drawn from the total of 35 for each permutation (without replacement), and the median and variance of each sample was calculated. This sampling was repeated 1000 times to create a distribution of the medians and variances measured for 5 *trans*-regulatory mutants, which was then compared to the observed median and

variance values for the 5 *cis*-regulatory mutants. P-values were calculated as the proportion of the 1000 random samples with a median less than or equal to the median or variance of the *cis*-regulatory mutants (Fig. 2-7).

Gene ontology analysis was performed using the *Saccharomyces* genome database (SGD) 'GO Term Finder' (Version 0.86) tool (Cherry et al. 2012). Both gene sets were tested for enrichment against a background set of all genes included for analysis in our RNA-seq experiment. GO process terms significant at an FDR of 0.01 were examined, and those terms with direct gene associations with the most significant enrichments (rather than parent terms) were reported.

To compare the downstream effects of perturbing *TDH3* expression via *cis*- and *trans*-regulatory mutations, for each gene significantly differentially expressed in the *TDH3* null mutant (Table 2), the $\log_2$ fold change estimates across all *cis*-regulatory mutants were converted to percent wild-type estimates. A linear model was then fit to each gene's expression level regressed on the percent wild-type expression level of *TDH3* across all *cis*-regulatory mutants using the base R 'lm' function. We then applied a Benjamini-Hochberg multiple testing correction to the p-values generated from the F-test of the linear regression to identify the 132 genes that were significantly linearly related to *TDH3* expression at an FDR of 10%. These linear models were then used to predict each of the 132 genes' expression levels in *trans*-regulatory mutants using *TDH3* expression level in each mutant as the predictor. The absolute difference between the predicted gene expression level and the actual observed gene expression level (i.e., the residual) was calculated for each mutant for each of the 132 genes. All scripts used to perform these analyses are available on Github.

Acknowledgments

We thank Mark Hill, Brian Metzger, Fabien Duveau, Mo Siddiq, Henry Ertl, Anna Redgrave, and other members of the Wittkopp lab for helpful discussions and feedback on drafts of the manuscript. We also thank Bin Z. He (University of Iowa) for sharing the protocol used for RNA-seq. The University of Michigan Advanced Genomics Core, Center for Statistical Consultation and Research, and High-Performance Computing Cluster provided services used to conduct this work. Funding for this work was provided by National Institutes of Health grants R35GM118073 and R01GM108826 to P.J.W. and T32GM07544 to P.V.Z.

References

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40

Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. 2014. Tempo and mode of regulatory evolution in Drosophila. *Genome Res.* 24:797–808.

Duveau F, Hodgins-Davis A, Metzger BPH, Yang B, Tryban S, Walker EA, Lybrook T, Wittkopp PJ. 2018. Fitness effects of altering gene expression noise in Saccharomyces cerevisiae. *Elife* [Internet] 7. Available from: https://elifesciences.org/articles/37272

Duveau F, Toubiana W, Wittkopp PJ. 2017. Fitness effects of cis -regulatory variants in the Saccharomyces cerevisiae TDH3 promoter.

Duveau F, Vande Zande P, Metzger BP, Diaz CJ, Walker EA, Tryban S, Siddiq MA, Yang B, Wittkopp PJ. 2021. Mutational sources of trans-regulatory variation affecting gene expression in Saccharomyces cerevisiae. Elife [Internet] 10. Available from: http://dx.doi.org/10.7554/eLife.67806

Featherstone DE, Broadie K. 2002. Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. *Bioessays* 24:267–274.

Gokhman D, Agoglia RM, Kinnebrew M, Gordon W, Sun D, Bajpai VK, Naqvi S, Chen Coral, Chan A, Chen Chider, et al. 2021. Human-chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. *Nat. Genet.* [Internet]. Available from: http://dx.doi.org/10.1038/s41588-021-00804-3

Hill MS, Vande Zande P, Wittkopp PJ. 2021. Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* 22:203–215.

Huie MA, Scott EW, Drazinic CM, Lopez MC, Hornstra IK, Yang TP, Baker HV. 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.

Kemmeren P, Sameith K, van de Pasch LAL, Benschop JJ, Lenstra TL, Margaritis T, O'Duibhir E, Apweiler E, van Wageningen S, Ko CW, et al. 2014. Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. *Cell* 157:740–752.

Kimura M, Ohta T. 1974. On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 71:2848–2852.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.

Lynch VJ, Wagner GP. 2008. Resurrecting the role of transcription factor change in developmental evolution. *Evolution* 62:2131–2154.

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

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

Metzger BPH, Duveau F, Yuan DC, Tryban S, Yang B, Patricia J, Arbor A, Arbor A, Wittkopp PJ, Biology E, et al. 2016. Contrasting frequencies and effects of cis- and trans-regulatory mutations affecting gene expression. :1–35.

Metzger BPH, Wittkopp PJ, Coolon JD. 2017. Evolutionary dynamics of regulatory changes underlying gene expression divergence among Saccharomyces species. *Genome Biol. Evol.* 177:1987–1996.

Metzger BPH, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ. 2015. Selection on noise constrains variation in a eukaryotic promoter. *Nature* 521:344–347.

Paaby AB, Rockman MV. 2013. The many faces of pleiotropy. *Trends Genet.* 29:66-73

Paaby AB, Rockman MV. 2013. Pleiotropy: what do you mean? Reply to Zhang and Wagner. *Trends Genet.* 29:384.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14:417–419.

Ringel AE, Ryznar R, Picariello H, Huang K-L, Lazarus AG, Holmes SG. 2013. Yeast Tdh3 (Glyceraldehyde 3-Phosphate Dehydrogenase) Is a Sir2-Interacting Factor That Regulates Transcriptional Silencing and rDNA Recombination.Pikaard CS, editor. *PLoS Genet.* 9:e1003871.

Schaefke B, Emerson JJ, Wang TY, Lu MYJ, Hsieh LC, Li WH. 2013. Inheritance of gene expression level and selective constraints on trans-and cis-regulatory changes in yeast. *Mol. Biol. Evol.* 30:2121–2133.

Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. *Trends Genet.* 34:532–544.

Sprouffske K, Wagner A. 2016. Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics* 17:172.

Stern DL, Orgogozo V. 2008. The loci of evolution: How predictable is genetic evolution? :2155–2177.

Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between Drosophila species. *Nat. Genet.* 40:346–350.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8:206–216.

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 J, Wagner GP. 2013. On the definition and measurement of pleiotropy. *Trends Genet.* 29:383–384.

Figures

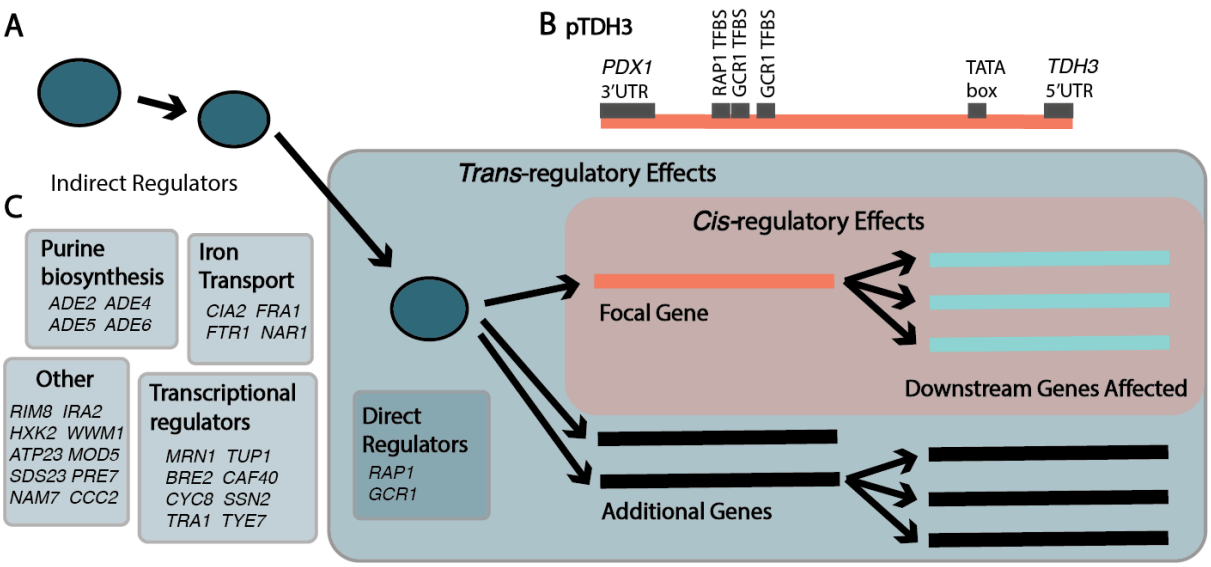

Figure 2-1: cis- and trans-regulatory mutations have different effects on gene expression

(A) *Trans*-regulatory mutations in either indirect or direct regulators (blue) influence expression of a focal gene (orange), which in turn influences expression of downstream genes (light blue). Mutations in *trans*-regulators can also influence expression of additional genes in the genome (black), suggesting the effects of *trans*-regulatory mutations (blue box) should be more widespread than the effects of mutations that affect the focal gene's expression in *cis* (orange box). (**B**) Schematic shows the *cis*-regulatory sequence (promoter) for the *S. cerevisiae TDH3* gene (*pTDH3*) used as a focal gene for this work. This promoter is in the intergenic sequence between *PDX1* and *TDH3* and includes transcription factor binding sites (TFBS) for the direct regulators (transcription factors) encoded by the *RAP1* and *GCR1* genes as well as a TATA box. (**C**) Previously identified (Duveau et al. 2021) indirect regulators of *TDH3* expression harboring *trans*-regulatory mutations tested in this work are shown.

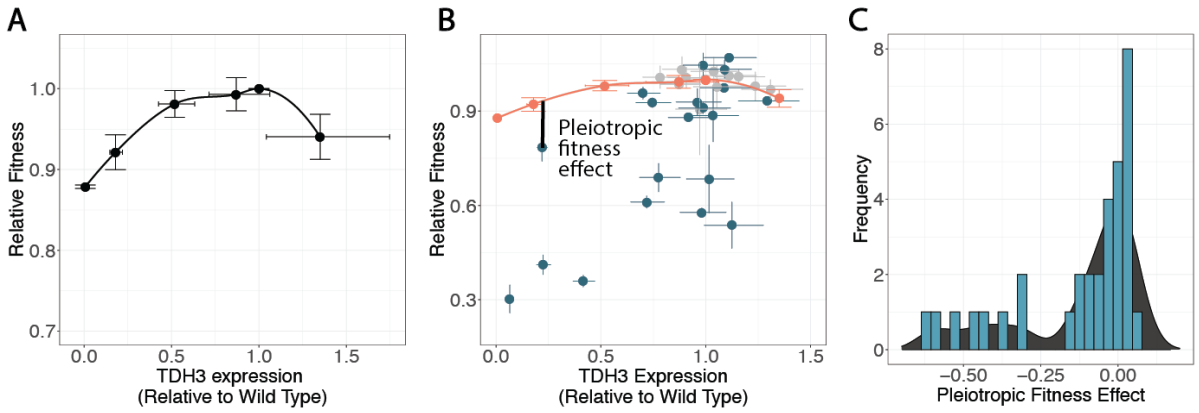

Figure 2-2: Pleiotropic effects of trans-regulatory mutations on fitness

(**A**) Relative fitness is shown for 5 *cis*-regulatory mutants and a wild-type strain based on the level of *TDH3* expression in each strain, with error bars representing 95% confidence intervals. *TDH3* expression from RNA-seq data is plotted on the x-axis, with error bars representing one standard error. A local polynomial regression (LOESS) fit line is also shown. (**B**) Relative fitness and *TDH3* expression of *trans*-regulatory mutants is shown, with the *cis*-regulatory mutants and fitted LOESS curve from (A) included for comparison in orange. The degree to which relative fitness of a *trans*-regulatory mutant deviates from the relative fitness predicted for a *cis*-regulatory mutant with similar effects on *TDH3* expression (orange line) is defined as the pleiotropic fitness effect of that *trans*-regulatory mutant (example shown with solid black line). *Trans*-regulatory mutants with significant pleiotropic fitness effects (95% confidence intervals for fitness that do not overlap the LOESS fit line for *cis*-regulatory mutants) are shown in blue. (**C**) Histogram summarizes pleiotropic fitness effects of all *trans*-regulatory mutants, as defined in panel B. A smoothed density distribution derived from this histogram is underlaid in black.

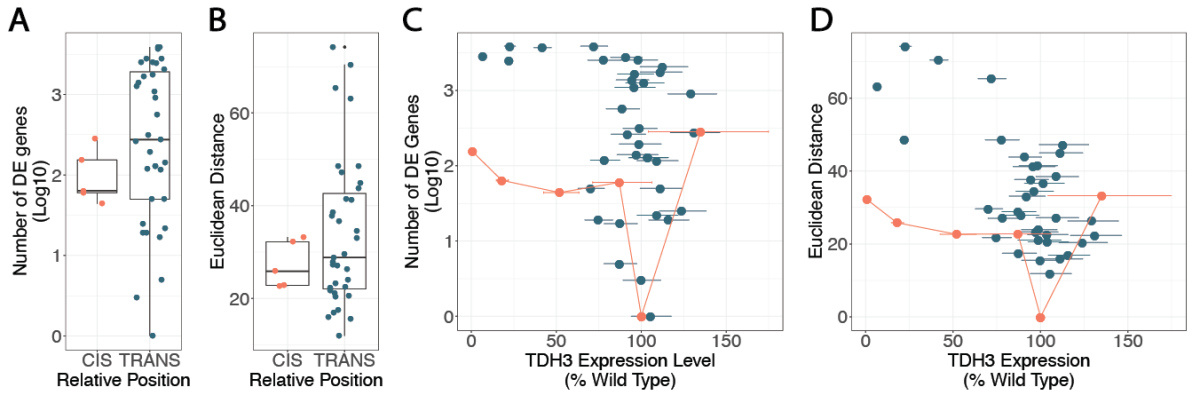

Figure 2-3: trans-regulatory mutants have broader impacts on expression than cis-regulatory mutants only when they have similar effects on the focal gene

(**A**) The number of significantly differentially expressed (DE) genes at a 10% FDR is shown for *cis*-regulatory mutants (orange) and *trans*-regulatory mutants (blue). (**B**) Euclidean distances among $\log_2$ fold changes in expression are shown for *cis*-regulatory (orange) and *trans*-regulatory (blue) mutants. Box plots in (A) and (B) show median and quartile values. (**C**) The number of significantly differentially expressed genes at a 10% FDR is shown for each mutant, plotted according to the mutant's impact on *TDH3* expression. (**D**) Euclidean distances among $\log_2$ fold changes in expression are shown for each mutant, plotted according to the mutant's impact on *TDH3* expression. In both (C) and (D), *cis*-regulatory mutants are shown in orange, with points connected by straight line segments. Error bars for *TDH3* expression are one standard error from RNA-seq data.

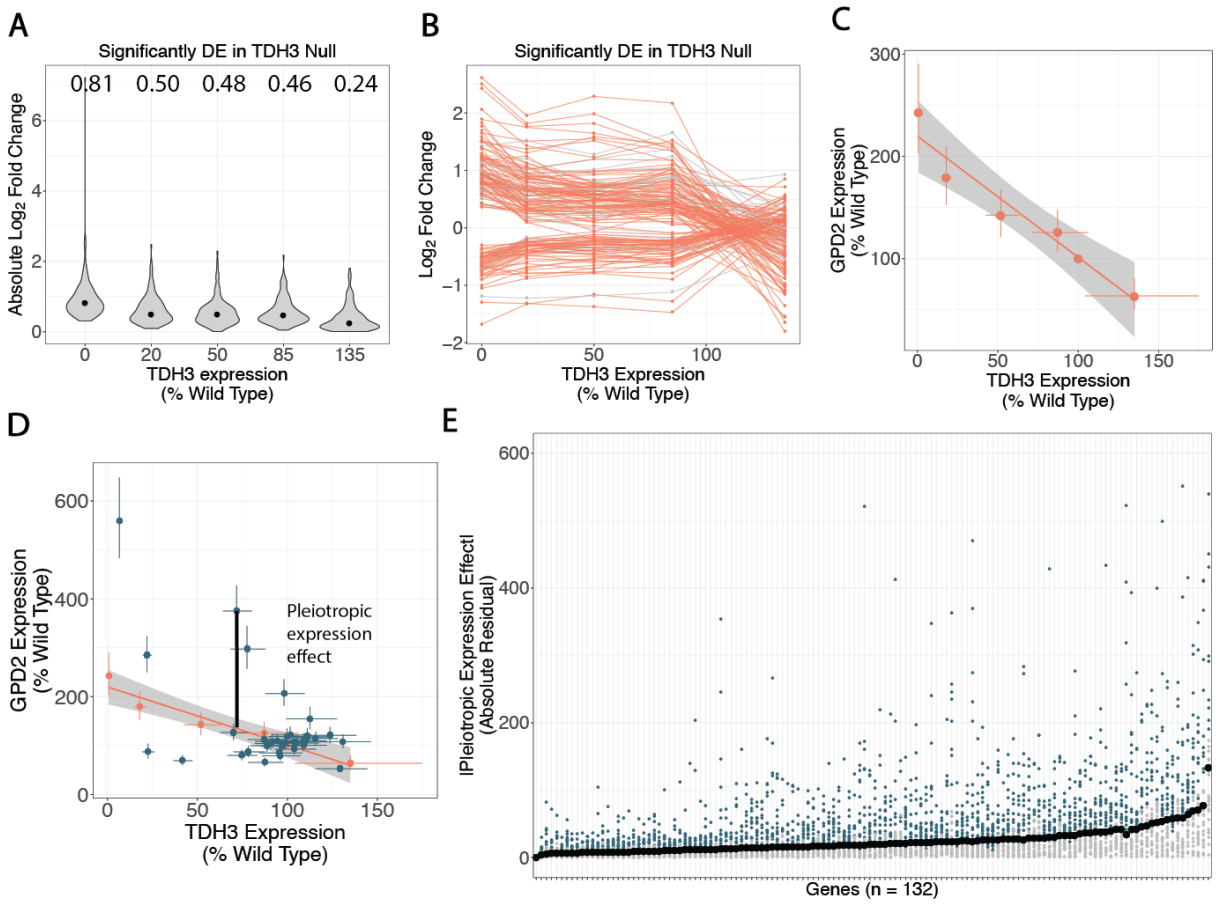

Figure 2-4: cis- and trans-regulatory mutants have distinct effects on expression of genes downstream of TDH3

(A) Violin plots show absolute $\log_2$ fold changes in the 5 *cis*-regulatory mutants for the 153 genes other than *TDH3* that were significantly differentially expressed (DE) in the *TDH3* null mutant. Median absolute $\log_2$ fold changes are shown above each plot and indicated with black dots. (**B**) Log_2 fold changes in *cis*-regulatory mutants are shown for the 153 genes downstream of *TDH3*, with expression changes for the same gene connected by line segments. Genes whose expression was not significantly linearly correlated with *TDH3* expression are shown in grey. (**C**) Expression of *TDH3* and *GPD2* is shown for the *cis*-regulatory mutants. The best fit linear regression line and 95% confidence interval (grey shaded area) are shown with error bars representing one standard error of the $\log_2$ fold change. (**D**) Expression of *GPD2* and *TDH3* is shown for the *trans*-regulatory mutants (blue), with the expression and linear regression from *cis*-regulatory mutants shown in (C) included in orange for comparison. Effects of *trans*-regulatory mutants on *GPD2* that are not explained by their impact on *TDH3* were considered pleiotropic expression effects of the *trans*-regulatory mutant (illustrated by a solid black line for one *trans*-regulatory mutant). (**E**) For each of the 132 genes downstream of *TDH3* with a significant linear relationship to *TDH3* expression in the *cis*-regulatory mutants (x-axis), the pleiotropic expression effect is shown as the absolute value of the residual from fitting *trans*-regulatory mutant expression levels to the gene-specific linear regression model defined by the *cis*-regulatory mutant expression data, as illustrate in D (y-axis). For each gene, each point represents a different *trans*- regulatory mutant. Genes are ordered on the x-axis by median

absolute residual (black points). Blue points indicate *trans*-regulatory mutants that lie outside of the 95% confidence interval for the linear model fit to the *cis*-regulatory mutants. 8 points representing *trans*-regulatory mutants with absolute residuals between 600 and 1000 are not shown for better resolution of the rest of the data

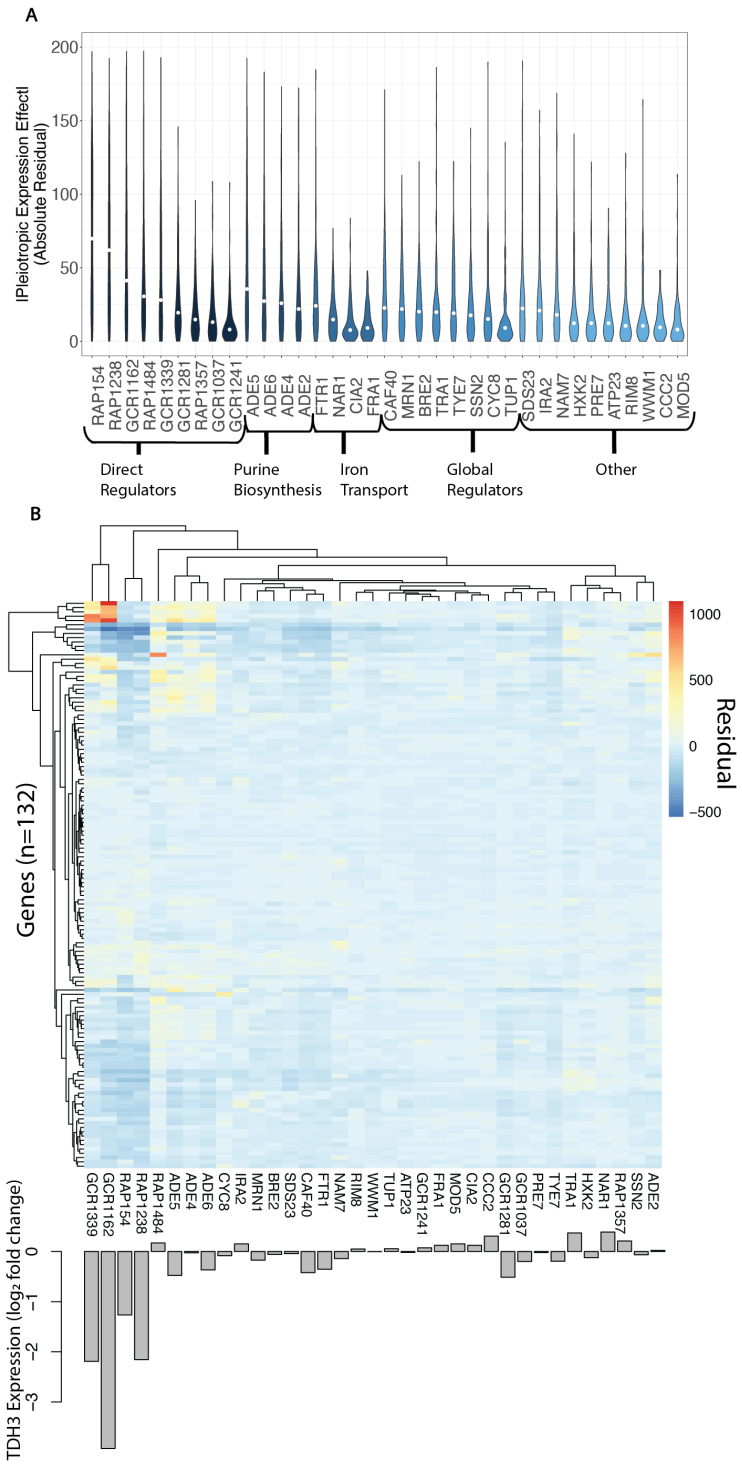

Figure 2-5: trans-regulatory mutations have diverse pleiotropic effects on expression of genes downstream of TDH3

(A) For each *trans*-regulatory mutant, violin plots show the distribution of absolute pleiotropic expression effects (measured as absolute residuals from a linear model, as described in Figure 4E) for the 132 genes that were significantly differentially expressed in the *TDH3* null mutant (i.e., downstream of *TDH3*) and had expression significantly correlated with *TDH3* in the *cis*-regulatory mutants. *Trans*-regulatory mutants are shown grouped according to the functional categories shown in Figure 1C and ordered within each group by the median pleiotropic expression effect. The y-axis is truncated at 200 to better visualize most of the data. (**B**) A heatmap of residuals from the 132 gene-specific linear models based on expression of genes downstream of *TDH3* in *cis*-regulatory mutants (rows) is shown with hierarchical clustering used to group *trans*-regulatory mutants based on these pleiotropic expression effects (columns). The $\log_2$-fold change in *TDH3* expression level relative to wildtype in each *trans*-regulatory mutant is shown in the bar chart below.

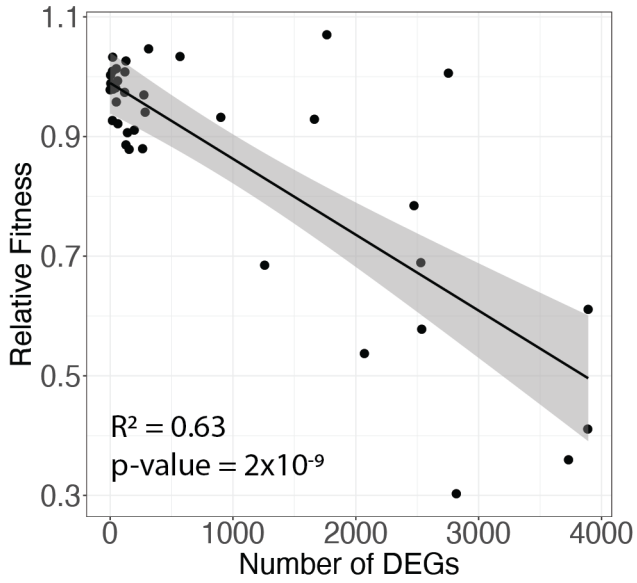

Figure 2-6: The number of differentially expressed genes correlates with relative fitness

The number of significantly differentially expressed genes (DEGs) at an FDR of 10% is shown plotted against relative fitness (relative growth rate) for all *cis*-regulatory and *trans*-regulatory mutants. The best-fit line from a linear regression of relative fitness on the number of differentially expressed genes in all mutants is shown, with the 95% confidence interval for the fit line shaded gray.

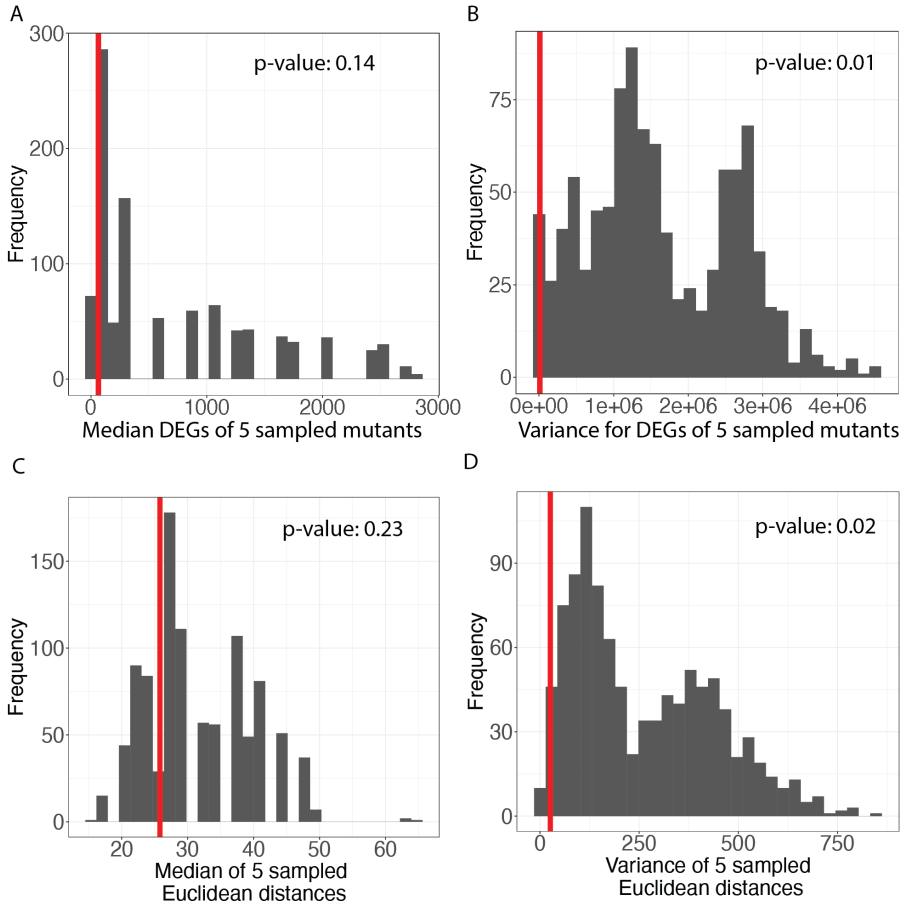

Figure 2-7: Permutation tests for comparing effects of cis- and trans-regulatory mutations on gene expression

(**A**) Histogram (in grey) shows the median number of significantly differentially expressed genes (DEGs) for sets of 5 *trans*-regulatory mutants randomly sampled from the total set of 35 without replacement (1000 permutations). Red line is the median of the 5 *cis*-regulatory mutants. (**B**) Histogram (in grey) shows the variance in the number of DEGs for sets of 5 *trans*-regulatory mutations randomly sampled from the total set of 35 without replacement (1000 permutations). Red line is the variance of the 5 *cis*-regulatory mutants. (**C,D**) The same information is shown as in A and B, but for Euclidean distances among $\log_2$-fold changes rather than the number of DEGs.

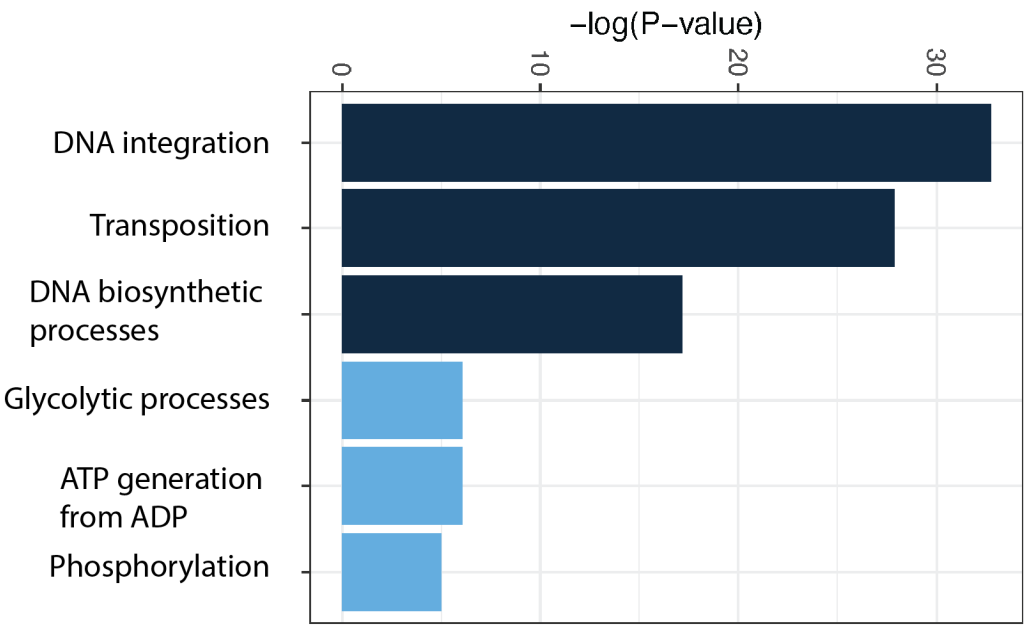

Figure 2-8: Gene Ontology (GO) terms enriched in genes differentially expressed in the TDH3 null mutant

Terms significantly enriched for genes significantly differentially expressed in the *TDH3* null mutant are shown. The three terms associated with transposition are shown in dark blue, and terms associated with glycolysis shown in light blue. Negative $\log_{10}$ P-values were calculated using the SGD online GO term finder tool, using the total set of genes analyzed in this study as the background gene list for enrichment.

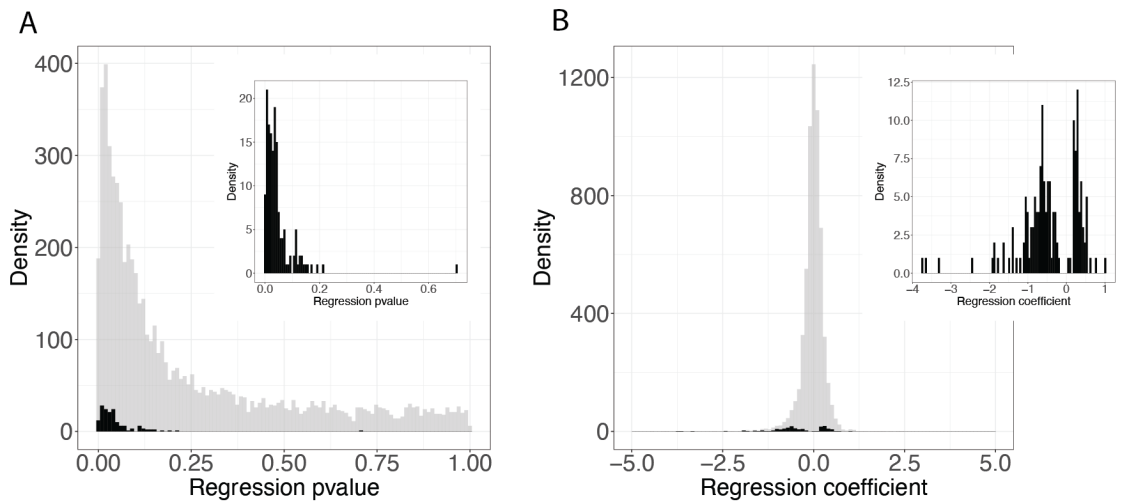

Figure 2-9: Linear relationships between TDH3 expression and expression of other genes in cis-regulatory mutants

(**A**) A linear regression was performed for each gene included in the study (n = 6,128) regressed on *TDH3* values for *cis*-regulatory mutants. Histograms show the p-value of the F-test calculating *TDH3* level's additional information as a predictor in the linear model. All genes are shown in grey. Genes significantly differentially expressed in the *TDH3* null mutant are shown in black and enlarged in the inset. (**B**) The coefficients (slopes) of the linear models in panel A calculated for all genes in the study (grey). As in A, genes that are significantly differentially expressed in the *TDH3* null mutant are shown in black and enlarged in inset.

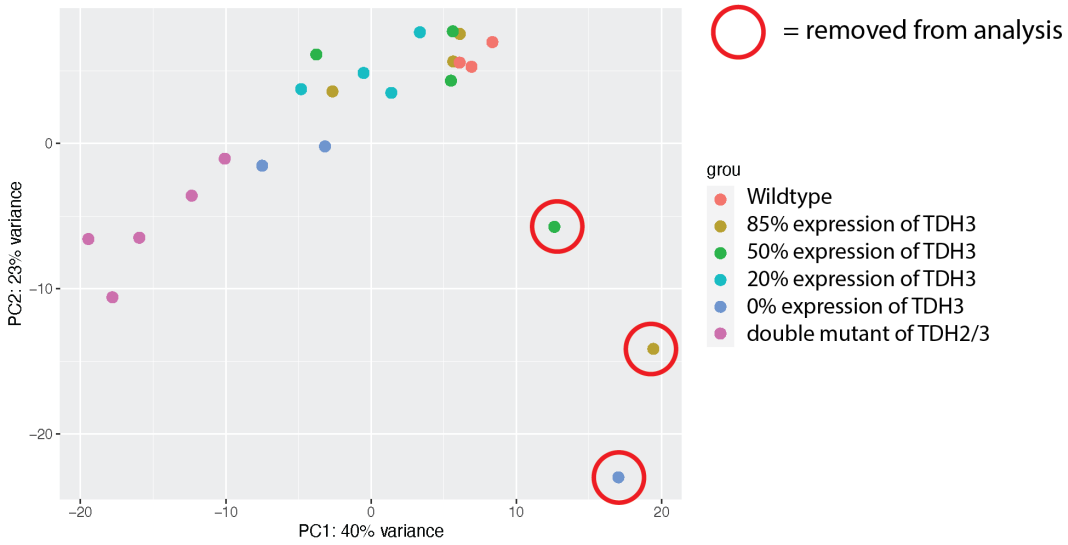

Figure 2-10: Principle Components Analysis (PCA) of cis-regulatory mutant strains and references showing removal of outlier samples

PCA analysis performed on a variance stabilizing transformed count matrix of all genes with more than 10 reads in all mating type **a** strains, which includes under-expression *cis*-regulatory mutants and reference. Outliers circled in red were excluded from the data set prior to DESeq2 modeling and differential expression analysis. *Trans*-regulatory mutants and the *cis*-regulatory overexpression strain and references are not shown for better visibility of the *cis*-regulatory mutants that were outliers.

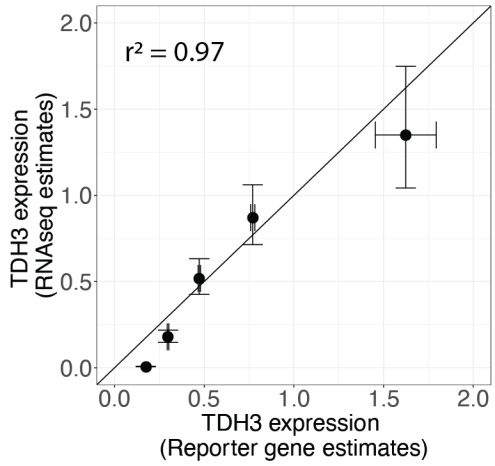

Figure 2-11: Measures of TDH3 expression in RNA-seq data were similar to measures using a fluorescent reporter gene for cis-regulatory mutants

RNA-seq estimates of *TDH3* expression correlated highly (Pearson's $r^2 = 0.97$) with estimates predicted by the same mutations in the *TDH3* promoter driving expression of a fluorescent reporter and measured using flow cytometry (see Methods). Error bars represent one standard error for the $\log_2$ fold change estimates from RNA-seq and flow cytometry.

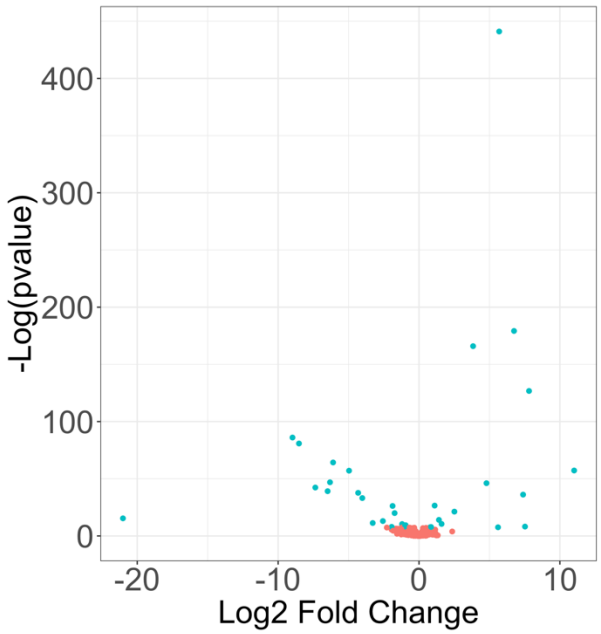

Figure 2-12: Genes differentially expressed between mating type alpha and mating type "a" reference strains

A volcano plot showing P-values and $\log_2$ fold change estimates between the mating type **a** reference strain and mating type α reference strain. Genes significantly differentially expressed at a 10% FDR are colored in blue, while non-significant genes are colored in orange. Significantly differentially expressed genes are listed in Table 3.

Table 1: Trans-regulatory mutant identities, growth rates, and effects on TDH3 expression

Each row contains a description of one *trans*-regulatory mutant used in this study. Columns include collection numbers for each strain, the gene in which the mutation is located, the exact position of the nucleotide change for each mutation, the mutation type, the effect on *TDH3* expression relative to the alpha wild type reference strain, and the growth rate relative to the same reference strain.

Collection	Gene	Position	Reference Nucleotide	Resulting Nucleotide	Mutation Type	*TDH3* Expression (Relative to Wild Type)	Growth Rate (Relative to Wild Type)
YPW2504	CYC8	1010	G	A	nonsynonymous	-0.084233803	NA: flocculant
YPW2506	WWM1	313	C	T	nonsense	-0.003769963	1.002461824
YPW2525	PRE7	83	G	A	nonsynonymous	-0.019909986	0.910057698
YPW2690	NAR1	536	G	A	nonsynonymous	0.388543289	0.969350403
YPW2728	TYE7	391	C	T	nonsense	-0.194962679	0.985029755
YPW2794	SDS23	676	G	A	nonsynonymous	-0.042533121	0.906479485
YPW2798	CCC2	2108	G	A	nonsynonymous	0.306681567	0.979599089
YPW2800	NAM7	1304	G	A	nonsynonymous	-0.140366862	1.006228554
YPW2911	ADE4	856	G	A	nonsynonymous	-0.02728248	0.577954779
YPW2919	ADE6	3327	G	A	nonsynonymous	-0.367210264	0.6888549
YPW3068	ADE5	1715	G	A	nonsynonymous	-0.477739095	0.610869021
YPW3088	RIM8	436	C	T	nonsense	0.049997138	0.886004651
YPW3092	SSN2	2911	C	T	nonsense	-0.06525204	NA: flocculant
YPW3161	FTR1	856	G	A	nonsynonymous	-0.354215414	1.008109509
YPW3191	ATP23	476	G	A	nonsense	-0.017727453	1.046129642
YPW3198	MOD5	765	G	A	nonsense	0.150842738	1.013285564
YPW3202	FRA1	2175	G	A	nonsense	0.123593606	1.033096543
YPW3203	TRA1	8560	G	A	nonsynonymous	0.369973614	0.932334074
YPW3228	CIA2	626	G	A	nonsynonymous	0.122862838	0.9739733
YPW3241	CAF40	674	G	A	nonsynonymous	-0.423146658	0.926848067
YPW3245	HXK2	1388	G	A	nonsynonymous	-0.123392389	0.88000086
YPW3247	BRE2	405	G	A	nonsense	-0.058319984	0.929150318
YPW3254	IRA2	7496	G	A	nonsense	0.151125465	1.070307468
YPW3256	MRN1	1664	G	A	nonsynonymous	-0.17371249	1.033420585
YPW3272	TUP1	2098	T	del	nonsynonymous	0.053700866	1.026132314
YPW3275	ADE2	3275	C	T	nonsynonymous	0.022035294	0.684375025
YPW3282	GCR1162	833, 1112, 1946, 2305, 2755	del, T, G, A, T	T, C, A, G, C	frameshift, nonsynonymous, nonsynonymous, synonymous, synonymous	-3.927013598	0.30252632
YPW3283	GCR1281	940, 1224, 2178, 2599	T, T, T, T	C, C, C, C	synonymous, nonsynonymous, nonsynonymous, synonymous	-0.513344386	0.957384576
YPW3284	GCR1037	737, 1183, 1224, 1258, 2038, 3079,	G, T, T, T, A, T	C, C, C, C, G, d	intron, silent, nonsynonymous, silent, silent, frameshift	-0.198930953	0.9883014
YPW3285	GCR1339	726, 737, 740, 840, 2574	T, G, C, A, A	G, C, G, G, G	intron, intron, intron, nonsynonymous, nonsynonymous	-2.191062688	0.784056797
YPW3286	GCR1241	1366	A	G	synonymous	0.07367076	0.977587627
YPW3287	RAP1357	2378, 1881, 284	A, G, T	G, A, C	silent, nonsynonymous, silent	0.210454342	1.009266995
YPW3288	RAP154	upstream922, 1100, 2042,2043	C, T, A, T	T, C, del, del,	promoter, nonsynonymous, frameshift nonsense	-1.265524056	0.35949695
YPW3289	RAP1238	95, 568, 693, 1365, 1338, 2121	T, A, A, G, A, A	C, G, G, A, G,	silent, all other nonsynonymous	-2.15553115	0.4112339
YPW3290	RAP1484	upstream85, 2128, 1060	A, A, A	G, G, G	noncoding, nonsynonymous, nonsynonymous	0.169982976	0.537104052

Table 2: Genes significantly differentially expressed in the TDH3 null mutant and their expression levels across cis-regulatory mutants

Rows are genes significantly differentially expressed in the *TDH3* null mutant. Columns include the gene's systematic name, the GO terms associated with that gene that were enriched in this gene set, and the expression level of the gene in each of the cis-regulatory mutants used in this study.

Gene	GO Term	0%TDH3	20% TDH3	50% TDH3	85% TDH3	135% TDH3
YAL005C		0.86066634	0.46867909	0.64915917	0.62505496	-0.2322649
YAL053W		0.70014478	0.63765153	0.7219827	0.87795873	-0.03986
YAR019C		0.73592328	0.6418794	0.38690003	0.42155652	-0.3494885
YBL004W		0.65422297	0.47528772	0.39420511	0.40090958	0.22571344
YBL023C		-0.5207478	-0.2853212	-0.3550784	-0.4728499	0.12688174
YBL100W-B	transposition, DNA-biosynthetic processes, integration	0.9549226	0.49362961	0.39937317	0.04841407	0.16901136
YBR007C		-0.5376707	-0.3188718	-0.3522927	-0.2521071	0.15077559
YBR054W		2.51147046	1.82159869	1.9868231	1.46512146	-1.8074294
YBR068C		1.22596614	1.04236262	0.94520549	1.0192456	-0.1848353
YBR093C		1.39689238	1.1236301	0.79792075	0.7527805	0.71096701
YBR191W-A		-1.6786788	-1.3289287	-1.164011	-1.2806877	0.29904333
YBR195C		-0.7358919	-0.181189	-0.067738	-0.1182993	0.10196586
YBR208C		-1.297964	-1.318449	-1.3772373	-1.4758394	0.13166646
YBR261C		-0.5293321	-0.2989313	-0.2524233	-0.2126064	0.21917058
YBR279W		-0.3803111	-0.296689	-0.2068958	-0.22538	0.04669047
YCL019W	transposition, DNA-biosynthetic processes, integration	1.17683614	0.80309918	0.50879323	0.40759858	-0.4221329
YCL056C		-0.5904766	-0.284652	-0.3991719	-0.2941421	-0.0290152
YCR012W	glycolytic processes	0.637829	0.46439832	0.47153713	0.38398315	-0.2846416
YCR013C		0.70993409	0.430882	0.53651839	0.39315451	-0.2193424
YCR021C		2.61590346	1.95957799	2.29181667	2.17487065	-1.6492935
YCR023C		0.56383548	0.27779257	0.33905855	0.27374732	-0.3238732
YCR051W		-0.6665915	-0.3894082	-0.4536655	-0.38827	0.34349723
YDL021W	glycolytic processes	0.85684961	0.09729501	-0.2005881	-0.2164709	-1.1500561
YDL048C		1.02970916	0.56110847	0.58283593	0.78878269	-0.4529256
YDL085C-A		-0.8264719	-0.7944528	-0.7671969	-0.7880634	0.10403448
YDL092W		-0.4939471	-0.4763423	-0.3802905	-0.3858608	0.53779105
YDL124W		0.94439258	0.61128381	0.58871702	0.46319703	-1.0344842
YDL130W		-0.6076937	-0.3845098	-0.275188	-0.3256289	0.2178517
YDL136W		-0.6407814	-0.4289455	-0.3304797	-0.3571794	0.34333329
YDL150W		-0.7587477	-0.2972928	-0.333881	-0.3790837	0.52289683
YDL154W		0.58582875	0.17550295	-0.0061721	-0.096088	0.05499779
YDL184C		-0.8967832	-0.679928	-0.5730351	-0.6134652	0.1793895
YDL206W		0.99685689	0.65005378	0.55600104	0.69961501	-0.7633393
YDL215C		-0.6777918	-0.785325	-0.7804942	-0.586021	0.45106268
YDR013W		-0.5462004	-0.4363923	-0.2837542	-0.3604916	0.13374363
YDR034C-D	transposition, DNA-biosynthetic processes, integration	1.23883682	1.08104378	0.56208386	0.42892389	-0.0327272
YDR046C		0.96786207	0.70310004	0.64349972	0.54353664	-0.4833774
YDR055W		0.92319318	0.54562398	0.81672432	0.86432579	-0.5518199
YDR147W		-0.5606684	-0.1685244	-0.2552964	-0.3842678	0.02213538
YDR222W		1.65311834	1.512342	1.55165996	1.34021699	-0.8308529
YDR247W		1.36913967	0.49823725	0.43531949	0.69531746	-0.6514854
YDR261C-D	transposition, DNA-biosynthetic processes, integration	1.42935682	1.51581413	1.27959632	1.65425259	-0.3066979
YDR261W-B	transposition, DNA-biosynthetic processes, integration	1.04882158	0.19274226	0.27898746	0.08342882	-0.2226545
YDR349C		0.51047609	0.36448651	0.3590008	0.36677682	0.02625158
YDR373W		-0.3598618	-0.2423316	-0.185691	-0.253708	0.0521549
YDR378C		-0.6252056	-0.4212149	-0.2939464	-0.4450057	0.12059366
YDR422C		0.37076387	0.29920349	0.09000073	0.1131553	-0.2405016
YDR472W		-0.4927066	-0.2973928	-0.2297901	-0.2777246	0.10823909
YDR497C		-0.4468258	-0.3818399	-0.2892402	-0.2960419	-0.5258938

YER009W		-0.3056796	-0.2740487	-0.1425791	-0.2320449	0.16835746
YER029C		-0.8404108	-0.5391244	-0.5674455	-0.6218083	0.22258765
YER037W		1.88899967	1.75286412	1.64919922	1.40175236	-1.1554391
YER053C-A		2.42470394	1.76621497	1.82214719	1.35693898	-1.0219239
YER062C		0.81458787	0.50062181	0.57965887	0.322569	-0.4834964
YER073W		1.12501631	0.85778382	0.62433256	0.95645546	0.41824078
YER092W		-0.5627956	-0.2917561	-0.2851196	-0.3021331	0.17476597
YER138C	transposition, DNA-biosynthetic processes, integration	0.92119869	0.69215076	0.36275624	0.4396383	0.06799732
YER150W		1.82971459	0.88344585	0.68761325	1.07093467	-1.3613543
YER158W-A		-0.7239621	-0.4593073	-0.6104178	-0.5445631	0.0062481
YER175C		1.00092359	0.33312463	0.48747156	0.49257478	-0.5885706
YER177W		0.41169458	0.31483614	0.34190112	0.32740319	-0.3108167
YFL002W-A	transposition, DNA-biosynthetic processes, integration	1.50686586	1.13182247	0.77658193	0.71411418	0.0253865
YGL059W		1.23762084	0.40180693	0.49103164	0.55404651	-0.6432454
YGL077C		-0.5965786	-0.3843474	-0.4729184	-0.5525173	0.03762174
YGR023W		1.52120796	0.89854953	0.83599049	0.93635066	-0.7178598
YGR027W-B	transposition, DNA-biosynthetic processes, integration	0.91484337	0.64116504	0.57765049	0.58437387	-0.0980371
YGR052W		1.09715656	0.55524713	0.69507861	0.68626911	-0.0635117
YGR097W		1.22883749	0.9547725	0.72897341	0.73625084	-0.2201387
YGR138C		1.52073516	1.30240259	1.58652534	1.45280752	-0.9832644
YGR192C	glycolytic processes	-7.2184431	-2.4832589	-0.9476816	-0.199752	0.43256327
YGR249W		2.05940792	1.42762796	1.22733858	1.52291153	0.27740313
YGR254W	glycolytic processes	0.95344727	0.41583639	0.47999643	0.43161383	-0.889351
YGR275W		-0.5027187	-0.1988632	-0.1557657	-0.2574876	0.13003787
YHR030C		1.14075097	1.02651822	1.20143	1.37454988	0.05881109
YHR094C		1.26872732	1.1976312	0.79388569	0.68940646	0.92859022
YHR099W		0.55753817	0.32617899	0.37181349	0.37287062	-0.2453025
YHR211W		0.88244982	0.67245168	0.14566263	0.15755503	-0.1499835
YHR214C-B	transposition, DNA-biosynthetic processes, integration	0.98152823	0.72758317	0.52029813	0.40384828	-0.1579098
YIL002W-A		-1.0539483	-0.883223	-0.7889473	-0.8972083	0.38549425
YIL053W		0.75948429	0.64882808	0.60393888	0.40320191	-0.1737059
YIL117C		1.01935741	0.92310396	0.98618865	1.04857293	0.3363456
YJL016W		0.93021892	0.86316934	0.72914943	0.69118167	-0.7475784
YJL052W	glycolytic processes	1.14736473	0.30939502	0.27993022	0.30027725	-1.5507751
YJL106W		1.17940156	0.94121445	0.74538347	0.81936365	-0.4667983
YJL107C		0.96144709	0.89317709	0.67133033	0.94970027	0.01240414
YJL159W		0.77844051	0.67937056	0.71780592	0.80249474	-0.2605412
YJL165C		0.64510747	0.67968242	0.68136453	0.90828424	-0.5183733
YJR027W	transposition, DNA-biosynthetic processes, integration	0.88165004	0.51054626	0.20409902	0.53267417	0.24208928
YJR029W	transposition, DNA-biosynthetic processes, integration	0.7759587	0.7366108	0.44515869	0.30437809	0.03811381
YJR057W		-0.6609933	-0.3383362	-0.3979509	-0.473859	0.1492504
YJR148W		-0.8245001	-0.7070821	-0.5257967	-0.4290103	0.46722147
YKL018C-A		-0.7521448	-0.4530152	-0.2949304	-0.372647	-0.2863031
YKL028W		-0.4465961	-0.2467237	-0.3503454	-0.3574153	0.19202553
YKL042W		-0.702744	-0.2596756	-0.3366526	-0.3567737	-0.1349093
YKL113C		-0.7519998	-0.4927659	-0.4944739	-0.5901704	0.13347236
YKR039W		-1.1951768	-1.2258048	-1.1849119	-1.1225647	-0.4061292
YLL026W		1.23260815	0.84063786	0.98383007	1.02798103	-0.7971515
YLL039C		0.67645737	0.38921617	0.37381477	0.46560265	-0.4415091
YLR058C		-0.5875071	-0.324415	-0.2225145	-0.2071374	-0.148967
YLR154W-A		1.17950125	1.24786287	0.55248206	1.01371498	-0.0471554
YLR154W-B		1.09850774	1.05661347	0.51228049	0.84567192	0.06864263
YLR168C		-1.0089094	-0.7342781	-0.655504	-0.534473	-0.0983332
YLR183C		-0.4309238	-0.2400174	-0.1106977	-0.1942807	-0.1559593
YLR198C		-0.7990229	-0.5872125	-0.5415412	-0.4937648	0.52437752

YLR227W-B	transposition, DNA-biosynthetic processes, integration	0.75309476	0.87452381	0.65078568	0.53751864	0.31281102
YLR256W-A	transposition	0.60528041	0.17546561	0.03192263	-0.0870855	-0.0987043
YLR342W		0.60149886	0.47332817	0.47687147	0.46304462	-0.2281693
YLR410W-B	transposition, DNA-biosynthetic processes, integration	0.91688031	0.56472791	0.64438237	0.46447923	-0.1127141
YLR431C		-0.5324455	-0.2461052	-0.1156486	-0.1689699	-0.0371495
YMR011W		-0.6844054	-0.6241923	-0.718308	-0.2455436	0.84385126
YMR173W		0.83141522	0.34797302	0.36667095	0.4094405	-0.4834893
YMR194W		-0.4329798	-0.2460524	-0.1737232	-0.2505416	0.16866355
YMR205C	glycolytic processes	0.42895711	0.24394723	0.25074132	0.26996543	-0.14892
YMR291W		0.84254655	0.36268789	-0.0587458	0.28300356	-0.4999755
YNL149C		-0.4581966	-0.3050828	-0.2548503	-0.2648983	0.56961344
YNL153C		-0.6368529	-0.3083021	-0.2931977	-0.2863009	0.33164859
YNL160W		1.77181042	0.91723952	0.76874567	0.45225038	-1.5539747
YNL216W		-0.4531155	-0.299457	-0.4229376	-0.3339431	-0.0598565
YNL217W		0.53357911	0.48605335	0.55183819	0.62889629	-0.0076583
YNR001C		0.7185517	0.09752488	0.1339201	0.22064893	-1.0869248
YNR033W		0.40597877	0.20583885	0.17058988	0.35067118	-0.1254372
YOL059W		1.28022919	0.84175963	0.51287373	0.330073	-0.6505222
YOL103W-B	transposition, DNA-biosynthetic processes, integration	0.76343481	0.42455728	0.29060327	0.25347599	0.10228976
YOL130W		0.38922891	0.47749275	0.38362897	0.51239699	0.16199552
YOR120W		1.68549372	0.6645684	0.64348851	1.01699522	-1.1220238
YOR123C		-0.5078559	-0.3780871	-0.2891009	-0.2978429	0.10196898
YOR142W-B	transposition, DNA-biosynthetic processes, integration	1.08032994	0.75944245	0.48121039	0.45816432	0.29347694
YOR185C		0.42146273	0.25032531	0.2119808	0.24557619	-0.8851427
YOR192C-B	transposition, DNA-biosynthetic processes, integration	1.19389331	0.78545801	0.77778425	0.59503152	-0.1155479
YOR194C		-0.4978772	-0.2550435	-0.2924639	-0.2107798	0.15553717
YOR208W		1.14076079	0.95517447	1.01530067	1.24680247	-0.2202613
YOR267C		0.78965039	0.55563098	0.57089915	0.54479677	-0.4423027
YOR289W		0.95355565	0.35701709	0.10436301	0.50966473	-1.0049502
YOR303W		1.24971178	0.79492582	0.71341413	0.84550794	0.11371954
YOR304C-A		-0.9003944	-0.4348218	-0.2437335	-0.402951	0.02448415
YOR347C	glycolytic processes	0.71123851	0.12474379	0.17189174	0.33221604	-0.7291648
YOR375C		-0.6612782	-0.851605	-0.6987726	-0.6761926	0.17610604
YPL006W		0.59448022	0.32238009	0.26435939	0.34490556	-0.3880468
YPL014W		1.39840873	1.01253195	1.19550022	1.17775826	-1.1492074
YPL075W		0.8148559	0.7046063	0.60311361	0.5206207	-0.3146591
YPL089C		0.73179837	0.73762565	0.65555204	0.89511814	-0.0109038
YPL109C		0.77860967	0.45735114	0.34032088	0.39319688	-0.396289
YPL110C		0.88650857	0.634204	0.63752351	0.61633787	0.20459414
YPL213W		-0.6926151	-0.4330097	-0.3941716	-0.6459169	0.15279932
YPL250C		1.59078856	0.95905662	0.98663081	1.10323877	0.48883344
YPR018W		-0.8210867	-0.4333412	-0.4491297	-0.5939671	0.35508964
YPR024W		0.36282346	0.28707249	0.38735899	0.29962013	-0.0776075
YPR036W-A		1.2857187	0.88458966	0.92582115	0.94722496	-0.4693108
YPR133C		-0.3597467	-0.2303225	-0.2250339	-0.2133372	0.33245168
YPR156C		0.78584778	0.66439269	0.71891206	0.6446663	-0.6981469
YPR157W		1.49814804	1.37416846	1.51727234	1.17039695	-0.6318488
YPR158C-D	transposition, DNA-biosynthetic processes, integration	0.80343069	0.27984836	0.282714	0.43478332	-0.1361258
YPR158W-B	transposition, DNA-biosynthetic processes, integration	0.98918972	0.61063099	0.26412448	0.38671415	0.42410433
YPR188C		-0.498461	-0.3884995	-0.2239871	-0.2965941	0.24175091

Table 3: Genes significantly differentially expressed between mating type alpha and mating type 'a' reference strains

Rows are genes significantly differentially expressed between mating type **a** and mating type α reference strains. Columns include gene systematic name, adjusted P-value for differential expression obtained from DESeq2, and functional category or common name of each gene. Un orf = uncharacterized open reading frame, mating = gene associated with mating type differences, dub orf = dubious open reading frame, as annotated in the Saccharomyces Genome Database.

Gene	Adjusted p-value	Functional category or Common name
YBR056W-A	0.000595218	un orf
YCL066W	2.14E-16	mating
YCL067C	1.41E-09	mating
YCR040W	6.41E-07	mating
YCR097W	0.05681645	mating
YCR097W-A	0.00314114	dub orf
YDR461W	8.42E-14	mating
YER160C	0.096872963	Ty gene
YFL026W	4.57E-75	mating
YFL027C	1.09E-09	Gyp8
YGL032C	1.38E-52	mating
YGL089C	2.58E-18	mating
YGL263W	0.000222955	Cos12
YGR240C-A	0.082307423	un orf
YHR054C	0.065441286	un orf
YIL015W	1.82E-188	mating
YIL117C	0.018482987	mating
YJL170C	5.16E-18	mating
YJR004C	1.15E-22	mating
YKL177W	4.84E-15	dub orf
YKL178C	1.11E-25	mating
YKL209C	1.80E-69	mating
YKR035W-A	6.08E-05	Did2
YKR091W	0.056163877	Srl3
YLR040C	1.93E-14	mating
YLR041W	1.60E-12	dub orf
YLR121C	0.06307448	Yps3
YLR154C-H	0.006771427	un orf
YNL145W	0.088266952	mating
YNL146C-A	1.90E-07	un orf
YOL016C	0.007109895	Cmk2
YOR208W	0.018333026	mating
YOR225W	0.0724697	dub orf
YPL088W	0.05681645	un orf
YPL187W	8.15E-33	mating

Chapter 3 Network Topology Generates Differences in the Pleiotropy of *Cis*- and *Trans*-Acting Mutations in *Saccharomyces cerevisiae*

Abstract

A mutation's degree of pleiotropy, or the number of traits it impacts, is predicted to increase the probability of the mutation being detrimental to fitness. For mutations that impact gene expression, pleiotropy is suggested to be an important factor explaining why some types of regulatory mutations are more likely to fix than others. Specifically, mutations that affect a gene's expression in cis are hypothesized to generally be less pleiotropic and thus more likely to fix than mutations that affect expression of the same gene in trans. Here, we use gene expression data from Saccharomyces cerevisiae gene deletion strains to test this hypothesis, estimating the pleiotropy of *cis*- and *trans*-acting mutations with pleiotropy measured as the number of genes that change expression in response to each *cis*- or *trans*-acting mutation. These data showed that *trans*-acting mutations did indeed tend to have higher pleiotropy than *cis*-acting mutations affecting expression of the same gene. We found that this pattern held for the vast majority of genes in the dataset and could be explained by the topology of the regulatory network controlling gene expression. Coupling this analysis with measures of fitness for the same gene deletions showed that *trans*-acting deletions tended to be more detrimental to fitness than the corresponding *cis*-acting deletion, supporting the hypothesis that differences in pleiotropy contribute to the apparent preferential fixation of *cis*-regulatory alleles over evolutionary time.

Introduction

A mutation is said to be pleiotropic if it affects more than one trait. For example, in a multicellular organism, a mutation in the coding sequence of a gene required for development of multiple tissues is considered pleiotropic if it impacts formation of all these tissues (Paaby and Rockman 2013). Mutations affecting activity of this gene in only one or a subset of these tissues, however, including mutations in modular, tissue-specific, enhancers controlling expression of the gene, are less pleiotropic. Because evolutionary theory suggests that mutations that are more pleiotropic are more likely to be deleterious (Kimura and Ohta 1974), mutations in noncoding sequences that affect a subset of a gene's functions have been suggested to be more likely to contribute to evolutionary change than coding mutations that impact all functions of a gene (Carroll 2005; Wray 2007; Carroll 2008; Stern and Orgogozo 2009). Indeed, many examples of such noncoding changes between species have now been identified (Stern and Orgogozo 2008; Martin and Orgogozo 2013, Courtier-Orgogozo et al. 2020).

Mutations in non-coding sequences are often *cis*-acting (and known as *cis*-regulatory) because they tend to affect expression of a coding sequence allele on the same chromosome, whereas mutations in coding sequences are often *trans*-acting (and known as *trans*-regulatory) because they tend to impact expression of genes by altering a diffusible RNA or protein that can impact expression of genes on any chromosome (Signor and Nuzhdin 2018). But the terms *cis*- and *trans*-acting are not strictly synonymous with non-coding and coding mutations. Rather, *cis*- and *trans*-acting describe the way by which a mutation affects expression of a particular gene, with *cis*-acting variation having allele-specific effects on expression and *trans*-acting variation affecting expression of both alleles of a gene in diploid cells.

Comparing gene expression between strains or species to allele-specific expression in F_1 hybrids produced by crossing these strains or species allows the relative contributions of *cis*- and *trans*-acting variation to be estimated without identifying specific genetic variants (Wittkopp et al. 2004). Using this approach, which is agnostic to the coding or noncoding status of the causative genetic variants as well as its tissue-specificity, to analyze regulatory evolution in fruit flies (*Drosophila*) showed a greater contribution of *cis*- than *trans*-acting variation to expression differences between than within species (Wittkopp et al. 2008; McManus et al. 2010; Coolon et al. 2014), suggesting that *cis*-regulatory divergence accumulates preferentially over evolutionary time. An even stronger pattern of increasing relative *cis*-regulatory contributions to expression divergence with increasing divergence time was observed when comparing strains and species of *Saccharomyces* yeast (Metzger et al. 2017), suggesting it is a common feature of regulatory evolution. Moreover, observing this evolutionary advantage of *cis*-regulatory variation in unicellular yeasts indicates that it cannot solely be explained by lower pleiotropy of tissue-specific noncoding mutations acting in *cis* relative to more pleiotropic coding mutations acting in *trans*.

Indeed, both *cis*- and *trans*-acting mutations can be found in either coding or noncoding sequences (Jakobson and Jarosz 2019; Lutz et al. 2019; Hill et al. 2021). In fact, because the categorization of mutations as *cis*- or *trans*-acting is relative to a focal gene, a single mutation can be classified as either *cis*- or *trans*-acting depending on which gene is considered the focal gene. For example, a noncoding difference between two species of yeast in the promoter of the *OLE1* gene has been shown to affect expression of *OLE1* in *cis* as well as expression of many other genes in *trans* (Lutz et al. 2019). But if the same mutation can be considered *cis*- or *trans*-

acting depending on the point of reference, how can we explain the observed preferential accumulation of *cis*-regulatory divergence?

Considering the effects of regulatory mutations in the context of gene regulatory networks offers an explanation for this pattern; one that is still based on differences in the relative pleiotropy of *cis*- and *trans*-acting mutations (Wittkopp 2005; Wittkopp 2007; Kopp and McIntyre 2012; Yang and Wittkopp 2017). If the pleiotropy of a mutation is measured as the number of genes whose expression it affects, *cis*-acting mutations are expected to be less pleiotropic than *trans*-acting mutations because although a *cis*-acting mutation affecting expression of a particular gene might also have downstream effects on expression of other genes (making it pleiotropic to a degree), mutations that affect expression of the same gene in *trans* should also have these same effects plus effects attributable to their impact on expression of other genes in parallel (making them more pleiotropic than *cis*-acting mutations).

This idea has recently been tested in the baker's yeast *Saccharomyces cerevisiae* by directly comparing the effects of *cis*- and *trans*-regulatory mutations affecting expression of the *TDH3* gene (Vande Zande and Wittkopp, in review). As predicted, the *trans*-regulatory mutations tended to cause changes in the expression of more genes than the *cis*-regulatory mutations, but only when *cis*- and *trans*-regulatory mutations had similar effects on expression of *TDH3*. The impacts of *cis*- and *trans*-regulatory mutations on expression of genes downstream of *TDH3* differed significantly, suggesting that feedforward and/or feedback loops involving genes downstream of *TDH3* complicate the simple model described above. Because this study examined *cis*- and *trans*-regulatory mutations defined by a single focal gene, and because all the *cis*-acting mutations examined were noncoding and all the *trans*-acting mutations coding (Vande

Zande and Wittkopp, in review), it remains to be seen whether these same relationships hold for other genes and for other mutations with *cis*- and *trans*-acting effects.

Here, we compare the pleiotropic effects of *cis*- and *trans*-acting gene deletion mutations for 1484 genes in *S. cerevisiae*. For each gene, deletion of the gene itself is a *cis*-acting mutation whereas deletions of other genes that impact its expression are *trans*-acting. With this approach, we compare the effects of the same mutations acting in both *cis* and *trans*, eliminating the often-confounded correlation of *cis*- and *trans*-acting variants with noncoding and coding mutations. Specifically, we compared the extent of pleiotropy (measured as the number of genes significantly differentially expressed in response to the mutation) for mutations that affect expression of a focal gene in *cis* or in *trans,* considering each deleted gene sequentially as a focal gene. At both the level of individual genes and in aggregate we find that *cis*-acting mutations tend to be less pleiotropic than *trans*-acting mutations. Analysis of a perturbation network describing how deletion of one gene impacts expression of other genes showed that the topology of this network is sufficient to explain this tendency of *cis*-regulatory mutations to be less pleiotropic than *trans*-regulatory mutations. Consistent with prior work, we observed a negative relationship between fitness and pleiotropy defined by impacts on gene expression, indicating that *trans*-acting deletions tend to be more detrimental to fitness than *cis*-acting deletions. These findings indicate that the larger pleiotropic impacts of *trans*-regulatory mutations as compared to *cis*-regulatory mutations does not require differences between coding and noncoding mutations.

Results

trans-acting mutations are more pleiotropic than cis-acting mutations

To quantify the pleiotropic effects on gene expression of *cis*- and *trans*-acting mutations, we utilized a large compendium of gene expression profiles, consisting of microarray data

measuring expression levels of 6,123 genes in 1,484 single gene deletion mutations in the baker's yeast *S. cerevisiae* (Kemmeren et al. 2014). Using the significance and fold change thresholds used in the original publication of these data (p-value < 0.05, fold change > 1.7, see Methods), we excluded gene deletion strains that did not show a significant decrease in expression of the deleted gene, resulting in measurements of 6,123 genes in a total of 1275 deletion strains. These data can be visualized as a perturbation network in which each gene in the dataset is represented by a node, and edges are directed from gene 1 to gene 2 when the expression of gene 2 is significantly changed upon deletion of gene 1 (Fig. 3-1A).

We considered each deleted gene, or node, in turn to be the focal gene. The deletion of the focal gene itself acts in *cis* to the focal gene because it causes an allele specific absence of expression for that gene. A complete deletion of the focal gene is expected to cause a more extreme change in expression than other mutations in *cis*-regulatory sequences that modify the focal gene's expression level, but using the deletion as a *cis*-acting mutation allowed us to compare *cis*- and *trans*-acting mutations that are identical in mutation type and use the same exact mutations as *cis*- or *trans*-acting when considering different focal genes. Furthermore, because deletions tend to cause more severe phenotypes than other reductions in expression (Keren et al. 2016), *cis*-acting deletions are expected to be more pleiotropic than other types of regulatory mutations and using deletions as *cis*-acting mutations is thus a conservative way to test the hypothesis that *cis*-regulatory mutations tend to be less pleiotropic than *trans*-acting mutations.

We then identified mutations that act in *trans* on expression of the same focal gene as deletion mutants in which the focal gene is significantly differentially expressed, or as nodes with an outgoing edge to the focal gene (Fig. 3-1B). The number of *trans*-regulators per focal

gene ranged from 0 to 228, with a large number of focal genes being influenced by a few *trans*-regulators, and a few being influenced by a large number (Fig. 3-1C). Because they are identified exclusively from expression data, these *trans*-regulators can influence expression of the focal gene either directly or indirectly. Therefore, the *trans*-regulators include not only transcription factors directly controlling transcription of the focal gene, but also genes that influence cellular systems such as metabolism or progression through the cell cycle that indirectly change expression the focal gene. As expected, effects of *trans*-acting deletions on expression of the focal gene tended to be smaller than the impact of the *cis*-acting deletion of the focal gene itself (Fig. 3-4).

We next quantified the pleiotropy of *cis*- and *trans*-acting deletions as the number of genes significantly differentially expressed in that deletion strain. Specifically, when the focal gene itself was deleted, the number of genes that were significantly differentially expressed as a result, except for the focal gene itself, was taken as the pleiotropy of the *cis*-acting mutation for that focal gene, visualized as the number of out-going edges from that node (Fig. 3-1D). For example, when the membrane peptide transporter gene *PTR2* was deleted, 1 other gene was significantly differentially expressed, so the pleiotropy of the *cis*-acting mutation to *PTR2* was 1. The distribution of *cis*-acting pleiotropy for all focal genes ranged from 1 to 1014, with once again many focal genes having a very small pleiotropic effect and a small proportion of genes affecting expression of a large number of other genes (Fig. 3-1E).

We then similarly estimated the pleiotropy of *trans*-acting mutations as the number of other genes differentially expressed in each *trans*-acting deletion for each focal gene (Fig. 3-1F). For example, there were 116 gene deletion mutants in which *PTR2* is significantly differentially expressed, so there are 116 *trans*-acting mutations to *PTR2*. The total number of genes that are

significantly differentially expressed in each of those 116 *trans*-acting mutations, minus the focal gene and *trans*-regulator itself, or the number of outgoing edges from each of those nodes, together make up the distribution of the pleiotropic effects of all *trans*-acting mutations to the *PTR2* (Fig. 3-1G, purple histogram).

To compare the effects of *cis*- and *trans*-acting deletions to each focal gene, we calculated the median pleiotropy of all *trans*-acting mutations for each focal gene (for *PTR2*, Fig. 3-1G, black line) and plotted these against the *cis*-acting pleiotropy for the same focal gene (Fig. 3-1H). Points falling above the x = y line are focal genes in which the median pleiotropy of all *trans*-acting mutations is larger than the pleiotropy of the *cis*-acting deletion. We next calculated the difference between the median pleiotropy for all *trans*-acting mutations to one focal gene and the pleiotropy of the *cis*-acting mutation for that focal gene. We examined the distribution of all of the differences between *cis*- and *trans*-acting mutations (Fig. 3-1I) and found that the median of this distribution was significantly greater than zero (Student's t-test p-value = 3×10^{-165}), indicating greater pleiotropy of *trans*- than *cis*-acting mutations. In fact, only ~3% of all focal genes with at least one *trans*-regulator (26/804) had a *cis*-acting mutation with higher pleiotropy than the median *trans*-acting mutation. When considering each *trans*-acting mutation individually rather than summarizing the *trans*-acting mutations in a median value (Fig. 3-5), only 4% of all *trans*-acting and *cis*-acting mutation pairs (365/8392) had more pleiotropic *cis*-acting deletions. These results were robust to changes in both the significance and fold-change cutoffs used to identify genes as significantly differentially expressed (Fig. 3-6). Relative to specific focal genes, *cis*-acting mutations were thus indeed less pleiotropic than *trans*-acting mutations, even though the *cis*-acting mutations had greater impacts on expression of the focal gene.

Network topology can explain differences in pleiotropy between cis- and trans-regulatory mutations

We hypothesized that the strong trend of *trans*-acting mutations being so frequently more pleiotropic than *cis*-acting mutations affecting expression of the same focal gene might be related to the overall topology of the perturbation network. As described above, the pleiotropic effects of both *cis*- and *trans*-acting mutations are drawn from the distribution of out-going edges for the entire perturbation network. This distribution of all out-going edges approximates a power-law distribution, in agreement with previous studies examining perturbation networks for yeast gene expression (Featherstone and Broadie 2002; Kemmeren et al. 2014) (Fig. 3-2A). This type of degree distribution is indicative of a scale-free network, in which a few nodes are highly connected while most have many fewer connections, and is a common network topology found in diverse contexts such as social relationships, the internet, and protein-protein interactions (Barabási and Oltvai 2004). Because each deletion is the *cis*-acting mutation at that focal gene, the overall distribution of out-going edges from each node is identical to the distribution of pleiotropy for all *cis*-acting mutations. In contrast, the total distribution of pleiotropy for *trans*-acting mutations for all focal genes is drawn from the same distribution as the *cis*-acting mutations, but each gene can be drawn as a *trans*-acting mutation multiple times, depending on the number of focal genes for which that deletion operates as a *trans*-acting mutation. This 'repeated sampling' of highly connected *trans*-acting deletions results in a distribution of pleiotropic effects of *trans*-acting mutations with a much higher median or average than than that for all *cis*-acting mutations (Fig. 3-2B).

This overall network topology explains how differences in the pleiotropic effects of *cis*- and *trans*-acting mutations might simply be accounted for by a statistical feature of the

perturbation network topology. Essentially, the structure of the perturbation network results in a type of sampling bias for highly pleiotropic *trans*-acting mutations. Any gene is likely to be the target of a highly connected *trans*-acting mutation because of the very fact that the *trans*-acting mutation is highly connected. Therefore, many focal genes will be influenced by highly pleiotropic *trans*-acting deletions that are more pleiotropic than themselves. In contrast, less pleiotropic *trans*-acting mutations influence the expression of few focal genes, and therefore do not frequently serve as potential *trans*-acting mutations. In this way, the perturbation network structure inherently generates a pattern in which each gene is likely to be influenced by genes that are more pleiotropic than itself. This phenomenon has been described in other contexts as an emergent feature of scale free networks (Feld 1991).

To test whether the network topology rather than the specific connections between nodes are responsible for the difference in the extent of effects between *cis*- and *trans*-acting mutations, we permuted the edges in the network in two different ways. In the first, we rearranged all network edges but maintained the overall degree distribution, so that the number of targets a gene was connected to did not change but the specific genes it was connected to did. This permutation did not affect the pattern of *trans*-acting mutations being more more pleiotropic than *cis*-acting mutations affecting expression of the same focal gene (Fig. 3-2C, 100 permutations, p-value = 0). In the second type of permutation, we randomly permuted all edges without maintaining the degree distribution, destroying the scale-free network topology by changing both the number of connections for each gene as well as which genes they were connected to. In this case, the difference between *cis*- and *trans*-acting mutations was removed (Fig. 3-2D, 100 permutations, p-value = 1), indicating that *trans*-regulatory mutations no longer tended to be more pleiotropic than *cis*-acting mutations affecting expression of the same focal gene.

Consequently, we conclude that the degree distribution of the *S. cerevisiae* perturbation network can explain the pattern of less pleiotropic effects of *cis*-acting mutations than *trans*-acting mutations that we observed, and that this trend does not require connections between specific regulators and target genes.

trans-acting deletions are more detrimental to fitness than cis-acting deletions affecting expression of the same focal gene

The differences in the pleiotropic effects of *cis*- and *trans*-acting deletions described above may result in a difference in the average fitness effects of these different types of mutations, potentially contributing to a preferential fixation of *cis*-regulatory alleles over time. Consistent with this hypothesis, pleiotropy, defined here as the number of genes differentially expressed as a result of a particular mutation, has been shown to be negatively correlated with fitness in *S. cerevisiae* (Featherstone and Broadie 2002, Vande Zande and Wittkopp, in review). This negative correlation suggests that the higher pleiotropy of *trans*-acting than *cis*-acting deletions may also produce a pattern of lower fitness of *trans*-acting deletions as compared to *cis*-acting deletions to the same focal gene. To test whether this was also true for the set of mutations examined in this study, we used measures of fitness for the *S. cerevisiae* deletion strains from another study (Maclean et al. 2017) and found that pleiotropy was indeed negatively correlated with fitness for all gene deletions present in both datasets ($n = 1105$, $R^2 = 0.25$, p-value $< 2.2 \times 10^{-16}$). We next examined the fitness cost of all *trans*-acting deletions relative to the *cis*-acting deletion for each focal gene present in both datasets (706/804 focal genes with at least one *trans*-regulator). As predicted, we found that the majority of pairs of *trans*-acting and *cis*-acting deletions fell above the x=y line (7378/11579, ~63%, Fig. 3-3B), indicating that the *trans*-acting deletion had a greater fitness cost than the *cis*-acting deletion. This pattern was stronger when the

effects of all *trans*-acting deletions affecting expression of a focal gene were represented by the median *trans*-acting fitness measure for each focal gene (503/599, ~83%, Fig. 3-3C). Taken together, these data show that for the majority of genes, *trans*-acting deletions tend to be more deleterious than *cis*-acting deletions, supporting the hypothesis that mutations affecting expression of a focal gene in *cis* should tend to be preferentially fixed over evolutionary time.

Discussion

This study demonstrates that network structure can result in *trans*-acting mutations tending to have more pleiotropic effects on gene expression than *cis*-acting mutations even when the same set of mutations is considered as *cis*- or *trans*-acting. Furthermore, the data analyzed are consistent with mutations that act in *trans* tending to be more detrimental to fitness than mutations that act in *cis*. As a statistical result of network structure, this pattern is essentially the result of a sampling bias for *trans*-acting mutations that are highly pleiotropic, as they are likely to be 'sampled' from all mutations influencing the expression of any particular focal gene. Consequently, this work suggests that the tendency of *trans*-regulatory mutations to be more pleiotropic and more deleterious than *cis*-regulatory mutations is not dependent on the type of sequence mutated (e.g., coding or non-coding) or the function of the focal gene. These findings have several important implications for interpreting the preferential fixation of *cis*-acting mutations relative to *trans*-acting mutations over evolutionary time suggested by prior work (Metzger et al. 2017).

First, the limited pleiotropy of *cis*-acting mutations is not dependent on them occurring in tissue specific regulatory elements, but occurs even when *cis*-acting mutations are deletions of the coding sequence of the focal gene. This means that the pattern of less pleiotropic *cis*-acting mutations as compared to *trans*-acting mutations is not limited to organisms with complex

regulatory elements and may help explain why we see trends of preferential fixation of *cis*-regulatory mutations in unicellular systems such as yeast that have relatively low regulatory complexity. This does not mean that mutations in tissue specific regulatory elements are not less likely to be deleterious than coding mutations. It does suggest, however, that because differences in tissue specificity between *cis*- and *trans*-acting mutations are not necessary to produce differences in their relative pleiotropy, expression divergence occurring most frequently via *cis*-regulatory mutations may not be limited to organisms or loci that are regulated in a tissue specific matter, such as those involved in morphological patterning (Carroll 2005).

Second, the patterns we observe are not specific to a functional category of focal gene or *trans*-regulatory genes, but instead are nearly universal. This may be surprising giving that some categories of genes have been shown to have different levels of connectivity in genetic networks (Luscombe et al. 2004). However, because of the relative nature of the differences in pleiotropy between *cis*- and *trans*-acting mutations for each focal gene, all genes except those few most highly connected 'hub genes' will have less pleiotropic *cis*-acting mutations than *trans*-acting mutations regardless of their connectivity. Furthermore, which transcription factors serve as 'hubs' is not permanent and can change in response to environmental perturbations (Luscombe et al. 2004). Therefore, it is likely that all genes show a pattern of less pleiotropic *cis*-acting mutations than *trans*-acting mutations in at least some environmental contexts. The independence of the differences in *cis*- and *trans*-pleiotropy from the connectivity of the focal gene itself may help explain why past studies have not found a significant relationship between the number of targets of a transcription factor and its *cis*-regulatory expression divergence (Kopp and McIntyre 2012; Yang and Wittkopp 2017). Of course, the most highly connected genes need not necessarily be transcription factors at all, but more broadly can be any gene whose deletion

will result in sweeping changes in genome-wide gene expression by disrupting important cellular functions (Rockman and Kruglyak 2006).

While our analyses are limited to gene expression patterns in *Saccharomyces cerevisiae* in one environmental condition, we expect that the pattern of lower relative pleiotropy and higher fitness for *cis*-acting mutations as compared to *trans*-acting mutations will hold true for any organism or environment in which the perturbation network of gene expression is approximately scale free. However, while it is likely that the overall topologies of various perturbation networks are similar, it remains to be seen how perturbation network structure relates to transcription factors binding regulatory elements of target genes (Flint and Ideker 2019). Integrating the various determinants of a mutation's impact on global gene expression and fitness is an important next step in understanding how gene expression evolves and in moving toward predictive models of gene expression evolution.

Materials and Methods

Expression data and inference of the perturbation network

The file containing microarray expression data used in this study to build a perturbation network for gene deletions in *S. cerevisiae* from Kemmeren *et al*. (2014) named "deleteome_all_mutants_ex_wt_var_controls.txt", was downloaded from http://deleteome.holstegelab.nl/ on 06/19/2020. This file included all M values and p-values for expression changes of each gene on the microarray (n = 6,123) for each gene deletion strain relative to the wild type control, where the M value is the $\log_2$ fold-change in expression and the p-value is obtained after Benjamini-Hochberg FDR correction for a statistically significant change in expression relative to a wild-type strain as calculated using the *limma* R package. *Limma* uses linear expression models and an empirical Bayes model to moderate standard errors

and calculate a moderated t-statistic and log-odds of differential expression (Smyth et al, 2005). The file also included experiments for strains grown in different media types, which were removed from the dataset for the analyses conducted in this paper. The authors of this study were aware of the aneuploid strains that have been identified in yeast gene deletion mutants and analyzed all expression profiles for evidence of aneuploidy. Any strains showing evidence of aneuploidy were remade and re-assayed or excluded from the dataset. The file was read into R (version 3.5.2) where all statistical analysis was performed. A gene was considered significantly differentially expressed in a deletion mutant if it showed a 1.7 or greater $\log_2$-fold change in expression (M value) and a *P*-value less than or equal to 0.05 after Benjamini-Hochberg FDR adjustment, resulting in a directed edge drawn from the deleted gene to the differentially expressed gene. In this way, we generated a binary, asymmetrical, adjacency matrix in which rows represent each deleted gene and columns represent the expression levels of all genes in the microarray (Supplementary Fig. 2). We then excluded 127 genes for which their deletion did not cause a statistically significant decrease in their expression. This matrix was the basis for all analyses on gene expression described in this paper. To test the robustness of patterns reported in the main text, we repeated analyses with different cut-offs used for both the fold change and p-value, with results shown in Fig. 3-6.

Assessing the impact of network topology

To determine whether the perturbation network was consistent with a scale free topology, we tested whether it was well-represented by the formula, $p(K) \sim= K^{-gamma}$, where K is the node out-degree, or number of edges proceeding from the node. p(K) was calculated based on the distribution of K values and the probability density function. The relationship between p(K) and K was then assessed using a least-squares regression of the log(p(K)) on the log(K). The linear

regression was highly significant ($R^2 = 0.73$, p-value = $< 2x10^{-16}$), with a coefficient (which translates to the -gamma value) of -0.75. This value is similar to the gamma of 0.7 calculated for a perturbation network of a smaller yeast gene expression dataset (Featherstone and Broadie 2002).

Two types of permutations to the network topology were conducted. In both, the number of nodes and edges were not changed, only the structure of the network itself. In the first type of permutation, conducted 100 times, we randomly reassigned the target of each directed edge without re-assigning the node the edge came from, changing the structure of the network without affecting the out-degree distribution. We did this by randomly shuffling the values of each row of the adjacency matrix described above. For the second type of permutation, also conducted 100 times, we randomly shuffled all edges of the adjacency matrix by first randomly shuffling the values in each row, and then randomly shuffling the values in each column. This resulted in a random network structure as demonstrated by the roughly normal distribution of out-degrees for each node (Fig. 3-2D).

Measures of fitness for gene deletions

Fitness measurements for gene deletion strains from Maclean et al. (2017) as reported in "Supplementary Data 6" of that publication. The reported fitness measures of gene deletions relative to the reference strain in YPD media were used directly, without any sort of significance cutoff to identify deletions that resulted in statistically significant decrease in fitness.

Statistical analyses

All statistical analyses and plots were produced using R (version 3.5.2) and code available on Github.

Acknowledgements

We thank Luis Zaman for pivotal discussions on the 'friendship paradox' in scale-free networks, and Jianzhi Zhang, Mark Hill, Mo Siddiq, Henry Ertl, Eden McQueen, Anna Redgrave for constructive feedback on the manuscript.

References

Barabási A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5:101–113.

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol.* 3:e245.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134:25–36.

Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. 2014. Tempo and mode of regulatory evolution in Drosophila. *Genome Res.* 24:797–808.

Courtier-Orgogozo V, Arnoult L, Prigent SR, Wiltgen S, & Martin A. 2020. Gephebase, a database of genotype–phenotype relationships for natural and domesticated variation in Eukaryotes. Nucleic acids research, 48(D1), D696-D703.

Featherstone DE, Broadie K. 2002. Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. *Bioessays* 24:267–274.

Feld SL. 1991. Why Your Friends Have More Friends Than You Do. *Am. J. Sociol.* 96:1464–1477.

Flint J, Ideker T. 2019. The great hairball gambit. Copenhaver GP, editor. *PLoS Genet.* 15:e1008519.

Hill MS, Vande Zande P, Wittkopp PJ. 2021. Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* 22:203–215.

Jakobson CM, Jarosz DF. 2019. Molecular Origins of Complex Heritability in Natural Genotype-to-Phenotype Relationships. *Cell Syst* 8:363–379.e3.

Kemmeren P, Sameith K, van de Pasch LAL, Benschop JJ, Lenstra TL, Margaritis T, O'Duibhir E, Apweiler E, van Wageningen S, Ko CW, et al. 2014. Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. *Cell* 157:740–752.

Keren L, Hausser J, Lotan-Pompan M, Vainberg Slutskin I, Alisar H, Kaminski S, Weinberger A,

Alon U, Milo R, Segal E, et al. 2016. Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell* 0:636–643.

Kimura M, Ohta T. 1974. On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 71:2848–2852.

Kopp A, McIntyre LM. 2012. Transcriptional network structure has little effect on the rate of regulatory evolution in yeast. *Mol. Biol. Evol.* 29:1899–1905.

Lutz S, Brion C, Kliebhan M, Albert FW. 2019. DNA variants affecting the expression of numerous genes in trans have diverse mechanisms of action and evolutionary histories.Fay JC, editor. *PLoS Genet.* 15:e1008375.

Luscombe NM, Babu M, Yu H, Snyder M, Teichmann SA, Gerstein M. 2004. Genomic Analysis of Regulatory Network Dynamics Reveals Large Topological Changes. *Nature* 431 (7006): 308–12.

Maclean CJ, Metzger BPH, Yang J-R, Ho W-C, Moyers B, Zhang J, Hernandez R. Deciphering the Genic Basis of Yeast Fitness Variation by Simultaneous Forward and Reverse Genetics. Available from: http://umich.edu/~zhanglab/publications/2017/Maclean_2017_MBE.pdf

Martin A, Orgogozo V. 2013. The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* 67:1235–1250.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res.* 20:816–825.

Metzger BPH, Wittkopp PJ, Coolon JD. 2017. Evolutionary dynamics of regulatory changes underlying gene expression divergence among Saccharomyces species. *Genome Biol. Evol.* 177:1987–1996.

Paaby AB, Rockman MV. 2013. The many faces of pleiotropy. *Trends Genet.* 29:66–73.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. *Nat. Rev. Genet.* 7:862–872.

Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. *Trends Genet.* 34:532–544.

Smyth GK, Michaud J, Scott HS. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* 21:2067–2075.

Stern DL, Orgogozo V. 2008. The loci of evolution: How predictable is genetic evolution? :2155–2177.

Stern DL, Orgogozo V. 2009. Is genetic evolution predictable? *Science* 323:746–751.

Wittkopp PJ. 2005. Genomic sources of regulatory variation in cis and in trans. *Cell. Mol. Life*

Sci. 62:1779–1783.

Wittkopp PJ. 2007. Variable gene expression in eukaryotes: a network perspective. *J. Exp. Biol.* 210:1567–1575.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* 430:85–88.

Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between Drosophila species. *Nat. Genet.* 40:346–350.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8:206–216.

Yang B, Wittkopp PJ. 2017. Structure of the Transcriptional Regulatory Network Correlates with Regulatory Divergence in Drosophila. *Mol. Biol. Evol.* 34:1352–1362.

Figures

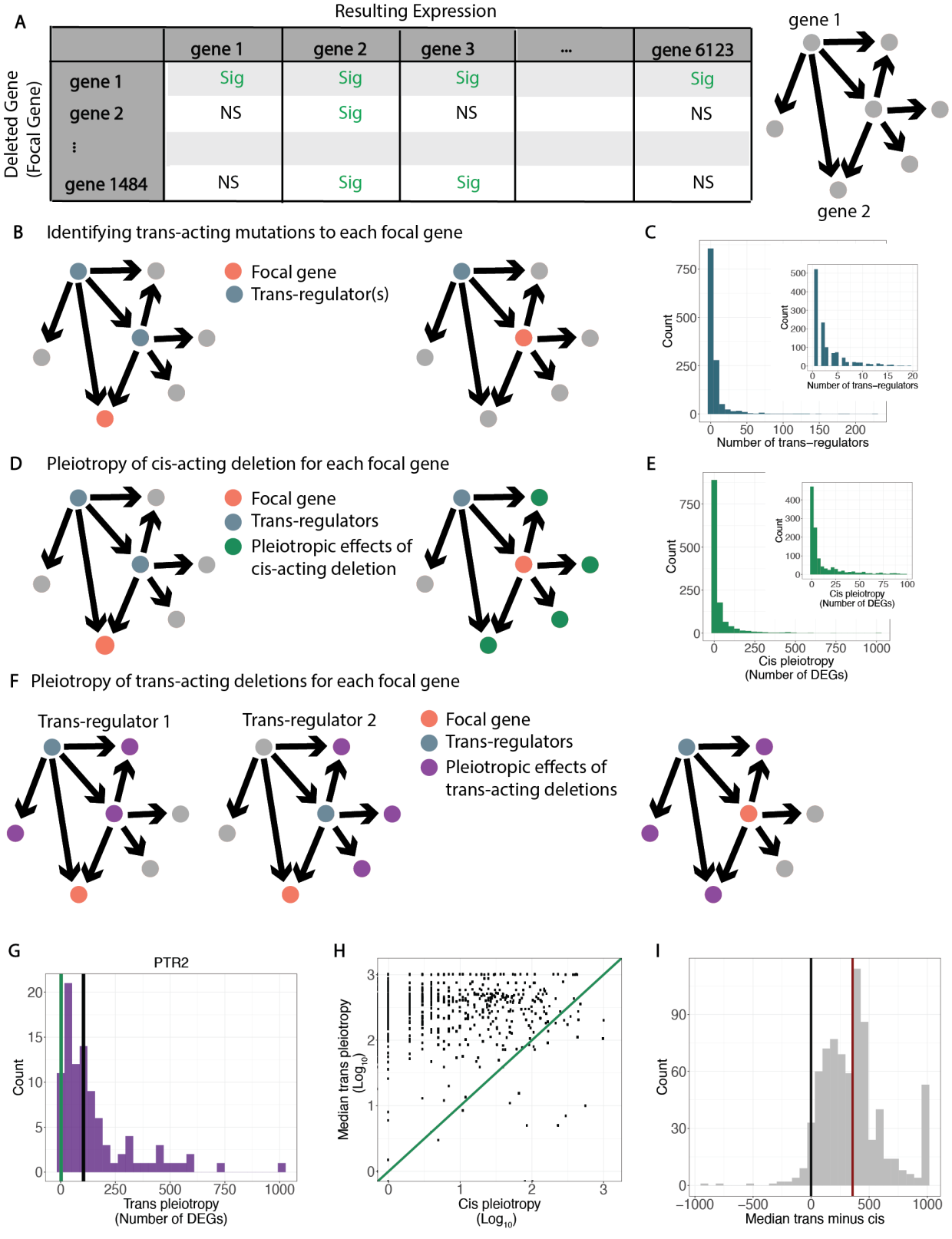
A
Resulting Expression
Deleted Gene (Focal Gene)
gene 1
gene 2
gene 3
...
gene 6123
gene 1 Sig Sig Sig Sig
gene 2 NS Sig NS NS
gene 1484 NS Sig Sig NS
gene 1
gene 2
B Identifying trans-acting mutations to each focal gene
Focal gene
Trans-regulator(s)
C
Count
Number of trans−regulators
D Pleiotropy of cis-acting deletion for each focal gene
Focal gene
Trans-regulators
Pleiotropic effects of cis-acting deletion
E
Count
Cis pleiotropy (Number of DEGs)
F Pleiotropy of trans-acting deletions for each focal gene
Trans-regulator 1
Trans-regulator 2
Focal gene
Trans-regulators
Pleiotropic effects of trans-acting deletions
G
PTR2
Count
Trans pleiotropy (Number of DEGs)
H
Median trans pleiotropy (Log10)
Cis pleiotropy (Log10)
I
Count
Median trans minus cis

Figure 3-1: cis-acting deletions are less pleiotropic than trans-acting deletions

(a) Schematic shows a cartoon example of the perturbation network constructed from gene expression data in *S. cerevisiae* gene deletion mutants. Directed edges are drawn when a deleted gene (gene 1) causes a significant change in expression of another gene (gene 2, among others). In the sample matrix, "Sig" indicates a gene that was significantly differentially expressed in that gene deletion, and "NS" indicates a gene that was not significantly differentially expressed in that gene deletion. *(b)* Two copies of the network schematic from Fig.1A are shown with the first illustrating a gene considered as the focal gene (orange node) that is influenced by two *trans*-regulatory mutations (blue nodes) and the second illustrating a gene considered as the focal gene that is influenced by one *trans*-regulatory mutation. *(c)* Histograms show the number of deletions that affect expression of each focal gene in *trans* (i.e., the number of *trans*-regulatory mutations). Inset histogram is a close-up of the larger histogram with the x-axis ranging from 0-20 for better resolution of that portion of the data. *(d)* The same schematics are shown as in panel B, but with the pleiotropic effects (i.e., genes significantly differentially expressed upon deletion of the focal gene) shown in green. *(e)* Histograms show the distribution of pleiotropic effects for *cis*-regulatory mutations, measured as the number of genes significantly differentially expressed upon deletion of the focal gene, for all focal genes. Inset histogram shows close-up of the x-axis ranging from 0 to 100 for better resolution of that portion of the data. *(f)* Same schematics are shown as in panel B, but with the pleiotropic effects of *trans*-acting mutations for the two focal genes shown in purple. *(g)* The number of differentially expressed genes in the deletion of *PTR2* (pleiotropy of its *cis*-regulatory deletion, green line) is smaller than the median (black line) number of differentially expressed genes for all *trans*-regulatory mutations to *PTR2* (purple histogram). *(h)* For all focal genes, the $\log_{10}$ median pleiotropy of all deletions that act in *trans* to that gene is plotted on the y-axis and the $\log_{10}$ pleiotropy of the deletion that acts in *cis* to that gene is plotted on the x-axis. An x=y line is shown in green. *(i)* A histogram of differences between *cis*-acting pleiotropy and the median *trans*-acting pleiotropy for all focal genes included in the study is shown. The median difference (red line) is significantly greater than zero (black line) (p-value = 3×10^{-165} , one-sided t-test).

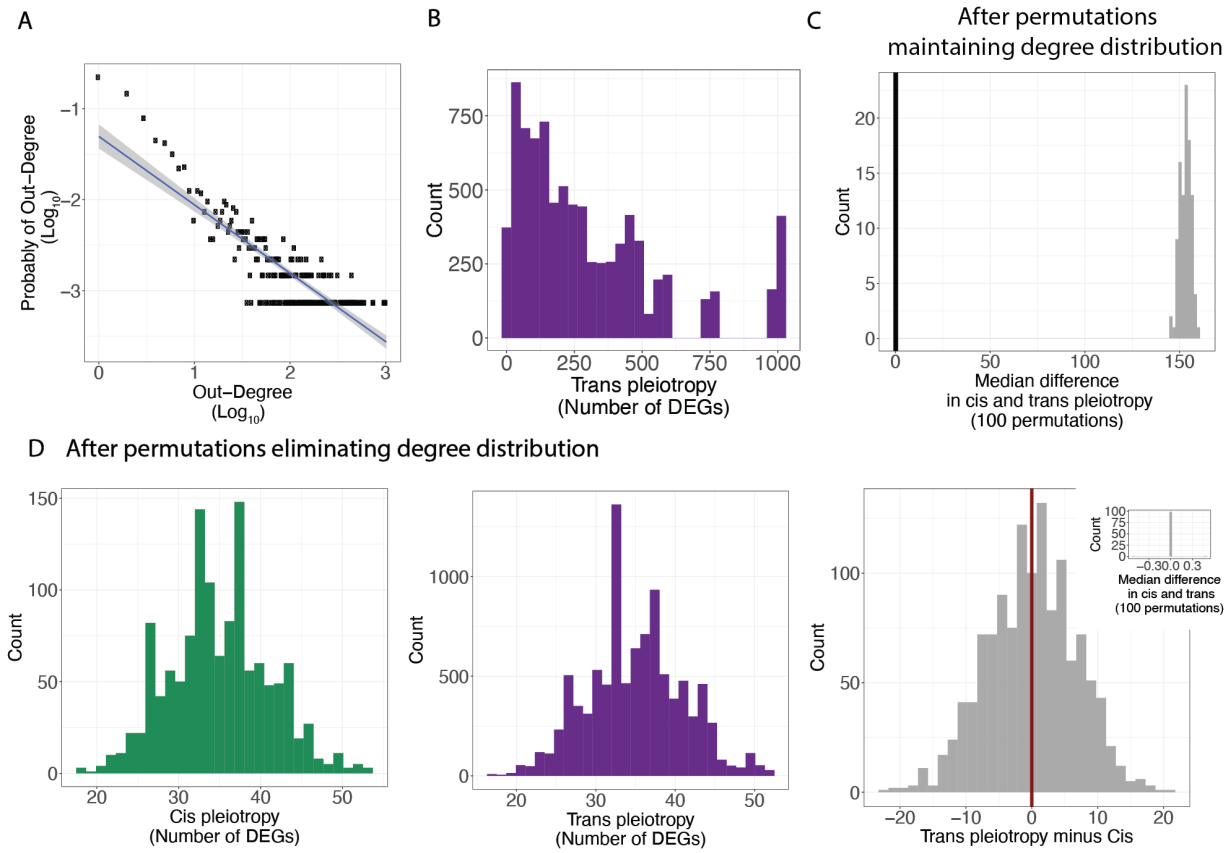

Figure 3-2: Network topology can explain trans-acting mutations tending to be more pleiotropic than cis-acting mutations

(a) The out-degree distribution, or number of outgoing edges from each node (gene), follows a power-law distribution, as indicated by a significant linear relationship (black line surrounded by 95% confidence intervals) between the log of the out-degree and the log of the probability of that out-degree (see Methods). *(b)* A histogram showing the distribution of all out-degrees for all *trans*-regulatory mutations for all focal genes. *(c)* After the network edges were permuted (100 permutations) to change connections between individual genes while maintaining the overall degree distribution, all 100 medians of the distributions of differences between the pleiotropy of the *cis*-acting mutation and the median pleiotropy of *trans*-acting mutations affecting expression of the same focal genes were higher than zero, showing the greater median pleiotropy of *trans*- than *cis*- acting mutations in all 100 permuted networks. *(d)* Permuting edges of the perturbation network without maintaining the degree distribution results in distributions of pleiotropy for *cis* (green) and *trans*-regulatory mutations (purple) that are not significantly different from each other (Welch two-sample t-test, p-value = 0.8221). This type of permutations was conducted 100 times. In 98 cases, median difference between the pleiotropy of the cis-acting deletion and the median pleiotropy of the trans-acting deletion was zero, in one case it was -0.5 and in one case 0.5, resulting in a p-value of 1 (inset).

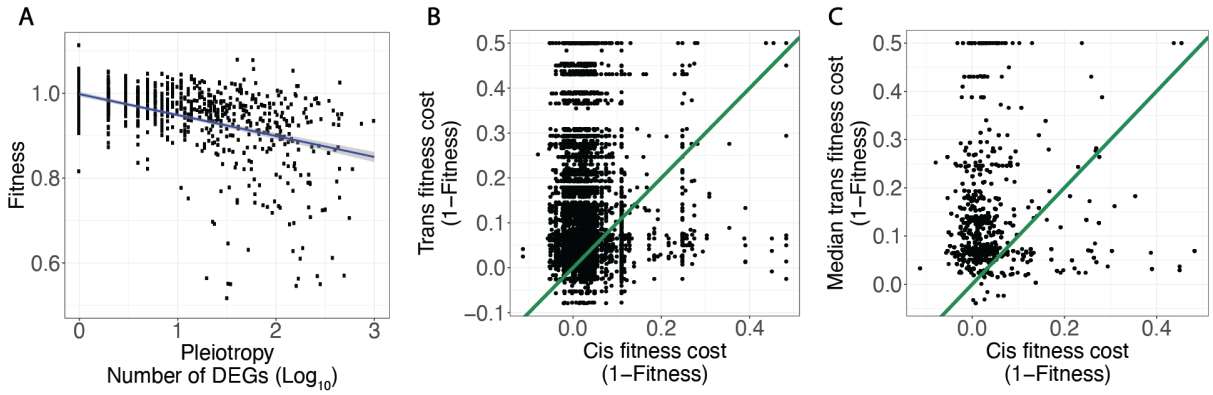

Figure 3-3: trans-acting deletions tend to decrease fitness more than cis-acting deletions affecting expression of the same focal gene

(a) The fitness of each gene deletion strain (measures taken from Maclean et al. 2017) is plotted against the $\log_{10}$ number of significantly differentially expressed genes attributable to the deletion of that gene (using data from Kemmeren et al. 2014). The fit line is a least-squares regression surrounded by the 95% confidence interval. *(b)* For every focal gene, one minute the fitness of each *trans*-acting deletion (signifying the fitness cost of the mutation) is plotted on the y-axis and the fitness cost of the *cis*-acting deletion is plotted on the x-axis. Each focal gene thus has one x value and many y values. An x=y line, corresponding to equal fitness of *cis*- and *trans*-acting deletions, is plotted in green. *(c)* For all focal genes, the median fitness cost of all *trans*-acting deletions to that gene is plotted on the y-axis and the fitness cost of the *cis*-acting deletion is plotted on the x-axis. An x=y line is shown in green.

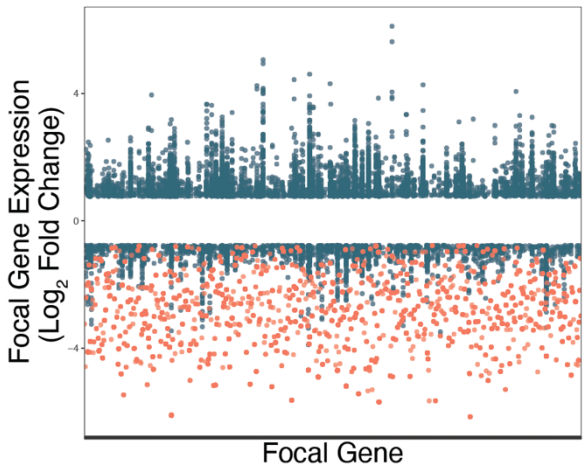

Figure 3-4: cis-acting mutations tend to have larger effects on expression of the focal gene than trans-acting mutations

(a) For each deletion mutant (focal gene, x-axis), the value of the focal gene's expression level relative to the wild-type control is plotted for the deletion mutant itself (*cis*-acting mutation, orange points) and all other deletion mutants in which the focal gene was significantly differentially expressed (*trans*-acting mutations, blue points).

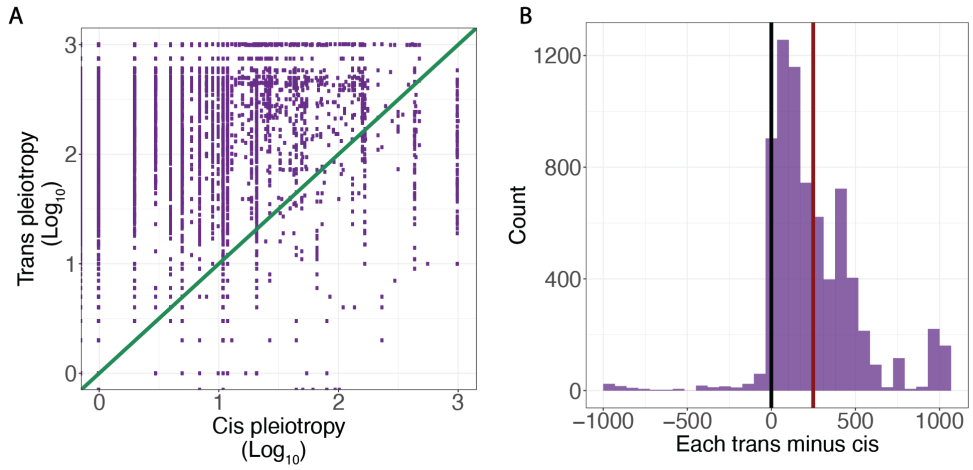

Figure 3-5: trans-acting mutations tend to be more pleiotropic than cis-acting mutations affecting expression of the same focal gene in pairwise comparisons

(a) For every focal gene, the pleiotropy of each *trans*-acting deletion is plotted on the y-axis and the pleiotropy of the *cis*-acting deletion is plotted on the x-axis. Therefore, for each focal gene there is one x value and many y values. An x=y line is plotted in green. *(b)* A histogram of the differences between the pleiotropy of *cis*- and *trans*-acting mutations for all pairs of one *trans*-regulator and the corresponding focal gene (all purple points in panel A). The median (red line) is significantly higher than zero (one-sided t-test, p-value = 0).

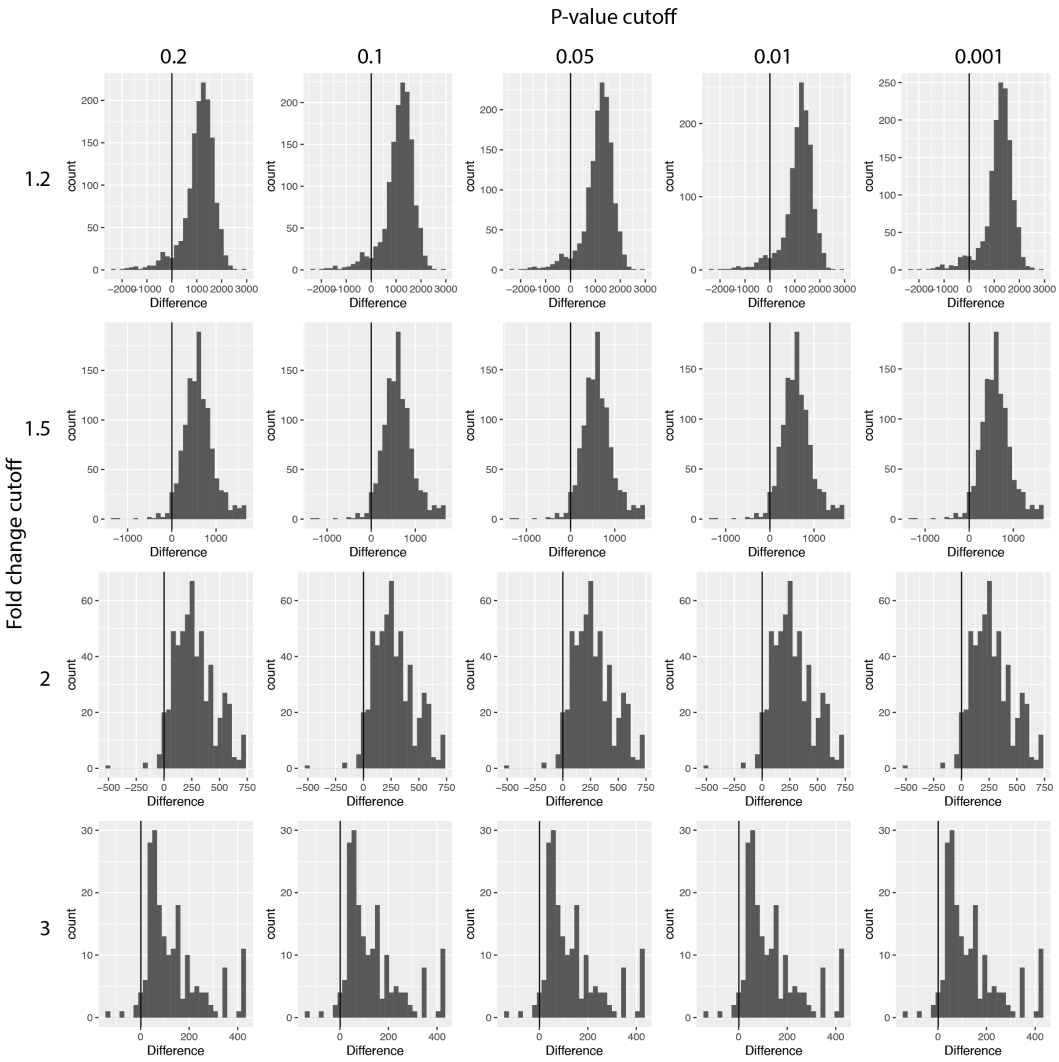

Figure 3-6: Differences between median pleiotropy of trans-acting deletions and cis-acting deletions for all focal genes are robust to changes in thresholds used for differential expression

Each panel shows the distribution of differences in pleiotropy between the *cis*-acting deletion and the median pleiotropy of the *trans*-acting deletions affecting expression of the same focal gene. The fold-change cutoffs (rows) and *P*-value cutoffs (columns) used to classify a gene as significantly differentially expressed in each case are shown. The similarity of distributions across columns demonstrates that fold-change cutoffs had a larger impact on the number of edges in the network than *P*-value cutoffs. Despite these effects, the median of the difference in pleiotropy between the cis- and trans-acting mutations was higher than zero for all cutoff combinations, showing that the *trans*-acting mutations tended to be more pleiotropic than the *cis*-acting mutation affecting expression of the same focal gene regardless of the criteria used to identify significant changes in gene expression.

Chapter 4 Differing Mechanisms of Active Compensation for Reduction of *TDH3* Activity by its Paralogs *TDH1* and *TDH2*

Abstract

Paralogous genes that retain overlapping functions can confer robustness to genetic networks by compensating for each other. Compensation occurs passively when the normal activity of one paralog can compensate for the loss of the other, or actively when a change in one paralog's expression, localization, or activity is required to compensate for loss of the other. Here we explore the mechanisms of active compensation for loss or reduction in expression of the *Saccharomyces cerevisiae* gene *TDH3* by its paralogs *TDH1* and *TDH2*. *TDH1* and *TDH2* are upregulated in dose-responsive manners to reductions in *TDH3* by a mechanism that also upregulates a wild type *TDH3* promoter. *TDH2* is not upregulated when *TDH3* expression is lowered via mutations in the transcription factors Rap1p and Gcr1p, which regulate both *TDH3* and *TDH2* expression, suggesting that active compensation by *TDH2* occurs via homeostatic feedback mechanisms involving the transcription factors Gcr1p and Rap1p. *TDH1* is upregulated in Gcr1p and Rap1p mutants, indicating a different mechanism of compensation by *TDH1*. Other glycolytic genes regulated by Rap1p and Gcr1p show a similar pattern to *TDH2*, indicating that the mechanism of active compensation by *TDH2* is not specific to the paralog, but a general homeostatic response to reductions in *TDH3* expression.

Introduction

Biological systems are often robust to genetic and environmental perturbations (Félix and Barkoulas 2015; Gibson and Lacek 2020). This robustness is explained at least in part by the presence of independent genes in the genome with overlapping functions (Diss et al. 2014). Such genes often arise evolutionarily through gene duplication events that give rise to two or more paralogous genes (Wagner 2000; Gu et al. 2003). Divergence of duplicate genes is often a prerequisite for their survival (Lynch and Force 2000; Zhang 2003), yet many paralogous genes retain overlapping functions that contribute to robustness (Kafri et al. 2006; Ihmels et al. 2007; Dean et al. 2008; DeLuna et al. 2008; Kafri et al. 2008; Hanada et al. 2009; Li et al. 2010; Kuzmin et al. 2020). As described in Diss et al. (2014), paralogs can contribute to phenotypic robustness through either passive or active mechanisms. In passive paralogous compensation, the normal activity of one of the paralogs is sufficient to minimize the phenotypic impact of losing the activity of the other paralog, whereas in active paralogous compensation, activity of one paralog is changed in some way in response to loss of activity of the other paralog such that its phenotypic impact is reduced. One such type of change in activity of the paralog is an increase in its expression level leading to more protein available to perform the function of the mutated gene. While some examples of active compensation by upregulation of a paralog exist (Rudnicki et al. 1992; DeLuna et al. 2010; Denby et al. 2012; Dong et al. 2016; Dohn and Cripps 2018; Rodriguez-Leal et al. 2019), the molecular mechanisms behind active compensation remain to be elucidated (Diss et al. 2014).

The *TDH1, TDH2,* and *TDH3* genes of *Saccharomyces cerevisiae* are an example of a set of paralogous genes that appear to be able to compensate for each other's function. All three of these genes encode proteins that act as glyceraldehyde-3-phosphate dehydrogenases (GAPDHs)

(McAlister and Holland 1985a; Linck et al. 2014), catalyzing a central step in both glycolysis and gluconeogenesis. The *TDH2* and *TDH3* proteins are most similar to each other, retaining 94% amino acid sequence identity (Holland and Holland 1980; Engel et al. 2014), with the *TDH1* and *TDH3* proteins having 89% amino acid sequence identity (Holland et al. 1983; Engel et al. 2014). *TDH2* and *TDH3* are also more similar in their expression pattern, as they are expressed during exponential growth, while *TDH1* is expressed primarily during stationary phase (Delgado et al. 2001; Bradley et al. 2019). The first characterization of the *TDH* paralogs reported a reduction in fitness when either *TDH1, TDH2,* or *TDH3* were disrupted by insertion mutations (McAlister and Holland 1985b). The triple mutant and the double *tdh2-Δ,tdh3-Δ* mutant were reported to be lethal, while the *tdh1-Δ,tdh3-Δ* mutant showed a larger reduction in fitness than would be expected for an additive effect of the single mutant fitness measures (McAlister and Holland 1985b). A more recent assay for negative genetic interactions in *S. cerevisiae* found that deletion of *TDH3* reduced fitness to ~90% of wild type, while deletion of either *TDH1* or *TDH2* did not have a significant impact on fitness (Costanzo et al. 2010). The double mutant *tdh1-Δ,tdh3-Δ* had a slightly significant negative genetic interaction, while the *tdh2-Δ,tdh3-Δ* double mutant demonstrated a strong negative genetic interaction, growing at only 7% relative to wild type (Costanzo et al. 2010). These nonadditive impacts on fitness suggest that the functional overlap of these paralogs allows them to compensate for each other. The mechanism by which they may compensate for each other and whether they do so in an active or passive manner, however, is unknown.

Here we test the hypothesis that a reduction in the *TDH3* protein is actively compensated for by one or both of its paralogs, *TDH1* or *TDH2,* via a mechanism of transcriptional upregulation. We use RNA-sequencing data from a series of *S. cerevisiae* strains bearing

mutations in the *TDH3* promoter that result in expression of *TDH3* ranging from 0% to 135% of wild type expression to show that *TDH1* and *TDH2* expression are upregulated in a dose-responsive manner when *TDH3* expression is reduced. A wild-type *TDH3* promoter driving expression of a reporter gene was also upregulated when expression of the native *TDH3* gene was reduced, suggesting that the cells respond to reduced *TDH3* expression by increasing activity of one or more factors that regulate expression of *TDH3*. Rap1p and Gcr1p are transcription factors known to directly regulate *TDH3* as well as expression of other glycolytic genes, thus we also tested whether the reduced expression of *TDH3* seen in strains with mutated versions of Rap1p or Gcr1p caused a upregulation of *TDH1* and/or *TDH2*. We found that *TDH1* but not *TDH2* was upregulated in these mutants, suggesting that Rap1p and Gcr1p are involved in the upregulation of *TDH2* in response to reduced *TDH3* activity. These data are consistent with a model in which feedback mechanisms that homeostatically regulate expression of *TDH3* also cause the upregulation of other genes regulated by Rap1p and Gcr1p, including *TDH2*. Prior work indicates that paralogs retaining some shared transcriptional regulators often compensate for each other (Kafri et al. 2005; He and Zhang 2006; Kafri et al. 2006), suggesting that the molecular mechanism we identify here for active compensation among *TDH* paralogs might also apply to other sets of paralogous genes.

Results

Active compensation for loss of TDH3 by paralogs TDH1 and TDH2

To determine whether the compensation for loss of *TDH3* activity by *TDH1* or *TDH2* might be mediated by changes in their expression, we used RNA-seq data describing changes in gene expression when *TDH3* was deleted (Chapter 2) to test for significant changes in expression of *TDH1* and/or *TDH2*. We found that both genes showed significantly higher expression in the

tdh3-Δ deletion strain than in the unmutated wild-type strain (Figure 4-1A, Wald test P-value for *TDH1* = $2x10^{-5}$, P-value for *TDH2* = 0.04). Using strains carrying mutations in the promoter of *TDH3* that resulted in *TDH3* expression levels 20%, 50%, and 85% of wild type, as well as a strain expressing *TDH3* at 135% of wildtype levels carrying a duplication of a mutant *TDH3* allele (Chapter 2), we also tested for changes in *TDH1* or *THD2* expression in response to more moderate alterations in *TDH3* activity using RNA-seq data. We found that *TDH2* expression was negatively correlated to *TDH3* expression among these strains, with *TDH2* expression increased when *TDH3* expression was decreased and *TDH2* expression decreased when *TDH3* expression was increased (Figure 4-1B). *TDH1*, on the other hand, showed more of a threshold-like relationship to *TDH3* expression: *TDH1* expression increased to a similar extent in the mutants expressing *TDH3* at 20%, 50%, and 85% of wild-type levels, but increased much more strongly in the *TDH3* null mutant strain and decreased in the *TDH3* overexpression strain (Figure 4-1C). Taken together, these data show that expression of *TDH1* and *TDH2* changes when *TDH3* expression is altered in ways expected to help compensate for the changes in *TDH3* expression.

All of the strains with modified *TDH3* expression also carried a wild-type *TDH3* promoter driving expression of a yellow fluorescent protein (*YFP*) (Duveau et al. 2017). We noticed that expression of this reporter gene was also increased when *TDH3* expression was decreased (Figure 4-1D), suggesting that factors regulating expression of *TDH3* itself might be involved in the mechanism of active compensation. The transcription factors Rap1p and Gcr1p are known to directly regulate expression of *TDH3* (Figure 4-1E, Huie et al. 1992; Yagi et al. 1994) as well as expression of other glycolytic genes, including *TDH1* and *TDH2* (MacIsaac et al. 2006; Hu et al. 2007; Venters et al. 2011; Lickwar et al. 2012). In fact, the mutations altering expression of *TDH3* in the mutant strains expressing *TDH3* at 20%, 50%, 85%, and 135% of

wild-type expression levels all contained mutations in either the Rap1p or Gcr1p binding sites (Figure 4-1G). Taken together, these observations suggest that changes in expression of *TDH1* and *TDH2* in response to changes in expression of *TDH3* might be caused by homeostatic feedback mechanisms involving Rap1p and/or Gcr1p.

TDH2 is not upregulated when TDH3 expression is reduced by mutations in RAP1 or GCR1

If the upregulation of *TDH1* and *TDH2* upon reduction of *TDH3* expression is the result of a homeostatic feedback mechanism involving Rap1p and Gcr1p, we would not expect to observe an increase in expression of *TDH1* and *TDH2* when *TDH3* expression is reduced via mutations in Rap1p and Gcr1p that affect their ability to regulate *TDH3* expression. Using RNA-seq data from a set of 9 strains of *S. cerevisiae* each carrying 1-6 mutations in the *RAP1* (4 mutants) or *GCR1* (5 mutants) gene previously shown to affect *TDH3* expression (Duveau et al. 2021), we asked whether the changes in *TDH3* expression observed in these strains were accompanied by compensatory changes in expression of *TDH1, TDH2,* and/or *YFP* driven by a wild type *TDH3* promoter. One *GCR1* mutant strain carries a single nucleotide deletion resulting in an early stop codon, suggesting it is likely to be a null mutation, whereas the other mutant alleles are more likely to be hypo- or hypermorphs. *GCR1* mutants showed *TDH3* expression levels ranging from ~7% to ~105% of wild type *TDH3* expression (Figure 4-2A), with the likely null mutant allele showing the largest reduction in expression of *TDH3*. *RAP1* null mutants are lethal (Giaever et al. 2002), suggesting that all of the *RAP1* mutants examined were either hypomorphs or gain of function alleles which caused changes in *TDH3* expression ranging from ~20% to ~115% (Figure 4-2A). We found that when *TDH3* expression was reduced in these strains, *TDH2* expression did not increase (Figure 4-2B). Expression of *TDH1*, on the other hand, was increased in the *RAP1* and *GCR1* mutants (Figure 4-2C). The expression of the *YFP* reporter

gene being driven by a wild type *TDH3* promoter closely matched expression of *TDH3* itself (Figure 4-2D), demonstrating that these two promoters function similarly when neither are mutated. These data are consistent with a model in which upregulation of *TDH2*, but not *TDH1*, is the result of homeostatic feedback mechanisms involving Rap1p and Gcr1p, which consequently does not occur when these transcription factors are mutated.

Upregulation of *TDH2* upon reduction of *TDH3* when *RAP1* and *GCR1* are not mutated could occur via an upregulation of Rap1p and/or Gcr1p themselves. Examining our RNA-sequencing data for strains bearing mutant alleles of the *TDH3* promoter, we found that *GCR1* transcripts were upregulated linearly in response to reductions in *TDH3* expression caused by mutations in the *TDH3* promoter (Figure 4-2E). Transcription of *RAP1,* however, was not upregulated upon reduction in *TDH3* expression (Figure 4-2E). Therefore, changes in *GCR1* abundance might be primarily responsible for the linear response to reductions in *TDH3* expression seen in *TDH2* and the intact *TDH3* promoter expressing *YFP*.

Upregulation of genes regulated by Gcr1p/Rap1p upon reduction in TDH3 is not limited to TDH2

Upregulation by Rap1p and Gcr1p upon reduction in *TDH3* might not be limited to *TDH2*, but might extend to other genes that are regulated by these transcription factors. To see if this is the case, we used our RNA-sequencing data to examine the expression levels of other genes which encode enzymes that catalyze steps in glycolysis and which are regulated by Rap1p and Gcr1p (Figure 4-3A). We found that the genes *PGK1*, *ENO1*, and *PFK2* were significantly upregulated in the *thd3-Δ* null mutant. Fold changes in expression of these genes showed a similar pattern to *TDH2*, with expression increasing when *TDH3* expression is lowered via mutations in the *TDH3* promoter (Figure 4-3B), but not when it is lowered via mutations in

RAP1 or *GCR1* (Figure 4-3C). The genes *FBA1, TPI1,* and *GPM1* showed a similar pattern of expression fold changes (Figure 4-3D,E), but of a smaller magnitude, and were not significantly upregulated in the *thd3-Δ* null mutant. This common pattern among glycolytic genes regulated by Gcr1p/Rap1p further suggests that reduction in *TDH3* expression leads to homeostatic feedback in which glycolytic genes regulated by the transcription factors Gcr1p and Rap1p are upregulated, including *TDH2*.

Discussion

These findings are consistent with a model in which reduction in *TDH3* expression triggers feedback mechanisms aimed at increasing its expression level via the transcription factors Rap1p and Gcr1p. *TDH3* expression cannot be increased because the transcription factor binding sites for Rap1p or Gcr1p have been destroyed (or because the locus is absent in the null mutant), but expression of other genes regulated by Gcr1p and Rap1p is increased, including the *TDH3* paralog *TDH2.* This results in active compensation for reduction in *TDH3* by an upregulation of *TDH2* that does not occur when *RAP1* or *GCR1* are mutated (Figure 4-3F).

It appears that the upregulation of *TDH2* by Gcr1p/Rap1p might be achieved by an upregulation of expression of the *GCR1* gene. Although transcriptional upregulation is not the only mechanism of activation of transcription factors (Hahn and Young 2011), *GCR1* has been shown to be both transcriptionally and post-transcriptionally regulated by glucose availability (Hossain et al. 2016). Reductions in *TDH3* expression hindering metabolic flux through glycolysis and leading to an abundance of glucose might therefore result in an upregulation of *GCR1* transcripts and protein available to upregulate the *TDH* genes. *RAP1*, on the other hand, performs roles in telomere maintenance and activation of ribosomal protein genes in addition to the activation of glycolytic genes (Sussel and Shore 1991; Shore 1994), and is not known to be

transcriptionally regulated in response to glucose availability. Rap1p and Gcr1p act in a complex to activate target gene expression, with Gcr1p being the major activator of the complex (Piña et al. 2003). It is likely, therefore, that upregulation of *GCR1* upon reduction in *TDH3* is primarily responsible for the upregulation of the Rap1p/Gcr1p complex's target genes.

Active compensation by *TDH1* appears to occur via a different mechanism, as indicated by its more threshold-like response to reduction in *TDH3* expression and its upregulation in strains bearing mutations in *RAP1* and *GCR1*. These differences in how *TDH1* and *TDH2* respond to reduction in *TDH3* expression may not be surprising considering the fact that the expression pattern of *TDH1* has diverged from that of the other two paralogs (McAlister and Holland 1985a). *TDH1* has been shown to be upregulated under various stress conditions causing slow growth (Linck et al. 2014), and may therefore be upregulated by a mechanism related to the slower growth of mutants with reduced *TDH3* expression level rather than feedback specifically involving Gcr1p or Rap1p.

The fact that the upregulation of *TDH1* and *TDH2* does not completely eliminate the fitness effect of deleting or reducing *TDH3* expression suggests that the functions of these paralogs have diverged to some extent and cannot completely compensate for each other. Such partial subfunctionalization is thought to occur relatively frequently (Harrison et al. 2007; Kuzmin et al. 2020), and suggests that the maintenance of these paralogs by natural selection is not exclusively due to their ability to compensate for each other. Although *TDH3* is best known for its roles in glycolysis and gluconeogenesis, it has also been implicated in transcriptional silencing (Ringel et al. 2013), RNA-binding (Shen et al. 2014) and possibly antimicrobial defense (Branco et al. 2014). These functions may not be able to be compensated for by *TDH1* or *TDH2* despite their high levels of protein conservation.

The mechanism of active compensation by *TDH2* presented here appears to be the result of shared regulators between paralogs that may be a consequence of the paralogs common ancestry rather than a result of natural selection for robustness. Natural selection undoubtedly plays a role in the maintenance of regulatory structures, but may not be selecting for robustness itself. Active compensation by *TDH1*, however, may not be due to similar regulatory structures and should be further explored. In this way, the case of active compensation by *TDH2* provides an example of the molecular mechanisms by which compensation can take place, while the case of *TDH1* hints at the variety of possible mechanisms of active compensation that could be at play for paralogous genes across the genome.

Materials and Methods

Strains used in this study

The *S. cerevisiae* strains used in this study include a set of 5 *cis*-regulatory mutants to the yeast gene *TDH3* and their wild type reference strain, as well as 9 *trans*-regulatory mutants bearing mutations in either the *RAP1* or *GCR1* gene, and their wild type reference strain. Strains bearing *cis*-regulatory mutations to *TDH3,* including the tdh3-**Δ** strain, were first described in (Duveau et al. 2017). They are haploid, mating type **a** strains of *S. cerevisiae* derived from S288C and constructed from the progenitor strain YPW1001, which contains a wild type P_{TDH3}-*YFP* construct and a *NatMX4* drug resistance marker at the *HO* locus. It also bears alleles of *MKT1*, *SAL1*, *CAT5* and *MIP1* decreasing petite frequency and the alleles of *RME1* and *TAO3* increasing sporulation efficiency, as previously described (Duveau et al. 2017). The collection numbers and specific mutations in each strain, as well as their impacts on *TDH3* expression, are detailed in Table 4. Strains bearing mutations in the genes *RAP1* or *GCR1* are first described in (Duveau et al. 2021). They are haploid, mating type alpha strains of *S. cerevisiae* derived from

S288C and constructed from the progenitor strain YPW1139, which also contains a wild type P_{TDH3}-*YFP* construct and a *KanMX* drug resistance marker at the *HO* locus and also bears alleles of *MKT1*, *SAL1*, *CAT5* and *MIP1* decreasing petite frequency and the alleles of *RME1* and *TAO3* increasing sporulation efficiency, as previously described (Duveau et al. 2017). The mating type and drug resistance marker are the only differences in the background of the strains bearing *cis*-regulatory mutations and strains bearing mutations in either *RAP1* or *GCR1,* and have been shown not to influence expression of genes mis-expressed upon reduction in *TDH3* expression (Chapter 2, Materials and Methods). Mutants analyzed with mutations in *GCR1* or *RAP1* were constructed from this progenitor strain by using mutagenic PCR to randomly introduce mutations within each gene and then using CRISPR-mediated allele-replacement to substitute the native locus with a mutant allele. The collection numbers and specific mutations in each strain, as well as their impacts on *TDH3* expression, are detailed in Table 4.

Gene expression data

RNA-sequencing data presented in this paper is a subset of the data described in Chapter 2, and which are available at GEO accession GSE175398. That dataset consists of RNA-sequencing data for *cis*-regulatory mutants and a larger set of *trans*-regulatory mutants affecting *TDH3* expression. Details of data collection and processing are available in Chapter 2: Materials and Methods, and are summarized here. Briefly, yeast cells were grown to mid log phase in glucose media, pelleted, and frozen at -80C. polyA RNA was extracted from frozen cell pellets using oligodT magnetic beads. RNA libraries were prepared for sequencing using a ⅓ volume TruSeq RNA Sample Preparation v2 kit (Illumina), and sequenced on a HiSeq 4000 by the University of Michigan Sequencing Core. Each genotype (all mutants and non mutated reference strains) was assayed in quadruplicate with each replicate consisting of a unique random array of

genotypes and controls in a 96 well plate. Reads were pseudomapped to the *S.cerevisiae* transcriptome (Ensemble, release 38, retrieved from ftp://ftp.ensemblgenomes.org/pub/release-38/fungi/fasta/saccharomyces_cerevisiae/cdna/), and DeSeq2 (Love et al. 2014) was used to estimate $\log_2$ fold changes and significance values reported in the text.

References

Bradley PH, Gibney PA, Botstein D, Troyanskaya OG, Rabinowitz JD. 2019. Minor Isozymes Tailor Yeast Metabolism to Carbon Availability. mSystems [Internet] 4. Available from: http://dx.doi.org/10.1128/mSystems.00170-18

Branco P, Francisco D, Chambon C, Hébraud M, Arneborg N, Almeida MG, Caldeira J, Albergaria H. 2014. Identification of novel GAPDH-derived antimicrobial peptides secreted by Saccharomyces cerevisiae and involved in wine microbial interactions. Appl. Microbiol. Biotechnol. 98:843–853.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, et al. 2010. The Genetic Landscape of a Cell. Science 327.

Dean EJ, Davis JC, Davis RW, Petrov DA. 2008. Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet. 4:e1000113.

Delgado ML, O'Connor JE, Azorı N I, Renau-Piqueras J, Gil ML, Gozalbo D. 2001. The glyceraldehyde-3-phosphate dehydrogenase polypeptides encoded by the Saccharomyces cerevisiae TDH1, TDH2 and TDH3 genes are also cell wall proteins. Microbiology 147:411–417.

DeLuna A, Springer M, Kirschner MW, Kishony R. 2010. Need-based up-regulation of protein levels in response to deletion of their duplicate genes. PLoS Biol. 8:e1000347.

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colón-González M, Chao S, Kishony R. 2008. Exposing the fitness contribution of duplicated genes. Nat. Genet. 40:676–681.

Denby CM, Im JH, Yu RC, Pesce CG, Brem RB. 2012. Negative feedback confers mutational robustness in yeast transcription factor regulation. Proc. Natl. Acad. Sci. U. S. A. 109:3874–3878.

Diss G, Ascencio D, DeLuna A, Landry CR. 2014. Molecular mechanisms of paralogous compensation and the robustness of cellular networks. J. Exp. Zool. B Mol. Dev. Evol. 322:488–499.

Dohn TE, Cripps RM. 2018. Absence of the Drosophila jump muscle actin Act79B is compensated by up-regulation of Act88F. Dev. Dyn. 247:642–649.

Dong OX, Tong M, Bonardi V, El Kasmi F, Woloshen V, Wünsch LK, Dangl JL, Li X. 2016. TNL-mediated immunity in Arabidopsis requires complex regulation of the redundant ADR1 gene family. New Phytol. 210:960–973.

Duveau F, Toubiana W, Wittkopp PJ. 2017. Fitness effects of cis -regulatory variants in the Saccharomyces cerevisiae TDH3 promoter.

Duveau F, Vande Zande P, Metzger BP, Diaz CJ, Walker EA, Tryban S, Siddiq MA, Yang B, Wittkopp PJ. 2021. Mutational sources of trans-regulatory variation affecting gene expression in Saccharomyces cerevisiae. Elife [Internet] 10. Available from: http://dx.doi.org/10.7554/eLife.67806

Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, et al. 2014. The reference genome sequence of Saccharomyces cerevisiae: then and now. G3 4:389–398.

Félix M-A, Barkoulas M. 2015. Pervasive robustness in biological systems. Nat. Rev. Genet. 16:483–496.

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

Gibson G, Lacek KA. 2020. Canalization and Robustness in Human Genetics and Disease. Annu. Rev. Genet. 54:189–211.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421:63–66.

Hahn S, Young ET. 2011. Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. Genetics 189:705–736.

Hanada K, Kuromori T, Myouga F, Toyoda T, Li W-H, Shinozaki K. 2009. Evolutionary persistence of functional compensation by duplicate genes in Arabidopsis. Genome Biol. Evol. 1:409–414.

Harrison R, Papp B, Pál C, Oliver SG, Delneri D. 2007. Plasticity of genetic interactions in metabolic networks of yeast. Proc. Natl. Acad. Sci. U. S. A. 104:2307–2312.

He X, Zhang J. 2006. Transcriptional reprogramming and backup between duplicate genes: is it a genomewide phenomenon? Genetics 172:1363–1367.

Holland JP, Holland MJ. 1980. Structural comparison of two nontandemly repeated yeast glyceraldehyde-3-phosphate dehydrogenase genes. J. Biol. Chem. 255:2596–2605.

Holland JP, Labieniec L, Swimmer C, Holland MJ. 1983. Homologous nucleotide sequences at the 5’ termini of messenger RNAs synthesized from the yeast enolase and glyceraldehyde-3-

phosphate dehydrogenase gene families. The primary structure of a third yeast glyceraldehyde-3-phosphate dehydrogenase gene. J. Biol. Chem. 258:5291–5299.

Hossain MA, Claggett JM, Edwards SR, Shi A, Pennebaker SL, Cheng MY, Hasty J, Johnson TL. 2016. Posttranscriptional Regulation of Gcr1 Expression and Activity Is Crucial for Metabolic Adjustment in Response to Glucose Availability. Mol. Cell 62:346–358.

Huie MA, Scott EW, Drazinic CM, Lopez MC, Hornstra IK, Yang TP, Baker HV. 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.

Hu Z, Killion PJ, Iyer VR. 2007. Genetic reconstruction of a functional transcriptional regulatory network. Nat. Genet. 39:683–687.

Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS. 2007. Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. Mol. Syst. Biol. 3:86.

Kafri R, Bar-Even A, Pilpel Y. 2005. Transcription control reprogramming in genetic backup circuits. Nat. Genet. 37:295–299.

Kafri R, Dahan O, Levy J, Pilpel Y. 2008. Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. Proc. Natl. Acad. Sci. U. S. A. 105:1243–1248.

Kafri R, Levy M, Pilpel Y. 2006. The regulatory utilization of genetic redundancy through responsive backup circuits. Proc. Natl. Acad. Sci. U. S. A. 103:11653–11658.

Kuzmin E, VanderSluis B, Nguyen Ba AN, Wang W, Koch EN, Usaj M, Khmelinskii A, Usaj MM, van Leeuwen J, Kraus O, et al. 2020. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. Science [Internet] 368. Available from: http://dx.doi.org/10.1126/science.aaz5667

Lickwar CR, Mueller F, Hanlon SE, McNally JG, Lieb JD. 2012. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature 484:251–255.

Li J, Yuan Z, Zhang Z. 2010. The cellular robustness by genetic redundancy in budding yeast. PLoS Genet. 6:e1001187.

Linck A, Vu X-K, Essl C, Hiesl C, Boles E, Oreb M. 2014. On the role of GAPDH isoenzymes during pentose fermentation in engineered Saccharomyces cerevisiae. FEMS Yeast Res. 14:389–398.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. 2006. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 7:113.

McAlister L, Holland MJ. 1985a. Differential expression of the three yeast glyceraldehyde-3-phosphate dehydrogenase genes. J. Biol. Chem. 260:15019–15027.

McAlister L, Holland MJ. 1985b. Isolation and characterization of yeast strains carrying mutations in the glyceraldehyde-3-phosphate dehydrogenase genes. J. Biol. Chem. 260:15013–15018.

Piña B, Fernández-Larrea J, García-Reyero N, Idrissi F-Z. 2003. The different (sur)faces of Rap1p. Mol. Genet. Genomics 268:791–798.

Ringel AE, Ryznar R, Picariello H, Huang K-L, Lazarus AG, Holmes SG. 2013. Yeast Tdh3 (Glyceraldehyde 3-Phosphate Dehydrogenase) Is a Sir2-Interacting Factor That Regulates Transcriptional Silencing and rDNA Recombination.Pikaard CS, editor. PLoS Genet. 9:e1003871.

Rodriguez-Leal D, Xu C, Kwon C-T, Soyars C, Demesa-Arevalo E, Man J, Liu L, Lemmon ZH, Jones DS, Van Eck J, et al. 2019. Evolution of buffering in a genetic circuit controlling plant stem cell proliferation. Nat. Genet. 51:786–792.

Rudnicki MA, Braun T, Hinuma S, Jaenisch R. 1992. Inactivation of MyoD in mice leads to up-regulation of the myogenic HLH gene Myf-5 and results in apparently normal muscle development. Cell 71:383–390.

Shen X, De Jonge J, Forsberg SKG, Pettersson ME, Sheng Z, Hennig L, Carlborg Ö. 2014. Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. PLoS Genet. 10:e1004842.

Shore D. 1994. RAP1: A protean regulator in yeast. Trends Genet. 10:408–412.

Sussel L, Shore D. 1991. Separation of transcriptional activation and silencing functions of the RAP1-encoded repressor/activator protein 1: isolation of viable mutants affecting both silencing and telomere length. Proc. Natl. Acad. Sci. U. S. A. 88:7749–7753.

Venters BJ, Wachi S, Mavrich TN, Andersen BE, Jena P, Sinnamon AJ, Jain P, Rolleri NS, Jiang C, Hemeryck-Walsh C, et al. 2011. A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces. Mol. Cell 41:480–492.

Wagner A. 2000. Robustness against mutations in genetic networks of yeast. Nat. Genet. 24:355–361.

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 J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298.

Figures

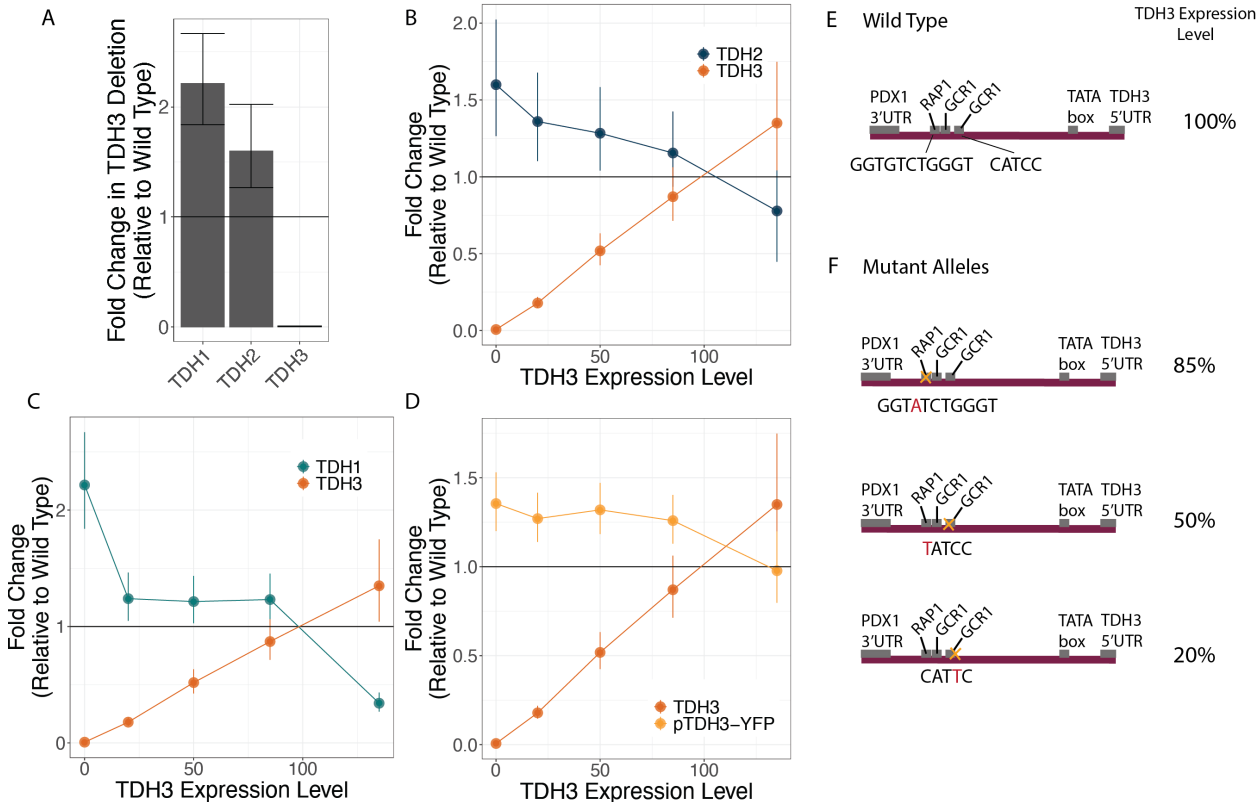

Figure 4-1: TDH1 and TDH2 actively compensate for changes in TDH3 expression

(A) Changes in expression of *TDH1*, *TDH2*, and *TDH3* in response to the deletion of *TDH3* are shown, measured as fold change in expression relative to a wild type. Error bars represent one standard error of the mean. Statistical significance of expression changes was assessed using Wald tests in DESeq2, with the P-value for *TDH1* = $2x10^{-5}$, *TDH2* = 0.04, and *TDH3* = $7x10^{-107}$. Changes in expression of *TDH3* and *TDH2* (B), *TDH1* (C), and a reporter gene with a wild type *TDH3* promoter driving expression of YFP (P_{TDH3}-*YFP*) (D) are shown for strains with *cis*-acting mutations causing 0%, 20%, 50%, 85%, and 135% of wild type *TDH3* expression. Error bars show one standard error of the mean. (F) A schematic of the wild type *TDH3* promoter is shown with the location and sequences of previously identified Rap1p and Gcr1p transcription factor binding sites indicated. (G) Schematics and sequences of the *TDH3* promoter in mutant strains bearing mutations in binding sites for Rap1p and Gcr1p that result in *TDH3* expression levels of 20%, 50%, and 85% relative to wild type are shown. No schematic is shown for the mutant strain expressing *TDH3* expression at 135% of wild type levels, which contains two copies of the *TDH3* gene separated by a copy of the *URA3* gene, with both copies of *TDH3* containing a mutation in the binding site for Rap1p (GGTGTCTGaGT).

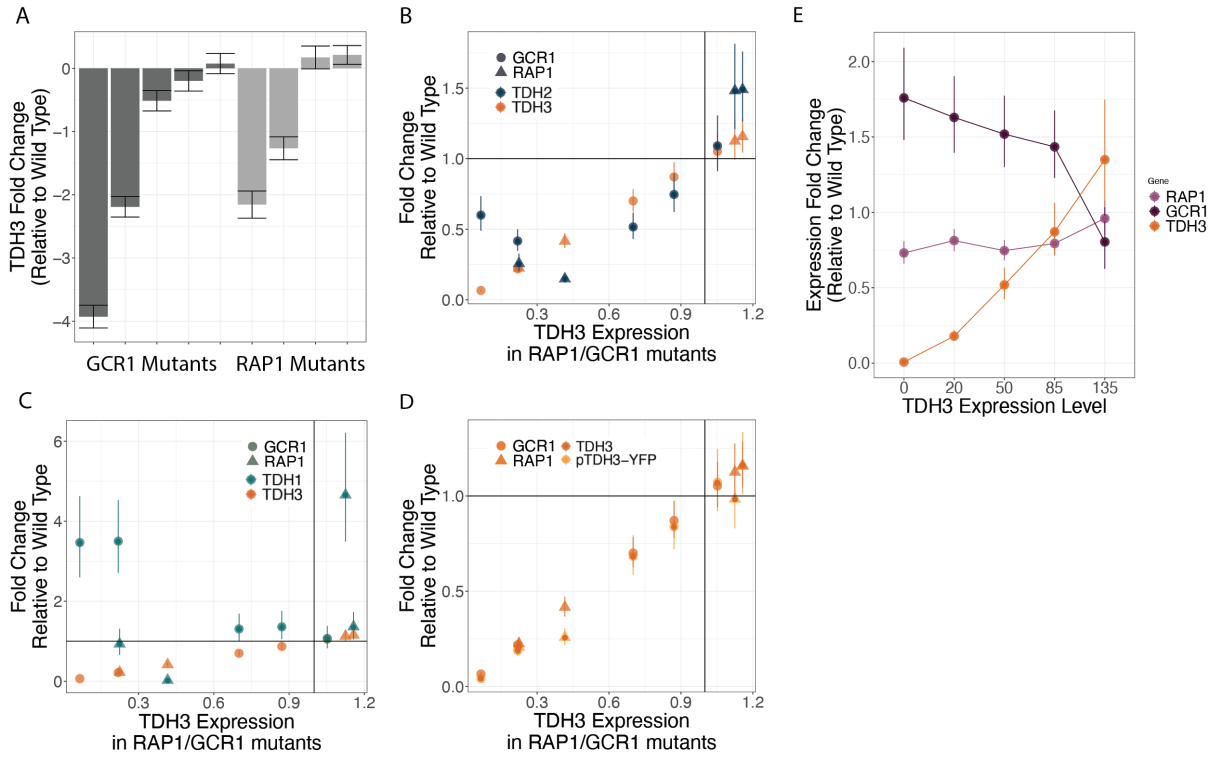

Figure 4-2: TDH2 is not upregulated when TDH3 expression is reduced by mutations in RAP1 or GCR1

(A) Changes in expression of *TDH1*, *TDH2*, and *TDH3* in response to various mutations in either GCR1 or RAP1, measured as fold change in expression relative to a wild type. Specific mutation identities are listed in Table 4. Error bars represent one standard error of the mean. Changes in expression of *TDH3* and *TDH2* (B), *TDH1* (C), and a reporter gene with a wild type *TDH3* promoter driving expression of YFP (P_{TDH3}-*YFP*) (D) are shown for strains with mutations in either RAP1 (triangles) or GCR1 (circles). Error bars show one standard error of the mean. (E) Changes in expression of *RAP1, GCR1*, and *TDH3* are shown for strains with *cis*-acting mutations causing 0%, 20%, 50%, 85%, and 135% of wild type *TDH3* expression, measured as fold change in expression relative to a wild type. Error bars represent one standard error of the mean.

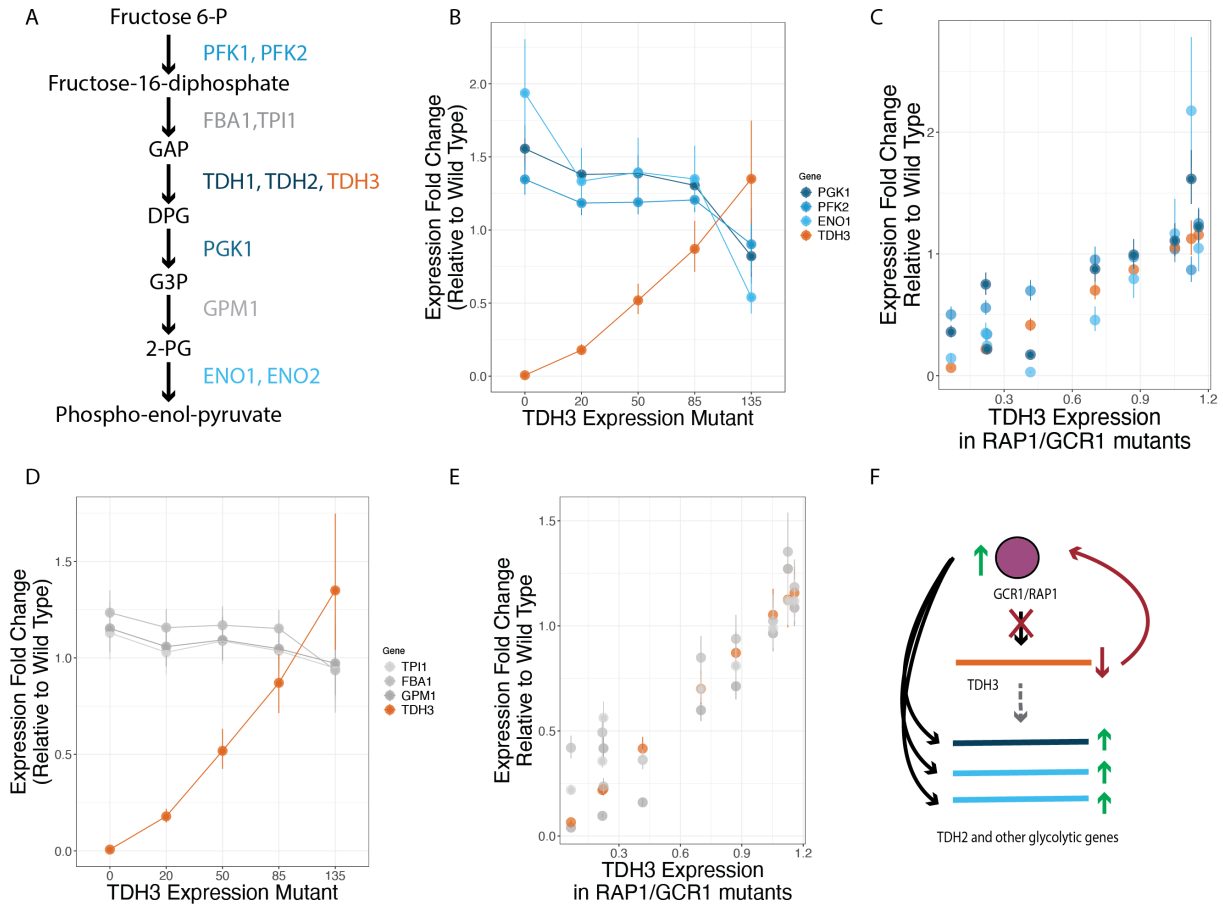

Figure 4-3: Multiple enzymes in the glycolysis pathway are upregulated upon reduction in TDH3 expression in a Rap1p/Gcr1p dependent manner

(A) A schematic of the glycolytic pathway surrounding the metabolic step catalyzed by *TDH1,2,* and *3*, showing other enzymes catalyzing adjacent reactions that are significantly upregulated upon reduction in *TDH3* in blue. Enzymes in this pathway that were not significantly upregulated are shown in grey. (B) Expression fold changes relative to wild type of the genes *PGK1, PFK2, ENO1,* and *TDH3* in yeast strains with varying levels of *TDH3* expression due to mutations in the native *TDH3* promoter, as estimated by RNA-sequencing data. Error bars are one standard error of the mean. (C) Expression fold changes relative to wild type of the genes *PGK1, PFK2, ENO1,* and *TDH3* in the 9 yeast strains with varying levels of *TDH3* expression due to mutations in the genes encoding *RAP1* or *GCR1*, as estimated by RNA-sequencing data. Error bars are one standard error of the mean. (D) Expression fold changes relative to wild type of the genes *TPI1, FBA1, GPM1,* and *TDH3* in yeast strains with varying levels of *TDH3* expression due to mutations in the native *TDH3* promoter, as estimated by RNA-sequencing data. Error bars are one standard error of the mean. (C) Expression fold changes relative to wild type of the genes *TPI1, FBA1, GPM1* and *TDH3* in 9 yeast strains with varying levels of *TDH3* expression due to mutations in the genes encoding *RAP1* or *GCR1*, as estimated by RNA-sequencing data. Error bars are one standard error of the mean. (F) A model for active compensation (grey dotted arrow) in which feedback from a reduction in *TDH3* expression (red

arrows) upregulates GCR1p/RAP1p complex (purple circle) which leads to upregulation of *TDH2* and other glycolytic genes regulated by GCR1/RAP1 (black arrows).

Table 4: TDH3 titration and Rap1p/Gcr1p mutant identities and effect on TDH3 expression

Collection	Strain name	Position	Reference Nucleotide	Resulting Nucleotide	Mutation Type	*TDH3* Expression (Relative to Wild Type)
YPW1177	tdh3-delta	deletion of locu	NA	NA	deletion	-7.218443
YPW1156	20% expression st	-482	C	T	promoter	-2.483259
YPW1200	50% expression st	-485	C	T	promoter	-0.9476816
YPW1188	85% expression st	-510	G	A	promoter	-0.199752
YPW3059	135% expression s	duplication, -50	G	A	duplication, promoter	0.4325633
YPW3282	GCR1162	833, 1112, 1946, 2305, 2755	del, T, G, A, T	T, C, A, G, C	frameshift, nonsynonymous, nonsynonymous, synonymous, synonymous	-3.927013598
YPW3283	GCR1281	940, 1224, 2178, 2599	T, T, T, T	C, C, C, C	synonymous, nonsynonymous, nonsynonymous, synonymous	-0.513344386
YPW3284	GCR1037	737, 1183, 1224, 1258, 2038, 3079,	G, T, T, T, A, T	C, C, C, C, G, del	intron, silent, nonsynonymous, silent, silent, frameshift	-0.198930953
YPW3285	GCR1339	726, 737, 740, 840, 2574	T, G, C, A, A	G, C, G, G, G	intron, intron, intron, nonsynonymous, nonsynonymous	-2.191062688
YPW3286	GCR1241	1366	A	G	synonymous	0.07367076
YPW3287	RAP1357	2378, 1881, 284	A, G, T	G, A, C	silent, nonsynonymous, silent	0.210454342
YPW3288	RAP154	upstream922, 1100, 2042,2043	C, T, A, T	T, C, del, del,	promoter, nonsynonymous, frameshift nonsense	-1.265524056
YPW3289	RAP1238	95, 568, 693, 1365, 1338, 2121	T, A, A, G, A, A	C, G, G, A, G, G	silent, all other nonsynonymous	-2.15553115
YPW3290	RAP1484	upstream85, 2128, 1060	A, A, A	G, G, G	noncoding, nonsynonymous, nonsynonymous	0.169982976

Chapter 5 Conclusions and Future Directions

Fitness has been at the center of evolutionary biology from the beginning of the field (Darwin et al. 1859). But what determines how a mutation will impact fitness? Is there any way to predict which mutations will be advantageous and which will be deleterious? These questions are at the center of research in evolutionary biology today (de Visser and Krug 2014; Das et al. 2020; Zheng et al. 2020). The work in this dissertation represents one small step forward in the journey toward answering them.

The interconnected nature of organisms results in mutations frequently influencing not just one trait, but several (Stearns 2010; Wagner and Zhang 2011; Boyle et al. 2017; Mehlhoff et al. 2020). The extent of these pleiotropic effects are expected to increase the probability that a mutation will be deleterious for the organism as a whole (Fisher 1930), potentially imposing a strong constraint on the evolution of complexity (Orr 2000). While attempts have been made to describe the extent of mutational pleiotropy (Featherstone and Broadie 2002; Dudley et al. 2005; Cooper et al. 2007; McGuigan et al. 2014; Kinsler et al. 2020), relating pleiotropy to fitness has remained difficult due to the imprecision of defining distinct traits and quantifying them (Paaby and Rockman 2013a; Paaby and Rockman 2013b; Zhang and Wagner 2013).

Mutations that influence gene expression are an ideal system in which to explore the relative pleiotropy of different types of mutations because the traits upon which they act –

expression of all genes across the genome – represent discreet units that can be precisely quantified. In addition, gene expression can be molecularly manipulated and analyzed at larger systems-level scales, enabling the molecular dissection of expression changes genome-wide that are due to changes in one particular gene's expression level. In addition, pleiotropy is hypothesized to differ between mutations occurring proximally, or in *cis* to a particular gene, and those occurring distally, or in *trans* (Wittkopp 2005; Wray 2007; Signor and Nuzhdin 2018; Hill et al. 2021). These potential differences in the pleiotropic effects of *cis*- and *trans*-regulatory mutations may result in *cis*-regulatory mutations being preferentially fixed relative to *trans*-regulatory mutations, leading to an increasing proportion of expression divergence due to *cis*-regulatory changes over evolutionary time (Emerson et al. 2010; Coolon et al. 2014; Metzger et al. 2017). The work presented in this dissertation is the first exploration of this hypothesis utilizing empirical data.

Here I utilize both molecular and systems level approaches to provide a quantitative description of the pleiotropic effects of regulatory mutations and their relationship to fitness. In the second chapter, a detailed analysis of *cis*- and *trans*-regulatory mutations to the gene *TDH3* in *Saccharomyces cerevisiae* enables the calculation of a distribution of pleiotropic fitness effects of *trans*-regulatory mutations. In addition, I compare the extents the impacts of *cis*- and *trans*-regulatory mutation on gene expression across the genome and demonstrate that the pleiotropic effects of *trans*-regulatory mutations affect genes that are downstream of the focal gene in addition to those outside of it. Going forward, similar studies should be conducted using other focal genes and biological systems to reveal how generalizable these trends are and to define more distributions of pleiotropic fitness effects that can be used to model the evolution of gene

expression while incorporating pleiotropic fitness costs. The third chapter fully embraces systems level analysis by demonstrating that the degree distribution of a network constructed from the effects of deletion mutants can explain the pattern of *trans*-regulatory mutations being on average more pleiotropic than *cis*-regulatory mutations to the same gene. This work can also be extended into other organisms as similar datasets are produced. Computationally, network simulations can demonstrate the range of degree distributions that will result in this pattern. Also, the degree to which accumulation of expression divergence due to *cis*-regulatory mutations can be accounted for by network topology as compared to the contributions of other factors, such as modularity of regulatory sequences, should be further explored. In the fourth chapter, I return to the *TDH3* system to describe how compensation for loss of *TDH3* by its paralogs can help to explain the molecular mechanism behind the pleiotropic effects of *trans*-regulatory mutations on expression of genes downstream of *TDH3*. Whether this type of compensation occurs for many paralogs in the *S. cerevisiae* genome is debatable (Kafri et al. 2005; He and Zhang 2006; Kafri et al. 2006), and can begin to be elucidated as datasets for other genes that have partially redundant paralogs are gathered.

Pleiotropic effects of *trans*-regulatory mutations relative to *cis*-regulatory mutations to the focal gene *TDH3*

It has been hypothesized that *trans*-regulatory mutations should influence the expression of more genes, and therefore be more pleiotropic, than *cis*-regulatory mutations based on their positions relative to each other in the regulatory network (Wittkopp 2005). This difference could explain the increase in expression divergence attributable to *cis*-regulatory variation observed over increasing evolutionary distances in both flies and yeast (Coolon et al. 2014; Metzger et al. 2017). Despite the intuitive nature of this model, it had not previously been tested with empirical

data (Signor and Nuzhdin 2018; Hill et al. 2021). My use of strains of *S. cerevisiae* containing one mutation that is either *cis*-regulatory or *trans*-regulatory to the yeast gene *TDH3* enables me to directly test this question by separating the effects of changing the focal gene from those that are pleiotropic to it occurring in *trans*-regulatory mutants. The separation the fitness effects of changing the expression of the focal gene from the pleiotropic fitness effects in *trans*-regulatory mutations creates a framework for quantifying the fitness effects of pleiotropy which can be included in models of gene expression evolution. This will enable more realistic predictions and descriptions of what classes of mutations are likely to generate population level and species level variation in gene regulation (Hill et al. 2021).

The findings presented here provide empirical support for theoretical expectations, showing that the pleiotropic fitness effects are mostly negative and that *trans*-regulatory mutations have more widespread effects than *cis*-regulatory mutations, but with added nuance of being dependent on the effect size of both classes of mutations on expression of the focal gene. While we suspect that these findings will be generalizable for many genes beyond *TDH3* (see Chapter 3), more work should be done to collect similar datasets for other focal genes to empirically test this. The ability to collect RNA-sequencing data for a larger number of genotypes (Jackson et al. 2020), and measure the genomic impact of titrated gene expression levels (Jost et al. 2020) using single cell sequencing and CRISPR is making such an undertaking more feasible than ever before.

Additionally, the work presented here raises the interesting point that *cis*- and *trans*-regulatory mutations have different effects on genes downstream of a change in the focal gene

itself in a way that is somewhat akin to epistasis (Domingo et al. 2019). If one conceptualizes the change in expression of the focal gene as a mutation itself, and the downstream consequences of that change as the phenotype, *trans*-regulatory mutations essentially have epistatic effects on those phenotypes. These epistatic effects heavily influence the phenotype that results from a change in the focal gene, particularly if there are multiple variants present. This is reminiscent of the 'omnigenic' model of heritability (Boyle et al. 2017), in which the highly connected nature of genetic networks results in many indirect effects of variants on complex traits, all of which influence whether or not a phenotype will result from a variant of major effect (Liu et al. 2019). It could be interesting to explore how these pleiotropic effects of *trans*-regulatory mutations influence overall phenotypes when there are compensatory mutations that occur in *cis*, which may restore expression of the focal gene but potentially not have the same downstream consequences (Goncalves et al. 2012; Coolon et al. 2014).

Network topology can explain differences in *cis* and *trans* pleiotropy and fitness

This finding that the effects of *cis*-regulatory mutations are not 'encapsulated' within the effects of *trans*-regulatory mutations raises the question of what then might explain the higher pleiotropy of *trans*-regulatory mutations relative to *cis*-regulatory mutations. The third chapter of this thesis takes this question on from a systems biology perspective, taking the structure of the regulatory network into account rather than describing the specific effects on one gene. Although gene expression data is not available for *cis*- and *trans*-regulatory mutations to many different focal genes (Hill et al. 2021), the effects of *cis*-regulatory mutations can be approximated by the effects of a deletion of the focal gene itself, with the caveat that they are likely the most severe effects of changing the expression of a gene in *cis* (Keren et al. 2016). Likewise, effects of *trans*-regulatory mutations can be approximated by the effects of gene deletions that cause

misexpression of the focal gene (Landry et al. 2007). Using these approximations to estimate the degree of pleiotropy for *cis*- and *trans*-regulatory mutations to ~1200 focal genes (Kemmeren et al. 2014), I show that for the vast majority of genes assayed, *trans*-regulatory mutations are more pleiotropic than *cis*-regulatory mutations, even when *cis*-regulatory mutations are of a greater effect size on the focal gene than *trans*-regulatory mutations. This pattern is not dependent on specific connections between regulators and target genes. Rather, it is the result of a sampling bias for highly pleiotropic *trans*-regulators that serve as *trans*-regulatory mutations for a large number of focal genes- a phenomenon that has been described in social networks with similar topologies and termed the 'friendship paradox' (Feld 1991).

Although the scale-free topology of many biological networks has raised much interest in applying network theory to biological systems (Barabási and Oltvai 2004; Barzel and Barabási 2013), there is also evidence that many empirical networks may not be strictly scale-free (Broido and Clauset 2019), and many questions remain about whether findings from network science will be informative for biological systems, especially at evolutionary timescales (Siegal et al. 2007). The work presented here is an interesting case in which the empirical data led to an explanation from network science rather than network science being applied to a biological system, and therefore suggests that network science can indeed inform biological systems.

Another interesting implication of the findings presented in chapter three is that the difference in pleiotropy between *cis* and *trans*-regulatory mutations could be independent of differences in the modularity (such as environmental or tissue specificity) of the elements mutated. This could explain why we see patterns of increasing proportions of expression

divergence due to *cis*-regulatory mutations in species without extensive, modular regulatory sequences such as yeast (Metzger et al. 2017). One possible consequence of this finding is that the accumulation of *cis*-regulatory divergence would occur for all genes, rather than just those with more modular regulatory elements (Wittkopp 2005) or for genes with roles in morphology as opposed to physiology (Carroll 2005; Stern and Orgogozo 2008). A connection between the modularity of regulatory elements and the accumulation of *cis*-regulatory divergence should be further explored to test whether or not this is the case. Any test of this hypothesis, however, must be careful to also take in to account alternative, non-mutually exclusive differences between *cis*- and *trans*-regulatory mutations such as effect size and dominance that may also influence their relative frequency of fixation (Gruber et al. 2012).

The presence of a few very pleiotropic *trans*-regulators as 'network hubs' as discussed in chapter three suggests that the expression or sequence divergence of these hub genes may be more constrained than other genes in the genome. Past attempts to draw a connection between divergence and network hubs has only been moderately successful, however (Featherstone and Broadie 2002; Siegal et al. 2007; Costanzo et al. 2010; Kopp and McIntyre 2012; Yang and Wittkopp 2017; Flint and Ideker 2019; Wollenberg Valero 2020). This could be partially due to the fact that different *trans*-regulators serve as highly connected 'hubs' in different environments or genetic backgrounds (Luscombe et al. 2004), and that different genes can serve as 'hubs' in different cellular networks. More work should be done to assess the connectivity of different regulators in different environments to test whether there is a connection between the number of conditions under which a *trans*-regulator serves as a hub gene and its evolutionary divergence in expression or sequence.

Compensatory upregulation of paralogs in cis-regulatory mutants

In the fourth chapter, I return to the *TDH3* system to characterize a phenomenon that can help explain the differences in downstream effects of *cis*- and *trans*-regulatory mutations to the same focal gene. Specifically, I test the hypothesis that the two paralogs of *TDH3*, *TDH1* and *TDH2*, actively compensate for loss or reduction of *TDH3*. In the case of *TDH2,* an upregulation of *TDH2* expression appears to be the result of feedback mechanisms that control the expression levels of all three paralogs involving the transcription factors Rap1p and Gcr1p. Therefore, upregulation of *TDH2* does not occur when *TDH3* expression is lowered by mutations in these Rap1p or Gcr1p themselves. Whether this type of compensation by co-regulated paralogs is common across the genome is an open question in the field (Kafri et al. 2005; He and Zhang 2006; Kafri et al. 2006; Li et al. 2010; Diss et al. 2014; Kuzmin et al. 2020; Kovács et al. 2021). Once again, collecting similar information for other paralogs can speak to the generality of these findings across the genome.

These findings also describe one reason why the downstream effects of lowering *TDH3* expression differ when a mutation occurs in *cis* or in *trans*. Mutations that induce 'breaks' in the regulatory network at different points relative to the focal gene will determine what feedback mechanisms are still intact and how they will be employed, and whether or not they will be advantageous (Kovács et al. 2021). Examining the effects of mutations that titrate gene expression to varying degrees at different points in the regulatory network allow us to examine how the network is rewired when specific regulatory connections are broken. An extension of this work could include a larger, more systematic collection of genome-wide gene expression data upon titration of the expression levels of many focal genes, which may now be possible with

single cell sequencing and methods for titrating gene expression levels using targeted interference with CRISPR (Jackson et al. 2020; Jost et al. 2020; Urbonaite et al. 2021).

Conclusion

Here I focus on understanding whether and why *cis*- and *trans*-regulatory mutations to the same focal gene have different pleiotropic effects on expression and fitness and therefore are likely to be acted on differently by natural selection. This work advances our understanding how pleiotropy may influence the evolution of gene expression and is also an example of the information exchange between molecular and systems biology that is necessary to address the questions of how mutations influence phenotype. The work in this dissertation makes another small step forward in answering questions of how genotypes are related to phenotypes and ultimately to fitness, which continue to be a major focus of evolutionary and molecular biology.

References

Barabási A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5:101–113.

Barzel B, Barabási AL. 2013. Universality in network dynamics. *Nat. Phys.* 9:673–681.

Boyle EA, Li YI, Pritchard JK. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 169:1177–1186.

Broido AD, Clauset A. 2019. Scale-free networks are rare. *Nat. Commun.* 10:1017.

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol.* 3:e245.

Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. 2014. Tempo and mode of regulatory evolution in Drosophila. *Genome Res.* 24:797–808.

Cooper TF, Ostrowski EA, Travisano M. 2007. A negative relationship between mutation pleiotropy and fitness effect in yeast. *Evolution* 61:1495–1499.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, et al. 2010. The Genetic Landscape of a Cell. *Science* 327.

Darwin, Charles, 1809-. 1859. On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. London : John Murray, 1859

Das SG, Direito SO, Waclaw B, Allen RJ, Krug J. 2020. Predictable properties of fitness landscapes induced by adaptational tradeoffs. *Elife* [Internet] 9. Available from: http://dx.doi.org/10.7554/eLife.55155

Diss G, Ascencio D, DeLuna A, Landry CR. 2014. Molecular mechanisms of paralogous compensation and the robustness of cellular networks. *J. Exp. Zool. B Mol. Dev. Evol.* 322:488–499.

Domingo J, Baeza-Centurion P, Lehner B. 2019. The Causes and Consequences of Genetic Interactions (Epistasis). *Annu. Rev. Genomics Hum. Genet.* 20:433–460.

Dudley AM, Janse DM, Tanay A, Shamir R, Church GM. 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol. Syst. Biol.* 1:2005.0001.

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

Featherstone DE, Broadie K. 2002. Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. *Bioessays* 24:267–274.

Feld SL. 1991. Why Your Friends Have More Friends Than You Do. *Am. J. Sociol.* 96:1464–1477.

Fisher RA. 1930. The genetical theory of natural selection. 272. Available from: https://psycnet.apa.org/fulltext/1930-04698-000.pdf

Flint J, Ideker T. 2019. The great hairball gambit.Copenhaver GP, editor. *PLoS Genet.* 15:e1008519.

Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, Flicek P, Brazma A, Odom DT, Marioni JC. 2012. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res.* 22:2376–2384.

Gruber JD, Vogel K, Kalay G, Wittkopp PJ. 2012. Contrasting Properties of Gene-Specific Regulatory , Coding , and Copy Number Mutations in Saccharomyces cerevisiae : Frequency , Effects , and Dominance. 8. Available from: http://dx.doi.org/10.1371/journal.pgen.1002497

He X, Zhang J. 2006. Transcriptional reprogramming and backup between duplicate genes: is it a genomewide phenomenon? *Genetics* 172:1363–1367.

Hill MS, Vande Zande P, Wittkopp PJ. 2021. Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* 22:203–215.

Jackson CA, Castro DM, Saldi G-A, Bonneau R, Gresham D. 2020. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *Elife* [Internet] 9. Available from: http://dx.doi.org/10.7554/eLife.51254

Jost M, Santos DA, Saunders RA, Horlbeck MA, Hawkins JS, Scaria SM, Norman TM, Hussmann JA, Liem CR, Gross CA, et al. 2020. Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. *Nat. Biotechnol.* 38:355–364.

Kafri R, Bar-Even A, Pilpel Y. 2005. Transcription control reprogramming in genetic backup circuits. *Nat. Genet.* 37:295–299.

Kafri R, Levy M, Pilpel Y. 2006. The regulatory utilization of genetic redundancy through responsive backup circuits. *Proc. Natl. Acad. Sci. U. S. A.* 103:11653–11658.

Kemmeren P, Sameith K, van de Pasch LAL, Benschop JJ, Lenstra TL, Margaritis T, O'Duibhir E, Apweiler E, van Wageningen S, Ko CW, et al. 2014. Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. *Cell* 157:740–752.

Keren L, Hausser J, Lotan-Pompan M, Vainberg Slutskin I, Alisar H, Kaminski S, Weinberger A, Alon Uri, Milo R, Segal Eran, et al. 2016. Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell* 0:636–643.

Kinsler G, Geiler-Samerotte K, Petrov DA. 2020. Fitness variation across subtle environmental perturbations reveals local modularity and global pleiotropy of adaptation. *Elife* [Internet] 9. Available from: http://dx.doi.org/10.7554/eLife.61271

Kopp A, McIntyre LM. 2012. Transcriptional network structure has little effect on the rate of regulatory evolution in yeast. *Mol. Biol. Evol.* 29:1899–1905.

Kovács K, Farkas Z, Bajić D, Kalapis D, Daraba A, Almási K, Kintses B, Bódi Z, Notebaart RA, Poyatos JF, et al. 2021. Suboptimal Global Transcriptional Response Increases the Harmful Effects of Loss-of-Function Mutations. *Mol. Biol. Evol.* 38:1137–1150.

Kuzmin E, VanderSluis B, Nguyen Ba AN, Wang W, Koch EN, Usaj M, Khmelinskii A, Usaj MM, van Leeuwen J, Kraus O, et al. 2020. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science* [Internet] 368. Available from: http://dx.doi.org/10.1126/science.aaz5667

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic properties influencing the evolvability of gene expression. *Science* 317:118–121.

Li J, Yuan Z, Zhang Z. 2010. The cellular robustness by genetic redundancy in budding yeast. *PLoS Genet.* 6:e1001187.

Liu X, Li YI, Pritchard JK. 2019. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* 177:1022-1034.e6.

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431:308–312.

McGuigan K, Collet JM, McGraw EA, Ye YH, Allen SL, Chenoweth SF, Blows MW. 2014. The Nature and Extent of Mutational Pleiotropy in Gene Expression of Male Drosophila serrata. *Genetics* 196:911–921.

Mehlhoff JD, Stearns FW, Rohm D, Wang B, Tsou E-Y, Dutta N, Hsiao M-H, Gonzalez CE, Rubin AF, Ostermeier M. 2020. Collateral fitness effects of mutations. *Proc. Natl. Acad. Sci. U. S. A.* [Internet]. Available from: http://dx.doi.org/10.1073/pnas.1918680117

Metzger BPH, Wittkopp PJ, Coolon JD. 2017. Evolutionary dynamics of regulatory changes underlying gene expression divergence among Saccharomyces species. *Genome Biol. Evol.* 177:1987–1996.

Orr HA. 2000. Adaptation and the cost of complexity. *Evolution* 54:13–20.

Paaby AB, Rockman MV. 2013a. The many faces of pleiotropy. *Trends Genet.* 29:66–73.

Paaby AB, Rockman MV. 2013b. Pleiotropy: what do you mean? Reply to Zhang and Wagner. *Trends Genet.* 29:384.

Siegal ML, Promislow DEL, Bergman A. 2007. Functional and evolutionary inference in gene networks: does topology matter? *Genetica* 129:83–103.

Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. *Trends Genet.* 34:532–544.

Stearns FW. 2010. One Hundred Years of Pleiotropy: A Retrospective. *Genetics* 186.

Stern DL, Orgogozo V. 2008. The loci of evolution: How predictable is genetic evoltuion? :2155–2177.

Urbonaite G, Lee JTH, Liu P, Parada GE, Hemberg M, Acar M. 2021. A yeast-optimized single-cell transcriptomics platform elucidates how mycophenolic acid and guanine alter global mRNA levels. *Commun Biol* 4:822.

de Visser JAGM, Krug J. 2014. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* 15:480–490.

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* 12:204–213.

Wittkopp PJ. 2005. Genomic sources of regulatory variation in cis and in trans. *Cell. Mol. Life Sci.* 62:1779–1783.

Wollenberg Valero KC. 2020. Aligning functional network constraint to evolutionary outcomes. *BMC Evol. Biol.* 20:58.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8:206–216.

Yang B, Wittkopp PJ. 2017. Structure of the Transcriptional Regulatory Network Correlates with Regulatory Divergence in Drosophila. *Mol. Biol. Evol.* 34:1352–1362.

Zhang J, Wagner GP. 2013. On the definition and measurement of pleiotropy. *Trends Genet.* 29:383–384.

Zheng J, Guo N, Wagner A. 2020. Selection enhances protein evolvability by increasing mutational robustness and foldability. *Science* 370. Available from: http://dx.doi.org/10.1126/science.abb5962

Appendix: Mutational Sources of *Trans*-regulatory Variation Affecting Gene Expression in *Saccharomyces cerevisiae* [2]

Abstract

Heritable variation in a gene's expression arises from mutations impacting *cis*- and *trans*-acting components of its regulatory network. Here, we investigate how *trans*-regulatory mutations are distributed within the genome and within a gene regulatory network by identifying and characterizing 69 mutations with *trans*-regulatory effects on expression of the same focal gene in *Saccharomyces cerevisiae*. Relative to 1766 mutations without effects on expression of this focal gene, we found that these *trans*-regulatory mutations were enriched in coding sequences of transcription factors previously predicted to regulate expression of the focal gene. However, over 90% of the *trans*-regulatory mutations identified mapped to other types of genes involved in diverse biological processes including chromatin state, metabolism and signal transduction. These data show how genetic changes in diverse types of genes can impact a gene's expression in *trans*, revealing properties of *trans*-regulatory mutations that provide the raw material for *trans*-regulatory variation segregating within natural populations.

[2] This appendix is published as: Duveau F, Vande Zande P, Metzger BP, Diaz CJ, Walker EA, Tryban S, Siddiq MA, Yang B, Wittkopp PJ. 2021. Mutational sources of trans-regulatory variation affecting gene expression in Saccharomyces cerevisiae. Elife. Available from: http://dx.doi.org/10.7554/eLife.67806. Specific contributions by PVZ include the generation of all strains bearing mutations in RAP1 and assaying their effects on TDH3 expression.

Introduction

The regulation of gene expression is a complex process, essential for cellular function, that impacts development, physiology, and evolution. Expression of each gene is regulated by its *cis*-regulatory DNA sequences (e.g., promoters, enhancers) interacting either directly or indirectly with *trans*-acting factors (e.g. transcription factors, signaling pathways) encoded by genes throughout the genome. Genetic variants affecting both *cis*- and *trans*-acting components of regulatory networks contribute to expression differences within and between species (Albert & Kruglyak, 2015; Barbeira et al., 2018; Ferraro et al., 2020; Gamazon et al., 2018; Oliver et al., 2005). This regulatory variation arises the same way as genetic variation affecting any other quantitative trait: new mutations generate variation in gene expression and selection favors the transmission of some genetic variants over others, giving rise to polymorphism within a species and divergence between species. Because new mutations are the raw material for this polymorphism and divergence, knowing how new mutations impact gene expression is essential for understanding how gene regulation evolves (reviewed in Hill et al., 2020). Targeted mutagenesis has been used to systematically examine the effects of individual mutations in *cis*-regulatory sequences for a variety of elements in a variety of species (Hornung et al., 2012; Kwasnieski et al., 2012; Maricque et al., 2017; Melnikov et al., 2012; Metzger et al., 2015; Patwardhan et al., 2009; Sharon et al., 2012), but such targeted approaches are not well-suited for surveying the effects of new *trans*-regulatory mutations because *trans*-regulatory mutations can be located virtually anywhere within the genome. Consequently, we know comparatively little about the genomic sources, molecular mechanisms of action and evolutionary contributions of individual *trans*-regulatory mutations.

Genetic mapping experiments and genome-wide association studies (GWAS) have shown that gene expression is a highly polygenic trait, with hundreds of genetic variants typically associated with natural variation in expression levels of each gene (Albert et al., 2018; Metzger & Wittkopp, 2019; Sinnott-Armstrong et al., 2021). Although these studies often lack the resolution to identify individual genetic changes affecting expression, most of this variation maps far from the gene whose expression it affects and is therefore likely to have *trans*-acting effects. *Trans*-acting variants segregating in natural populations are most often expected to affect transcription factors (Albert et al., 2018; Lewis et al., 2014), but they can also alter genes encoding signaling proteins, chromatin modifiers, metabolic enzymes, or any other gene product that can influence the availability, accessibility, or activity of transcription factors (Lutz et al., 2019; Mehrabian et al., 2005; Schadt et al., 2005; Yvert et al., 2003). Indeed, the recently proposed omnigenic model emphasizes the interconnectedness of regulatory networks controlling transcription to help explain the highly polygenic nature of diverse quantitative traits.

Despite the vast potential target size for *trans*-regulatory mutations (Hill et al., 2020), regions of the genome most likely to harbor mutations affecting a particular gene's expression might be predictable from knowledge of its regulatory network. Among eukaryotes, the set of genes and interactions regulating gene expression in *trans* is perhaps best understood in the baker's yeast *Saccharomyces cerevisiae* (Hughes & Boer, 2013): networks of regulatory connections (Teixeira et al., 2018) have been inferred from experiments that profile the transcriptional effects of gene deletions (Hughes et al., 2000; Jackson et al., 2020; Kemmeren et al., 2014), map binding sites for transcription factors (Rhee & Pugh, 2011; Zheng et al., 2010; Zhu et al., 2009), identify protein-protein interactions (Gavin et al., 2002; Liu et al., 2020;

Tarassov et al., 2008), and test pairs of genes for genetic interactions (Costanzo et al., 2016; Leeuwen et al., 2016). However, the extent to which the genomic sources of *trans*-regulatory mutations can be predicted from such networks is generally unknown (Flint & Ideker, 2019). In addition, the extent to which the genomic distribution of new mutations predicts the genomic distribution of natural polymorphisms is also unclear because mutations that are strongly deleterious might rarely be found circulating within a population as standing genetic variation. For example, mutations in coding sequences might often impact gene expression but might also tend to be more pleiotropic and thus more deleterious than mutations in non-coding regions of these genes. Comparing the genomic distribution of mutations that have not experienced natural selection to the genomic distribution of polymorphisms that have can reveal such differences between the possible and actual sources of variation in gene expression in the wild.

Systematic studies of new mutations identifying and characterizing the effects of individual genetic changes are thus an important complement to GWAS describing the polygenic variation segregating within a species. Recently, a chemical mutagen was used to induce mutations throughout the genome of *S. cerevisiae,* and hundreds of mutant genotypes were collected that all altered expression of the same gene, providing the biological resources needed to systematically characterize properties of new *trans*-regulatory mutations and to test the predictive power of inferred regulatory networks (Gruber et al., 2012; Metzger et al., 2016). Here, we use genetic mapping, candidate gene sequencing and functional validation to identify 69 *trans*-regulatory mutations that alter expression of the focal gene from this set of mutants and contrast their properties with a comparable set of 1766 mutations that did not affect expression of the focal gene.

Using this collection of individual *trans*-regulatory mutations, we determined how *trans*-regulatory mutations affecting expression of a single gene were distributed within the genome and within a regulatory network. For example, we asked how frequently *trans*-regulatory mutations were located in coding or non-coding sequences because *trans*-regulatory variants are often predicted to affect coding sequences (Hill et al., 2020) but some non-coding variants have been shown to be associated with *trans*-regulatory effects on gene expression (Consortium, 2020; Yao et al., 2017; Yvert et al., 2003). We also asked whether genes encoding transcription factors were the primary source of *trans*-regulatory variation, which is often assumed (Albert et al., 2018; Lewis et al., 2014) despite case studies identifying *trans*-regulatory variants in genes encoding proteins with other functions (Lutz et al., 2019; Mehrabian et al., 2005; Schadt et al., 2005; Yvert et al., 2003). To determine how well an inferred regulatory network can predict genomic sources of expression changes, we mapped the *trans*-regulatory mutations to a network of transcription factors predicted by functional genomic data to regulate expression of the focal gene and examined the molecular functions and biological processes impacted by *trans*-regulatory mutations that did not map to genes in this network. By systematically examining the properties and identity of new *trans*-regulatory mutations, this work fills a key gap in our understanding of how expression differences arise and may help predict sources of *trans*-regulatory variation segregating in natural populations. Indeed, we found that the genomic distribution of new *trans*-regulatory mutations overlaps significantly with the genomic distribution of *trans*-regulatory variants segregating among wild isolates of *S. cerevisiae* that affect expression of the same gene (Metzger & Wittkopp, 2019), suggesting that the mutational process generating new *trans*-regulatory variation significantly shaped the regulatory variation we see in the wild.

Results

Genetic mapping of trans-regulatory mutations

To characterize properties of new *trans*-regulatory mutations affecting expression of a focal gene, we took advantage of three previously collected sets of haploid mutants that all showed altered expression of the same reporter gene (Figure A-1A, Gruber et al., 2012; Metzger et al., 2016). This reporter gene (P_{TDH3}-*YFP*) encodes a yellow fluorescent protein whose expression is regulated by the *S. cerevisiae TDH3* promoter, which natively drives constitutive expression of a glyceraldehyde-3-phosphate dehydrogenase involved in glycolysis and gluconeogenesis (McAlister & Holland, 1985). Mutations in these mutants were caused by exposure to the chemical mutagen ethyl methanesulfonate (EMS), which induces primarily G:C to A:T point mutations randomly throughout the genome (Shiwa et al., 2012). The dose of EMS used in these studies was chosen so that most mutants with a detectable change in P_{TDH3}-*YFP* expression should have only one mutation causing this change in expression among the mutations they carry (Metzger et al., 2016; Gruber et al., 2012). Together, these collections contain ~1500 mutants isolated irrespective of their fluorescence levels ("unenriched" mutants) and ~1200 mutants isolated after enriching for cells with the largest changes in fluorescence (Figure A-1A, see Figure A-2 for a diagram showing the number of mutants and mutations included at each step of the study). When we started this work, expression level of P_{TDH3}-*YFP* in these mutant genotypes had been described (Gruber et al., 2012; Metzger et al., 2016), but the specific mutations present within each mutant as well as which mutation(s) alter(s) P_{TDH3}-*YFP* expression in each genotype were unknown.

From these collections, we selected 82 EMS-treated mutants for genetic mapping to identify individual causal mutations (Figure A-1A, Figure A-2). Sanger sequencing of the

reporter gene in these mutants showed that none had mutations in the *TDH3* promoter or any other part of the reporter gene, indicating that they harbored mutations affecting P_{TDH3}-*YFP* expression in *trans*. 39 of these mutants were selected based on previously published fluorescence data, with 11 mutants selected from the collections enriched for large effects (red points in Figure A-1B,C) and 28 mutants selected from the unenriched collection (red points in Figure A-1D). Each selected mutant showed changes in average YFP fluorescence greater than 1% relative to the un-mutagenized progenitor strain. Another 197 mutants from the unenriched collection (blue points in Figure A-1D) were subjected to a secondary fluorescence screen, from which an additional 43 mutants with a change in fluorescence greater than 1% (red points in Figure A-1E) were chosen. Overall, the 82 mutants were selected randomly from the 528 EMS mutants that showed statistically significant fluorescence changes greater than 1% relative to wild-type ($P < 0.05$, see Methods and Figure A-2 legend for a description of the statistical tests). A 1% change in YFP fluorescence has previously been shown to correspond to a ~3% change in YFP mRNA abundance (see Methods and Duveau et al., 2018), although changes in fluorescence caused by *trans*-regulatory mutations in these mutants could affect either transcription driven by the *TDH3* promoter or post-transcriptional regulation of YFP synthesis or stability.

To identify mutations within the 82 selected EMS mutants, and to determine which of these mutation(s) were most likely to affect YFP expression in each mutant, we performed bulk-segregant analysis followed by whole-genome sequencing (BSA-Seq) as described in Duveau et al. (2014) with minor modifications (see Methods). Briefly, each mutant strain was crossed to a common mapping strain expressing the P_{TDH3}-*YFP* reporter gene, and large populations of random haploid spores were isolated after inducing meiosis in the resulting diploids (Figure A-3A). For each of the 82 segregant populations, a low fluorescent bulk and a high fluorescent bulk

of ~1.5 x 10^5 cells each were isolated using fluorescence-activated cell sorting (FACS) (Figure A-3B). Genomic DNA extracted from each bulk was then sequenced to an average coverage of ~105x (ranging from 75x to 134x among samples, Supplementary File 1) to identify the mutations present within each mutant genotype and to quantify the frequency of mutant and non-mutant alleles in both bulks (Figure A-3C). A mutation causing a change in fluorescence is expected to be found at different frequencies in the two populations of segregant cells. Conversely, a mutation with no effect on fluorescence that is not genetically linked to a mutation affecting fluorescence is expected to be found at similar frequencies in these two populations.

Using a stringent approach for calling sequence variants (see Methods), we identified a total of 1819 mutations in the BSA-Seq data from the 76 mutants from Metzger et al. (2016) (Supplementary File 2, Figure A-2), among which 1768 mutations (97.2%) were single nucleotide changes (Figure A-3D). Of these single nucleotide changes, 96.3% were one of the two types of point mutations (G:C to A:T transitions) known to be primarily induced by EMS (Shiwa et al., 2012). 48 small indels and 3 aneuploidies, which could have arisen spontaneously or been introduced by EMS, were also identified. Of these 3 mutants with aneuploidies, 2 were found to have an extra copy of chromosome I and 1 was found to have an extra copy of chromosome V based on ~1.5-fold higher sequencing coverage of these chromosomes relative to the rest of the genome in the BSA-seq data from segregant populations (shown in Supplementary File 3). We identified an average of 23.9 mutations per strain, which is within the 95% confidence interval of 21 to 45 mutations per strain estimated previously from the frequency of canavanine resistant mutants (Metzger et al., 2016). Surprisingly, the number of mutations per strain did not follow a Poisson distribution: we observed more strains with a number of mutations far from the average than expected for a Poisson process (P-value < 10^{-5}, resampling

test; Figure A-4), which could be explained by cell-to-cell heterogeneity in DNA repair after exposure to the mutagen (Liu et al., 2019; Uphoff et al., 2016).

At least one mutation was significantly associated with fluorescence in 46 of the mutants analyzed based on likelihood ratio tests (*G*-tests described in Methods, Supplementary File 2), with a total of 67 mutations associated with fluorescence identified among these mutants (Figure A-2), including all 3 aneuploidies (Supplementary File 3). 29 mutants had a single mutation associated with fluorescence, 13 mutants had two associated mutations, and 4 mutants had three associated mutations. However, 8 of the 13 mutants with two associated mutations and all 4 mutants with three associated mutations showed linkage (genetic distance below 25 cM) between at least two of the mutations associated with fluorescence (Supplementary File 4, Figure A-2), suggesting that only one of the linked mutations might impact fluorescence in each of these mutants. To determine whether one linked mutation was more likely to impact fluorescence than the others, we compared the magnitude of allele-frequency difference between the high and low fluorescence pools (estimated by the *G*-value) for each mutation. For 9 of the 12 mutants with linked mutations, we found that the mutation with the highest *G*-value was significantly more strongly associated with fluorescence than the linked mutation(s) (resampling test: $P < 0.05$, Supplementary File 4), suggesting that this mutation was responsible for the fluorescence change. For the other 3 mutants, none of the linked mutations showed stronger evidence of impacting fluorescence than the others (resampling test: $P > 0.05$, Supplementary File 4).

The remaining 36 mutants did not have any mutations significantly associated with fluorescence (Supplementary File 2, Figure A-2). These mutants tended to show smaller changes in fluorescence than mutants with one or more associated mutations (Figure A-5), suggesting that our power to map mutations causing 1% changes in fluorescence might have been lower than

anticipated. These 36 mutants might also harbor multiple mutations with small effects on expression, each of which was below our detection threshold. Consistent with this possibility, we observed a small but significant correlation ($r^2 = 0.127$, $P = 0.03$) between the total number of mutations in these 36 EMS mutants and their expression level (Figure A-6). It is also possible that we failed to find associated mutations in some of these mutants because their change in fluorescence was initially overestimated by the "winner's curse" (Xiao & Boehnke, 2009). Accordingly, 71% of mutants selected for mapping after two independent fluorescence screens had at least one mutation significantly associated with fluorescence compared to only 30% of mutants selected after a single fluorescence screen. Some changes in fluorescence observed in these 36 mutants might also have been caused by non-genetic variation and/or undetected mutations.

Additional trans-regulatory mutations identified by sequencing candidate genes

We noticed in the BSA-seq data that three mutations increasing fluorescence more than 5% relative to the un-mutagenized progenitor strain mapped to two genes (*ADE4* and *ADE5*) in the same biochemical pathway (*de novo* purine biosynthesis) (Supplementary File 2). We therefore used Sanger sequencing to test whether these genes or other genes in this pathway were also mutated in 15 additional EMS mutants with fluorescence at least 5% higher than the progenitor strain. We first looked for mutations in *ADE4*, then *ADE5* if no mutation was found in *ADE4*, and then *ADE6* if no mutation was found in the other genes. At least one nonsynonymous mutation was identified by Sanger sequencing in one of these three genes in 14 of the 15 EMS mutants (green points in Figure A-1C,E; Supplementary File 5, Figure A-2). For the remaining mutant (brown point in Figure A-1E), we sequenced a fourth purine biosynthesis gene, *ADE8*, but again found no mutation. In two additional EMS mutants with smaller increases in

fluorescence (2.1% and 4.6%, purple points in Figure A-1D,E) and a reddish color characteristic of *ADE2* loss of function mutants (Roman, 1956), we found nonsynonymous mutations in *ADE2* by Sanger sequencing (Supplementary File 5, Figure A-2). Follow-up experiments showed that mutations in *ADE2*, *ADE5*, and *ADE6* did not increase YFP fluorescence driven by two other promoters (P_{RNR1} and P_{STM1}), suggesting that mutations in the purine biosynthesis pathway affected expression of P_{TDH3}-*YFP* through mechanisms mediated by the *TDH3* promoter rather than YFP (Figure A-7). Taken together, these data suggest that genes in the purine biosynthesis pathway are the predominant mutational source of large increases in *TDH3* expression.

Functional testing confirms effects of trans-regulatory mutations identified by genetic mapping and candidate gene sequencing

To determine whether mutations statistically associated with fluorescence in the BSA-seq data actually affected expression of P_{TDH3}-*YFP*, we introduced 34 of the 67 associated mutations individually into the fluorescent progenitor strain using scarless genetic engineering approaches (Supplementary File 6, Figure A-2). We also used scarless genome editing to create single-site mutants for 11 of the 17 additional mutations identified in purine biosynthesis genes by Sanger sequencing (Supplementary File 5, Supplementary File 6, Figure A-2). Fluorescence of these engineered strains (called "single-site mutants" hereafter) was then quantified by flow cytometry in parallel with fluorescence of the EMS mutant carrying the same associated mutation as well as the un-mutagenized progenitor strain, with four replicate populations analyzed for each genotype. Fluorescence values were then transformed into estimates of YFP abundance as described in the Methods.

Of the 24 mutations without linked variants in EMS mutants that were tested in single-site mutants, 23 (96%) caused a significant change in expression ($P < 0.05$, permutation test,

Supplementary File 6), suggesting a ~4% false positive rate in our BSA-Seq experiment. In addition, all 11 single-site mutants with mutations in purine biosynthesis genes identified by Sanger sequencing showed statistically significant effects on fluorescence relative to the un-mutagenized progenitor strain (all increased fluorescence, $P < 0.05$, permutation test, Supplementary File 6). The remaining 10 mutations tested in single-site mutants were from 5 of the EMS mutants with two linked mutations associated with fluorescence. Each of these mutations was introduced separately into a single-site mutant to independently measure its effect on expression. For 4 of these 5 pairs of linked mutations, only one of the two single-site mutants showed a significant change in expression relative to the progenitor strain (Figure A-3E). In each case, the single-site mutant and the EMS mutant showed changes in expression in the same direction relative to the progenitor strain (Figure A-3E). The mutation affecting expression was always the mutation with the larger *G*-value in the BSA-Seq data, consistent with the results of the statistical tests described above (Supplementary File 4). In the last case (YPW54 in Figure A-3E), both mutations affected expression in the single-site mutants, consistent with our inability to statistically predict which mutation was more likely to impact expression from the BSA-Seq data for this mutant as well as both mutations being nonsynonymous changes in the same gene (*CHD1*) (Supplementary File 4). The BSA-seq data also accurately predicted whether a mutation increased or decreased fluorescence for 27 (93%) of the 29 mutations with significant effects on fluorescence in single-site mutants (Figure A-3F). For the other two mutations, effects on expression in the same direction were observed in the single-site mutants and the corresponding EMS mutants (Supplementary File 5), suggesting that the different growth conditions used for the mapping experiment (see Methods) might have modified the effects of these mutations.

Comparing P_{TDH3}-*YFP* expression in the 40 single-site mutants that significantly altered fluorescence to that in the 40 EMS mutants from which these mutations were identified showed that expression was very similar overall between single-site and EMS mutants sharing the same mutation (Figure A-3G, linear regression: $r^2 = 0.944$, $P = 2.4 \times 10^{-25}$), although significant differences in expression were observed for some pairs (Figure A-3G, Figure A-8). The linear correlation between the expression of single-site mutants and EMS mutants remained strong when mutations identified by sequencing candidate genes (triangles in Figure A-3G) were excluded ($r^2 = 0.854$, $P = 5.5 \times 10^{-13}$). These data suggest that (1) the vast majority of the mutations we identified by genetic mapping and candidate gene sequencing do indeed have *trans*-regulatory effects on expression of P_{TDH3}-*YFP* and (2) the majority of EMS mutants analyzed had a single mutation that was primarily, if not solely, responsible for the observed change in P_{TDH3}-*YFP* expression.

Properties of trans-regulatory mutations affecting expression driven by the TDH3 promoter

In all, 69 mutations showed evidence of affecting P_{TDH3}-*YFP* expression in *trans* (Figure A-2), including 3 aneuploidies and 66 point mutations. 52 of these mutations were identified by genetic mapping (Supplementary File 7) and 17 were identified by sequencing candidate genes (Supplementary File 5). 12 of the mutations identified by genetic mapping were genetically linked to one or more other mutations but showed stronger evidence of affecting P_{TDH3}-*YFP* expression than the linked mutation(s) in statistical and/or functional tests described above (Supplementary File 3). To identify trends in the properties of these 69 *trans*-regulatory mutations, we compared them to 1766 mutations considered non-regulatory regarding P_{TDH3}-*YFP* expression because they showed no significant association with expression of the reporter gene in the BSA-Seq experiment (*G*-test: $P > 0.01$, Figure A-2). To be conservative, 8 mutations

that showed a marginally significant association with expression (*G*-test: $0.001 < P < 0.01$) as well as 15 mutations associated with expression only because of genetic linkage were excluded from further analyses.

First, we asked whether the mutational spectra of *trans*-regulatory mutations differed from non-regulatory mutations (Figure A-9A). We found that G:C to A:T transitions most commonly introduced by EMS occurred at similar frequencies in the two groups (*G*-test, $P = 0.84$). No indels were associated with expression in the BSA-seq data (Supplementary File 7), which was not statistically different from the frequency of indels among non-regulatory mutations (0% vs 2.7%, *G*-test, $P = 0.056$). By contrast, aneuploidies were highly over-represented in the set of *trans*-regulatory mutations since all three extra copies of a chromosome observed in the BSA-Seq data were found to be associated with fluorescence (*G*-test, $P = 8.6 \times 10^{-6}$); a similar overrepresentation was observed when considering only mutations identified by BSA-Seq (Figure A-10A; *G*-test, $P = 3.5 \times 10^{-6}$). We also found a significant difference in the genomic distribution of the two sets of mutations (*G*-test, $P = 2.4 \times 10^{-3}$), with non-regulatory mutations appearing to be randomly distributed throughout the genome but *trans*-regulatory mutations enriched on chromosomes VII and XIII (Figure A-9B, Figure A-11). However, these two chromosomes contain the purine biosynthesis genes in which multiple *trans*-regulatory mutations were identified, and there was no significant difference in genomic distributions between *trans*-regulatory and non-regulatory mutations when mutations in purine biosynthesis genes were excluded (*G*-test, $P = 0.35$) or when mutations identified by direct sequencing of candidate genes were excluded (Figure A-10B; *G*-test, $P = 0.22$).

Trans-regulatory mutations are often assumed to be located in coding sequences, but they can also be located in non-coding, presumably *cis*-regulatory, sequences of *trans*-acting genes

(Hill et al., 2020). We therefore asked whether *trans*-regulatory mutations affecting P_{TDH3}-*YFP* expression were more often found in coding or non-coding regions of the genome than expected by chance. Of the 1766 non-regulatory mutations, 1257 (71.3%) were coding mutations located in exons, and 506 (28.7%) were non-coding mutations located in intergenic (n = 500) or intronic (n = 6) regions (Figure A-9C). This paucity of mutations in introns is consistent with the rarity of introns in *S. cerevisiae*, and the overall frequency of non-coding mutations (28.7%) is similar to the fraction of the *S. cerevisiae* genome (30.6% of 12.1 Mb) considered non-coding (www.yeastgenome.org). By contrast, of the 66 *trans*-regulatory point mutations, only one was located in a non-coding sequence (Figure A-9C). This non-coding mutation was located in the intergenic sequence between *IOC2* and *KIN2*, presumably affecting expression of one or both genes with a downstream effect on P_{TDH3}-*YFP* expression. The 3 aneuploidies were excluded from this and subsequent analyses because they affected both coding and non-coding sequences of a large number of genes. The underrepresentation of non-coding changes among regulatory mutations was statistically significant (1.5% of *trans*-regulatory mutations are non-coding *vs* 28.4% of non-regulatory mutations; *G*-test, $P = 4.3 \times 10^{-9}$), even when excluding mutations identified by sequencing candidate genes (Figure A-10C; *G*-test, $P = 9.1 \times 10^{-7}$).These observations suggest that new *trans*-regulatory mutations affecting P_{TDH3}-*YFP* expression by more than 3% (*i.e.* fluorescence changes greater than 1%) are more likely to alter coding than non-coding sequences. This enrichment in coding sequences might be because coding sequences tend to have a higher density of functional sites than non-coding sequences.

Finally, we examined how *trans*-regulatory mutations located in coding sequences impacted the amino acid sequences of the corresponding proteins. Among mutations identified in coding sequences, 100% of the 65 *trans*-regulatory mutations changed the amino acid sequence

of proteins compared to only 70% of 1257 non-regulatory mutations (Figure A-9D, *G*-test, $P = 1.4 \times 10^{-4}$). Limiting this analysis to the 48 *trans*-regulatory mutations identified by BSA-seq also showed an enrichment of mutations changing the amino acid sequence of proteins (Figure A-10D, *G*-test, $P = 5.6 \times 10^{-6}$). This difference was primarily driven by mutations that introduced stop codons (nonsense mutations) rather than mutations that substituted one amino acid for another (nonsynonymous mutations): 20% of *trans*-regulatory mutations in coding sequences were nonsense mutations *versus* 3% of non-regulatory mutations (Figure A-9D; *G*-test, $P = 4.8 \times 10^{-6}$), and 80% of *trans*-regulatory mutations were nonsynonymous versus 67% of non-regulatory mutations (Figure A-9D; *G*-test, $P = 0.07$). A similar pattern was observed when considering only *trans*-regulatory mutations identified by BSA-Seq (Figure A-10D). Nonsense mutations always altered an arginine, glutamine, or tryptophan codon (Figure A-9E), consistent with the structure of the genetic code and the types of mutations induced by EMS (figure S8 in Metzger et al., 2016). For nonsynonymous mutations, two types of amino acid changes were particularly enriched among *trans*-regulatory mutations (Figure A-9E; Figure A-12): 26.2% of *trans*-regulatory mutations changed glycine to aspartic acid *versus* 5.2% of non-regulatory mutations (permutation test, $P < 10^{-4}$), and 10.8% of *trans*-regulatory mutations changed glycine to glutamic acid versus 2.7% of non-regulatory mutations (permutation test, $P = 0.0042$). As a consequence, mutations altering glycine codons were strongly over-represented in general among *trans*-regulatory mutations (49.2% of *trans*-regulatory mutations *vs* 14.5% of non-regulatory mutations in coding sequences; permutation test, $P < 10^{-4}$). This over-representation remained significant after excluding mutations identified by Sanger sequencing (Figure A-10E, A-13; 41.7% of *trans*-regulatory mutations altering glycine vs 14.5% of non-regulatory mutations, $P = 10^{-4}$). This pattern may be observed because glycine is the smallest amino acid,

making its substitution likely to modify protein structure (Bhate et al., 2002; Miller, 2007). Indeed, glycine is one of the three amino acids with the lowest experimental exchangeability (Yampolsky & Stoltzfus, 2005) and mutations affecting glycine codons are enriched among mutations causing human diseases (Khan & Vihinen, 2007; Molnár et al., 2016; Vitkup et al., 2003).

Regulatory mutations are enriched in a predicted TDH3 regulatory network

Because of the key role transcription factors play in the regulation of gene expression, and because transcription factors have been shown to be a source of *trans*-regulatory variation in natural populations (Albert et al., 2018; Lewis et al., 2014), we asked whether *trans*-regulatory mutations affecting P_{TDH3}-*YFP* expression were enriched in genes encoding transcription factors. We found that 5 (7.7%) of the 65 *trans*-regulatory coding mutations mapped to the coding sequence of one of the 212 genes predicted to encode a transcription factor in the YEASTRACT database (Teixeira et al., 2018), but this was not significantly more than the 5.6% of non-regulatory coding mutations mapping to these genes (*G*-test: $P = 0.52$). *Trans*-regulatory coding mutations were also not significantly enriched in transcription factor genes when we excluded the 17 mutations identified by Sanger sequencing (*G*-test: $P = 0.22$). Not all transcription factors are expected to regulate expression of *TDH3*, however, so we also tested for enrichment of *trans*-regulatory mutations among transcription factors specifically predicted to regulate *TDH3*.

Using information consolidated in the YEASTRACT database (Teixeira et al., 2018) that supports evidence of a transcription factor binding to a gene's promoter and regulating its expression, we constructed a network (Figure A-14) of potential direct regulators of *TDH3* as well as potential direct regulators of these direct regulators (1st and 2nd level regulators of *TDH3*) and asked how often the *trans*-regulatory mutations we identified mapped to these genes. We

found that 4 *trans*-regulatory mutations mapped to three genes in this network, with 2 mutations affecting the 1st level regulator *TYE7*, 1 mutation affecting the 1st level regulator *GCR2*, and 1 mutation affecting the 2nd level regulator *TUP1* (Supplementary File 7). This number of mutations mapping to genes in the predicted *TDH3* regulatory network was 12-fold greater than expected by chance (6.1% for *trans*-regulatory *vs* 0.5% for non-regulatory mutations; *G*-test, $P = 0.0037$), or 16-fold greater than expected by chance when excluding mutations identified by Sanger sequencing (8.2% for *trans*-regulatory *vs* 0.5% for non-regulatory mutations; *G*-test, $P = 0.0024$). Therefore, the inferred regulatory network had predictive power as expected, but the vast majority of *trans*-regulatory coding mutations (61 of 65, or 94%) mapped to genes outside of this network. Only one of these other *trans*-regulatory mutations mapped to a transcription factor. This mutation was a nonsynonymous substitution affecting *ROX1*, which is predicted in the YEASTRACT database to directly regulate expression of the indirect *TDH3* regulator *TUP1*. In other words, *ROX1* is predicted by existing functional genomic data to be a 3rd level regulator of *TDH3* (Figure A-14). With no other transcription factors harboring a *trans*-regulatory mutation in our dataset, this result suggests that mutations in transcription factors located more than three levels away from *TDH3* in its transcriptional regulatory network are unlikely to be sources of new expression changes driven by the *TDH3* promoter.

Deleterious effects of mutations in two direct regulators of TDH3

Transcription factors encoded by the *TYE7* and *GCR2* genes found to harbor *trans*-regulatory mutations affecting expression of P_{TDH3}-*YFP* are known to regulate the expression of glycolytic genes (including *TDH3*) by forming a complex with transcription factors encoded by the *RAP1* and *GCR1* genes (Shively et al., 2019). Rap1p (Yagi et al., 1994) and Gcr1p (Huie et al., 1992) are both known to bind directly to the *TDH3* promoter (Figure A-15A), and mutations

in these binding sites cause large decreases in *TDH3* expression (Metzger et al., 2015). These observations strongly suggest that mutations in *RAP1* and *GCR1* should also cause detectable changes in *TDH3* expression, yet no mutations were observed in these genes in our set of *trans*-regulatory mutations. To investigate why we did not recover *trans*-regulatory mutations in *RAP1* or *GCR1*, we used error-prone PCR to generate mutant alleles of these genes with mutations in either the promoter or coding sequence of *RAP1* or the second exon of *GCR1*, which includes 99.7% of the *GCR1* coding sequence (Figure A-15B). Hundreds of these *RAP1* and *GCR1* mutant alleles were then introduced individually into the un-mutagenized strain carrying the P_{TDH3}-*YFP* reporter gene using CRISPR/Cas9-guided allelic replacement. Sequencing the mutated regions of *RAP1* and *GCR1* in a random subset of transformants showed that each strain harbored an average of 1.8 mutations in the *RAP1* gene (Figure A-15C) or 2.4 mutations in the *GCR1* gene (Figure A-15D). As expected for PCR-based mutagenesis, the number of mutations per strain appeared to follow a Poisson distribution both for *RAP1* mutants (Figure A-15C, Chi-square goodness of fit, $P = 0.14$) and *GCR1* mutants (Figure A-15D, Chi-square goodness of fit, $P = 0.79$).

Among the *RAP1* mutant strains, only 9.1% (43 of 470 strains) showed a significant change in P_{TDH3}-*YFP* expression greater than 3% (corresponding to a ~1% change in fluorescence) relative to the un-mutagenized progenitor strain (Figure A-15E), suggesting that most EMS mutants harboring coding mutations in *RAP1* would have been excluded from our mapping study. In addition, the strongest decrease in P_{TDH3}-*YFP* expression observed among *RAP1* mutants (17%) was substantially smaller than the strongest decrease in expression caused by mutating the RAP1 binding site in the *TDH3* promoter (57.5% reported in Duveau et al., 2018), suggesting that even this most severe phenotype was not caused by a null allele of *RAP1*.

To test this hypothesis, we used site-directed mutagenesis to alter 5 amino acids (one at a time) in Rap1p expected to disrupt DNA binding based on the crystal structure of Rap1p complexed with DNA (Konig et al., 1996). In each case, we obtained by PCR a DNA fragment containing either a synonymous mutation in the codon corresponding to the amino acid (which should not affect the DNA binding of Rap1p) or one of two nonsynonymous mutations, with one nonsynonymous mutation more likely to alter protein function than the other (Yampolsky & Stoltzfus, 2005). We then used CRISPR/Cas9 allele replacement to introduce each mutation into the yeast genome and sequenced 10 independent clones from each transformation to determine if the mutation was introduced in the *RAP1* coding sequence as intended. All five synonymous mutations were observed in several of the clones sequenced, but 7 of the 10 nonsynonymous mutations were never recovered (Supplementary File 8). This outcome suggests that nonsynonymous mutations altering the DNA binding of Rap1p are lethal or nearly lethal, making them unlikely to have been recovered in a mutagenesis screen. Indeed, Rap1p1 is known to be an essential, pleiotropic transcription factor playing critical roles in regulating expression of glycolytic genes like *TDH3* as well as ribosomal proteins and genes required for mating (reviewed in Piña et al., 2003). Taken together, these data indicate that *RAP1* mutations are unlikely to be common sources of variation in expression driven by the *TDH3* promoter.

For the *GCR1* mutant strains, 37.7% showed a significant change in P_{TDH3}-*YFP* expression greater than 3% relative to the un-mutagenized progenitor strain (Figure A-15F). Several of these mutant alleles decreased the expression driven by the *TDH3* promoter by ~80%, which is similar to the previously reported effects of mutations in the Gcr1p binding sites of the *TDH3* promoter (Metzger et al., 2015), suggesting that they were null alleles. Indeed, resequencing these large effect alleles revealed that one of them had a single nucleotide insertion

in the 28th codon of the *GCR1* ORF, which led to a frame shift eliminating 96% of amino acids (757 of 785) from Gcr1p. Because Gcr1p regulates expression of many glycolytic genes (Uemura et al., 1997) and *GCR1* deletion has been reported to cause severe growth defects in fermentable carbon source environments (Clifton et al., 1978; Hossain et al., 2016; López & Baker, 2000), we hypothesized that the fitness effects of mutations in *GCR1* might also have caused them to be underrepresented in the population from which the EMS mutants analyzed were derived. To test this hypothesis, we measured the relative fitness of 62 of the 220 *GCR1* mutants, including all mutants with decreased P_{TDH3}-*YFP* expression. *GCR1* mutants causing the largest changes in P_{TDH3}-*YFP* expression showed strong defects in growth rate; however, several *GCR1* mutants with changes in P_{TDH3}-*YFP* expression greater than 3% did not strongly affect fitness (Figure A-15G). This observation suggests that some of the coding mutations in *GCR1* decreasing P_{TDH3}-*YFP* expression could have been sampled among the EMS mutants used for mapping. We therefore conclude that mutations in *GCR1* were most likely not recovered in our set of regulatory mutations because of the wide diversity of mutations that can affect *TDH3* expression and the limited number of EMS mutants included in the mapping experiment.

Properties of genes harboring regulatory mutations

With only 5 of the 65 *trans*-regulatory point mutations in coding sequences mapping to transcription factors, we used gene ontology (GO) analysis to examine the types of genes harboring *trans*-regulatory mutations affecting P_{TDH3}-*YFP* expression more systematically. In all, these 65 mutations mapped to 42 different genes, with 9 genes affected by more than one mutation, 4 of which were genes involved in the *de novo* purine biosynthesis pathway (Figure A-16A). Several gene ontology terms were significantly enriched among genes affected by *trans*-regulatory mutations relative to genes affected by non-regulatory mutations. Supplementary File

9 includes all enriched GO terms, whereas Figure 6B only includes enriched GO terms that are not parent to other GO terms in the GO hierarchy. Excluding mutations identified by sequencing candidate genes had a negligible impact on the outcome of the GO term analysis, with more than 96% of overlap between the GO terms found to be enriched before and after excluding mutations identified by Sanger sequencing (Supplementary File 8). Of the 33 GO terms enriched for *trans*-regulatory mutations shown in Figure 6B, 11 terms (including 13 of the 42 genes with *trans*-regulatory mutations) were related to chromatin structure (Figure A-16B), which is known to play an important role in the regulation of gene expression (Li et al., 2007). An additional 5 GO terms (including 6 genes with *trans*-regulatory mutations) were related to metabolism, and 4 terms (including 9 genes with *trans*-regulatory mutations) were related to transcriptional regulation (Figure A-16B). Three GO terms related to glucose signaling, including regulation of transcription by glucose, carbohydrate transmembrane transport and glucose metabolic process, were also significantly enriched for genes affected by *trans*-regulatory mutations (Figure A-16B). When we broadened this category of genes based on a review of glucose signaling (Santangelo, 2006), the enrichment included 5 genes implicated in glucose signaling (Supplementary File 10; 12.2% of genes affected by *trans*-regulatory mutations were involved in glucose signaling *vs* 2.7% of genes affected by non-regulatory mutations; Fisher's exact test: $P = 6.2 \times 10^{-3}$).

At the pathway level, we found that genes involved in glycolysis and *de novo* purine biosynthesis were also significantly enriched for *trans*-regulatory mutations (Figure a-16B), with the latter driven by the mutations in *ADE2*, *ADE4*, *ADE5* and *ADE6* genes described above (Supplementary File 11). Genes involved in iron homeostasis also emerged as an over-represented group, with 5 GO terms (including 7 genes) being related to the regulation of

intracellular iron concentration (Figure A-16B). Diverse cellular processes implicated in iron homeostasis were represented among genes harboring *trans*-regulatory mutations, such as iron transport (*FTR1*, *CCC2*), iron trafficking and maturation of iron-sulfur proteins (*CIA2*, *NAR1*), transcriptional regulation of the iron regulon (*FRA1*) and post-transcriptional regulation of iron homeostasis (*TIS11*). Remarkably, nearly half of all *trans*-regulatory point mutations in coding sequences (31 of 65) were located in genes involved either in purine biosynthesis or iron homeostasis. Moreover, 6 of the 8 genes harboring more than one *trans*-regulatory mutation (Figure A-16A) were involved in one of these two processes. Mutations in purine biosynthesis genes tended to cause large increases in expression, whereas mutations in iron homeostasis genes tended to cause large decreases in expression (Supplementary File 11). Although the mechanistic relationship between these pathways and *TDH3* expression is not known, changing cellular conditions, including concentrations of metabolites (Pinson et al., 2009) or iron within the cell (reviewed in Outten & Albetel, 2013), can affect the regulation of gene expression. Ultimately, our data suggest that although mutations affecting P_{TDH3}-*YFP* expression map to genes with diverse functions, genes involved in a small number of well-defined biological processes are particularly likely to harbor such *trans*-regulatory mutations.

Trans-regulatory mutations are enriched in genomic regions harboring natural variation affecting TDH3 expression

Because new mutations affecting gene expression provide the raw material for regulatory variation segregating within a species, we asked whether the *trans*-regulatory mutations we observed were enriched in genomic regions associated with naturally occurring *trans*-regulatory variation affecting expression driven by the *TDH3* promoter. Specifically, we compared the genomic locations of *trans*-regulatory mutations identified in the current study to the locations of

trans-acting quantitative trait loci (QTL) affecting expression of P_{TDH3}-*YFP* identified from crosses between the progenitor strain of the EMS mutants (BY) and 3 other *S. cerevisiae* strains (SK1, YPS1000, M22) (Metzger & Wittkopp, 2019) (Figure A-17A).

Non-regulatory mutations were observed in eQTL regions as often as expected by chance (66.7% of non-regulatory mutations *vs* 65.1% of the whole genome in eQTL regions; *G*-test: $P = 0.15$), but the 66 *trans*-regulatory mutations were significantly enriched in eQTL regions (Figure A-17B; 88% of *trans*-regulatory mutations *vs* 66.7% of non-regulatory mutations in eQTL regions; *G*-test: $P = 9.6 \times 10^{-5}$). The overrepresentation of *trans*-regulatory mutations in eQTL regions remained statistically significant when we considered only the 44 *trans*-regulatory mutations identified from the collection of EMS mutants not enriched for large effects (Figure A-17B; *G*-test: $P = 0.027$) or when we excluded the 17 *trans*-regulatory mutations identified by sequencing candidate genes (Figure A-18; *G*-test: $P = 8.4 \times 10^{-3}$). The enrichment of *trans*-regulatory mutations in eQTL regions was thus not driven solely by the effect size of these mutations or by the fact that several of the *trans*-regulatory mutations with large effects were located in the same genes. We also found that differences in sequencing coverage across the genome were unlikely to account for this enrichment (Figure A-18). When we considered eQTL regions identified from each cross separately, we observed a significant enrichment of *trans*-regulatory mutations in eQTL regions identified in SK1 x BY and YPS1000 x BY crosses, but not in eQTL regions identified in the M22 x BY cross (Figure A-17B; *G*-tests: $P = 0.016$ for SK1 x BY, $P = 6.5 \times 10^{-3}$ for YPS1000 x BY, $P = 0.70$ for M22 x BY). Overall, the enrichment of *trans*-regulatory mutations in eQTL regions suggests that biases in the mutational sources of regulatory variation have shaped genetic sources of expression variation segregating in wild populations.

Discussion

By systematically isolating and characterizing 69 *trans*-regulatory mutations that all affect expression of the same focal gene, this study reveals how *trans*-regulatory mutations are distributed within a genome and within a regulatory network. For example, we found that these *trans*-regulatory mutations were widely spread throughout the genome, with all except one located in coding sequences. These data also allowed us to determine how well a regulatory network inferred from integrating functional genomic and genetic data can predict sources of *trans*-regulatory variation. Like many biological networks, transcriptional regulatory networks have been inferred with the promise of explaining relationships between genetic variants and the higher order trait of gene expression, but the predictive power of such networks remains sparsely tested (Flint & Ideker, 2019).

We found that although the *trans*-regulatory mutations in coding regions were not enriched in transcription factors generally, they were overrepresented among transcription factors inferred to be regulators of *TDH3*. None of these transcription factors are known to directly bind to the *TDH3* promoter, however, and mutations in *RAP1* and *GCR1,* which have well characterized binding sites in the *TDH3* promoter, were notably missing from our set of *trans*-regulatory mutations affecting P_{TDH3}-*YFP* expression. Targeted mutagenesis of *RAP1* and *GCR1* suggested that most mutations in these genes (particularly *RAP1*) cause severe growth defects that might have prevented their recovery in mutagenesis screens. Over 90% of the *trans*-regulatory mutations examined were located in genes outside of this transcription factor network encoding proteins with diverse molecular functions involved in chromatin remodeling, nonsense-mediated mRNA decay, translation regulation, purine biosynthesis, iron homeostasis, and glucose sensing. Surprisingly, nearly half of the *trans*-regulatory mutations mapped to genes

involved in either the purine biosynthesis or iron homeostasis pathways. Although not anticipated, finding so many *trans*-regulatory mutations in genes that are not transcription factors is consistent with the transcriptomic effects of gene deletions showing that transcription factors tend not to affect expression of more genes than other types of proteins (Featherstone & Broadie, 2002). Consequently, it seems that regulatory networks describing the relationships between transcription factors and target genes might capture only a small fraction of the potential sources of *trans*-regulatory variation.

Understanding the properties of *trans*-regulatory mutations is important because these mutations provide the raw material for natural *trans*-regulatory variation. We found that mutations affecting P_{TDH3}-*YFP* expression were enriched in genomic regions associated with expression variation among wild isolates of *S. cerevisiae*, suggesting that mutational sources of regulatory variation have had a lasting effect on sources of genetic variation affecting gene expression segregating in natural populations. This pattern is not necessarily expected if the *trans*-regulatory mutations we characterized captured only a small subset of the loci that can contribute to segregating *trans*-regulatory variation for this gene. Differences between the distribution of new *trans*-regulatory mutations and segregating *trans*-regulatory variants are also expected to arise when natural selection favors the maintenance of mutations at some loci more than others. Such differences in fitness can arise independently of a mutation's impact on *TDH3* expression because *trans*-acting mutations can also have pleiotropic effects on expression of other genes. A third reason why differences between the mutational sources of *trans*-regulatory variation characterized here and *trans*-regulatory variation segregating in the wild can occur would be because of epistatic interactions among variants that are not captured by studying the effects of mutations individually. Ultimately, explaining the variation in gene expression we see

in natural populations will require studies like this elucidating the mutational input as well as studies describing the fitness, pleiotropic, and epistatic effects of these mutations in native environments.

To the best of our knowledge this work provides the largest collection of individual mutations with *trans*-regulatory effects on expression of a single gene available to date, but it still only interrogates a single gene in a single species. Moreover, although the methods used were sensitive enough to identify genetic changes impacting expression of the focal gene as little as 1.6%, many mutations important for natural variation might have even smaller individual effects on a focal gene's expression and are thus missing from this study (Rockman, 2012). The chemical mutagen (EMS) used to generate the mutants analyzed in this work also captures only a subset of the type of mutations that arise naturally, and the use of a YFP reporter gene to measure activity of the *TDH3* promoter precluded recovery of *trans*-regulatory mutations that can impact native *TDH3* expression post-transcriptionally. The focal gene chosen for this work, *TDH3*, might also have properties that cause its spectrum of *trans*-regulatory mutations to differ from other genes in *S. cerevisiae*. For example, *TDH3* is one of the most highly expressed genes in *S. cerevisiae* (Ghaemmaghami et al., 2003), and it is one of the ~8% of genes in the *S. cerevisiae* genome that contains both a TATA box and a large nucleosome-free region in its promoter (Tirosh & Barkai, 2008). The metabolic functions of the TDH3p protein encoded by the *TDH3* gene might also cause its regulatory network to have properties that differ from genes encoding proteins with other types of functions (Luscombe et al., 2004).

It is tempting to extend these results from *S. cerevisiae* to other eukaryotes, but such extrapolation must take into account differences in genomes and gene regulatory mechanisms among species. For example, compared to species like fruit flies, mice, and humans, the baker's

yeast *S. cerevisiae* has a much higher proportion of its genome (69.4%, www.yeastgenome.org) that codes for proteins and much more compact *cis*-regulatory sequences (the median promoter length is 455 bp (Kristiansson et al., 2009)). Consequently, new *trans*-regulatory mutations in coding sequences might be more likely to arise in *S. cerevisiae* than in these other species. Most *S. cerevisiae* genes also lack introns (Parenteau et al., 2019) and DNA methylation is less prevalent than in many other eukaryotic species (Tang et al., 2012), so these potential sources of *trans*-regulatory variation in other species are unlikely to be captured when studying regulatory mutations in *S. cerevisiae*. Nonetheless, we think some observations, such as that genes with diverse functions can harbor *trans*-regulatory mutations, are likely to also apply to other eukaryotic species. Ultimately, we believe that this work provides an important foundation for understanding how the *trans*-regulatory mutations that give rise to *trans*-regulatory variation segregating in natural populations are structured within a genome and a regulatory network.

Materials and methods

Mutant strains selected for mapping

To identify mutations associated with expression changes, we selected 82 haploid mutant strains for bulk segregant analysis (Figure A-2A) from three collections of mutants obtained in Gruber et al. (2012) and Metzger et al. (2016) via ethyl methanesulfonate (EMS) mutagenesis of two progenitor strains expressing a *YFP* reporter gene (Yellow Fluorescent Protein) under control of the *TDH3* promoter (P_{TDH3}-*YFP*). 71 mutants were selected from a collection of 1498 lines founded from cells isolated randomly (unenriched) after mutagenesis in Metzger et al. (2016), 5 mutants were selected from 211 lines founded from cells enriched for fluorescence changes after mutagenesis in Metzger et al. (2016) and the last 6 mutants were selected from 1064 lines founded from cells enriched for fluorescence changes in Gruber et al. (2012). Mutants

from Metzger et al. (2016) were obtained by mutagenesis of the progenitor strain YPW1139 (*MATα ura3d0*), while mutants from Gruber et al. (2012) were obtained by mutagenesis of the progenitor strain YPW1 (*MAT***a** *ura3d0 lys2d0*). Both progenitors were derived from S288c genetic background (see Metzger et al. 2016 and Gruber et al. 2012 for details on construction of YPW1139 and YPW1 strains). In YPW1139, P_{TDH3}-*YFP* is inserted at the *ho* locus with a *KanMX* drug resistance marker. In YPW1, P_{TDH3}-*YFP* is inserted at position 199270 on chromosome I near a pseudogene. YPW1139 harbors *RME1(ins-308A)* and *TAO3(1493Q)* alleles (Deutschbauer & Davis, 2005) that increase sporulation frequency relative to YPW1 alleles, as well as *SAL1*, *CAT5* and *MIP1* alleles that decrease the frequency of the petite phenotype (Dimitrov et al., 2009). We previously showed that the few genetic differences between YPW1 and YPW1139 did not affect the magnitude of effects of *TDH3* promoter mutations on fluorescence (Metzger et al., 2016). Fluorescence levels of the three collections were measured in Gruber et al. (2012) and in Metzger et al. (2016). From these data, we selected 39 mutants for BSA-Seq that showed statistically significant fluorescence changes greater than 1% relative to the progenitor strain. Among these mutants, 6 were selected from the Gruber et al. (2012) collection (Z-score > 2.58, $P < 0.01$), 5 were selected from mutants enriched for large effects in Metzger et al. (2016) (permutation test, $P < 0.05$) and 28 were selected from unenriched mutants in Metzger et al. (2016) (permutation test, $P < 0.05$). The remaining 43 mutants included in BSA-Seq experiments were selected from mutants in Metzger et al. (2016) for which we collected new fluorescence measures using flow cytometry. This second fluorescence screen included 197 lines from the unenriched collection that were chosen because they showed statistically significant fluorescence changes (permutation test, $P < 0.05$) greater than 1% relative to the progenitor strain in the initial screen published in Metzger et al. (2016). The 43 mutants

selected from this 2nd screen showed statistically significant fluorescence changes (permutation test, $P < 0.05$) greater than 1% relative to the progenitor strain.

Measuring YFP expression by flow cytometry

Fluorescence levels of mutant strains were quantified by flow cytometry using the same approach as described in Metzger et al. (2016) and Duveau et al. (2018). For assays involving strains stored in individual tubes at -80°C, all strains were thawed in parallel on YPG plates (10 g yeast extract, 20 g peptone, 50 ml glycerol, 20 g agar per liter) and grown for 2 days at 30°C. Strains were then arrayed using pipette tips in 96 deep well plates containing 0.5 ml of YPD medium (10 g yeast extract, 20 g peptone, 20 g D-glucose per liter) per well at positions defined in Supplementary File 12. The reference strain YPW1139 was inoculated at 20 fixed positions on each plate to correct for plate and position effects on fluorescence. The non-fluorescent strain YPW978 was inoculated in one well per plate to quantify the autofluorescence of yeast cells. Plates were incubated at 30°C for 20 hours with 250 rpm orbital shaking (each well contained a sterile 3 mm glass bead to maintain cells in suspension). Samples from each plate were then transferred to omnitrays containing YPG-agar using a V & P Scientific pin tool. For assays involving strains already arrayed in 96-well plates at -80°C (*i.e. RAP1* and *GCR1* mutants), strains were directly transferred on YPG omnitrays after thawing. After 48 hours of incubation at 30°C, samples from each omnitray were inoculated using the pin tool in four replicate 96-well plates containing 0.5 ml of YPD per well and cultivated at 30°C with 250 rpm shaking for 22 hours. Then, 15 µl of cell cultures were transferred to a 96-well plate with 0.5 ml of PBS per well (phosphate-buffered saline) and samples were immediately analyzed on a BD Accuri C6 flow cytometer connected to a HyperCyt autosampler (IntelliCyt Corp). A 488 nm laser was used for excitation and the YFP signal was acquired with a 530/30 optical filter. Each well was

sampled for 2 seconds, yielding fluorescence and cell size measurements for at least 5000 events per well. Flow cytometry data were analyzed using custom R scripts (Source Code 1) as described in Duveau et al. (2018). First, events that did not correspond to single cells were filtered out using *flowClust* clustering functions. Second, fluorescence intensity was scaled by cell size in several steps. For Figures A-1B-D, these values of fluorescence relative to cell size were directly used for subsequent steps of the analysis. For other figures, these values were transformed using a log-linear function to be linearly related with YFP abundance. Transformations of fluorescence values were performed using the relationship between fluorescence levels and *YFP* mRNA levels established in Duveau et al. (2018) from five strains carrying mutations in the promoter of the P_{TDH3}-*YFP* reporter gene. The YFP mRNA levels quantified in these five strains are expected to be linearly related with YFP protein abundance based on a previous study that compared mRNA and protein levels for a similar fluorescent protein (GFP) across a broad range of expression levels (Kafri et al., 2016). For this reason and because mutations recovered in this study may alter YFP expression at the post-transcriptional level, the transformed values of fluorescence were considered to provide estimates of YFP abundance instead of mRNA levels. The median expression among all cells of each sample was then corrected to account for positional effects estimated from a linear model applied to the median expression of the 20 control samples on each plate. To correct for autofluorescence, the mean of median expression measured among all replicate populations of the non-fluorescent strain was then subtracted from the median expression of each sample. Finally, a relative measure of expression was calculated by dividing the median expression of each sample by the mean of the median expression among replicates of the reference strain. Samples for which the relative expression differed from the median expression among replicate populations by more

than five times the median absolute deviation measured among replicate populations were considered as outliers and ignored. Figures show the mean relative expression among replicate populations of each genotype. Permutation tests used to compare the expression level of each single site mutant to the expression level of the EMS mutant carrying the same mutation are described in the legend of Figure A-8A.

Two-level permutation tests

We developed a permutation-based approach to determine which EMS mutant strains from Metzger et al. (2016) showed a significant change in YFP expression relative to their progenitor strain. This permutation approach was motivated by the fact that Student tests and Mann-Whitney-Wilcoxon tests applied to these data appeared to be overpowered. Indeed, the flow cytometry assay from Metzger et al. (2016) included 146 instances of the progenitor strain YPW1139 that were placed at random plate positions and with fluorescence measured in four replicate populations for each position. When comparing the mean expression of the four replicate populations of YPW1139 grown at a given plate position to the mean expression of all other replicate populations of YPW1139, the *P*-value was below 0.05 in 25.3% of cases when using Student tests and in 13.7% of cases when using Mann-Whitney-Wilcoxon tests. The fact that more than 5% of *P*-values were below 0.05 indicated that the tests were overpowered, which was because expression differences between YPW1139 populations grown at different plate positions were in average larger than expression differences between replicate populations grown at the same position. For this reason, we compared the expression of each mutant strain to the expression of the 146 x 4 populations of the YPW1139 progenitor strain using permutation tests with two levels of resampling as described below. In these tests, we compared 10,000 times the expression levels of each tested strain measured in quadruplicates to the expression levels of

YPW1139 measured in quadruplicates at a randomly selected plate position among the 146 available positions (a new position was picked at each iteration). For each iteration of the comparison, we calculated the difference *D* between 1) the absolute difference observed between the mean expression of the tested strain and the mean expression of YPW1139 and 2) a randomized absolute difference of mean expression between two sets of 4 expression values obtained by random permutation of the 4 expression values measured for the tested strain and of the 4 expression values measured for YPW1139 at the selected plate position. Finally, for each tested strain the proportion of *D* values that were negative (after excluding *D* values equal to zero) corresponded to the *P*-value of the permutation test. When we applied this test to YPW1139 as a tested strain, we found that the *P*-value was below 0.05 for 6.1% of the 146 plate positions containing YPW1139, indicating that the permutation test was not overpowered.

BSA-Seq procedure

To identify mutations associated with fluorescence levels in EMS-treated mutants, we used bulk-segregant analysis followed by Illumina sequencing (BSA-Seq). BSA-Seq data corresponding to the 6 mutants from Gruber et al. (2012) were collected together with the BSA-Seq dataset published in Duveau et al. (2014). For the other 76 mutants (from Metzger et al., 2016), BSA-Seq data were collected in this study in several batches (see Supplementary File 13) using the experimental approach described in Duveau et al. (2014) (with few modifications). First, each EMS-treated mutant (*MATα ura3d0 ho::*P_{TDH3}*-YFP ho::KanMX*) was crossed to the mapping strain YPW1240 (*MAT***a** *ura3d0 ho::*P_{TDH3}*-YFP ho::NatMX4 mata2::yEmRFP-HygMX*) that contained the *FASTER MT* system from Chin et al. (2012) used to tag diploid and *MAT***a** cells with a fluorescent reporter. Crosses were performed on YPD agar plates and replica-plated on YPD + G418 + Nat medium (YPD agar with 350 mg/L geneticin (G418) and 100 mg/L

Nourseothricin) to select diploid hybrids. After growth, cells were streaked on another YPD + G418 + Nat agar plate, one colony was patched on YPG agar for each mutant and the diploid strain was kept frozen at -80°C. Bulk segregant populations were then collected for batches of 8 mutants in parallel as follows. Diploid strains were thawed and revived on YPG plates, grown for 12 hours at 30°C on GNA plates (50 g D-glucose, 30 g Difco nutrient broth, 10 g yeast extract and 20 g agar per liter) and sporulation was induced for 4 days at room temperature on KAc plates (10 g potassium acetate and 20 g agar per liter). For each mutant, we then isolated a large population of random spores (> 10^8 spores) by digesting tetrads with zymolyase, vortexing and sonicating samples in 0.02% triton-X (exactly as described in Duveau et al., 2014). ~3 x 10^5 *MATα* spores were sorted by FACS (BD FACSAria II) based on the absence of RFP fluorescence signal measured using a 561 nm laser and 582/15 optical filter. Spores were then resuspended in 2 ml of YPD medium. After 24 hours of growth at 30°C, 0.4 ml of cell culture was transferred to a 5 ml tube containing 2 ml of PBS. Three populations of 1.5 x 10^5 segregant cells were then collected by FACS: 1) a low fluorescence population of cells sorted among the 2.5% of cells with lowest fluorescence levels ("low bulk"), 2) a high fluorescence population of cells sorted among the 2.5% of cells with highest fluorescence levels ("high bulk") and 3) a control population of cells sorted regardless of their fluorescence levels. YFP signal was measured using a 488 nm laser and a 530/30 optical filter. To exclude budding cells and enrich for single cells, ~70% of all events were filtered out based on the area and width of the forward scatter signal prior to sorting. In addition, the median FSC.A (area of forward scatter, a proxy for cell size) was maintained to similar values in the low fluorescence bulk and in the high fluorescence bulk by drawing sorting gates that were parallel to the linear relationship between FSC.A and fluorescence intensity in the FACSDiva software. After sorting, cells were

resuspended in 1.6 ml of YPD medium and grown for 30 hours at 30°C. Each sample was then stored at -80°C in 15% glycerol in two separate tubes: one tube containing 1 ml of culture (for DNA extraction) and one tube containing 0.5 ml of culture (for long-term storage). Extraction of genomic DNA was performed for 24 samples in parallel using a Gentra Puregene Yeast/Bact kit (Qiagen). Then, DNA libraries were prepared from 1 ng of genomic DNA using Nextera XT DNA Library Prep kits (Illumina) for low fluorescence bulks and for high fluorescence bulks (control populations were not sequenced). Tagmentation was carried out at 55°C for 5 minutes. Dual indexing of the libraries was achieved using index adapters provided in the Nextera XT Index kit (index sequences used for each library are indicated in Supplementary File 14). Final library purification and size selection was achieved using Agencourt AMPure XP beads (30 µl of beads added to 50 µl of PCR-amplified libraries followed by ethanol washes and resuspension in 50 µl of Tris-EDTA buffer). The average size of DNA fragments in the final libraries was 650 bp, as quantified from a subset of samples using high sensitivity assays on a 2100 Bioanalyzer (Agilent). The concentration of all libraries was quantified with a Qubit 2.0 Fluorometer (Thermo Fisher Scientific) using dsDNA high sensitivity assays. Libraries to be sequenced in the same flow lane were pooled to equal concentration in a single tube and sequenced on a HiSeq4000 instrument (Illumina) at the University of Michigan Sequencing Core Facility (150-bp paired-end sequencing). The 2 x 76 libraries were sequenced in 4 distinct sequencing runs (45300, 45301, 54374 and 54375) that included 36 to 54 samples (libraries sequenced in each run are indicated in Supplementary File 14). In addition, 4 control libraries were sequenced in run 45300, corresponding to genomic DNA from 1) YPW1139 progenitor strain, 2) YPW1240 mapping strain, 3) a bulk of low fluorescence segregants from YPW1139 x YPW1240 cross and

4) a bulk of high fluorescence segregants from YPW1139 x YPW1240 cross. 18 libraries sequenced in run 54374 were not analyzed in this study.

Analysis of BSA-Seq data

Demultiplexing of sequencing reads and generation of FASTQ files were performed using Illumina *bcl2fastq* v1.8.4 for sequencing runs 45300 and 45301 and *bcl2fastq2* v2.17 for runs 54374 and 54375. The next steps of the analysis were processed on the Flux cluster administered by the Advanced Research Computing Technology Services of the University of Michigan (script available in Source Code 4). First, low quality ends of reads were trimmed with *sickle* (https://github.com/najoshi/sickle) and adapter sequences were removed with *cutadapt* (Martin, 2011). Reads were then aligned to the S288c reference genome (https://www.yeastgenome.org/, R64-1-1 release to which we added the sequences corresponding to P_{TDH3}-*YFP*, *KanMX* and *NatMX4* transgenes, available in Supplementary File 12) using *bowtie2* (Langmead & Salzberg, 2012) and overlaps between paired reads were clipped using *clipOverlap* in *bamUtil* (https://github.com/statgen/bamUtil). The sequencing depth at each position in the genome was determined using *bedtools genomecov* (https://github.com/arq5x/bedtools2). For variant calling, BAM files corresponding to the low fluorescence bulk and to the high fluorescence bulk of each mutant were processed together using *freebayes* (https://github.com/ekg/freebayes; Garrison & Marth, 2012) with options *--pooled-discrete --pooled-continuous*. That way, sequencing data from both bulks were pooled to increase the sensitivity of variant calling and allele counts were reported separately for each bulk. To obtain a list of mutations present in each mutant strain, false positive calls in the VCF files generated by *freebayes* were then filtered out with the Bioconductor package *VariantAnnotation* in R (Source Code 2). Filtering was based on the values of several parameters

such as quality of genotype inference (QUAL > 200), mapping quality (MQM > 27), sequencing depth (DP > 20), counts of reference and alternate alleles (AO > 3 and RO > 3), frequency of the reference allele (FREQ.REF > 0.1), proportion of reference and alternate alleles supported by properly paired reads (PAIRED > 0.8 and PAIREDR > 0.8), probability to observe the alternate allele on both strands (SAP < 100) and at different positions of the reads (EPP < 50 and RPP < 50). The values of these parameters were chosen to filter out a maximum number of calls while retaining 28 variants previously confirmed by Sanger sequencing. We then used likelihood ratio tests (*G*-tests) in R to determine for each variant site whether the frequency of the alternate allele (*i.e.* the mutation) was statistically different between the low fluorescence bulk and the high fluorescence bulk (Source Code 2). A point mutation was considered to be associated with fluorescence (directly or by linkage) if the *P*-value of the *G*-test was below 0.001, corresponding to a *G* value above 10.828. Since this *G*-test was performed for a total of 1819 mutations, we expected that 1.82 mutations would be associated with fluorescence due to type I error (false positives) at a *P*-value threshold of 0.001. This expected number of false positives was considered acceptable since it represented only 2.7% of all mutations that were associated with fluorescence. To determine if an aneuploidy was associated with fluorescence level, we compared the sequencing coverage of the aneuploid chromosome to genome-wide sequencing coverage in the low and high fluorescence bulks using *G*-tests. The *G* statistics was computed from the number of reads mapping to the aneuploid chromosome and the number of reads mapping to the rest of the genome in the low and high fluorescence bulks. Aneuploidies with $G >$ 10.828, which corresponds to *P*-value < 0.001, were considered to be present at statistically different frequencies in both bulks. A custom R script was used to annotate all mutations identified in BSA-Seq data (Source Code 3), retrieving information about the location of

mutations in intergenic, intronic or exonic regions, the name of genes affected by coding mutations or the name of neighboring genes in case of intergenic mutations and the expected impact on amino acid sequences (synonymous, nonsynonymous or nonsense mutation and identity of the new amino acid in case of a nonsynonymous mutation).

Sanger sequencing of candidate genes

As an alternative approach to BSA-Seq, additional mutations were identified by directly sequencing candidate genes in a subset of EMS-treated mutants (Supplementary File 4). More specifically, we sequenced the P_{TDH3}-*YFP* transgene in 95 mutant strains from Metzger et al. (2016) that showed decreased fluorescence by more than 10% relative to the progenitor strain. We sequenced the *ADE4* coding sequence in 14 mutants from Metzger et al. (2016) that were not included in the BSA-Seq assays and that showed increased fluorescence by more than 5% relative to the progenitor strain. Two of the sequenced mutants had a mutation in the *ADE4* coding sequence. We then sequenced the *ADE5* coding sequence in the remaining 12 mutants and found a mutation in five of the sequenced mutants. We continued by sequencing the *ADE6* coding sequence in the remaining seven mutants. Five of the sequenced mutants had a single mutation and one mutant had two mutations in the *ADE6* coding sequence. We sequenced the *ADE8* coding sequence in the last mutant but we found no candidate mutation in this mutant. Finally, we sequenced the *ADE2* coding sequence in two mutants that showed a reddish color when growing on YPD plates. For all genes, the sequenced region was amplified by PCR from cell lysates, PCR products were cleaned up using Exo-AP treatment (7.5 µl PCR product mixed with 0.5 µl Exonuclease-I (NEB), 0.5 µl Antarctic Phosphatase (NEB), 1 µl Antarctic Phosphastase buffer and 0.5 µl H_2O incubated at 37°C for 15 minutes followed by 80°C for 15 minutes) and Sanger sequencing was performed by the University of Michigan Sequencing Core

Facility. Oligonucleotides used for PCR amplification and sequencing are indicated in Supplementary File 14.

Site-directed mutagenesis

34 mutations identified by BSA-Seq and 11 mutations identified by sequencing candidate genes were introduced individually in the genome of the progenitor strain YPW1139 to quantify the effect of these mutations on fluorescence level. “Scarless” genome editing (i.e. without insertion of a selection marker) was achieved using either the delitto perfetto approach from Stuckey et al. (2011) (for 19 mutations) or CRISPR-Cas9 approaches derived from Laughery et al. (2015) (for 26 mutations). Compared to delitto perfetto, CRISPR-Cas9 is more efficient and it can be used to introduce mutations in essential genes. However, it requires specific sequences in the vicinity of the target mutation (see below). The technique used for the insertion of each mutation is indicated in Supplementary File 15. The sequences of oligonucleotides used for the insertion and the validation of each mutation can be found in Supplementary File 14.

In the delitto perfetto approach, the target site was first replaced by a cassette containing the *Ura3* and *hphMX4* selection markers (pop-in) and then this cassette was swapped with the target mutation (pop-out). The *Ura3-hphMX4* cassette was amplified from pCORE-UH plasmid using two oligonucleotides that contained at their 5’ end 20 nucleotides for PCR priming in pCORE-UH and at their 3’ end 40 nucleotides corresponding to the sequences flanking the target site in the yeast genome (for homologous recombination). The amplicon was transformed into YPW1139 cells using a classic LiAc/polyethylene glycol heat shock protocol (Gietz & Schiestl, 2007). Cells were then plated on synthetic complete medium lacking uracil (SC-Ura) and incubated for two days at 30°C. Colonies were replica-plated on YPD + Hygromycin B (300 mg/l) plates. A dozen [Ura+ Hyg+] colonies were streaked on SC-Ura plates to remove residual

parental cells and the resulting colonies were patched on YPG plates to counterselect petite cells. Cell patches were then screened by PCR to confirm the proper insertion of *Ura3-hphMX4* at the target site. One positive clone was grown in YPD and stored at -80°C in 15% glycerol. For the pop-out step, a genomic region of ~240 bp centered on the mutation was amplified from the EMS-treated mutant containing the desired mutation. The amplicon was transformed into the strain with *Ura3-hphMX4* inserted at the target site. Cells were plated on a synthetic complete medium containing 0.9 g/l of 5-fluoroorotic acid (SC + 5-FOA) to counterselect cells expressing *Ura3*. After growth, a dozen [Ura-] colonies were streaked on SC + 5-FOA plates and one colony from each streak was patched on a YPG plate. Cell patches were screened by PCR using oligonucleotides that flanked the sequence of the transformed region and amplicons of expected size (~350 bp) were sequenced to confirm the insertion of the desired mutation and the absence of PCR-induced mutations. When possible two independent clones were stored at -80°C in 15% glycerol, but in some cases only one positive clone could be retrieved and stored.

A "one-step" CRISPR-Cas9 approach was used to insert mutations impairing a NGG or CCN motif in the genome (22 mutations), which corresponds to the protospacer adjacent motif (PAM) targeted by Cas9. First, a DNA fragment containing the 20 bp sequence upstream of the target PAM in the yeast genome was cloned between SwaI and BclI restriction sites in the pML104 plasmid. This DNA fragment was obtained by hybridizing two oligonucleotides designed as described in Laughery *et al.* 2015. The resulting plasmid contained cassettes for expression of Ura3, Cas9 and a guide RNA targeted to the mutation site in yeast cells. In parallel, a repair fragment containing the mutation was obtained either by PCR amplification of a ~240 bp genomic region centered on the mutation in the EMS-treated mutant or by hybridization of two complementary 70-mer oligonucleotides containing the mutation and its flanking genomic

sequences. The Cas9/sgRNA plasmid and the repair fragments were transformed together (~150 nmol of plasmid + 20 µmol of repair fragment) into the progenitor strain YPW1139 using LiAc/polyethylene glycol heat shock protocol (Gietz & Schiestl, 2007). Cells were then plated on SC-Ura medium and incubated at 30°C for 48 hours. This medium selected cells that both internalized the plasmid and integrated the desired mutation in their genome. Indeed, cells with the Cas9/sgRNA plasmid stop growing as long as their genomic DNA is cleaved by Cas9 but their growth can resume once the PAM sequence is impaired by the mutation, which is integrated into the genome via homologous recombination with the repair fragment (Laughery et al., 2015). A dozen [Ura+] colonies were then streaked on SC-Ura plates and one colony from each streak was patched on a YPG plate. Cell patches were screened by PCR using oligonucleotides that flanked the mutation site and amplicons of expected size (~350 bp) were sequenced to confirm the insertion of the desired mutation and the absence of secondary mutations. Then, one or two positive clones were patched on SC + 5-FOA to counterselect the Cas9/sgRNA plasmid, grown in YPD and stored at -80°C in 15% glycerol.

A "two-steps" CRISPR-Cas9 approach was used to insert mutations located near but outside a PAM sequence (4 mutations). Each step was performed as described above for the "one-step" CRISPR-Cas9 approach. In the first step, Cas9 was targeted by the sgRNA to a PAM sequence (the initial PAM) located close to the mutation site (up to 20 bp). The repair fragment contained two synonymous mutations that were not the target mutation: one mutation that impaired the initial PAM and one mutation that introduced a new PAM as close as possible to the target site. This repair fragment was obtained by hybridization of two complementary 90-mer oligonucleotides and transformed into YPW1139. In the second step, Cas9 was targeted to the new PAM. The repair fragment contained three mutations: two mutations that reverted the

mutations introduced in the first step and the target mutation. This repair fragment was obtained by hybridization of two complementary 90-mer oligonucleotides and transformed into the strain obtained in the first step. Positive clones were sequenced to confirm the insertion of the target mutation and the absence of other mutations.

We used CRISPR/Cas9-guided allele replacement to introduce individual mutations in five codons of the *RAP1* coding sequence that encode for amino acids predicted to make direct contact with DNA when RAP1 binds to DNA (Konig et al., 1996). For each codon, we tried to insert one synonymous mutation, one nonsynonymous mutation predicted to have a weak impact on RAP1 protein structure and one nonsynonymous mutation predicted to have a strong impact on RAP1 protein structure based on amino acid exchangeability scores from Yampolsky & Stoltzfus (2005) (see Supplementary File 8 for the list of mutations). Each mutation was introduced in the genome of strain YPW2706. This strain is derived from YPW1139 and contains two identical sgRNA target sites upstream and downstream of the *RAP1* gene (see below for details on YPW2706 construction). Therefore, we could use a single Cas9/sgRNA plasmid to excise the entire *RAP1* gene in YPW2706 by targeting Cas9 to both ends of the gene. We used gene SOEing (Splicing by Overlap Extension) to generate repair fragments corresponding to the *RAP1* gene (promoter and coding sequence) with each target mutation. First, a left fragment of *RAP1* was amplified from YPW1139 genomic DNA using a forward 20-mer oligonucleotide priming upstream of the RAP1 promoter and a reverse 60-mer oligonucleotide containing the target mutation and the surrounding *RAP1* sequence. In parallel, a right fragment of RAP1 overlapping with the right fragment was amplified from YPW1139 genomic DNA using a forward 60-mer oligonucleotide complementary to the reverse oligonucleotide used to amplify the left fragment and a reverse 20-mer oligonucleotide priming

in *RAP1* 5'UTR sequence. Then, equimolar amounts of the left and right fragments were mixed in a PCR reaction and 25 cycles of PCR were performed to fuse both fragments. Finally, the resulting product was further amplified using two 90-mer oligonucleotides with homology to the sequence upstream of *RAP1* promoter and to the *RAP1* 5'UTR but without the sgRNA target sequences. Consequently, transformation of the repair fragment together with the Cas9/sgRNA plasmid in YPW2706 cells was expected to replace the wild type allele of *RAP1* by an allele containing the target mutation in *RAP1* coding sequence and without the two flanking sgRNA target sites. For each of the 15 target mutations, we sequenced the *RAP1* promoter and coding sequence in 10 independent clones obtained after transformation. All synonymous mutations were retrieved in several clones, while several of the nonsynonymous mutations were not found in any clone, suggesting they were lethal (Supplementary File 8).

RAP1 and GCR1 mutagenesis using error-prone PCR

We used a mutagenic PCR approach to efficiently generate hundreds of mutants with random mutations in the *RAP1* gene (promoter and coding sequence) or in the second exon of *GCR1* (representing 99.7% of *GCR1* coding sequence). DNA fragments obtained from the mutagenic PCR were introduced in the yeast genome using CRISPR/Cas9-guided allele replacement as described above. The sequences of all oligonucleotides used for *RAP1* and *GCR1* mutagenesis can be found in Supplementary File 14.

First, we constructed two yeast strains for which the *RAP1* gene (strain YPW2706) or the second exon of *GCR1* (strain YPW3082) were flanked by identical sgRNA target sites and PAM sequences. To generate strain YPW2706, we first identified a sgRNA target site located downstream of the *RAP1* coding sequence (41 bp after the stop codon in the 5'UTR) in the S288c genome. Then, we inserted the 23 bp sequence corresponding to this sgRNA target site

and PAM upstream of the *RAP1* promoter (immediately after *PPN2* stop codon) in strain YPW1139 using the delitto perfetto approach (as described above). To generate strain YPW3082, we first identified a sgRNA target site located at the end of the *GCR1* intron (22 bp upstream of exon 2) in the S288c genome. Then, we inserted the 23 bp sequence corresponding to this sgRNA target site and PAM immediately after the *GCR1* stop codon in strain YPW1139 using the delitto perfetto approach (as described above).

Second, we constructed plasmid pPW437 by cloning the 20mer guide sequence directed to *RAP1* in pML104 as described in Laughery *et al.* 2015 and we constructed plasmid pPW438 by cloning the 20mer guide sequence directed to *GCR1* in pML104 as described in Laughery et al. 2015. These two sgRNA/Cas9 plasmids can be used, respectively, to excise the *RAP1* gene or *GCR1* exon 2 from the genomes of YPW2706 and YPW3082.

Third, we generated repair fragments with random mutations in *RAP1* or *GCR1* genes using error-prone PCR. We first amplified each gene from 2 ng of YPW1139 genomic DNA using a high-fidelity polymerase (KAPA HiFi DNA polymerase) and 30 cycles of PCR. PCR products were purified with the Wizard SV Gel and PCR Clean-Up System (Promega) and quantified with a Qubit 2.0 Fluorometer (Thermo Fisher Scientific) using dsDNA broad range assays. 2 ng of purified PCR products were used as template for a first round of mutagenic PCR and mixed with 25 µl of DreamTaq Master Mix 2x (ThermoFisher Scientific), 2.5 µl of forward and reverse primers at 10 µM, 5 µl of 1 mM dATP and 5 µl of 1 mM dTTP in a final volume of 50 µl. The imbalance of dNTP concentrations (0.3 µM dATP, 0.2 µM dCTP, 0.2 µM dGTP and 0.3 µM dTTP) was done to bias the mutagenesis toward misincorporation of dATP and dTTP. For *RAP1* mutagenesis, the forward oligonucleotide primed upstream of the *RAP1* promoter (in *PPN2* coding sequence) and the reverse oligonucleotide primed in the *RAP1* terminator and

contained a mutation in the PAM adjacent to the sgRNA target site. For *GCR1* mutagenesis, the forward oligonucleotide primed at the end of the *GCR1* intron and contained a mutation in the PAM adjacent to the sgRNA target site and the reverse primer primed in the *GCR1* terminator. The PCR program was 95°C for 3 minutes followed by 32 cycles with 95°C for 30 seconds, 52°C for 30 seconds, 72°C for 2 minutes and a final extension at 72°C for 5 minutes. For *RAP1* mutagenesis, the product of the first mutagenic PCR was diluted by a factor of 33 and used as template for a second round of mutagenic PCR (1.5 µl of product in a 50 µl reaction) similar to the first round but with only 10 cycles of amplification. For *GCR1* mutagenesis, the product of the first mutagenic PCR was diluted by a factor of 23 and used as template for a second round of mutagenic PCR (2.2 µl of product in a 50 µl reaction) with 35 cycles of amplification. Using this protocol, we expected to obtain on average 1.6 mutations per fragment for *RAP1* mutagenesis and 1.8 mutations per fragment for *GCR1* mutagenesis (see below for calculations of these estimates).

pPW437 was transformed with *RAP1* repair fragments into YPW2706 and pPW438 was transformed with *GCR1* repair fragments into YPW3082 as described above for CRISPR/Cas9 site directed mutagenesis. To select cells that replaced the wild type alleles with alleles containing random mutations, transformed cells were plated on SC-Ura and incubated at 30°C for 48 hours. To confirm the success of each mutagenesis and to estimate actual mutation rates, we then sequenced the *RAP1* genes in 27 random colonies from the *RAP1* mutagenesis and we sequenced the second exon of *GCR1* in 18 random colonies from the *GCR1* mutagenesis. Next, 500 colonies from *RAP1* mutagenesis and 300 colonies from *GCR1* mutagenesis were streaked onto SC-Ura plates. After growth, one colony from each streak was patched on YPG and grown four days at 30°C. Then, patches were replica-plated with velvets onto SC + 5-FOA to eliminate

sgRNA/Cas9 plasmids. Finally, 488 clones from RAP1 mutagenesis and 355 clones from GCR1 mutagenesis were arrayed in 96-well plates containing 0.5 ml of YPD (same plate design as used for the flow cytometry assays) and grown overnight at 30°C. 0.2 ml of cell culture from each well was then mixed with 46 µl of 80% glycerol in 96-well plates and stored at -80°C. The fluorescence of these strains was quantified by flow cytometry as described above to assess the impact of RAP1 and GCR1 mutations on P_{TDH3}-*YFP* expression (expression data for each mutant can be found in Supplementary File 16).

In our mutagenesis approach, we introduced a mutation that impaired the target PAM sequence in all *RAP1* and *GCR1* mutants. To determine the effect of this mutation alone, we generated strains YPW2701 and YPW2732 that carried the PAM mutation in the *RAP1* terminator or in the *GCR1* intron, respectively, without any other mutation in *RAP1* or *GCR1*. The fluorescence level of these two strains was not significantly different from the fluorescence level of the progenitor strain YPW1139 in flow cytometry assays.

Estimation of RAP1 and GCR1 mutation rates

The expected number of mutations per PCR amplicon (Nmut) depends on the error rate of the Taq polymerase (), on the number of DNA duplications (D) and on the length of the amplicon (L): Nmut=DL. The published error rate for a classic polymerase similar to DreamTaq is ~3 x 10^{-5} errors per nucleotide per duplication (McInerney et al., 2014). Amplicon length was 3057 bp for *RAP1* mutagenesis and 2520 pb for *GCR1* mutagenesis. The number of duplications of PCR templates was calculated from the amounts of double stranded DNA quantified using Qubit 2.0 dsDNA assays before (I) and after (O) each mutagenic PCR reaction as follows: D=lnOIln 2. For the first round of *RAP1* mutagenesis, D=ln65501.93ln 2 =11.7. For the second round of *RAP1* mutagenesis, D=ln300068.1ln 2 =5.5. Therefore, the total number of duplications

was 17.2 and the expected number of mutations per amplicon Nmut was 1.6 on average. For the first round of *GCR1* mutagenesis, D=ln68702.15ln 2 =11.6. For the second round of *GCR1* mutagenesis, D=ln65351.65ln 2 =12.0. Therefore, the total number of duplications was 23.6 and the expected number of mutations per amplicon Nmut was 1.8 on average.

Effects of mutations in purine biosynthesis genes on expression from different promoters

We compared the individual effects of three mutations in the purine biosynthesis pathway (*ADE2-C1477a*, *ADE5-G1715a* and *ADE6-G3327a*) on YFP expression driven by four different yeast promoters (P_{TDH3}, P_{RNR1}, P_{STM1} and P_{GPD1}). Each mutation was introduced individually in the genomes of four parental strains described in Hodgins-Davis et al. (2019) carrying either P_{TDH3}-*YFP* (YPW1139), P_{RNR1}-*YFP* (YPW3758), P_{STM1}-*YFP* (YPW3764) or P_{GPD1}-*YFP* (YPW3757) reporter gene at the *ho* locus. Site-directed mutagenesis was performed as described in the corresponding section (see above). The fluorescence of the four parental strains, of a non-fluorescent strain (YPW978) and of the 12 mutant strains (4 reporter genes x 3 mutations) was quantified using a Sony MA-900 flow cytometer (the BD Accuri C6 instrument used for other fluorescence assays was not available due to Covid-19 shutdown) in three replicate experiments performed on different days. For each experiment, all strains were grown in parallel in culture tubes containing 5 ml of YPD and incubated at 30°C for 16 hours. Each sample was diluted to 1-2 x 10^7 cells/mL in PBS prior to measurement. At least 5 x 10^4 events were recorded for each sample using a 488 nm laser for YFP excitation and a 525/50 optical filter for the acquisition of fluorescence. At least 5 x 10^4 events were recorded for each sample. Flow cytometry data were then processed in R using functions from the *FlowCore* package and custom scripts available in Source Code 1. After log-transformation of flow data, events considered to correspond to single cells were selected on the basis of their forward scatter height and width (FSC-H and FSC-W).

Fluorescence values of single cells were then normalized to account for differences in cell size. Finally, the median fluorescence among cells was computed for each sample and averaged across replicates of each genotype.

Statistical comparisons of trans-regulatory and nonregulatory mutations

We established a set of 69 *trans*-regulatory mutations that included 52 mutations with a *P*-value below 0.01 in the *G*-tests comparing the frequencies of mutant and reference alleles in low and high fluorescence bulks (see above) as well as 17 mutations identified by Sanger sequencing in the coding sequence of purine biosynthesis genes. In parallel, we established a set of 1766 nonregulatory mutations regarding P_{TDH3}-*YFP* expression that included mutations with a *P*-value above 0.01 in the *G*-tests comparing the frequencies of mutant and reference alleles in low and high fluorescence bulks (see above) and mutations that did not affect P_{TDH3}-*YFP* expression in single-site mutants. We performed statistical analysis to compare properties of *trans*-regulatory and nonregulatory mutations using RStudio v1.2.5019 (R scripts are in Source Code 2). We used *G*-tests (*likelihood.ratio* function in *Deducer* package) to compare the following properties between *trans*-regulatory and nonregulatory mutations: i) the frequency of G.C to A:T transitions, ii) the frequency of indels, iii) the frequency of aneuploidies, iv) the distribution of mutations among chromosomes, v) the frequency of mutations in coding, intronic and intergenic sequences, vi) the frequency of synonymous, nonsynonymous and nonsense changes among coding mutations, vii) the frequency of coding mutations in transcription factors, viii) the frequency of coding mutations in the predicted *TDH3* regulatory network (see below), ix) the proportion of mutations in eQTL regions (see below). We used resampling tests to compare the frequencies of different amino acid changes caused by *trans*-regulatory and nonregulatory mutations in coding sequences. We computed for each possible amino acid change

the observed absolute difference between i) the proportion of coding *trans*-regulatory mutations causing the amino acid change and ii) the proportion of nonregulatory mutations causing the amino acid change. Then, we computed similar absolute differences for 10,000 randomly permuted sets of *trans*-regulatory and nonregulatory mutations. The *P*-value for each amino acid change was calculated as the proportion of resampled absolute differences greater or equal to the observed absolute difference.

TDH3 regulatory network

The network of potential *TDH3* regulators shown on Figure A-14 was established using data available in July 2019 on the YEASTRACT (www.yeastract.com) repository of regulatory associations between transcription factors and target genes in *Saccharomyces cerevisiae* (Teixeira et al., 2018). We used the tool "Regulation Matrix" to obtain three matrices in which rows corresponded to the 220 transcription factor genes in YEASTRACT and columns corresponded to the 6886 yeast target genes included in the database. In the first matrix obtained using the option "Only DNA binding evidence", an element had a value of 1 if the transcription factor at the corresponding row was reported in the literature to bind to the promoter of the target gene at the corresponding column and a value of 0 otherwise. The two other matrices were obtained using the option "Only Expression evidence" with either "TF acting as activator" or "TF acting as inhibitor". An element had a value of 1 only in the "TF acting as activator" matrix if perturbation of the transcription factor at the corresponding row was reported to increase expression of the target gene at the corresponding column. An element had a value of 1 only in the "TF acting as inhibitor" matrix if perturbation of the transcription factor at the corresponding row was reported to decrease expression of the target gene at the corresponding column. An element had a value of 1 in both matrices if perturbation of the transcription factor at the

corresponding row was reported to affect expression of the target gene at the corresponding column in an undetermined direction. Finally, an element had a value of 0 in both matrices if perturbation of the transcription factor at the corresponding row was not reported to alter expression of the target gene at the corresponding column in the literature. We then used a custom R script (Source Code 2) to generate a smaller matrix that only contained first level and second level regulators of *TDH3* and *TDH3* itself. A transcription factor was considered to be a first level regulator of *TDH3* if a regulatory association with *TDH3* was supported both by DNA binding evidence and expression evidence. A transcription factor was considered to be a second level regulator of *TDH3* if a regulatory association with a first level regulator of *TDH3* was supported both by DNA binding evidence and expression evidence. The network shown on Figure A-14 was drawn using Adobe Illustrator based on regulatory interactions included in the matrix of *TDH3* regulators (in Supplementary File 12). To determine whether mutations in the *TDH3* regulatory network constituted a significant mutational source of regulatory variation affecting P_{TDH3} activity, we compared the proportions of *trans*-regulatory and non-regulatory mutations that were located in a *TDH3* regulator gene (first or second level) using a *G*-test (*likelihood.ratio* function in R package *Deducer*).

Competitive fitness assays

We performed competitive growth assays to quantify the fitness of 62 strains with random mutations in the second exon of *GCR1*. These 62 strains corresponded to all *GCR1* mutants that showed a significant decrease of P_{TDH3}-*YFP* expression as quantified by flow cytometry as well as *GCR1* mutants for which *GCR1* exon 2 was sequenced and the location of mutations was known. The 62 strains were thawed on YPG plates as well as reference strains YPW1139 and YPW2732 and strain YPW1182 that expressed a GFP (Green Fluorescent

Protein) reporter instead of YFP. After three days of incubation at 30°C, strains were arrayed in four replicate 96-well plates containing 0.5 ml of YPG per well. In parallel, the [GFP+] strain YPW1182 was also arrayed in four replicate 96-well plates. The eight plates were incubated on a wheel at 30°C for 32 hours. We then measured the optical density at 620 nm of all samples using a Sunrise plate reader (Tecan) and calculated the average cell density for each plate. Samples were then transferred to 1.2 ml of YPD in 96-well plates to reach an average cell density of 10^6 cells/ml for each plate. 21.25 µl of samples from plates containing [YFP+] strains were mixed with 3.75 µl of [GFP+] samples in four 96-well plates containing 0.45 ml of YPD per well. The reason why [YFP+] and [GFP+] strains were mixed to a 17:3 ratio is because we anticipated that some of the *GCR1* mutants may grow slower than the [GFP+] competitor in YPD. Samples were then grown on a wheel at 30°C for 10 hours and the optical density was measured again after growth to estimate the average number of generations for each plate. The ratio of [YFP+] and [GFP+] cells in each sample was quantified by flow cytometry before and after the 10 hours of growth. Samples were analyzed on a BD Accuri C6 flow cytometer with a 488 nm laser used for excitation and two different optical filters (510/10 and 585/40) used to separate YFP and GFP signals. FCS data were analyzed with custom R scripts using *flowCore* and *flowClust* packages (Source Code 1) as described in Duveau et al. (2018). First, we filtered out artifactual events with extreme values of forward scatter or fluorescence intensity. Then, for each sample we identified two clusters of events corresponding to [YFP+] and [GFP+] cells using a principal component analysis on the logarithms of FL1.H and FL2.H (height of the fluorescence signal captured through the 510/10 and 585/40 filters, respectively). Indeed, [YFP+] cells tend to have lower FL1.H value and higher FL2.H value than [GFP+] cells and these two parameters are positively correlated. The competitive fitness of [YFP+] cells relative to [GFP+] cells was

calculated as the exponential of the slope of the linear regression of logeYFPGFPon the number of generations of growth (where YFP corresponds to the number of [YFP+] cells and GFP corresponds to the number of [GFP+] cells). We then divided the fitness of each sample by the mean fitness among all replicates of the reference strain YPW1139 to obtain a fitness value relative to YPW1139. The fitness of each strain was calculated as the mean relative fitness among the four replicate populations for that strain. These fitness data can be found in Supplementary File 16.

Gene ontology (GO) analysis

GO term analyses were performed on www.pantherdb.org website in June 2020 (Mi et al., 2019). In "Gene List Analysis", we used "Statistical overrepresentation test" on a query list corresponding to the 42 genes affected by *trans*-regulatory coding mutations. GO enrichment was determined based on a reference list of the 1251 genes affected by non-regulatory coding mutations using Fisher's exact tests. Four separate analyses were performed for GO biological processes, GO molecular functions, GO cellular components and PANTHER pathways. GO terms that are significantly enriched in the list of *trans*-regulatory mutations (mutations associated with fluorescence level) relative to non-regulatory mutations (mutations not associated with fluorescence level) at $P < 0.05$ are listed in Supplementary File 9.

Enrichment of mutations in eQTL regions

Genomic regions containing expression quantitative trait loci (eQTL) associated with P_{TDH3}-*YFP* expression variation in three different crosses (BYxYPS1000, BYxSK1 and BYxM22) were obtained from Table S11 in Metzger & Wittkopp (2019). A custom R script was used to determine the number of *trans*-regulatory and non-regulatory mutations located inside

and outside these eQTL intervals (Source Code 2). *G*-tests were performed to determine whether the proportion of *trans*-regulatory mutations in eQTL intervals was statistically different from the proportion of non-regulatory mutations in the same eQTL intervals.

Data archiving

De-multiplexed sequencing data are available in FASTQ format from NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under BioProject number PRJNA706682. Flow cytometry data (FCS files) are available on the Flow Repository (https://flowrepository.org/) under the following experiments ID: FR-FCM-Z3WV for the secondary screen of fluorescence in EMS mutants shown in Figure A-2E, FR-FCM-Z3JY for the quantifications of fluorescence in single site mutants and in the corresponding EMS mutants (Figure A-3E-G), FR-FCM-Z3J2 for the quantifications of fluorescence in *RAP1* mutant strains (Figure A-15E), FR-FCM-Z3J3 for the quantifications of fluorescence in *GCR1* mutant strains (Figure A-15F-G) and FR-FCM-Z3J5 for the quantifications of fitness in the same *GCR1* mutants strains (Figure A-15G).

Acknowledgments

We thank Gaël Yvert and Mark Hill for helpful comments on the manuscript, the University of Michigan sequencing core and University of Michigan flow cytometry core for research support, and the National Institutes of Health (R01GM108826 and R35GM118073 to P.J.W.), European Molecular Biology Organization (EMBO ALTF 1114-2012 to F.D.), National Science Foundation (MCB-1929737 to P.J.W.), NIH Genetics Training grant (T32GM007544 to P.V.Z.), NIH Genome Sciences Training Grant (T32HG000040 to B.P.H.M and M.A.S.), and the Michigan Life Sciences Fellow program (M.A.S.) for funding.

References

Albert, F. W., Bloom, J. S., Siegel, J., Day, L., & Kruglyak, L. (2018). Genetics of *trans*-regulatory variation in gene expression. *eLife*, *7*, e35471. https://doi.org/10.7554/eLife.35471

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews. Genetics*, *16*(4), 197-212. https://doi.org/10.1038/nrg3891

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., Stahl, E. A., Huckins, L. M., Nicolae, D. L., Cox, N. J., & Im, H. K. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, *9*(1), 1825. https://doi.org/10.1038/s41467-018-03621-1

Bhate, M., Wang, X., Baum, J., & Brodsky, B. (2002). Folding and Conformational Consequences of Glycine to Alanine Replacements at Different Positions in a Collagen Model Peptide. *Biochemistry*, *41*(20), 6539-6547. https://doi.org/10.1021/bi020070d

Chin, B. L., Frizzell, M. A., Timberlake, W. E., & Fink, G. R. (2012). FASTER MT : Isolation of Pure Populations of a and α Ascospores from *Saccharomyces cerevisiae*. *G3: Genes, Genomes, Genetics*, *2*(4), 449-452. https://doi.org/10.1534/g3.111.001826

Clifton, D., Weinstock, S. B., & Fraenkel, D. G. (1978). Glycolysis mutants in *Saccharomyces cerevisiae*. *Genetics*, *88*(1), 1-11.

Consortium, T. Gte. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318-1330. https://doi.org/10.1126/science.aaz1776

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., Leeuwen, J. van, Dyk, N. van, Lin, Z.-Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., … Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, *353*(6306). https://doi.org/10.1126/science.aaf1420

Deutschbauer, A. M., & Davis, R. W. (2005). Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nature Genetics*, *37*(12), 1333-1340. https://doi.org/10.1038/ng1674

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*(1), 365-383. https://doi.org/10.1534/genetics.109.104497

Duveau, F., Hodgins-Davis, A., Metzger, B. P., Yang, B., Tryban, S., Walker, E. A., Lybrook, T., & Wittkopp, P. J. (2018). Fitness effects of altering gene expression noise in *Saccharomyces cerevisiae*. *eLife*, *7*, e37272. https://doi.org/10.7554/eLife.37272

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: Genes, Genomes, Genetics*, *4*(7), 1205-1216. https://doi.org/10.1534/g3.114.011783

Featherstone, D. E., & Broadie, K. (2002). Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *24*(3), 267-274. https://doi.org/10.1002/bies.10054

Ferraro, N. M., Strober, B. J., Einson, J., Abell, N. S., Aguet, F., Barbeira, A. N., Brandt, M., Bucan, M., Castel, S. E., Davis, J. R., Greenwald, E., Hess, G. T., Hilliard, A. T., Kember, R. L., Kotis, B., Park, Y., Peloso, G., Ramdas, S., Scott, A. J., … Battle, A. (2020). Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science*, *369*(6509). https://doi.org/10.1126/science.aaz5900

Flint, J., & Ideker, T. (2019). The great hairball gambit. *PLOS Genetics*, *15*(11), e1008519. https://doi.org/10.1371/journal.pgen.1008519

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E. M., Aguet, F., Quan, J., Nicolae, D. L., Eskin, E., Kellis, M., Getz, G., McCarthy, M. I., Dermitzakis, E. T., Cox, N. J., & Ardlie, K. G. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics*, *50*(7), 956-967. https://doi.org/10.1038/s41588-018-0154-4

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]*. http://arxiv.org/abs/1207.3907

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., … Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415*(6868), 141-147. https://doi.org/10.1038/415141a

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*(6959), 737-741. https://doi.org/10.1038/nature02046

Gietz, R. D., & Schiestl, R. H. (2007). High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols*, *2*(1), 31-34. https://doi.org/10.1038/nprot.2007.13

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 Genetics*, *8*(2), e1002497. https://doi.org/10.1371/journal.pgen.1002497

Hill, M. S., Vande Zande, P., & Wittkopp, P. J. (2020). Molecular and evolutionary processes generating variation in gene expression. *Nature Reviews Genetics*, 1-13. https://doi.org/10.1038/s41576-020-00304-w

Hodgins-Davis, A., Duveau, F., Walker, E. A., & Wittkopp, P. J. (2019). Empirical measures of mutational effects define neutral models of regulatory evolution in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, *116*(42), 21085-21093. https://doi.org/10.1073/pnas.1902823116

Hornung, G., Oren, M., & Barkai, N. (2012). Nucleosome organization affects the sensitivity of gene expression to promoter mutations. *Molecular Cell*, *46*(3), 362-368. https://doi.org/10.1016/j.molcel.2012.02.019

Hossain, M. A., Claggett, J. M., Edwards, S. R., Shi, A., Pennebaker, S. L., Cheng, M. Y., Hasty, J., & Johnson, T. L. (2016). Posttranscriptional Regulation of Gcr1 Expression and Activity Is Crucial for Metabolic Adjustment in Response to Glucose Availability. *Molecular Cell*, *62*(3), 346-358. https://doi.org/10.1016/j.molcel.2016.04.012

Hughes, T. R., & Boer, C. G. de. (2013). Mapping Yeast Transcriptional Networks. *Genetics*, *195*(1), 9-36. https://doi.org/10.1534/genetics.113.153262

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., … Friend, S. H. (2000). Functional Discovery via a Compendium of Expression Profiles. *Cell*, *102*(1), 109-126. https://doi.org/10.1016/S0092-8674(00)00015-5

Huie, M. A., Scott, E. W., Drazinic, C. M., Lopez, M. C., Hornstra, I. K., Yang, T. P., & Baker, 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. *Molecular and Cellular Biology*, *12*(6), 2690-2700.

Jackson, C. A., Castro, D. M., Saldi, G.-A., Bonneau, R., & Gresham, D. (2020). Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *eLife*, *9*, e51254. https://doi.org/10.7554/eLife.51254

Kafri, M., Metzl-Raz, E., Jona, G., & Barkai, N. (2016). The Cost of Protein Production. *Cell Reports*, *14*(1), 22-31. https://doi.org/10.1016/j.celrep.2015.12.015

Kemmeren, P., Sameith, K., van de Pasch, L. A. L., Benschop, J. J., Lenstra, T. L., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C. W., van Heesch, S., Kashani, M. M., Ampatziadis-Michailidis, G., Brok, M. O., Brabers, N. A. C. H., Miles, A. J., Bouwmeester, D., van Hooff, S. R., van Bakel, H., … Holstege, F. C. P. (2014). Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. *Cell*, *157*(3), 740-752. https://doi.org/10.1016/j.cell.2014.02.054

Khan, S., & Vihinen, M. (2007). Spectrum of disease-causing mutations in protein secondary structures. *BMC Structural Biology*, *7*(1), 56. https://doi.org/10.1186/1472-6807-7-56

Konig, P., Giraldo, R., Chapman, L., & Rhodes, D. (1996). The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell*, *85*(1), 125-136. https://doi.org/10.1016/s0092-8674(00)81088-0

Kristiansson, E., Thorsen, M., Tamás, M. J., & Nerman, O. (2009). Evolutionary Forces Act on Promoter Length : Identification of Enriched *Cis*-Regulatory Elements. *Molecular Biology and Evolution*, *26*(6), 1299-1307. https://doi.org/10.1093/molbev/msp040

Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C., & Cohen, B. A. (2012). Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(47), 19498-19503. https://doi.org/10.1073/pnas.1210678109

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923

Laughery, M. F., Hunter, T., Brown, A., Hoopes, J., Ostbye, T., Shumaker, T., & Wyrick, J. J. (2015). New Vectors for Simple and Streamlined CRISPR-Cas9 Genome Editing in *Saccharomyces cerevisiae*. *Yeast (Chichester, England)*, *32*(12), 711-720. https://doi.org/10.1002/yea.3098

Leeuwen, J. van, Pons, C., Mellor, J. C., Yamaguchi, T. N., Friesen, H., Koschwanez, J., Ušaj, M. M., Pechlaner, M., Takar, M., Ušaj, M., VanderSluis, B., Andrusiak, K., Bansal, P., Baryshnikova, A., Boone, C. E., Cao, J., Cote, A., Gebbia, M., Horecka, G., … Boone, C. (2016). Exploring genetic suppression interactions on a global scale. *Science*, *354*(6312). https://doi.org/10.1126/science.aag0839

Lewis, J. A., Broman, A. T., Will, J., & Gasch, A. P. (2014). Genetic Architecture of Ethanol-Responsive Transcriptome Variation in *Saccharomyces cerevisiae* Strains. *Genetics*, *198*(1), 369-382. https://doi.org/10.1534/genetics.114.167429

Li, B., Carey, M., & Workman, J. L. (2007). The Role of Chromatin during Transcription. *Cell*, *128*(4), 707-719. https://doi.org/10.1016/j.cell.2007.01.015

Liu, J., François, J.-M., & Capp, J.-P. (2019). Gene Expression Noise Produces Cell-to-Cell Heterogeneity in Eukaryotic Homologous Recombination Rate. *Frontiers in Genetics*, *10*, 475. https://doi.org/10.3389/fgene.2019.00475

Liu, Z., Miller, D., Li, F., Liu, X., & Levy, S. F. (2020). A large accessory protein interactome is rewired across environments. *eLife*, *9*, e62365. https://doi.org/10.7554/eLife.62365

López, M. C., & Baker, H. V. (2000). Understanding the Growth Phenotype of the Yeast *gcr1* Mutant in Terms of Global Genomic Expression Patterns. *Journal of Bacteriology*, *182*(17), 4970-4978.

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*(7006), 308-312. https://doi.org/10.1038/nature02782

Lutz, S., Brion, C., Kliebhan, M., & Albert, F. W. (2019). DNA variants affecting the expression of numerous genes in *trans* have diverse mechanisms of action and evolutionary histories. *PLoS Genetics*, *15*(11), e1008375. https://doi.org/10.1371/journal.pgen.1008375

Maricque, B. B., Dougherty, J. D., & Cohen, B. A. (2017). A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of *cis*-regulatory activity in neural cells. *Nucleic Acids Research*, *45*(4), e16-e16. https://doi.org/10.1093/nar/gkw942

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10-12. https://doi.org/10.14806/ej.17.1.200

McAlister, L., & Holland, M. J. (1985). Isolation and characterization of yeast strains carrying mutations in the glyceraldehyde-3-phosphate dehydrogenase genes. *The Journal of Biological Chemistry*, *260*(28), 15013-15018.

McInerney, P., Adams, P., & Hadi, M. Z. (2014). Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Molecular Biology International*, *2014*, 1-8. https://doi.org/10.1155/2014/287430

Mehrabian, M., Allayee, H., Stockton, J., Lum, P. Y., Drake, T. A., Castellani, L. W., Suh, M., Armour, C., Edwards, S., Lamb, J., Lusis, A. J., & Schadt, E. E. (2005). Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nature Genetics*, *37*(11), 1224-1233. https://doi.org/10.1038/ng1619

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. *Nature Biotechnology*, *30*(3), 271-277. https://doi.org/10.1038/nbt.2137

Metzger, B. P. H., Duveau, F., Yuan, D. C., Tryban, S., Yang, B., & Wittkopp, P. J. (2016). Contrasting Frequencies and Effects of *cis*- and *trans*-Regulatory Mutations Affecting

Gene Expression. *Molecular Biology and Evolution*, *33*(5), 1131-1146. https://doi.org/10.1093/molbev/msw011

Metzger, B. P. H., & Wittkopp, P. J. (2019). Compensatory *trans*-regulatory alleles minimizing variation in TDH3 expression are common within *Saccharomyces cerevisiae*. *Evolution Letters*, *3*(5), 448-461. https://doi.org/10.1002/evl3.137

Metzger, B. P. H., Yuan, D. C., Gruber, J. D., Duveau, F., & Wittkopp, P. J. (2015). Selection on noise constrains variation in a eukaryotic promoter. *Nature*, *521*(7552), 344-347. https://doi.org/10.1038/nature14244

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14 : More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, *47*(D1), D419-D426. https://doi.org/10.1093/nar/gky1038

Miller, B. G. (2007). The mutability of enzyme active-site shape determinants. *Protein Science : A Publication of the Protein Society*, *16*(9), 1965-1968. https://doi.org/10.1110/ps.073040307

Molnár, J., Szakács, G., & Tusnády, G. E. (2016). Characterization of Disease-Associated Mutations in Human Transmembrane Proteins. *PLoS ONE*, *11*(3). https://doi.org/10.1371/journal.pone.0151760

Oliver, F., Christians, J. K., Liu, X., Rhind, S., Verma, V., Davison, C., Brown, S. D. M., Denny, P., & Keightley, P. D. (2005). Regulatory Variation at Glypican-3 Underlies a Major Growth QTL in Mice. *PLOS Biology*, *3*(5), e135. https://doi.org/10.1371/journal.pbio.0030135

Outten, C. E., & Albetel, A.-N. (2013). Iron sensing and regulation in *Saccharomyces cerevisiae*: Ironing out the mechanistic details. *Current Opinion in Microbiology*, *16*(6), 662-668. https://doi.org/10.1016/j.mib.2013.07.020

Parenteau, J., Maignon, L., Berthoumieux, M., Catala, M., Gagnon, V., & Abou Elela, S. (2019). Introns are mediators of cell response to starvation. *Nature*, *565*(7741), 612-617. https://doi.org/10.1038/s41586-018-0859-7

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. *Nature Biotechnology*, *27*(12), 1173-1175. https://doi.org/10.1038/nbt.1589

Piña, B., Fernández-Larrea, J., García-Reyero, N., & Idrissi, F.-Z. (2003). The different (sur)faces of Rap1p. *Molecular Genetics and Genomics: MGG*, *268*(6), 791-798. https://doi.org/10.1007/s00438-002-0801-3

Pinson, B., Vaur, S., Sagot, I., Coulpier, F., Lemoine, S., & Daignan-Fornier, B. (2009). Metabolic intermediates selectively stimulate transcription factor interaction and modulate phosphate and purine pathways. *Genes & Development*, *23*(12), 1399-1407. https://doi.org/10.1101/gad.521809

Rhee, H. S., & Pugh, B. F. (2011). Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*, *147*(6), 1408-1419. https://doi.org/10.1016/j.cell.2011.11.013

Rockman, M. V. (2012). The QTN Program and the Alleles That Matter for Evolution: All That's Gold Does Not Glitter. *Evolution*, *66*(1), 1-17. https://doi.org/10.1111/j.1558-5646.2011.01486.x

Roman, H. (1956). A system selective for mutations affecting the synthesis of adenine in yeast. *Compt. Rend. Trav. Lab. Carlsberg, Ser. physiol.*, *26*, 299-314.

Santangelo, G. M. (2006). Glucose Signaling in *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews*, *70*(1), 253-282. https://doi.org/10.1128/MMBR.70.1.253-282.2006

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, J., Zhu, H., Kash, S. F., Drake, T. A., … Lusis, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, *37*(7), 710-717. https://doi.org/10.1038/ng1589

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. *Nature Biotechnology*, *30*(6), 521-530. https://doi.org/10.1038/nbt.2205

Shively, C. A., Liu, J., Chen, X., Loell, K., & Mitra, R. D. (2019). Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32), 16143-16152. https://doi.org/10.1073/pnas.1818015116

Shiwa, Y., Fukushima-Tanaka, S., Kasahara, K., Horiuchi, T., & Yoshikawa, H. (2012). Whole-Genome Profiling of a Novel Mutagenesis Technique Using Proofreading-Deficient DNA Polymerase. *International Journal of Evolutionary Biology,* 2012:860797. https://doi.org/10.1155/2012/860797

Sinnott-Armstrong, N., Naqvi, S., Rivas, M., & Pritchard, J. K. (2021). GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife*, *10*, e58615. https://doi.org/10.7554/eLife.58615

Stuckey, S., Mukherjee, K., & Storici, F. (2011). In Vivo Site-Specific Mutagenesis and Gene Collage Using the Delitto Perfetto System in Yeast *Saccharomyces cerevisiae*. *Methods in molecular biology (Clifton, N.J.)*, *745*, 173-191. https://doi.org/10.1007/978-1-61779-129-1_11

Tang, Y., Gao, X.-D., Wang, Y., Yuan, B.-F., & Feng, Y.-Q. (2012). Widespread Existence of Cytosine Methylation in Yeast DNA Measured by Gas Chromatography/Mass Spectrometry. *Analytical chemistry*, *84*(16), 7249-7255. https://doi.org/10.1021/ac301727c

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M. S., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., & Michnick, S. W. (2008). An in Vivo Map of the Yeast Protein Interactome. *Science*, *320*(5882), 1465-1470. https://doi.org/10.1126/science.1153878

Teixeira, M. C., Monteiro, P. T., Palma, M., Costa, C., Godinho, C. P., Pais, P., Cavalheiro, M., Antunes, M., Lemos, A., Pedreira, T., & Sá-Correia, I. (2018). YEASTRACT : An upgraded database for the analysis of transcription regulatory networks in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, *46*(D1), D348-D353. https://doi.org/10.1093/nar/gkx842

Tirosh, I., & Barkai, N. (2008). Two strategies for gene regulation by promoter nucleosomes. *Genome Research*, *18*(7), 1084-1091. https://doi.org/10.1101/gr.076059.108

Uemura, H., Koshio, M., Inoue, Y., Lopez, M. C., & Baker, H. V. (1997). The role of Gcr1p in the transcriptional activation of glycolytic genes in yeast *Saccharomyces cerevisiae*. *Genetics*, *147*(2), 521-532.

Uphoff, S., Lord, N. D., Okumus, B., Potvin-Trottier, L., Sherratt, D. J., & Paulsson, J. (2016). Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation. *Science*, *351*(6277), 1094-1097. https://doi.org/10.1126/science.aac9786

Vitkup, D., Sander, C., & Church, G. M. (2003). The amino-acid mutational spectrum of human genetic disease. *Genome Biology*, *4*(11), R72. https://doi.org/10.1186/gb-2003-4-11-r72

Xiao, R., & Boehnke, M. (2009). Quantifying and correcting for the winner's curse in genetic association studies. *Genetic epidemiology*, *33*(5), 453-462. https://doi.org/10.1002/gepi.20398

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. *The Journal of Veterinary Medical Science*, *56*(2), 235-244. https://doi.org/10.1292/jvms.56.235

Yampolsky, L. Y., & Stoltzfus, A. (2005). The exchangeability of amino acids in proteins. *Genetics*, *170*(4), 1459-1472. https://doi.org/10.1534/genetics.104.039107

Yao, C., Joehanes, R., Johnson, A. D., Huan, T., Liu, C., Freedman, J. E., Munson, P. J., Hill, D. E., Vidal, M., & Levy, D. (2017). Dynamic Role of *trans* Regulation of Gene Expression in Relation to Complex Traits. *The American Journal of Human Genetics*, *100*(4), 571-580. https://doi.org/10.1016/j.ajhg.2017.02.003

Yvert, G., Brem, R. B., Whittle, J., Akey, J. M., Foss, E., Smith, E. N., Mackelprang, R., & Kruglyak, L. (2003). *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genetics*, *35*(1), 57-64. https://doi.org/10.1038/ng1222

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M., & Snyder, M. (2010). Genetic analysis of variation in transcription factor binding in yeast. *Nature*, *464*(7292), 1187-1191. https://doi.org/10.1038/nature08934

Zhu, C., Byers, K. J. R. P., McCord, R. P., Shi, Z., Berger, M. F., Newburger, D. E., Saulrieta, K., Smith, Z., Shah, M. V., Radhakrishnan, M., Philippakis, A. A., Hu, Y., De Masi, F., Pacek, M., Rolfs, A., Murthy, T., Labaer, J., & Bulyk, M. L. (2009). High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Research*, *19*(4), 556-566. https://doi.org/10.1101/gr.090233.108

Figures

A **Collections of mutants analyzed in this study.**

Collection	*Initial sorting*	*Ploidy*	*Mutants total*	*Analyzed by BSA-Seq*	*Analyzed by Sanger seq*	*Decreased fluorescence*	*Increased fluorescence*	*Figure panels*
Gruber *et al.* 2012	Large effects	1*n*	1064	6	0	3	3	1B
Metzger *et al.* 2016	Large effects	1*n*	211	5*	11*	1	14	1C
Metzger *et al.* 2016	Unenriched	1*n*	1498	71	6	35	42	1D,E

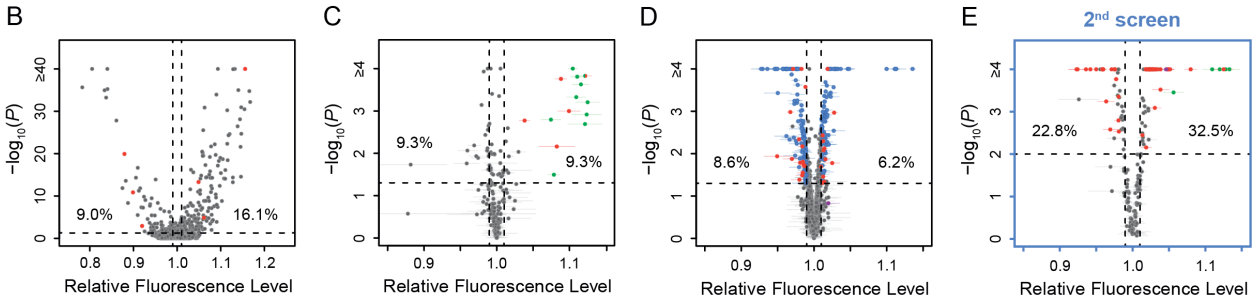

Figure A-1: Mutant strains analyzed with altered expression of a PTDH3-YFP reporter gene.

(A) Summary of the three previously published collections of *S. cerevisiae* mutants obtained by ethyl methanesulfonate (EMS) mutagenesis of a haploid strain expressing a yellow fluorescent protein (YFP) under control of the *TDH3* promoter. *One mutant is included in both columns because it was analyzed both by BSA-Seq and Sanger sequencing. **(B-D)** Previously published fluorescence levels (x-axis) and statistical significance of the difference in median fluorescence between each mutant and the un-mutagenized progenitor strain (y-axis) are shown for mutants analyzed in **(B)** Gruber et al. (2012) and **(C,D)** Metzger et al. (2016). **(B)** Collection of 1064 mutants from Gruber et al. (2012) enriched for mutations causing large fluorescence changes. *P*-values were computed using *Z*-tests in this study, based on one measure of fluorescence for each mutant and 30 measures of fluorescence for the progenitor strain. **(C)** Collection of 211 mutants from Metzger et al. (2016) enriched for mutations causing large fluorescence changes. **(D)** Collection of 1498 mutants from Metzger et al. (2016) obtained irrespective of their fluorescence levels (unenriched mutants). **(E)** A new fluorescence dataset for 197 unenriched mutants from Metzger et al. (2016) (blue in panel D) that were reanalyzed in a 2nd screen as part of this study. **(C-E)** 4 replicate populations were analyzed for each mutant. Error bars show 95% confidence intervals of fluorescence levels measured among these replicates. *P*-values were obtained using the permutation tests described in Methods. **(B-E)** Mutants analyzed by BSA-Seq are highlighted in red. All of these mutants showed fluorescence changes greater than 0.01 (vertical dotted lines) and *P*-value below 0.05 (horizontal dotted lines); percentages of all mutants that met these selection criteria in each collection are also shown. Mutants selected for Sanger sequencing of the *ADE4*, *ADE5*, and/or *ADE6* candidate genes are highlighted in green. The mutant analyzed with both BSA-seq and Sanger sequencing is both red and green in panel C). Two mutants selected for Sanger sequencing of the *ADE2* gene are highlighted in purple, one in **D** and one in **E**.

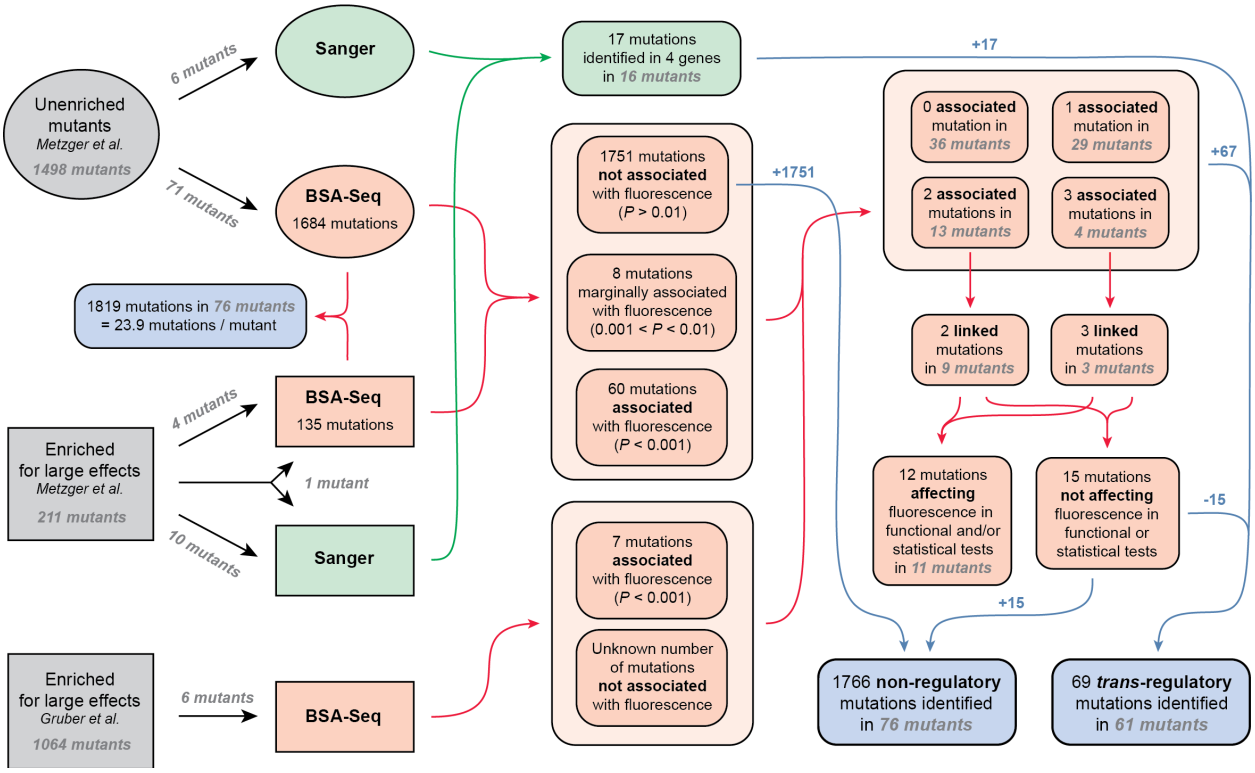

Figure A-2: Diagram showing the number of mutant strains and mutations considered at each step of the study.

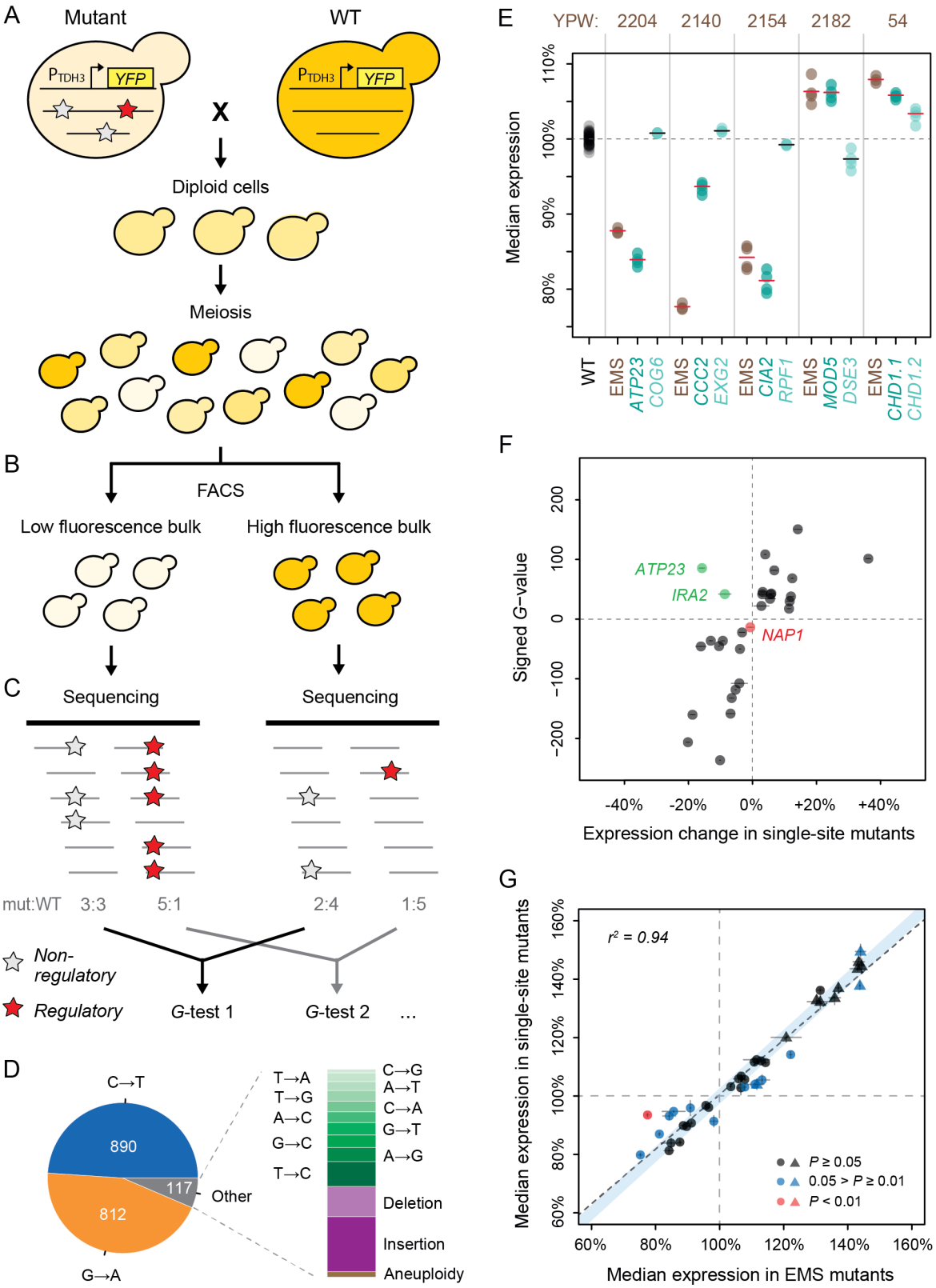

Figure A-3: Genetic mapping and functional testing of trans-regulatory mutations affecting PTDH3-YFP expression.

(A-C) Overview of the BSA-Seq approach. **(A)** Crossing scheme used to map mutations in each EMS mutant strain by crossing to an un-mutagenized strain expressing P_{TDH3}-*YFP*. Stars indicate hypothetical mutations. **(B)** Isolation of two bulks of haploid segregants with high and low fluorescence levels (see Methods). **(C)** Estimation of allele frequencies in each bulk using high-throughput sequencing. A mutation without effect on fluorescence is found at similar frequencies in the two bulks (white stars). A mutation affecting fluorescence or genetically linked to a mutation affecting fluorescence is found at different frequencies between the two bulks (red stars). **(D)** Type of mutations identified in BSA-Seq data for the 76 mutants from Metzger et al. (2016). **(E)** Median expression of P_{TDH3}-*YFP* is shown for the wild-type (WT) progenitor strain (black), for 5 EMS mutants (brown) with two linked mutations associated with fluorescence in BSA-Seq data and for 10 single-site mutants (turquoise) carrying one of the two linked mutations in the 5 EMS mutants. Single-site mutants are grouped in pairs next to the EMS mutant carrying the same mutations and are named after the gene that they affect. Expression levels are expressed relative to the wild-type progenitor strain. For each strain, dots represent the median expression measured for each replicate population and tick marks represent the mean of median expression from replicate populations. **(F)** Effects of mutations associated with fluorescence in BSA-Seq experiments tested in single-site mutants. X-axis: Effect of each mutation on expression measured in a single site mutant and relative to the wild-type progenitor strain. Error bars are 95% confidence intervals obtained from at least 4 replicate populations. Y-axis: *G* statistics of the tests used to compare the frequencies of each mutation between the two bulks in BSA-Seq experiments, with a negative sign if the mutation was more frequent in the low fluorescence bulk and a positive sign if the mutation was more frequent in the high fluorescence bulk. One single-site mutant (*NAP1*, red) showed no significant change in expression relative to the wild-type progenitor strain (*t*-test, *P*-value > 0.05); the mutation it carries is therefore considered to be a false positive in the BSA-seq data. For two other single-site mutants (*ATP23* and *IRA2*, green), the expression changes were not in the same direction as predicted by the signed *G*-values. **(G)** P_{TDH3}-*YFP* expression levels in single-site mutants and in EMS mutants sharing the same mutation. Data points represent median expression levels of 40 EMS mutants (x-axis) and 40 single-site mutants (y-axis) measured by flow cytometry in four replicate populations. Circles: mutations identified by BSA-Seq. Triangles: mutations identified by sequencing candidate genes. Error bars: 95% confidence intervals of expression levels obtained from replicate populations. Data points are colored based on the *P*-values of permutation tests used to assess the statistical significance of expression differences between each single site mutant and the EMS mutant carrying the same mutation (see Figure 2 - figure supplement 5 for details). The light blue area represents the 95% confidence interval of expression differences between genetically identical samples across the whole range of median expression values. This confidence interval was calculated from a null distribution described in Figure 2 - figure supplement 5A. **(E-G)** Expression levels are expressed on a scale linearly related to *YFP* mRNA levels and relative to the median expression of the wild-type progenitor strain (see Methods).

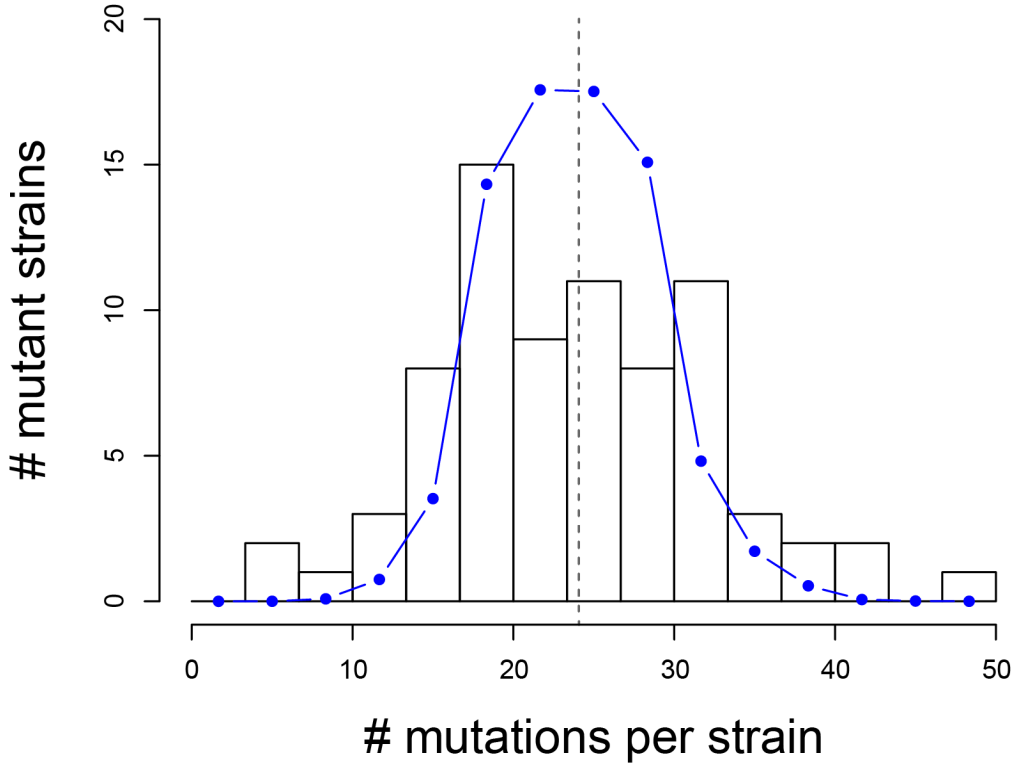

Figure A-4: Number of mutations per strain identified from BSA-Seq data.

Data from 76 EMS mutants from Metzger et al. (2016) are shown. Vertical dotted line: mean number of mutations per strain (23.9). Blue dots and line: Poisson distribution with $\lambda = 23.9$ and $k = 76$ representing the expected numbers of mutations per line if mutations had the same probability of occurring in all mutant lines.

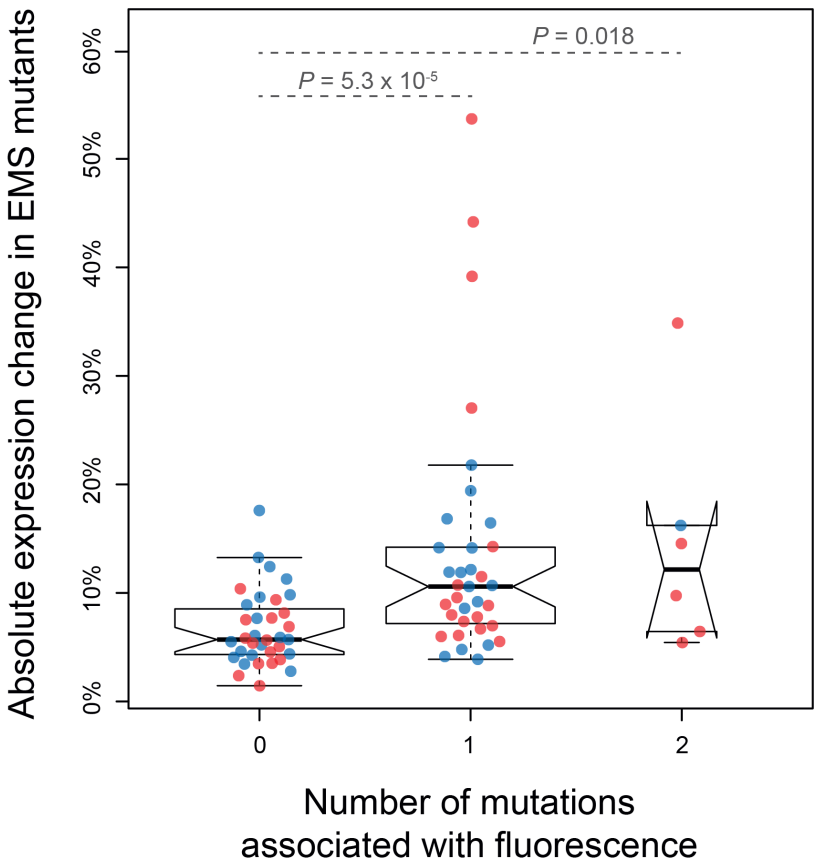

Figure A-5: Magnitude of expression changes in EMS mutants depending on the number of mutations associated with fluorescence in BSA-Seq experiments.

Individual data points represent absolute differences between the median expression levels of EMS mutants and of the un-mutagenized progenitor strain averaged among four replicate populations. Mutations that were associated with fluorescence only because of genetic linkage (i.e., without additional evidence of affecting expression) were not counted (see Supplementary File 4). Blue dots: mutants with decreased expression relative to the progenitor strain. Red dots: mutants with increased expression relative to the progenitor strain. Using Mann-Whitney-Wilcoxon tests, the magnitude of expression changes was found to be significantly lower for mutants without any mutation associated with fluorescence than for mutants with 1 ($P = 5.3 \times 10^{-5}$) or 2 ($P = 0.018$) mutations associated with fluorescence.

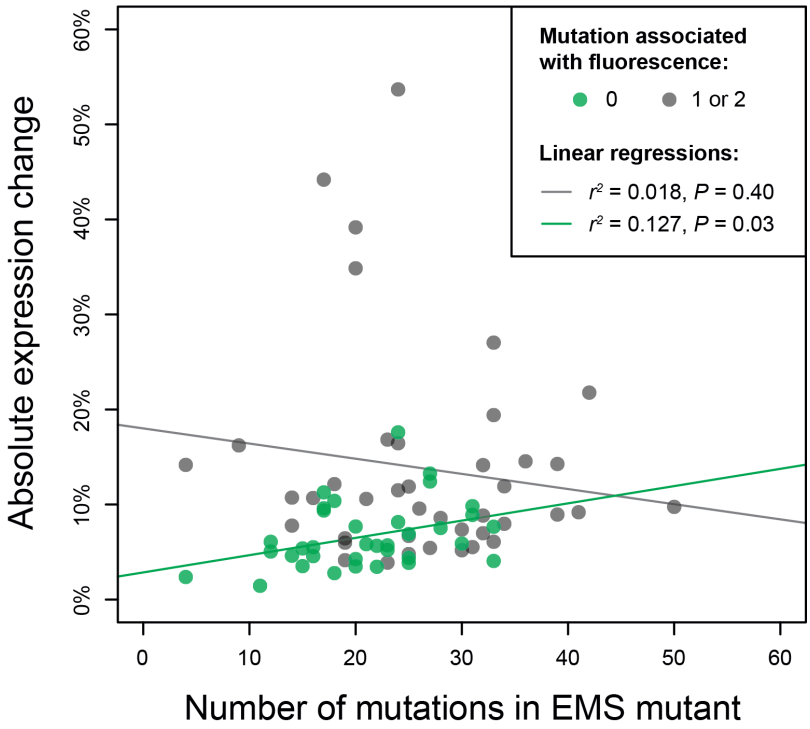

Figure A-6: Relationship between the number of mutations per EMS mutant strain and the absolute expression change relative to the progenitor strain.

This relationship is shown for EMS mutants without any mutation associated with fluorescence in BSA-Seq data (green dots and green regression line) as well as for EMS mutants with at least one mutation associated with fluorescence in BSA-Seq data (gray dots and gray regression line). Mutations that were associated with fluorescence only because of genetic linkage and without other evidence of affecting expression were excluded (see Supplementary File 4). *F*-tests were used to assess the statistical significance of linear regressions. A significant relationship was observed between the number of mutations per mutant strain and the absolute expression change only when no mutation was associated with fluorescence ($r^2 = 0.127$, *P*-value = 0.03). This observation supports the hypothesis that several mutations with small effects could collectively contribute to the expression change observed in mutants for which no mutation was associated with fluorescence. The small effects of these mutations would explain why they were not associated with fluorescence in the BSA-Seq analyses.

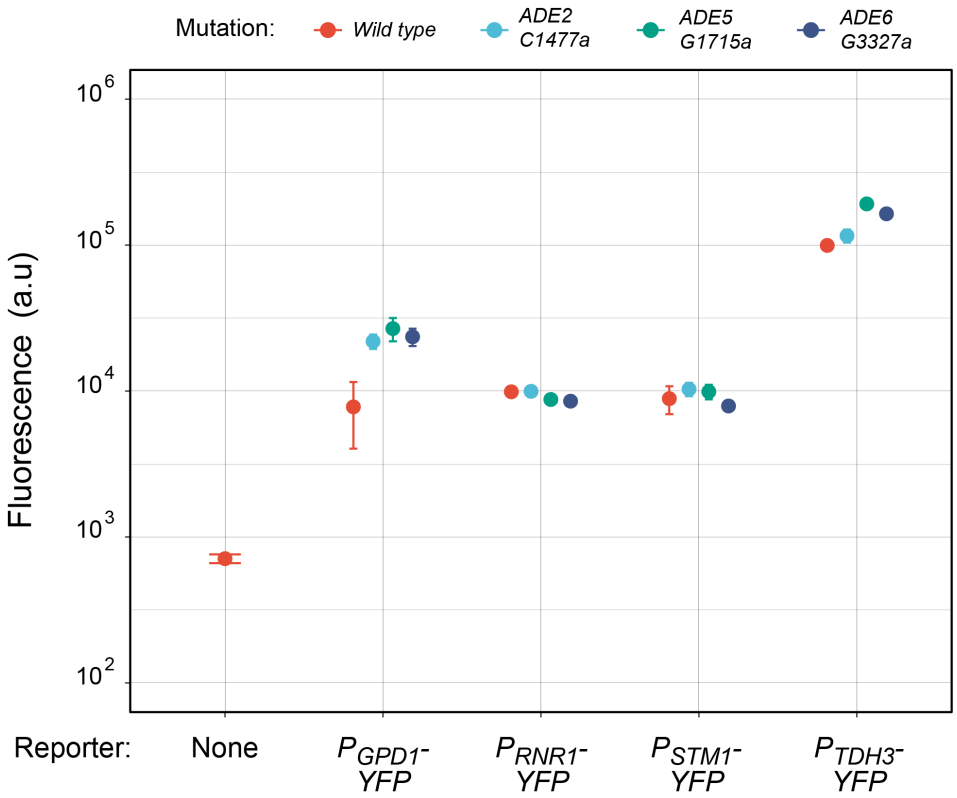

Figure A-7: Effects of individual mutations in purine biosynthesis genes on YFP expression levels differ among promoters.

Each dot indicates the median fluorescence level of at least 5 x 10^4 cells for each genotype averaged across 3 experimental replicates. Error bars represent median absolute deviation across replicates. Dots are grouped along the x-axis based on the yeast promoter used to drive YFP expression (P_{GPD1}, P_{RNR1}, P_{STM1} and P_{TDH3}), with "None" corresponding to the autofluorescence measured in a strain without a fluorescent reporter gene. The color of each dot indicates which mutation was introduced in one of the genes involved in *de novo* purine synthesis (*ADE2*, *ADE5* or *ADE6*), with the specific mutation introduced indicated in the key. The goal of this experiment was to determine whether the regulatory mutations identified in purine synthesis genes altered P_{TDH3}-*YFP* expression at the transcriptional or post-transcriptional level. If the mutations acted post-transcriptionally, their effect on fluorescence level should be the same among strains with different promoters driving YFP expression because they all produce the same *YFP* transcript. However, we observed that the mutations in purine synthesis genes increased fluorescence level when YFP expression was driven by the *TDH3* or the *GPD1* promoter but not when YFP expression was driven by the *RNR1* or the *STM1* promoter, indicating that the effects of these mutations on YFP expression were promoter specific.

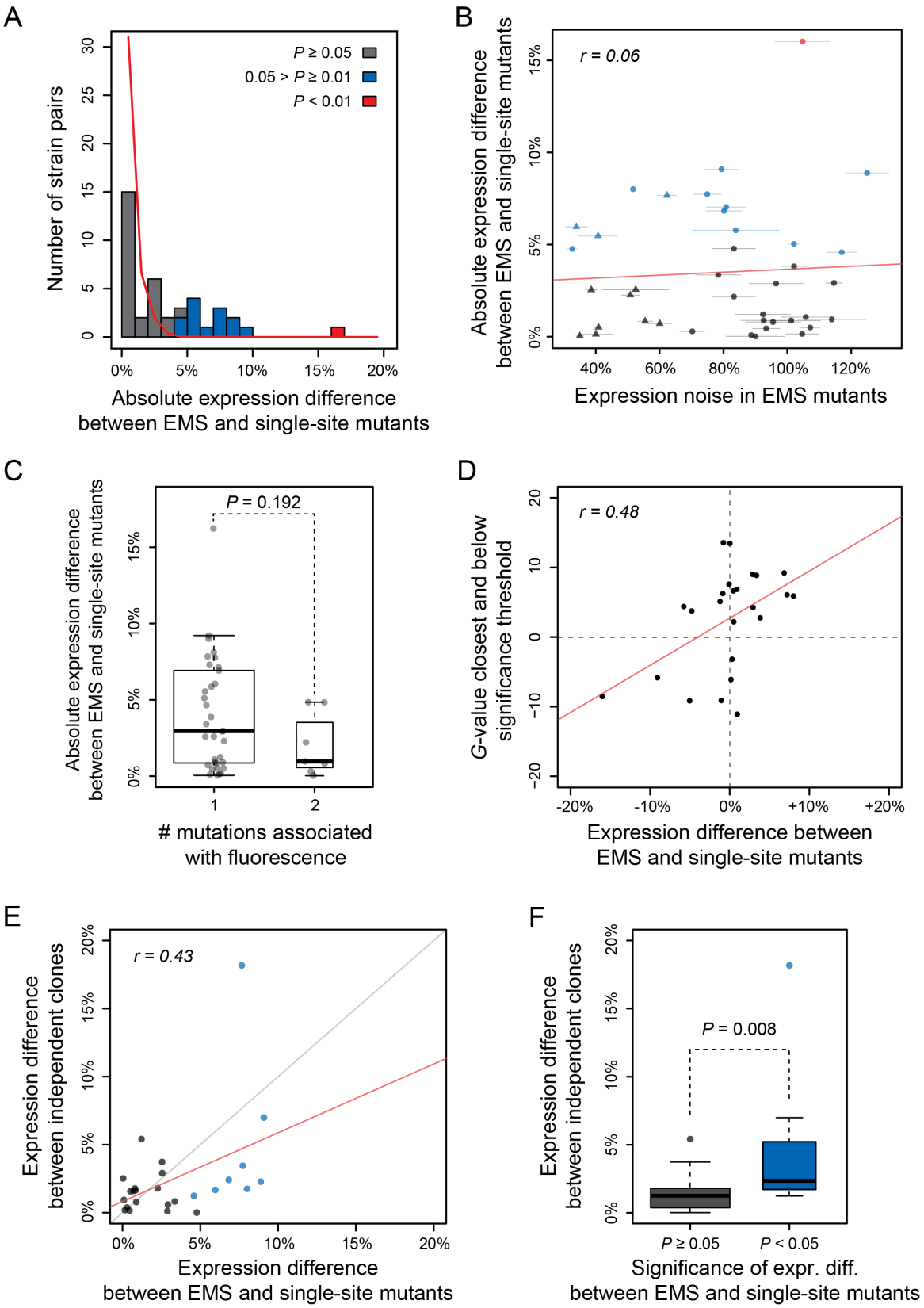

Figure A-8: Factors contributing to expression differences observed between EMS and single-site mutants.

(A) Distribution of absolute expression differences observed between EMS and single-site mutants (bars). To assess the statistical significance of these expression differences, we estimated the magnitude of expression differences expected to arise by chance between genetically identical strains grown at different positions of a 96-well plate (red line). This null distribution was obtained from the differences in expression measured for 10,440 pairs of the un-mutagenized progenitor strain grown at different well positions in four replicate populations. We next randomly permuted 10^5 times the expression values between i) each pair of EMS and single-site mutants and ii) random pairs of the progenitor strain to calculate the one-sided *p*-value for each pair of mutants (*i.e.* the proportion of randomized expression differences greater than the observed expression difference). After Benjamini-Hochberg correction for multiple testing, we found that the expression difference between the single-site mutant and the EMS mutant carrying the same mutation was statistically significant (adjusted *p*-value < 0.05) for 14 out of the 40 pairs of mutants (35%, red and blue bars), but highly significant (adjusted *p*-value < 0.01) for only 1 pair (2.5%, red bar). Because mutant strains were exposed to the same micro-environmental and technical variation as the control samples used to establish the null distribution, these sources of variation are unlikely to explain the significant differences of expression observed between EMS and single-site mutants. Panels **B-F** test three other hypotheses to explain expression differences observed between single-site and EMS mutants. **(B)** Hypothesis 1: expression differences between EMS and single-site mutants are explained by differences in expression noise (*i.e.* the variability of expression observed among genetically identical cells grown in the same environment) among mutants. To test this hypothesis, we compared the expression noise measured by flow cytometry for each EMS mutant (x-axis) to the absolute difference of median expression levels between this EMS mutant and the corresponding single-site mutant (y-axis). We observed no significant correlation between the two parameters (r = 0.06, *P*-value = 0.71), indicating that expression noise is unlikely to explain expression differences between EMS and single-site mutants. Expression noise was calculated for each sample as the standard deviation of expression among cells divided by the median expression and it is reported as the average value among 4 replicate populations relative to the expression noise of the wild-type progenitor strain. Dot colors: *P*-values as shown in panel A. Dot shapes: circles represent mutations identified by BSA-seq; triangles represent mutations identified by sequencing candidate genes. Error bars: 95% confidence intervals calculated from 4 replicate populations. **(C-D)** Hypothesis 2: expression differences between EMS and single-site mutants are explained by additional mutations present in the EMS mutants. **(C)** Testing effects of additional mutations associated with fluorescence: boxplot comparing the magnitude of expression differences when only one mutation was associated with fluorescence and when more than one mutation was associated with fluorescence in BSA-Seq experiments. The fact that no statistical difference was observed between the two classes (Mann-Whitney-Wilcoxon test, *P* = 0.192) suggests that expression differences between EMS and single-site mutants were not likely to be caused by additional mutations associated with fluorescence in the BSA-Seq data. **(D)** Testing effects of additional mutations with statistical support for an association with fluorescence below the significance threshold. Expression difference between EMS and single-site mutants (x-axis) was compared to the highest *G*-value that was below our significance threshold for considering a mutation to be associated with fluorescence in the BSA-Seq data from each mutant (y-axis). A significant correlation was observed between the two parameters (Pearson's *r* = 0.48; *P* = 0.02), suggesting that some mutations with associations below our detection threshold in the BSA-Seq experiments might contribute to expression differences

observed between EMS and single-site mutants. Dots represent individual pairs of EMS and single-site mutants sharing the same mutation (with random jitter). The red line represents the linear regression of the y-axis parameter on the x-axis parameter. **(E-F)** Hypothesis 3: expression differences between EMS and single-site mutants are explained by secondary mutation(s) or epigenetic changes that occurred during construction of single-site mutants. To test this hypothesis, we isolated two independent clones for 26 single-site mutants after transformation of the progenitor strain and measured the expression difference between the two clones. **(E)** A positive correlation was observed between the expression difference between EMS and single-site mutants (x-axis) and the expression difference between the two independent clones for each single-site mutant (y-axis). This positive correlation indicates that mutations with larger expression differences between the single-site and EMS mutants tended to also show larger expression differences between independent transformants. Dot colors: *P*-values as shown in panel A. **(F)** Boxplot also showed that the average magnitude of expression differences between independent clones was higher for single site mutants with a statistically significant expression difference between the single-site and EMS mutant sharing the same mutation (Mann-Whitney-Wilcoxon test, $P = 0.008$). Results from E and F suggest that secondary mutation(s) and/or epigenetic changes that unintentionally occurred in some of the single-site mutant clones likely contributed to expression differences between some EMS and single-site mutants. It is important to emphasize, however, that these expression differences were small in magnitude and that overall the expression level of single-site mutants was strongly correlated with the expression level of EMS mutants (Figure 3).

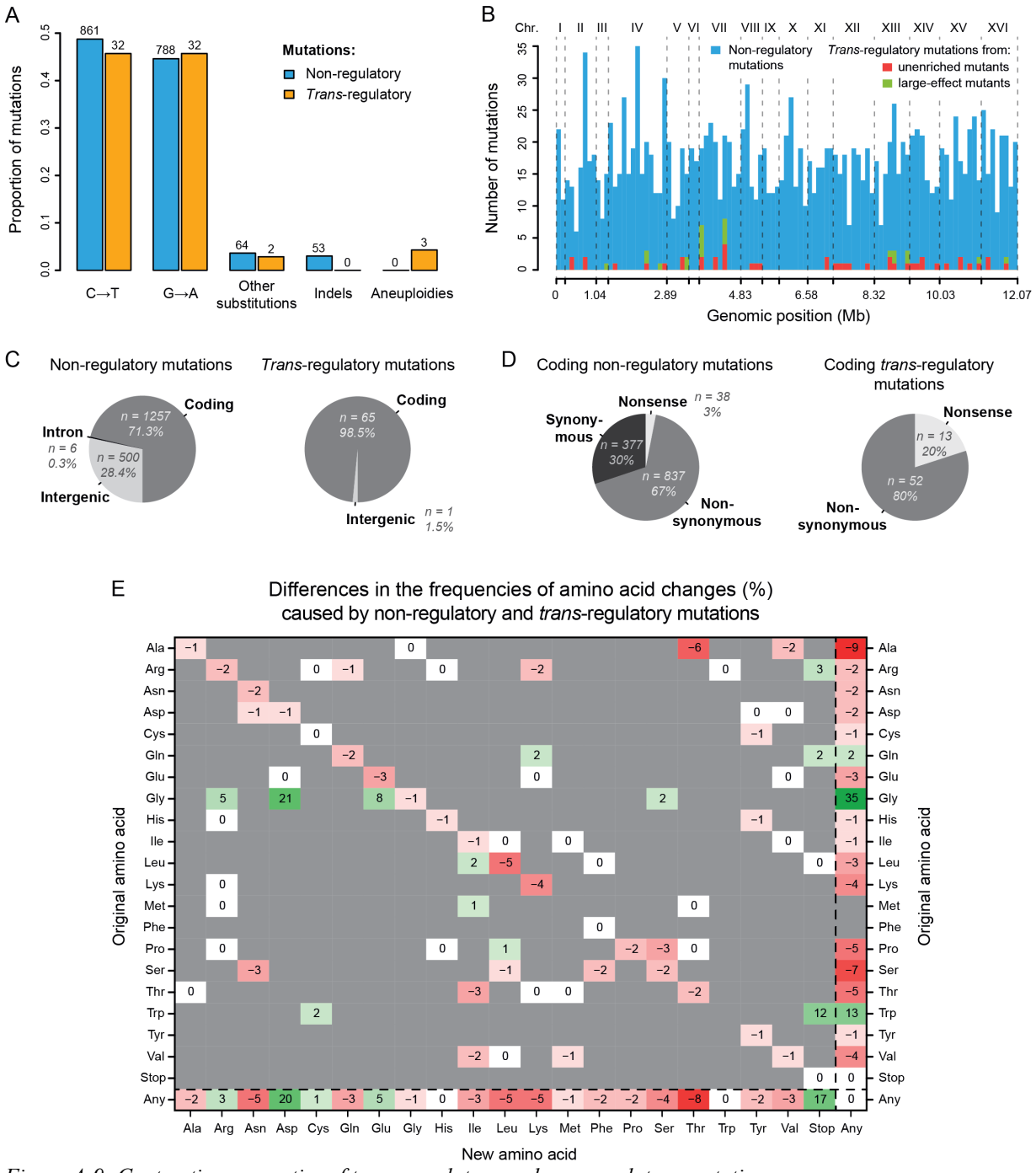

Figure A-9: Contrasting properties of trans-regulatory and non-regulatory mutations.

(A) Proportions of different types of mutations in a set of 1766 non-regulatory mutations (blue) and in a set of 69 *trans*-regulatory mutations (orange). Numbers of mutations are indicated above bars. **(B)** Distributions of non-regulatory and *trans*-regulatory point mutations along the yeast genome. 1766 non-regulatory mutations are shown in blue, 44 *trans*-regulatory mutations that were identified from the collections of unenriched mutants in Metzger et al. (2016) are shown in red and 22 *trans*-regulatory mutations that were identified from the collections of mutants enriched for large expression changes in Gruber et al. (2012) and in Metzger et al. (2016) are

shown in green. **(C)** Proportions of non-regulatory (left) and *trans*-regulatory (right) mutations affecting either coding sequences, introns or intergenic regions. **(D)** Proportions of coding non-regulatory (left) and coding *trans*-regulatory (right) mutations that either introduce an early stop codon (nonsense), that substitute one amino acid for another (nonsynonymous) or that do not change the amino acid sequence (synonymous). **(E)** Frequency of all amino acid changes induced by *trans*-regulatory mutations as compared to non-regulatory mutations. Each entry of the table represents the difference of frequency (percentage) between non-regulatory and *trans*-regulatory mutations that are changing the amino acid shown on the y-axis into the amino acid shown on the x-axis. For instance, the -6 on the first row indicates that the proportion of mutations changing an Alanine into a Threonine is 6% lower among *trans*-regulatory mutations than among non-regulatory mutations. Shades of red: amino acid changes underrepresented in the set of *trans*-regulatory mutations. Shades of green: amino acid changes overrepresented in the set of *trans*-regulatory mutations. White: amino acid changes equally represented in the *trans*-regulatory and non-regulatory sets of mutations. Grey: amino acid changes not observed in the sets of *trans*-regulatory and non-regulatory mutations. **(B-E)** The three aneuploidies were excluded for these plots. **(D,E)** Non-coding mutations were excluded for these plots.

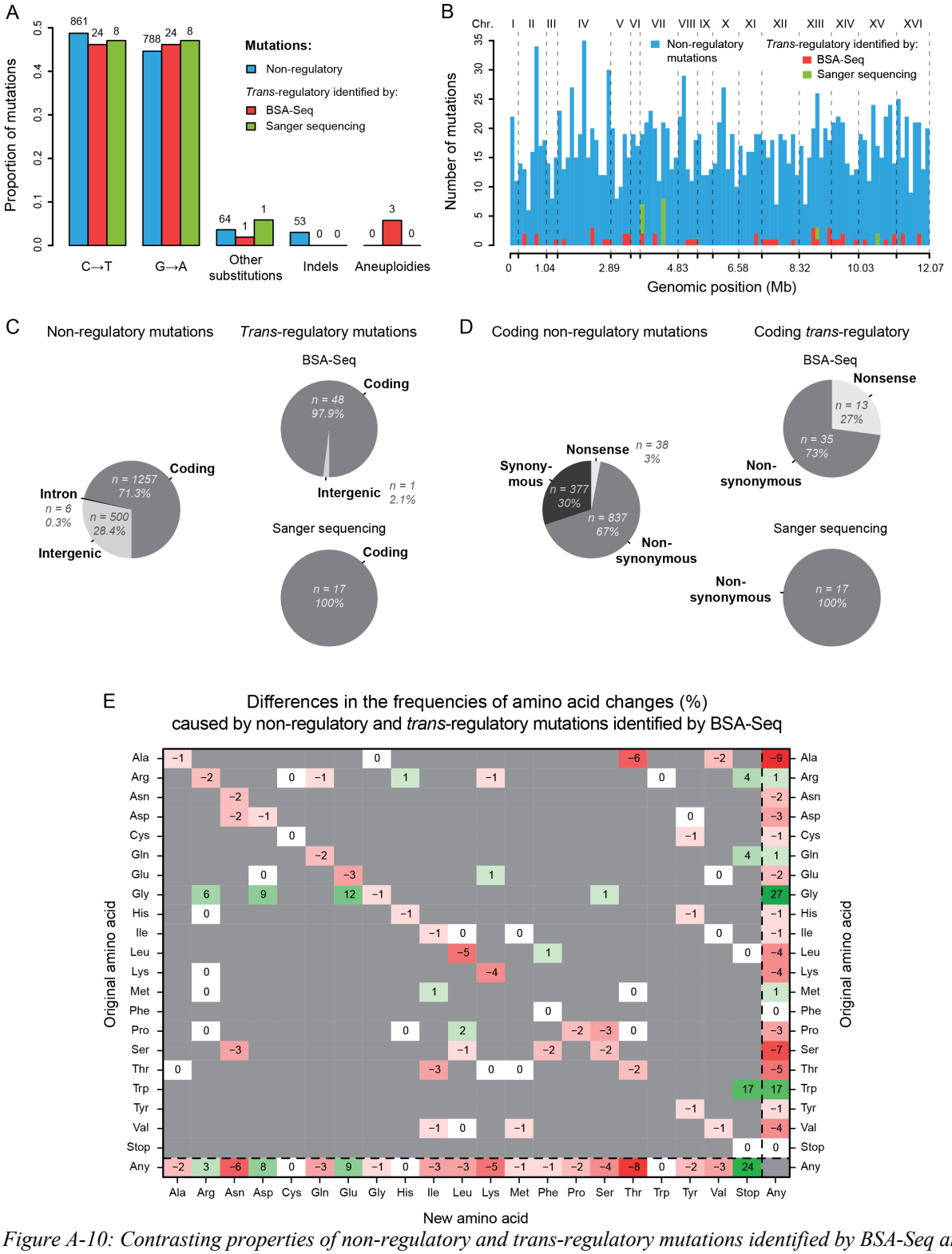

Figure A-10: Contrasting properties of non-regulatory and trans-regulatory mutations identified by BSA-Seq and of trans-regulatory mutations identified by Sanger sequencing of candidate genes.

(A) Proportions of different types of mutations observed among 1766 non-regulatory mutations (blue), among 52 *trans*-regulatory mutations identified by BSA-Seq (red) and among 17 trans-

regulatory mutations identified by Sanger sequencing of candidate genes. Numbers of mutations are indicated above bars. **(B)** Distributions of non-regulatory and *trans*-regulatory point mutations along the yeast genome. 1766 non-regulatory mutations are shown in blue, 49 *trans*-regulatory mutations that were identified by BSA-Seq are shown in red and 17 *trans*-regulatory mutations that were identified by Sanger sequencing are shown in green. **(C)** Proportions of non-regulatory mutations (left), *trans*-regulatory mutations identified by BSA-Seq (upper right) and *trans*-regulatory mutations identified by Sanger sequencing (bottom right) that affect either coding sequences, introns or intergenic regions. **(D)** Proportions of coding non-regulatory mutations (left), coding *trans*-regulatory mutations identified by BSA-Seq (upper right) and coding *trans*-regulatory mutations identified by Sanger sequencing (bottom right) that either introduce an early stop codon (nonsense), that substitute one amino acid for another (nonsynonymous) or that do not change the amino acid sequence (synonymous). **(E)** Frequency of all amino acid changes induced by *trans*-regulatory mutations identified by BSA-Seq as compared to non-regulatory mutations. Each entry of the table represents the difference of frequency (percentage) between non-regulatory and *trans*-regulatory mutations that are changing the amino acid shown on the y-axis into the amino acid shown on the x-axis. Shades of red: amino acid changes underrepresented in the set of *trans*-regulatory mutations identified by BSA-Seq. Shades of green: amino acid changes overrepresented in the set of *trans*-regulatory mutations identified by BSA-Seq. White: amino acid changes equally represented in the *trans*-regulatory and non-regulatory sets of mutations. Grey: amino acid changes not observed in the sets of *trans*-regulatory and non-regulatory mutations. **(B-E)** The three aneuploidies were excluded for these plots. **(D,E)** Non-coding mutations were excluded for these plots.

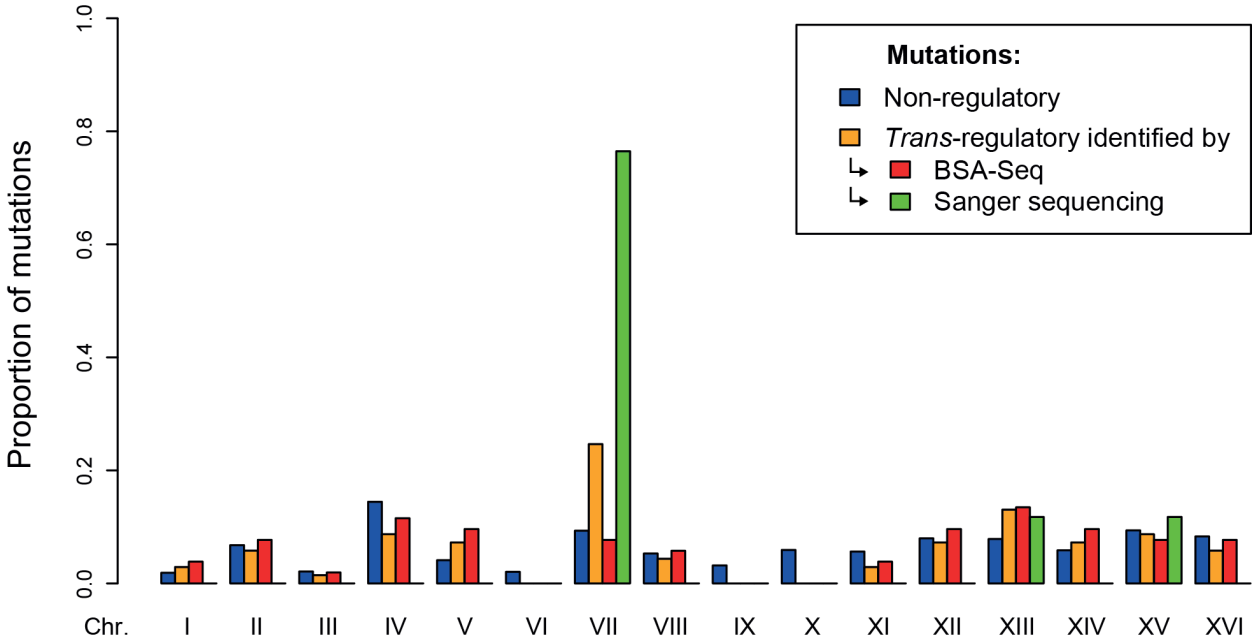

Figure A-11: Distributions of trans-regulatory and non-regulatory mutations among chromosomes.

1766 non-regulatory mutations are shown in blue and 69 *trans*-regulatory mutations are shown in orange, among which 52 mutations were identified by BSA-Seq (shown in red) and 17 mutations were identified by Sanger sequencing of candidate genes (shown in green). *Trans*-regulatory mutations were significantly enriched on chromosome VII that contained the purine biosynthesis genes *ADE5* and *ADE6* in which several mutations were identified (24.3% of *trans*-regulatory mutations located on chromosome VII *vs* 9.3% of non-regulatory mutations; *G*-test, $P = 3.4 \times 10^{-4}$). *Trans*-regulatory mutations were also enriched on chromosome XIII that contained the purine synthesis gene *ADE4*, although this enrichment was not statistically significant (13.0% of *trans*-regulatory mutations located on chromosome XIII *vs* 7.8% of non-regulatory mutations; *G*-test, $P = 0.15$).

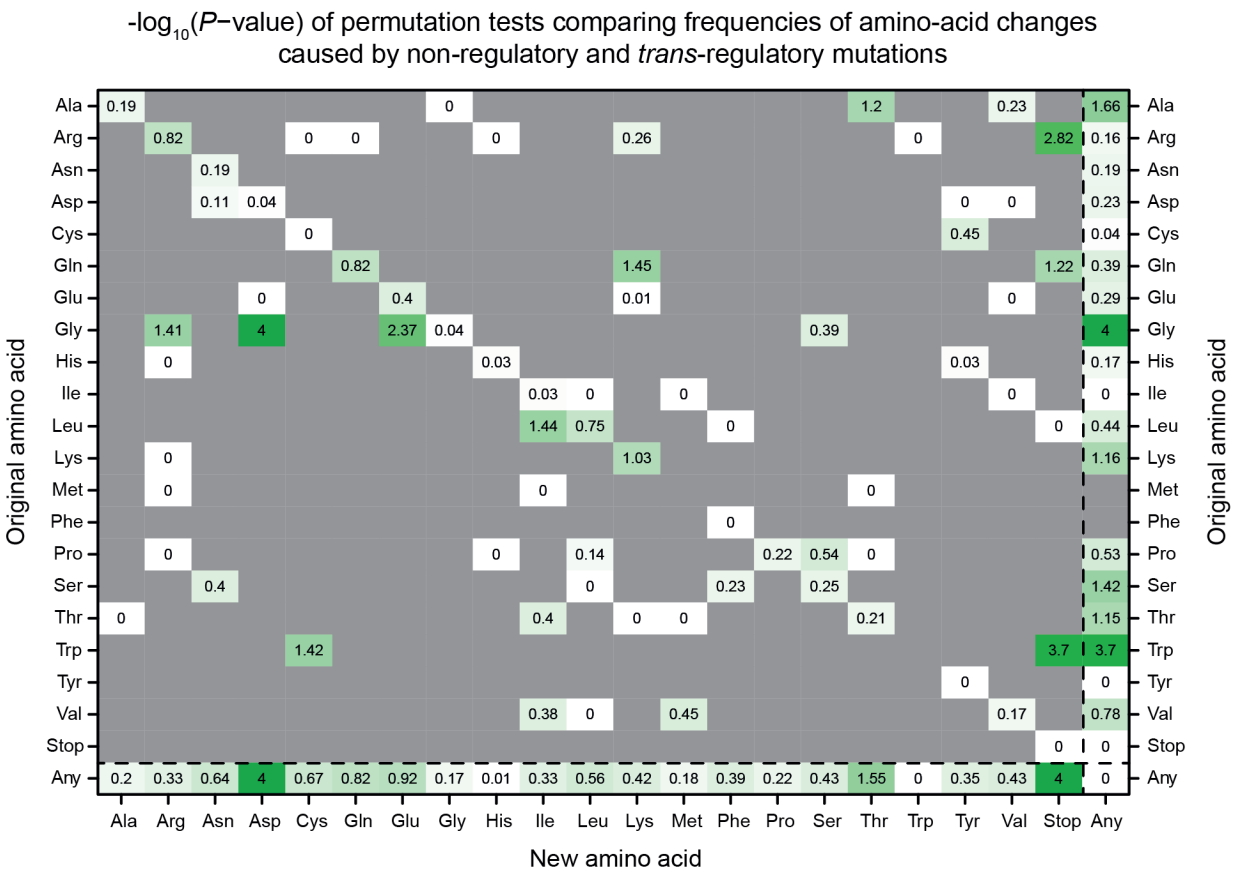

Figure A-12: Statistical significance of the enrichment and depletion of amino acid changes induced by trans-regulatory mutations.

Permutations tests were used to assess the statistical significance of the frequency differences between non-regulatory and *trans*-regulatory mutations shown on Figure 3E. Each number represents the negative logarithm (base-10) of the *P*-value obtained using a permutation test to compare the frequency of changing the amino acid on the y-axis to the amino acid shown on the x-axis between non-regulatory and *trans*-regulatory mutations. Green color intensity scales with the negative logarithm of *P*-values. White: amino acid changes equally represented in the *trans*-regulatory and non-regulatory sets of mutations. Grey: amino acid changes not observed in the sets of *trans*-regulatory and non-regulatory mutations.

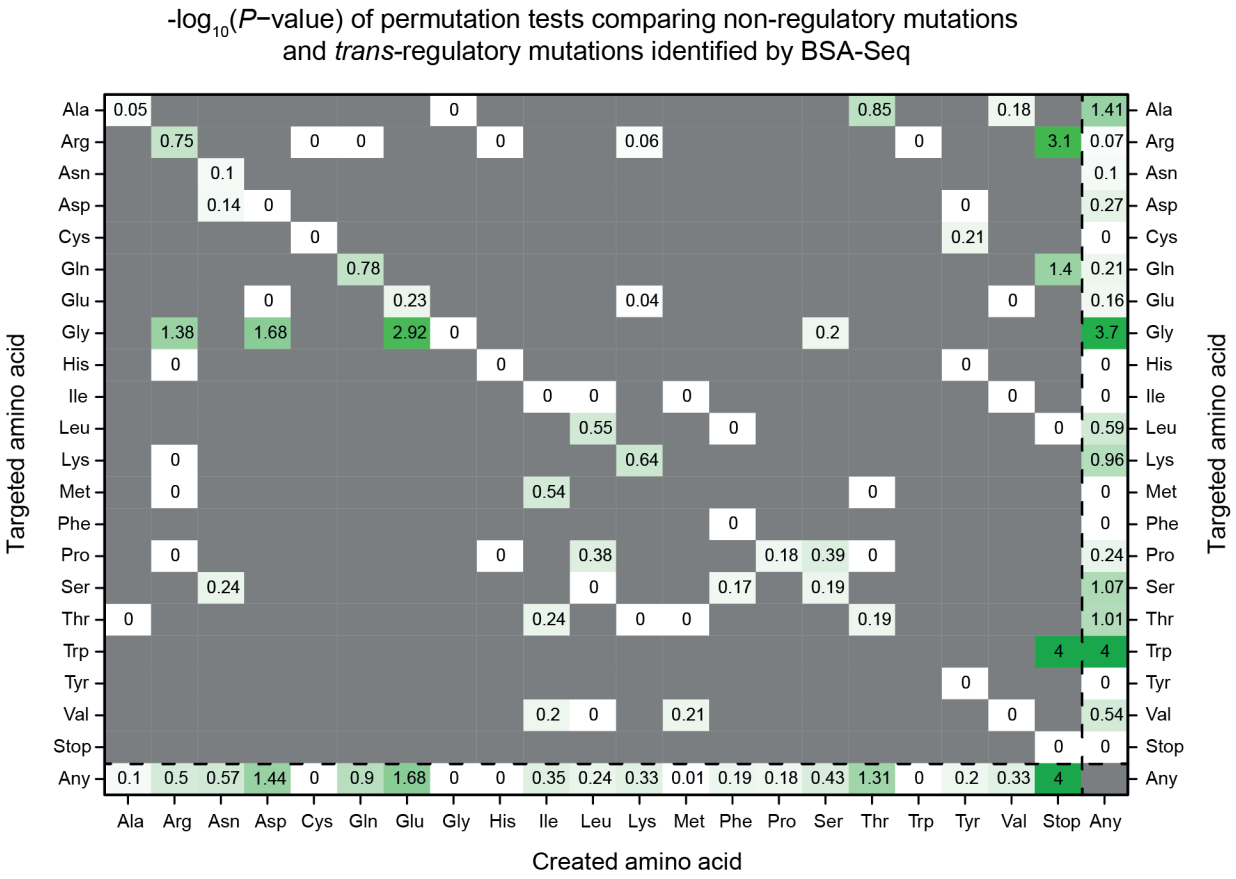

Figure A-13: Statistical significance of the enrichment and depletion of amino acid changes induced by trans-regulatory mutations identified by BSA-Seq.

Permutations tests were used to assess the statistical significance of the frequency differences between non-regulatory and *trans*-regulatory mutations shown on Figure 3 - figure supplement 1E. Each number represents the negative logarithm (base-10) of the *P*-value obtained using a permutation test to compare the frequency of changing the amino acid on the y-axis to the amino acid shown on the x-axis between non-regulatory and *trans*-regulatory mutations. Green color intensity scales with the negative logarithm of *P*-values. White: amino acid changes equally represented in the *trans*-regulatory and non-regulatory sets of mutations. Grey: amino acid changes not observed in the sets of *trans*-regulatory and non-regulatory mutations.

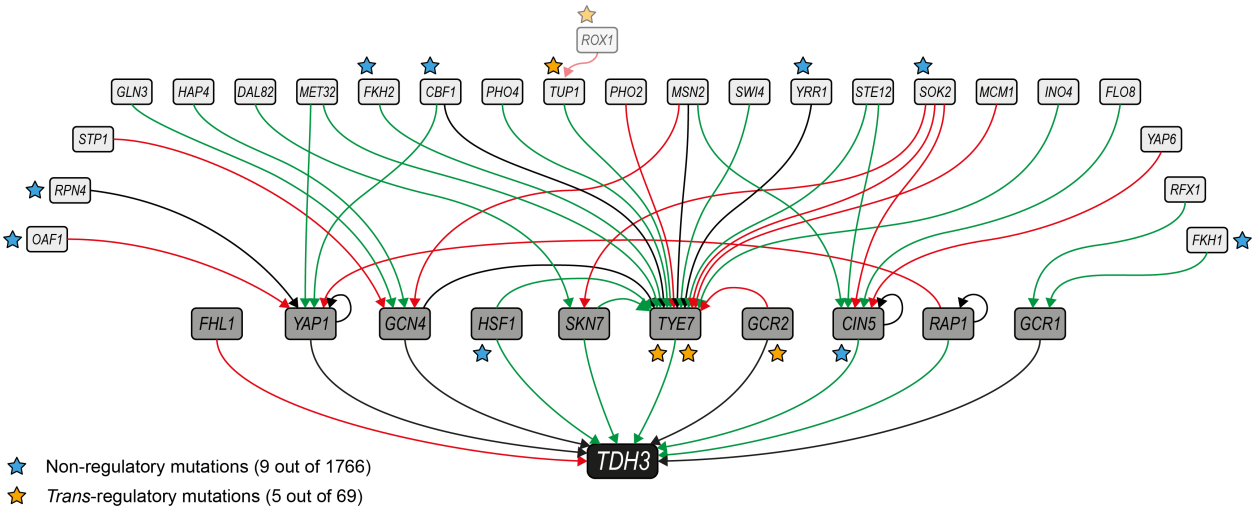

Figure A-14: Mutations mapping to a predicted TDH3 regulatory network.

The network of inferred interactions between *TDH3* and transcription factors regulating its expression was established using the YEASTRACT repository (Teixeira et al., 2018). First level regulators (dark grey boxes) are transcription factors with evidence of binding to the *TDH3* promoter and regulating its expression. Second level regulators (light grey boxes) are transcription factors with evidence of binding to the promoter of at least one first level regulator and regulating its expression. Green arrows: evidence for activation of expression. Red arrows: evidence for inhibition of expression. Black arrows: unknown direction of regulation. Non-regulatory and *trans*-regulatory mutations identified in the network are represented by blue and orange stars, respectively, near the affected genes. *ROX1*, inferred to be a third level regulator, is also shown because a *trans*-regulatory mutation was identified in its coding sequence.

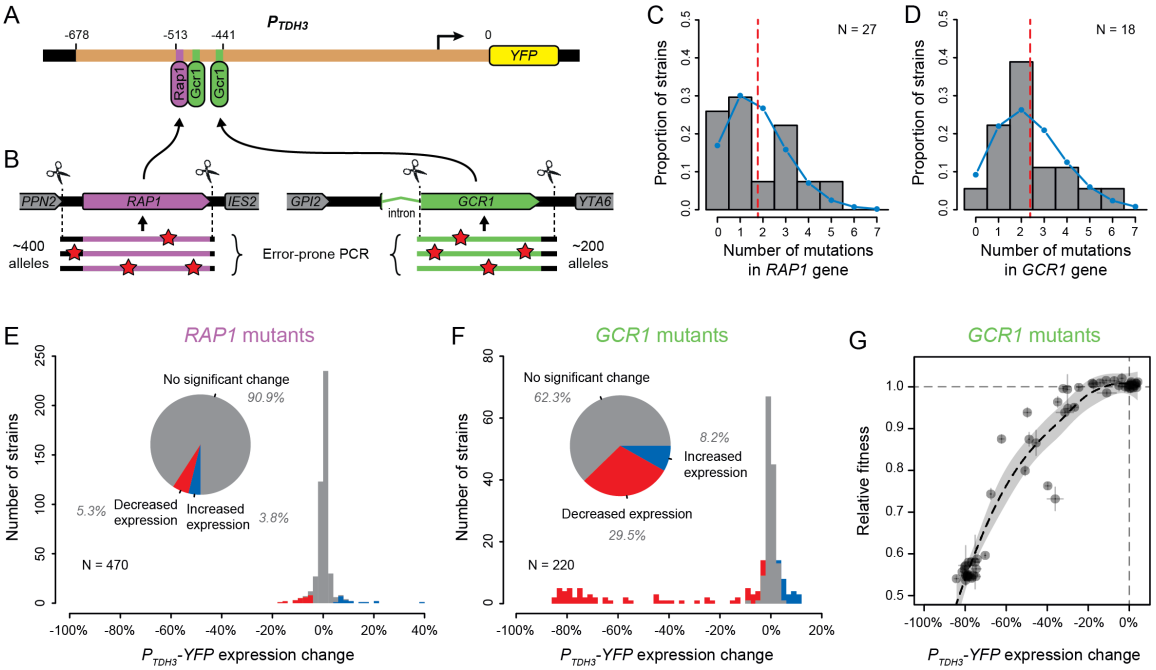

Figure A-15: Impact of mutations in two direct regulators of the TDH3 promoter.

(A) Schematics of the P_{TDH3}-*YFP* reporter gene with locations of three known binding sites for transcription factors Rap1p (purple) and Gcr1p (green) shown in the *TDH3* promoter. **(B)** Regions of *RAP1* (purple) and *GCR1* (green) genes that were subjected to random mutagenesis using error-prone PCR. 470 *RAP1* mutants and 220 *GCR1* mutants were obtained by integration of random PCR fragments at the native *RAP1* or *GCR1* loci using CRISPR/Cas9 allelic replacement. **(C-D)** Distributions of the number of mutations per strain identified by Sanger sequencing the mutated regions of **(C)** *RAP1* in 27 strains or **(D)** *GCR1* in 18 strains. These data are shown in histograms. Blue curves: Poisson distribution with the same mean as observed in data. Red dotted line: Mean number of mutations among sequenced strains. **(E-F)** Distributions of P_{TDH3}-*YFP* expression changes relative to the un-mutagenized reporter strain measured in four replicate samples for **(E)** the 470 *RAP1* mutants or **(F)** the 220 *GCR1* mutants. Fluorescence measures were transformed to be linearly related with *YFP* mRNA levels (see Methods). Red bars: Mutants with significant decrease in median expression greater than 3% relative to the un-mutagenized strain (permutation test, $P < 0.05$). Blue bars: Mutants with significant increase in median expression greater than 3% relative to the un-mutagenized strain (permutation test, $P < 0.05$). Pie charts: Proportions of mutants with significant increase in expression (blue), significant decrease in expression (red) and no significant change in expression (gray) relative to the un-mutagenized strain. **(G)** Relationship between changes in P_{TDH3}-*YFP* expression levels (x-axis) and fitness (y-axis) measured in 62 *GCR1* mutants. Expression changes and fitness are both expressed relative to the un-mutagenized strain. Gray dotted lines: Expression change and fitness of the un-mutagenized strain. Error bars: 95% confidence intervals of expression changes and fitness measures obtained from four replicate populations of each mutant. The black dotted line represents a LOESS regression of fitness on median expression with a smoothing parameter of 1 and 95% confidence intervals of the estimates shown as a gray shaded area.

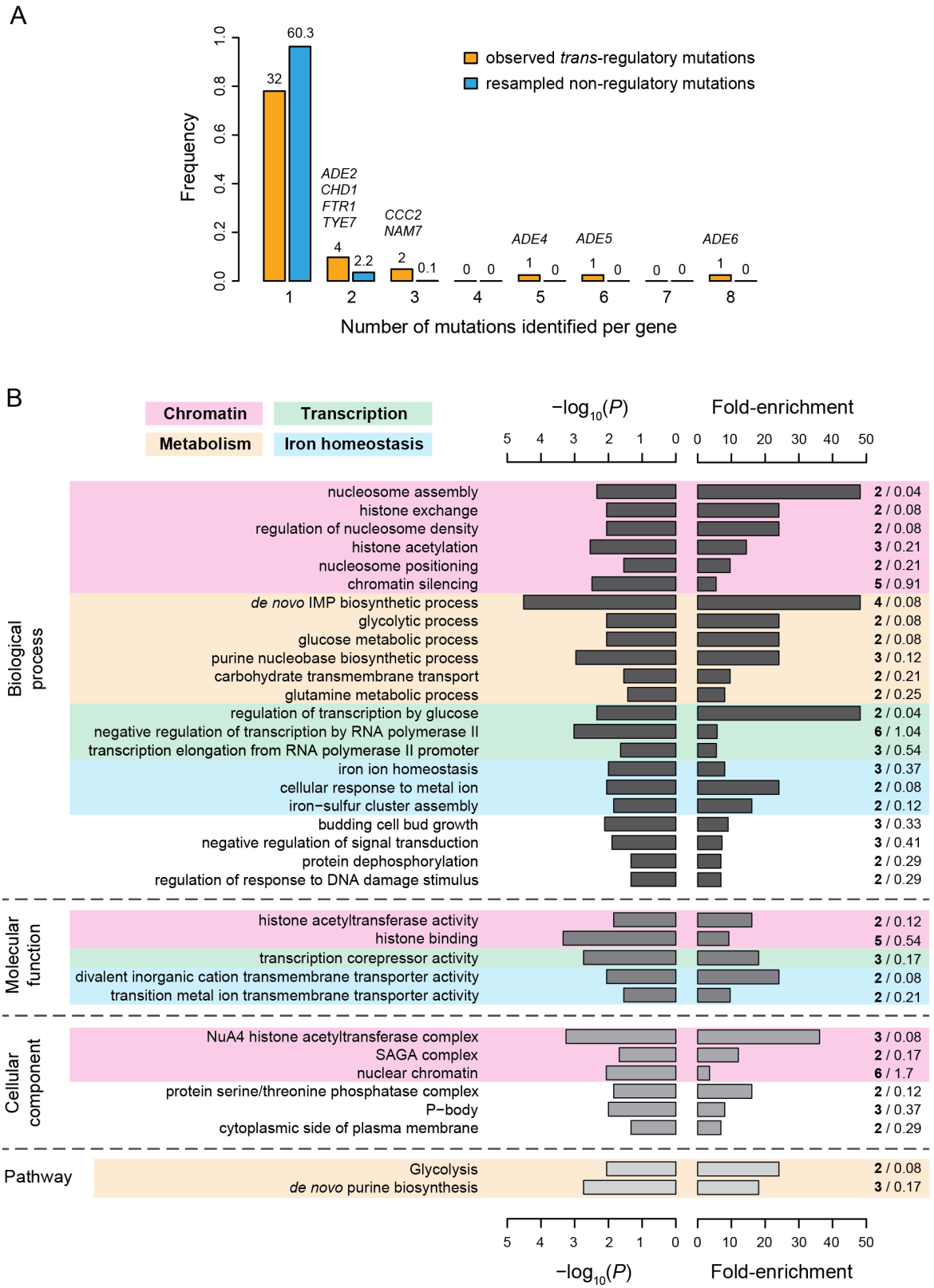

Figure A-16: Properties of genes with coding mutations altering PTDH3-YFP expression level.

(A) Proportion of genes with one or more mutations identified among EMS mutants. Mutations in intergenic regions were excluded from this analysis. Orange bars include genes harboring one or more of the 65 *trans*-regulatory mutations identified in coding sequences. Blue bars include genes harboring one or more of 65 non-regulatory mutations randomly chosen among the set of 1095 non-regulatory mutations observed in coding sequences. The number of genes hit by 1 to 8 mutations is indicated above the corresponding bar. For blue bars, this number represents the mean number of genes obtained from 1000 random sets of 65 non-regulatory mutations. The names of genes with at least 2 *trans*-regulatory mutations identified among mutants are indicated above the bars. *FTR1* and *CCC2* are involved in iron homeostasis, *ADE2,4,5,6* are involved in *de novo* purine biosynthesis, *NAM7* is involved in nonsense mediated mRNA decay, *CHD1* is involved in chromatin regulation and *TYE7* encodes a transcription factor regulating *TDH3* expression. **(B)** Summary of gene ontology (GO) enrichment analysis performed with PANTHER tool (http://www.pantherdb.org/). Fisher's exact tests were used to evaluate the overrepresentation of GO terms among the 42 genes affected by one or more of the 66 *trans*-regulatory mutations in coding sequences relative to the 1043 genes affected by one or more of the 1251 non-regulatory mutations in coding sequences. The descriptions shown on the left correspond to GO terms with a P value < 0.05 (left bars), a fold-enrichment > 3 (right bars) and that are not parents to other GO terms in the ontology hierarchy (*i.e.* GO terms that are the most specific). A more complete list of enriched GO terms can be found in Supplementary File 8. Shades of gray represent different categories of GO terms (from darkest to lightest: biological processes, molecular functions and cellular components) or PANTHER pathways (lightest gray). Fold-enrichment was calculated as the observed number of genes with a particular GO term in the set of genes affected by *trans*-regulatory mutations (bold numbers on the right) divided by an expected number of genes obtained from the number of genes with the same GO term in the set of genes affected by non-regulatory mutations (regular numbers on the right). Four groups of GO terms and pathways involved in similar processes are represented by colored areas: chromatin (pink), metabolism (orange), transcription (green) and iron homeostasis (blue).

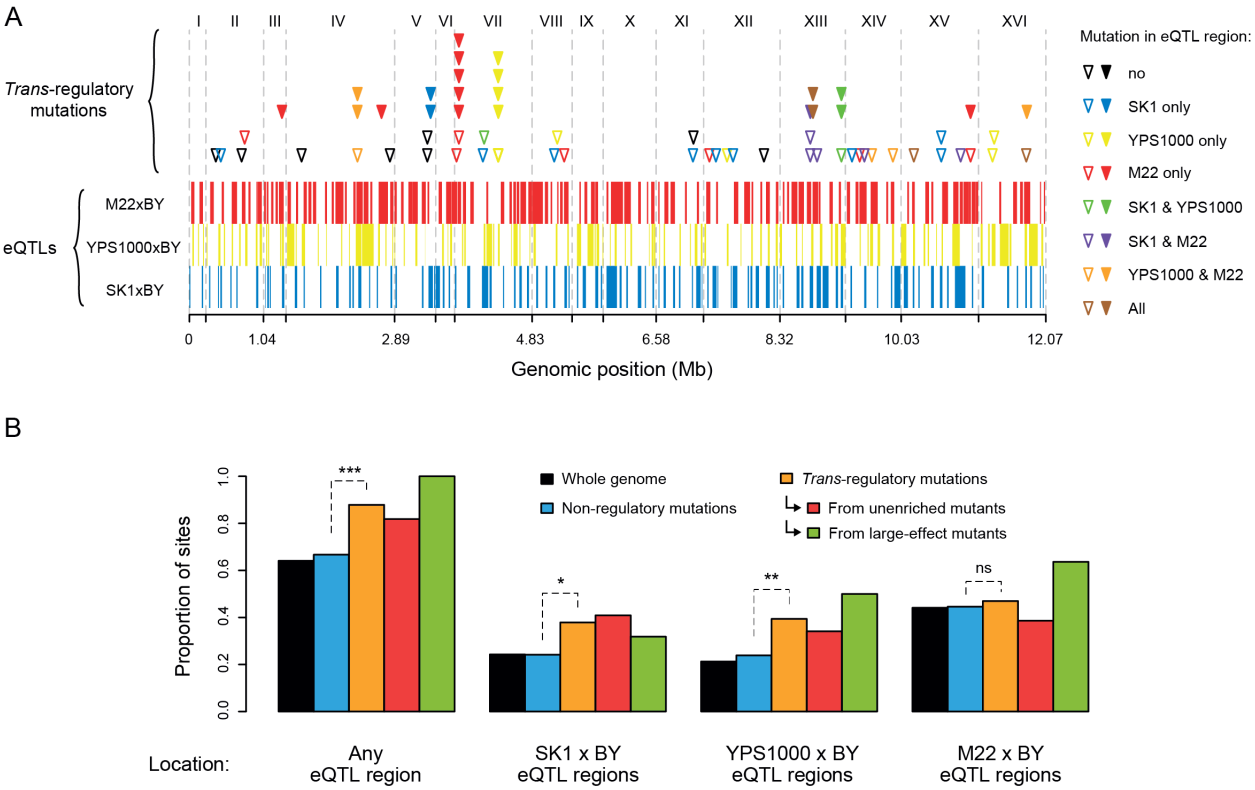

Figure A-17: Overrepresentation of trans-regulatory mutations in eQTLs regions.

(A) Overlap of 66 *trans*-regulatory point mutations and 317 eQTL regions along the yeast genome. eQTL regions were identified by BSA-Seq in Metzger and Wittkopp 2019 from three crosses of a laboratory strain (BY) to each of three strains expressing P_{TDH3}-*YFP* in the genetic background of different *S. cerevisiae* isolates: SK1 (eQTL regions represented by blue bars), YPS1000 (eQTL regions represented by yellow bars) and M22 (eQTL regions represented by red bars). Triangles indicate the genomic locations of *trans*-regulatory mutations, with open triangles representing mutations identified in mutants from the unenriched collection and filled triangles representing mutations identified in mutants enriched for large effects. Triangles are colored depending on the overlap between mutations and eQTL regions: black if the mutation is outside of any eQTL region, blue if the mutation lies in an eQTL region only identified from SK1xBY, yellow if the mutation lies in an eQTL region only identified from YPS1000xBY, red if the mutation lies in an eQTL region only identified from M22xBY, green if the mutation lies in two overlapping eQTL regions identified from SK1xBY and YPS1000xBY, purple if the mutation lies in two overlapping eQTL regions identified from SK1xBY and M22xBY, orange if the mutation lies in two overlapping eQTL regions identified from M22xBY and YPS1000xBY and brown if the mutation lies in three overlapping eQTL regions identified from the three crosses. **(B)** Proportions of non-regulatory and *trans*-regulatory mutations located in eQTL regions. Black bars: proportions of sites among the 12.07 Mb yeast genome. Blue bars: proportions of the 1759 non-regulatory point mutations. Orange bars: proportions of the 66 *trans*-regulatory mutations (excluding aneuploidies). Red bars: proportions of the 44 *trans*-regulatory mutations identified in mutants from the unenriched collection. Green bars: proportions of the 22 *trans*-regulatory mutations identified in mutants enriched for large effects. The proportions of non-regulatory and *trans*-regulatory mutations in eQTL regions were compared using *G*-tests (***: $P < 0.001$, **: $0.001 < P < 0.01$, *: $0.01 < P < 0.05$, ns: $P > 0.05$).

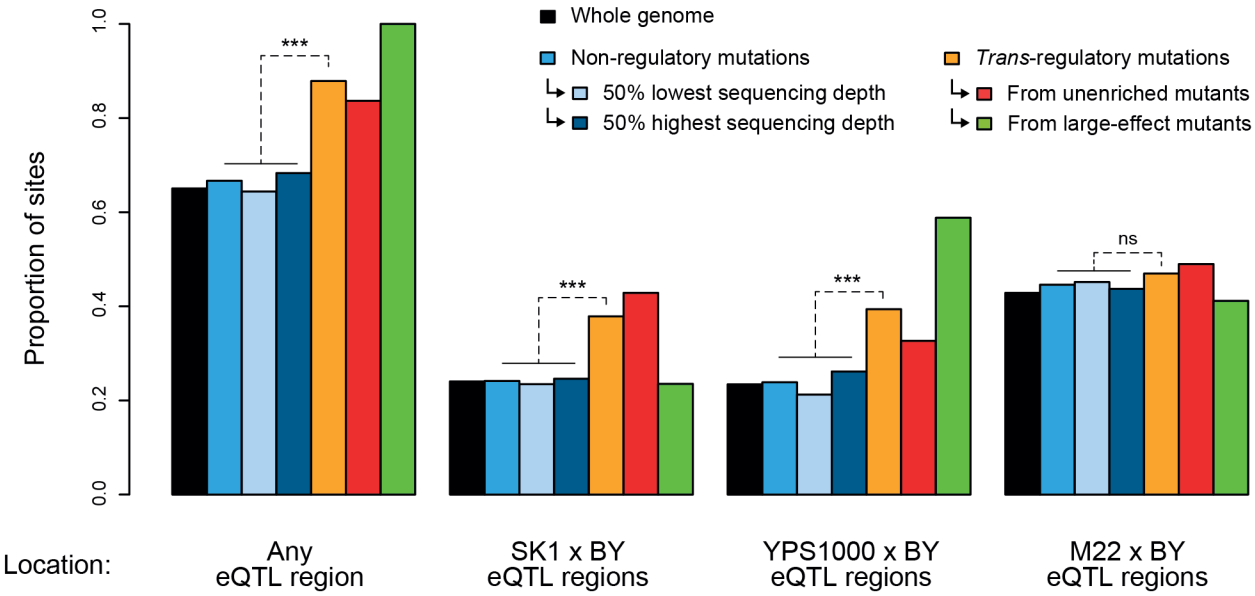

Figure A-18: Proportions of different categories of non-regulatory mutations and trans-regulatory mutations located in eQTLs regions.

Black bars: proportions of all sites among the 12.07 Mb yeast genome. Medium blue bars: proportions of the 1759 non-regulatory point mutations. Light blue bars: proportions of non-regulatory mutations at sites for which the total sequencing depth was below the median sequencing depth of the corresponding library in BSA-Seq data. Dark blue bars: proportions of non-regulatory mutations at sites for which the total sequencing depth was equal or above the median sequencing depth of the corresponding library in BSA-Seq data. Orange bars: proportions of the 66 *trans*-regulatory mutations (excluding aneuploidies). Red bars: proportions of the 49 *trans*-regulatory mutations identified by BSA-Seq. Green bars: proportions of the 17 *trans*-regulatory mutations identified by Sanger sequencing of candidate genes. The proportions of non-regulatory and *trans*-regulatory mutations in eQTL regions were compared using *G*-tests (***: $P < 0.001$, **: $0.001 < P < 0.01$, *: $0.01 < P < 0.05$, ns: $P > 0.05$).

Supplementary Data

All supplementary data files are available at eLife online: https://elifesciences.org/articles/67806

SupplementaryFiles.zip. Compressed folder containing Supplementary Files 1-16.

Supplementary File 1. Sequencing depth in BSA-seq data.

Supplementary File 2. List of all mutations identified by BSA-Seq or Sanger sequencing in this study.

Supplementary File 3. Statistical associations between aneuploidies and fluorescence level.

Supplementary File 4. Linked mutations associated with fluorescence level in BSA-Seq experiments.

Supplementary File 5. Mutations identified by Sanger sequencing of candidate genes.

Supplementary File 6. Mutations tested in single-site mutants.

Supplementary File 7. Mutations associated with fluorescence level in BSA-Seq experiments.

Supplementary File 8. Targeted mutagenesis of RAP1 residues making direct contact with DNA.

Supplementary File 9. List of GO terms overrepresented in genes hit by causative mutations relative to genes hit by neutral mutations.

Supplementary File 10. Mutations located in the coding sequence of glucose signaling genes.

Supplementary File 11. *Trans*-regulatory effects of mutations in purine biosynthesis genes or iron homeostasis genes.

Supplementary File 12. Files used as inputs for analyses performed with the PBS script (Source Code 4) and R scripts (Source Code 1-3).

Supplementary File 13. List of DNA libraries grouped by sequencing runs.

Supplementary File 14. List of oligonucleotides used in this study.

Supplementary File 15. Construction of single-site mutant strains.

Supplementary File 16. Phenotypes of RAP1 mutants (expression) and GCR1 mutants (expression and fitness).

Source Code 1. R scripts used for the analysis of flow cytometry data.

Source Code 2. R scripts used for the analysis of BSA-Seq data and for comparing the properties of *trans*-regulatory and non-regulatory mutations.

Source Code 3. R script used to annotate variants identified in BSA-Seq data.

Source Code 4. PBS script used to process FASTQ files.

SourceData.bz2. Compressed folder including 34 Source Data files in .txt format that contain quantitative data displayed on all figures.