The Characteristics of Genetic Variants that Impact Gene Expression

by

David Chih-Hsiang Yuan

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
2014

Doctoral Committee:

Associate Professor Patricia J. Wittkopp, Chair
Associate Professor Anuj Kumar
Assistant Professor Andrzej T. Wierzbicki
Professor Jianzhi George Zhang

“We sat on a crate of oranges and thought what good [people] most biologists are, the tenors of the scientific world—temperamental, moody, lecherous, loud-laughing, and healthy. ... The true biologist deals with life, with teeming boisterous life, and learns something from it, learns that the first rule of life is living...a thing every starfish knows in the core of his soul and in the vesicles between his rays. He must, so know the starfish and the student biologist who sits at the feet of living things, proliferate in all directions. ... Your true biologist will sing you a song as loud and off-key as will a blacksmith... Sometimes [they] may proliferate a little too much in all directions, but ... [they are] very good company, and at least [they do] not confuse a low hormone productivity with moral ethics.”

John Steinbeck & Edward F. Ricketts

Sea of Cortez: A Leisurely Journal of Travel and Research

Dedications

To my mother, who taught me life and gave me a lifetime of love.

To my father, who worked tirelessly to provide us a better life.

To my wife, who loved, supported, and encouraged me through our adventures.

And to our son, who brings us so much joy and wonder for the future.

Acknowledgements

There are many people I would like to thank and acknowledge, first and foremost of whom is my advisor, Trisha Wittkopp. Her curiosity and passion for science have inspired and motivated me, and her patience, support, and guidance have been invaluable for my training. I am grateful for the opportunity to be part of her laboratory in the past five years; the open and friendly environment she fostered and the group of people she brought together have made it an enriching and enjoyable place for training and discovery. I also would like to thank my committee members Anuj Kumar, Andrzej Wierzbicki, and Jianzhi George Zhang for their invaluable feedback and guidance on my research and dissertation.

I would like to thank past and present members of the Wittkopp laboratory for discussion, feedback, and friendship. In particular, Jonathan Gruber taught me the foundations of yeast genetics as well as coding in R and, along with Gizem Kalay and Trisha, guided, supported, and motivated me through some early obstacles; Brian Metzger and Fabien Duveau have been amazing research partners, being always at the ready to discuss cool ideas, brainstorm projects, and solve problems; Rich Lusk provided valuable suggestions on data analysis, especially for Chapter 2; Joe Coolon has provided valuable advice and feedback as well as friendship since I joined the lab;

Elizabeth Walker made our lives in the lab so much smoother by being an amazing lab manager; Zhiyuan Yao contributed to the construction of the *cis*-regulatory mutants; Lisa Sramkoski, Gizem Kalay, Arielle Cooley, Emma Stewart, Kraig Stevenson, and Ulises Rosas made the lab such a fun place to be. There are several other people whose influence and support made my work in this dissertation possible. Donna Jurdy, Ruby and Bob MacDonald, Dave Jacobs, Volker Hartenstein, Chuck Taylor, Nagayasu Nakanishi, and Chris Winchell, in their teachings and friendships, each cast unique and important influence on me as a scientist. My parents, Virginia Mung and Joseph Yuan, gave me life, love, education, freedom to pursue my interests, and a home filled with affection, wisdom, and happiness. Last but certainly not least, my amazing wife, Claire Spafford, has been a wellspring of love and support. Her kind heart and cheery spirit have filled my life with warmth and joy, while her patience and encouragement have carried me through challenges. She also gave us our amazing son, Isidro V Yuan. I don’t know how we ever lived without him in our lives.

The work in this dissertation was supported by the University of Michigan, Rackham Graduate School, Department of Molecular, Cellular, and Developmental Biology, Program in Biomedical Sciences, National Science Foundation, and National Institutes of Health Genetics Training Program (GM 007544-34).

Table of Contents

List of Figures

Abstract

Mutation is the root of all genetic variation. A new mutation may exhibit little or no effect. A mutation in the coding region of a gene may affect the gene product's structure and function. A mutation can also impact gene expression. Gene expression is the process through which information encoded in the DNA is converted to the molecular machinery carrying out specific biological functions, ultimately generating higher-order phenotypes. Mutations that modify gene expression can therefore contribute to phenotypic variation. In this dissertation, I characterize the effects of mutations and natural genetic variants that impact gene expression to better understand how mutations contribute to gene expression variation. To characterize an array of genetic variants, I used a fluorescent reporter controlled by the promoter of the *Saccharomyces cerevisiae* gene *TDH3*. I first investigate the effect of *cis*- and *trans*-regulatory mutations. Prior studies suggest that *cis*-acting expression quantitative trait loci may exhibit larger effect on average than *trans*-acting ones, yet little is known empirically regarding the difference in effect of *cis*- and *trans*-regulatory mutations. I directly compared *cis*-regulatory mutants to previously isolated *trans*-regulatory mutants and found that *cis*-regulatory mutations intrinsically have larger effects than *trans*-regulatory mutations. Next, I investigate the contribution of mutations to *cis*-regulatory variation in *TDH3* expression. Determining the evolutionary contribution of mutation and selection among regulatory variation is challenging due to lack of functional annotation among regulatory sequences. Instead of choosing putatively non-functional sites as the neutral model, I used an empirical null distribution of functional effects of *cis*-regulatory mutations for comparison. The results suggest that mutation underlies *TDH3 cis*-regulatory variation in mean expression level while selection may have favored decreased expression noise via epistasis. Lastly, I investigate the effect of genotype-by-environment (GxE) interactions on mutations to gain further mechanistic insight into how the mutation process contributes to expression variation. I characterized *cis*-regulatory mutations in environments reflecting conditions in which *TDH3* functions and found GxE interactions to be common. These results and their implications on the evolutionary impact of the mutation process are discussed.

Chapter 1

Introduction

Understanding the origins of biodiversity is a fundamental motivation in biology, and hence how observable traits change has long been of interest. To address this question, biologists have worked to uncover the components and mechanisms that produce such traits as well as processes that alter them. As a result, the types of genetic changes and molecular mechanisms that underlie many trait or phenotypic changes have become better-understood over the last few decades. In several cases, the causal genetic changes, particularly those that act via modification of gene expression, have even been identified at the nucleotide level. A complementary question following this concerns the origins and characteristics of such genetic changes. Where do genetic changes come from? What are their characteristics regarding the phenotypes they impact? How do they contribute to evolution? To answer these questions, the mutation process—the original source of all genetic variation—needs to be interrogated.

In this chapter, I will introduce the mutation process and the role it plays in evolution. I will then discuss mutation in the context of variation in gene expression, which has been implicated to play an important role in phenotypic evolution. A few case studies in which mutation alters phenotype via modification of gene expression will be presented. I will follow by reviewing prior studies that characterize the mutation process as well as show how such understanding can be applied to infer the evolutionary processes responsible for gene expression variation observed in extant or natural populations. To present an approach that addresses some limitations of the studies discussed, I introduce my efforts to characterize a mutation spectrum and how it can be used to answer fundamental evolutionary questions regarding variation observed in nature. Next, to further characterize mutations in a more biologically and evolutionarily realistic manner, I will introduce how environment may impact the effect of mutations. As the effect of a mutation may determine its evolutionary fate, the variability of effect due to environmental condition is an important mutation characteristic. I will review prior studies that characterize such variability due to interaction between genetic variation and the environment. Finally, I introduce my efforts to characterize the mutation spectrum in different environments in order to understand how mutational effect on gene expression may vary.

Mutation process

In its most elemental form, the process of biological evolution that underlies the origins of biodiversity entails a change, within a population, in the relative abundance of individual organisms with different genotypes, which can often exhibit different phenotypes. This change in abundance among genotypes is driven by evolutionary forces acting at the population level: genetic drift (random fluctuations in the frequencies of genotypes) and selection (non-random differential survival of a genotype over others), with selection acting on phenotypes that are heritable via genotypes. For evolution to occur, there must be a variety of genotypes within a population. The ultimate source of this genotypic variation is mutation.

Mutation is an inevitable process associated with information storage, usage, and propagation in the genome. If a genome storing millions or billions of nucleotides is likened to a book composed of letters, errors can arise in several ways from extracting, interpreting, and propagating the information in the book: the book may be physically damaged as to cause misreading of the letters, or copying of the book may not be of perfect fidelity. Similarly, genomes can be damaged from environmental insults. For instance, ultraviolet radiation may fuse adjacent thymine bases to form pyrimidine dimers. Such damages can subsequently lead to sequence errors that may hamper DNA function in replication, transcription, and, indirectly, translation. DNA replication itself as well as DNA repair mechanisms can also introduce sequence errors. Furthermore, certain chemical and physical properties of DNA may render it prone to spontaneous reactions such as the deamination of cytosine into uracil. Such error or change in DNA sequence—mutation—is thus a continual occurrence inherent to biological organisms.

Mutations exist as several types. Point mutations or single nucleotide substitutions are those that affect a single nucleotide and are a common form of spontaneous mutation. Indels—insertions or deletions—may involve a single or many nucleotides. When occurring in coding sequence in multiples other than three, they shift the reading frame of the codons and are hence frame-shift mutations. Duplications or copy number variation involve changes in the number of a functional unit of DNA, e.g. gene or chromosome. Stretches of DNA may also be rotated to form inversions or exchanged to form recombinants. Functionally, a silent mutation is one that exhibits no effect; an example is a synonymous mutation in the coding sequence of a gene that alters the codon but not the amino acid in the protein product due to degeneracy of the genetic code (although a synonymous mutation may still exhibit effect by altering the transcribed RNA, e.g. disrupting transcript stability or creating or deleting a splicing site). In contrast, a non-synonymous mutation changes the codon to that of a different amino acid (missense mutation) or a stop codon (nonsense mutation). Different types of mutations vary in terms of their phenotypic effects as well as rates of occurrence. For example, mutations that occur in regulatory elements may impact the target gene's expression, whereas those in coding sequence may alter the activity of the gene product. In terms of mutation rate, the relative sizes of a gene's coding sequence and regulatory region may impact how frequently mutations may occur in each. These characteristic differences across mutations can have different bearing on evolution.

While mutations are often defined as errors, it is important to recognize that they also provide a continuous influx of novel DNA sequences into a population. Some new mutations may have little or no effect. They may arise and remain, by random chance, to become natural variants in a population without causing appreciable detriment to the survival and reproduction of individual organisms at a given time. Yet as the physical, cellular, or even genetic environment changes, such natural variants may exhibit effect that now impacts the survival and reproduction of the organisms harboring them relative to other individuals in the population and be subjected to evolutionary forces. This process is illustrated in Figure 1.1. While mutation alone is not sufficient to drive such evolutionary change, its role as the source of genetic variation is key to evolution.

Gene expression, evolution, and mutation

Gene expression is the process through which information encoded in the DNA is transformed into molecules that carry out specific biological functions. Functional interactions among such molecules result in many observable traits or “higher-order” phenotypes. For example, the complex and precise spatiotemporal expression of genes encoding developmental transcription factors lead to specification of body plans and development of body parts [Spitz & Furlong 2012]. Gene expression is therefore critical in converting genotype to phenotype. How might genetic changes or mutations that alter gene expression impact higher-order phenotypes?

Phenotypic changes across many organisms—e.g. pigmentation in fruit flies [Wittkopp *et al*. 2002], pelvic spine in stickleback fish [Chan *et al*. 2010], beak morphology in Darwin’s finches [Abzhanov *et al*. 2004]—have been associated with variation in gene expression. In some of these cases, the mechanism underlying the change in gene expression has been shown to involve changes in the *cis*-regulatory element—DNA sequences that contain functional elements critical to the expression of the gene responsible for or associated with the phenotype (Figure 1.2). A recent genome-wide study also found the majority of genetic changes underlying adaptive evolution to be regulatory rather than coding changes in natural populations of sticklebacks [Jones *et al*. 2012]. Expression variation through genetic changes in regulatory DNA elements is therefore an important mechanism of phenotypic evolution [Wray 2007; Stern & Orgogozo 2008; Wittkopp & Kalay 2011].

Cis-regulatory elements are typically found outside of the coding sequence of a gene; e.g. upstream or downstream of the coding region or within introns. Their function is effected by transcription factors binding to cognate binding sites contained within the *cis*-regulatory element in level-, time-, and space-dependent manners. An exact combination of bound transcription factors may impart certain physical properties to the DNA and ultimately lead to the recruitment of the transcription machinery, eliciting transcriptional expression (Figure 1.2). Changes in the *cis*-regulatory element may affect this relationship, thereby altering spatiotemporal expression and the eventual phenotype. The impact of such change is particularly well-illustrated during development, in which small changes in expression of key developmental genes can have large impacts on the developing phenotype. An example is beak morphology of Darwin's finches, in which expression levels of *Bmp4* and CaM correlate with beak length, depth, and width across species adapted to different foraging specializations [Abzhanov *et al*. 2006; Abzhanov *et al*. 2004]. Another example is in *Drosophila* pair-rule gene *eve*, in which *cis*-regulatory genetic changes in its stripe 2 enhancer across species alter expression level such that the level in one species is not viable in another [Ludwig *et al*. 2006].

To date, the field has taken a few phenotypic changes of interest and uncovered the underlying regulatory changes to great detail as mechanisms of phenotypic evolution. For instance, loss of pelvic spines in natural populations of threespine sticklebacks has been shown to involve regulatory mutations at the *Pitx1* locus [Chan *et al*. 2010]. Interspecific wing [Gompel *et al*. 2005] and abdominal [Wittkopp *et al*. 2002] pigmentation variation in *Drosophila* have also been shown to involve *cis*-regulatory changes at the *yellow* locus. Such genetic changes represent extant natural genetic variants that survived evolutionary processes acting at the phenotypic level. These discoveries bring up questions regarding the origin of such genetic variants. How do they come about? As discussed earlier, mutation is the source of genetic change. What are the characteristics of mutations in *cis*-regulatory elements? What effect do such mutations have on gene expression? Unfortunately, predicting the impact of regulatory changes on downstream phenotypes has remained challenging. This is due in part to the complex relationship between *cis*-regulatory elements and transcription factors in the regulation of gene expression. In addition, the spectrum of possible regulatory variants before being subjected to evolutionary processes is not well-characterized in terms of its phenotypic consequence. Such regulatory variants represent the novel DNA sequences produced by the mutation process.

As discussed previously, mutation generates genetic variation on which evolutionary forces—i.e. genetic drift and selection—can act. Understanding such mutation spectra is key to discerning the forces that shape natural variation. This is an important task of interest to evolutionary biology. Because the mutation spectrum is a random sample of possible mutations, it represents genetic variation independent of or prior to the effects of selection. Using it as a basis of comparison allows us test if variation observed in nature—such as gene expression variation—is consistent with that produced by the mutation process alone or has been influenced by selection. Characterizing and understanding this spectrum thus offers tools to answer a fundamental question in evolutionary biology.

Mutation spectrum

Characterizing the mutation spectrum requires a large number of mutations obtained randomly in the absence of natural selection. Sampling of natural isolates for genetic variants is not an ideal method to achieve this due to the low-frequency nature of the mutation process. More importantly, such variants may themselves be the results of natural selection. To circumvent these issues, prior studies that characterized newly-arisen mutations have employed mutation accumulations lines which have been subjected to serial artificial population bottlenecks—in which as few as one individual was used to reestablish the population each generation—in order to minimize the effect of selection. Notably, such studies across four model systems have offered genome-wide views of the mutation spectra at the DNA level:

Arabidopsis thaliana [Ossowski *et al*. 2010], *Caenorhabditis elegans* [Denver *et al*. 2009], *Drosophila melanogaster* [Keightley *et al*. 2009], and *Saccharomyces cerevisiae* [Lynch *et al*. 2008].

New mutations genome-wide can be detected and characterized by coupling whole-genome sequencing to long-term mutation accumulation experiments. Using this method, these studies offered global views of the mutation spectrum in terms of rate of occurrence and distributions of type and genomic location. Specifically, the genome-wide mutation rates were on the magnitude of 10^{-9} to 10^{-10} mutation per site per generation. Single nucleotide substitutions were the most frequently-detected mutations, although this may be influenced by the ability to map longer stretches of indels and structural rearrangements during analysis of genome sequencing data. Across all four organisms, G:C → A:T transitions and G:C → T:A transversions were most commonly observed, with the rate of former up to twice that of the latter. The mutations detected tended to be uniformly distributed in the genome. Altogether, these studies obtained and characterized more newly-arisen spontaneous mutations per line than previously possible: On average, 20 mutations were observed in each mutation accumulation line of *A. thaliana* [Ossowski *et al*. 2010], 39 in *C. elegans* [Denver *et al*. 2009], 58 in *D. melanogaster* [Keightley *et al*. 2009], and 33 in *S. cerevisiae* [Lynch *et al*. 2008].

The mutation accumulation experiments agnostically surveyed newly-arisen mutations in the genome. Properties of such mutations—including frequency, range of effects, and other characteristics that affect the activity of a focal gene—are of interest because they can deepen the level of detail on how the mutation process shapes genetic variation, complement the studies that revealed mechanisms underlying specific phenotypic changes, and enable tests of selection among gene expression variation. However, the mutation accumulation experiments cannot provide much detail regarding the mutation spectrum that affects gene expression beyond frequency. This is because these studies focused on capturing and identifying any newly-arisen mutations regardless of location as opposed to characterizing a range of mutations impacting expression at a specific locus. The frequencies of intergenic mutations—many mutations that affect gene expression are found outside of coding sequence—were 0.55 (54 out 98 mutations mapped total) in *A. thaliana* [Ossowski *et al*. 2010], 0.42 (165 out of 391) in *C. elegans* [Denver *et al*. 2009], and *D. melanogaster* 0.45 (78 out of 174) [Keightley *et al*. 2009]. [Landry *et al*. 2007] used mutational accumulation in *S. cerevisiae* to describe global expression variation contributed by the mutation process. Specifically, four mutation accumulation lines were evolved for 4000 generations, followed by microarray profiling to estimate divergence of expression phenotype. However, this study focused on exploring the mutational effect without identification of the causal mutations. It remains unclear the frequencies, range of effects, and other characteristics of mutations that impact the activity of a focal gene.

To achieve a more comprehensive view of the mutation spectrum at both the DNA and phenotype levels, an alternative method to obtain newly-arisen mutations independent of selection is necessary. Specifically, mutations need to be characterized individually and impact a specific locus. This is difficult to achieve with mutation accumulation experiments because more than one mutation is likely to accumulate in a line or genotype. The likelihood of obtaining many mutations that impact a specific locus is also low using mutation accumulation. Furthermore, the activity of a gene needs to be measurable in a high-throughput and quantitative manner. This can be achieved by coupling chemical mutagenesis, a fluorescent reporter, and flow cytometry. Chemical mutagenesis allows the mutagen to be titrated to induce, on average, one causal mutation that impacts gene expression per genome. Fluorescent reporter can be engineered in a transgene to assay activity of a *cis*-regulatory element. Specifically, this regulatory activity can be rapidly and precisely quantified in flow cytometry as reporter fluorescence.

Overview of experimental system in yeast

The yeast species *Saccharomyces cerevisiae* is a suitable model organism in which to build and characterize a spectrum of mutations that impact gene expression using mutagenesis, fluorescent reporter, and flow cytometry. As majority of yeast genes lack introns [Parenteau *et al*. 2008], their *cis*-regulatory elements are compact and typically consist of the promoter in the intergenic region upstream of the coding sequence; this simplifies the transgenic design for the interrogation of regulatory activity. Like other eukaryotes, yeast regulatory regions contain modular functional elements, many of which are well-characterized. Abundant genetic tools also exist for yeast, and its amenability for genetic manipulation enables large collections of *cis*-regulatory variant strains to be feasibly constructed. Furthermore, high-throughput assessment of gene expression on individual yeast cells using flow cytometry makes it possible to rapidly and precisely quantify the effect of mutations on regulatory activity.

[Gruber *et al*. 2012] used this method previously to obtain and characterize a spectrum of newly-arisen mutations that impact the expression of a focal gene in *S. cerevisiae*. In this study, ethyl methanesulfonate (EMS) was used to elevate the mutation rate in order to survey a large number of newly-arisen mutations. EMS preferentially induces G:C → A:T transitions, which, as previously mentioned, is tied with G:C → T:A transversions as the most common type of substitutions in yeast observed by [Lynch *et al*. 2008]. The activity of the promoter from *TDH3*, the gene-specific promoter chosen for analysis, was assayed by fusing its *cis*-regulatory element to the coding sequence of yellow fluorescent protein and integrating this transgene into the yeast genome (Figure 1.3). Reporter fluorescence was used as the marker to quantify *TDH3* promoter activity using flow cytometry. This study was thus able to capture a large number of newly-arisen mutations in the genome that affect the expression of *TDH3*. As summarized in Figure 1.4, 221 mutants were obtained, characterized, and subdivided into four categories: coding (16), copy number variation (22), *cis*- (4), and *trans*-regulatory (179). On average, compared to *cis*-regulatory mutants, *trans*-regulatory mutants exhibited smaller effect size on expression level, had larger mutation target size, and tended to be recessive. While this represents one of the largest collections of regulatory mutations affecting expression of a focal gene, the comparison of mutational effect between *cis*- and *trans*-regulatory mutations was based on only 4 (3 unique) *cis*-regulatory mutations. Few other studies to date has empirically compared the mutational effect of *cis*- versus *trans*-regulatory mutations, despite the contribution of both to phenotypic evolution [e.g. Wittkopp *et al*. 2008; Emerson *et al*. 2010]. This serves as the starting point of my experimental efforts to continue the characterization of mutations that impact gene expression.

TDH3, the yeast gene whose promoter was chosen for analysis, encodes glyceraldehyde-3-phosphate dehydrogenase (GAPDH) isozyme 3 (also known as triose-phosphate dehydrogenase). Its native locus resides on chromosome VII. Null mutants for *TDH3* are viable but exhibit increased heat sensitivity and decreased competitive fitness as well as dessication and chemical resistance [Deutschbauer *et al*. 2005; Sinha *et al*. 2008; Ratnakumar *et al*. 2011; Grant *et al*. 1999]. *Tdh3p* is one of three GAPDHs in yeast involved in glycolysis and gluconeogenesis, each one of which has a different specific activity [McAlister & Holland 1985; McAlister & Holland 1985]. GAPDHs are also involved in osmolarity regulation in yeast, serving as substrates for a key enzyme in the metabolic network that leads to production of glycerol as solute [O'Rourke *et al*. 2002; Hyduke & Palsson 2010]. All three GAPDHs are localized in cytosol as well as the cell wall [Delgado *et al*. 2001].

The promoter of *TDH3* (P_{TDH3}) was chosen for analysis for several reasons. Wild-type *TDH3* expression level, as measured by protein abundance, is the 42nd highest among ~6000 in the yeast genome [Ghaemmaghami *et al*. 2003]. This facilitates reliable detection and precise quantification of expression by reporter fluorescence, which is not possible with genes expressed at median levels [Huh *et al*. 2003]. P_{TDH3} is also a well-characterized promoter with known transcription factors [Holland *et al*. 1987; Pavlović & Hörz 1988; Kuroda *et al*. 1994; Yagi *et al*. 1994], an attribute that is useful in interpreting and understanding the effect of mutations across P_{TDH3}. Last but not least, the diverse function of the *Tdh3p* protein makes P_{TDH3} well-suited to test the impact of environmental conditions on mutational effect.

Building a *cis*-regulatory mutation spectrum

To gain a better understanding of the mutation spectrum that impacts gene expression, I expanded the collection of *cis*-regulatory mutations in [Gruber *et al*. 2012] into a collection of hundreds of mutations. This enables a more precise and detailed characterization of *cis*-regulatory mutations. In addition, as *cis*-regulatory natural variants can be more readily identified, this also enables a comparison of the impact on expression of *cis*-regulatory mutation spectrum versus natural variants in the same region of sequence in order to understand what forces shaped *cis*-regulatory variation observed in the wild.

To build a large collection of *cis*-regulatory mutations, I used the identical genetic system described in [Gruber *et al*. 2012] but using site-directed, instead of chemical, mutagenesis to generate mutations in the *cis*-regulatory element. To facilitate comparison of these *cis*-regulatory mutations to those in *trans* from [Gruber *et al*.

2012], I mimicked the action of EMS (G:C → A:T transitions) during site-directed mutagenesis. Besides high frequency of occurrence observed in mutation accumulation experiments, this type of substitution is also the most frequently observed mutations among natural isolates of yeast [Maclean *et al*. *In prep*.]. Targeting G and C sites across the 678bp P_{TDH3} resulted in 236 mutations, each engineered individually in a mutant strain for a total of 236 strains. The effects of the mutations were then quantified using flow cytometry and compared to that of *trans*-regulatory mutants. This work is discussed in Chapter 2.

Natural variation

In previous subsections, I discussed how elucidating the mutation spectrum in the absence of selection provides a reference to better infer the history of variation observed in nature. So, what kind of expression variation exists in natural populations? Gene expression has been widely observed to vary within populations across many organisms. For instance, expression variation has been reported for a significant fraction of loci in the human genome as well as in yeast, fly, and mouse [Gilad *et al*. 2008; Rockman & Kruglyak 2006].

Two related questions logically follow the observation of expression variation at the population level: what are its functional and evolutionary origins? The genetic changes associated with such variation may be mapped with a number of tools. For instance, quantitative trait locus mapping can be used to identify region or locus associated with expression variation. Genetic variants found in the region may then be directly tested to identify the causal genetic change. Once the genetic change producing expression variation has been identified, understanding the underlying molecular mechanisms may still be challenging as functional annotation of regulatory sequence is relatively lacking. For well-characterized *cis*-regulatory elements (e.g. promoters of well-studied genes), functional elements such as the TATA-box, untranslated regions, activating and repressing regions, and transcription factor binding sites may be empirically known but not necessarily the impact of changes in them. This is in contrast to estimating the effect of a change in the coding sequence; for instance, using the genetic code as well as other functional characterizations of transcripts and proteins, a change in the coding sequence can be identified as silent, missense, nonsense, or/and altering transcript processing. Lack of understanding of function further precludes answering evolutionary questions. However, efforts have been made to investigate regulatory variation from functional and evolutionary perspectives.

Function and evolution of regulatory variation

Substantial portions of the eukaryotic genome are comprised of non-coding sequences which contain regulatory information critical for gene expression. While such sequences have historically been less well-characterized than coding sequences, recent studies have suggested that a significant fraction of them may be functional [Zhen & Andolfatto 2012]. Identifying the functionally important regions and elements in the non-coding parts of the genome as well as the evolutionary forces—specifically, detecting and quantifying selection—that shaped their form remain an area of active interest [Romero *et al*. 2012]. Many such efforts have entailed the use of evolutionary constraint via level of divergence and polymorphism frequency [Zhen & Andolfatto 2012].

Evolutionary constraint relies on the logic that functional regions of the genome are less amenable to change than non-functional regions. For instance, a non-synonymous change in the protein-coding sequence is expected to be more deleterious than a synonymous change; the frequency of former among extant variants is thus expected to be lower than that of the latter. On the other hand, in a non-functional region, both types of changes should be equally devoid of effect and hence occur at the same frequency. Applied to non-coding sequence, level of divergence at non-coding regions of interest has been compared to that at neutral reference sites to identify conserved non-coding sequences [e.g. Andolfatto 2005; Siepel *et al*. 2005; Gaffney & Keightley 2006]. Applied across greater evolutionary distances, phylogenetic footprinting studies have used the logic to identify regulatory elements throughout the genome [e.g. Duret & Bucher 1997; Boffeli *et al.* 2003]. However, this method requires the assumption that genetic changes are mostly or all deleterious, which is not realistic. Variable mutation rate and bias between regions compared can also confound the comparison. In addition, the choice of neutral reference sites is critical. As how much of the genome is truly “non-functional” and therefore neutral is currently not empirically known, this may lead to biased estimation of constraint.

As purifying or negative selection—selection against and removal of deleterious mutations—tends to decrease polymorphism at functional sites, the distribution of polymorphism frequencies can be used as an alternative or in addition to level of divergence to investigate function and detect selection. This approach has been used to reveal that purifying selection has acted on most of the polymorphisms underlying expression variation in yeast [Ronald & Akey 2007; Emerson *et al*. 2010]. Furthermore, when coupled with level of divergence, this method can provide more information about the direction and intensity of selection [e.g. McDonald & Kreitman 1991; Keightley & Eyre-Walker 2007; Bustamante *et al*. 2001]. Such tests are called or based on the McDonald-Kreitman test, which compares measures of constraint based on polymorphism within and divergence between species or population [McDonald & Kreitman 1991]. While originally developed for coding sequence, this test has also been applied to non-coding sequence by, for instance, comparing non-coding DNA of interest to functional elements (e.g. transcription factor binding sites) in non-coding DNA, which serves as the neutral reference [e.g. Jenkins *et al*. 1995; Ludwig & Kreitman 1995]. The choice of the neutral reference site may thus still be a source of bias.

A common impediment to the methods described to detect function and selection with non-coding DNA lies in the choice of the neutral reference site. This is further hampered by the lack of functional annotation of non-coding DNA. Application of such tests to an individual regulatory element is difficult. Instead of choosing neutral reference sites with compromises or caveats, generating an empirical set of random genetic changes as the null for comparison is an alternative. Doing so closer mimics the random mutation process, and the distribution of functional consequences of the mutations can also be empirically determined. This type of logic has been applied in several studies, including: variation in *cis*-regulatory sequence and transcription factor binding affinity across *Drosophila* species [Moses 2009], bristle number in *D. melanogaster* as a quantitative trait locus [Rice & Townsend 2012], and variation in *cis*-regulatory sequence and transcriptional expression across mammals [Smith *et al.* 2013].

In [Moses 2009], functional effects of *cis*-regulatory variation across *Drosophila* species was compared to a null distribution of effects of random substitutions to detect selection, with function being transcription factor binding affinity. The null distribution in this study was actually generated *in silico* and hence not empirical, although the transcription factors of interest (*Bcd* and *Kr*), their bindings sites (*hb* anterior activator and *eve* stripe 2 enhancer), and their binding affinities are highly characterized as to allow more realistic predictions of function. However, this particular method cannot be readily applied to other systems as most *cis*-regulatory elements and their binding relationship with transcription factors are not annotated to such degree. The assumption that selection favors increased binding affinity employed in this study also cannot be generalized across all other systems.

A test for a different type of variation but using a similar null model as basis of comparison was used in [Rice & Townsend 2012]. In this study, the authors developed a population genetics-based model to describe distributions of phenotypes resulting from neutral evolution and different selection schemes using data from prior mutation accumulation experiments as the null distribution of mutational effect size on sternopleural and abdominal bristle numbers in *Drosophila*. Quantitative trait locus (QTL) data from prior artificial selection experiments on bristle number was then compared to the model; the selection schemes producing the observed QTL matched those employed during artificial selection, while the neutral model produced none of the observed QTL. In other words, QTL of interest can be compared to a range of effect size models produced using the null distribution and varying selection schemes to infer the presence and strength of selection. However, variation resulting from artificial selection may be less complex than and not representative of that produced by natural selection [Nei 2013]. This particular approach also focused on DNA regions associated with quantitative traits instead of regulatory variation. Nevertheless, it does further highlight the use of empirical null distributions to characterize neutral variation and ultimately be used as the basis to detect selection.

Using an empirical null distribution of expression variation generated from a large number of *cis*-regulatory variants, [Smith *et al*. 2013] tested mammalian *cis*-regulatory elements for selection. This study used saturation mutagenesis of the enhancers of interest and transcriptional output of these mutant enhancers to assay the effect distribution of *cis*-regulatory mutations. The enhancer activity measured was semi-quantitative, and the effect of each mutation was classified as silent, up-, or down-regulatory. The metrics K_u/K_n and K_d/K_n were calculated, with K_u being substitution frequency of up-regulatory mutations, K_d down-regulatory, and K_n silent, which was assumed to be selectively neutral. The metrics were then similarly obtained for species within the same phylogenetic order as that of the enhancer mutagenized to create the null distribution and compared to infer selection. Of the three liver enhancers tested—ALDOB, ECR11, and LTV1—purifying selection against down-regulatory mutations was overall detected, while positive selection for up-regulatory selection was detected only for LTV1. While this study demonstrates the use of an empirically-derived null distribution to detect selection among *cis*-regulatory variation, the semi-quantitative nature of the mutational effect distribution makes it challenging to interrogate gene expression variation, which is quantitative by nature. Furthermore, measurement of the effect of mutations and natural variants was carried out in non-endogenous environment, which complicates the interpretation of the results across the mammalian order.

These studies represent some of the efforts so far to investigate variation in functional and evolutionary contexts with noted limitations and caveats. A complementary approach to investigate *cis*-regulatory variation requires generating the mutation spectrum as the basis for the null distribution and testing natural variants in a comparable system and context. The genetic system used in Chapter 2 provides tools for this task.

The effects of *cis*-regulatory mutations and natural variants

TDH3 expression variation has been observed among natural isolates of yeast. Work by Brian Metzger in the Wittkopp laboratory has identified *cis*-regulatory variants among these natural isolates, some of which are putatively associated with expression variation. In Chapter 3, the effects of these *cis*-regulatory variants are tested using the identical genetic system as in Chapter 2. Specifically, each *cis*-regulatory variant was engineered individually into the transgene with *TDH3* promoter driving reporter yellow fluorescent protein in the same genomic location, each one carrying a single natural variant. Flow cytometry was again used to quantify the effect of natural variants on expression. As many of these variants exist in the natural isolates not individually but in the context of one to several other variants as haplotypes, their effects were also tested in haplotypes. This entailed constructing another collection, again using the same genetic system except haplotypes instead of individual variants were tested for effect on expression. By comparing the effect on expression of the natural variants individually and in haplotypes, epistasis or genetic interaction among variants can also be detected.

Finally, a null distribution representing the spectrum of effects of random mutations was generated by a resampling or bootstrap approach using the *cis*-regulatory mutations described in Chapter 2. The distribution was created this way as to closer represent the mutation process in nature in which new mutations arise out of the spectrum of possible mutations. The effects of *TDH3* *cis*-regulatory natural variants was then compared against this null distribution to infer the contribution of mutation and selection in shaping natural variation. This work represents an approach to detect selection among regulatory variation for a specific regulatory element based on empirical data.

The interactions between mutation and the environment

The impact of environment on the effect of mutations should not be ignored when characterizing mutations with the goal of understanding their evolutionary contribution. Just as genetic changes may interact with each other epistatically, they may also interact with the environment. For microorganisms such as yeast, environmental perturbations can pose considerable challenge to the stability of their cellular environment. A host of responses at the gene expression level exists to maintain equilibrium. From a functional perspective, *cis*-regulatory elements may contain specific functional elements critical for gene expression in different environments. A *cis*-regulatory mutation may exhibit variable effect in one environment versus another depending on its location relative to such functional elements. This can greatly impact its evolutionary trajectory and underscores the importance of *cis*-regulatory mutations in the evolution of responses to fluctuating environmental factors [Landry *et al*. 2006; Smith & Kruglyak 2008; Espinosa-Soto *et al*. 2011]. Investigating how the environment can modify the effects of mutations may shed light on another mechanism through which the mutation process contributes to phenotypic variation.

Changes in gene expression in response to environmental variation is an area of active research. The ability of microorganisms to respond, survive, and modify their environments has garnered much interest for application to a myriad of environmental, energy, and health problems. Genome-wide expression variation in yeast has been surveyed across a range of environments and stressors, including carbon sources, heat shock, oxidative and reductive stress, hyper- and hypo-osmolarity, extreme pH, and mutagens [Gasch *et al*. 2000; Gasch & Werner-Washburne 2002; Kvitek *et al*. 2008]. Initial studies employing microarrays discovered drastic expression changes in response to environmental perturbation. While some loci observed were environment-specific, a large set of genes were commonly involved and designated part of the environmental stress response [Gasch 2007]. The environment has thus been shown as an important influence on gene expression in yeast.

In light of the environmental impact on gene expression, how do different genetic variants that affect expression respond to environmental change? Genotype-by-environment (GxE) interactions—how genetic variants manifest different phenotypes in different environments—have also been a subject of interest [Maranville *et al*. 2012; Grishkevich & Yanai 2013]. Genome-wide surveys for GxE interactions have revealed such interactions to be common. In *C. elegans*, for instance, 197 expression quantitative trait loci have been mapped in different temperatures [Li *et al*. 2006]. Similar numbers of loci have also been identified in yeast across different nutrient environments [Landry *et al*. 2006]. More specifically, different loci have been found to be associated with expression variation in different carbon sources in yeast, suggesting environmental differences may impact the relative importance of genetic variants [Smith & Kruglyak 2008]. While these studies revealed diverse loci in the genome underlying expression variation across environments, few have examined a wide range of locus-specific genetic variants and their impact on gene expression in different environments. Such study can further complement our characterization of the mutation process as the environment is an important agent of selection. Chapter 4 thus investigates GxE interactions for a spectrum of genotypes at a specific locus.

Based on functional knowledge of *TDH3* as well as prior studies on environmental conditions that influence activity of P_{TDH3}, I chose several environments in which to test the spectrum of *cis*-regulatory mutations in order to assess the effect of GxE interactions on regulatory activity. Specifically, the effects of mutations were characterized in glycolytic versus gluconeogenic conditions as the *Tdh3p* protein participates in metabolic pathways in both. Variable activity in different regions of P_{TDH3} has also been observed between these two conditions [Kuroda *et al*. 1994], making *TDH3* a candidate to be influenced by GxE interactions. The effects of mutation were also characterized in condition that induces osmotic stress, as *Tdh3p* is involved in the regulation of cellular osmolarity [O'Rourke *et al*. 2002; Hyduke & Palsson 2010]. Results from these experiments as well as their implications on how the mutation process contributes to evolution are discussed in Chapter 4.

Summary

The work described in this dissertation improves our understanding of the mutation process and evolution through the characterization of a *cis*-regulatory mutation spectrum (Chapter 2), the development and implementation of an empirical method to detect selection among gene expression variation (Chapter 3), and the investigation of genotype-by-environment interaction that can influence the effect and evolutionary trajectory of newly-arisen mutations (Chapter 4). Through the course of the following chapters, I describe the construction and characterization of a *cis*-regulatory mutation spectrum for the yeast glycolytic gene *TDH3*. As a counterpart for this collection of possible genetic variants, I examine extant *TDH3 cis*-regulatory variants from natural isolates and their effects on gene expression. To better understand the evolutionary forces that shaped the *cis*-regulatory variation surveyed, I generate a null distribution representing variation produced by the neutral, random mutation process based on the collection of *cis*-regulatory mutations and use it to test for selection among the natural variants. Finally, to investigate another contributor to expression variation, I examine how interaction between mutations and the environment can influence the effect of mutations on gene expression. In Chapter 5, I discuss overall conclusions from these experiments, their implications, and future directions to further increase our understanding of the mechanisms that underly gene expression variation and phenotypic evolution.

References

Abzhanov A, Protas M, Grant BR, Grant PR, and Tabin CJ. *Bmp4* and Morphological Variation of Beaks in Darwin's Finches. *Science* 305:1462-1465 (2004).

Andolfatto P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437:1149-1152 (2005).

Boffeli D, McAuliffe J, Ovcharenko D, Lewis K, Ovcharenko I, Pachter L, and Rubin E. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299:1391-1394 (2003).

Bustamante CD, Wakeley J, Sawyer S, and Hartl DL. Directional Selection and the Site-Frequency Spectrum. *Genetics* 159:1779-1788 (2001).

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, *et al*. Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a *Pitx1* Enhancer. *Science* 327:302-305 (2010).

Delgado ML, O'Connor JE, Azorin I, Renau-Piqueras J, Gil ML, and Gozalbo D. 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 (2001).

Denver DR, Dolan P, Wilhelm L, Sung W, Lucas-Lledó J, Howe D, Lewis S, Okamoto K, Thomas W, and Lynch M. A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *PNAS* 106:16310-16314 (2009).

Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, and Giaever G. Mechanisms of Haploinsufficiency Revealed by Genome-Wide Profiling in Yeast. *Genetics* 169:1915-1925 (2005).

Duret L and Bucher P. Searching for regulatory elements in human noncoding sequences. *Current Opinion in Structural Biology* 7:399-406 (1997).

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

Espinosa-Soto C, Martin OC, and Wagner A. Phenotypic plasticity can facilitate adaptive evolution in gene regulatory circuits. *BMC Evolutionary Biology* 11:5-18 (2011).

Gaffney DJ and Keightley PD. Genomic selective constraints in murid non-coding DNA. *PLoS Genetics* 2:1912-1923 (2006).

Gasch AP. Comparative genomics of the environmental stress response in ascomycete fungi. *Yeast* 24:961-976 (2007).

Gasch AP and Werner-Washburne M. The genomics of yeast responses to environmental stress and starvation. *Functional and Integrative Genomics* 2:181-192 (2002)

Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, and Weissman JS. Global analysis of protein expression in yeast. *Nature* 425:737-741 (2003).

Gilad Y, Rifkin S, and Pritchard J. Revealing the architecture of gene regulations: the promise of eQTL studies. *Trends in Genetics* 24:408-415 (2008).

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, and Carroll S. Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* 433:481-487 (2005).

Grant CM, Quinn KA, and Dawes IW. Differential Protein S-Thiolation of Glyceraldehyde-3-Phosphate Dehydrogenase Isoenzymes Influences Sensitivity to Oxidative Stress. *Molecular Cell Biology* 19:2650-2656 (1999).

Grishkevich V and Yanai I. The genomic determinants of genotype x environment interactions in gene expression. *Trends in Genetics* 29:479-487 (2013).

Gruber JD, Vogel K, Kalay G, and Wittkopp PJ. Contrasting Properties of Gene-Specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects, and Dominance. *PLoS Genetics* 8:e1002497 (2012).

Holland MJ, Yokoi T, Holland JP, Myambo K, and Innis MA. The *GCR1* gene encodes a positive transcriptional regulator of the enolase and glyceraldehyde-3-phosphate dehydrogenase gene families in *Saccharomyces cerevisiae*. *Molecular and Cell Biology* 7:813-820 (1987).

Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, and O'Shea EK. Global analysis of protein localization in budding yeast. *Nature* 425:686-691 (2003).

Hyduke DR and Palsson BØ. Towards genome-scale signalling-network reconstructions. *Nature Reviews Genetics* 11:297-307 (2010).

Jenkins DL, Ortori CA, and Brookfield JF. A test for adaptive change in DNA sequences controlling transcription. *Proceedings of the Royal Society Biological Sciences* 261:203-207 (1995).

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, *et al*. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484:55-61 (2012).

Keightley PD and Eyre-Walker A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177:2251-2261 (2007).

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, and Blaxter ML. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research* 19:1195-1201 (2009).

Kuroda S. Otaka S, and Fujisawa Y. Fermentable and nonfermentable carbon sources sustain constitutive levels of expression of yeast triosephosphate dehydrogenase 3 gene from distinct promoter elements. *Journal of Biological Chemistry* 269:6153-6162 (1994).

Kvitek DJ, Will J, and Gasch A. Variations in Stress Sensitivity and Genomic Expression in Diverse *S. cerevisiae* Isolates. *PLoS Genetics* 4:e1000223 (2008).

Landry CR, Oh J, Hartl D, and Cavalieri D. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* 366:343-351 (2006).

Landry CR, Lemos B, Rifkin S, Dickinson WJ, and Hartl D. Genetic Properties Influencing the Evolvability of Gene Expression. *Science* 317:118-121 (2007).

Li Y, Álvarez O, Gutteling E, Tijsterman M, Fu J, Riksen J, Hazendonk E, Prins P, Plasterk R, and Jansen R. Mapping Determinants of Gene Expression Plasticity by Genetical Genomics in *C. elegans*. *PLoS Genetics* 2:e222 (2006).

Ludwig MZ and Kreitman M. Evolutionary dynamics of the enhancer region of even-skipped in *Drosophila*. *Molecular Biology and Evolution* 12:1002-1011 (1995).

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, and Thomas WK. A genome-wide view of the spectrum of spontaneous mutations in yeast. *PNAS* 105:9272-9277 (2008).

Maclean C, Metzger BPH, Yang J, Ho W-C, Moyers B, and Zhang J. Deep sequencing, population genetics and high-throughput phenotypic analysis of diverse *Saccharomyces cerevisiae* strains. *In preparation*.

Maranville JC, Luca F, Stephens M, and Di Rienzo A. Mapping gene-environment interactions at regulatory polymorphisms: Insights into mechanisms of phenotypic variation. *Transcription* 3:56-62 (2012).

McDonald JH and Kreitman M. Adaptive Protein Evolution at the *Adh* Locus in *Drosophila*. *Nature* 351:652-654 (1991).

McAlister L and Holland MJ. Isolation and characterization of yeast strains carrying mutations in the glyceraldehyde-3-phosphate dehydrogenase genes. *Journal of Biological Chemistry* 260:15013-15018 (1985).

McAlister L and Holland MJ. Differential expression of the three yeast glyceraldehyde-3-phosphate dehydrogenase genes. *Journal of Biological Chemistry* 260:15019-15027 (2985).

Moses AM. Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. *BMC Evolutionary Biology* 9:286-299 (2009).

Nei, M. Mutation-driven evolution. Oxford University Press, Oxford (2013).

O'Rourke SM, Herskowitz I, and O'Shea EK. Yeast go the whole HOG for the hyperosmotic response. *Trends in Genetics* 18:405-412 (2002).

Ossowski S, Schneeberger K, Lucas-Lledó J, Warthmann N, Clark R, Shaw R, Weigel D, and Lynch M. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327:92-94 (2010).

Parenteau J, Durand M, Véronneau S, Lacombe A-A, Morin G, Guérin V, Cecez b, Gervais-Bird J, Chu-Shin Koh C-S, Brunelle D, Wellinger RJ, Chabot B, and Elela SB. Deletion of Many Yeast Introns Reveals a Minority of Genes that Require Splicing for Function. *Molecular Biology of the Cell* 19:1932-1941 (2008)

Pavlović B and Hörz W. The chromatin structure at the promoter of a glyceraldehyde phosphate dehydrogenase gene from *Saccharomyces cerevisiae* reflects its functional state. *Molecular and Cellular Biology* 8:5513-5520 (1988).

Ratnakumar S, Hesketh A, Gkargkas K, Wilson M, Rash BM, Hayes A, Tunnacliffe A, and Oliver SG. Phenomic and transcriptomic analyses reveal that autophagy plays a major role in desiccation tolerance in *Saccharomyces cerevisiae*. *Molecular Biosystems* 7:139-149 (2011).

Rice DP and Townsend JP. A Test for Selection Employing Quantitative Trait Locus and Mutation Accumulation Data. *Genetics* 190:1533-1545 (2012).

Rockman MV and Kruglyak L. Genetics of global gene expression. *Nature Reviews Genetics* 7:862-872 (2006).

Romero IG, Ruvinsky I, and Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics* 13:505-516 (2012).

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

Siepel A, Bejerano G, Pedersen J, Hinrichs A, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier L, Richards S, Weinstock G, Wilson R, Gibbs R, Kent W, Miller W, and Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 15:1034-1050 (2005).

Sinha H, David L, Pascon RC, Clauder-Munster S, Krishnakumar S, Nguyen M, Shi G, Dean J, Davis RW, Oefner PJ, McCusker JH, and Steinmetz LM. Sequential Elimination of Major-Effect Contributors Identifies Additional Quantitative Trait Loci Conditioning High-Temperature Growth in Yeast. *Genetics* 180:1661-1670 (2008).

Smith EN and Kruglyak L. Gene-Environment Interaction in Yeast Gene Expression. *PLoS Biology* 6:e83 (2008).

Smith JD, Mcmanus KF, and Fraser HB. A Novel Test for Selection on *cis*-Regulatory Elements Reveals Positive and Negative Selection Acting on Mammalian Transcriptional Elements. *Molecular Biology and Evolution* 30:2509-2518.(2013).

Spitz F and Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* 13:613-626 (2012).

Stern DL and Orgogozo V. The loci of evolution: How predictable is genetic evolution? *Evolution* 62:2155-2177 (2008).

Wittkopp P, Vaccaro K, and Carroll S. Evolution of *yellow* Gene Regulation and Pigmentation in *Drosophila*. *Current Biology* 12:1547-1556 (2002).

Wittkopp PJ, Haerum B, and Clark A. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genetics* 40:346-350 (2008).

Wittkopp PJ and Kalay G. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* 13:59-69 (2011).

Wray GA. The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews Genetics* 8:206-216 (2007).

Yagi S, Yagi K, Fukuoka J, and Suzuki M. The UAS of the yeast GAPDH promoter consists of multiple general functional elements including RAP1 and GRF2 binding sites. *Journal of Veterinary Medical Science* 56:235-244 (1994).

Zhen Y and Andolfatto P. Methods to Detect Selection on Noncoding DNA in *Evolutionary Genomics: Statistical and Computational Methods, Volume 2, Methods in Molecular Biology*, vol. 856, edited by Anisimova M. Humana Press, New York (2012).

Figure 1.1

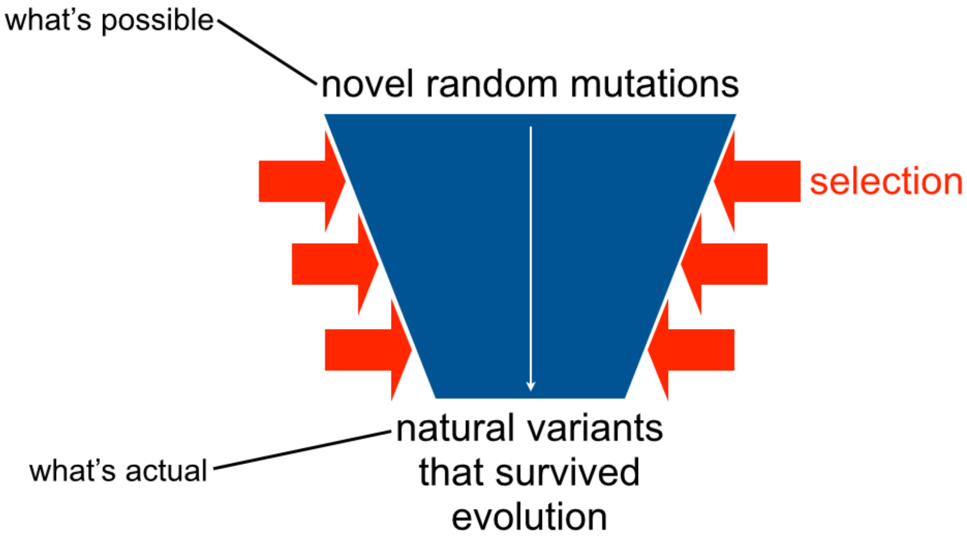

Figure 1.1 Schematic of the evolutionary process from newly-arisen mutations to extant natural variants. Novel random mutations (top) arise out of the spectrum of possible mutations. If/when selection is present (red), a subset of what were newly-arisen mutations may produce phenotypes more favorable for the survival and reproduction of the individuals harboring them. After iteration(s) of this process, a non-random subset of mutations survive to become natural variants (bottom), which represent what actually exist in extant natural populations at a given point in time.

Figure 1.2

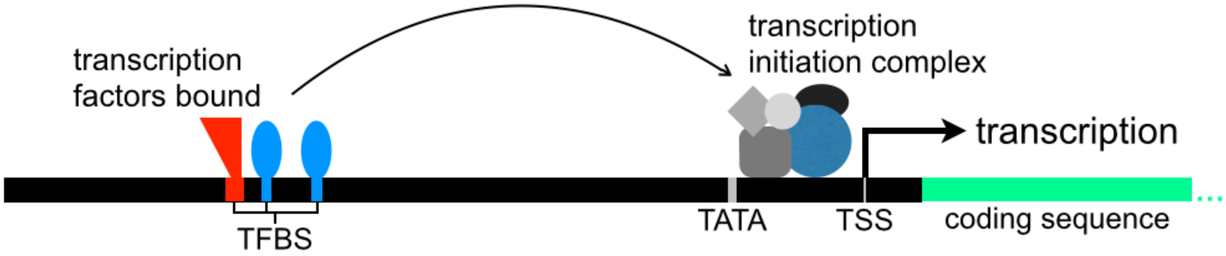

Figure 1.2 Schematic of a *cis*-regulatory element. The *cis*-regulatory element (black) of a gene contains function elements—such as transcription factor binding sites (TFBS) and TATA box—important for the transcription initiation of that gene. Proper binding of transcription factors lead to recruitment of the transcription machinery and ultimately transcription initiation. (TSS: transcription start site.)

Figure 1.3

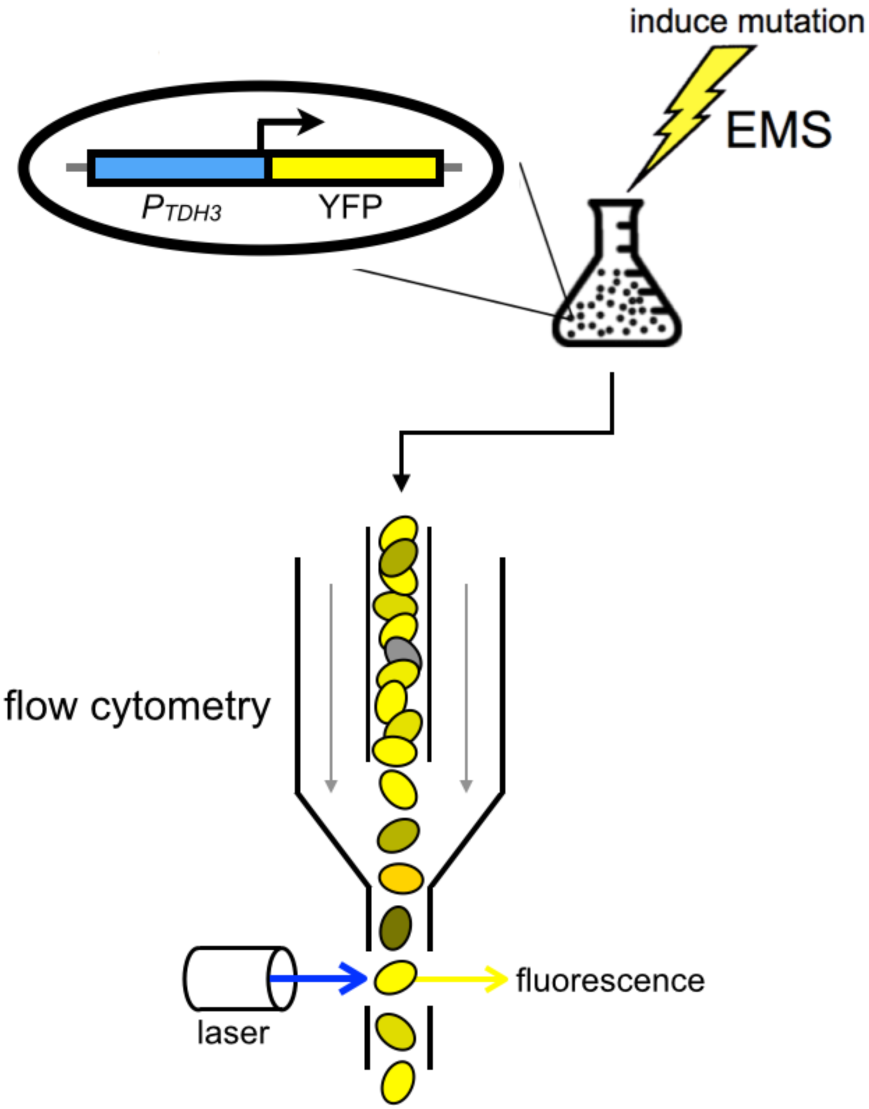

Figure 1.3 Overview of the experimental method used by [Gruber *et al*. 2012] to characterize newly-arisen mutations that impact gene expression. The *cis*-regulatory element of interest, the *Saccharomyces cerevisiae TDH3* promoter (P_{TDH3}), is engineered into a transgene driving expression of reporter yellow fluorescent protein (YFP) integrated into the *S. cerevisiae* genome. Population of a strain carrying wild-type P_{TDH3} in this transgene is mutagenized with ethyl methanesulfonate (EMS) to obtain newly-arisen mutants. The effects of newly-arisen mutations are subsequently quantified in flow cytometry as reporter fluorescence.

Figure 1.4

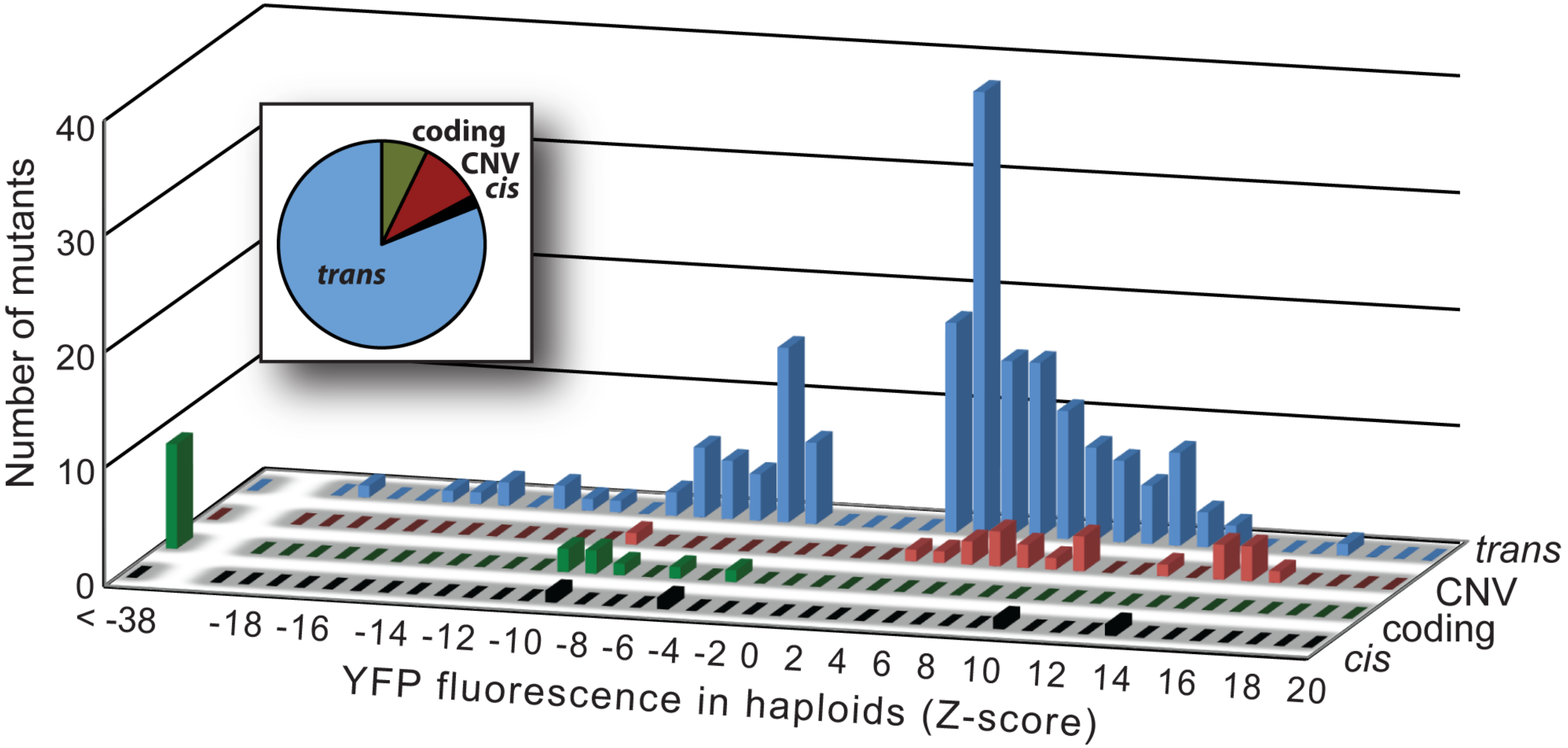
Number of mutants
0
10
20
30
40
coding
CNV
cis
trans
< -38
-18
-16
-14
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
14
16
18
20
YFP fluorescence in haploids (Z-score)
trans
CNV
coding
cis

Figure 1.4 The distribution of effect among mutants characterized by [Gruber *et al*. 2012]. The effects of mutations are shown as reporter YFP fluorescence of the 4 *cis*-regulatory (black), 16 coding (green), 22 copy number variation (CNV; red), and 179 *trans*-regulatory (blue) mutants. Inset pie chart shows the relative frequencies of different mutants. Figure from [Gruber *et al*. 2012].

Chapter 2

The relative effects of *cis*- and *trans*-regulatory mutations on gene expression

Abstract

Regulatory divergence has been shown to play an important role in phenotypic evolution. This has motivated researchers to better understand the characteristics of genetic changes that influence gene regulation. Such genetic changes can be broadly divided into two functional modes: *cis*- and *trans*-acting. *Cis*-acting genetic changes are expected to have smaller mutational target size as such mutations are located within *cis*-regulatory regions of a gene (e.g. promoter), whereas *trans*-acting ones—located anywhere in the genome—are presumed to be more pleiotropic and hence potentially more deleterious [Wray 2007; Stern & Orgogozo 2008; Wittkopp & Kalay 2012]. Expression changes in *cis* have been observed to predominate divergence between species, whereas those in *trans* are more prevalent within species [Wittkopp *et al*. 2008; McManus *et al*. 2010; Tirosh *et al*. 2009; Emerson *et al*. 2010]. Additionally, results from the mapping of expression quantitative trait loci have

suggested that those acting in *cis* may be of larger effect size those acting in *trans* [Schadt *et al*. 2003; Dixon *et al*. 2007; Hubner *et al*. 2007; West *et al*. 2007]. This may reflect a difference in intrinsic properties of *cis*- and *trans*-regulatory mutations. Alternatively, this may be the result of selection eliminating *trans*-regulatory mutations of large effect. To further characterize and compare *cis*- and *trans*-regulatory genetic changes, this study uses large collections of both types of mutations affecting a common locus to investigate their effect on gene expression. Using a reporter gene containing the promoter of the glycolytic gene *TDH3* driving expression of yellow fluorescent protein that has been integrated into the genome of *Saccharomyces cerevisiae*, *cis*- and *trans*-regulatory mutants were tested for their effect on expression, which was quantified as reporter fluorescence using flow cytometry. *Cis*-regulatory mutants exhibited, on average, larger effect on expression than *trans*-regulatory mutants and were more likely to decrease expression. These differences in effect can impact the evolutionary trajectories of newly-arisen *cis*- and *trans*-regulatory mutations and may help explain their observed distributions.

Introduction

The regulation of gene expression is critical in converting genotypes to phenotypes. The evolutionary importance of genetic changes that affect this process has been recognized for over 40 years and was alluded to even earlier [Monod & Jacob 1961; Britten & Davidson 1971; King & Wilson 1975]. This is now supported by case studies in which gene regulatory divergence has been shown to underlie phenotypic change [Wittkopp *et al*. 2002; Chan *et al*. 2010; more reviewed in Wray 2007, Carroll 2008, and Stern & Orgogozo 2008]. The genetic changes themselves—the genotypes—driving observed expression variation can be classified as *cis*- or *trans*-regulatory. The functional effect of a *cis*-regulatory change is allele-specific, with the causal genetic change typically located proximally (e.g. in promoter or intron) to or at least on the same chromosome as the impacted gene. In contrast, the functional effect of a *trans*-regulatory change is not allele-specific because the causal genetic change impacts diffusible *trans*-acting factors. As gene expression itself is a phenotype amenable to precise and high-throughput quantification, its profiling within and between divergent populations or species has been used to gain evolutionary insight. A key finding from such investigations has been that the proportion of expression differences attributable to *cis*-regulatory change appears greater than that to *trans*-regulatory changes between, rather than within, species [Wittkopp *et al*. 2008]. This is well-documented in both flies and yeast [e.g. McManus *et al*. 2010; Tirosh *et al*. 2009; Emerson *et al*. 2010]. This preferential accumulation of *cis*-regulatory change over evolutionary time suggests that natural selection may be sensitive to properties of *cis*- and *trans*-regulatory mutations and favor particular molecular mechanisms to achieve divergence in expression.

Pleiotropy is often used to explain this pattern of regulatory divergence [Wray 2007; Stern & Orgogozo 2008]. *Cis*-regulatory elements are comprised of modular functional elements, thus the effect of a mutation within a module will be restricted to only that module and potentially incur less selective penalty. By contrast, since *trans*-acting factors typically regulate multiple genes, the impact of a *trans*-regulatory mutation will be more wide-spread or pleiotropic, thereby amplifying any potential deleterious effect across multiple downstream loci. A prediction from this difference is that extant *trans*-regulatory variation may be of smaller effect than *cis-regulatory* variation, since *trans*-regulatory mutations of large effect may be selectively disadvantageous. Consistent with this idea, selection against *trans*-regulatory expression changes have been shown in natural populations of *Caenorhabditis elegans* [Denver *et al*. 2005]. In terms of effect size, findings from expression quantitative trait loci (eQTL) mapping appear consistent with this as well; *cis*-acting eQTL appear to exhibit, on average, larger effects on expression than *trans*-acting eQTL [Schadt et al. 2003; Dixon et al. 2007; Hubner et al. 2007; West et al. 2007].[1] To date, little is known about the relative effects of *cis*- and *trans*-regulatory mutations.

Prior work [Gruber *et al*. 2012] by provided a glimpse into this question. With the goal of characterizing mutations that impact gene expression, a collection of random

[1] A number of confounding experimental factors may influence such eQTL results, however, some of which have been reviewed in [Alberts *et al*. 2007; Gilad *et al*. 2008; Rockman & Kruglyak 2008].

newly-arisen regulatory mutants that exhibited altered reporter fluorescence was isolated independent of selection using ethyl methanesulfonate (EMS) mutagenesis. While many classes of mutants—i.e. *cis*, *trans*, coding, and duplication—were captured, the majority turned out to be *trans*-regulatory. The 4 (3 unique) *cis*- and 179 *trans*-regulatory mutants in this study exhibited similar patterns regarding mutational effect size as discussed above, raising the possibility that *cis*-regulatory mutations may intrinsically have larger effect than those in *trans*. Here I expand the collection of *cis*-regulatory mutations to investigate the effect on gene expression of *cis*- and *trans*-regulatory mutations with the goal of providing a more thorough empirical dataset to test the hypothesis that *cis*-regulatory mutations have larger intrinsic effect than *trans*-regulatory mutations.

Results

To systematically characterize the effect of *cis*-regulatory mutations on gene expression, I used a reporter transgene, integrated into the *Saccharomyces cerevisiae* genome, containing the *cis*-regulatory element of interest driving expression of yellow fluorescent protein (Figure 2.1A). The *cis*-regulatory element I used contained the promoter of the yeast gene *TDH3* (P_{TDH3}). This gene has robust expression [Ghaemmaghami *et al*. 2003], which enables precise quantification of expression from P_{TDH3} alleles using reporter fluorescence [Huh *et al*. 2003]. To generate a large collection of *cis*-regulatory mutants, site-directed mutagenesis was used to engineer single-nucleotide substitutions individually into P_{TDH3} in the P_{TDH3}-YFP reporter transgene. Specifically, I made G:C → A:T transitions. These substitutions represent two of the most frequently observed spontaneous mutations in the yeast genome [Lynch et al. 2008] as well as the predominant type of mutation induced by EMS used by [Gruber *et al*. 2012] to obtain the *trans*-regulatory mutants [Coulondre & Miller 1977; Greene *et al*. 2003]. This resulted in 236 *cis*-regulatory mutations spread throughout the 678bp P_{TDH3} (Figure 2.1B). Each of these *cis*-regulatory mutations was engineered into a strain individually, resulting in 236 mutant strains. The effect of each *cis*-regulatory mutation on P_{TDH3} activity was then quantified using flow cytometry as reporter fluorescence, which is well-correlated with reporter protein abundance [Huh *et al*. 2003].

The effects exhibited by these *cis*-regulatory mutants were compared to those by the *trans*-regulatory mutants isolated previously in the EMS screen for mutants affecting fluorescence of the same P_{TDH3}-YFP transgene [Gruber *et al*. 2012]. Briefly, these mutant genotypes were isolated by exposing a population of wild-type P_{TDH3}-YFP strain to EMS, which was titrated to introduce one mutation that impacts P_{TDH3} activity in each strain on average[2]. To obtain the 179 *trans*-regulatory mutants mentioned earlier, fluorescence-activated cell sorting was used following mutagenesis to isolate those with significantly altered reporter fluorescence out of the pool of EMS-mutagenized strains.

[2] Subsequent efforts to map the causal mutation using bulk segregant analysis have identified one candidate single-nucleotide substitution each in 11 *trans*-regulatory mutant strains, so far 7 of which have been experimentally tested and verified to recapitulate the phenotype of the original mutants [personal communications J. Gruber, B. Metzger, and F. Duveau]. This provides confidence that the titration of the mutagen indeed resulted in the *trans*-regulatory mutants to have one mutation that may impact expression on average.

Cis-regulatory mutants exhibit lower mean expression level

The effect on reporter fluorescence of the *TDH3 cis*-regulatory mutations are shown in Figure 2.2. The majority (177/236) of the mutations did not significantly alter mean expression level (based on t-tests with Bonferroni-adjusted significance threshold of $p = 0.0002$). The mutations that exhibited effects of the largest magnitude were found in previously-characterized transcription factor binding sites (TFBS). To put the distribution of effects found in these *cis*-regulatory mutants in the context of previously published work, I directly compared them to the collection of *trans*-regulatory mutants affecting P_{TDH3}-YFP activity described in [Gruber *et al*. 2012] (Figure 2.3). I found that the *cis*-regulatory mutants, on average, exhibited lower mean expression level than genotypes reported as *trans*-acting mutants in [Gruber *et al*. 2012] (t-test, *p*-value < 0.05). This remained the case after I excluded *cis*-regulatory mutants with mutations in known TFBS (t-test, *p*-value < 0.05).

Cis-regulatory mutants show higher enrichment for large effect mutations

The *trans*-regulatory mutants just shown represent those that exhibited significant difference from wild-type controls from a separate experiment. In other words, they are *trans*-acting mutants of particularly large effect. By contrast, the *cis*-regulatory mutants represent all of those inducible by EMS within P_{TDH3} regardless of significance and agnostic of effect size. A more appropriate comparison to the collection of *cis*-regulatory mutants is thus with the pool of EMS-mutagenized strains (i.e. the population of wild-type P_{TDH3}-YFP strain that was exposed to EMS) from which the *trans*-regulatory mutants were originally isolated. This pool is thought to provide a reasonable approximation of newly-arisen *trans*-regulatory mutants induced by EMS without filtering for statistical significance because most (179/221) of the regulatory mutants isolated from it were *trans*-acting [Gruber *et al*. 2012].

To obtain a more unbiased comparison of effects from *cis*- and *trans*-regulatory mutations, I compared the collection of *cis*-regulatory mutants with systematic mutations to the collection of EMS-treated genotypes with random and (nearly all) potentially *trans*-acting mutations (Figure 2.4). The *cis*-regulatory mutants exhibited decreased fluorescence more often than increased fluorescence (+5.9% vs. +0.5% respectively; Fisher's exact test, *p*-value < 0.05). The EMS mutants also exhibited decreased fluorescence more often than increased fluorescence (+1.4% vs. +1.1% respectively; Fisher's exact test, *p*-value < 0.05). However, *cis*-regulatory mutants were more likely to decrease fluorescence than *trans*-regulatory mutants (Fisher's exact test, *p*-value < 0.05). This is consistent with differences observed between the systematic *cis*-regulatory mutants and the statistically significant *trans*-regulatory mutants from [Gruber *et al*. 2012] (see Figure 2.3). This suggests that *cis*-regulatory mutations are more likely to decrease expression than *trans*-regulatory mutations. For the phenotype of decreased fluorescence, *cis*-regulatory mutants also exhibited larger magnitude of mutational effect than *trans*-regulatory mutants (mean ΔYFP -0.0991. vs. -0.0809; t-test, *p*-value < 0.05). This suggests that a phenotype of decreased expression level of large effect may be more likely to be achieved through a *cis*-

regulatory mutation than a *trans*-regulatory mutation, at least for the *TDH3* gene in *S. cerevisiae*.

Discussion

Properties of a mutation can influence its evolutionary fate

As mutation is the source of genetic variation, understanding the characteristics of newly-arisen mutations provides insight into a fundamental component of the evolutionary process. The fate of a new mutation can be broadly influenced by random chance as well as its characteristics. Genetic drift acts through the former to eventually fix or eliminate neutral and nearly-neutral mutations, the rate of which hinges on the size of the population in which the mutations arise. Selection can also fix or eliminate a mutation based on its characteristics. In this study, mutations were characterized and compared by their mode of action and effect on gene expression with the motivation being to understand mutational properties that can influence how they contribute to phenotypic evolution. Below, I discuss (1) the choice of EMS-type mutations for this study, (2) potential mechanism underlying the observed difference in *cis*- and *trans*-regulatory mutational effects, and (3) evolutionary implications of the results.

Approximation of common point mutations in yeast using EMS

This study represents one of the first to empirically compare and contrast the characteristics of large collections of both *cis*- and *trans*-regulatory mutants. Before discussing the results and implications, a consideration regarding the type of mutations analyzed should be addressed. An aspect of the study that is not wholly representative of nature is the choice of EMS-type mutations targeted for analysis. While mutagenesis for an array of mutation types can be performed for the *cis*-regulatory mutants due to their discrete mutational target, this is less so the case for *trans*-regulatory mutants due to their wide-spread genomic distribution. Mutations acting in *trans* to a focal gene may be found anywhere in the genome outside of that gene's *cis*-regulatory and coding regions. Such mutations may occur, for example, in the *cis*-regulatory or coding region of transcription factors regulating the focal gene. Site-directed mutagenesis is therefore not feasible given the large number of loci and sites that need to be targeted. As elevation of mutation rate is required to obtain a large number of mutants in a feasible manner, an alternative to chemical mutagenesis is mutation accumulation (MA). However, the number of mutations obtainable from MA experiments is typically on the order of 30 to 60 (e.g. 33 in [Lynch *et al*. 2008], 58 in [Keightley *et al*. 2009]). Obtaining a spectrum of mutations that impact a focal gene is also difficult. Using (and simulating) EMS to induce one of the most common types of mutations observed in the yeast genome should thus provide a reasonable estimate of the effects of spontaneous single-nucleotide substitutions in yeast. Other types of mutations, such as duplications and long stretch indels, however, may exhibit different characteristics and distributions of effect.

Functional mechanism of cis-regulatory mutations may lead to large effect size

The comparison of mutational effects on gene expression in this study revealed that *cis*-regulatory mutants exhibited higher enrichment for phenotypes of larger effect size and lower expression level than *trans*-regulatory mutants. Mechanistically, how might this be accounted for given the functional differences between *cis*- and *trans*-regulatory mutations? Mutations in the *cis*-regulatory element impact transcription through, for instance, altering transcription factor or nucleosome binding sites. This may help explain the predominance of down-regulatory phenotypes among *cis*-regulatory mutations as changes in binding sites more likely lead to disruption rather than improvement of their function, thereby hampering transcription activation. This general relationship may be assumed since control of transcription in eukaryotes tends to be effected through activation from a basal state rather than repression (as in bacterial systems) [Struhl 1999]. However, transcription factors that function as repressors—whether intrinsically or in context-dependent manners—certainly exist in eukaryotes. *Cis*-regulatory mutations that affect the binding of such transcription factors would be expected to increase expression.

While mutations that affect a *trans*-acting factor also disrupt transcription of a target gene, the action may be less direct. The regulation of expression of a target gene can be represented by a gene regulatory network consisting of layers of *cis*-regulatory elements regulating the coding sequence of *trans*-acting factors. *Trans*-regulatory mutations that impact the target gene may occur anywhere in the regulatory network outside of the target gene and its *cis*-regulatory element. Given this level of separation from the target locus, there may be buffering of the effect of such mutations through feedback, compensation, or redundancy. For instance, mechanisms may exist to maintain expression of *trans*-acting factors at a certain level and compensate, to the extent possible, for the effect of a mutation that alters its expression [Kafri *et al*. 2005]. Such buffering is a key property of regulatory networks in proper maintenance of gene expression, an inherently noisy process. In addition, the mutational target size of *trans*-regulatory mutations is much larger than that of *cis*-regulatory mutations [Wittkopp 2005; Landry *et al*. 2007]; the former comprises the coding and regulatory sequence of all *trans*-acting factors in a target gene's regulatory network, whereas the latter comprises only the *cis*-regulatory sequence of the target gene. The probability of a *cis*-regulatory mutation hitting a functional element, such as a TFBS, and incurring more drastic effects is higher than that of a *trans*-regulatory mutation. Empirically testing these hypotheses will require mapping a large number of *trans*-regulatory mutations and determining their genomic locations and mechanisms of action.

Evolutionary implications

Cis-regulatory expression change has been implicated to play a larger role in divergence between species than *trans*-regulatory expression change [Wittkopp *et al.*

2008; McManus *et al*. 2010; Tirosh *et al*. 2009; Emerson *et al*. 2010]. One reason for this might be that *cis*-regulatory mutations are less pleiotropic than *trans*-regulatory mutations, and hence *trans*-regulatory mutations (especially those of large effect) are more likely to be selected against [Wittkopp *et al*. 2008; Wray 2007; Stern & Orgogozo 2008]. This explanation suggests that the apparent larger effect of *cis*-acting expression variation observed in extant populations compared to that of *trans*-acting expression variation may be the result of selection rather than intrinsic properties of the underlying genetic changes. However, my empirical data suggest that newly-arisen *cis*- and *trans*-regulatory mutations intrinsically differ in effect size. A phenotypic change of large effect may be more readily achieved through a single *cis*-regulatory mutation than a single *trans*-regulatory mutation.

Even if newly-arisen *cis*- and *trans*-regulatory mutations have similar distributions of effect sizes, negative selection against *trans*-regulatory mutations and/or positive selection for *cis*-regulatory mutations could explain the increasing contribution of *cis*-regulatory variation with evolutionary time. My results suggest, however, that this may not be the case. At least for single-nucleotide substitutions, *cis*-regulatory mutations are more likely to exhibit larger effect than *trans*-regulatory mutations on average. This is consistent with the observation that *trans*-regulatory mutations arise frequently and tend to be of limited effect [Wittkopp *et al*. 2008]. While this study does not examine the pleiotropic consequences of *cis*- and *trans*-regulatory mutations, the distribution of *trans*-regulatory mutational effects suggests the possibility that negative selection against *trans*-regulatory mutations may not be as big a contributor as positive selection for *cis*-regulatory mutations to the observed distributions of *cis*- and *trans*-regulatory variation.

Materials and Methods

Strains

All strains were constructed and tested as haploids. The wild-type control against which all P_{TDH3} mutant strains were compared is the *S. cerevisiae* strain P_{TDH3}-YFP, MAT**a**, *lys2Δ0*, *ura3Δ0*, *CAN1*^{S} previously constructed based on BY4724 by G. Kalay and J. Gruber [Gruber *et al*. 2012]. P_{TDH3}-YFP is a reporter transgene containing the wild-type form of P_{TDH3}, coding sequence of YFP (Venus variant), and the yeast *CYC1* (cytochrome c isoform 1) terminator placed into a pseudogene on chromosome I. The 678bp P_{TDH3} sequence consists of the 5' intergenic region up to but not including the start codon of the *TDH3* coding sequence; this region contains TFBS, TATA box, 5' untranslated region (UTR) of *TDH3*, and 3' UTR of *PDX3*, the gene upstream of *TDH3* in the native locus on chromosome VII. *CYC1* terminator is the canonical yeast terminator and polyadenylation sequence to provide efficient transcription termination and transcript stability [Russo *et al*. 1989; Zaret *et al*. 1984]. This strain was treated with EMS to obtain the pool of EMS mutants, from which the large effect *trans*-regulatory mutants were isolated, as described in [Gruber *et al*. 2012].

An intermediate strain based on the wild-type control strain was created from which mutant strains were subsequently constructed: *URA3*-yfp, MAT**a**, *lys2Δ0*, *ura3Δ0*, *CAN1*^{S}. It was derived from the control strain by exchanging P_{TDH3} in the reporter transgene with a PCR construct containing the *URA3* promoter & coding sequence and a mutation that disrupts the start codon of the YFP coding sequence; both *URA3* and the YFP start codon mutation were utilized to aid the screening process during mutant strain construction. Transformation was carried out following the lithium acetate protocol using selection for *ura*+ phenotype and loss of YFP fluorescence [Gietz & Schiestl 2007] and confirmed using Sanger sequencing.

Mutant strain construction

Site-directed mutagenesis was performed using two overlapping primers containing the desired mutation for each *cis*-regulatory mutation followed by PCR sewing. The resultant PCR construct also contains sequence that restores the YFP start codon. Mutant strains were then constructed by transforming the PCR constructs into the *URA3*-yfp intermediate strain following the lithium acetate method using selection for *ura*- (with 5-FOA) and gain of YFP fluorescence [Gietz & Schiestl 2007]. Putative transformants were confirmed using Sanger sequencing. Mutant strains are of the genotype P_{TDH3}^{mut}-YFP, MAT**a**, *lys2Δ0*, *ura3Δ0*, *CAN1*^{S}.

Flow cytometry

Prior to quantifying mutational effects in flow cytometry, strains were arrayed into a 96-well format in randomized order and stored at -80˚C in glycerol stocks. Arrayed strains were revived from glycerol stock onto YPG (glycerol) solid media to reduce formation of petites. This was carried out for all strains simultaneously to control for potential age-related effects and new additional mutations. Strains were subsequently transferred into deep 96-well plate liquid culture, with each strain grown in 500µl of liquid YPD (glucose) media for 20 hours to stationary phase while shaking at 250rpm with 3mm glass bead in a 30˚C incubator.

The effects exhibited by *cis*-regulatory and large effect *trans*-regulatory mutants were quantified as reporter fluorescence using a BD Accuri C6 flow cytometer coupled with an IntelliCyt HyperCyt Autosampler. 9 biological replicates were independently inoculated, grown, and analyzed. Immediately prior to quantification of fluorescence, 20µl of YPD culture was transferred into 500µl of SC-R (glucose) and passed into the flow cytometer. Flow rate of 14µl/min and core size of 10µm was used, with a blue laser (λ = 480 nm) for excitation of YFP and fluorescence data collected from the FL1 channel using a 533/30nm filter. Each well in a 96-well plate was sampled for 2-3s, with 20,000 events recorded on average. Data was then processed using flowClust and flowCore packages in R to remove artifacts such as debris and other non-cell events [Lo *et al.* 2009; Hahne *et al.* 2009]. Following processing, samples with fewer than 1000 cells were excluded from further analysis. Using the remaining data, YFP fluorescence was calculated as $(\log_{10}FL1.A)^2/(\log_{10}FSC.A)^3$ for each event for each sample. Mean expression level for each strain is then estimated as mean across 9 biological replicates. Random variation during growth that may affect gene expression was controlled for using 20 replicates of the wild-type control strain in the same position on each plate. Autofluorescence estimated from a non-fluorescent strain was used to correct all fluorescence values.

Quantification for the EMS mutants is described in [Gruber *et al*. 2012]. Briefly, the control (unmutagenized wild-type P_{TDH3}-YFP strain, stained with Cy5) and EMS-treated cells were analyzed simultaneously in a FACSaria flow cytometer/cell sorter for 9 consecutive runs. Data was processed using flowClust and flowCore packages in R similar to above, and the remaining data used to calculated YFP fluorescence as $(\log_{10}FL1.A)^2/(\log_{10}FSC.A)^3$. Random variation during growth that may affect gene expression was controlled for using the unmutagenized control strain across in each of the 9 runs.

Statistical analysis

All statistical analyses were carried out in R 3.0.2. To generate the distribution of *cis*-regulatory mutants comparable to that of the EMS mutants (Figure 2.4A), I sampled from the flow cytometry data (post-processing with debris and non-cell events removed) on the *cis*-regulatory mutants with the same number of cells randomly sampled from each mutant strain across 9 replicates. A distribution of control wild-type strain was similarly generated. This resulted in 2,227,069 cells in the *cis*-regulatory mutant distribution and 64,603 cells in the control wild-type distribution (compared to 2,855,692 cells in the EMS mutant distribution and 3,405,833 cells in the corresponding control wild-type distribution). T-test on the difference in mutational effect between *cis*- and *trans*-regulatory mutants was carried out using the function t.test in R. Fisher's exact test on enrichment differences among *cis*-regulatory and EMS mutants was carried out using the function t.test and fisher.test in R.

References

Alberts R, Terpstra P, Li Y, Breitling R, Nap J-P, and Jansen RC. Sequence polymorphisms cause many false *cis* eQTLs. *PLoS One* 7:e622 (2007).

Britten RJ and Davidson EH. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarter Review of Biology* 46:111-138 (1971).

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

Chan YF et al. Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a *Pitx1* Enhancer. *Science* 327:302-327 (2010).

Coulondre C and Miller JH. Genetic studies of the *lac* repression : IV. Mutagenic specificity in the *lacI* gene of *Escherichia coli*. *Journal of Molecular Biology* 117:577-606 (1977).

Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, and Thomas WK. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nature Genetics* 37:544-548 (2005).

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, and Cookson WOC. A genome-wide association study of global gene expression. *Nature Genetics* 39:1201-1207 (2007).

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

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, and Weissman JS. Global analysis of protein expression in yeast. *Nature* 425:737-741 (2003).

Gietz RD and Schiestl RH. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols* 2:31-34 (2007).

Gilad Y, Rifkin SA, and Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics* 24:408-415 (2008).

Greene EA, Codomo CA, Taylor NE, Henikoff JG, Till BJ, Reynolds SH, Enns LC, Burtner C, Johnson JE, Odden AR, Comai L, and Henikoff S. Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis*. *Genetics* 164:731-740 (2003).

Gruber JD, Vogel K, Kalay G, and Wittkopp PJ. Contrasting Properties of Gene-Specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects, and Dominance. *PLoS Genetics* 8:e1002497 (2012).

Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, Sarkar D, Spidlen J, Strain E, and Gentleman R. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* 10:106-113 (2009).

Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics* 37:243-253 (2007).

Holland MJ, Yokoi T, Holland JP, Myambo K, and Innis MA. The *GCR1* gene encodes a positive transcriptional regulator of the enolase and glyceraldehyde-3-phosphate dehydrogenase gene families in *Saccharomyces cerevisiae*. *Molecular and Cell Biology* 7:813-820 (1987).

Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, and O'Shea EK. Global analysis of protein localization in budding yeast. *Nature* 425:686-691 (2003).

Kafri R, Bar-Even A, and Pilpel Y. Transcription control reprogramming in genetic backup circuits. *Nature Genetics* 37:295-299 (2005).

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, and Blaxter ML. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research* 19:1195-1201 (2009).

King MC and Wilson AC. Evolution at two levels in humans and chimpanzees. *Science* 188:107-116 (1975).

Kuroda S, Otaka S, and Fujisawa Y. Fermentable and nonfermentable carbon sources sustain constitutive levels of expression of yeast triosephosphate dehydrogenase 3 gene from distinct promoter elements. *Journal of Biological Chemistry* 269:6153-6162 (1994).

Landry C, Lemos B, Rifkin S, Dickinson WJ, and Hartl D. Genetic properties influencing the evolvability of gene expression. *Science* 317:118-121 (2007).

Lo K, Hahne F, Brinkman R, and Gottardo R. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics* 10:145-152 (2009).

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DK, and Thomas WK. A genome-wide view of the spectrum of spontaneous mutations in yeast. *PNAS* 105:9272–9277 (2008).

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

Monod J and Jacob F. General conclusion—teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor Symposia on Quantitative Biology* 26:389-401 (1961).

Pavlović B and Hörz W. The chromatin structure at the promoter of a glyceraldehyde phosphate dehydrogenase gene from *Saccharomyces cerevisiae* reflects its functional state. *Molecular and Cellular Biology* 8:5513-5520 (1988).

Rockman MV and Kruglyak L. Genetics of global gene expression. *Nature Reviews Genetics* 7:862-872 (2006).

Russo P and Sherman F. Transcription terminates near the poly(A) site in the *CYC1* gene of the yeast *Saccharomyces cerevisiae*. *PNAS* 86:8348-8352 (1989).

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, and Friend SH. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422:297-302 (2003).

Stern DL and Orgogozo V. The loci of evolution: how predictable is genetic evolution? *Evolution* 62:2155-77 (2008).

Struhl K. Fundamentally Different Logic of Gene Regulation in Eukaryotes and Prokaryotes. *Cell* 98:1-4 (1999).

Tirosh I, Reikhav S, Levy AA, and Barkai N. A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation. *Science* 324:659-662 (2009).

West MAL, Kim K, Kleibenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, and St. Clair DA. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* 175:1441-1450 (2007).

Wittkopp PJ, True JR, and Carroll SB. Reciprocal functions of the *Drosophila yellow* and *ebony* proteins in the development and evolution of pigment patterns. *Development* 129:1849-1858 (2002).

Wittkopp, PJ. Genomic sources of regulatory variation in *cis* and in *trans*. *Cellular and Molecular Life Sciences* 62:1779-1783 (2005).

Wittkopp PJ, Haerum BK, and Clark AG. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genetics* 40:346-350 (2008).

Wittkopp PJ and Kalay G. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* 13:59-69 (2012).

Wray G. The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews Genetics* 8:206-216 (2007).

Yagi S, Yagi K, Fukuoka J, and Suzuki M. The UAS of the yeast GAPDH promoter consists of multiple general functional elements including RAP1 and GRF2 binding sites. *Journal of Veterinary Medical Science* 56:235-244 (1994).

Zaret KS and Sherman F> Mutationally altered 3' ends of yeast *CYC1* mRNA affect transcript stability and translational efficiency. *Journal of Molecular Biology* 176:107-135 (1984).

Figure 2.1

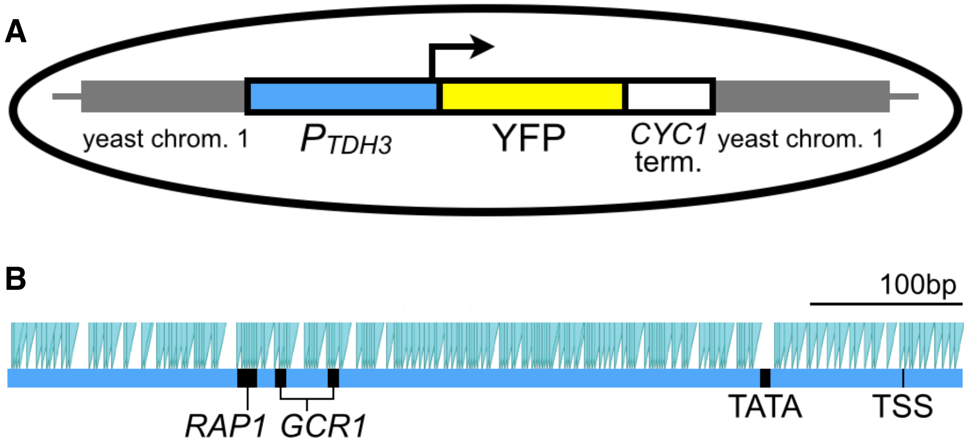

Figure 2.1 Overview of the reporter gene. **A.** The reporter gene containing the *TDH3* promoter (P_{TDH3}) driving expression of the reporter yellow fluorescent protein (YFP) flanked by the *CYC1* terminator integrated into chromosome 1 of the yeast genome. **B.** 236 single-nucleotide substitutions (▽) across 678bp of P_{TDH3}. Known functional elements, including binding sites for transcription factors *RAP1* and *GCR1*, are indicated [Holland *et al*. 1987; Pavlović & Hörz 1988; Kuroda *et al*. 1994; Yagi *et al*. 1994]

Figure 2.2

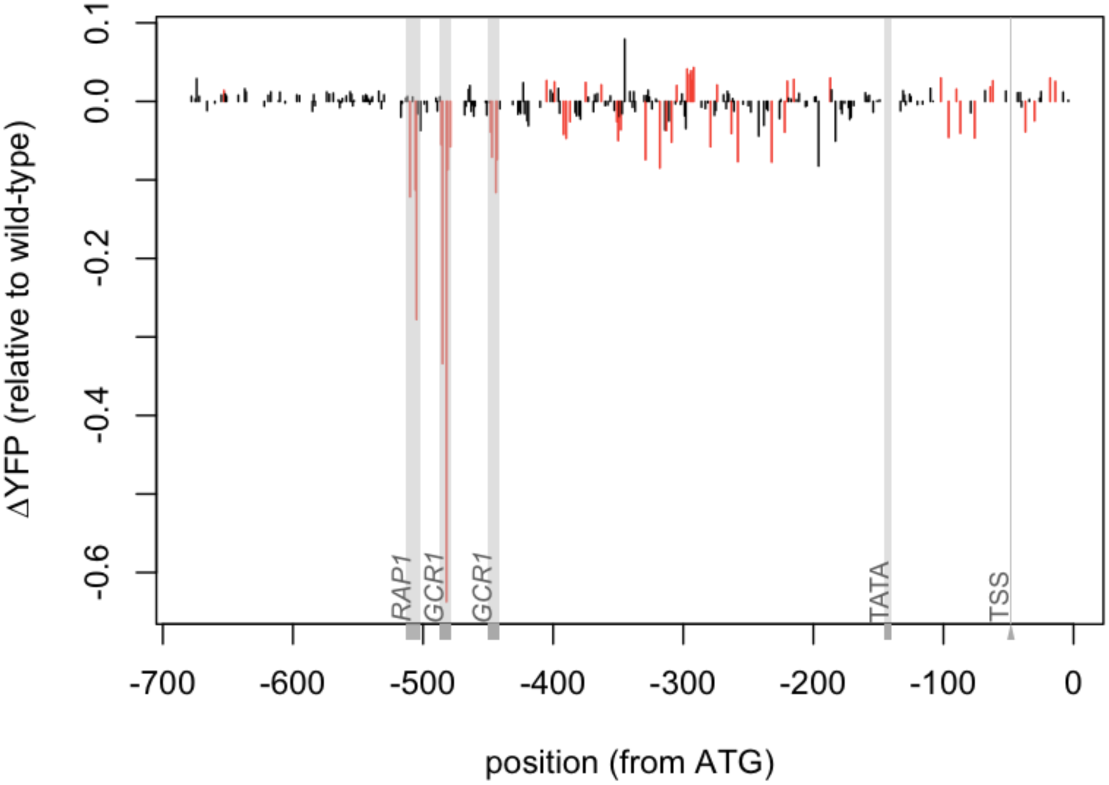

Figure 2.2. The effect of *cis*-regulatory mutations on fluorescence relative to that of wild-type plotted across P_{TDH3}. Red indicates significant difference from wild-type. Known functional elements are indicated.

Figure 2.3

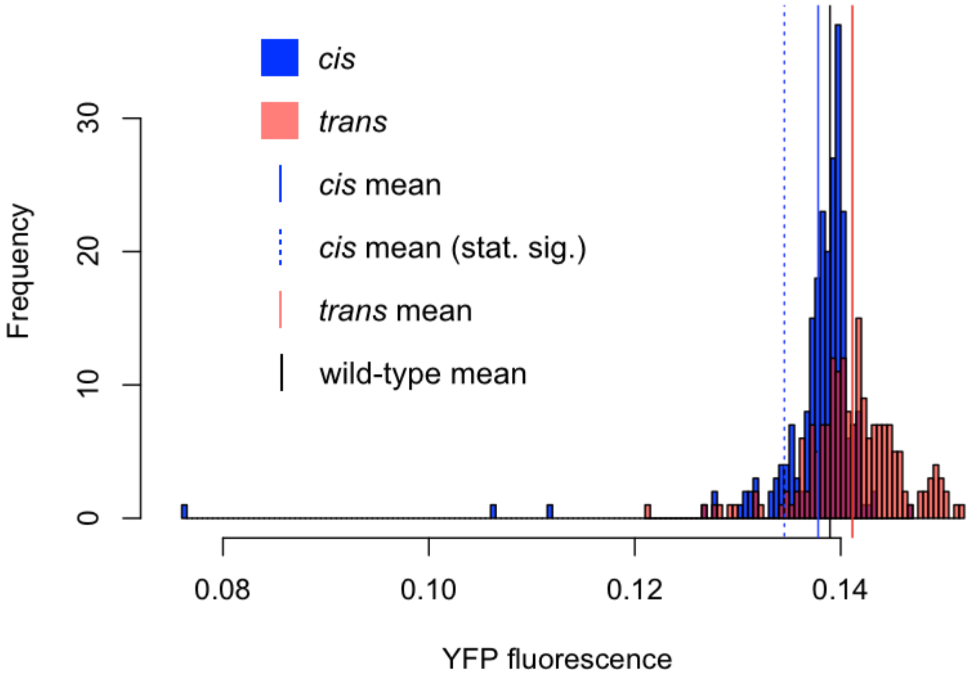

Figure 2.3. The distributions of expression phenotype as YFP fluorescence of *cis*-regulatory mutants (blue) plotted against that of *trans*-regulatory mutants from [Gruber *et al*. 2012]. Mean values are indicated as vertical lines. "*cis* mean (stat. sig.)" denotes mean YFP fluorescence of statistically significant *cis*-regulatory mutants with significantly altered expression from that of wild-type (i.e. red in Figure 2.2).

Figure 2.4

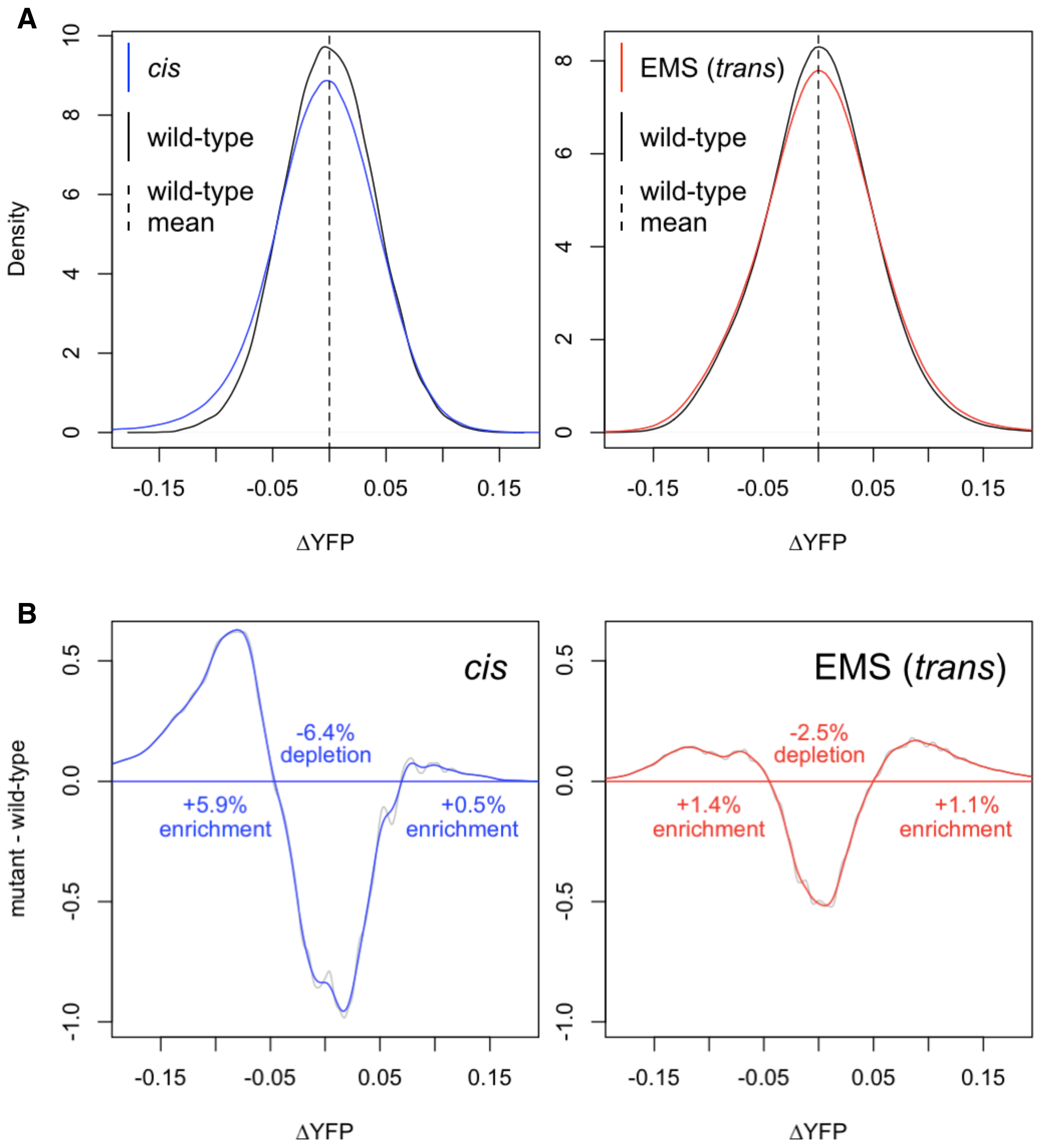

Figure 2.4 Comparison of *cis*-regulatory and EMS mutants. **A.** Distributions of effect on expression as YFP fluorescence of *cis*-regulatory mutants (left, blue) and EMS mutants (right, red) compared against that of wild-type (black) plotted as density curves. **B.** The difference between the distributions of mutant (left *cis*, right EMS) and wild-type effects on YFP. Corresponding blue (*cis*) and red (EMS) curves are spline fit to data (gray). Values above horizontal line represent phenotypes more abundant in the mutants (enrichment), while values below less abundant in mutants (depletion).

Chapter 3

Contrasting *cis*-regulatory effects of mutations and polymorphisms to test for selection[1]

Abstract

Genotypic and phenotypic variation segregating within a species reflects the combined activities of mutation, selection, and genetic drift. The mutation process introduces new alleles, while selection and drift change the frequency of these alleles based on their relative fitness and chance events. Disentangling the contributions of these processes to variation in the wild is an ongoing challenge for evolutionary biologists. This is especially true for non-coding sequences that play critical *cis*-regulatory roles in the control of gene expression [Fay and Wittkopp 2008; Zhen and Andolfatto 2012]. Here, we use an empirical approach to infer the relative impacts of mutation and selection on variation segregating in the *Saccharomyces cerevisiae* *TDH3* promoter (P_{TDH3}). We estimated the distribution of mutational effects for both

[1] This chapter represents a compilation of data and analyses currently being prepared for publication by Brian P. H. Metzger, David C. Yuan (co-first authors), Jonathan D. Gruber, Fabien Duveau, and Patricia J. Wittkopp. I contributed to designing the mutation spectrum project, creating the P_{TDH3}-YFP mutant and natural variant strains, performing flow cytometry experiments, and writing the manuscript.

gene expression level and noise (the variability in expression level among genetically identical individuals) using 236 point mutations in P_{TDH3} and compare them to distributions of effects for natural variants segregating among 85 diverse strains of *S. cerevisiae*. On average, the effects of natural variants on mean expression level showed little evidence of epistasis and were consistent with the mutation process. By contrast, natural variants exhibited significantly lower noise than expected from the mutational distribution, but only when natural variants were tested in the promoter context in which they occurred. Our data thus suggest that selection acted primarily to favor variants which decrease noise. This may reflect the relative rarity of mutations with large effects on expression level compared to the relative commonality of mutations with large effects on expression noise and shows how the distribution of mutational effects can shape *cis*-regulatory sequence evolution.

Introduction

Natural selection is often credited with shaping patterns of genotypic and phenotypic variation, but the mutation process that creates the variation upon which selection acts can also introduce biases [Stoltzfus & Yampolsky 2009]. Probabilistic mutation models that predict neutral patterns of variation in coding sequences are well-established and widely used as null models to test for evidence of selection, but comparable mutation models for non-coding sequences that control gene expression are still in their infancy [Fay and Wittkopp 2008; Zhen and Andolfatto 2012]. These models are challenging to construct because *cis*-regulatory sequences have complex relationships with gene expression that depend upon interactions with diverse *trans*-acting factors. As an alternative approach, we generated empirical distributions of mutational effects which solely reflect the action of the mutation process for a single *cis*-regulatory element [Patwardhan *et al.* 2009; Patwardhan *et al.* 2012; Melnikov *et al.* 2012; Kwasnieski *et al.* 2012]. We used these distributions as explicit null models on which to test for the action of natural selection by comparison to the effects of naturally-occurring variation.

Results and Discussion

Prior theoretical work suggests that both mean expression and expression noise can be acted on by natural selection and thus influence the evolutionary trajectory of a regulatory mutation [Wang & Zhang 2011]. While selection is typically thought to be stronger on mean expression, the relative strength of selection acting on mean expression and expression noise remains unknown. To determine this, we first generated null distributions of mutational effects for both mean expression and expression noise using a previously established quantitative assay of *cis*-regulatory activity by integrating the *TDH3* promoter (P_{TDH3}) and yellow fluorescent protein (YFP) reporter into the *S. cerevisiae* genome (P_{TDH3}-YFP) [Gruber *et al.* 2012].

We used site-directed mutagenesis to systematically introduce G:C → A:T transitions into individual haploid strains and used flow cytometry to quantify YFP fluorescence of individual cells. As illustrated in Figure 3.1A, we determined the effect of each mutation on mean expression level (μ) and expression noise (σ^2/μ^2, where σ is the standard deviation of μ) using measures of YFP fluorescence from ~10,000 single cells in each of 9 biological replicate populations. These mutations represent greater than 10% of all possible point mutations within P_{TDH3} and reflect the most frequent type of single nucleotide polymorphisms (SNPs) observed among natural isolates of *S. cerevisiae* [Maclean *et al. In prep*.] and the most frequent type of spontaneous point mutations in *S. cerevisiae* mutation accumulation lines [Lynch *et al.* 2008]. While P_{TDH3} has well-defined regulatory elements, including binding sites for the transcription factors *RAP1* and *GCR1* [Holland *et al*. 1987; Pavlović & Hörz 1988; Kuroda *et al*. 1994; Yagi *et al*. 1994], sequence conservation with other *Saccharomyces* species extends beyond these known transcription factor binding sites, suggesting the presence of additional unidentified functional elements (Figure 3.1B). In addition, P_{TDH3} has a degree of sequence polymorphism (frequency of polymorphic sites = 0.05) that is intermediate to that of synonymous (0.09) and non-synonymous sites (0.01) for the *TDH3* coding sequence, which is typically seen as evidence of functional constraint and the action of natural selection [Zhen and Andolfatto 2012]. However, the cause of this constraint is unknown.

We found that 59 of the 236 mutations tested caused statistically significant changes in mean expression (Figure 3.1C) and 40 caused statistically significant changes in expression noise (Figure 3.1D). Consistent with prior studies [e.g. Hornung *et al*. 2012], a significant negative correlation was observed between the effects of mutations on mean and noise. For both phenotypes, mutations showed an approximately exponential distribution of effects but contained an excess of large effects and fewer effects of intermediate size then commonly assumed in modeling mutational distributions [Eyre-Walker & Keightley 2007; Halligan & Keightley 2009]. Mutations with the largest effects on both mean expression and expression noise were located in previously identified transcription factor binding sites in P_{TDH3} (Figure 3.1C and D). We detected no statistically significant difference in the effects on mean fluorescence level or noise between G → A and C → T mutations (based on t-tests with significance threshold of $p = 0.05$), consistent with prior work showing similar distributions of phenotypic effects for all classes of nucleotide substitutions [Patwardhan *et al*. 2012]. As expected, mutations in sites that were more conserved amongst species had greater effects on both mean and noise than less conserved sites (Figure 3.1E and F).

Having characterized the mutational potential for P_{TDH3} activity, we sought to characterize variation in *TDH3* expression that exists in nature in order to assess the role of mutation and selection in shaping natural variation using our characterization of mutational effects in P_{TDH3} as an empirical baseline. First, to determine the relative contributions of *cis*- and *trans*-regulatory changes to *TDH3* expression divergence, we used pyrosequencing of total and allele-specific expression in 48 strains of *S. cerevisiae* collected from a wide range of environments relative to a common laboratory reference strain [Wittkopp *et al*. 2004]. These 48 strains encompass most of the sequence variation currently known to exist within this species [Liti *et al.* 2009; Maclean *et al*. *In prep.*]. Among all strains, we observed over five-fold variation in *TDH3* mRNA levels, with statistically significant effects of *cis*-regulatory variation in 36 of the 48 strains (Figure 3.2A). These *cis*-regulatory differences explained 1% to 75% of the total difference in *TDH3* regulation from the reference strain.

Next, we determined the effects of individual naturally-occurring *cis*-regulatory variants using our P_{TDH3}-YFP system. We first aligned homologous sequences from the 48 strains used for pyrosequencing as well as an additional 37 strains recently isolated from the wild. Among these 85 strains, we found 42 segregating sites: 36 SNPs and 6 insertions/deletions ranging from 1 to 32bp (Figure 3.2B). We tested for statistical association between these *cis*-regulatory variants and variation in *TDH3* mRNA levels estimated by pyrosequencing and identified six potential causal variants. This further suggests that *cis*-regulatory variants may contribute to *TDH3* expression variation. To empirically determine the effect of all variants in P_{TDH3}, we engineered each polymorphism individually into the haploid strain carrying the P_{TDH3}-YFP reporter gene and used flow cytometry to measure YFP fluorescence. Two of the 42 polymorphisms tested had significant effects, with both altering mean fluorescence (Figure 3.2C) and noise (Figure 3.2D).

We observed variation in P_{TDH3} activity among the naturally-occurring *cis*-regulatory variants, but is this variation the result of mutation or selection? To test for evidence of selection, we compared the effects of P_{TDH3} polymorphisms to the effects observed in our collection of systematic P_{TDH3} mutations for both mean expression and expression noise. In the absence of selection, the effects of naturally-occurring variants will be consistent with the effects of a random subset of mutations. However, the presence of selection will bias these effects. To test our null hypothesis of no difference in distributions of effect, we simulated the evolutionary process in the absence of selection by drawing subsets of mutations from our mutational distribution and comparing their average effect to that of naturally-occurring variation. We found that the effects on expression noise of individual polymorphisms tested were consistent with a random sample of the effects on expression noise drawn from the mutational distribution (t-test $p > 0.05$; Figure 3.3B), providing no evidence for selection. For mean expression, however, we observed a marginally significant difference between the effects of new mutations and polymorphisms observed in the wild (t-test $p = 0.051$; Figure 3.3A), with polymorphisms having on average smaller effects than systematic mutations. This suggests purifying selection against mutations with the largest effects on mean P_{TDH3} activity in the wild; however, this difference is non-significant when we limited the polymorphism dataset to only SNPs (t-test $p > 0.05$) or only G:C → A:T transitions (t-test $p > 0.05$) to more closely match the types of mutations used to determine the mutational distribution. Our comparison of mutations and polymorphisms on a common reference haplotype therefore provides weak evidence that natural selection has shaped segregating *cis*-regulatory variation affecting activity of the P_{TDH3}.

In the wild new mutations do not arise all on the same genetic background; instead, they arise randomly on haplotypes already existing in the population. Consequently, epistatic interactions can alter the actual effects of a new mutation. To determine the impact of epistasis within P_{TDH3}, we compared the effects of each polymorphism in the context of the reference allele to its effects on the promoter allele on which it most likely arose. We first constructed a haplotype network of segregating *S. cerevisiae* P_{TDH3} variation and determined the order the mutations most likely occurred using P_{TDH3} sequences from additional *sensu stricto* species. We then engineered each distinct P_{TDH3} haplotype into the haploid strain carrying the P_{TDH3}-YFP reporter gene and used flow cytometry to quantify mean expression and expression noise. The effect of each polymorphism in the context of its ancestral haplotype was determined by comparing YFP fluorescence for pairs of haplotypes that differed at a single site (Figure 3.4A). We found that 8 of the 45 polymorphisms significantly altered mean expression and 16 of 45 polymorphisms significantly altered expression noise in their respective haplotype contexts.[2]

[2] There are more polymorphisms here than segregating sites because a particular polymorphism may appear in multiple haplotypes and hence its effect would be tested multiple times.

By directly comparing the effects of individual polymorphisms on the reference and naturally-occurring haplotypes, we found evidence of epistasis affecting mean expression for 2 of the 43 polymorphisms tested (Figure 3.4B) and affecting expression noise for 3 of the 43 polymorphisms (Figure 3.4C). All 3 polymorphisms with significant epistatic effects on noise showed lower expression noise on the naturally-occurring genetic backgrounds than in the context of the reference allele. This is particularly striking given that all systematic mutations tested in the reference allele with significant effects on expression noise increased expression noise (Figure 3.1D). This suggests that P_{TDH3} haplotypes that minimize expression noise in response to new mutations may have been selectively advantageous in the wild.

To account for epistasis when testing for evidence of selection, we used the same sampling strategy as before but compared the null distribution to the distribution of effects observed from polymorphisms on the background they were inferred to occur on. We observed no significant difference for effects on mean expression between polymorphisms and the systematic mutations (t-test $p > 0.05$; Figure 3.4D). This difference was also not significant when restricting the polymorphisms tested to only SNPs or only G:C → A:T transitions to better match the mutational distribution (t-tests $p > 0.05$). Consequently, our data provide no evidence that selection has shaped the distribution of polymorphisms in terms of their mean effects on P_{TDH3} activity in natural populations. For expression noise, however, we observed that polymorphisms in their native context had significantly smaller effects than the systematic mutations (t-test $p < 0.05$; Figure 3.4E); a significant difference was also observed when considering only SNPs or only G:C → A:T transitions (t-tests $p > 0.05$). In addition, a significant difference was still observed after removing the effects of systematic mutations in previously characterized binding sites, suggesting that the action of selection was not restricted to them (t-test $p > 0.05$). This suggests that selection has favored genetic variants and haplotypes that decrease or maintain low levels of expression noise in natural populations.

By characterizing the effects of individual mutations and polymorphisms on *cis*-regulatory activity and comparing the two datasets, we empirically tested for selection among *cis*-regulatory variation observed in nature. For *TDH3*, our data suggest that the mutation process underlies the *cis*-regulatory variation in mean expression among the natural isolates sampled. However, our data suggest that selection has resulted in lowered expression noise via epistatic interactions among polymorphisms in the *TDH3* promoter. This finding underscores the importance of genetic background when estimating the effects of mutations and genetic variants. The selection detected on expression noise but not mean expression may be explained by the larger effect size of *TDH3 cis*-regulatory mutations for expression noise compared to that for mean expression (see Figure 3.1C and D). As a phenotype, expression noise may be a larger mutational target than mean expression for selection to act on. Our results illustrate how this difference in mutational effect between the mean and the variability of a phenotype can shape *cis*-regulatory evolution.

Material and Methods

Quantifying cis-regulatory activity of the TDH3 promoter

Construction of the P_{TDH3}-YFP transgene and strain was discussed in [Gruber *et al.* 2012]. Briefly, the P_{TDH3}-YFP transgene consists of the 678bp P_{TDH3} fused to reporter yellow fluorescent protein (YFP) coding sequence and the *CYC1* (cytochrome c isoform 1) terminator. The P_{TDH3} sequence includes the 5' intergenic region up to the start codon of *TDH3* coding sequence and contains the 5' untranslated region (UTR) of *TDH3* and the 3' UTR of the upstream gene *PDX1*. The transgene was integrated in a pseudogene on chromosome 1 of strain BY4724 at position 199270.

All strains were revived from glycerol stocks onto YPG (glycerol) at the same time to control for age-related effects on expression. Strains were inoculated from YPG solid media into 500µl of YPD (dextrose) liquid media and grown for 20 hours at 30˚C in 2ml 96-well plates with 3mm glass beads and shaken at 250rpm. Immediately prior to flow cytometry, 20µl of the overnight culture was transferred into 500µl of SC-R (dextrose) media.

Flow cytometry data was collected on an Accuri C6 using an intellicyt hypercyt autosampler. Flow rate was 14µl/min and core size was 10µm. A blue laser (λ = 480nm) was used for excitation of YFP. Data was collected from FL1 using a

533/30nm filter. Each culture was sampled for 2-3 seconds, resulting in approximately 20,000 recorded events.

Samples were processed using the flowClust [Lo *et al.* 2009] and flowCore [Hahne *et al.* 2009] packages within R and custom R scripts. All data was $\log_{10}$ transformed and artifacts were removed by excluding events with extreme FSC.H, FSC.A, SSC.H, SSC.A and width values. Samples were clustered based on FSC.A and Width to remove non-viable cells and cellular debris. Samples were then clustered on FSC.H and FSC.A to remove doublets. Finally, samples were clustered on FL1.A and FSC.A to obtain homogeneous populations. In all cases, two clusters were used and samples containing fewer than 1000 events after processing were discarded. For each sample, YFP expression was calculated as the median $\log_{10}(\mathrm{FL1.A})^2/\log_{10}(\mathrm{FSC.A})^3$ to control for effects of YFP amount on cell size and expression noise was calculated as $\mathrm{sd}(\mathrm{YFP})^2/\mathrm{mean}(\mathrm{YFP})^2$.

To control for variation in growth conditions, all plates contained 20 replicates of the wild-type reference strain. YFP expression from these controls was used to fit a linear model[3] and parameter estimates of the effect for each plate were subtracted from all samples on that plate. This correction was applied to both mean and standard deviation independently. A non-fluorescent strain was used to estimate auto fluorescence and correct all YFP expression values. For each strain, mean expression

[3] The parameters fitted were: day in which a plate was run, replicate that plate belongs to, plate, row/column/block(half plate) in a plate a sample was in, stack and depth in stack a plate was cultured in, and order in which a plate was run within a replicate.

and expression noise were reported as the average of nine biological replicates with independent overnight growths.

Determining the distribution of mutational effects

Mutant P_{TDH3} constructs, each containing a single *cis*-regulatory mutation, were created by site-directed mutagenesis and transformed into a *URA3*-YFP intermediate using the lithium acetate method and selection on 5-FOA [Gietz *et al.* 2007]. Site-directed mutagenesis was performed by PCR sewing of overlapping PCR products created with primers containing the targeted mutation. All mutants were confirmed by Sanger sequencing. The effect of each strain was compared to that of the wild-type P_{TDH3} strain to estimate effect on mean expression and expression noise. The wild-type P_{TDH3} strain was recreated three times to test for effects of background mutations on YFP expression.

Characterizing segregating variation in TDH3

A common allelic difference for pyrosequencing was created by inserting the *URA3* gene into the *TDH3* coding region in both BY4741 and BY4742. 80bp oligonucleotides were designed containing a synonymous (T243G, A81A) mutation with 40bp homology on each side of the target site. Successful transformants were confirmed by Sanger sequencing. Hygromycin B and KanMX4 resistance markers were inserted into the *HO* locus of the BY4741 and BY4742 derived strains respectively (Y360 and Y361 respectively) and used used to create a diploid (Y362).

Natural isolates of *S. cerevisiae* were obtained from J. Zhang [Maclean *et al. In prep*.]. Diploid strains were heterozygous for a KanMX4 resistance marker at the *HO* locus. All mating type α cells contained the same marker. All mating type a cells contained a Hygromycin B resistance marker at the *HO* locus.

Hybrids between each natural variant and Y360 were created by mixing cells in equal numbers on solid YPD. After 24 hours, cultures were streaked for singles and then patched to YPD containing G418 and Hygromycin B to select for diploids. All cultures were grown in 500μl of YPD liquid media for 20 hours at 30˚C in 2ml 96-well plates with 3mm glass beads shaken at 250rpm. Cultures were diluted to an OD_{600} of 0.1 and, for diploids cultures, mixed with an equivalent number of cells of Y362. Cultures were grown for an additional 4 hours and then centrifuged. YPD liquid was removed and cultures were placed in a dry ice/ethanol bath until frozen and then stored at -80˚C. For each strain, four biological replicates were used.

DNA and RNA were co-extracted from each culture using a modified protocol of Promega's SV Total RNA Isolation System. Cultures were thawed on ice and 175μl of SV RNA lysis buffer (with BME), 350μl of ddH20 and 50μl of 400 micron RNase free beads were added. Plates were vortexed until pellets were completely resuspended. Cultures were centrifuged and 175μl of supernatant was mixed with 25μl of RNase-free 95% ethanol and loaded onto a binding plate. To extract RNA, 100µl of RNase-free 95% ethanol was added to the flowthrough, loaded onto a second binding plate, washed twice with 500µl of SV RNA wash solution, and allowed to dry. To extract DNA, the first binding plate was washed twice with 700µl of cold 70% ethanol and allowed to dry. For both binding plates, 100µl of ddH$_2$0 was added to each well and incubated at room temperature for 7.5 minutes. Flowthrough was collected and diluted to 100µl. RNA was converted to cDNA by mixing 5µl of extracted RNA with 2µl RNase free water, 1µl DNase buffer, 1µl RNasin Plus, and 1µl DNase 1 and incubated at 37˚C for 1 hour followed by 65˚C for 15 minutes. 3µl of oligo dT (T19VN) was added and slow cooled to 37˚C. 4µl of First Strand Buffer, 2µl dNTPs, 0.5µl RNasin Plus, and 0.5µl of SuperScript II were added and incubated for 1 hour. 30µl of ddH20 was added. 1µl of cDNA or gDNA was used in a PCR. Pyrosequencing was performed as previously described using a PSQ 96 pyro sequencing machine and Qiagen pyroMark Gold Q96 reagents [Wittkopp 2011].

TDH3 promoter sequences were obtained by PCR and Sanger sequencing of diploid strains. Heterozygous strains were sporulated and haploid derivatives were sequenced to determine phase. The *TDH3* promoter haplotype network was created using all individual's haplotypes with TCS 1.21 [Clement *et al.* 2000]. Promoter sequences for all strains and *sensu stricto* species were aligned using Pro-Coffee [Notredame *et al.* 2000] and manually adjusted around repetitive elements and insertions/deletions. For all segregating sites within *S. cerevisiae*, the ancestral state was determined by parsimony and maximum likelihood in MEGA 5.1 [Tamura *et al*. 2011] using the *sensu stricto* phylogeny from [Hittinger 2013].

To determine *cis*-regulatory activity of full haplotypes, native *TDH3* promoters were PCR amplified and used to transform a *URA3*-YFP strain. Transformants were confirmed by Sanger sequencing. Intermediate haplotypes were created on individual haplotype background using site-directed mutagenesis as described for the *cis*-regulatory mutants but with naturally-occurring haplotypes instead of the laboratory reference haplotype. Descendant haplotypes were compared to their direct ancestral haplotypes to determine the effect of a mutation on the promoter background on which it arose.

Naturally-occurring haplotypes contain a single-bp difference from the reference haplotype. This difference has a significant effect on both mean expression and expression noise but is consistent across backgrounds. The effect of this difference on both mean expression and expression noise was subtracted from each of systematic *cis*-regulatory mutants when comparing effects of naturally-occurring variants on the reference background or on the haplotype on which they arose. In addition, several natural haplotypes are missing 6bp from the 5' end of the *TDH3* promoter. This difference has no significant effect on either mean expression or expression noise.

Association Test

Association testing was performed using Tassel 4.0 [Bradbury *et al*. 2007]. *Cis*-regulatory effects from pyrosequencing data were used in a general linear model and all segregating variants within the *TDH3* promoter were tested.

Epistasis

Epistasis was tested in R by fitting a linear model describing mean expression or expression noise as a function of genotype with and without an interaction for each mutation independently. In each case, the effect of a mutation was compared to the reference haplotype as well as between inferred ancestral and descendant naturally occurring haplotypes. Significance of interaction terms were tested using analysis of variance (ANOVA).

Comparing effects of new mutations and segregating variants

Mutational effects were sampled from all possible effects until an equal number of effects had been sampled as observed within segregating variation. Sampling was repeated 1,000,000 times and each time the mean was calculated. The mean effect of segregating variation was compared to this distribution to assess statistical significance.

References

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, and Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633-2635 (2007)

Clement M, Posada D, and Crandall K. TCS: a computer program to estimate gene genealogies. *Molecular Ecology* 9:1657-1660 (2000).

Eyre-Walker A and Keightley PD. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* 8:610-618 (2007).

Fay JC and Wittkopp PJ. Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* 100:191-199 (2008).

Gietz RD and Schiestl RH. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols* 2:31-34 (2007).

Gruber JD, Vogel K, Kalay G, and Wittkopp PJ. Contrasting Properties of Gene-Specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects, and Dominance. *PLoS Genetics* 8:e1002497 (2012).

Hahne F, LeMeur N,Brinkman RR, Ellis B, Haaland P, Sarkar D, Spidlen J, Strain E, and Gentleman R. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* 10:106-113 (2009).

Halligan DL and Keightley PD. Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. *Annual Review of Ecology, Evolution, and Systematics* 40:151-172 (2009).

Hittinger CT. *Saccharomyces* diversity and evolution: a budding model genus. *Trends in Genetics* 29:309-317 (2013).

Holland MJ, Yokoi T, Holland JP, Myambo K, and Innis MA. The *GCR1* gene encodes a positive transcriptional regulator of the enolase and glyceraldehyde-3-phosphate dehydrogenase gene families in *Saccharomyces cerevisiae*. *Molecular and Cell Biology* 7:813-820 (1987).

Hornung G, Bar-Ziv R, Rosin D, Tokuriki N, Tawfik DS, Oren M, and Barkai N. Noise–mean relationship in mutated promoters. *Genome Research* 22:2409-2417 (2012).

Kuroda S. Otaka S, and Fujisawa Y. Fermentable and nonfermentable carbon sources sustain constitutive levels of expression of yeast triosephosphate dehydrogenase 3 gene from distinct promoter elements. *Journal of Biological Chemistry* 269:6153-6162 (1994).

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, and Cohen BA. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *PNAS* 109:19498-19503 (2012).

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, *et al*. Population genomics of domestic and wild yeasts. *Nature* 458:337-341 (2009).

Lo K, Hahne F, Brinkman RR, and Gottardo R. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics 2009* 10:145-152 (2009).

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, and Thomas WK. A genome-wide view of the spectrum of spontaneous mutations in yeast. *PNAS* 105:9272-9277 (2008).

Maclean C, Metzger BPH, Yang J, Ho W-C, Moyers B, and Zhang J. Deep sequencing, population genetics and high-throughput phenotypic analysis of diverse *Saccharomyces cerevisiae* strains. *In preparation*.

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

Notredame C, Higgins DG, and Heringa J. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal of Molecular Biology* 302:205-217 (2000).

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, and Shendure J. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology* 27:1173-1175 (2009)

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, *et al*. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nature Biotechnology* 30:265-270 (2012).

Pavlović B and Hörz W. The chromatin structure at the promoter of a glyceraldehyde phosphate dehydrogenase gene from *Saccharomyces cerevisiae* reflects its functional state. *Molecular and Cellular Biology* 8:5513-5520 (1988).

Stoltzfus A and Yampolsky LY. Climbing Mount Probable: Mutation as a Cause of Nonrandomness in Evolution. *Journal of Heredity* 100:637-647 (2009).

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* 28:2731-2739 (2011).

Wang Z and Zhang J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *PNAS* 108:E67-76 (2011).

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

Wittkopp PJ. Using pyrosequencing to measure allele-specific mRNA abundance and infer the effects of *cis*- and *trans*-regulatory differences in *Molecular Methods for Evolutionary Genetics, Methods in Molecular Biology*, vol. 772, edited by Orgogozo V and Rockman MV. Humana Press, New York (2011).

Yagi S, Yagi K, Fukuoka J, and Suzuki M. The UAS of the yeast GAPDH promoter consists of multiple general functional elements including RAP1 and GRF2 binding sites. *Journal of Veterinary Medical Science* 56:235-244 (1994).

Zhen Y and Andolfatto P. Methods to Detect Selection on Noncoding DNA in *Evolutionary Genomics: Statistical and Computational Methods, Volume 2, Methods in Molecular Biology*, vol. 856, edited by Anisimova M. Humana Press, New York (2012).

Figure 3.1A-D

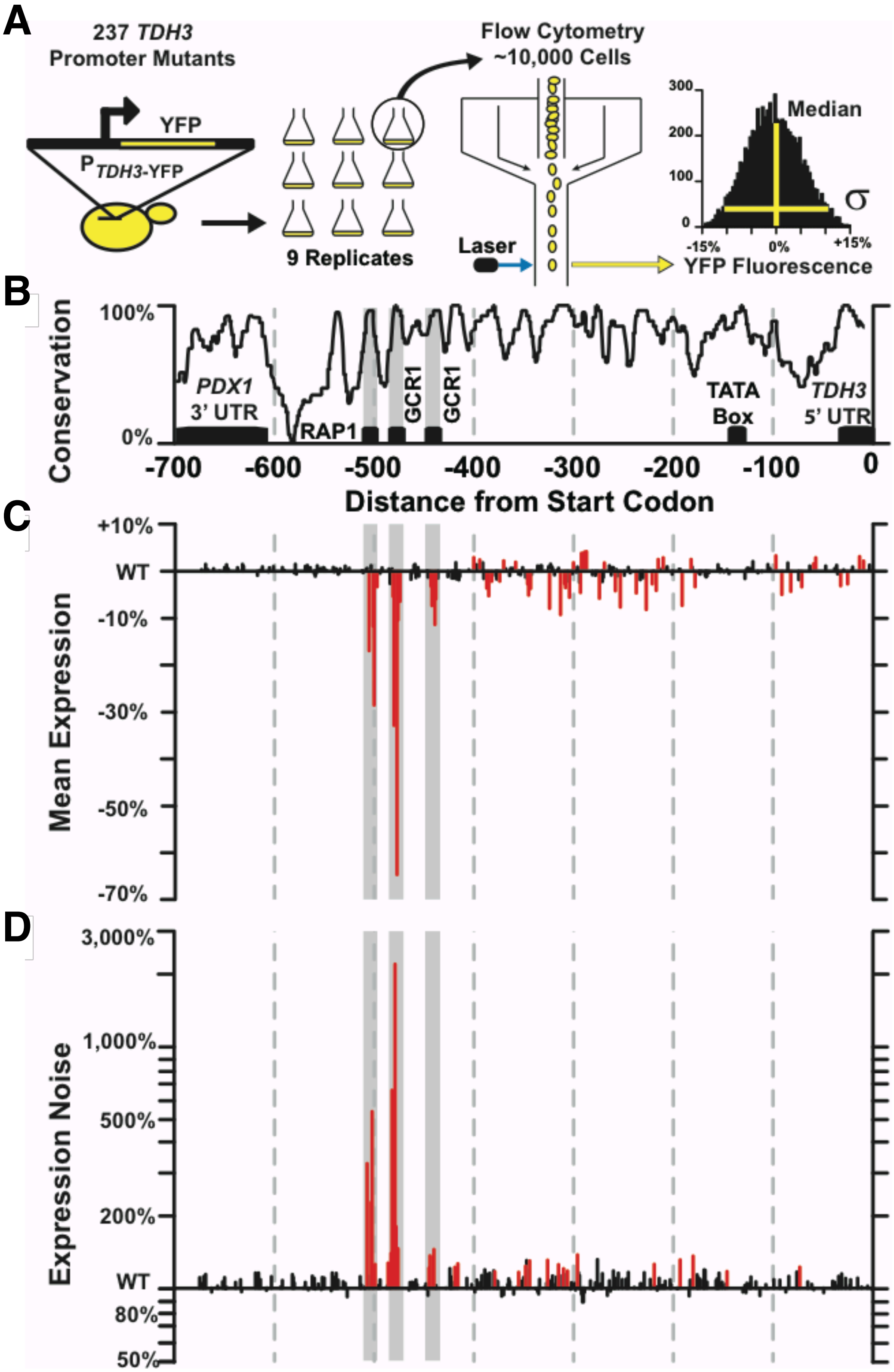
A
237 TDH3 Promoter Mutants
YFP
P TDH3-YFP
9 Replicates
Flow Cytometry ~10,000 Cells
Laser
300
200
100
0
-15%
0%
+15%
Median
σ
YFP Fluorescence
B
Conservation
100%
0%
PDX1 3' UTR
RAP1
GCR1
GCR1
TATA Box
TDH3 5' UTR
-700
-600
-500
-400
-300
-200
-100
0
Distance from Start Codon
C
Mean Expression
+10%
WT
-10%
-30%
-50%
-70%
D
Expression Noise
3,000%
1,000%
500%
200%
WT
80%
50%

Figure 3.1E-F

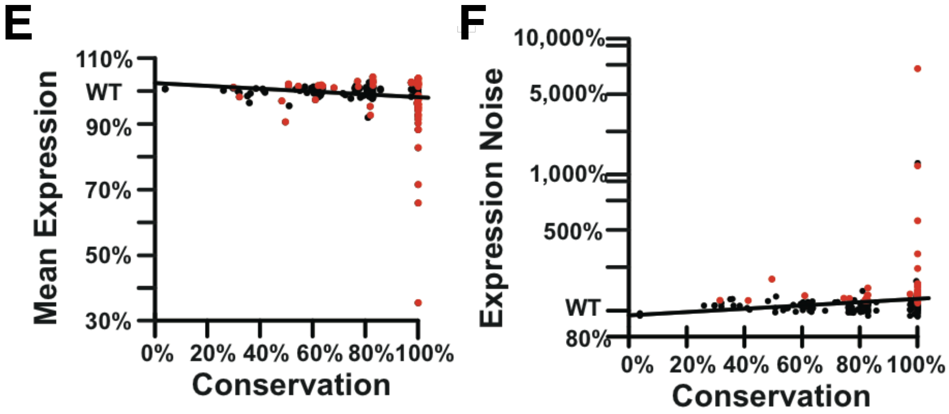

Figure 3.1 Most mutations within the *TDH3* promoter (P_{TDH3}) have little effect on *cis*-regulatory activity. **A**. Methodology for obtaining quantitative measurements of mean expression and expression noise using flow cytometry. **B**. *S. cerevisiae* P_{TDH3} with known functional elements shown as black boxes. Sequence conservation across the *sensu stricto* species is shown in black. Previously identified binding site positions are outlined in gray. **C**. Lines show the position and effect on mean expression relative to wild-type for each promoter mutation. Red lines are significantly different from wild-type (based on t-tests with Bonferroni-corrected significance threshold of $p = 0.0002$). **D**. Same as **C** but for expression noise. **E**. Mutations have significantly larger effects on mean expression at more conserved sites (based on ANOVA significance threshold of $p = 0.05$). **F**. Mutations have significantly larger effects on expression noise at more conserved sites (ANOVA $p < 0.05$).

Figure 3.2

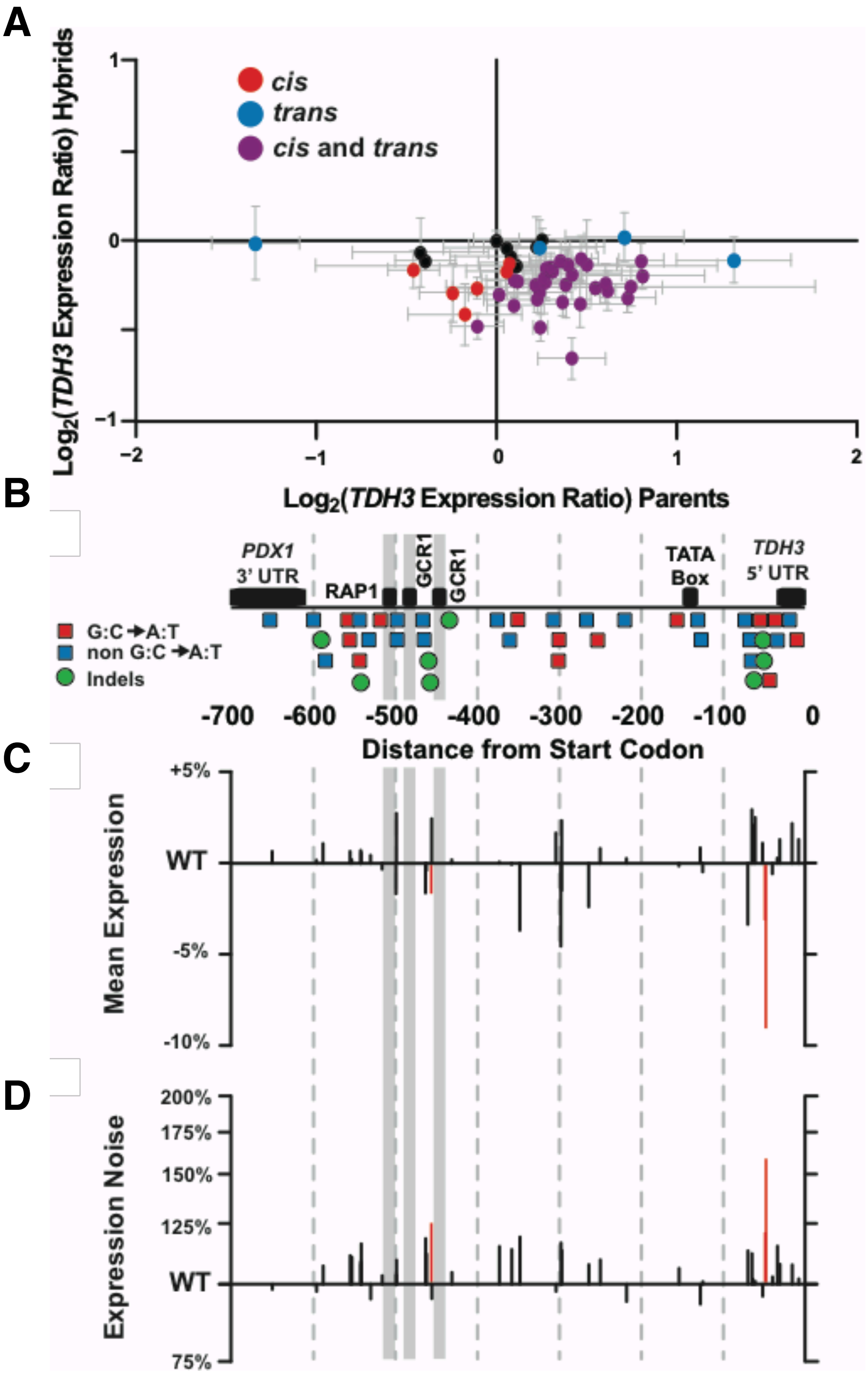
A
cis
trans
cis and trans
Log2(TDH3 Expression Ratio) Hybrids
Log2(TDH3 Expression Ratio) Parents
1
0
−1
−2
−1
0
1
2
B
PDX1
3' UTR
RAP1
GCR1
GCR1
TATA
Box
TDH3
5' UTR
G:C → A:T
non G:C → A:T
Indels
-700
-600
-500
-400
-300
-200
-100
0
Distance from Start Codon
C
Mean Expression
+5%
WT
-5%
-10%
D
Expression Noise
200%
175%
150%
125%
WT
75%

Figure 3.2 Extensive variation in *S. cerevisiae TDH3* expression. **A**. Log$_2$ ratio of total expression divergence between natural isolates and the reference strain (x-axis). Log$_2$ ratio of total *cis*-regulatory expression divergence between natural isolates and the reference strain (y-axis). Error bars represent 95% C.I. **B**. Location of segregating variation within the *TDH3* promoter indicated as boxes and circles in color. Black boxes indicate known functional elements. Previously identified binding sites positions are outlined in gray. **C**. Lines show the position and effect on mean expression relative to wild-type for each natural variant. Red lines are statistically different from wild-type (based on t-tests with Bonferroni-corrected significance threshold of $p = 0.001$). **D**. Same as **C** but for expression noise

Figure 3.3

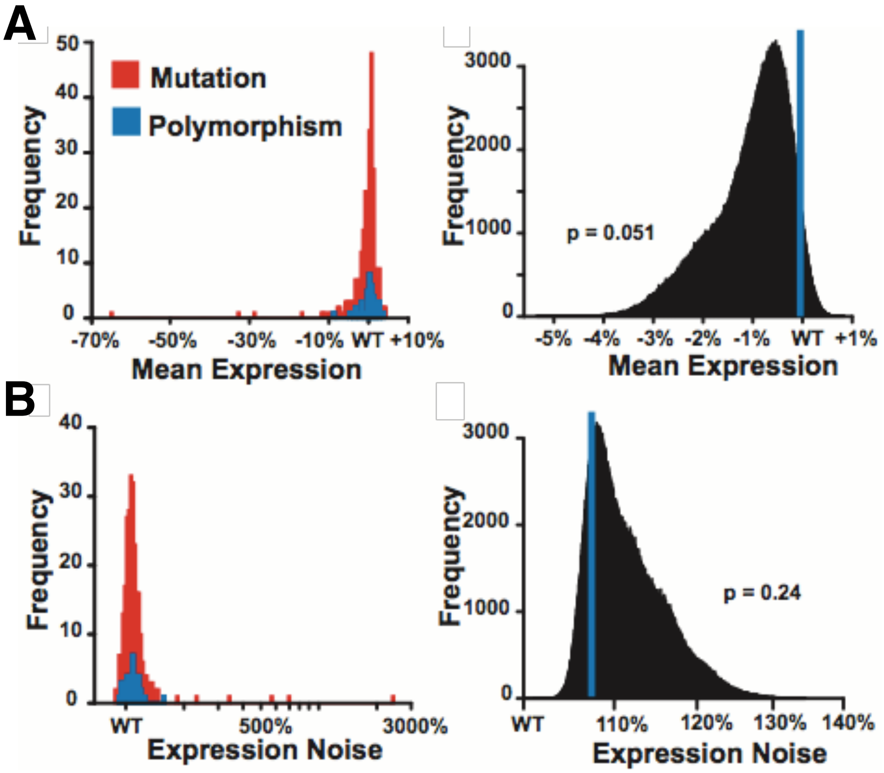

Figure 3.3 Comparison of the effect of *TDH3 cis*-regulatory mutations and that of individual polymorphisms. **A**. Left panel shows the distribution of mutational effect on mean expression level (Red indicates mutational distribution, blue indicates distribution of naturally-occurring variants). Right panel shows the simulation of evolution without selection by repeatedly drawing mutations from the mutational distribution and calculating mean expression, resulting in the distribution shown in black. Blue line indicates the average effect of the naturally-occurring variants. *P*-value is the percentage of pulls with a larger effect on mean expression than the naturally-occurring variants. **B**. Same as **A** but for expression noise.

Figure 3.4

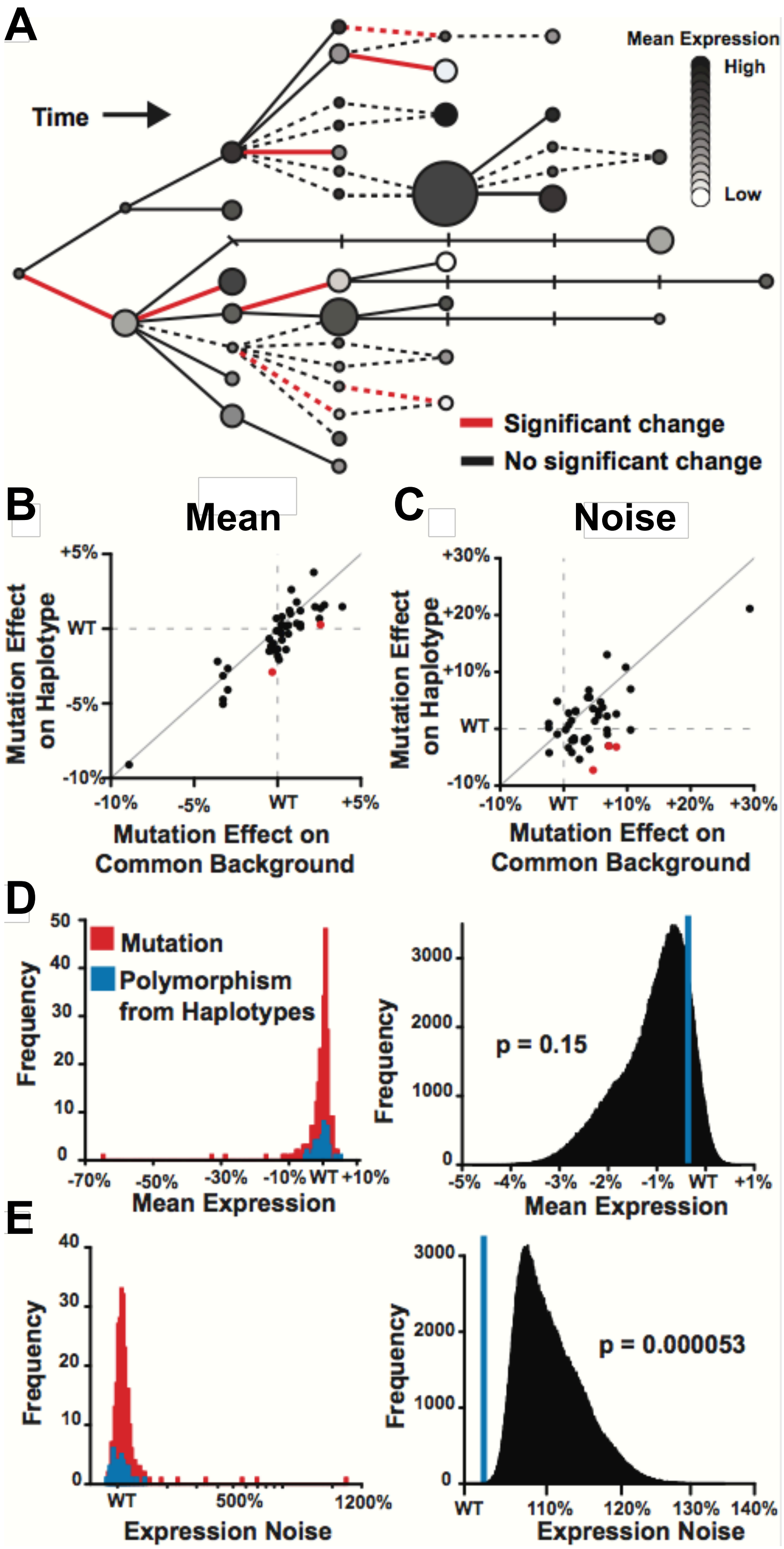
A
Time
Mean Expression
High
Low
Significant change
No significant change
B
Mean
Mutation Effect on Haplotype
+5%
WT
-5%
-10%
-10%
-5%
WT
+5%
Mutation Effect on Common Background
C
Noise
Mutation Effect on Haplotype
+30%
+20%
+10%
WT
-10%
-10%
WT
+10%
+20%
+30%
Mutation Effect on Common Background
D
Frequency
50
40
30
20
10
0
Mutation
Polymorphism from Haplotypes
-70%
-50%
-30%
-10%
WT
+10%
Mean Expression
Frequency
3000
2000
1000
0
p = 0.15
-5%
-4%
-3%
-2%
-1%
WT
+1%
Mean Expression
E
Frequency
40
30
20
10
0
WT
500%
1200%
Expression Noise
Frequency
3000
2000
1000
0
p = 0.000053
WT
110%
120%
130%
140%
Expression Noise

Figure 3.4 Comparison of the effect of *TDH3 cis*-regulatory mutations and that of polymorphisms in haplotypes. **A**. *TDH3* promoter haplotype network. Each circle represents a unique haplotype with size of circle proportional to frequency within the species. The inferred ancestral haplotype is shown at the left. Related haplotypes are connected by solid lines if both haplotypes were sampled or dashed lines if one of the haplotypes was not sampled. Shading is proportional to the haplotype's mean expression level. Natural variants with significant effects on the backgrounds they occurred are shown with red lines connecting to their background haplotypes. **B**. Comparison of the effects of each natural variant on mean expression when estimated on a single background (x-axis) versus the background on which it arose (y-axis). Mutations with significant epistasis are indicated in red. **C**. Same as **B** but for expression noise. **D**. Same as Figures 3.3A and 3.3B but using effects of naturally-occurring variants estimated from haplotypes (blue). **E**. Same as Figures 3.3C and 3.3D but using effects of naturally-occurring variants estimated from haplotypes (blue).

Chapter 4

Phenotypic plasticity and genotype-by-environment interactions among *TDH3 cis*-regulatory mutations

Abstract

Mutation is the original source of genetic variation and hence provides the raw material for evolution. A newly-arisen mutation may be fixed by genetic drift or, depending on its relative fitness, selection. Characterizing the effects of mutations in multiple environments is important because genetic variants can exhibit phenotypic plasticity (phenotypic variation in response to environmental variation) or genotype-by-environment (GxE) interaction (different level of phenotypic plasticity due to genetic variation). Such characterization requires determining the phenotypic effects of a large collection of mutations in different environments. I previously characterized the effects on gene expression of a large collection of *cis*-regulatory mutations in yeast using a reporter gene driven by the promoter of the *Saccharomyces cerevisiae* gene *TDH3* integrated into the *S. cerevisiae* genome. This system is well-suited for characterizing the impact of environment on effects of a mutation at a specific locus. *TDH3* is a

glycolytic gene, and its function and regulation are well-characterized. Prior work has shown that *TDH3* expression is regulated by different functional elements in different environments. Here I characterize the spectrum of *TDH3 cis*-regulatory mutations in several environments, specifically testing for phenotypic plasticity and GxE interactions in the presence of fermentable and non-fermentable carbon sources as well as osmotic stress. Both mean expression level as well as expression noise were found to exhibit phenotypic plasticity among environments, while over half of the *cis*-regulatory mutations tested showed different effects in different environments. Most of these GxE interactions were associated with differences in carbon source for mean expression level. The evolutionary implications of these results are discussed in light of the role mutation plays in generating genetic and ultimately phenotypic variation.

Introduction

A common goal across much of biology is to understand the genetic basis of phenotypes ranging from morphology to complex human disease. The genetic component is not solely responsible for the manifested phenotype, however; there often exists an interplay between genotype and environment. For instance, cancer development may involve not just the genotype with predisposed risk but also environmental influences. Indeed, a recent study identified interactions between genetic variants and environmental factors influencing the risk for breast cancer [Nickels *et al*. 2013]. From an evolutionary perspective, the environment is a key actor shaping variation and diversity. Not only is environment the agent through which selection acts, it can also influence how phenotypes are manifested from genotypes. The interaction of genotype with environment thus has wide-reaching implications.

Variation in how different genotypes respond to environmental differences is termed genotype-by-environment (GxE) interaction. This is related to but distinct from phenotypic plasticity, which is variation of phenotype in response to environmental change. GxE interactions occur when genetic variation influences how genotypes manifest as phenotypes in different environments. This distinction is illustrated in Figure 4.1. In Figure 4.1A, genetic variation among different genotypes manifests as different phenotypes that are invariable in different environments. By contrast, Figure 4.1B illustrates phenotype that varies by environment—phenotypic plasticity—while little or no genetic variation exists. Figure 4.1C illustrates both genetic variation and phenotypic plasticity. Genetic variation produces phenotypic variation. The phenotype also varies by environment, but all genotypes exhibit the same pattern of phenotypic plasticity. By contrast, genotypes in Figure 4.1D exhibit different patterns of phenotypic plasticity. Genetic variation produces variations in both phenotype and how that phenotype changes with the environment, illustrating the effect of GxE interactions.

GxE interactions have been documented for many organisms. In *Drosophila melanogaster*, for instance, interactions between larval-rearing environment and genetic variation at candidate loci explain variation in olfactory behavior [Sambandan *et al*. 2008]. GxE interactions can also impact viability and fitness. Variation in expression of environmental stress response genes in yeast correlates with variation in viability in different environments [Kvitek *et al*. 2008]. In addition, [Remold & Lenski 2001] observed random mutations exhibiting GxE interactions affecting fitness in *Escherichia coli*. Variable phenotypes resulting from GxE interactions can lead to complex evolutionary trajectories of the underlying genetic variants [Landry *et al*. 2006]. For instance, GxE interactions can unmask cryptic genetic variation and contribute to evolution of phenotypes otherwise invariable in steady conditions [Gibson & Dworkin 2004]. GxE interactions can thus influence the evolutionary process.

Evolution of phenotypes may involve changes in gene regulation [e.g. Wittkopp *et al*. 2002; Chan *et al*. 2010]. An ongoing challenge is to understand how such changes mechanistically alter gene expression and ultimately the higher-level phenotype. Molecular mechanisms in transcription (e.g. genetic changes impacting transcription factor binding to *cis*-regulatory element) is often used as a functional model of regulatory evolution, but how does the environment impact such mechanisms? Identifying and understanding the environmental conditions that influence the effect of regulatory variants can shed light on mechanisms contributing to phenotypic variation and evolution [Maranville *et al*. 2012]. To this end, prior studies have characterized phenotypic plasticity and GxE interactions in gene expression.

Phenotypic plasticity involving differential gene expression has been documented in, for instance, variation in wing morphology of ants associated with expression variation due to different environmental cues [Abouheif & Wray 2002]. Genetic variation underlying GxE interactions in gene expression has also been investigated. For instance, [Landry *et al*. 2006] profiled genome-wide transcriptional expression across six *Saccharomyces cerevisiae* natural isolate strains in different nutrient environments and found more genes exhibiting genetic variation for phenotypic plasticity—GxE interactions—than not. To further characterize the genetic architecture underlying GxE interactions in gene expression, [Smith & Kruglyak 2008] expanded to 109 segregant strains and mapped thousands of loci in the yeast genome showing GxE interactions for different nutrient environments. Their results were consistent with findings of [Li *et al*. 2006] for *Caenorhabditis elegans* cultured in different temperatures. [Hodgins-Davis *et al*. 2012] also observed rampant GxE interactions in *S. cerevisiae* gene expression across ecologically relevant copper concentrations. These studies shed light on the global pattern of loci involved in GxE interaction, but the resolution does not allow detailed characterization of how regulatory changes respond to environmental changes at the level of single nucleotides.

Genetic changes, such as those that impact gene regulation, result from the mutation process and are the raw material for evolution. The evolutionary trajectory of a newly-arisen mutation may depend on genetic drift (i.e. random chance events) or selection. Selection acts on the effect of a mutation on phenotype, with fitness being the ultimate phenotype. Since phenotypic plasticity may alter the effect of mutations systematically (e.g. by the same magnitude and direction across all mutations), its impact on the evolutionary fate of different mutations in different environments may be negligible. This is not the case with GxE interactions, as the change in effect of different mutations may differ by environment. Characterizing the impact of phenotypic plasticity and GxE interaction on the effect of mutations can improve our understanding of the evolutionary contribution of the mutation process.

Using a reporter gene driven by the promoter of the *S. cerevisiae* gene *TDH3* integrated into the *S. cerevisiae* genome (Figure 4.2), I previously characterized a collection of *cis*-regulatory mutations that impact *TDH3* expression with the goal of understanding the contribution of the mutation process to expression variation [Metzger *et al*. *In prep*]. These *cis*-regulatory mutations represent a spectrum of potential newly-arisen mutations and were characterized in the standard laboratory nutrient environment: glucose. Glucose is the carbon source metabolized in glycolysis, in which the *Tdh3p* protein (glyceraldehyde-3-phosphate dehydrogenase, isozyme 3) plays a role [McAlister & Holland 1985]. *Tdh3p* also participates in gluconeogenesis, the “opposite” metabolic pathway converting non-fermentable carbon sources (e.g. glycerol or ethanol) into fermentable glucose [McAlister & Holland 1985]. Prior studies identifying functional elements in the *TDH3* promoter (P_{TDH3}) revealed regions in P_{TDH3} active in different carbon sources. Specifically, a region 528bp upstream from the coding sequence may be active primarily during metabolism of fermentable carbon sources while another region 309bp from the coding sequence is active when metabolizing non-fermentable carbon sources [Kuroda *et al*. 1994] (Figure 4.2B). The effect of mutations in these regions may vary depending on carbon source context. Besides carbon source metabolism, *Tdh3p* is involved in the regulation of cellular osmolarity and has been found in the cell wall [Delgado *et al*. 2001; O’Rourke *et al*. 2002; Hyduke & Palsson 2010]. Thus, osmotic stress may also influence expression of *TDH3*.

TDH3 is active in glycolytic, gluconeogenic, and high osmolarity environments, but do its *cis*-regulatory mutations have the same effect in all environments? If not, is the effect of the environment on one genotype the same on all genotypes? Here I characterize the effects of *TDH3 cis*-regulatory mutations in different environments to test for phenotypic plasticity and GxE interactions, with the goal of gaining insight on how environmental variation may affect regulatory variation. Based on *TDH3*’s function in carbon metabolism and osmolarity regulation along with the environment-dependent activity of P_{TDH3}, I investigate the effects of fermentable and non-fermentable carbon sources and osmotic stress on *TDH3 cis*-regulatory activity.

Results

P_{TDH3} activity is plastic across environments

To examine variation in *cis*-regulatory activity of *TDH3* across environments, I used a reporter transgene system consisting of the 678bp *P_{TDH3}* driving expression of reporter yellow fluorescent protein (YFP) whose fluorescence, as quantified in flow cytometry, was used to assay *P_{TDH3}* activity (Figure 4.2A). A wild-type reference strain —one carrying the wild-type *P_{TDH3}* in the reporter gene—was cultured in media with fermentable (glucose and galactose) or non-fermentable (glycerol and ethanol) carbon source as well as media that induces osmotic stress (sorbitol) prior to quantification of *P_{TDH3}* activity. Both mean expression level (YFP fluorescence) and expression noise (variance of YFP fluorescence normalized by mean expression level) were estimated from the flow cytometry data, as variation in both phenotype and its variability can be important contributors to evolution [Raser & O'Shea 2004; Wang & Zhang 2011].

As shown in Figure 4.3, the reference strain exhibited phenotypic plasticity for *P_{TDH3}* activity across the environments tested. This was the case for both mean expression level (Figure 4.3A) as well as expression noise (Figure 4.3B). Specifically, *TDH3* mean expression level as measured by reporter fluorescence was significantly higher in sorbitol (osmotic stress) and glycerol and ethanol (non-fermentable carbon sources) and significantly lower in galactose (fermentable) compared to that in glucose (fermentable), based on t-tests with a significance threshold of $p = 0.05$.

When pooling the data by fermentability of the carbon source, *TDH3* mean expression level was significantly higher in non-fermentable carbon sources (t-test $p < 0.05$). Expression noise was also significantly higher in glycerol and ethanol and significantly lower in galactose compared to that in glucose (t-tests, significance threshold of $p = 0.05$). However, expression noise in sorbitol was not significantly different from that in glucose (t-test $p > 0.05$). Again, higher expression noise was observed in non-fermentable carbon sources compared to that in fermentable carbon sources (t-test $p < 0.05$). P_{TDH3} activity thus exhibits phenotypic plasticity in the environments tested.

TDH3 cis-regulatory mutants exhibit expression plasticity across environments

To determine whether the effects of *TDH3 cis*-regulatory mutants vary in different environments, the 236 mutant strains—each carrying a single G → A or C → T transition in P_{TDH3} (Figure 4.2B)—were tested similarly as above. The effect of the *cis*-regulatory mutations on mean expression level relative to that of the wild-type reference strain is shown in Figure 4.4. Both qualitatively and quantitatively, mutation effect on mean expression level (ΔYFP) differed between fermentable and non-fermentable carbon sources. The large effects exhibited by mutations in known transcription factor binding sites (TFBS) in fermentable carbon sources were absent in glycerol. This result is consistent with prior work carried out in glycerol showing that this region—which contains the known TFBS—is functional only in fermentable carbon sources [Kuroda *et al*. 1994]. The distributions of both mean expression level and expression noise among the *cis*-regulatory mutants significantly differed in all environments tested from that in glucose (t-test $p < 0.05$) with the exception of expression noise in glycerol (t-test $p > 0.05$) (Figure 4.5). A majority of all 236 *cis*-regulatory mutants exhibited significantly different mean expression levels across the environments tested compared to glucose while up to 80 exhibited significantly different expression noise (significance threshold based on false discovery rate of 0.05; Figure 4.5). The number of significant differences in expression noise differed dramatically among environments, with 80 observed in ethanol and only 1 or 2 observed in each of the other three environments other than glucose. However, the power to detect differences among expression noise may be less than that among mean expression level because expression noise has higher variance than mean expression level.

Cis-regulatory mutations in the TDH3 promoter exhibit GxE and GxGxE interactions

Do *cis*-regulatory mutations in P_{TDH3} exhibit genotype-by-environment interactions? This possibility was tested using a linear model that includes an interaction term between genotype (mutation) and environment (see Materials and Methods). The results are summarized in Figure 4.6. Using the glucose environment as reference, over half of the *cis*-regulatory mutations tested were found to exhibit significant GxE interaction on mean expression level. The highest number of interactions was found in non-fermentable carbon sources compared to glucose with 41 (out of 236) in glycerol and 136 in ethanol, while 4 were found in galactose and 6 in sorbitol (FDR $q < 0.05$). In contrast, GxE interaction was observed much less frequently for expression noise, with 1 in galactose, 2 in glycerol, and 1 in ethanol (FDR $q < 0.05$). Figure 4.7 illustrates mutations that exhibited GxE interaction and their positions along P_{TDH3}. All three known TFBS included mutations affected by GxE interaction, with the upstream *GCR1* site having the most overlap.

GxE interaction is a mechanism that can generate phenotypic variation through the environmental context in which genotypes exert their effects on phenotype. An additional level of such interaction—genotype-by-genotype-by-environment (GxGxE) interaction—was also observed among the *cis*-regulatory mutations. GxGxE interaction incorporates interactions of genetic variants with other genetic variants (epistasis) as well as the environment (GxE interaction). Despite the potential evolutionary impact of such interaction [Wade & Goodnight 1998], few studies have investigated its influence on the effect of individual mutations [Flynn *et al*. 2013]. The wild-type reference strain from which all *cis*-regulatory mutants were constructed differs from many other common yeast genetic backgrounds by an A → G transition -293bp upstream from the start of the coding sequence (G_{-293}). To test the effect of this genetic variant on gene expression, a subset of *cis*-regulatory mutants were reconstructed to exclude this background genetic variant (i.e. $G_{-293} \rightarrow A_{-293}$) and characterized across environments.

As shown in Figure 4.8, the effect of G_{-293} on mean expression level in glucose was consistent (i.e. systematically decreasing expression) among most of the *cis*-regulatory mutations tested but not in other environments. Using a linear model to test for interaction among G_{-293}, genotype, and environment revealed up to 19 out of 29 *cis*-regulatory mutations tested to exhibit GxGxE interactions with G_{-293} on mean expression level (FDR $q < 0.05$), with significant interactions observed more frequently between non-fermentable carbon sources and glucose (Figure 4.8). However, no significant interaction was detected on expression noise.

Discussion

This study examines the impact of environment on the effect of mutations on gene expression. The mutation process generates genetic variation that can influence phenotypic variation in natural populations. While newly-arisen mutations may persist in a population due to genetic drift, the fixation or elimination of a mutation by selection depends on the effect of that mutation. What can influence the effect of a mutation? The characteristics of the mutation itself are certainly important, and these include its mechanism of action, allele specificity, degree of dominance, interaction with other genetic variants, etc. The environment can be another influence. Environmental cues may elicit specific mechanisms that can impact how the effect of a mutation is manifested. By characterizing the effect of a spectrum of mutations in different environments, this study aims to deepen our understanding how the mutation process contributes to phenotypic variation.

Potential mechanism of variable expression across environments

Rather than testing mutants in a large number of environments in an exploratory fashion, the environments in this study were chosen based on functional and regulatory characterizations of *TDH3*, the gene whose promoter was used for analysis. Specifically, [Kuroda *et al*. 1994] revealed variable activity in different regions of P_{TDH3} during metabolism of fermentable versus non-fermentable carbon sources. The three known TFBS—for which empirical data exist for both function and binding of transcription factors—fall within the fermentable region (Figure 4.1B).

In glucose and galactose, mutations in these TFBS had the most dramatic effect on mean expression level. In glycerol, however, the same mutations did not exhibit effect of such magnitude. This is consistent with findings by [Kuroda *et al*. 1994], whose experiments were carried out with glycerol. In ethanol, the other non-fermentable carbon source tested, these mutations nevertheless had large effects as they did in fermentable carbon sources. This suggests that some functional elements in the fermentable region may be active when metabolizing non-fermentable carbon source other than glycerol. However, the profile and distribution of mutation effect on mean expression level in ethanol were still distinct from that of glucose and galactose, with mutations in the non-fermentable region exhibiting larger effects in ethanol. This regional difference may also explain the GxGxE interaction observed with the background genetic variant in the wild-type reference strain. This variant, G_{-297}, is located within the non-fermentable region; consistently, GxGxE interaction was observed much more frequently in non-fermentable carbon sources (Figure 4.8).

Curiously, while prior studies showed the non-fermentable region as critical for expression, no mutation within this region exhibited the same magnitude of effect in glycerol and ethanol as those in the known TFBS did in glucose and galactose. Given the lack of functionally-characterized TFBS in this region, the functional elements here may act through means other than transcription factors (e.g. histone binding), although it is possible that, by targeting only G and C sites, some functional sites were missed. Nevertheless, the difference in mechanism underlying the effect of different mutations suggest the importance of environment on how mutations manifest their effects.

While osmotic stress increased mean expression level overall, the profile of mutation effect and the pattern of GxE interactions in sorbitol resembled those in galactose. This suggests that osmotic stress-specific regulation of *TDH3* may depend on factors outside of P_{TDH3}. This may be consistent with the role *Tdh3p* protein plays in hyperosmotic response. When facing increased environmental osmolarity, *S. cerevisiae* synthesizes glycerol as solute to balance cellular osmolarity [Blomberg *et al*. 1992; O'Rourke *et al*. 2002]. The glycerol biosynthetic pathway is coupled with decreased production of pyruvate (a byproduct of glycolysis) effected in part by the enzyme Hog1 decreasing the activity of *Tdh3p* (a substrate of Hog1) [Hyduke & Palsson 2010]. Thus, increased *TDH3* expression in sorbitol may be a compensatory response to decreased number of *Tdh3p* available for carbon metabolism due to sequestration by Hog1. The similar profiles of mutation effect in sorbitol, glucose, and galactose suggest lack of functional elements in *P_{TDH3}* specific to osmotic stress. Rather, the overall increased expression in sorbitol may be effected by, for instance, feedback mechanisms sensitive to *Tdh3p* abundance.

Observed GxE interactions consistent with prior findings among genes with paralogs

Both phenotypic plasticity and GxE interactions affected *TDH3 cis*-regulatory mutants tested in this study, with GxE interactions observed for over half of the genotypes tested. Given the regulation and range of function of *TDH3*, the effect of GxE interactions may not be surprising. This might not be the case for genes without environment-specific regulation or whose function is restricted to a single environment. Interestingly, *TDH3* has two paralogs yet all three *TDH* paralogs are functionally-related, suggesting possible subfunctionalization. In addition, a prior study examining expression divergence of duplicated genes in yeast did not find evidence of neofunctionalization for *TDH3* [Tirosh & Barkai 2007].

The three *TDH*s in *S. cerevisiae* encode distinct polypeptides, differ in promoter sequences, are expressed differently throughout the cell cycle, and perform related but distinct functions [McAlister & Holland 1985; McAlister & Holland 1985; Boucherie *et al*. 1995; Delgado *et al*. 2001]. As paralogs functionally diverge following duplication, those that become subfunctionalized may perform more specialized functions in pathways related to the original. They may even retain the original function, at least initially. This opens up the potential for GxE interaction for such paralogs, as the new function may be under novel regulation. Consistent with this, [Landry *et al*. 2006] found GxE interactions more prevalent among genes with paralogs. It may then be interesting to examine GxE interactions in a gene that has been neofunctionalized and retains little to no functional connection to its ancestor.

Impact of environment on the evolutionary trajectory of newly-arisen mutations

Given that mutation provides the raw material for evolution, how might the mutation process specifically contribute to evolution? This involves the interplay among mutation, genetic drift, and selection within a population. Genetic drift is the change in frequency of an allele due to random chance, the rate of which is dependent on population size. As the mutation process inputs novel genetic variation into a population, some of this genetic variation may immediately or quickly be lost due to lethality, another degree of fitness cost, or death of host individual simply by chance. Others, however, may remain and reach an appreciable frequency due to genetic drift and/or selection. Any selective force that exists or arises may speed up the rate of fixation of genetic variants manifesting phenotypes that lead to higher fitness.

Results of this study show how the phenotype of a mutation may vary with environment. Specifically, different mutations may exhibit different levels of environment-dependent variation in phenotype. How can this influence the evolutionary fate of mutations? Mutations of little or no effect—consequently with little or no bearing on fitness—drifting in an environment may quickly become advantageous when environmental conditions shift accordingly. For *TDH3* or a gene with similar degree of GxE interaction, there may exist layers of genotype- and environment-dependent phenotypic variation. An intriguing question then is how much extant standing genetic variation in a population exhibit GxE interaction? This is likely gene-dependent. GxE interaction may be more prevalent in genes with paralogs or environment-specific function and/or regulatory schemes. The evolutionary trajectory of mutations impacting such genes will likely be complex and environment-dependent.

The evolutionary impact of GxE interaction may help explain divergent genotypes underlying certain conserved phenotypes observed in nature. GxE interaction enables genotypes to harbor multiple levels of phenotypic variation. Within this framework, the phenotypic ranking of mutations changes across environments such that a particular phenotype can be achieved via different mutational paths in different environments. If selection is globally widespread and favors a particular magnitude of phenotypic change, populations in different environments may fix different mutations but result in similar phenotypes. This may have contributed to certain conserved phenotypes produced by divergent genotypes, an example of which is the *endo16* developmental gene in natural populations of the purple sea urchin *Strongylocentrotus purpuratus*.

Cis-regulatory element of *endo16* exhibits extensive sequence divergence despite conserved expression pattern [Romano & Wray 2003]. The *S. purpuratus* range encompasses diverse habitats across 20˚ of latitude, thus local adaptation may have led to the observed pattern of sequence divergence [Wray 2006]. Specifically, selection for a particular expression pattern during development in different local environments may have occurred on different mutational paths. Further investigation requires coupling expression with fitness across environments to test if a particular range of fitness may be achieved with different mutations in different environments.

In summary, this study shows that interaction between genetic variants and the environment can influence phenotype and provides another mechanism through which the mutation process contributes to phenotypic variation. Such interaction may have functional and evolutionary consequences and should be accounted for when assessing the phenotypic effects of genetic variants.

Materials and Methods

Strains

All strains were constructed and tested as haploids. The wild-type control, *S. cerevisiae* strain P_{TDH3}-YFP, MAT**a**, *lys2Δ0*, *ura3Δ0*, *CAN1*^{S}, was previously constructed based on BY4724 by G. Kalay and J. Gruber [Gruber *et al.* 2012]. P_{TDH3}-YFP is a reporter gene containing the wild-type form of P_{TDH3}, coding sequence of YFP (Venus variant), and the yeast *CYC1* (cytochrome c isoform 1) terminator placed into a pseudogene on chromosome I. The 678bp P_{TDH3} sequence consists of the 5' intergenic region up to but not including the start codon of the *TDH3* coding sequence;

this region contains TFBS, TATA box, 5’ untranslated region (UTR) of *TDH3*, and 3’ UTR of *PDX3*, the gene upstream of *TDH3* in the native locus on chromosome VII. *CYC1* terminator is the canonical yeast terminator and polyadenylation sequence to provide efficient transcription termination and transcript stability [Russo *et al*. 1989; Zaret *et al*. 1984].

An intermediate strain based on the wild-type control strain was created from which mutant strains were subsequently constructed: *URA3*-yfp, MAT**a**, *lys2Δ0*, *ura3Δ0*, *CAN1*^{S}. It was derived from the control strain by exchanging P_{TDH3} in the reporter transgene with a PCR construct containing the *URA3* promoter & coding sequence and a mutation that disrupts the start codon of the YFP coding sequence; both *URA3* and the YFP start codon mutation were utilized to aid the screening process during mutant strain construction. Transformation was carried out following the lithium acetate protocol using selection for *ura*+ phenotype and loss of YFP fluorescence [Gietz & Schiestl 2007] and confirmed using Sanger sequencing.

For the *cis*-regulatory mutant strains, site-directed mutagenesis was performed using two overlapping primers containing the desired mutation for each *cis*-regulatory mutation followed by PCR sewing. The resultant PCR construct also contains sequence that restores the YFP start codon. Mutant strains were then constructed by transforming the PCR constructs into the *URA3*-yfp intermediate strain following the lithium acetate method using selection for *ura*- (with 5-FOA) and gain of YFP fluorescence [Gietz & Schiestl 2007]. Putative transformants were confirmed using

Sanger sequencing. Mutant strains are of the genotype $P_{TDH3}$$^{\mathrm{mut}}$-YFP, MAT**a**, *lys2Δ0*, *ura3Δ0*, *CAN1*$^{\mathrm{S}}$. Each *cis*-regulatory mutant strain contains either a G → A or C → T transition in P_{TDH3}. This type of substitution represents one of the most common types of spontaneous single-nucleotide substitutions in the yeast genome [Lynch et al. 2008]. Construction of the G_{-293} → A_{-293} strains was similar to that of the *cis*-regulatory mutant strains but each containing an additional G_{-293} → A_{-293} transition in P_{TDH3} to test the effect of background G_{-293} in the wild-type reference strain. 29 such strains were constructed and represent a subset of the *cis*-regulatory mutants comprising a range of expression phenotypes.

Flow cytometry

Strains were arrayed into 96-well format in randomized order for glycerol stock. Prior to quantifying mutation effects in flow cytometry, strains were revived from glycerol stock onto YPG (glycerol) solid media to reduce formation of petites. This was carried out for all strains simultaneously to control for potential age-related effects and new additional mutations. Strains were subsequently transferred into deep 96-well plate liquid culture, with each strain grown in 500µl of liquid YPD (glucose), YPGal (galactose), and YPD+sorbitol (1M) media for 20 hours and YPG (glycerol) and YPEtOH (3%) for 42 hours to stationary phase while shaking at 250rpm with 3mm glass bead in a 30˚C incubator. Immediately prior to quantification of fluorescence, 20-25µl of YP culture was transferred into 500µl of SC-R (with respective carbon source or sorbitol) and passed into the flow cytometer. 9 biological replicates in YPD and 6 biological replicates in the other environments were independently inoculated and grown for flow cytometry.

Promoter activity was quantified as reporter fluorescence in flow cytometry using a BD Accuri C6 flow cytometer coupled with an IntelliCyt HyperCyt Autosampler. Flow rate of 14µl/min and core size of 10µm were used in the flow cytometer, with a blue laser (λ = 480 nm) for excitation of YFP and fluorescence data collected from the FL1 channel using a 533/30nm filter. Each well in a 96-well plate was sampled for 2-3s, with 20,000 events recorded on average. Data was then processed using flowClust (3.0.0) and flowCore (1.26.3) packages in R (3.0.2) to remove artifacts such as debris and other non-cell events [Lo *et al*. 2009; Hahne *et al*. 2009]. Samples with fewer than 1000 events following processing were excluded from further analysis. Using the remaining data, YFP fluorescence was calculated as $(\log_{10}\text{FL1.A})^2/(\log_{10}\text{FSC.A})^3$ for each event for each sample, and expression noise was calculated as sd(YFP fluorescence)^2/mean(YFP fluorescence)^2. Mean expression level and expression noise for each strain is then estimated as mean across 9 biological replicates. Random variation during growth that may affect gene expression was controlled for using 20 replicates of the wild-type reference control strain in the same position on each plate. YFP fluorescence or expression noise from these control strains was used to fit linear models to estimate effects of several parameters (YFP or expression noise ~ day + replicate + plate + row + column + block + stack + depth + order[1]) using the

[1] The parameters were: day in which a plate was run, replicate that plate belongs to, plate, row/column/block(half plate) in a plate a sample was in, stack and depth in stack a plate was cultured in, and order in which a plate was run within a replicate.

function lm in R. The function step in R was used to choose the model with the best explanatory power and estimate the effect(s) with statistical significance, which was consistently plate, for adjustment. Autofluorescence was accounted for by subtracting estimate of YFP fluorescence of a non-fluorescent strain from all samples. The resultant fluorescence measurement was used to calculate mutation effect. The wild-type reference control strain in all five environments tested were run together in an additional plate to estimate the baseline difference in expression across environments and used for adjustments accordingly.

Modeling

GxE and GxGxE interactions were tested with linear models using the lm function in R (3.0.2). For GxE interaction, the model used was phenotype ~ strain + environment + strain*environment, with the phenotype term being either mean expression level or expression noise. For GxGxE interaction, the model used was phenotype ~ strain + background + environment + strain*environment + strain*background*environment, with the background term indicating the presence or absence of the A_{-293} → G_{-293} background genetic variant. *p*-values from the regression results are then adjusted to false discovery rate (FDR) *q*-values using the function p.adjust in R to account for multiple testing.

References

Abouheif E and Wray GA. Evolution of the Gene Network Underlying Wing Polyphenism in Ants. *Science* 297:249-252 (2002).

Blomberg A and Adler L. Physiology of osmotolerance in fungi. *Advances in Microbial Physiology*. 33;145-212 (1992).

Boucherie H, Bataille N, Fitch IT, Perrot M, and Tuite MF. Differential synthesis of glyceraldehyde-3-phosphate dehydrogenase polypeptides in stressed yeast cells. *FEMS Microbiology Letters* 125:127-133 (1995).

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, *et al*. Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a *Pitx1* Enhancer. *Science* 327:302-305 (2010).

Delgado ML, O'Connor JE, Azorin I, Renau-Piqueras J, Gil ML, and Gozalbo D. 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 (2001).

Eng KH, Kvitek DJ, Keles S, and Gasch AP. Transient Genotype-by-Environment Interactions Following Environmental Shock Provide a Source of Expression Variation for Essential Genes. *Genetics* 184:587-593 (2010).

Flynn KM, Cooper TF, Moore FBG, and Cooper VS. The Environment Affects Epistatic Interactions to Alter the Topology of an Empirical Fitness Landscape. *PLoS Genetics* 9:e1003426 (2013).

Gibson G and Dworkin I. Uncovering cryptic genetic variation. *Nature Reviews Genetics* 5:681-690 (2004).

Gietz RD and Schiestl RH. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature Protocols* 2:31-34 (2007).

Gruber JD, Vogel K, Kalay G, and Wittkopp PJ. Contrasting Properties of Gene-Specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects, and Dominance. *PLoS Genetics* 8:e1002497 (2012).

Hahne F, LeMeur N,Brinkman RR, Ellis B, Haaland P, Sarkar D, Spidlen J, Strain E, and Gentleman R. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* 10:106-113 (2009).

Hodgins-Davis A, Adomas AB, Warringer J, and Townsend JP. Abundant Gene-by-Environment Interactions in Gene Expression Expression Norms to Copper within *Saccharomyces cerevisiae*. *Genome Biology and Evolution* 4:1061-1079 (2012).

Hoffjan S, Nicolae D, Ostrovnaya I, Roberg K, Evans M, Mirel DB, Steiner L, Walker K, Shult P, Gangnon RE, Gern JE, Martinez FD, Lemanske RF, and Ober C. Gene-environment interaction effects on the development of immune responses in the 1st year of life. *American Journal of Human Genetics* 76:696–704 (2005).

Holland MJ, Yokoi T, Holland JP, Myambo K, and Innis MA. The *GCR1* gene encodes a positive transcriptional regulator of the enolase and glyceraldehyde-3-phosphate dehydrogenase gene families in *Saccharomyces cerevisiae*. *Molecular and Cell Biology* 7:813-820 (1987).

Hyduke DR and Palsson BØ. Towards genome-scale signalling-network reconstructions. *Nature Reviews Genetics* 11:297-307 (2010).

Kuroda S. Otaka S, and Fujisawa Y. Fermentable and nonfermentable carbon sources sustain constitutive levels of expression of yeast triosephosphate dehydrogenase 3 gene from distinct promoter elements. *Journal of Biological Chemistry* 269:6153-6162 (1994).

Kvitek DJ, Will J, and Gasch A. Variations in Stress Sensitivity and Genomic Expression in Diverse *S. cerevisiae* Isolates. *PLoS Genetics* 4:e1000223 (2008).

Landry CR, Oh J, Hartle DL, and Cavalieri D. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* 366:343-351 (2006).

Li Y, Álvarez O, Gutteling E, Tijsterman M, Fu J, Riksen J, Hazendonk E, Prins P, Plasterk R, and Jansen R. Mapping Determinants of Gene Expression Plasticity by Genetical Genomics in *C. elegans*. *PLoS Genetics* 2:e222 (2006).

Lo K, Hahne F, Brinkman RR, and Gottardo R. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics 2009* 10:145-152 (2009).

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, and Thomas WK. A genome-wide view of the spectrum of spontaneous mutations in yeast. *PNAS* 105:9272-9277 (2008).

McAlister L and Holland MJ. Isolation and characterization of yeast strains carrying mutations in the glyceraldehyde-3-phosphate dehydrogenase genes. *Journal of Biological Chemistry* 260:15013-15018 (1985).

McAlister L and Holland MJ. Differential expression of the three yeast glyceraldehyde-3-phosphate dehydrogenase genes. *Journal of Biological Chemistry* 260:15019-15027 (1985).

Maranville JC, Francesca L, Stephens M, and Di Rienzo A. Mapping gene-environment interactions at regulatory polymorphisms. *Transcription* 3:56-62 (2012).

Metzger BPH, Yuan DC, Gruber JD, Duveau F, and Wittkopp PJ. Contrasting *cis*-regulatory effects of mutations and polymorphisms to test for selection. *In preparation*.

Nickels S, Truong T, Hein R, Stevens K, Buck K, Behrens S, Eilber U, Schmidt M, Häberle L, Vrieling A et al. Evidence of gene–environment interactions between common breast cancer susceptibility Loci and Established Environmental Risk Factors. *PLoS Genetics* 9:e1003284 (2013).

O'Rourke SM, Herskowitz I, and O'Shea EK. Yeast go the whole HOG for the hyperosmotic response. *Trends in Genetics* 18:405-412 (2002).

Pavlović B and Hörz W. The chromatin structure at the promoter of a glyceraldehyde phosphate dehydrogenase gene from *Saccharomyces cerevisiae* reflects its functional state. *Molecular and Cellular Biology* 8:5513-5520 (1988).

Raser JM and O'Shea EK. Control of Stochasticity in Eukaryotic Gene Expression. *Science* 304:1811-1814 (2004).

Remold SK and Lenski RE. Contribution of individual random mutations to genotype-by-environment interactions in *Escherichia coli*. *PNAS* 98:11388–11393 (2001).

Romano LA and Wray GA. Conservation of *Endo16* expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Development* 130:4187-4199 (2003).

Russo P and Sherman F. Transcription terminates near the poly(A) site in the *CYC1* gene of the yeast *Saccharomyces cerevisiae*. *PNAS* 86:8348-8352 (1989).

Sambandan D, Carbone M, Anholt R, and Mackay T. Phenotypic plasticity and genotype by environment interaction for olfactory behavior in *Drosophila melanogaster*. *Genetics* 179:1079-1088 (2008).

Smith EN and Kruglyak L. Gene-Environment Interaction in Yeast Gene Expression. *PLoS Biology* 6:e83 (2008).

Tirosh I and Barkai N. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biology* 8:R50 (2007).

Wittkopp P, Vaccaro K, and Carroll S. Evolution of yellow Gene Regulation and Pigmentation in Drosophila. *Current Biology* 12:1547-1556 (2002).

Wade MJ and Goodnight CJ. Perspective: The Theories of Fisher and Wright in the Context of Metapopulations: When Nature Does Many Small Experiments. *Evolution* 52:1537-1553 (1998).

Wang Z and Zhang J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *PNAS* 108:E67-76 (2011).

Wray GA. The evolution of embryonic gene expression in sea urchins. *Integrative and Comparative Biology* 46:233-242 (2006).

Yagi S, Yagi K, Fukuoka J, and Suzuki M. The UAS of the yeast GAPDH promoter consists of multiple general functional elements including RAP1 and GRF2 binding sites. *Journal of Veterinary Medical Science* 56:235-244 (1994).

Zaret KS and Sherman F. Mutationally altered 3' ends of yeast *CYC1* mRNA affect transcript stability and translational efficiency. *Journal of Molecular Biology* 176:107-135 (1984).

Figure 4.1

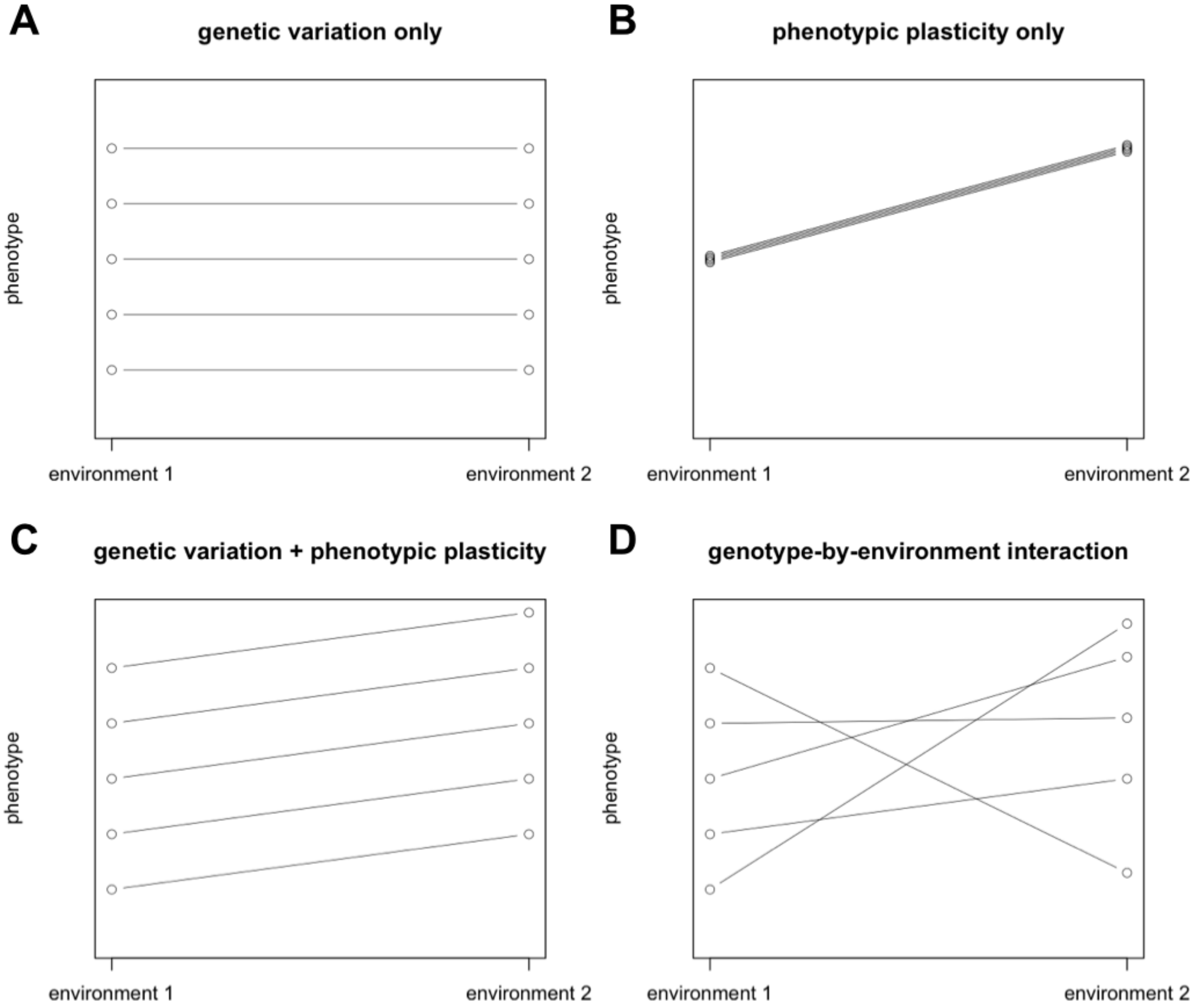

Figure 4.1 Reaction norm diagrams illustrating phenotypes (vertical axis) produced by different genotypes (open circles) in different environments (horizontal axis).
A. Genetic variation across genotypes produce different phenotypes that are invariable between environments 1 and 2. **B**. Little or no genetic variation exists among genotypes but all produce similar phenotypic variation between environments 1 and 2 and hence exhibit phenotypic plasticity. **C**. Both genetic variation and phenotypic plasticity exist but not genetic variation for phenotypic plasticity; there is no interaction. All genotypes produce similar phenotypic variation in response to environmental change. **D**. Genetic variation for phenotypic plasticity, or GxE interaction, producing the characteristic non-parallel reaction norms. Modified from [Landry *et al*. 2006].

Figure 4.2

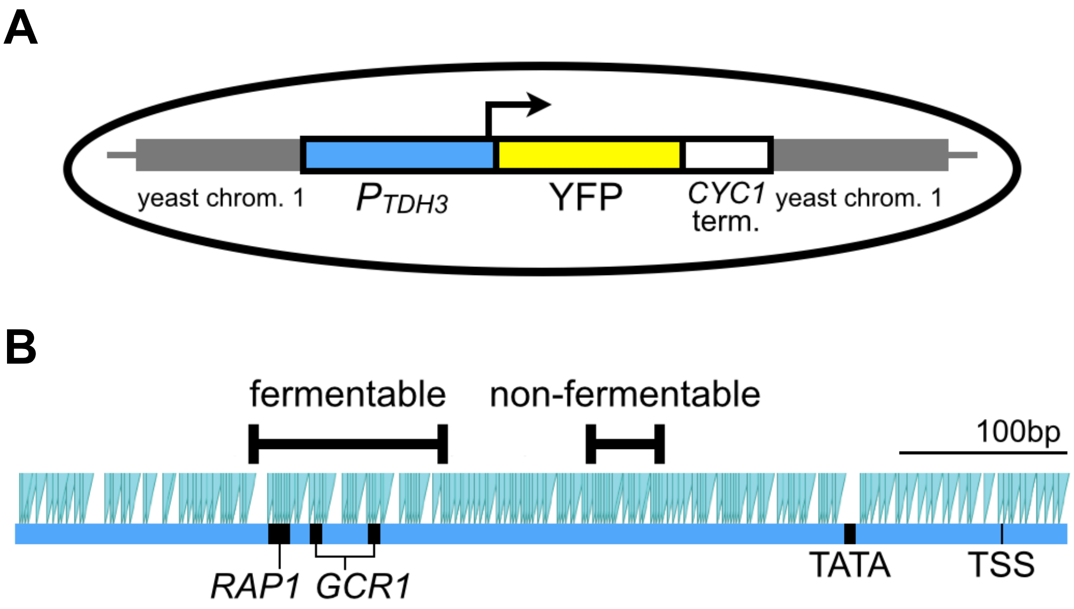

Figure 4.2 Overview of the reporter gene. **A**. The reporter gene containing the *TDH3* promoter (P_{TDH3}) driving expression of the reporter yellow fluorescent protein (YFP) flanked by the *CYC1* terminator integrated into chromosome 1 of the yeast genome. **B**. 236 single-nucleotide substitutions (◥) across 678bp of P_{TDH3}. Horizontal brackets indicate regions of the promoter previously characterized to be active during metabolism of fermentable and non-fermentable carbon sources [Kuroda *et al*. 1994]. TATA box, transcription start site (TSS), and known binding sites for transcription factors *RAP1* and *GCR1* are indicated [Holland *et al*. 1987; Pavlović & Hörz 1988; Kuroda *et al*. 1994; Yagi *et al*. 1994].

Figure 4.3

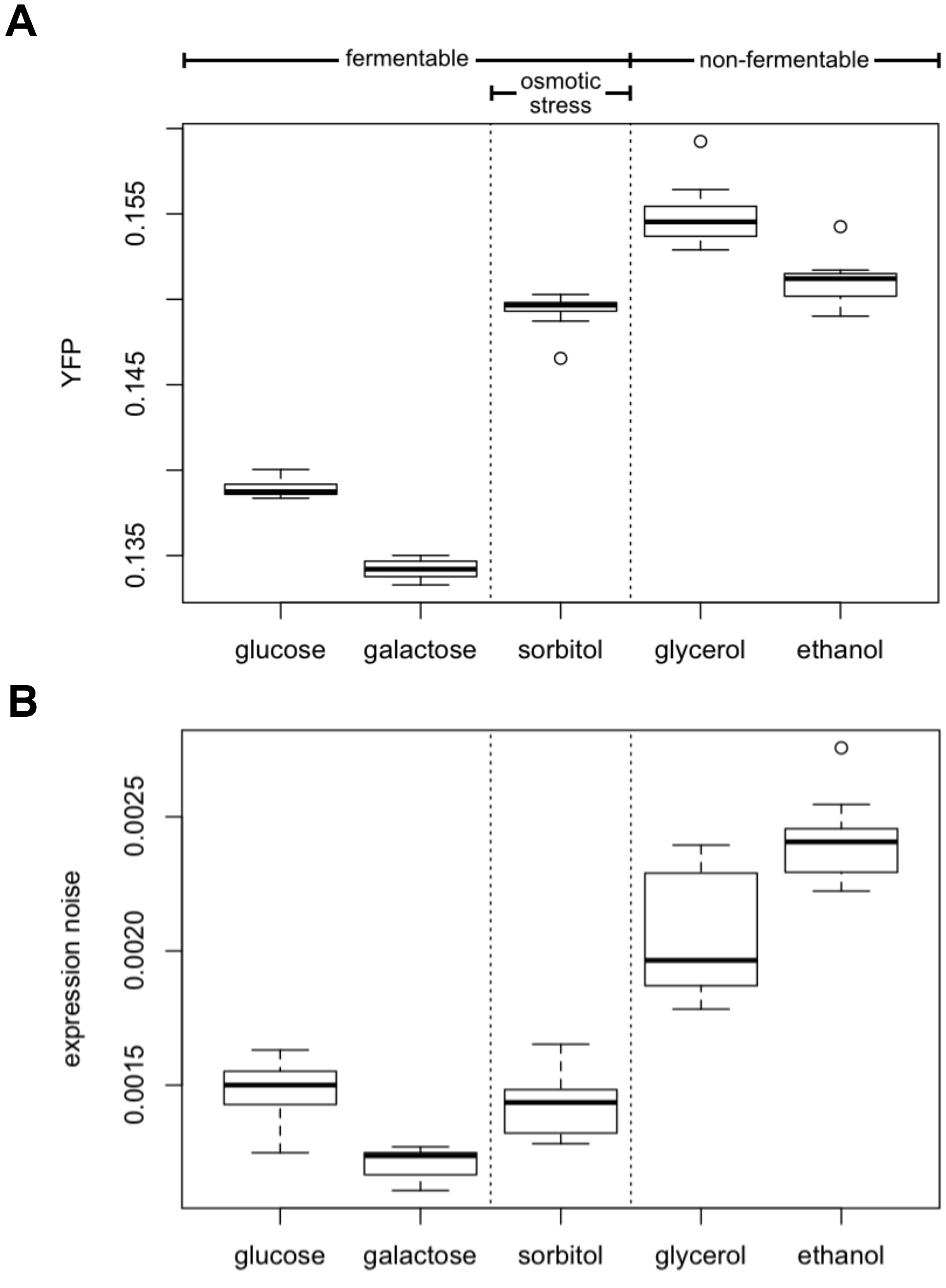

Figure 4.3 Boxplots illustrating phenotypes of the P_{TDH3}-YFP wild-type reference strain across the environments tested. **A**. Mean expression level (YFP). **B**. Expression noise (variance in YFP normalized by YFP).

Figure 4.4

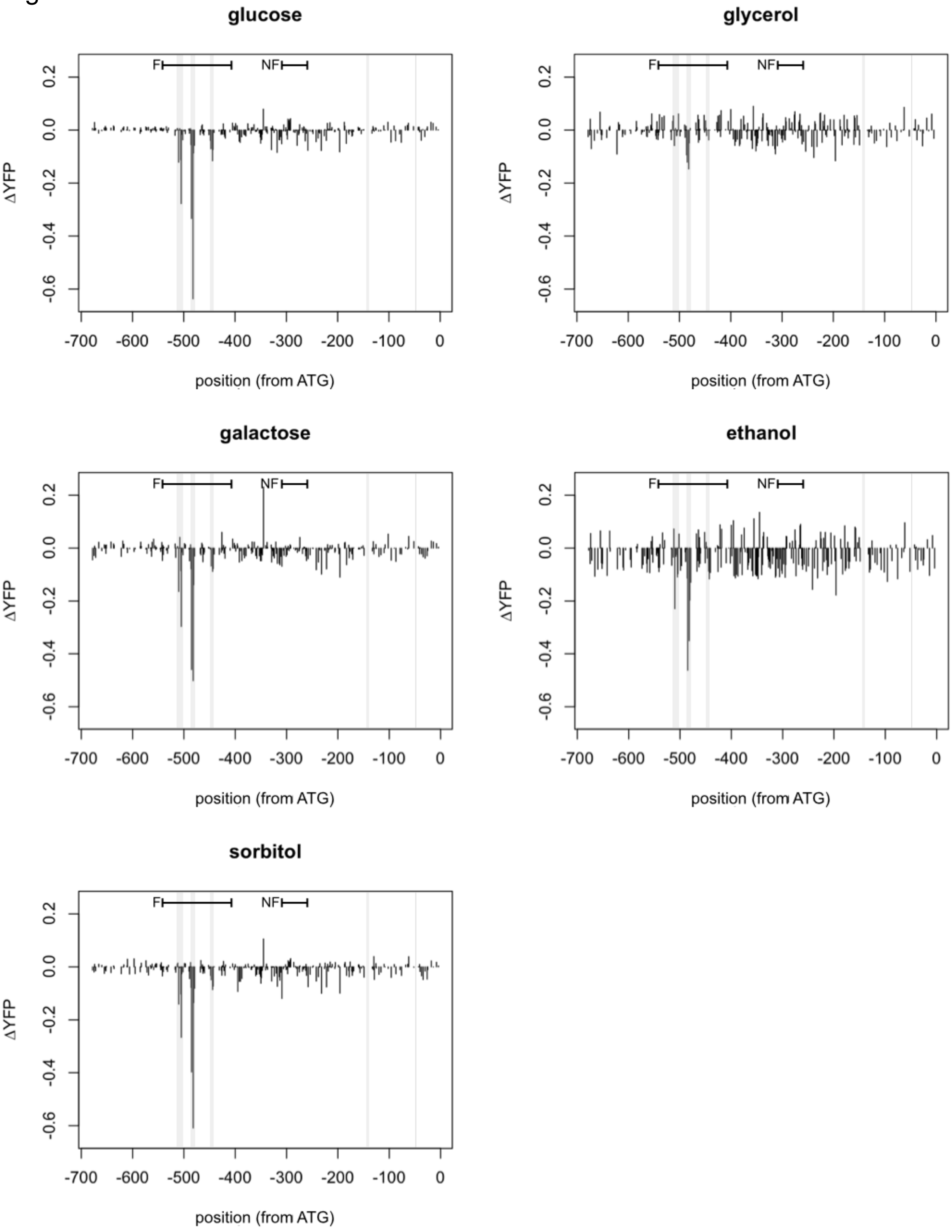
glucose
glycerol
galactose
ethanol
sorbitol
F
NF
ΔYFP
0.2
0.0
-0.2
-0.4
-0.6
-700
-600
-500
-400
-300
-200
-100
0
position (from ATG)

Figure 4.4 Effect on mean expression level of the 236 *cis*-regulatory mutations across environments. Each panel plots change in YFP fluorescence relative to the wild-type reference strain (vertical axis) against position (bp) of the mutation in P_{TDH3} from the coding sequence (horizontal axis). Each vertical line indicates the effect of the mutation at that position. Known functional elements shown in Figure 4.2B are indicated as gray bars and are, from left to right, known transcription factor binding sites for *RAP1*, *GCR1*, and *GCR1*, followed by TATA box and TSS. Horizontal brackets indicate regions previously characterized to be important during metabolism of fermentable (“F”) or non-fermentable (“NF”) carbon sources.

Figure 4.5A

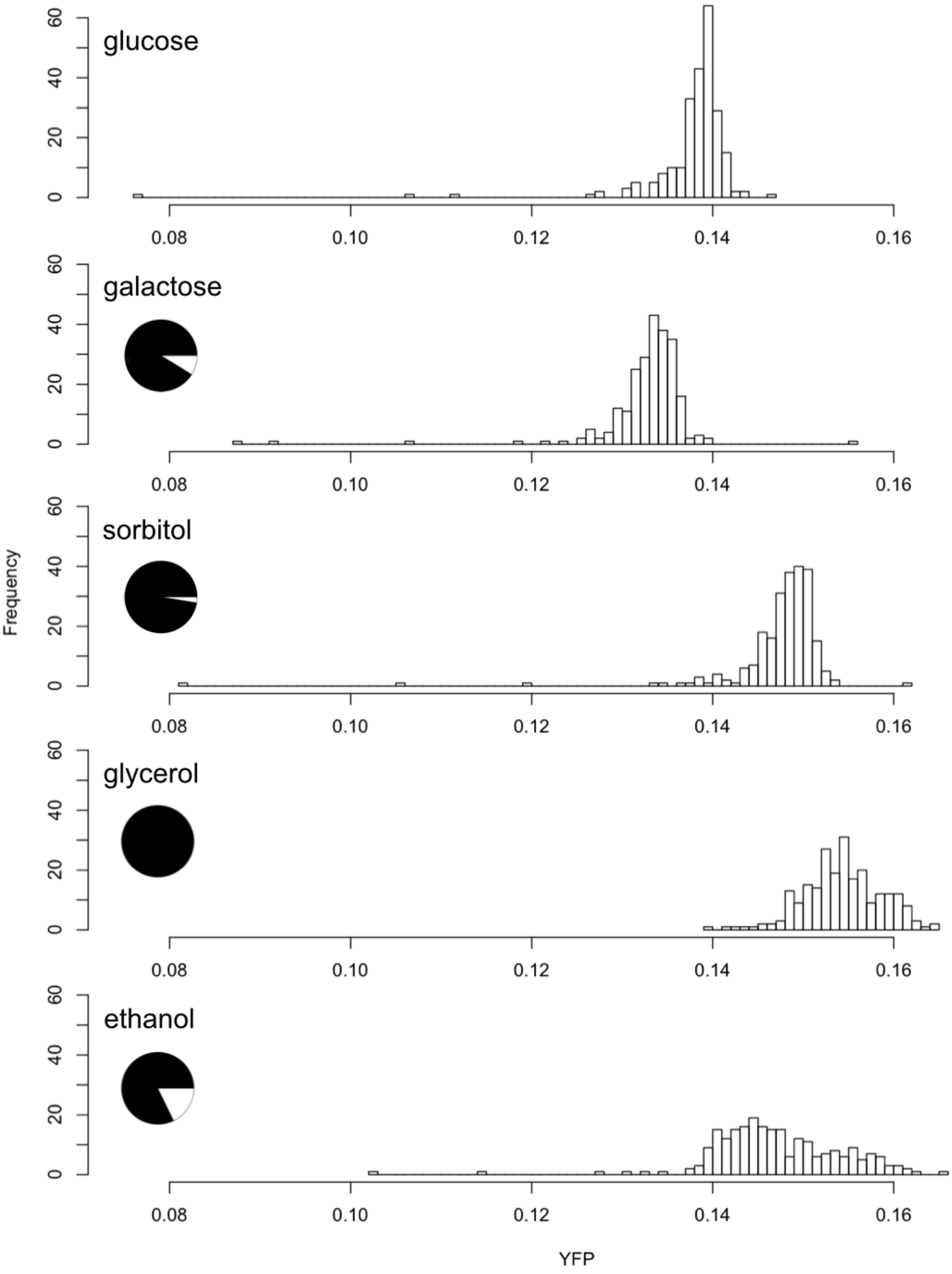
glucose
galactose
sorbitol
glycerol
ethanol
Frequency
0
20
40
60
0.08
0.10
0.12
0.14
0.16
YFP

Figure 4.5B

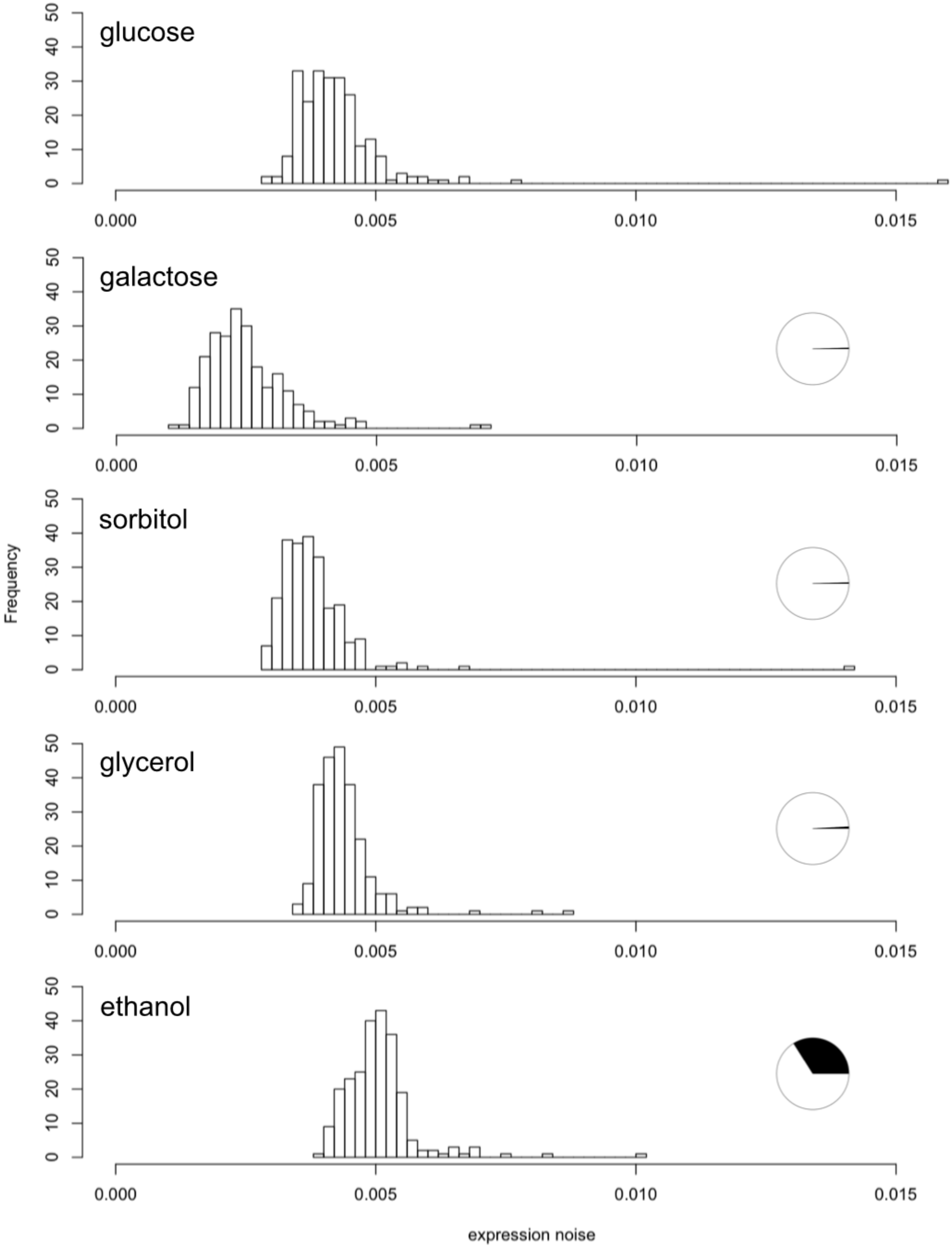
glucose
galactose
sorbitol
glycerol
ethanol
Frequency
0
10
20
30
40
50
0.000
0.005
0.010
0.015
expression noise

Figure 4.5 Distributions of phenotypes of the 236 *cis*-regulatory mutations across environments. In each panel, frequency (vertical axis) is plotted against phenotype value (horizontal axis) as histograms. Inset pie charts show relative frequency of *cis*-regulatory mutations (black) that exhibited phenotype significantly different in an environment compared to that in glucose (FDR $q < 0.05$). **A**. Mean expression level (YFP). All distributions are significantly different from that in glucose (t-tests $p < 0.05$). Numbers of individual *cis*-regulatory mutations that exhibited mean expression level significantly different from that in glucose are: 215 (galactose), 230 (sorbitol), 236 (glycerol), and 194 (ethanol). **B**. Expression noise. All distributions are significantly different from that in glucose (t-tests $p < 0.05$) except glycerol (t-test $p > 0.05$). Numbers of individual *cis*-regulatory mutations that exhibited expression noise significantly different from that in glucose are: 1 (galactose), 1 (sorbitol), 2 (glycerol), and 80 (ethanol).

Figure 4.6A

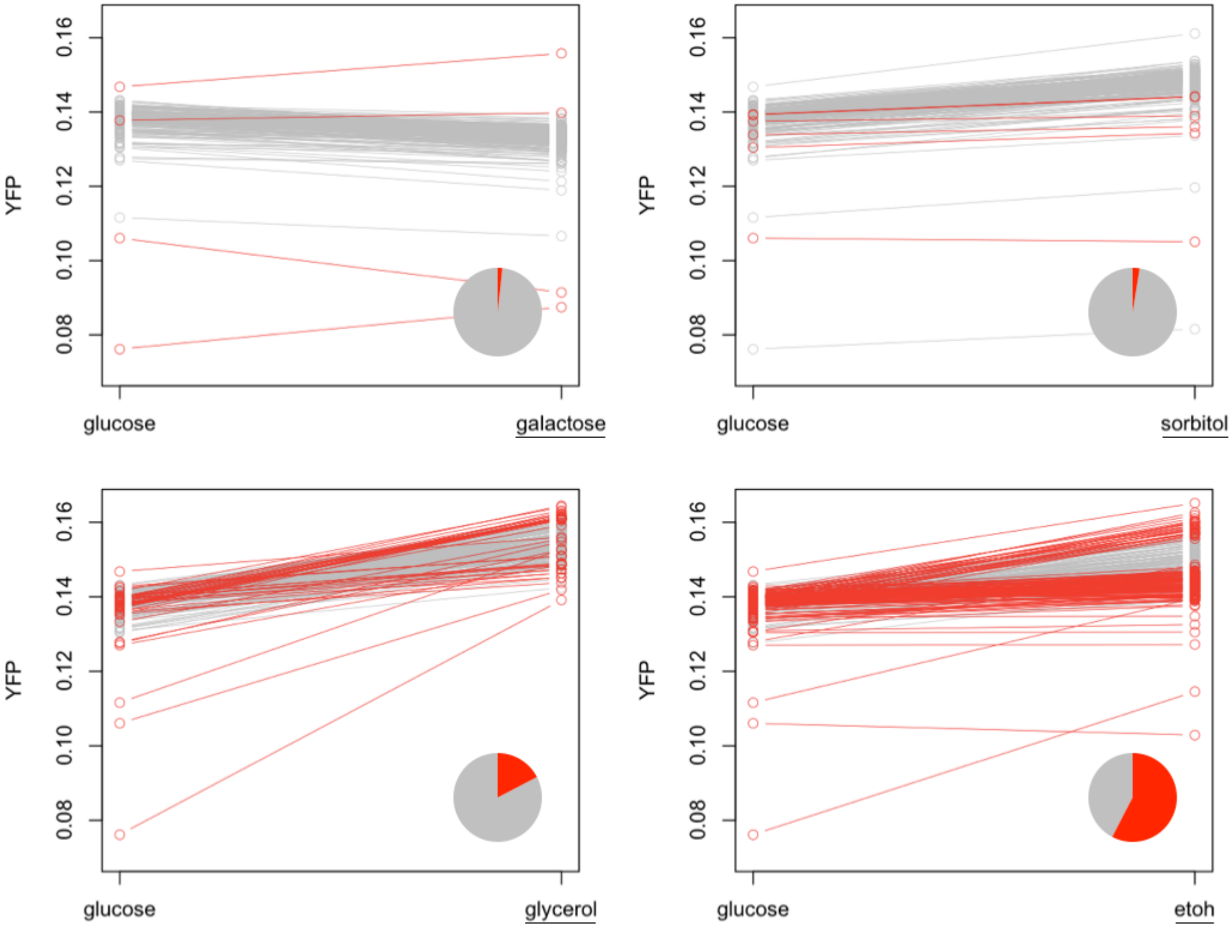
YFP
0.08
0.10
0.12
0.14
0.16
glucose
galactose
YFP
0.08
0.10
0.12
0.14
0.16
glucose
sorbitol
YFP
0.08
0.10
0.12
0.14
0.16
glucose
glycerol
YFP
0.08
0.10
0.12
0.14
0.16
glucose
etoh

Figure 4.6B

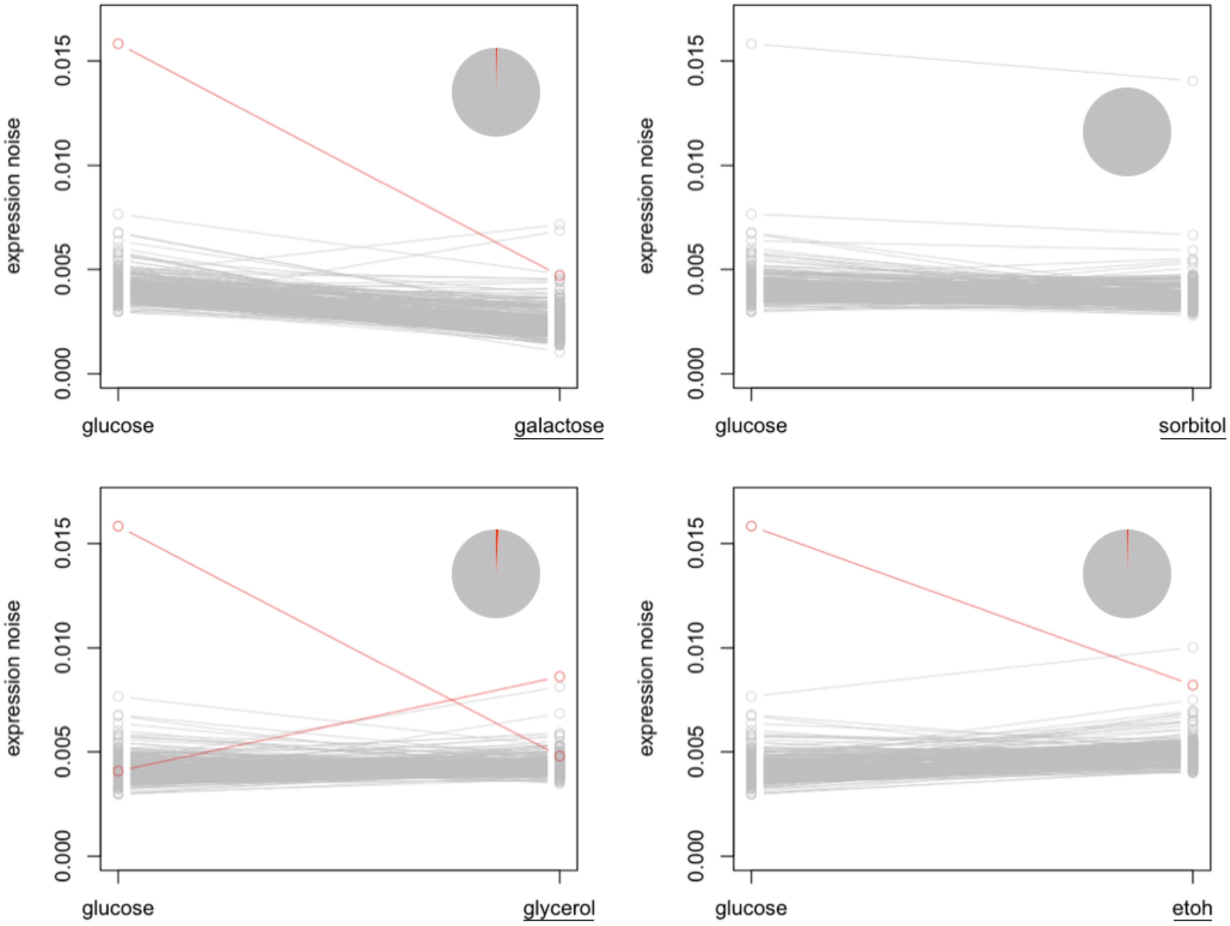
expression noise
0.000
0.005
0.010
0.015
glucose
galactose
expression noise
0.000
0.005
0.010
0.015
glucose
sorbitol
expression noise
0.000
0.005
0.010
0.015
glucose
glycerol
expression noise
0.000
0.005
0.010
0.015
glucose
etoh

Figure 4.6C

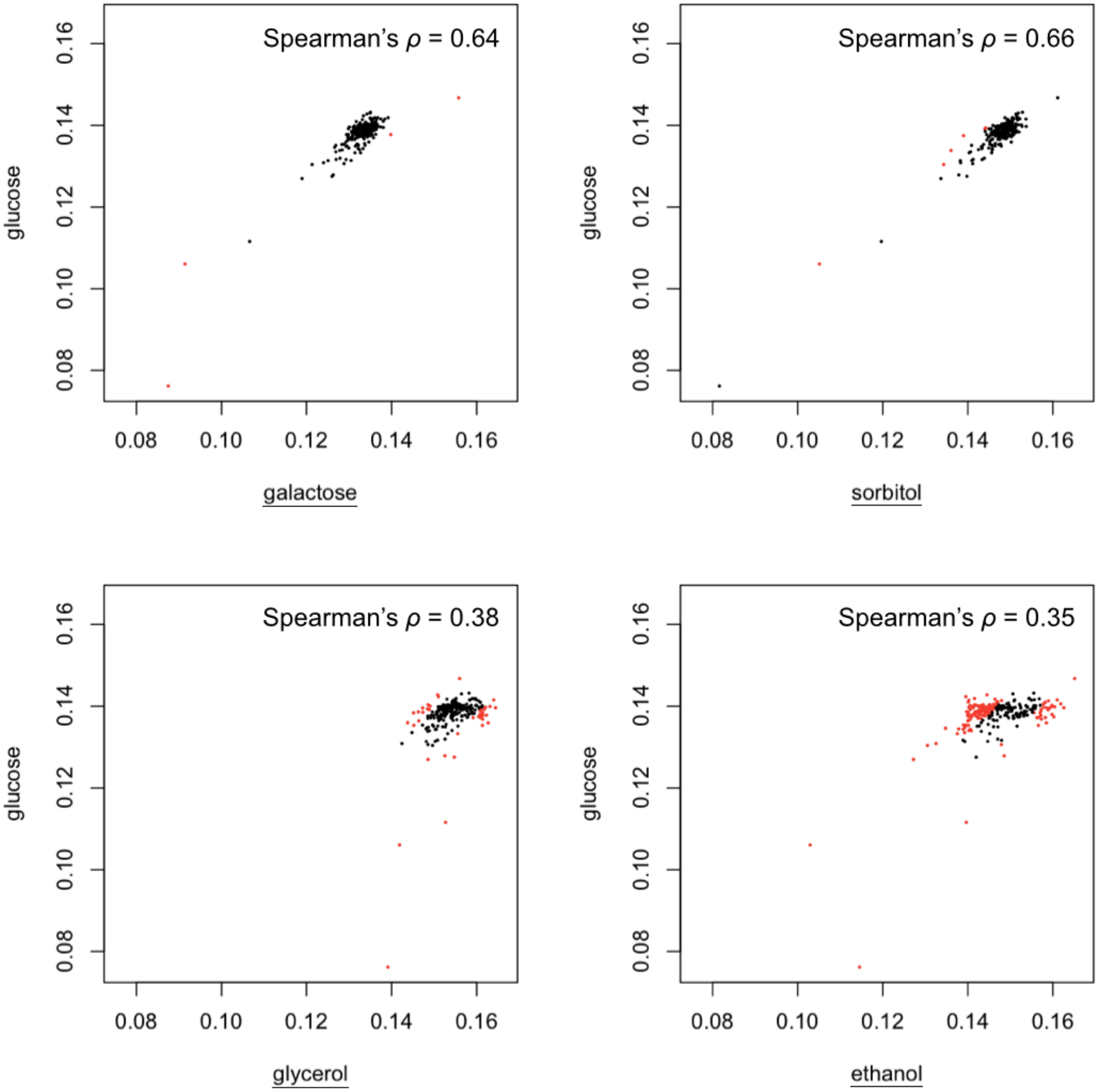
Spearman's ρ = 0.64
glucose
galactose
Spearman's ρ = 0.66
glucose
sorbitol
Spearman's ρ = 0.38
glucose
glycerol
Spearman's ρ = 0.35
glucose
ethanol
0.08
0.10
0.12
0.14
0.16

Figure 4.6D

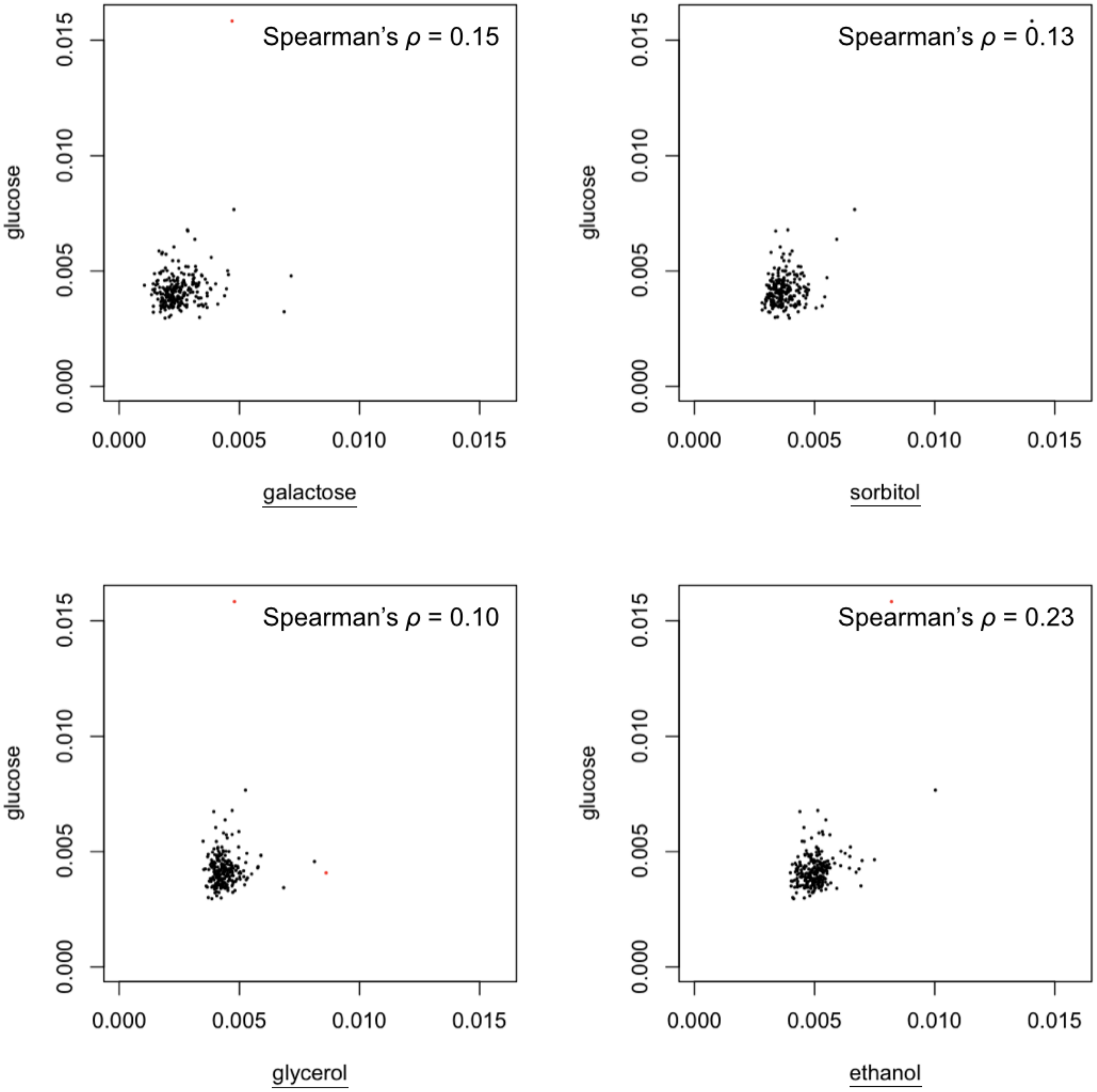
Spearman's ρ = 0.15
glucose
0.000
0.005
0.010
0.015
galactose
Spearman's ρ = 0.13
glucose
sorbitol
Spearman's ρ = 0.10
glucose
glycerol
Spearman's ρ = 0.23
glucose
ethanol

Figure 4.6 Phenotypes of the 236 *cis*-regulatory mutations across environments tested. Reaction norm diagrams are shown in **A**. for mean expression level (YFP) and **B**. for expression noise (vertical axis) across environments (horizontal axis). Red lines and red sectors (inset pie charts) indicate those that exhibited significant GxE interactions (FDR $q < 0.05$), gray lines and gray sectors those that did not (FDR $q > 0.05$). Scatter plots are shown in **C**. of mean expression level (YFP) and **D**. of expression noise across environments (horizontal axis) against that in glucose (vertical axis). Mutations that exhibited significant GxE interactions are in red. In each environment, rank correlation (Spearman's ρ) of the phenotype to that in glucose are shown.

Figure 4.7

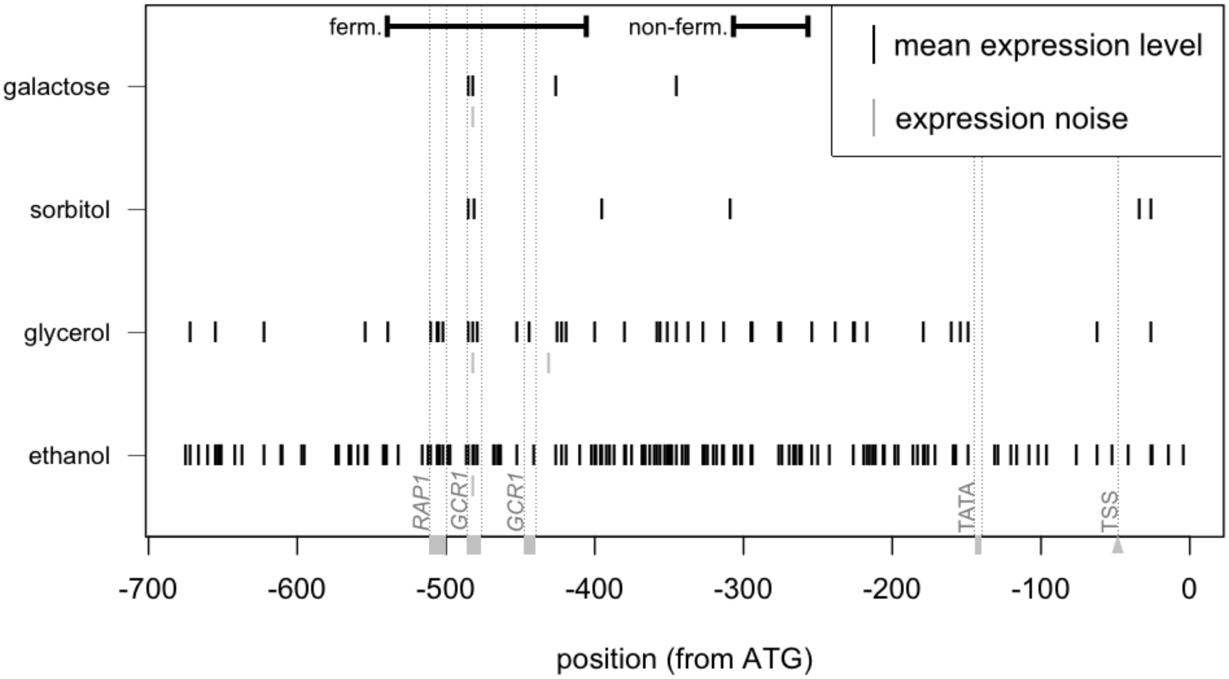

Figure 4.7 Sites along P_{TDH3} (horizontal axis) that exhibited significant GxE interaction (FDR $q < 0.05$) across environments tested (vertical axis). Black vertical bars indicate *cis*-regulatory mutations exhibiting GxE interaction for mean expression level, gray vertical bars for expression noise. Known functional elements are indicated with dotted lines; regions important during metabolism of fermentable ("ferm.") and non-fermentable ("non-ferm.") carbon sources are indicated with horizontal brackets.

Figure 4.8

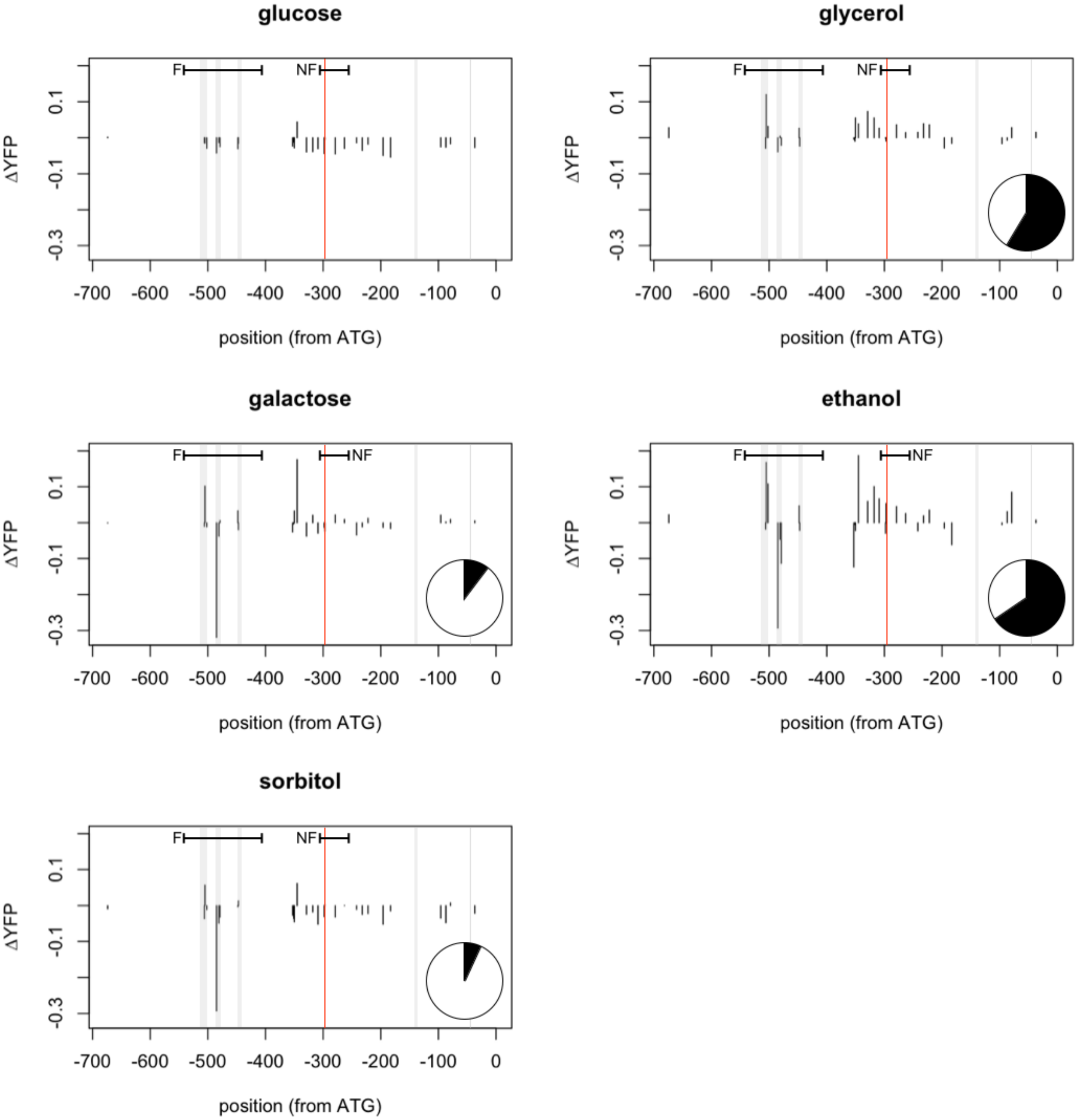
glucose
glycerol
galactose
ethanol
sorbitol
F
NF
ΔYFP
0.1
-0.1
-0.3
-700
-600
-500
-400
-300
-200
-100
0
position (from ATG)

Figure 4.8 Effect of background genetic variant (G_{-297}) on mean expression level across environments. In each panel, the change in YFP fluorescence (vertical axis) of the *cis*-regulatory mutant with G_{-297} relative to that of the same mutant with the variant found in other common genetic backgrounds (A_{-297}) is plotted against position (bp) of the *cis*-regulatory mutation in P_{TDH3} from the coding sequence (horizontal axis). Each vertical line indicates the effect of G_{-297} on P_{TDH3} activity in a *cis*-regulatory mutant with a mutation at that position. Red line indicates position -297. Known functional elements are indicated as gray bars and are, from left to right, known transcription factor binding sites for *RAP1*, *GCR1*, and *GCR1*, followed by TATA box and TSS. Horizontal brackets indicate regions previously characterized to be important during metabolism of fermentable ("F") or non-fermentable ("NF") carbon sources. Inset pie charts show relative frequency of mutations (black) that exhibited significant genotype-by-genotype-by-environment (GxGxE) interaction (FDR $q < 0.05$); specifically, out of 29 mutations tested, GxGxE interaction was detected in 3 (galactose), 2 (sorbitol), 17 (glycerol), and 19 (ethanol).

Chapter 5

Conclusion

Variation is a hallmark of biological organisms. The “endless forms most beautiful and most wonderful” described by Charles Darwin [1859] have motivated many to understand the origins of and mechanisms underlying variation of such forms and phenotypes. The evolution of many phenotypes has been shown to involve modification of gene expression [e.g. Wittkopp *et al*. 2002; Chan *et al*. 2010]. Gene expression is an important process through which information encoded in the DNA is transformed into gene products that carry out specific biological functions, ultimately leading to higher-level phenotypes. Genetic changes, or mutations, that modify gene expression are therefore important effectors of phenotypic variation. In the studies described in this dissertation, I set out to understand how mutation may contribute to variation in gene expression. To characterize an array of mutations and natural genetic variants, I used a transgene containing a *cis*-regulatory element driving expression of a reporter in the model system *Saccharomyces cerevisiae*. In this chapter, I discuss implications of the results as well as prospectives for future work.

Genetic changes that modify gene expression can be broadly divided into two categories: *cis*- and *trans*-regulatory. *Cis*-regulatory changes include those in the *cis*-regulatory elements (e.g. promoters or enhancers) of the target gene, while *trans*-regulatory changes may be found in the *cis*-regulatory elements or the coding sequence of *trans*-acting factors to the target gene. Both *cis*- and *trans*-regulatory changes contribute to evolution, but *cis*-regulatory expression change has been implicated to play a larger role in interspecies divergence than *trans*-regulatory expression change [Wittkopp *et al*. 2008; McManus *et al*. 2010; Tirosh *et al*. 2009; Emerson *et al*. 2010]. This unequal distribution of *cis*- and *trans*-regulatory changes may result from negative selection against *trans*-regulatory changes or positive selection for *cis*-regulatory changes [Wittkopp *et al*. 2008].

Trans-regulatory changes may be selected against due to their potentially deleterious effect from pleiotropy [Wray 2007; Stern & Orgogozo 2008]. While a *cis*-regulatory change is likely to impact only its target gene, a *trans*-regulatory change—such as one altering expression of a transcription factor—is likely to impact more than one target gene. Such change in one gene may be advantageous to that gene or the trait it is responsible for but harmful to others. Consistent with this hypothesis, *cis*-acting expression quantitative trait loci (eQTL) appear to exhibit larger effects on average than *trans*-acting eQTL among expression variation in extant populations [Schadt *et al*. 2003; Dixon *et al*. 2007; Hubner *et al*. 2007; West *et al*. 2007]. Little is known, however, about the distributions of effect size among newly-arisen *cis*- and *trans*-regulatory mutations. Before selection has acted, do *cis*- and *trans*-regulatory mutations have similar effects?

To answer this question, I compared the effects of *cis*- and *trans*-regulatory mutants in Chapter 2. I generated a collection *cis*-regulatory mutations in the promoter of the *S. cerevisiae* gene *TDH3* and quantified their effects using fluorescence of a reporter controlled by this promoter. The effects of these *cis*-regulatory mutants were compared to the effects of a collection of *trans*-regulatory mutants previously generated by [Gruber *et al*. 2012] using the same reporter system and promoter. The results suggest that *cis*-regulatory mutations have intrinsically larger effect on gene expression than *trans*-regulatory mutations. Thus, the difference in effect size between *cis*- and *trans*-regulatory changes observed in extant populations may reflect, at least in part, the mutation process. Negative selection against *trans*-regulatory mutations may not play as big a role in the disproportionate contribution of *cis*-regulatory changes in interspecies divergence as positive selection for *cis*-regulatory mutations.

Empirically testing this hypothesis may reveal further detail of the mechanism underlying phenotypic evolution. While selection for *cis*-regulatory changes may be identified by comparing extant *cis*-regulatory expression variation against a neutral or null expectation [Zhen & Andolfatto 2012], detecting negative selection against *trans*-regulatory mutations is difficult because the signal left by deleterious genetic variants may be weak [Ezawa *et al*. 2013; Zhai *et al*. 2009]. An alternative may be to investigate pleiotropy. While pleiotropy in gene expression has been examined using eQTL mapping [e.g. Brem *et al*. 2002], little is known empirically about the pleiotropic effects of *cis*- and *trans*-regulatory mutations that impact the same focal gene. Determining the difference in pleiotropic effects between *cis*- and *trans*-regulatory mutations can lend support to an expectation that underlies many hypotheses involving *cis*- and *trans*-regulatory changes, including how they contribute to evolution.

The *cis*-regulatory mutations from Chapter 2 represent a spectrum of newly-arisen mutations. In nature, only a subset of such mutations survive to remain in a population. Specifically, as the mutation process introduces new genetic variants into a population, the frequencies of those genetic variants may change due to genetic drift and/or selection. While genetic drift is based on random chance events, selection acts on the phenotypic effects of the genetic variants. By comparing the phenotypes of variants observed in nature to those produced solely by the mutation process, the effects of mutation and selection can be disentangled. In other other words, such comparison can reveal the contribution of mutation to gene expression variation in natural populations.

To determine the relative contributions of mutation and selection to expression variation in nature, I compared the effects on expression of natural variants and newly-arisen mutations in Chapter 3. I first characterized the effects on expression of natural variants in the *TDH3* promoter. As many natural variants exist in the context of each other as haplotypes in the wild, I characterized the natural variants both individually and as haplotypes. Using the collection of *TDH3 cis*-regulatory mutations, I constructed empirical distributions of mean expression level and expression noise to be used as the null expectations due to the mutation process alone. To determine whether *TDH3 cis*-regulatory natural variation reflects the mutation process or selection, I compared the distributions of mean expression level and expression noise to those of the null. The results suggest that, for mean expression level, natural variation is consistent with the underlying mutation process. In addition, epistasis is rare among natural variants for mean expression level. However, when tested as haplotypes, the natural variants exhibited significantly lower expression noise than that expected of the mutation process alone. This suggests that selection has favored, via epistatic interactions, *cis*-regulatory mutations with decreased expression noise.

Determining the evolutionary significance of variation observed in nature is a fundamental task in evolutionary biology. However, this is challenging for regulatory variation due to lack of functional annotation among regulatory DNA sequence [Zhen & Andolfatto 2012]. The challenge lies in choosing the neutral sites against which to compare the natural sequence variation. Rather than choosing putatively “non-functional” sites as the neutral model, I constructed an empirical null representing the distribution of functional effects of random genetic variants. This distribution thus reflects the mutation process alone and is locus-specific. The results of the comparison between *cis*-regulatory *TDH3* natural variants and the empirical null suggest that selection can act on expression noise. This may be due to the relatively small effect size of mutations on mean expression level compared to that of expression noise. For *TDH3*, the effect of a mutation on expression noise may be more readily acted on by selection than the effect on mean expression level.

The empirical neutral models used in Chapter 3 consist of distributions of phenotypes of the *cis*-regulatory mutations because selection acts at the phenotypic level. Ultimately, genetic variants are selected for or against on the basis of their effect on fitness. Comparing the distribution of fitness effects between the natural variants and the mutation spectrum is therefore a more direct test for selection. Since expression noise has been shown to impact fitness [Wang & Zhang 2011], it would also be interesting to examine the distribution of fitness effects of the *cis*-regulatory mutation spectrum. Indeed, such distribution is of fundamental interest across evolutionary biology [Eyre-Walker & Keightley 2007]. Constructing and comparing such distributions will require engineering the *TDH3* promoter mutants and natural variants into the native *TDH3* locus instead of the reporter transgene.

As mentioned earlier, both *cis*- and *trans*-regulatory changes can contribute to evolution. Can the relative contributions of mutation and selection to *trans*-regulatory expression variation in the wild be determined? Part of the challenge here lies in the construction of the null distribution of *trans*-regulatory mutation effects. While a large number of *cis*-regulatory mutations can be readily constructed due to its more discrete location and limited mutation target size, constructing the equivalent set of *trans*-regulatory mutations presents several obstacles. For instance, site-directed mutagenesis cannot be feasibly used due to the number and unknown identities of sites that must be targeted. Methods to elevate mutation rate, such as chemical mutagenesis, must then be used to obtain newly-arisen mutations. This necessitates identification of mutations in the genome, although this can be readily accomplished [Duveau *et al*. *In prep*.]. Last but not least, newly-arisen mutations should be isolated under little to no influence of selection. One way to achieve this may be to use chemical mutagenesis as in [Gruber *et al*. 2012], isolate a spectrum of mutagenenized cells as random mutants regardless of effect on expression, and use the distribution of effect of this spectrum as the empirical null. This distribution then may represent the combined spectrum of all mutations, including *cis*- and *trans*-regulatory. An alternative may be to use a synthetic regulatory network with known *trans*-acting factors regulating the target gene [e.g. Cantone *et al*. 2009]. Since each gene within such networks can be insulated from the endogenous genes outside the network [Blount *et al*. 2012], newly-arisen regulatory mutations may be much more readily identified. However, even if the synthetic regulatory network is truly insulated, characterization of mutations using such network may still not be representative of nature.

Epistatic interactions among natural variants in the *TDH3* promoter characterized in Chapter 3 contributed to variation in expression noise. Interaction between genetic variants and the environment may also contribute to phenotypic variation. Genotype-by-environment (GxE) interaction has been observed and characterized for gene expression in QTL studies [e.g. Landry *et al*. 2006; Li *et al*. 2006; Smith & Kruglyak]. GxE interactions for individual mutations affecting phenotypes such as sporulation efficiency [Gerke *et al*. 2010] and fitness [Remold & Lenski 2001] have also been characterized. Understanding how environment interacts with individual newly-arisen mutations to influence gene expression can provide further mechanistic insight into how mutation contributes to gene expression variation [Maranville *et al*. 2012].

To determine the effects of GxE interaction on gene expression, I tested the *TDH3 cis*-regulatory mutations in different environments in Chapter 4. Specifically, the mutants were tested in glycolytic, gluconeogenic, and high osmolarity environments to reflect conditions in which the *Tdh3p* protein performs different functions. The results suggest that GxE interactions are common among *TDH3 cis*-regulatory mutants, with variation in mean expression level and expression noise most frequently observed between glycolytic and gluconeogenic environments. This variation in the effect of mutations on gene expression is consistent with the *cis*-regulation of *TDH3* reflecting the dual function of *Tdh3p* in the metabolism of different carbon sources [Kuroda *et al*. 1994]. Combined epistatic and GxE (GxGxE) interactions were also observed to affect mean expression level, again, most frequently between glycolytic and gluconeogenic environments. These results suggest that the environment is an important determinant of the effects of different *TDH3 cis*-regulatory mutations and that genetic variants can readily interact with the environment and each other to generate phenotypic variation.

Understanding how genetic changes impact gene expression can reveal the genetic and molecular mechanisms underlying phenotypic evolution, as variation in some observable traits has been associated with modification of gene expression. The results in Chapter 3 suggest abundant interactions among components of gene regulation as a mechanism underlying gene expression variation. Such interactions may render the effect of mutations dependent on external environment and genetic background, thereby creating complex evolutionary trajectories for newly-arisen mutations. The abundance of GxE interactions may also underlie the biological function of cryptic genetic variation, which is standing genetic variation that does not contribute to the "normal" range of phenotypes but may do so after genetic or environmental perturbation [Gibson & Dworkin 2004]. Cryptic genetic variation may be important for adaptation, and thus it may be interesting to characterize the effect of GxE interactions among cryptic genetic variation within a population.

GxE interactions may cause mutations to exhibit different effects in different environments, such that the same set of mutations may achieve different rankings in effect in different environments. In other words, the same phenotype may be achieved in different environments with different mutations. This may help explain divergent genotypes underlying conserved phenotypes such as that observed at the sea urchin *endo16* locus [Romano & Wray 2003]. Further support of this idea may be gained by determining the impact on fitness of GxE interactions and testing if a particular range of fitness may be achieved through different mutations in different environments.

In this dissertation, I described and discussed studies investigating the effect of mutations that differ by *cis* or *trans* mode of action, the contribution of mutations to natural variation in gene expression, and the interactions between mutations and environment as a source of and a mechanism underlying variation in gene expression.

The overarching question to these studies is how do mutations contribute to variation in gene expression? This is largely motivated by the importance of changes in gene expression as a mechanism of phenotypic evolution and the role mutation plays as the original source of such genetic changes. To characterize the mutation process and the range of variation in gene expression it produces, examining genetic variants in wild populations does not provide a large and systematic sample. While much of the data presented in the studies here were derived from mutations generated *de novo*, it is still important to consider variants and conditions that exist in nature. For instance, evolutionary insight was gained by comparing the effects of natural variants on gene expression to those of the mutation spectrum. Investigating the effect of GxE interactions on the effect of mutations was also motivated by natural conditions in which the *Tdh3p* protein functions. For the future, characterizing the impact of mutations on phenotypes more directly acted on by evolutionary forces in nature (i.e. fitness) will complement my characterizations of mutations and natural variants in the reporter system.

References

Blount BA, Weenink T, and Ellis T. Construction of synthetic regulatory networks in yeast. *FEBS Letters* 586:2112-2121 (2012).

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

Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, Santini S, di Bernardo M, di Bernardo D, and Cosma MP. A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell* 137:172-181 (2009).

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, *et al*. Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a *Pitx1* Enhancer. *Science* 327:302-305 (2010).

Darwin, C. On the Origin of Species by Means of Natural Selection. (1859).

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, and Cookson WOC. A genome-wide association study of global gene expression. *Nature Genetics* 39:1201-1207 (2007).

Duveau F, Metzger BPH, Gruber JD, Mack K, Sood N, Brooks T, and Wittkopp PJ. Mapping small effect mutations in *Saccharomyces cerevisiae*: impacts of experimental design and mutational properties. *In preparation*.

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

Eyre-Walker A and Keightley PD. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* 8:610-618 (2007).

Ezawa K, Landan G, and Graur D. Detecting negative selection on recurrent mutations using gene genealogy. *BMC Genetics* 14:37 (2013).

Gerke J, Lorenz K, Ramnarine S, and Cohen B. Gene–Environment Interactions at Nucleotide Resolution. *PLoS Genetics* 6:e1001144 (2010).

Gibson G and Dworkin I. Uncovering cryptic genetic variation. *Nature Reviews Genetics* 5:681-690 (2004).

Gruber JD, Vogel K, Kalay G, and Wittkopp PJ. Contrasting Properties of Gene-Specific Regulatory, Coding, and Copy Number Mutations in *Saccharomyces cerevisiae*: Frequency, Effects, and Dominance. *PLoS Genetics* 8:e1002497 (2012).

Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics* 37:243-253 (2007).

Kuroda S. Otaka S, and Fujisawa Y. Fermentable and nonfermentable carbon sources sustain constitutive levels of expression of yeast triosephosphate dehydrogenase 3 gene from distinct promoter elements. *Journal of Biological Chemistry* 269:6153-6162 (1994).

Landry CR, Oh J, Hartl D, and Cavalieri D. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* 366:343-351 (2006).

Li Y, Álvarez O, Gutteling E, Tijsterman M, Fu J, Riksen J, Hazendonk E, Prins P, Plasterk R, and Jansen R. Mapping Determinants of Gene Expression Plasticity by Genetical Genomics in *C. elegans*. *PLoS Genetics* 2:e222 (2006).

Maranville JC, Luca F, Stephens M, and Di Rienzo A. Mapping gene-environment interactions at regulatory polymorphisms: Insights into mechanisms of phenotypic variation. *Transcription* 3:56-62 (2012).

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

Remold SK and Lenski RE. Contribution of individual random mutations to genotype-by-environment interactions in *Escherichia coli*. *PNAS* 98:11388–11393 (2001).

Romano LA and Wray GA. Conservation of *Endo16* expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Development* 130:4187-4199 (2003)

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, and Friend SH. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422:297-302 (2003).

Smith EN and Kruglyak L. Gene-Environment Interaction in Yeast Gene Expression. *PLoS Biology* 6:e83 (2008).

Stern DL and Orgogozo V. The loci of evolution: how predictable is genetic evolution? *Evolution* 62:2155-77 (2008).

Tirosh I, Reikhav S, Levy AA, and Barkai N. A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation. *Science* 324:659-662 (2009).

Wang Z and Zhang J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *PNAS* 108:E67-76 (2011).

West MAL, Kim K, Kleibenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, and St. Clair DA. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* 175:1441-1450 (2007).

Wittkopp P, Vaccaro K, and Carroll S. Evolution of yellow Gene Regulation and Pigmentation in Drosophila. *Current Biology* 12:1547-1556 (2002).

Wittkopp PJ, Haerum B, and Clark A. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genetics* 40:346-350 (2008).

Wray G. The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews Genetics* 8:206-216 (2007).

Zhai W, Nielsen R, and Slatkin M. An Investigation of the Statistical Power of Neutrality Tests Based on Comparative and Population Genetic Data. *Molecular Biology and Evolution* 26:273-283 (2009).

Zhen Y and Andolfatto P. Methods to Detect Selection on Noncoding DNA in *Evolutionary Genomics: Statistical and Computational Methods, Volume 2, Methods in Molecular Biology*, vol. 856, edited by Anisimova M. Humana Press, New York (2012).