The evolution and function of regulatory regions in yeast

Caitlin F. Connelly

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Joshua Akey, Chair

Maitreya Dunham

Joe Felsenstein

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

The evolution and function of regulatory regions in yeast

Caitlin F. Connelly

Chair of the Supervisory Committee:

Professor Joshua Akey

Genome Sciences

Gene regulatory changes have long been theorized to be a source of evolutionary novelty. In more recent years, we have learned that noncoding changes can have a large effect on phenotype and disease and can be of evolutionary importance in a variety of model systems. However, the study of regulatory regions is still hampered by challenges in identifying functional noncoding regions and testing their effects, and as such many basic questions remain unanswered about these regions. Namely, what is the location of functional regulatory regions, how does evolution affect these regulatory regions, what is their effect if altered on downstream traits such as gene expression, and what are the best ways to test these effects? In this dissertation, I use yeast as a model organism to assess the effects of noncoding variants and chromatin architecture on gene expression. I first address the question of how best to test for effects of genetic mutations on phenotypes by asking how well association mapping works in yeast and find that population

structure complicates the use of yeast in association mapping without careful choice of which strains to use. Secondly, I assess the evolutionary pressures acting on noncoding regions in yeast, specifically the differences in pressures on potentially functional sites, motif binding sites, and the other noncoding sites, and find that there is strong purifying pressure acting at motif binding sites. I then use data from yeast strains to associate noncoding variants with differences in gene expression. Finally, I use methods to map chromatin accessibility to ask how differences in accessibility of open chromatin regions upstream of genes affect differences in gene expression between species. I find that differences in chromatin accessibility are numerous, that these differences appear to be driven primarily by genetic changes in *cis*, and that the effects of chromatin accessibility on gene expression are modest. In conclusion, my thesis work has mapped regulatory regions in yeast, revealed the functional effects of these regions and the variants that lie within them, and suggested alternative methods for associating genetic variation with phenotypes in yeast.

Table of Contents

List of Figures……..8
List of Tables……... 9
Acknowledgements ……10
1 – Introduction ……11
1.1 The study of gene regulation ……...11
1.2 The significance of noncoding variation……..12
1.3 Mechanisms influencing gene expression variation…… 13
1.4 Genetic mapping of functional regulatory variation……14
1.5 Computational methods to identify functional non-coding regions…… 15
1.6 Use of molecular intermediates to identify candidate regions……16
1.7 Evolution of noncoding regions……...19
1.8 Yeast as a model…….. 19
1.9 Research goals…… 20
2 – On the prospects of whole-genome association mapping in Saccharomyces cerevisiae….....21
2.1 Summary……..21
2.2 Introduction……..22
2.3 Results……..23
2.3.1 GWA studies of quantitative traits in yeast can have high type I error rates …….23
2.3.2 Association studies of Mendelian traits and cis-acting QTL are feasible……...29
2.3.3 Association mapping of mtDNA copy number…….. 31
2.3.4 Strategies for enabling GWA studies in yeast ……... 35
2.4 Discussion……36
2.5 Materials and Methods……37
3- Population genomics and transcriptional consequences of regulatory motif variation in globally diverse Saccharomyces cerevisiae strains……... 42
3.1 Summary……..42
3.2 Introduction……..43
3.3. Results……45
3.3.1 Regulatory motif variation across *S. cerevisiae* strains…….. 45

3.3.2 Evolutionary forces shaping patterns of polymorphism and divergence of regulatory sequences……..... 48
3.3.3 Patterns of motif polymorphism are significantly correlated with transcriptional variation among strains……... 51
3.3.4 Features associated with transcriptional divergence……...55
3.4 Discussion……56
3.5 Materials and Methods……59
4 – Evolution and genetic architecture of chromatin accessibility and function in yeast……...64
4.1 Summary……..64
4.2 Introduction……..65
4.3 Results……..67
4.3.1 Differences in chromatin accessibility within and between species……...67
4.3.2 Genetic architecture of chromatin differences……71
4.3.3 Disrupted motifs are associated with cis effects……73
4.3.4 Differential footprints for certain DNA binding factors found at trans effects loci……... 76
4.3.5 Effects on gene expression……..76
4.4 Discussion……79
4.5 Materials and Methods……81
5 – Summary and future directions ……....89
5.1 Summary……..89
5.2 Explorations of different strain sources……... 89
5.3 Regulatory changes in different environments……90
5.4 Integrating more levels of information……91
5.5 Experimental methods……93
5.6 Predictions of causative alleles……94
5.7 Concluding remarks……94
References……96
Appendix A- Supplementary material for Chapter 2……...110

A.1. Tables……110
Appendix B- Supplementary material for Chapter 3……...112
B.1. Figures……112
B.2. Tables……114
Appendix C- Supplementary material for Chapter 4……...123
C.1. Figures……123
C.2. Tables……124

List of Figures

Figure 1.1. Methods to identify functional regulatory regions…… 17
Figure 2.1. GWA studies of quantitative traits in yeast result in elevated type I error rates…... 25
Figure 2.2. Heterogeneity of observed type I error rates across the genome……...27
Figure 2.3. Type I error rates from simulations using three yeast datasets…… 29
Figure 2.4. Type I error rates for GWA studies of Mendelian traits. …… 30
Figure 2.5. mtDNA copy number across 36 *S. cerevisiae* strains. …… 33
Figure 2.6. GWA study of mtDNA copy number. …… 34
Figure 2.7. Type I error rates from simulations using *S. paradoxus*……... 36
Figure 3.1. Examples of highly divergent regulatory regions across *S. cerevisiae* strains…….47
Figure 3.2. Evolutionary forces acting at intergenic regions……...49
Figure 3.3. Effects of variants at specific motifs on gene expression……..53
Figure 3.4. Examples of motifs effecting gene expression……55
Figure 4.1 Patterns of chromatin accessibility within and between *S. cerevisiae* and *S. paradoxus*…… 70
Figure 4.2 Schematic of approach to detect *cis* and *trans* effects on chromatin accessibility….72
Figure 4.3 *Cis* and *trans* effects on chromatin accessibility…… 73
Figure 4.4 Motifs contributing to *cis* and *trans* effects……75
Figure 4.5 Gene expression and chromatin accessibility……77
Figure 5.1 Molecular intermediates and rates which can be measured genome-wide……91
Figure B.1. Genome-wide phylogeny of the 37 strains……... 112
Figure B.2. Evolutionary pressures at noncoding sites using varying cutoffs for motif calling..113
Figure C.1. Enrichment of FAIRE signal in NFRs and intergenic regions……. 123

List of Tables

Table 2.1. Summary of power and Type I error rates in the two yeast datasets……. 24
Table 3.1. High confidence regulatory polymorphisms……. 52
Table 3.2. Motifs associated with consistent expression differences……. 54
Table A.1. Mitochondrial DNA Copy Number in SGRP Strains……... 110
Table A.2. Mitochondrial DNA Copy Number at the Most Highly Associated SNP…… 111
Table B.1. Highly differentiated intergenic regions…… 114
Table B.2. Motifs under purifying selection…… 118
Table B.3. Regions significant for being under purifying selection by the MK test……... 121
Table B.4. Power to detect associations as a function of effect size in our data set…… 122
Table C.1. Power and false positive rate for cis and trans tests……..124
Table C.2. Summary of different criteria used to investigate the relationship between chromatin and gene expression QTL……. 124

Acknowledgements

I would like to thank past and present members of the Akey lab, Joshua Akey, Laura Scheinfeldt, Shameek Biswas, Thomas Nicholas, Jenny Madeoy, Marnie Johansson, Tim O'Connor, Jacob Tennessen, Dan Skelly, Leslie Emery, Wenqing Fu, Ben Vernot, Jenny Andrie, Rachel Gittelman, Sunjin Moon, and Dayna Akey. I'd also like to thank my supervisory committee, John Stamatoyanopolous, Maitreya Dunham, Joe Felsenstein, and Jon Wakefield. Finally, I'd like to thank my family: my parents, Laurie and Pat Connelly, my brother and sister-in-law, Mike and Allison Connelly, and my husband, Peter Sudmant.

Chapter 1

Introduction

1.1 The history of the study of gene regulation

Gene regulatory changes have long been proposed to be a source of evolutionary novelty (Britten and Davidson 1971, King and Wilson 1975). Both of these highly-cited papers based their claims in part on the findings that many classes of protein coding genes that were first studied appeared to function very similarly across evolutionarily diverse taxa. Indeed, the conservation of protein function across diverse taxa was one of the landmark findings of the first years of protein and DNA sequencing. In particular, King and Wilson compared the protein coding regions of genes between humans and chimpanzees and argued that the amount of DNA differences was not significant compared to the amount of observed phenotypic difference (1975). Similarly, Britten and Davidson cite the fact that many enzymes are shared between prokaryotes and mammals (Britten and Davidson 1971). Therefore, these papers conclude, since there are limited differences between protein coding genes, there must be differences in how the genes are regulated. In the years since these landmark papers, we have learned a great deal about both gene regulation and protein coding gene evolution. In fact, there are numerous examples of the importance of protein coding gene evolution to evolution (Golding and Dean 1998, Harms and Thornton 2010). One well-studied example is the opsin protein family, which through a process of duplication and divergence has produced opsin proteins in different species that influence which wavelengths of light those species can see. For example, opsins that distinguish red and green colors in humans are functionally distinguished by 7 amino acid substitutions (Asenjo *et al.* 1994). Another example is a protein substitution in hemoglobin which has been under selection because it gives heterozygotes an advantage against malaria (Allison 1954).

Clearly, both protein coding and regulatory changes are important to evolutionary change (Hoekstra and Coyne 2007). However, regulatory change continues to be less well understood and less well studied due to a number of challenges.

1.2 Contemporary examples of the significance of noncoding variation

We have learned for a growing number of specific examples that regulatory changes, or genetic changes to noncoding regions, can have a significant effect on phenotypes and ultimately our understanding of evolution and human health and disease. The classical example of noncoding changes causing evolutionary change in humans is the example of lactase persistence. Variants upstream of the lactase gene allow the enzyme to be expressed into adulthood and have been under strong positive selection in humans in multiple populations (Enattah *et al*. 2002, Tishkoff *et al*. 2007). In human diseases, studies of a handful of diseases have identified noncoding variants which are associated with disease, such as promoter mutations in APOE which are associated with Alzheimer's disease (Ward and Kellis 2012). In addition, the majority of GWAS hits are located in noncoding regions, suggesting that noncoding changes may be underlying these hits (Maurano *et al*. 2012) In most cases, however, the causal alleles are not known, and most GWAS hits have a very small effect on phenotypes.

Looking between species at whether differences in gene regulation underlie species differences and diversity within species as proposed by Britten and King, examples of noncoding variants causing phenotypic differences between or within species have been accumulating in diverse lineages. For example, noncoding variants have been identified causing pigmentation differences in *Drosophila,* skeletal reduction in stickleback fish, skin wrinkling in domesticated dog, loss of neck feathers in chicken, and blond hair color in humans (Wittkopp *et al.* 2002, Shapiro *et al.* 2004, Olsson *et al.* 2011, Mou *et al.* 2011, Guenther *et al.* 2014).

Regulatory changes are also thought to be evolutionarily advantageous for several theoretical reasons. Specifically, changes to a gene's regulation may be less pleiotropic than changes to the coding region of a gene (Breuker *et al.* 2006). In addition, the theory of the evolution of development (evo-devo) holds that certain types of changes may be more likely to happen through regulatory change, specifically changes in body plan, or the addition or loss of body parts over time (Carroll 2008). This theory is supported by many of the examples of noncoding variation identified so far and discussed in the last paragraph, such as the loss of specific bones in fish, gain of expression of a protein in the skin which causes skin wrinkling, and loss of expression of genes associated with neck feathers. However, as multiple researchers have now pointed out, the concept of "regulatory change" is complicated by the fact that these changes in expression could also be due to protein coding changes to regulatory proteins such as transcription factors (Hoekstra and Coyne 2007).

The role of these 'trans' regulators is also important to evolution. For example, one of the major developmental genes identified in Drosophila, *fushi tarazu*, one of the pair rules genes which specifies formation of alternating segments, has completely changed its function in a closely related species (Heffer and Pick 2013). Another potential example of this is the finding that lower levels of hemoglobin in Tibetan populations who live at high altitude has been linked to the region of the transcription factor EPAS1 (Yi *et al.* 2010), although the most-differentiated SNP was in an intron, suggesting that it may be the regulation of this transcription factor which causes the phenotype. Overall, there is still much to learn about how regulatory changes and protein coding changes to regulatory proteins affects downstream phenotypes.

1.3 Mechanisms influencing gene expression variation.

Basic research into how gene expression is regulated have identified many ways in which

gene regulation can be influenced by genetic variation in noncoding regions. In eukaryotic cells, all genomic DNA in the nucleus is packaged into chromatin, the architecture of which plays a prominent role in regulating gene expression. Mechanistically, noncoding variants can act in several ways to affect this architecture and modulate gene expression. They can alter the binding of sequence-specific binding proteins such as transcription factors, which can result in remodeling of the local chromatin architecture (Farnham 2009). Variants can also influence nucleosome positioning and chromatin structure, through intrinsic sequence preferences for nucleosome positioning and recruitment of chromatin remodeling complexes (Jiang and Pugh 2009). Finally, variants at key positions can interfere with the basal transcriptional machinery (Lee *et al.* 2002).

1.4 Genetic mapping of functional regulatory variation

Genetics approaches have previously been used to identify variation associated with gene expression variation genome-wide. These expression QTL (eQTL) studies have led to many insights into the genetic architecture of gene expression differences. One question which these studies have attempted to address is whether gene expression differences are the result of genetic changes that are local to or distant from the gene whose expression is being tested, often referred to as *cis* and *trans* effects. One study design to map eQTLs uses the offspring of a cross between divergent individuals or species to map regions of the genome associated with expression differences for each gene (Jansen and Nap 2001, Ehrenreich *et al.* 2009). A second method which can identify whether differences in gene expression are due to changes in *cis* or in *trans* without mapping to a specific region for *trans* effects is to test for allele-specific differences in gene expression in a diploid cross between two strains and to measure expression in the parents (Wittkopp *et al.* 2004, Wittkop *et al.* 2008, Wilson *et al.* 2008). These expression QTL (eQTL)

studies have identified numerous linked regions contributing to gene expression differences between strains or species and have emphasized the importance of *cis* alleles in affecting gene expression (Skelly *et al.* 2009). For example, in a cross between divergent yeast strains, Brem *et al.* found that 32% of the genes (185 transcripts) showing linkage to a genomic location were in *cis* (2002). Recent studies of eQTLs in humans support an important role for *cis*-eQTL; Pickrell *et al.* found that the vast majority of associations detected in 69 human LCL cell lines were *cis* (2010). *Trans* eQTLs can also have a large effect (*Fu et al.* 2009, Fehrmann *et al.* 2011, Fairfax *et al.* 2012). For example, the auxotrophic markers in yeast are linked to differences in expression of a large number of genes (Brem *et al.* 2002). *Trans* eQTLs appear to be less common, though they are also harder to detect. Researchers have also attempted to learn about what types of variation may be contributing to these eQTLs; in humans, they have found that the strongest hits were more likely to be indels (Lappalainen *et al.* 2013). For sites within coding regions, non-synonymous sites are overrepresented (Lappalainen *et al.* 2013). eQTL studies have succeeded in generally identifying regions of the genome linked to expression differences; however, it is still challenging to narrow down regions to causal variants.

1.5 Computational methods to identify functional non-coding regions

Numerous methods have been used to identify potentially functional non-coding regions on a finer scale. One approach has been to look for sequence motifs representing the preferred sites for known DNA binding proteins (Bailey and Elkan 1995, Hughes *et al.* 2000, Kel *et al.* 2003, Xie *et al.* 2005, Stormo 2000, Vavouri and Elgar 2005). However, it is often unclear how many of the predicted sites are functional. Computational methods have also been used to identify potential binding sites for nucleosomes (Segal *et al.* 2006, Miele *et al.* 2008, Yuan and Liu 2008). Another approach has been to identify regions that are among the most highly

conserved in the genome (Siepel *et al.* 2005, Pollard *et al.* 2010). These methods do not identify noncoding regions specifically, but a substantial fraction of the identified regions are noncoding. While these represent genomic regions likely to be functionally important, experimental data suggests that a large fraction of binding sites for specific regulatory factors are not constrained between species, in part due to lineage-specific use of regulatory elements (ENCODE Project Consortium 2007, Borneman *et al.* 2007). Consequently, between species conservation-based methods likely miss many functional elements.

1.6 Use of molecular intermediates to identify candidate regions

Another approach used to identify potentially important binding sites is to use more direct experimental evidence. There has been great interest in mapping transcription factor binding using ChIP-seq, and to assess what effects changes to binding have on gene expression (Figure 1.1). Researchers have found that there is significant variation in binding even between closely related species (Bradley *et al.* 2010, Schmidt *et al.* 2010). However, more recent work has shown that the functional effects of these differences in binding are much less than the observed differences in binding (Cusanovich *et al.* 2014, Kasowski *et al.* 2013). For example, knocking down 59 transcription factors in LCL cell lines, Cusanovich *et al.* were able to ask which genes changed their expression levels when a particular transcription factor was knocked down and how this corresponded to the binding sites of those transcription factors identified by ChIP-seq (2014). They found that between 46.4% and 99.1% of binding sites did not show differential expression after knocking out that transcription factor (Cusanovich *et al.* 2014). This is in agreement with earlier studies that found 46% and 45% correlation between binding of NFKB and PolII with expression of downstream genes (Kasowski *et al.* 2010). For the yeast transcription factor Ste12, 28% of target genes had a correlation coefficient greater than 0.335

between the level of Ste12 binding and levels of gene expression across segregants (Zheng *et al.* 2010). Overall, we are still learning more about how to identify a functional binding site.

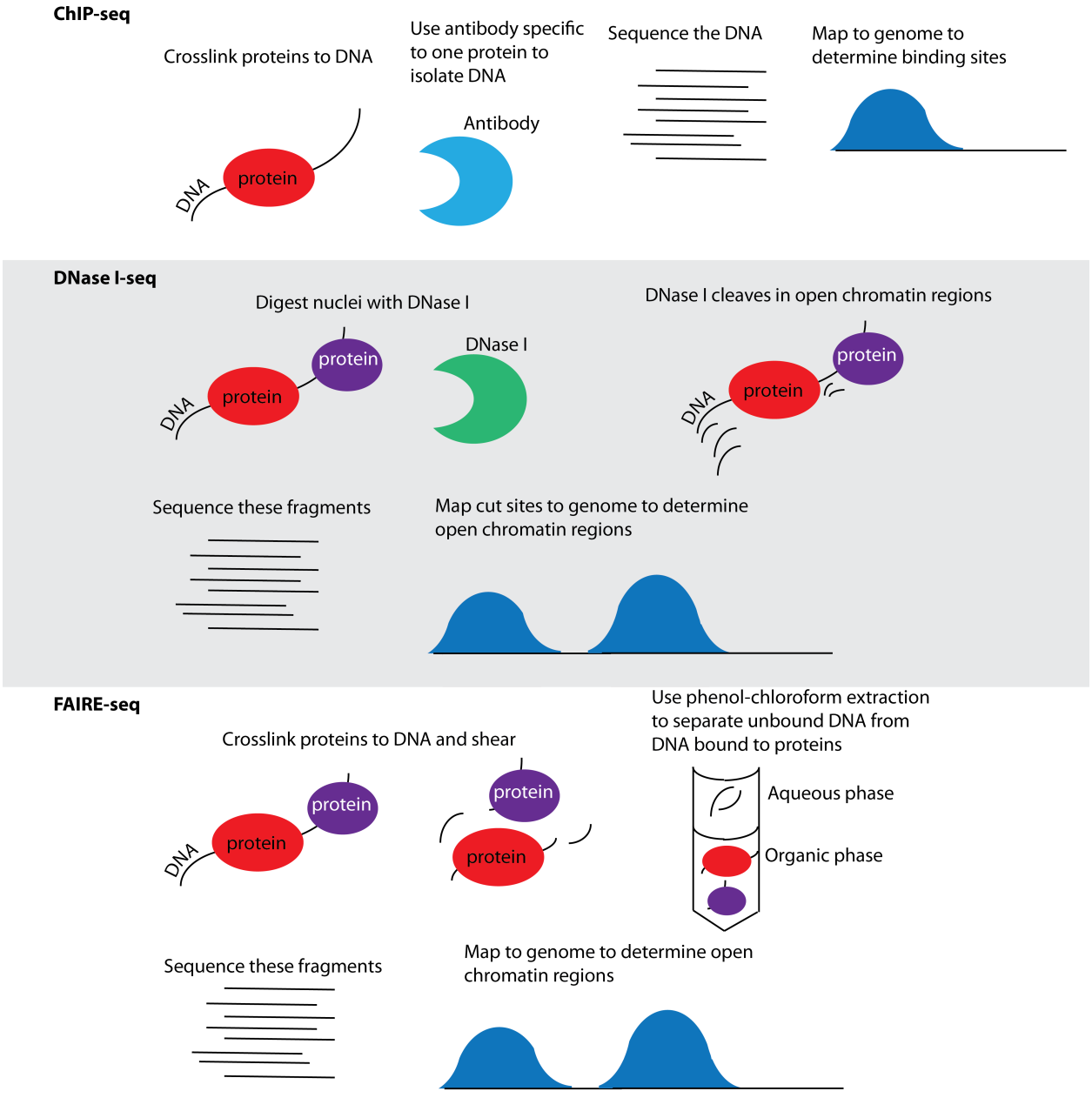

Figure 1.1 Methods to identify functional regulatory regions. Top, ChIP-seq is used to map the binding sites of DNA-binding proteins. A single experiment can identify genome-wide binding sites; however, only one protein can be mapped at a time. Middle, DNase I-seq is used to identify open chromatin regions. To do so, nuclei are digested with DNase I, fragments of digested DNA are sequenced, and cut sites are mapped to the genome. Bottom, FAIRE-seq also identifies open chromatin regions. Proteins are cross-linked to DNA and sheared, then unbound fragments of DNA are isolated by phenol-chloroform extraction and sequenced.

Another approach is to use one of several methods which have been used to map chromatin architecture and accessibility. The objective of these methods is to detect open chromatin regions which are likely to be active regulatory regions. One widely used method is the DNase I assay, which identifies regions which are more sensitive to cleavage by the enzyme DNase I (Galas and Schmitz 1978, Dorschner *et al.* 2004, Sabo *et al.* 2006, ENCODE Project Consortium 2007, Hesselberth *et al.* 2009). By mapping the sites of these cleavages, one can identify regions of open chromatin structure that are accessible to DNA binding proteins, which are called DNase I hypersensitive sites in the case of the DNase I assay. One advantage of these types of methods is that in contrast to ChIP-Chip and ChIP-Seq, which can only probe the locations of regulatory sequences for a specific transcription factor, mapping open chromatin can reveal information about interactions between DNA and any protein which binds DNA. In recent years, other methods with a similar aim have been developed, namely FAIRE-seq (Formaldehyde-Associated Isolation of Regulatory Elements) and ATAC-seq (Giresi *et al.* 2007, Buenrostro *et al.* 2013, Figure 1.1)

QTL studies similar in approach to eQTL studies have also identified DNase I hypersensitivity QTLs and attempted to correlate functional differences in DNaseI hypersensitivity QTLs with eQTLs. Degner *et al.* found that 16% of DNase I hypersentivity QTLs are associated with eQTLs (2012). Looking between different primate species, researchers found that DNase I hypersensitivity site gains between species were enriched near genes whose expression had change, but that this was true for only 58 of 1182 genes with increased expression in humans compared to chimpanzees (Shibata *et al.* 2012). Overall, it seems that functional changes in binding are more predictive than looking at computationally predicted

binding site changes, and differences in chromatin accessibility can also affect downstream expression differences.

1.7 Evolution of noncoding regions

Sequencing of whole genomes of multiple species led to the observation that many motifs are rapidly gained and lost between species (Dermitzakis and Clark 2002, Moses *et al.* 2006, Borneman *et al.* 2007, Doniger and Fay 2007) despite expression levels being generally well conserved. This has led to interest in what types of evolutionary pressures are acting at noncoding regions. Some papers have addressed this by looking at noncoding regions generally (Andolfatto 2005). Others have looked at selective pressures acting specifically at functionally validated or conserved binding sites (He *et al.* 2011, Raijman *et al.* 2008) or within eQTL-linked regions (Ronald and Akey 2007). Depending on the species used, researchers have found evidence for widespread positive selection or negative selection acting at noncoding regions. Patterns of variation at noncoding regions are less well characterized within species, although there has been more work on this in recent years. For example, in humans DNase I hypersensitive sites have been found to be vary less between individuals than other noncoding regions, suggesting constraint on these regions (Vernot *et al.* 2012). Researchers have also found differences in how noncoding variants accumulate over time. Specifically, it has been shown that *cis*-eQTLs accumulate preferentially between species compared to *trans*-eQTLs (Wittkopp *et al.* 2008) This has been hypothesized to be because *trans*-eQTLs may have more pleiotropic effects which would be selected against over longer periods of time.

1.8 Yeast as a model

Yeast is an excellent system for studying noncoding variation because of its small genome size, the availability of diverse sequenced strains and species, its relative ease of experimental

manipulation, and its well-annotated noncoding regions. In addition, application of next generation sequencing technologies in yeast is particularly powerful because its small genome size allows sufficient sequence coverage to obtain increased resolution for protein binding mapping or accessibility mapping (Hesselberth *et al.* 2009). In the past few years, there has also been great interest in using yeast as a model system for association mapping, for the reasons mentioned above. Two large projects have produced genome-wide data for fairly large collections of strains; Liti *et al.* produced 37 fully sequenced strains at fairly low coverage (2006), and Schacherer *et al.* called SNPs for 63 strains (2009). However, many of the methods developed for association mapping have been developed for human populations, and it is not clear how well they will work for yeast.

1.9 Research Goals

I. Assess the feasibility of association mapping in yeast, both for association mapping genome-wide and for mapping of *cis* associations.

II. Using computational methods, identify functional noncoding regions, and assess the evolutionary forces acting upon these and their affect on gene expression differences between strains.

III. Using experimental methods, map chromatin accessibility and assess the architecture of chromatin accessibility and its effects on gene expression differences between species.

Chapter 2

On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*

This chapter has been published: Connelly CF, Akey JM (2012) On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*. Genetics 191:1345-53.

2.1 Summary

Advances in sequencing technology have enabled whole-genome sequences to be obtained from multiple individuals within species, particularly in model organisms with compact genomes. For example, 36 genome sequences of *Saccharomyces cerevisiae* are now publicly available, and SNP data is available for even larger collections of strains. One potential use of these resources is mapping the genetic basis of phenotypic variation through genome-wide association (GWA) studies, with the benefit that associated variants can be studied experimentally with greater ease than in outbred populations such as humans. Here, we evaluate the prospects of GWA studies in *S. cerevisiae* strains through extensive simulations and a GWA study of mitochondrial copy number. We demonstrate that the complex and heterogeneous patterns of population structure present in yeast populations can lead to a high type I error rate in GWA studies of quantitative traits, and that methods typically used to control for population stratification do not provide adequate control of the type I error rate. Moreover, we show that while GWA studies of quantitative traits in *S. cerevisiae* may be difficult depending on the particular set of strains studied, association studies to map *cis*-acting quantitative trait loci (QTL) and Mendelian phenotypes are more feasible. We also discuss sampling strategies that could enable GWA studies in yeast and illustrate the utility of this approach in *Saccharomyces paradoxus*. Thus, our

results provide important practical insights into the design and interpretation of GWA studies in yeast, and other model organisms that possess complex patterns of population structure.

2.2 Introduction

Association studies have become a dominant paradigm in human genetics, with over 2000 studies published to data (Hindorff *et al.* 2009). The application of genome-wide association (GWA) studies to model organisms is a potentially powerful approach to rapidly identify causal variants that mediate heritable phenotypic variation (Risch and Merikangas 1996; Edwards *et al.* 2009; Atwell *et al.* 2010). However, a complication of GWA studies in many model organisms is the potential for type I errors produced by population structure, which lead to increased type I error rates (Lander and Schork 1994). Numerous methods have been developed to address this issue, which take structure into account when testing for associations (Pritchard *et al.* 2000; Patterson *et al.* 2006; Price *et al.* 2006; Kang *et al.* 2008). However, although these approaches work well in the cases of modest levels of population structure typical of human populations, it is unclear how effective they will be in different model systems.

Saccharomyces cerevisiae is a powerful model organism for genome-wide studies because of its small genome size, which has facilitated extensive sequencing of species and strains. For example, the Saccharomyces Genome Resequencing Project has performed whole- genome sequencing on 36 laboratory and wild strains of the model yeast *Saccharomyces cerevisiae* (Liti *et al.* 2009). Genetic diversity between these strains is high, and the average pair- wise distance between polymorphisms is 168 base pairs. Moreover, linkage disequilibrium (LD) is low; the half-life of LD is on average 3 kb. In theory, these characteristics should make *S. cerevisiae* strains a powerful resource for fine-scale mapping of associated loci. Indeed, association studies are beginning to be pursued using these and other strains (Mehmood *et al.* 2011; Muller *et al.*

2011).

However, the genomic patterns of population structure in *S. cerevisiae* strains are complex. Among the SGRP strains, the patterns of structure are complex and vary considerably across the genome; more than half the strains are considered mosaics (Liti *et al.* 2009). Thus, the prospect of GWA studies in yeast remains ambiguous. To address this issue, we performed a comprehensive set of analyses including simulations and an empirical GWA study of mtDNA copy number. We find that GWA studies in *S. cerevisiae* strains may be challenging for complex traits, although potentially feasible for Mendelian traits and in the identification of *cis*-acting variation. In addition, we suggest alternative study designs to mitigate the confounding effects of population structure, which will be broadly applicable to model organisms.

2.3 Results

2.3.1 GWA studies of quantitative traits in yeast can have high type I error rates

In order to evaluate the effects of population structure on GWA studies in yeast, we first performed extensive simulations conditional on the SGRP sequence data. In the simulations, we identified all SNPs where the minor allele was present in at least ten strains (MAF ranged from 0.27 – 0.50), as association studies with rare variants would have very low power in this data set, and selected tagSNPs from these (n = 18,637). Next, we randomly selected a "causal" SNP and assigned phenotypes to each strain based on the allele that the strain carried at this causal site. We initially focused on simulating quantitative trait loci (QTL) with large effects, which explain approximately 17% of phenotypic variation (see Methods for details). We then performed a GWA study on the 18,637 tagSNPs using three different analyses: a simple t-test, a t-test on the residuals of phenotypes regressed on the first principal component (Price *et al.* 2006), and a more

sophisticated linear mixed-model implemented in EMMA (Kang *et al.* 2008), which was designed for association mapping in model organisms. EMMA has been shown to dramatically reduce the type I error rates due to population structure (Kang *et al.* 2008; Atwell *et al.* 2010). This process was repeated 1,000 times and we calculated the average type I error rate for both t-tests and the linear mixed-model of EMMA.

As expected, t-tests lead to a substantial inflation of the type I error rate (Figure 2.1B). For example, at a nominal Type I error level of 0.05, the observed false positive rate was 0.20, which is four times higher than expected. Including the first principle component as a covariate was effective in reducing the type I error rate; however, the power was also dramatically reduced, rendering this method ineffective for this dataset (see Table 2.1). Surprisingly, EMMA, while performing considerably better than t-tests, also exhibited an elevated type I error rate, with nearly twice as many false positives relative to that expected at a nominal level of 0.05 (Figure 2.1).

Table 2.1. Summary of power and Type I error rates in the two yeast datasets.

Dataset	Expected Type I Error Rate	Power			Observed Type I Error Rate		
		t-test	t-test, PCA corrected	EMMA	t-test	t-test, PCA corrected	EMMA
Liti et al. (n = 36)	$\alpha = 0.05$	0.804	0.277	0.729	0.197	0.019	0.098
	$\alpha = 1x10^{-5}$	0.033	0.001	0.014	$5.0x10^{-4}$	$2.95x10^{-6}$	$2.0x10^{-4}$
Schacherer et al. (n=63)	$\alpha = 0.05$	0.961	0.495	0.930	0.161	0.043	0.084
	$\alpha = 1x10^{-5}$	0.137	0.008	0.152	$3.0x10^{-4}$	$1.57x10^{-5}$	$2.0x10^{-4}$

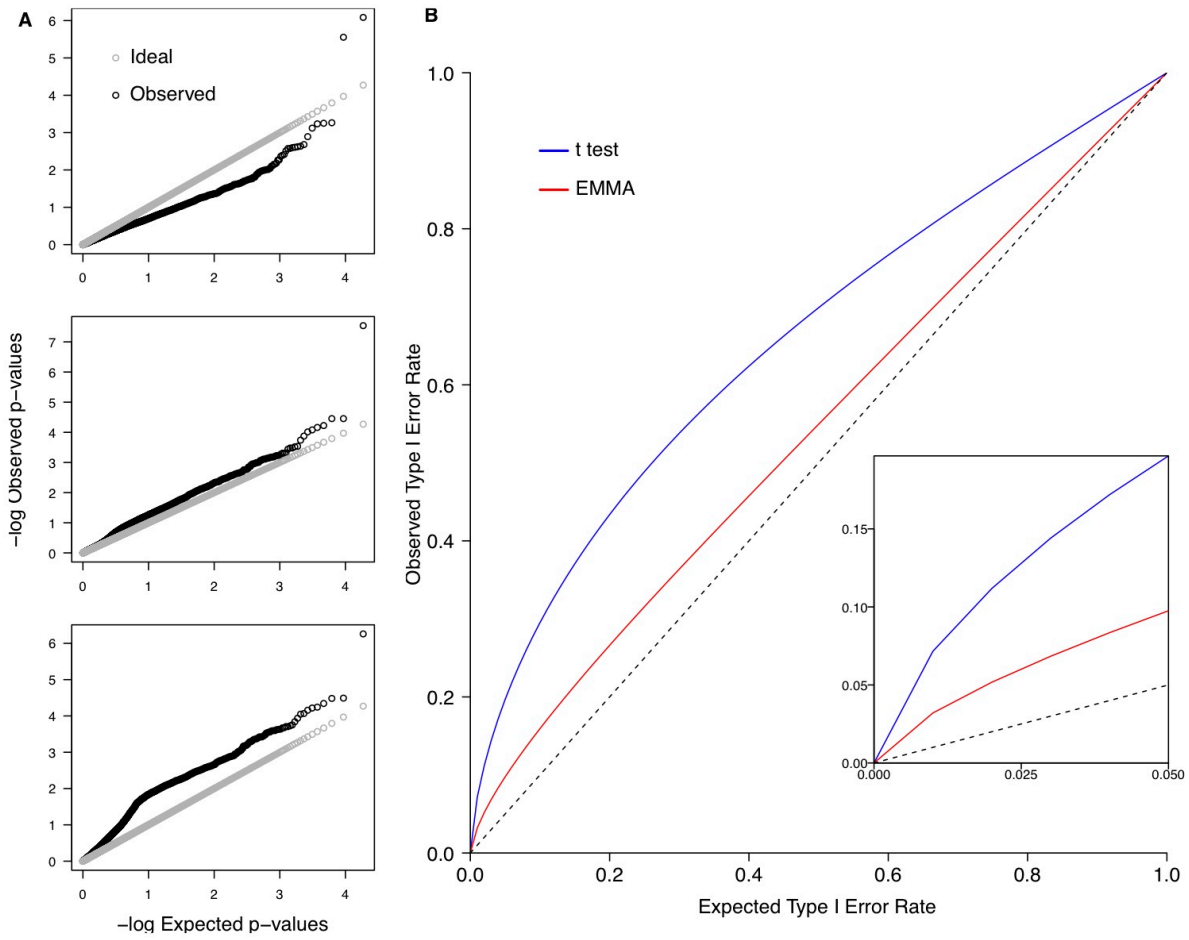

Figure 2.1. GWA studies of quantitative traits in yeast result in elevated type I error rates.

A) Representative qq plots of simulations using EMMA, which had at least one SNP that reached genome-wide significance. The observed p-values are shown in black and the expected p-values are shown in gray. B) The mean observed type I error rate from 1000 simulations is plotted versus the expected type I error rate for association tests done using a simple t test (blue) and EMMA (red). The theoretical expectation in the absence of population structure is shown as a dashed line. Inset, detail of the observed vs. expected type I error rates at low expected error rates.

To explore the cause of the elevated type I error rates, we investigated whether there was variation in the type I error rate at individual SNPs across the 1,000 simulations. In other words, we were interested in identifying regions of the yeast genome that were particularly susceptible to generating type I errors. We found that the error rate at individual SNPs at an expected rate of 0.05 varied from 0.038 to 0.168, with a mean of 0.096 (Figure 2.2, Table A.1). The error rate of

0.168 is much higher than would be expected by chance assuming type I errors are randomly distributed across the genome (p=2.80 x 10^{-12}). We also found that SNPs with the highest type I error rate (defined as those in the top 1% of all SNPs tested) are highly correlated (average pairwise $r^2 = 0.58$; also see Figure 2.2) even though they are distributed across the yeast genome. Strikingly, the genealogy of SNPs with the highest type I error rates differs substantially from the average genome-wide genealogy (Figure 2.2). Specifically, many of the mosaic strains identified by Liti *et al.* (2009) cluster with the wine/European strains, for SNPs with the highest type I error rate (Figure 2.2). Thus, the distinct pattern of structure in these regions compared to the genome-wide average likely makes it difficult to fully account for stratification in association test statistics.

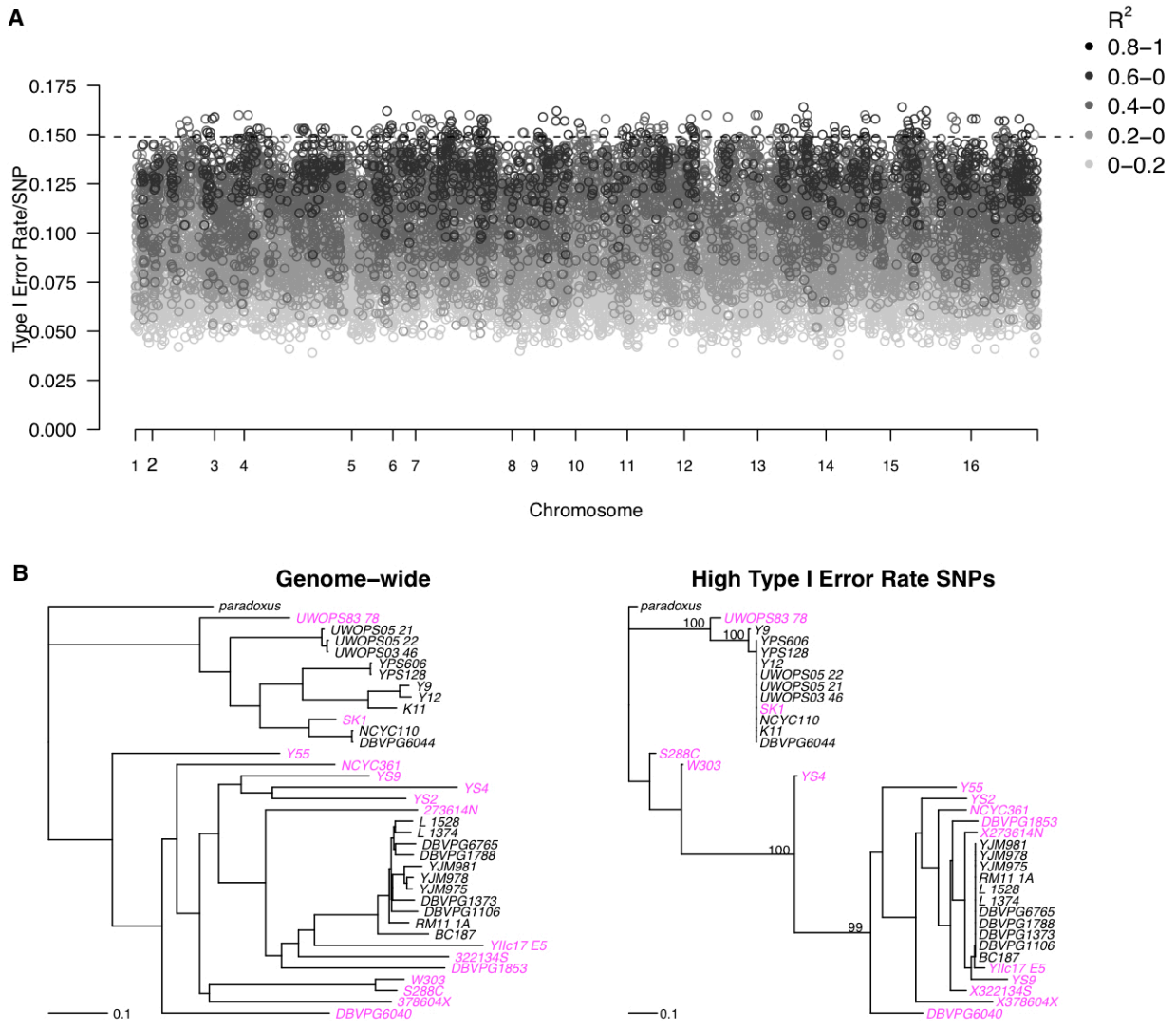

Figure 2.2. Heterogeneity of observed type I error rates across the genome.

A) Observed type I error rate at each SNP across the 1000 simulations. Dotted line shows the cutoff for high type I error rate SNPs (see part B). Correlation (r2) between SNP alleles with the highest type I error rate and all other SNPs is represented by the color (see legend). B) Neighbor-joining trees constructed from the genome-wide data (18,637 tagSNPs, left) and from the high type I error rate SNPs. Strains identified as mosaics by Liti et al.. (2009) are shown in pink. The number of bootstrap replicates supporting each clade out of 100 total are labeled.

In order to see how representative our data from the SGRP strains was of other yeast datasets, we analyzed a second dataset consisting of 63 *S. cerevisiae* strains (Schacherer *et al.* 2009). We performed simulations using the same methods as above (see Methods). We again found an elevated type I error rate using EMMA, though it was modestly lower than in the

original dataset of 36 strains (Table A.1). Similarly to the SGRP dataset, including one principle component as a covariate was effective in reducing type I error rates at the expense of power (Table A.1). Finally, we also analyzed the *S. paradoxus* strains sequenced by the SGRP. These 35 strains differ from the *S. cerevisiae* strains in that the patterns of structure are more consistent; there are no mosaic strains. We hypothesize that this consistent pattern of structure is easier to correct for using EMMA, and indeed we found that the type I error rate was lower than that found in the SGRP strains (Figure 2.3), though still elevated above the expected rate.

We were also interested in more directly comparing the SGRP and Schacherer *et al.* datasets in order to gain insights into the differences in type I error rates between the two. In order to compare the two datasets, we resampled 36 strains from the 63 strains from Schacherer *et al.* (2009). We observed a similar small reduction in the type I error rate in this resampled set of 36 strains compared to the 63 strains, suggesting that it is the particular set of strains in the Schacherer *et al.* dataset that is responsible for the reduced type I error rate, not the larger sample size (Figure 2.3). Additionally, the Schacherer *et al.* strains appear to have the least amount of structure or recognizable subpopulations, which may account for their reduced error rate (Figure 2.3). For example, the average pairwise divergence at noncoding sites is 2.57×10^{-3} and 4.82×10^{-3}, respectively in the Schacherer *et al.* and SGRP strains, respectively. Therefore, the patterns of structure found in these two datasets result in elevated type I error rates, and differences in the amount and patterns of structure may explain the differences in error rates.

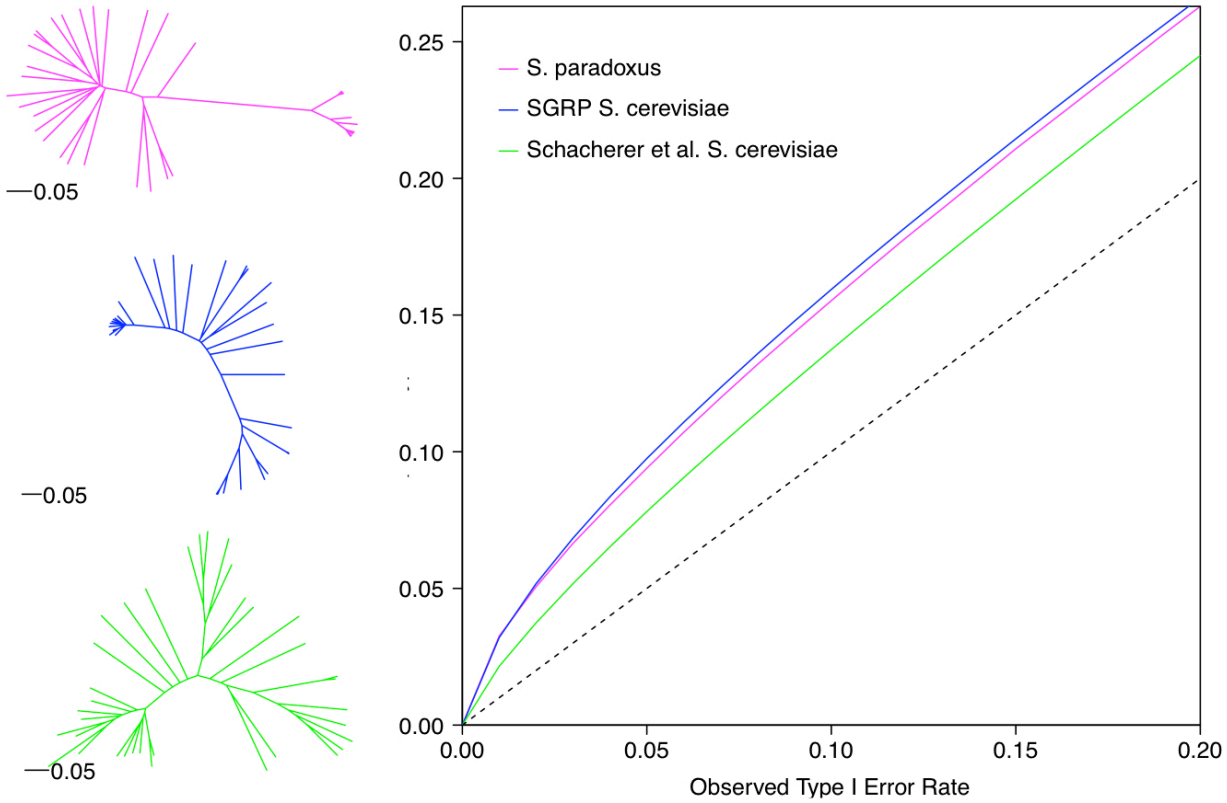

Figure 2.3. Type I error rates from simulations using three yeast datasets.

A. Neighbor-joining trees constructed from the tagSNPs from three datasets. Top, SGRP *S. paradoxus* strains. Middle, SGRP *S. cerevisiae* strains. Bottom, 36 *S. cerevisiae* strains sampled from Schacherer *et al.*. (2009). B. The mean observed type I error rate from 1000 simulations is plotted versus the expected type I error rate for association tests for the three datasets. Colors correspond to the genealogies on the left.

2.3.2 Association studies of Mendelian traits and cis-acting QTL are feasible

The results presented above demonstrate that GWA studies of quantitative traits in existing *S. cerevisiae* strains result in high type I error rates. Here, we evaluate whether other association strategies, such as the analysis of *cis*-acting QTL or focusing on Mendelian traits results in more interpretable results. To this end, we first repeated the simulations as described above for a Mendelian phenotype. In brief, for both the SGRP and Schacherer *et al.* datasets, a causal SNP was randomly selected and strains were assigned phenotypes based on the allele that they carried. A GWA study was then performed with EMMA using the 18,637 tagSNPs. As in

the previous simulations, the type I error rate was elevated; for the SGRP strains, at an expected rate of 0.05 the observed mean rate was 0.07, and similar patterns were observed for the Schacherer *et al.* strains. However, as shown in Figure 2.4, in both datasets, the causal SNP responsible for a Mendelian trait was on average easily distinguishable from the background distribution of p-values. Thus, GWA studies of Mendelian phenotypes are feasible even in yeast strains with complex patterns of population structure.

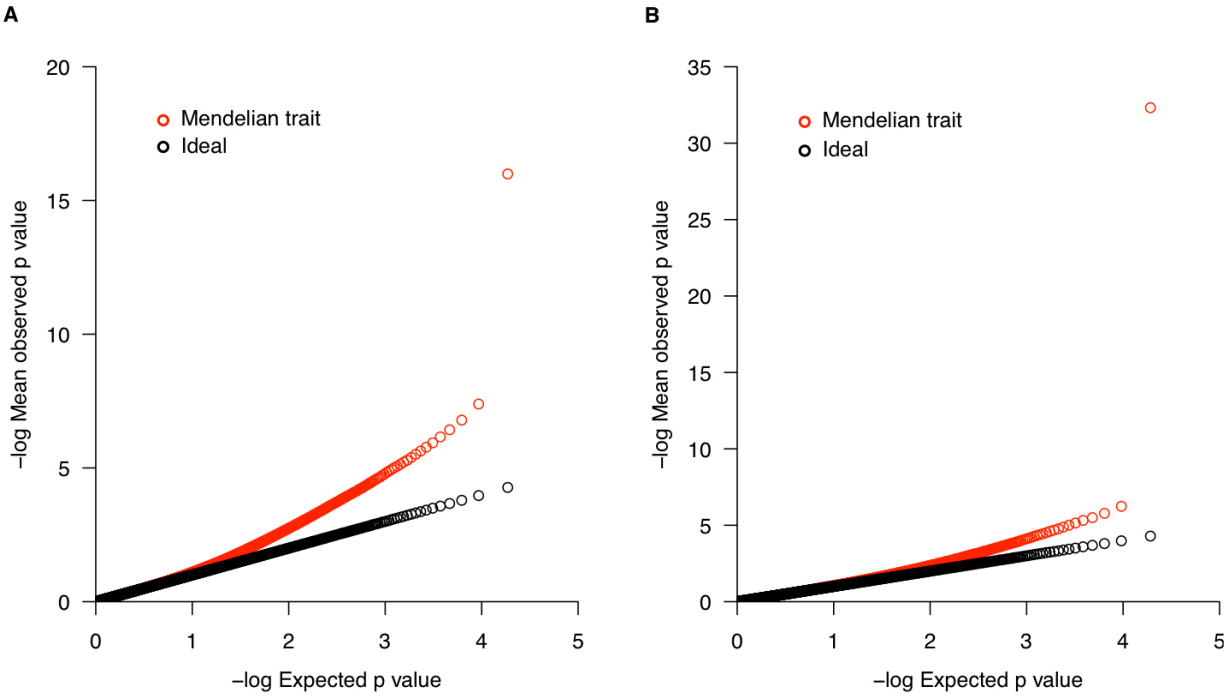

Figure 2.4. Type I error rates for GWA studies of Mendelian traits.

Panels A and B show qqplots averaged over simulations for the SGRP and Schacherer *et al.*. datasets, respectively. The average for 1,000 simulations of a Mendelian trait is shown in red, and the theoretical expectation is shown in black.

Next, we investigated the feasibility of *cis*-based association mapping, which in contrast to a GWA study only performs an association test at a locus of interest. For example, such study designs are popular in mapping *cis*-regulatory QTL, where gene or protein expression levels have been measured for a large number of genes (Skelly *et al.* 2011). Here, tests of association

are restricted to polymorphisms located in or adjacent to the gene whose transcript or protein abundance is being analyzed. To assess *cis*-based association mapping in the SGRP and Schacherer *et al.* datasets, we randomly selected a causal SNP nearby to each gene, assigned a quantitative phenotype as described above, and then used EMMA to perform an association test on 1,000 randomly selected unlinked polymorphisms. The goal of this analysis is to determine whether population structure also results in a higher type I error rate in *cis*-based association tests. We found that the type I error rate was again elevated above nominal levels. Specifically, at the expected rate of 0.05, we found that the observed rate for *cis*-association tests was 0.099 for the SGRP strains and 0.082 for the Schacherer *et al.* strains, similar to those for genome-wide tests for each dataset (see Methods). However, even though the type I error rate is similar to that observed for a GWA study, *cis*-based associations are more interpretable because of the limited number of hypothesis tests performed. For example, a *cis*-based association study of a single gene expression trait would require ~1-5 hypothesis tests (depending on local levels of LD and density of polymorphisms) versus ~19,000 hypothesis tests in a GWA study. In addition, the relative ease in functionally validating putative regulatory polymorphisms in yeast makes *cis*-associations an attractive approach. Thus, while some caution is warranted in interpreting *cis*-based association studies in yeast strains, they are a reasonable approach for identifying putative regulatory QTL.

2.3.3 Association mapping of mtDNA copy number

To complement the simulations, we next carried out a GWA study on mtDNA copy number in 36 *S. cerevisiae* strains. To measure copy number, we isolated total DNA containing both nuclear and mitochondrial DNA from each strain. To control for changes in mtDNA copy number throughout growth, we isolated DNA from the same growth stage for all strains and

obtained two biological replicates for each strain. We used qPCR to quantify the amount of mtDNA per cell in the strains by comparing the amounts of nuclear and mtDNA (see Methods).

Using this assay, we found that relative copy number per strain varied from 1 in BY4716 to 3.28 in BC187 (Figure 2.5). These differences in copy number appear to approximately correlate with the genealogy of the strains (Figure 2.5). Using a linear mixed model that explicitly takes into account biological and technical sources of variation, we found that there was a significant level of variation across all the strains ($p = 0.003$). The effect of strain in this model accounted for 28.5% of total phenotypic variation. Interestingly, the strain with the lowest copy number, BY4716, which is a haploid derivative of S288C, is a long-used laboratory strain and the source of the *S. cerevisiae* reference sequence. The low copy number could be the result of relaxed selection for respiratory abilities during its many generations in the lab environment.

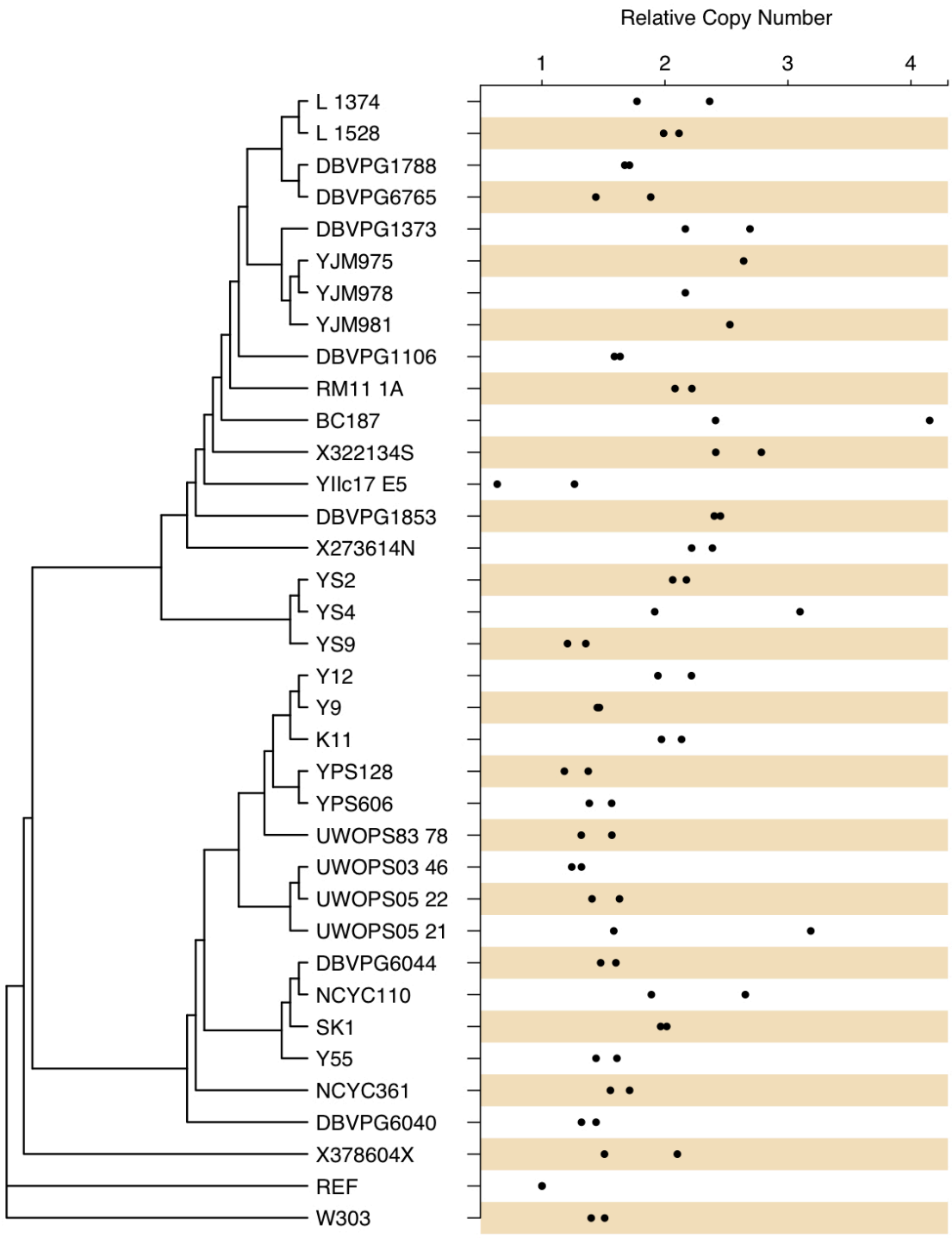

Figure 2.5. mtDNA copy number across 36 *S. cerevisiae* strains.

Measurements are normalized to BY4716, which is set to 1. For each strain, two biological replicates are plotted. The bottom panel depicts the genealogical relationship among strains using a Neighbor-joining tree.

We next performed a GWA study for mtDNA copy on 27,101 tagSNPs (see Methods) using a simple t-test. After correcting for multiple testing, we found seventy-three SNPs significant at $p < 0.05$ (Figure 2.6A, Table A.2). We next repeated the GWA analysis with EMMA and after correcting for multiple testing, we identified one significant ($p < 1.19 \times 10^{-5}$) SNP, a synonymous variant in *HPR1* on chromosome IV (Figure 2.6A, Table A.2). Strains with the C allele at this SNP had a mean copy number 0.5-fold higher than strains with the T allele (Figure 2.6B).

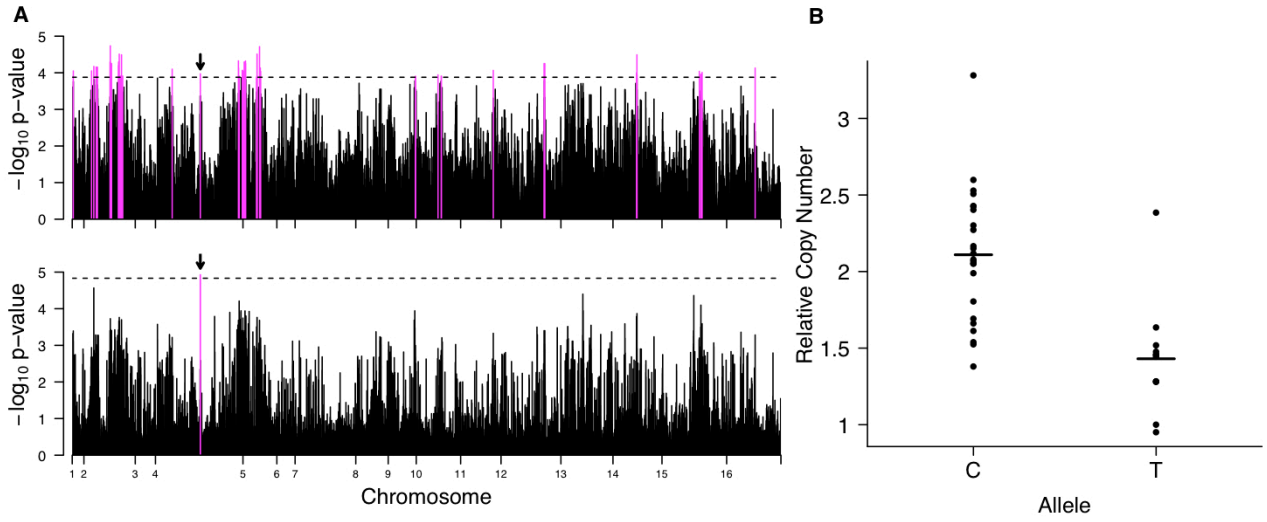

Figure 2.6. GWA study of mtDNA copy number.

A) The top panel shows the -log10 p-values obtained using a t-test and the bottom panel using the program EMMA. For each, the dotted line corresponds to $p < 0.05$, as determined by permutations. Chromosome numbers on the x-axis mark the first SNP marker on that chromosome. Those SNPs which passed the threshold for genome-wide significance are shown in magenta. The most highly-associated SNP identified by EMMA is labeled with an arrow in both panels. B) Distribution of copy numbers by allele at the most highly associated marker, which is a synonymous variant in the gene *HPR1*. Copy numbers are normalized to BY4716, as in Figure 3. The mean copy number is marked as a horizontal line for each allele.

HPR1 mediates mitotic recombination (Merker and Klein 2002) and recombination is thought to play a role in maintaining mtDNA and in proper segregation of mtDNA into newly-forming buds (Zelenaya-Troitskaya *et al.* 1998; Lecrenier *et al.* 2000; Ling and Shibata 2002; Lockshon *et al.* 1995). Thus, it is biologically plausible that *HPR1* is a novel regulator of

mtDNA copy number. However, the background distribution of *p*-values is high (Figure 2.6A), and our simulation results demonstrate the type I error rate is high. Moreover, the functional significance of the synonymous *HPR1* variant is unclear, and it is not in significant LD with adjacent polymorphisms (data not shown). Thus, these results highlight the difficulty in interpreting GWA results of quantitative traits in these strains.

2.3.4 Strategies for enabling GWA studies in yeast

Finally, we investigated sampling strategies that may facilitate GWA studies in yeast. Specifically, one possible approach to ameliorate the confounding effects of population stratification is to intelligently choose strains that have either low levels of structure or a consistent pattern of structure across the genome. To assess the type I error rate in such a genealogy, we used a subset of the SGRP *S. paradoxus* strains. As seen in Figure 2.6, the *S. paradoxus* strains contain a clade of 24 European strains with low levels of structure (Figure 2.7). We note that while the average pairwise divergence in the *S. paradoxus* European strains (θ_π = 0.0010) is somewhat lower than in the global *S. cerevisiae* population (θ_π= 0.0056), there is still a substantial amount of divergence (Liti *et al.* 2009). We used these strains to evaluate type I error rates in GWA studies as described above. We found that the type I error rate was only slightly elevated above nominal levels (Figure 2.7). To determine if the lower type I error rate was simply a function of limited power, we randomly sampled 24 *S. paradoxus* strains from the four divergent clades and repeated the GWA simulations. As shown in Figure 2.7, the type I error rate is significantly higher in the 24 randomly selected *S. paradoxus* strains, demonstrating that restricting association tests to a more homogenous subpopulation can decrease type I errors.

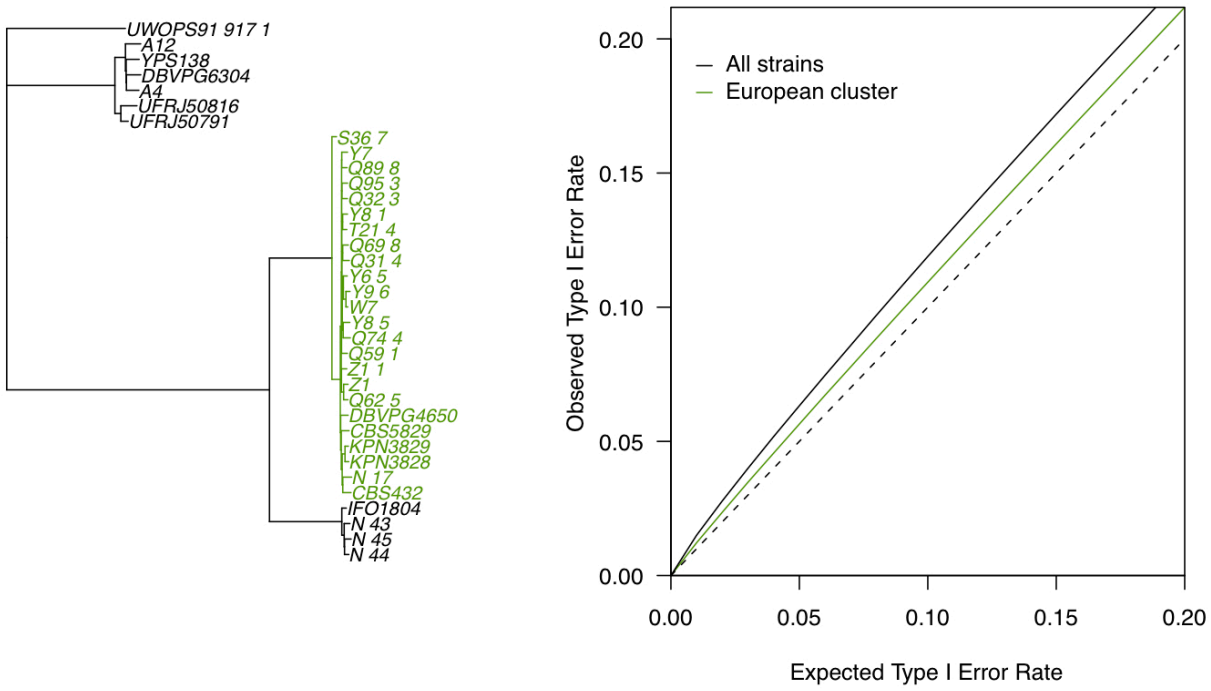

Figure 2.7. Type I error rates from simulations using *S. paradoxus*.

Left, a genealogy of the 34 *S. paradoxus* strains. Strains in green form the European cluster, used in the simulations as an example of strains with a consistent genome-wide genealogy. Right, the mean observed type I error rate from 100 simulations is plotted versus the expected type I error rate for association tests done using 24 *S. paradoxus* strains (green) and a random subset of 24 *S. paradoxus* strains (black). Association tests were done using EMMA. The ideal expectation is shown as a dashed line.

2.4 Discussion

Our findings have biological and methodological implications for further study. The full genome sequencing and SNP genotype data available in 86 strains is a powerful resource to test hypotheses about the population history, domestication, and evolution of *S. cerevisiae*. There is also considerable interest in using these strains to map QTL (Liti *et al.* 2009). However, our results suggest that GWA studies of quantitative traits may be challenging in the strains studied to date. Specifically, the complicated and heterogeneous patterns of structure across the yeast genome in the strains analyzed here result in high type I error rates, a problem that will only be exacerbated in analyses of high-dimensional molecular phenotypes, such as gene expression

levels. Several approaches could alleviate this problem, such as sequencing a greater number of strains and choosing a subset which that mitigates the problems of structure while maintaining a high level of variation. We have shown that strains with a simpler population structure show much lower levels of false positives. Alternatively, a deeper understanding of the empirical patterns of population structure across the *S. cerevisiae* genome in globally diverse strains will facilitate the development of new statistical models that can be uniquely tailored to ameliorating type I errors. Moreover, a strategy similar to the mouse collaborative cross (Threadgill *et al.* 2002) could be undertaken which would produce a large number of segregants that harbor substantial amounts of genetic diversity, but are free from the confounding effects of population structure.

Our results also provide important practical information on the design and interpretation of GWA studies in other model organisms, For example, EMMA has been used to perform GWA studies in dogs (Bokyo *et al.* 2010), *Arabidopsis* (Atwell *et al.* 2010), and *Caenorhabditis elegans* (Rockman and Kruglyak 2009). We suggest that GWA studies in model organisms with potentially strong and complicated patterns of structure carefully consider study design and sampling strategy, and empirically assess the effects of population structure on Type I error rates using the simple simulation framework described in our study. Fortunately, the decreasing costs of sequencing will allow individuals to be intelligently chosen to minimize the confounding effects of population structure and enable interpretable association studies in model organisms.

2.5 Materials and Methods

Sequence and SNP data

We obtained the publicly available genome sequences of the SGRP *S. cerevisiae* (n=36) and *S.*

paradoxus (n=36) strains (from ftp://ftp.sanger.ac.uk/pub/dmc/yeast/latest/; Liti *et al.* 2009). We also obtained SNP genotypes from a second set of 63 *S. cerevisiae* strains (Schacherer *et al.* 2009). We note that because the SNP data for the Schacherer *et al.* dataset was obtained using tiling micro-arrays with a resolution of 4 bp, it is not possible to combine the two datasets. Additionally, 13 *S. cerevisiae* strains are shared between the Liti *et al.* dataset and the Schacherer *et al.* dataset.

Simulations

For each dataset, we identified SNPs with a minor allele frequency (MAF) of ≥ 25%. We focused on common variants as the power to detect an association with more rare variation is limited given the relatively small number of strains. We then eliminated nearby SNPs in high LD using the program PLINK using a minimum r^2 of 0.8 (Purcell *et al.* 2007). Using this reduced set of SNPs, we randomly selected a "causal" SNP. We then generated quantitative phenotype data based on the allele of each strain, using a fixed effect equal to the standard deviation. For the lowest frequency variants (25%), this corresponds to a percent variance explained of 17%, using the formula: $p(1\text{-}p)k^2/(p(1\text{-}p)k^2 + 1 - 1/n) =\sim 1/(1 + 1/(p(1\text{-}p)k^2)$ where k is the fixed effect of 1.0 times the standard deviation, p is the frequency of the polymorphism with the fixed effect, and n is the number of individuals (in this case 36) (Yu *et al.* 2006). We examined several different sizes of fixed effects, and found that smaller effects were difficult to detect at all. Therefore, we chose to use a relatively strong fixed effect in order to maximize our ability to detect a significant association. We generated 1,000 simulated quantitative phenotype sets and tested for genome-wide association between the simulated phenotypes and all tagSNPs using 3 methods: a t-test, a t-test performed on the residuals of phenotypes after regressing out the first principle component derived from the SNP data (Price *et al.* 2006), and EMMA (Kang *et al.* 2008;

simulation programs and results are available on our website http://akeylab.gs.washington.edu/downloads.shtml). For EMMA, we used the entire set of tagSNPs to generate the *K* matrix. For Mendelian simulations, we used the same methods as above, but set the fixed effect equal to ten times the standard deviation, which corresponds to a percent variance explained of 95%. For the *cis*-based association simulations, for each gene we randomly picked a tagSNP within 1 kb up or downstream of each gene to be a "causal" variant. We then simulated a fixed effect using the alleles at this causal site, as for the quantitative trait simulations above. To assess the type I error rate at each gene, we then picked 1,000 random SNPs to test for association, and tested for association using EMMA, as above.

Yeast strains and culture

Two strains, *BY4716* and *RM11-1a*, were obtained from Leonid Kruglyak. The remaining thirty-four yeast strains were received from the Saccharomyces Genome Resequencing Project (Liti *et al.* 2009). These strains were confirmed as haploid through mating tests or made haploid, by integrating a *KanMX* cassette at the HO locus, sporulating the transformants, and isolating haploid spores. The remaining strains (DBVPG6044, YIIc17_E5, NCYC110, and Y9) are presumed to be diploids. Strains were grown in YEPD media to an OD of 0.8 – 1.0. Genomic DNA was isolated using the smash-n-grab method (Rose *et al.* 1990).

Measuring mtDNA copy number

PCR primers and probes were designed to two genes (*COX3* and *ATP6*) in the mitochondrial DNA and one gene in nuclear DNA (*GID8*). Primers for qPCR were designed by sequencing a large region of each gene in all strains and designing primers and TaqMan probes to regions where there were no polymorphisms between strains. We used three reporter dyes for the three probes: VIC for the *GID8* probe, FAM for the *COX3* probe, and TET for the *ATP6* probe. We

ran a standard curve for each probe to confirm that the efficiency of the probes was high. We found that the efficiency was approximately 1.0 for all three probes across a wide range of concentrations. We then ran two technical replicates for each biological replicate in 20ul reactions using ABI Master mix including ROX dye. qPCR reactions were run on an ABI3300 using ABI's standard 3300 protocol. We normalized the copy number in the mtDNA using the nuclear *GID8* probe, and calculated the relative mtDNA copy number in each strain by normalizing to the strain with the lowest copy number (BY4716). The mean copy number in the diploid strains was 1.55 versus the mean in all strains of 1.86 overall; however, this difference was not significant (t-test, p=0.14). We therefore included the diploid strains in our association test.

Association study of mtDNA copy number

To test for significant differences between strains, we fit a linear model to the qPCR data, where normalized copy was modeled as a function of biological replicate, technical replicate, mitochondrial probe, and strain. We tested for the significance of each variable using ANOVA and found that there was a significant ($p < 0.05$) effect for the biological replicate, probe, and strain effects, explaining 4.44%, 58.8%, and 28.5% of variance, respectively. We used the average of all technical and biological replicates for each strain to test for association. We then identified tagSNPs using the program HaploBlockFinder using a minimum LD (r^2) of 0.8 (Zhang and Lin 2003). We then tested for association between the mean mtDNA copy number and the 27,101 tagSNPs as above, using both a simple t-test and EMMA. We corrected for multiple testing by permuting the phenotype data 1,000 times, recording the lowest p-value from each genome-wide test, and used this distribution to determine genome-wide significance.

Neighbor-joining trees

Neighbor-joining trees were generated using the Neighbor program in PHYLIP (Felsenstein 1989). To assess confidence, we performed 100 bootstraps and built consensus trees using the program Consense in PHYLIP.

Resampling simulations in *S. paradoxus* We first randomly sampled 24 random strains out of 36, identified all SNPs with a MAF ≥ 25%, identified tagSNPs as above, and performed 1,000 simulations as above, using a fixed effect of 1.0 times the standard deviation. We next used the 24 strains in the European cluster as identified by Liti *et al.* (2009), identified SNPs with the same MAF cutoff (n=8822), and performed 1,000 simulations as described above.

Chapter 3

Population genomics and transcriptional consequences of regulatory motif variation in globally diverse *Saccharomyces cerevisiae* strains

This chapter has been published: Connelly CF, Skelly DA, Dunham MJ, Akey JM (2013) Population genomics and transcriptional consequences of regulatory motif variation in globally diverse Saccharomyces cerevisiae strains. Mol Biol Evol 30(7):1605-13.

3.1 Summary

Noncoding genetic variation is known to significantly influence gene expression levels in a growing number of specific cases; however, the patterns of genome-wide noncoding variation present within populations, the evolutionary forces acting on noncoding variants, and the relative effects of regulatory polymorphisms on transcript abundance are not well characterized. Here, we address these questions by analyzing patterns of regulatory variation in motifs for 177 DNA binding proteins in 37 strains of *S. cerevisiae*. Between *S. cerevisiae* strains, we found considerable polymorphism in regulatory motifs across strains (mean=0.005) as well as diversity in regulatory motifs (mean 0.91 motifs differences per regulatory region). Population genetics analyses reveal that motifs are under purifying selection, and there is considerable heterogeneity in the magnitude of selection across different motifs. Finally, we obtained RNA-Seq data in 22 strains and identified 49 polymorphic DNA sequence motifs in 30 distinct genes that are significantly associated with transcriptional differences between strains. In 22 of these genes, there was a single polymorphic motif associated with expression in the upstream region. Our results provide comprehensive insights into the evolutionary trajectory of regulatory variation in yeast and the characteristics of a compendium of regulatory alleles.

3.2 Introduction

Noncoding genetic variation makes a significant contribution to phenotypic diversity and disease susceptibility by modulating gene expression (Rockman and Kruglyak 2006, Skelly *et al.* 2009). Examples of noncoding variants causing phenotypic differences within and between species are rapidly accumulating in diverse lineages (Wray 2007). For example, noncoding variants have been identified that cause pigmentation differences in *Drosophila* (Wittkopp *et al.* 2002), skeletal reduction in stickleback fish (Shapiro *et al.* 2004), skin wrinkling in the domesticated dog (Akey *et al.* 2010, Olsson *et al.* 2011), and loss of neck feathers in chicken (Mou *et al.* 2011). Although the precise molecular mechanisms that causal noncoding variants act through remain poorly defined, many regulatory variants likely alter the binding of sequence-specific DNA binding proteins. These proteins affect gene expression by interacting with the transcriptional machinery, cooperatively binding to other activating or repressing proteins, or modulating chromatin structure (Lee and Young 2000, Farnham 2009).

Yeast is an excellent system in which to study noncoding variation because of the availability of whole-genome sequences from diverse strains and species. For example, whole-genome sequences are available for 37 *S. cerevisiae* strains, which are functionally and geographically diverse (Liti *et al.* 2009). In addition, sequence motifs for the majority of known DNA binding factors in yeast have been characterized (Bryne *et al.* 2008). Motif usage across species has been studied extensively in yeast. Previous work on the evolution of noncoding regions has shown that motifs rapidly turn over between species, including yeast (Dermitzakis and Clark 2002, Moses *et al.* 2006, Borneman *et al.* 2007, Doniger and Fay 2007). In some cases, genes whose coexpression has been conserved across species may have acquired different

regulators in different species, as in the case of ribosomal protein modules in yeast (Wapinski *et al.* 2010). Despite their frequent turnover, often the presence of specific motifs is conserved. For example, Doniger and Fay found that 55% of genome-wide binding sites fit a model of conservation when looking across four yeast species (2007). Motifs that are conserved within and between species are correlated with several characteristics, such as being upstream of essential genes, closer to transcription initiation sites, and within open chromatin regions (Francesconi *et al.* 2011).

More generally, previous analyses of noncoding regions in diverse species have found strong signatures of both positive and negative selection (Mustonen and Lassig 2005, Chen *et al.* 2010, He *et al.* 2011). These studies have had several limitations, however; for example, they have focused on small collections of known binding sites (Mustonen and Lassig 2005), motifs involved in key developmental modules (He *et al.* 2011), or motifs ascertained based on their conservation (Chen *et al.* 2010).

Gene expression variation has also been studied extensively in yeast. These studies have revealed that specific classes of genes are more likely to diverge between species (Thompson and Regev 2009), and such loci share architectural features such as containing a TATA box in their promoter and harboring more binding sites for regulatory proteins (Tirosh *et al.* 2006). In addition, expression QTL (eQTL) studies have identified a significant role for *cis*-acting variation in gene expression differences between strains or species (Brem *et al.* 2002, Ronald and Akey 2007, Ehrenreich *et al.* 2009, Tirosh *et al.* 2009, Emerson *et al.* 2010). Differences in predicted motifs have been associated with expression differences for some of the genes with *cis*-linkages in a cross between two strains (Chen *et al.* 2010). In addition, Zheng *et al.* identified several hundred genes showing significant gene expression variation associated with differences

in protein binding for the factor *STE12* (Zheng *et al.* 2010). Thus, variation in DNA- binding motifs can be an important causal source of gene expression variation.

In this study, we describe a comprehensive genome-wide analysis of polymorphisms located in 177 DNA sequence motifs across 37 *S. cerevisiae* strains (Liti *et al.* 2009). We expand the number of motifs studied from previous studies, and identify motifs genome-wide in an unbiased manner without regard to conservation. We perform extensive population genomics analyses that reveal DNA sequence motifs are subject to purifying selection, and quantify the strength of selection for each motif. Furthermore, we used RNA-Seq data that was previously collected for 22 of these strains and performed association analyses between polymorphisms in motifs and differences in gene expression. We identified six polymorphic motifs associated with widespread and consistent changes in gene expression, 49 polymorphic motifs associated with transcriptional variation at individual genes, and a compendium of high confidence regulatory alleles.

3.3. Results

3.3.1 Regulatory motif variation across S. cerevisiae strains

We first examined patterns of motif differences across 37 globally and functionally diverse *S. cerevisiae* strains whose genomes have been sequenced (Liti *et al.* 2009, see Figure S1), by independently calling motifs in all strains (see Methods). We found substantial divergence in motif content across strains. The average pairwise number of motif differences per intergenic region is 0.91 motifs (range 0-27, see Figure 3.1A), and as expected pairwise motif differences recapitulate the known phylogeny (data not shown). Across all strains, a median of eight motifs were called in each intergenic region, and a median of four motifs per intergenic region were

variable in at least one of the 37 strains (range 0-137). One example of a highly divergent region is the region upstream of *AAH1*, an adenine deaminase, which is regulated by nutrient levels (Figure 3.1B). Another highly variable region is upstream of *FLO1*, which is involved in flocculation, a phenotype known to have diverged between laboratory and wild strains (Liu *et al.* 1996). Interestingly, a cluster containing both lab and wild strains shows a divergent motif pattern in this region (Figure 3.1C). A list of additional genes with highly polymorphic motif patterns is provided in Table B.1.

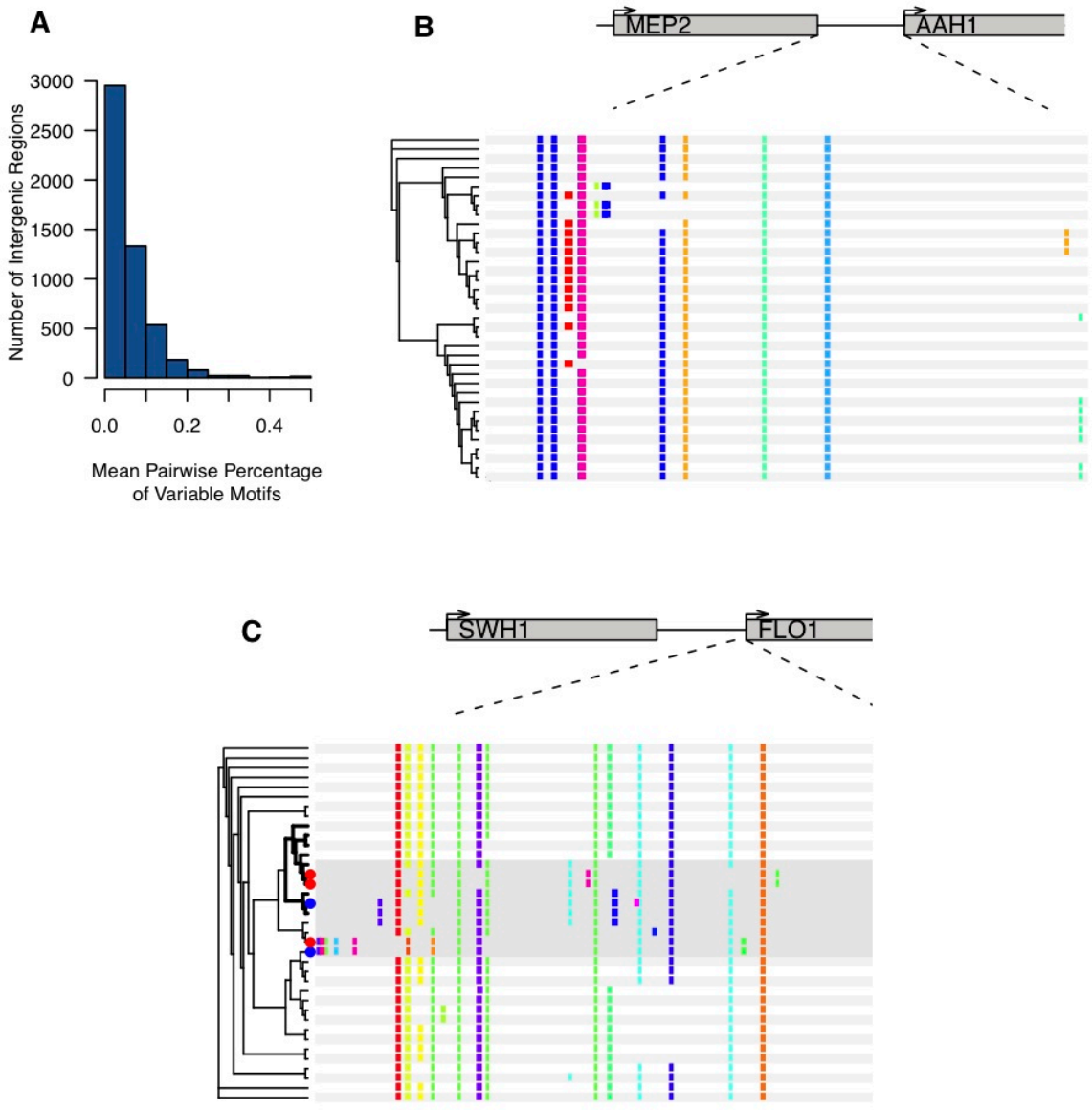

Figure 3.1. Examples of highly divergent regulatory regions across *S. cerevisiae* strains.

A. Histogram of the mean pairwise percentage of variables motifs across 37 *S. cerevisiae* strains for each of the 5,468 intergenic regions. B. Predicted motif calls for 37 *S. cerevisiae* strains are plotted for the intergenic region upstream of the gene *AAH1*. Each row is a strain, and colored boxes represent motif calls. Different colors represent distinct motifs. A phylogeny for the strains is shown to the left, as constructed from the motif calls for that region. C. Predicted motif calls for a section of the intergenic region upstream of the gene *FLO1*. The region shown represents 1000 bp upstream of the gene, out of 7218 total upstream bases. A divergent clade is highlighted in grey, and within this clade strains wild strains are marked with a red dot in the phylogeny, while laboratory strains are marked with a blue dot.

3.3.2 Evolutionary forces shaping patterns of polymorphism and divergence of regulatory sequences

To quantify the strength of selection acting on intergenic regions more systematically, we used the McDonald Kreitman framework to assess deviations from neutral expectations across intergenic regions (McDonald and Kreitman 1991). For measures of divergence, we used *S. paradoxus* as an outgroup. We initially characterized the evolutionary forces acting at four classes of sites: nonsynonymous, noncoding sites within predicted motifs, noncoding sites outside predicted motifs, and experimentally determined motifs (MacIsaac *et al.* 2006; see Methods). Specifically, we counted polymorphic and diverged sites across all intergenic and genic regions that could be aligned between *S. cerevisiae* and *S. paradoxus* (approximately 4,700 regions). As putatively neutral sites, we used synonymous sites. We found that purifying selection acts on all four classes of sites ($p<2.2 \times 10^{-16}$). We next estimated the –log(Neutrality Index), denoted as -log(NI) (Rand and Kann 1996), in order to compare the magnitude of purifying selection across site types. A value for –log(NI) of zero is consistent with neutrality, negative values suggest negative selection, and positive values indicate positive selection. As expected, the -log(NI) was lowest for experimentally determined motifs, which appear to be under strong purifying selection. We also found that –log(NI) was lower at noncoding sites inside predicted motifs compared to noncoding sites outside of motifs and nonsynonymous sites, suggesting that a higher proportion of sites falling within predicted motifs are under purifying selection than in the other classes of sites (Figure 3.2A). The observation that -log(NI) at noncoding sites outside predicted motifs was similar to that at nonsynonymous sites is unexpected because noncoding sites outside motifs are generally thought to be subject to less functional constraint. However, this result may be due in part to the high threshold we used to

call motifs; lowering the threshold resulted in the -log(NI) at noncoding sites outside motifs becoming closer to neutral expectations (Figure B.2).

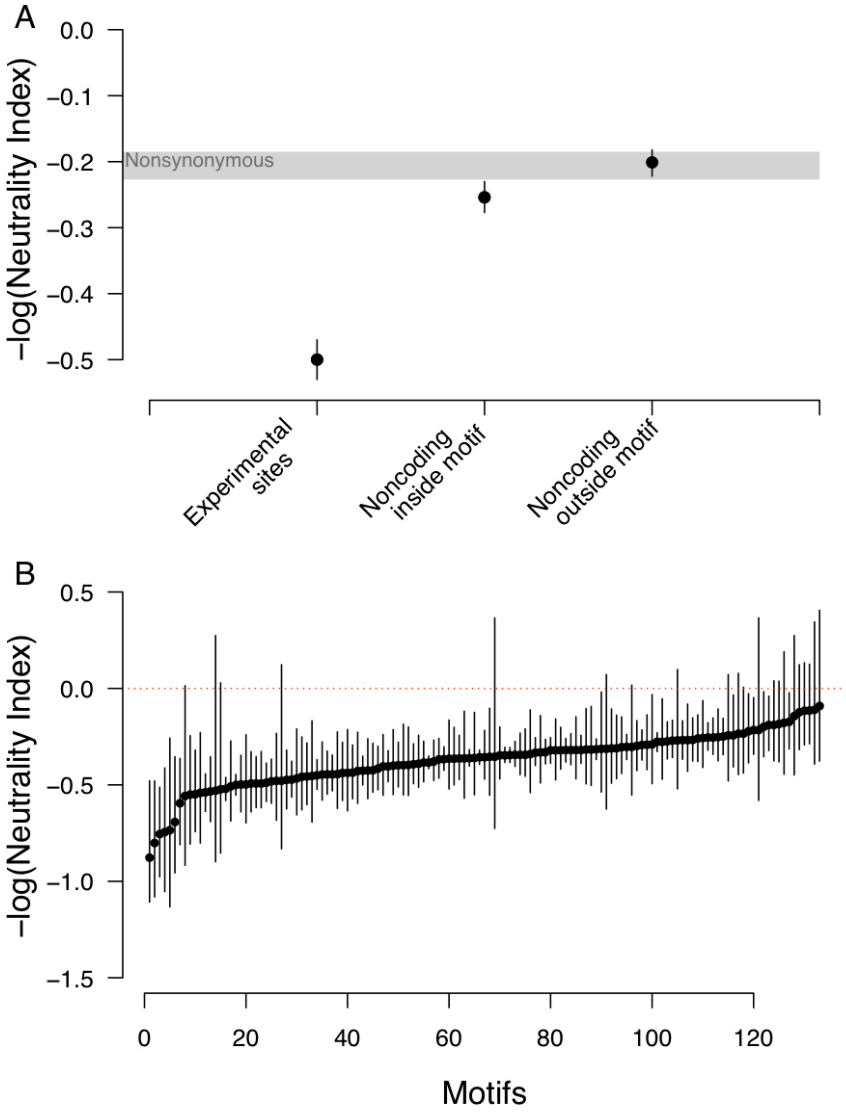

Figure 3.2. Evolutionary forces acting at intergenic regions.

A. –log(Neutrality Index) scores for two classes of sites (noncoding sites falling within predicted motifs and noncoding sites falling outside predicted motifs) are plotted. –log(NI) scores were obtained by summing information across all sites of a particular class and using synonymous sites within genes as putatively neutral sites. Confidence intervals were obtained by bootstrapping (see Methods). For noncoding sites, three different cutoffs were used for calling motifs. The most stringent cutoff results are colored black, a less stringent cutoff is colored blue, and the most stringent cutoff is colored red (see Methods). 95% CI for nonsynonymous sites are shown as in grey. B. –log(NI) scores for each of 133 motifs, sorted from lowest –log(NI) to highest –log(NI). –log(NI) scores were obtained by summing information across all sites genome-wide falling within a particular motif, and comparing to all synonymous sites. Motifs with low numbers of polymorphic and divergent sites were excluded due to low power to detect differences with such low counts (less than 15 total sites). Confidence intervals were obtained by bootstrapping (see Methods).

To identify heterogeneity of selective constraint across DNA binding motifs, we calculated a motif specific estimate of the -log(NI). As shown in Figure 3.2B, selective constraint varies widely across motifs, with some motifs under very strong purifying selection. Out of 133 motifs with sufficient data (see Methods), we identified 112 whose –log(NI) was significantly less than zero (Table B.2). As expected from the above analysis, a sizable number of motifs (63) had a –log(NI) significantly lower than that at nonsynonymous sites.

Moreover, we examined constraint acting at the level of individual intergenic regions. To this end, we compared polymorphism and divergence at sites that fell within predicted motifs in each region to synonymous sites in the genes flanking each region. We found that many intergenic regions had negative –log(NI), as expected from the motif-specific results described above, although the power of this analysis is lower given the reduced number of polymorphisms and divergent sites within each region. Using the MK test, we identified 152 regions that have significant evidence for purifying selection at FDR=0.10. 11 of these regions were significant after a more stringent Bonferroni correction for multiple testing (Table B.3). We did not find any regions significant for positive selection at FDR=0.10 or after a Bonferroni correction; however, four regions had a suggestive p-value ($p \leq 0.05$, uncorrected). Three of these regions flanked genes of unknown function; the remaining region flanks *ADH4*, an alcohol dehydrogenase gene, which has been linked to increased ethanol production (Mizuno *et al.* 2006). Interestingly, many *S. cerevisiae* strains were domesticated for use in fermentation, and thus positive selection for changes in the regulation of *ADH4* may have occurred between *S. cerevisiae* and *S. paradoxus* which made *S. cerevisiae* favorable for use in domestication.

3.3.3 Patterns of motif polymorphism are significantly correlated with transcriptional variation among strains

To assess the relationship between motif and gene expression variation, we obtained RNA-Seq data that had been collected on a subset of the 22 strains of *S. cerevisiae* analyzed above (Skelly *et al.* 2012). We performed extensive normalization of the data to account for batch effects and unknown sources of variation (see Methods). By analyzing the complete distribution of p-values using the positive false discovery rate approach of Storey and Tibshirani (Storey and Tibshirani 2003), we estimate that 79.0% of genes are differentially expressed across the 22 strains. Of these, 5,472 genes are significantly differentially expressed at a FDR = 0.10.

We investigated the relationship between motif polymorphism and transcriptional variation using two complimentary approaches. First, we tested for associations between the presence or absence of motifs and expression levels at downstream genes. Specifically, we performed association tests correcting for population structure for 13,089 motifs located upstream of 3,505 distinct genes (Connelly and Akey 2012). We note that with a small sample size of 22 strains, we have limited power to detect variants, except those with large effect sizes (Table B.4). We found 49 polymorphic motifs located upstream of 30 distinct genes that were correlated with significant changes in gene expression (FDR = 0.10). Of the 49 associated polymorphic motifs, 21 resulted in increased expression with the presence of the motif (i.e. acted as an activator) and 28 resulted in decreased expression with the presence of the motif. Interestingly, 22 of these genes contained only a single polymorphic motif associated with expression variation in the upstream region (Table 3.1). In addition, one gene did not contain any additional promoter variants located outside of motifs that are in strong linkage disequilibrium ($r^2 > 0.8$) with the polymorphic motif (Figure 3.3). Moreover, we found evidence for one case of a bi-directional promoter, where

polymorphism in the REB1 motif was associated with changes in expression of both flanking genes. Thus, the statistical and bioinformatics data strongly suggest that these 22 polymorphic motifs are enriched for causal regulatory polymorphisms.

Table 3.1. High confidence regulatory polymorphisms

Motif	Downstream Gene	Log(Difference in Expression)	Distance Upstream of Gene	q Value
HCM1	*YJL155C*	-0.26	300	0.04
HCM1	*YEL044W*	-0.27	955	0.04
MOT3	*YKL059C*	0.29	509	0.04
PHO2	*YJR108W*	0.57	144	0.04
PHO2	*YGL169W*	-0.32	381	0.04
REB1	*YNL239W*	0.47	239	0.04
SPT2	*YEL001C*	0.24	675	0.04
YAP5	*YOR108W*	-0.59	436	0.04
CRZ1	*YAL049C*	0.39	146	0.06
HAL9	*YPL255W*	-0.32	494	0.06
HAP2	*YNR049C*	0.27	28	0.06
HAP2	*YOR071C*	-0.44	2828	0.06
PHO2	*YPR119W*	-0.12	331	0.06
RAP1	*YPL108W*	0.48	407	0.06
REB1	*YNL240C*	0.59	352	0.06
STE12	*YKL108W*	-0.28	91	0.06
FHL1	*YJL094C*	-0.77	148	0.07
HAP2	*YBR222C*	0.39	245	0.07
HAP2	*YGL117W*	-0.71	1245	0.07
MOT3	*YLR152C*	-0.67	242	0.07
ABF2	*YPL167C*	-0.24	116	0.09
YAP3	*YLR007W*	-0.16	366	0.09

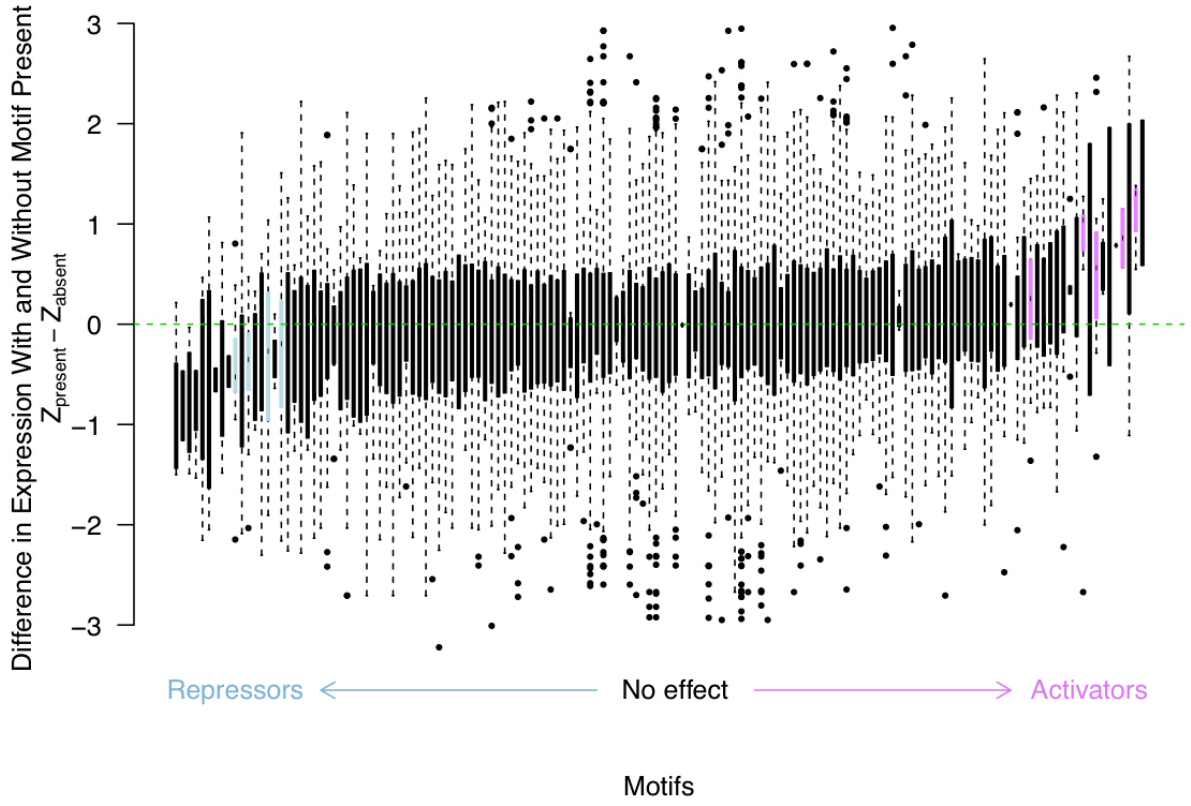

Figure 3.3. Effects of variants at specific motifs on gene expression.

For each motif, a boxplot of the difference in expression Z scores between strains containing the motif and strains not containing the motif at all genes with a variable motif are plotted. Motifs are sorted by mean difference in expression. Motifs significant in our test for genome-wide differences in expression for showing lower expression when the motif is present are colored in blue, and motifs significant for showing greater expression when the motif is present are colored in magenta.

In addition, we tested the hypotheses that levels of sequence conservation varied between polymorphic motifs that were or were not associated with differences in gene expression. Using phastCons scores from the 12-species yeast alignment (Siepel *et al.* 2005), we compared the mean conservation score at 1,039 polymorphic motifs nominally associated with expression differences among strains ($p \leq 0.05$) to a null distribution constructed by drawing the same number of randomly chosen motifs not associated with gene expression differences. We found conservation was significantly higher at motifs associated with expression differences ($p=0.024$).

Secondly, we tested whether motifs were acting consistently as activators or repressors across a majority of genes upstream of which they were polymorphic. Specifically, for the i_{th} motif, we identified all genes whose upstream intergenic region contained a variable motif *i*. We discarded genes where the variable motif was only observed in a single strain. Next, we converted gene expression values for this set of genes to a Z score and tested for differences between the distribution of expression values when motif *i* was present or absent (see Methods). At a FDR = 0.10, we found that polymorphisms in 9 out of the 148 motifs were significantly associated with consistent transcriptional differences (5 motifs were significantly associated with increased expression and 4 motifs were significantly associated with decreased expression; Table 3.2 and Figure 3.4).

Table 3.2. Motifs associated with consistent expression differences.

Motif	Number Genes Containing Upstream Motif	Number Polymorphic Motifs	Average Effect Size (sd)[a]
GAL4	5	2	0.86
SUT2	17	3	1.08
LEU3	48	7	0.47
RGT1	75	14	-0.43
UGA3	87	19	-0.40
MET4	88	28	-0.32
SWI4	104	17	0.58
RDR1	106	25	0.25
TOS8	265	70	-0.28

[a] Measured as the average difference in expression Z-scores between motif presence and absence averaged across genes.

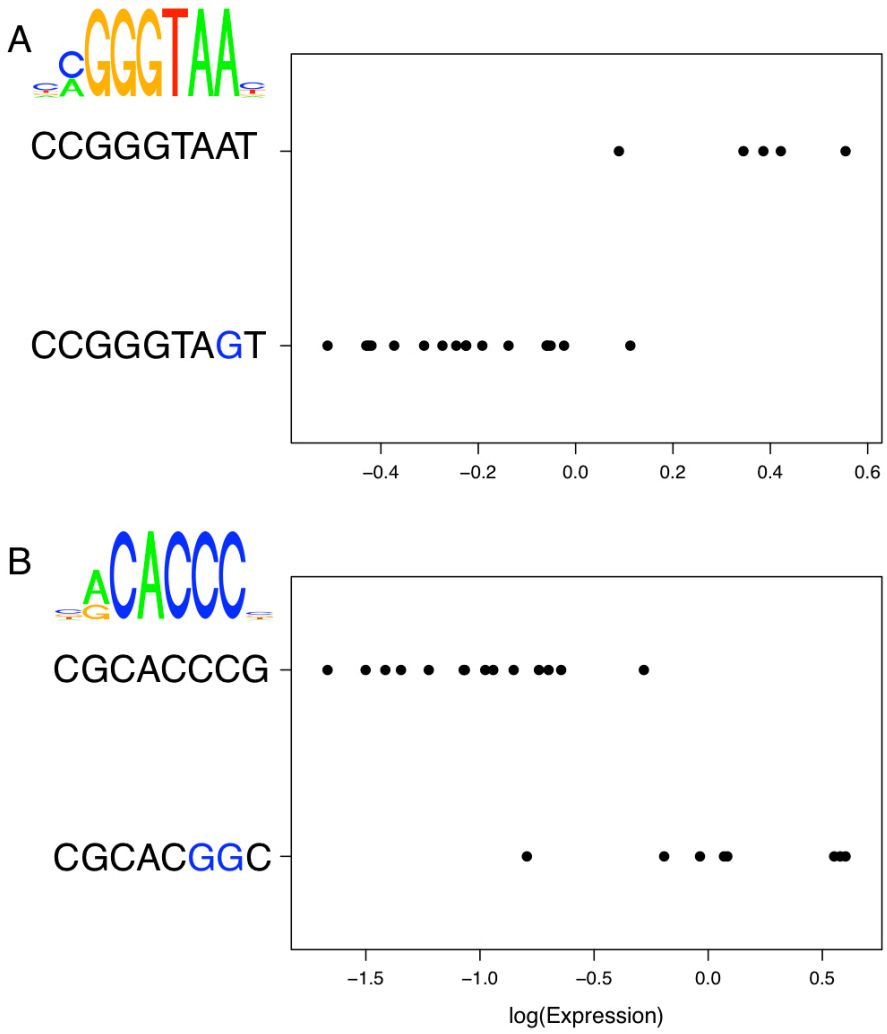

Figure 3.4. Examples of motifs effecting gene expression.

A. *NAR1* expression in strains containing the two labeled sequences at the motif REB1 in the upstream intergenic region. Substitutions to the consensus motif sequence for REB1 are marked in blue. A sequence logo for REB1 representing the PSSM is shown in the upper left corner. B. *YER186C* expression in strains containing the two labeled sequences at the motif AFT2 in the upstream intergenic regions. Substitutions to the consensus motif sequence are marked in blue. A sequence logo for AFT2 representing the PSSM is shown to the upper left of the plot.

3.3.4 Features associated with transcriptional divergence

Finally, we investigated what characteristics were associated with high expression divergence. As a measure of expression divergence between strains, we calculated the average pairwise difference in expression between strains. We first tested whether the absolute value of

the –log(NI) for motifs in each region was associated with expression divergence. We used the absolute value so that any region under either positive or negative selection would have a value greater than zero and regions under no selection would be closer to zero. We found a negative correlation (ρ = -0.07, p=2.43 x 10^{-6}, Spearman rank-sum test), demonstrating that regions under stronger selection showed less expression divergence. We also tested whether nucleotide diversity (π) within motifs was associated with expression divergence. We found a positive correlation (ρ = 0.10, p=6.84 x 10^{-14}, Spearman rank-sum test), illustrating that higher nucleotide diversity was associated with higher expression divergence between strains. This correlation was still significant after controlling for the presence or absence of TATA box and for the nucleosome occupancy upstream of each gene (see Methods).

3.4 Discussion

Interpreting noncoding variation is challenging yet vital for identifying causal regulatory variation, delimiting the contribution of expression variation to phenotypic diversity and evolutionary diversification, and elucidating the molecular mechanisms through which noncoding variation acts. By focusing on interpretable noncoding variation, namely variants within known motifs for DNA binding proteins, we were able to perform detailed evolutionary and statistical analyses on the evolutionary pressures acting at these motifs and the functional consequences of putative regulatory variation.

We first addressed the evolutionary pressures affecting motif diversity and divergence, and found that motifs are generally subject to purifying selection. These results are broadly consistent with previous analyses demonstrating purifying selection acting on yeast promoter and 3' UTR regions (Mustonen and Lassig 2005, Ronald and Akey 2007, Chen *et al.* 2010).

Similarly, studies in humans have found decreased nucleotide diversity in open chromatin regions (Thurman *et al.* 2012, Vernot *et al.* 2012) and have correlated transcription factor occupied sites with higher conservation across multiple species (Neph *et al.* 2012). Similarly, we found that experimentally validated sites were subject to stronger purifying selection. Interestingly, we found that the level of purifying selection acting on all predicted motifs was still quite strong. We also found that the selection on experimentally determined sites and on predicted sites was stronger than that on nonsynonymous variants. One possible explanation for this is that a smaller proportion of nonsynonymous sites will actually affect gene function compared to the proportion needed to disrupt a motif. It is also possible that by testing only motifs which can be aligned between the two species, we may be biased towards detecting conserved motifs. In comparison to other species, it is interesting that similar studies in *Drosophila* have found widespread evidence of adaptive evolution in noncoding regions (Andolfatto 2005) whereas we found little evidence for adaptive evolution. We speculate that these differences in the tempo and mode of noncoding evolution between species may be due, at least in part, to differences in effective population size. We also found that while a majority of motifs are under purifying selection, a subset are evolving neutrally. This may suggest that the position weight matrices for these motifs are ineffective at identifying functional binding sites or that, alternatively, these motifs are in general less constrained.

To investigate the effects of motif changes on transcriptional variation, we characterized gene expression differences among 22 strains. We identified six motifs (*MET4, RGT1, SUT2, SWI4, TOS8,* and *UGA3*) acting consistently as activators or repressors across a majority of genes they regulated. These transcription factors are involved in diverse processes, but it appears that they are broadly active as activators or repressors in phosphate-limiting conditions. We also

identified 30 genes where one or more motifs were associated with gene expression variation. Approximately one third of these genes contained multiple motifs associated with expression variance at that gene, making it difficult to identify the causal variant, though it is also possible that there may be multiple motif changes contributing to gene expression differences at these loci, as observed previously (Prud'homme *et al.* 2006, Tao *et al.* 2006). In addition, we were able to identify 22 genes with only one motif associated with expression differences. Although for all but one of these there were other SNPs in strong LD with the associated motif in the intergenic region, SNPs that fall within motifs are a strong strong candidate for being a causal SNP because of their potential functional role.

We found that conservation scores across species were significantly higher at motifs associated with expression differences than at motifs not associated with expression differences, suggesting that cross-species conservation is useful for fine-scale mapping causal regulatory variation. In addition, measures of constraint within species that combine information across multiple motifs in a region were useful for predicting more general patterns of expression divergence. Specifically, we found that regions with less constrained motifs as measured by the -log(NI) and nucleotide divergence were more likely to have higher expression divergence, although the magnitude of the correlation was modest.

There are several limitations to our study design. Because we are using computationally predicted motifs, not all are actually used *in vivo*; however, by using stringent cutoffs for calling motifs (see Methods) we attempted to collate a high confidence set of predicted motifs on which to perform our analyses. The evolutionary analyses also suggest that we are identifying active sites that are under constraint. In addition, our study only tests the effects of motif variation on gene expression in one experimental condition. Finally, since our sample size was small, we are

underpowered to detect associations attributable to rare variants or variants with small effect sizes (Table S4).

In summary, our approach demonstrates the utility of using motif predictions in conjunction with functional genomics data for identifying functional noncoding sequence variation and DNA binding proteins that have significant effects on gene expression. In the future, it will be important to integrate additional types of data, such as *in vivo* DNA binding protein information and ChIP-Seq data, to facilitate the interpretation of noncoding variation, the identification of causal noncoding variants, and the correlation of transcriptional variation to phenotypic diversity. Such integrative genomics analyses are likely to play a key role in ultimately developing predictive models to distinguish functionally important noncoding variation from functionally and phenotypically benign variants.

3.5 Materials and Methods

Sequence data and alignments

We obtained sequence data and whole genome alignments for 37 *S. cerevisiae* strains and the *S. paradoxus* reference sequence (CBS432-0809) from the *Saccharomyces* Genome Resequencing Project (Liti *et al.* 2009). The alignment between *S. cerevisiae* and *S. paradoxus* was done by repeating masking the reference *S. cerevisiae* genome and *CBS432*. The programs LASTZ (Harris 2007) and TBA (Blanchette 2004) were used to construct the alignment. Substitution scoring parameters for LASTZ alignments were inferred using two *S. cerevisiae* strains (the reference strain and *RM11_1A*). For all further analyses, we excluded intergenic regions that aligned to more than one contiguous block in *S. cerevisiae*.

Motif Analysis

We searched all intergenic regions for the 37 strains and *S. paradoxus* for each of the 177 known DNA binding motifs on both strands using Position-specific site matrices obtained from JASPAR and converted to PWMs (Bryne *et al.* 2008). Note that these matrices come from experimental studies and are not ascertained based on conservation across species. In all further analyses, we did not include sites with missing data in 1 or more of the strains, or sites that were called due to indels to mitigate alignment errors. Motifs were called if they had 90% of the observed maximum weight matrix score.

For the experimentally determined sites, we used binding sites identified by ChIP-chip (MacIsaac *et al.* 2006) that were significant at $p<0.001$ and not subject to conservation criteria. This list consists of 9708 motif sites.

MK Test measurements

We calculated the neutrality index (NI) as: $NI = \frac{D_n P_s}{D_s P_n}$ (Rand and Kann 1996). Here, *D* is the count of polymorphic sites between *S. cerevisiae* and *S. paradoxus*, and *P* is the count of polymorphic sites (frequency greater than 5 percent) between the *S. cerevisiae* strains, *n* = neutral sites (synonymous sites), and *s* = putative selected sites. When calculating the neutrality index for each intergenic region, we used the synonymous sites from immediately flanking genes. For bootstrapping, we resampled 1000 times from the data for each intergenic region.

RNA mapping and normalization

Raw RNA reads were obtained from Skelly *et al.* (2012). We mapped RNA-Seq reads to the S288c reference genome (UCSC sacCer2) using the program BFAST version 0.6.4e (Homer *et al.* 2009) with options –K 100 and –M 500 to bfast match. We aligned colorspace reads using a main index with mask 1111111111111111 (hash width 14) and secondary indexes with masks 111110111011101010010101101111, 101111010111010010110000110100011111111, and

10111001101001100100111101010001011111 (all using hash width 14). We output the results in SAM format and converted to BAM format using samtools (Li *et al.* 2009). We computed read depth across genes using bedtools version 2.15.0 (Quinlan and Hall 2010).

We normalized counts for each gene by the number of total read counts for that strain. We then carried out a median normalization step to normalize across flow cells (Pickrell *et al.* 2010). After this step, we removed any genes that had no counts across any strain. Finally, we fit a linear model of the form log(normalized_counts) ~ batch + flow cell + strain + significant surrogate variables. We used the R package sva to calculate surrogate variables, which revealed four significant surrogate variables (Leek and Storey 2007). Further tests used the residuals from this model.

Assessing differential expression across strains

We used a random effects model to test for a strain effect using the R package lme4, using the Maximum Likelihood method and calculating p-values using the Chi-square distribution (R Development Core Team 2011). We assigned q-values by permuting the strain assignments 1000 times and repeating the analysis, calculating the empirical p-value from this distribution, then using the R package qvalue to assign q values (Storey and Tibshirani 2003).

Testing for motif effects on expression across all genes

For each gene, we converted the normalized expression values to Z scores. For each motif, we then identified genes that had a variable motif upstream. We tested for differential expression by combining expression Z scores from all strains with the motif present, and compared them to Z scores from strains with the motif absent, combining these Z scores across the genes identified above. We tested for differential expression using a t-test, and determined q values by permuting the labels of present/absent for each gene 1000 times.

Testing for motif effects on expression at one gene

For each gene with a variable motif, we tested the hypothesis that there was a difference in expression between strains with the motif present and strains with the motif absent, using the program EMMA to control for population structure (Kang *et al.* 2008). P-values were again assigned by permuting the motif presence/absence labels 1000 times, and calculating q values as above.

Simulations

The simulations were done as previously described (Connelly *et al.* 2010). Briefly, we chose 1000 random SNPs which fell within genes or 1000 bp up- or downstream of genes and which had a minor allele frequency of at least 3 out of 22 as causal SNPs and simulated data based on the genotype at each SNP. We generated simulated data of three effect sizes, 25 percent of variance in phenotype explained by the genotype, 50 percent of variance explained by the genotype, and 75 percent of the variance explained by the genotype. This was equal to a fixed effect of k=1.64, 2.85, and 4.885 times the standard deviation, respectively, solving for k using the formula percent variance explained = $p(1 - p)k^2/(p(1 - p)k^2 + 1 - 1/n) =\sim 1/(1 + 1/(p(1 -p)k^2)$ where k is the fixed effect of x times the standard deviation, p is the frequency of the polymorphism with the fixed effect, and n is the number of individuals (Yu *et al.* 2006). To assess power, we tested for association between the simulated data and the genotype at the causal variant for each of the 1000 simulations using EMMA (Kang *et al.* 2008). To assess the type I error rate, we chose 1000 random SNPs and asked how often they showed association with any of the 1000 simulated datasets.

Motif conservation

We obtained phastCons scores from the UCSC genome browser for each position in the S288c genome (Siepel *et al.* 2005). We used the p-values from the gene-specific test above to identify motifs nominally associated with expression differences among strains ($p \leq 0.05$, n=1039). To assess significance of polymorphic motif conservation scores, we generated a null distribution by calculating mean conservation from 1000 randomly selected motifs that are not associated with expression differences ($p > 0.05$).

Nucleotide diversity within motifs and expression divergence

We obtained calls for the presence or absence of a TATA box upstream of each gene (Tirosh *et al.* 2006). For a measurement of nucleosome occupancy, we used the genome-wide nucleosome occupancy data (Lee *et al.* 2007), and calculated nucleosome occupancy 100 bp upstream of transcription start sites (Zhang *et al.* 2005), a similar approach to that taken by Tirosh and Barkai (2008). We used a linear model to test for the effect of nucleosome occupancy, TATA box presence, and nucleotide diversity within motifs on expression divergence.

Chapter 4

Evolution and genetic architecture of chromatin accessibility and function in yeast

This chapter has been published: Connelly CF, Wakefield J, Akey JM (2014) Evolution and genetic architecture of chromatin accessibility and function in yeast. PLoS Genet 10:e1004427.

4.1 Summary

Chromatin accessibility is an important functional genomics phenotype that influences transcription factor binding and gene expression. Genome-scale technologies allow chromatin accessibility to be mapped with high-resolution, facilitating detailed analyses into the genetic architecture and evolution of chromatin structure within and between species. We performed Formaldehyde-Assisted Isolation of Regulatory Elements sequencing (FAIRE-Seq) to map chromatin accessibility in two parental haploid yeast species, *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* and their diploid hybrid. We show that although broad-scale characteristics of the chromatin landscape are well conserved between these species, accessibility is significantly different for 947 regions upstream of genes that are enriched for GO terms such as intracellular transport and protein localization exhibit. We also develop new statistical methods to investigate the genetic architecture of variation in chromatin accessibility between species, and find that *cis* effects are more common and of greater magnitude than *trans* effects. Interestingly, we find that *cis* and *trans* effects at individual genes are often negatively correlated, suggesting widespread compensatory evolution to stabilize levels of chromatin accessibility. Finally, we demonstrate that the relationship between chromatin accessibility and

gene expression levels is complex, and a significant proportion of differences in chromatin accessibility might be functionally benign.

4.2 Introduction

Changes in gene regulation have long been hypothesized to be an important mechanism of evolutionary diversification (Britten and Davidson 1971, King and Wilson 1975, Wray 2007) and to contribute to phenotypic variation (Shapiro *et al.* 2004, Akey *et al.* 2010, Mou *et al.* 2011, Skelly *et al.* 2009). An increasing catalog of adaptive regulatory changes has been identified, such as lactase persistence (Enattah *et al.* 2002, Tishkoff *et al.* 2007) and the effect of the Duffy blood group chemokine receptor on malaria resistance in humans (Tournamille *et al.* 1995, Hamblin *et al.* 2000) and beak morphology in Darwin's finches (Abzhanov *et al.* 2004). Furthermore, it has also been suggested that a substantial fraction of SNPs associated with human diseases through genome-wide association studies may act through regulatory changes with genes (Visel *et al.* 2009, Maurano *et al.* 2012).

On a genome-wide scale, molecular studies have uncovered pervasive transcriptional variation within and between species (Primig *et al.* 2000, Sandberg *et al.* 2000, Brem *et al.* 2002, Khaitovich *et al.* 2005, Pickrell *et al.* 2010, Tsankov *et al.* 2010). A substantial amount of gene expression variation is heritable, and thousands of regulatory QTL have been mapped in numerous organisms (Brem *et al.* 2002, Wittkopp *et al.* 2004, Tirosh *et al.* 2009, Emerson *et al.* 2010, Skelly *et al.* 2011). In general, regulatory variation can act in *cis* or *trans*. *Cis*-acting regulatory QTL influence transcript levels in an allele-specific manner, typically from variation located within or near the gene being studied. In contrast, *trans*-acting regulatory QTL does not result in allelic differences in expression and arises from variation that is usually located at a

position distinct from the gene being studied (Skelly *et al.* 2009). Although both *cis* and *trans* regulatory variation make important contributions to heritable variation of transcript abundance, *cis*-acting variants are thought to be more numerous, have larger effect sizes, and accumulate at a faster rate between species (Wittkopp *et al.* 2004, Wittkopp *et al.* 2008).

Despite the progress in mapping *cis* and *trans*-acting regulatory QTL, the mechanisms they act through are less well understood. Chromatin structure is a fundamentally important determinant of gene regulation, and changes in the position and number of nucleosomes can affect transcript abundance (Han and Grunstein 1988, Gross *et al.* 1993, Birney *et al.* 2010, Gossett *et al.* 2012). New technologies have enabled genome-wide maps of chromatin architecture to be constructed across different cell types (Thurman *et al.* 2012, Stergachis *et al.* 2013), individuals (McDaniell *et al.* 2010, Kasowski *et al.* 2013, Lee *et al.* 2013), and species (Tsankov *et al.* 2010, Shibata *et al.* 2012). Although these studies have revealed extensive variation in chromatin structure, many outstanding issues remain, including how much of variation in chromatin accessibility is heritable, the relative contributions of *cis* and *trans*-acting regulatory variation to differences in chromatin architecture (McDaniell *et al.* 2010), and how often variation in chromatin structure results in gene expression variation (Tirosh *et al.* 2009, Degner *et al.* 2012).

Here, we describe a genome-wide analysis of chromatin accessibility between two closely related *Saccharomyces sensu stricto* yeast species, *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*, and their hybrid. *S. cerevisiae* is the yeast model species and has been extensively studied. *S. paradoxus* is the most closely related species to *S. cerevisiae*, with an estimated divergence time of 5 million years ago (Kellis *et al.* 2003). Chromatin structure in *S. cerevisiae* has been studied previously (Hogan *et al.* 2006, Hesselberth *et al.* 2009) and across

a single genome, open chromatin regions are weakly associated with increased expression (Simon *et al.* 2013). In addition, nucleosome locations have been compared across multiple yeast species, including *S. cerevisiae* and *S. paradoxus*, and *cis* changes, such as anti-nucleosomal sequences and binding sites for general regulatory factors, were found to contribute to differences in nucleosome location (Tsankov *et al.* 2010). Within species, the genetic architecture of chromatin accessibility has been studied using QTL mapping (Lee *et al.* 2013); however, this has not been addressed between species.

In this study, we assessed chromatin accessibility using FAIRE-Seq and found considerable divergence in chromatin structure between *S. cerevisiae* and *S. paradoxus*. Moreover, we developed a novel statistical approach to identify *cis* and *trans*-acting effects on chromatin accessibility in hybrids and found *cis* effects on chromatin structure are more common than *trans* effects, are of greater magnitude, and that the direction of *cis* and *trans* effects are often in opposite directions suggesting compensatory evolution. Finally, we show that the relationship between chromatin structure and transcript levels in *S. cerevisiae* and *S. paradoxus* is complex, and a significant proportion of differences in chromatin accessibility might be functionally benign.

4.3 Results

4.3.1 Differences in chromatin accessibility within and between species

We first assessed differences in chromatin structure between haploid strains of *S. cerevisiae* and *S. paradoxus*. We generated FAIRE-Seq (Formaldehyde-Assisted Isolation of Regulatory Elements) data (Simon *et al.* 2013) for two biological replicates for two strains of *S. cerevisiae* (DBVPG1373, a wine strain, and UWOPS05_217_3, a wild isolate) and one strain of the sister species *S. paradoxus*, CBS432 (see Methods). FAIRE isolates DNA that is not bound

to proteins, resulting in increased signal in regions with increased chromatin accessibility. We sequenced FAIRE DNA samples to approximately 10x coverage using short read sequencing (see Methods). As expected, sequencing reads were enriched in intergenic regions (mean of 2.4x enrichment compared to coding regions).

We first asked which specific areas of the genome have undergone changes in chromatin accessibility between species. We focused on the nucleosome-free region (NFR) found upstream of the transcription start site of many yeast genes because this region is known to harbor important regulatory information; this was also where the dominant FAIRE signal was found in our data (Rando and Chang 2009), (see Figure C.1). We computationally identified the nucleosome-free region from the FAIRE data (see Methods) by identifying the peak in FAIRE signal found upstream of each gene and extended the region in either direction until a background level of signal was observed. We then merged NFR calls across the two species (see Methods). We also carried out extensive filtering to eliminate peaks whose differences might be caused by duplications between species or mapping issues (see Methods). In total, we identified 3498 NFRs that passed our filtering and had an average size of 253 bp.

We first compared one strain of *S. paradoxus*, CBS432, and one strain of *S. cerevisiae*, UWOPS05_217_3. Overall, the locations of NFRs called were well-conserved across species, and on average the location of 42% of NFRs overlapped between the two species. As a complementary analysis, we compared levels of chromatin accessibility in the set of all 3,498 NFRs, and found them to be strongly correlated (R^2=0.68 between species, $p < 2.2 \times 10^{-16}$) suggesting that broad-scale patterns of accessibility are conserved over time.

Next, we tested each of the 3,498 NFRs for differences in chromatin accessibility between the two parental haploid species, *S. cerevisiae* and *S. paradoxus,* and used the R

package DESeq to test for significant differences. We found 947 NFRs showed significant differences in FAIRE signal (FDR=0.05, Figure 4.1, see Methods). Furthermore, by analyzing the distribution of *p*-values (Storey and Tibshirani 2003), we estimate that π_0 (the proportion of NFRs with no differences in chromatin accessibility) is 0.53, suggesting that 47 percent of NFRs are differentially accessible between species. These 947 NFRs were upstream of 1149 distinct genes and on average resulted in a 2.17-fold difference in FAIRE signal between the two species. 483 of the NFRs showed higher accessibility in UWOPS05_217_3, while 464 NFRs showed higher accessibility in CBS432. We carried out a test for GO enrichment at the genes downstream of differentially accessible peaks and found that several GO biological process terms were enriched compared to the genome as a whole (corrected $p < 0.05$), specifically intracellular transport, protein localization, protein transport, and establishment of protein localization (Medina *et al.* 2010).

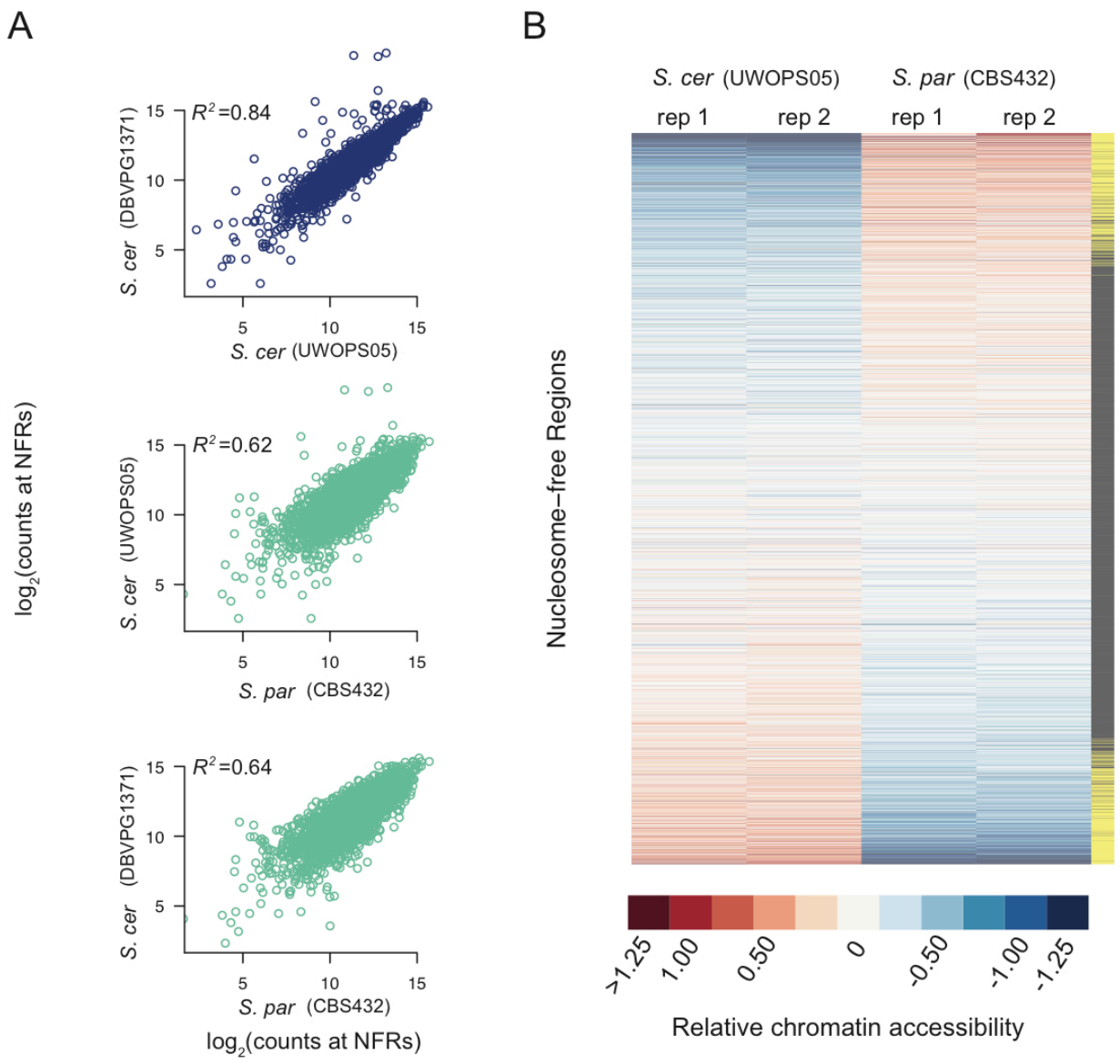

Figure 4.1. Patterns of chromatin accessibility within and between *S. cerevisiae* and *S. paradoxus*.

A. Scatterplots of relative chromatin accessibility between *S. cerevisiae* strains DBVPG1373 and UWOPS05_217_3 (top), *S. cerevisiae* strain UWOPS05_217_3 and *S. paradoxus* strain CBS432 (middle), and *S. cerevisiae* strain DBVPG1373 and *S. paradoxus* strain CBS432 (bottom). Note, comparisons within and between species are shown as blue and light green, respectively. B. Heatmap representation of chromatin accessibility at all NFRs in *S. cerevisiae* strain UWOPS05_217_3 versus *S. paradoxus* strain CBS432. Each row is a NFR, and columns are the two biological replicates of *S. cerevisiae* strain UWOPS05_217_3 and *S. paradoxus* strain CBS432. Rows are sorted by average difference in signal at NFRs between *S. cerevisiae* and *S. paradoxus*. The far right column indicates if the difference in chromatin accessibility between species is significant (yellow rectangles).

To assess the robustness of these results, we also generated FAIRE-Seq data for a second strain of *S. cerevisiae* (DBVPG1373, a wine strain). Divergence at synonymous sites between

these species is estimated to be 0.29 (Kellis *et al.* 2003). Levels of chromatin accessibility in NFRs was highly similar between the two *S. cerevisiae* strains (Figure 4.1; $R^2 = 0.84$; $p < 2.2 \times 10^{-16}$), and was of similar magnitude between species (Figure 4.1; mean $R^2 = 0.63$; $p < 2.2 \times 10^{-16}$). Similarly, of the 947 NFRs that showed differential accessibility between UWOPS05_217_3 and CBS432, 515 were also significantly different between DBVPG1373 and CBS432. Thus, patterns of chromatin accessibility are highly reproducible between genetically diverse strains of *S. cerevisiae* and *S. paradoxus*.

4.3.2 Genetic architecture of chromatin differences

To better understand the genetic architecture of the widespread differences in chromatin accessibility observed between *S. cerevisiae* and *S. paradoxus*, we developed novel statistical tests for the presence of *cis* and *trans* effects (see Methods; Figure 4.2). Simulations showed that these tests had high power and maintained correct false positive rates over a range of parameters (see Methods; Table C.1). Briefly, we tested for allele-specific chromatin accessibility within the hybrid to identify *cis* effects and tested for differences between the ratio of chromatin accessibility in the two parental species and the ratio of chromatin accessibility observed in the hybrid to identify *trans* effects (Figure 4.2). Over 99% of all NFRs identified in the parental strains contained one or more variants (median = 32) and could therefore be assessed for *cis* and *trans* effects. We identified 2256 NFRs showing a significant *cis* effect (posterior probability > 0.95, see Figure 4.3A) and 1020 NFRs showing a significant *trans* effect (posterior probability > 0.95, see Figure 4.3B). Interestingly, 782 NFRs showed both significant *cis* and significant *trans* effects. *Cis* effects were both more numerous as well as of greater magnitude on average (1.8-fold difference in chromatin accessibility compared to 1.6-fold difference in chromatin accessibility (Mann Whitney test, $p < 2.2 \times 10^{-16}$, Figure 4.3C). Strikingly, we found that *cis* and

trans effects were negatively correlated (r = -0.32, $p < 1 \times 10^{-16}$), which suggests a widespread role for compensatory evolution to stabilize chromatin structure (Figure 4.3D).

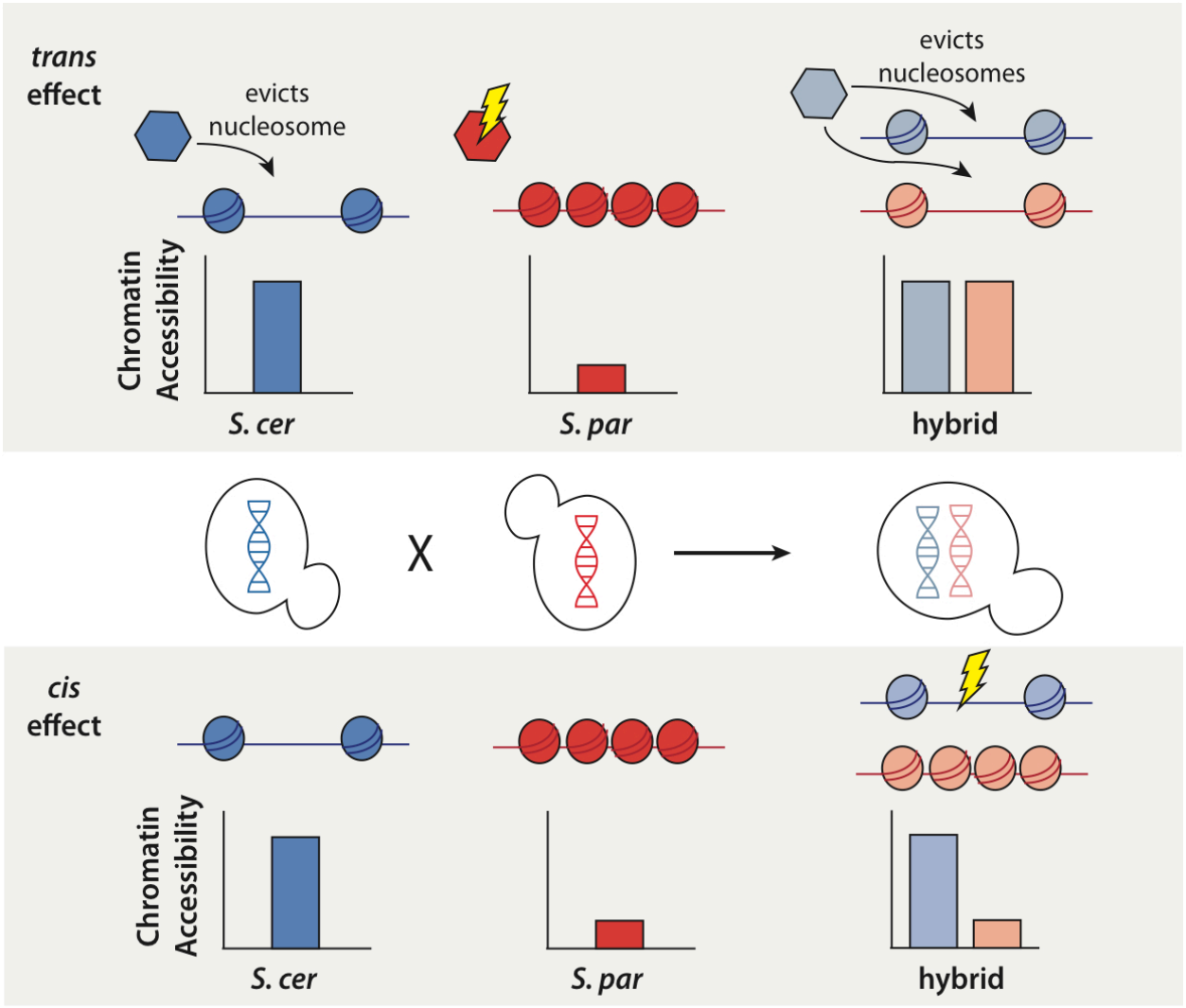

Figure 4.2. Schematic of approach to detect *cis* and *trans* effects on chromatin accessibility.

Top, an example of a NFR showing only a *trans* effect on chromatin accessibility. A *trans* effect is detected as a case where there is a difference in chromatin accessibility between the two parental haploid species, but there is no difference in chromatin accessibility between the two alleles in the hybrid. As shown above, this could be explained by a case where a nucleosome remodeler (shown as a hexagon) acts to evict nucleosomes and increase accessibility in *S. cerevisiae*, but a mutation in *S. paradoxus* has rendered it inactive and it is unable to evict the nucleosomes. In the diploid hybrid, the chromatin remodeler from *S. cerevisiae* is able to evict nucleosomes from both the *S. cerevisiae* and *S. paradoxus* chromosomes. Bottom, an example of a NFR showing only a *cis* effect on chromatin accessibility. A *cis* effect is detected as a difference between the accessibility detected between the two alleles in the diploid, and the lack of a *trans* effect is shown by the same difference being detected between the parental species. In this case, there has been a mutation at the NFR on the *S. cerevisiae* allele, leading to a difference in the number of nucleosomes binding in the region.

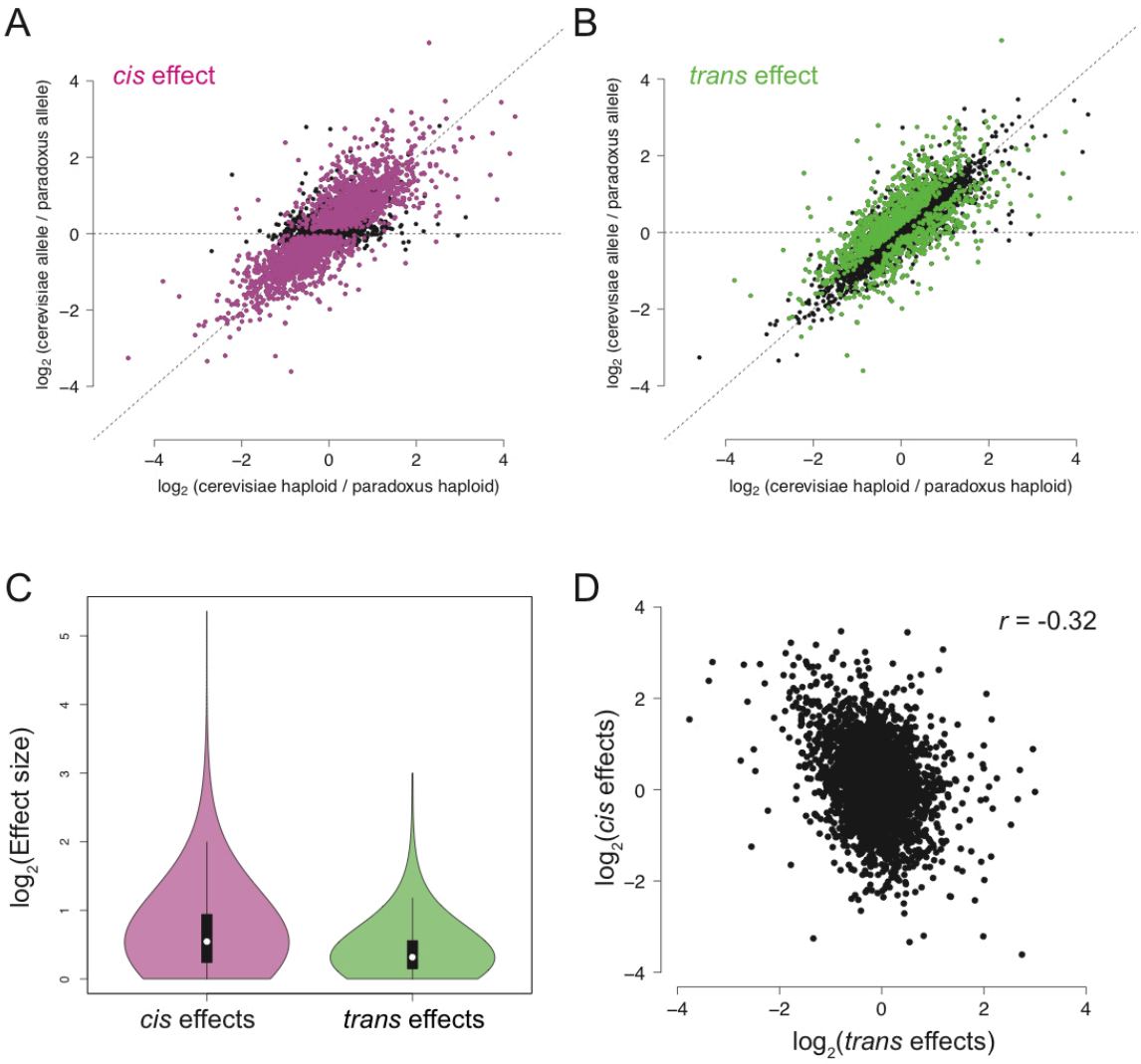

Figure 4.3. *Cis* and *trans* effects on chromatin accessibility.

A. For each NFR, the relative chromatin accessibility in the haploid is plotted versus the relative chromatin accessibility in the diploid. NFRs with a significant *cis* effect are shown in pink. B. Reproduction of the plot from (A), but NFRs with a significant *trans* effect are shown in green. C. Violin plot showing the effect size distribution of *cis* and *trans* effects. D. Scatter plot of relative *cis* and *trans* effect sizes. Positive effects indicate higher accessibility in *S. cerevisiae* and negative effects indicate higher accessibility in *S. paradoxus*.

4.3.3 Disrupted motifs are associated with cis effects

To test the hypothesis that *cis*-acting chromatin QTL result from variation in regulatory motifs, we identified motifs independently in the two species and computationally inferred whether sequence differences abrogated motif usage. Specifically, we define disrupted motifs as

those that were called in only one of the two species (see Methods). Disrupted motifs were strongly enriched in NFRs with significant *cis*-acting chromatin QTL ($p = 2.4 \times 10^{-7}$). We also found that overall nucleotide divergence was higher at NFRs with significant *cis* effects compared to regions without significant *cis* effects (Mann Whitney test, $p = 3.48 \times 10^{-6}$). Note, this observation parallels previous findings that polymorphism is higher for genes that show significant allele-specific expression in *S. cerevisiae* hybrids (Ronald *et al.* 2005).

We next asked if any of the 106 motifs were overrepresented for being disrupted in the set of significant *cis*-acting chromatin QTL. We found two overrepresented motifs, *GCN4* and *GZF3* (FDR = 0.10; Figure 4.4A). *GCN4* is an activator of amino acid biosynthetic genes, which itself is a tightly regulated pathway (Hinnebusch and Nataraian 2002). *GZF3* is a negative regulator of nitrogen catabolic gene expression (Stanbrough *et al.* 1995). While it is not immediately clear why disruption of these two genes is associated with changes in chromatin structure, it is interesting that both play an important role in metabolism, which is a highly regulated process.

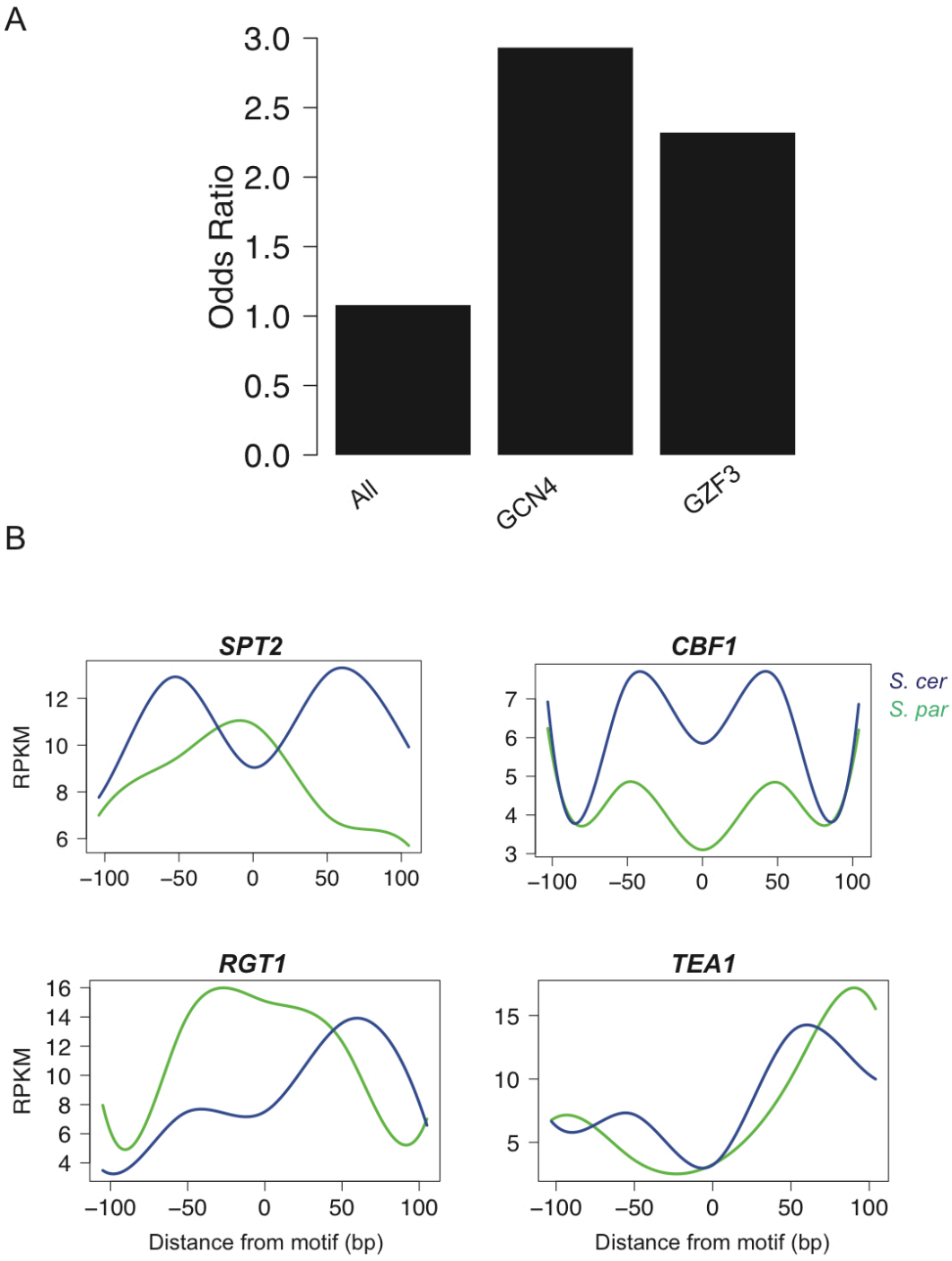

Figure 4.4. Motifs contributing to *cis* and *trans* effects.

A. The odds ratio of observing a disrupted motif compared to a non-disrupted motif in NFRs with a significant *cis* effect. Odds ratios are shown for all motifs, as well as the two individual motifs (*GCN4* and *GZF3*) that were found to be significant by permutations (FDR = 0.10). B. Pattern of accessibility for four motifs found within *trans* effect NFRs that vary between *S. cerevisiae* and *S. paradoxus*.

4.3.4 Differential footprints for certain DNA binding factors found at trans effects loci

To identify factors contributing to *trans* effects, we searched for cases where there was no disruption to the motif but the occupancy of the site changed between species. Such patterns could result from mutations that either alter the binding specificity of a *trans*-acting regulatory protein or change its regulation. We used the FAIRE data surrounding each motif to determine occupancy, analogous to a DNase I footprint (Hesselberth *et al.* 2009). We then tested whether there was a significant difference in the pattern of occupancy between species by fitting splines to the mean occupancy across conserved sites in *trans* regions and testing whether the splines were significantly different in a 100bp window surrounding the motif using bootstrapping (see Methods). We identified four motifs whose pattern of occupancy had significantly ($p < 0.05$) changed between species (Figure 4.4B). *SPT2*, a transcription factor that interacts with histones and the SWI/SNF complex, showed a clear footprint in *S. paradoxus*, but nearly the opposite pattern in *S. cerevisiae*, implying decreased occupancy in *S. cerevisiae* at these *trans* regions. Similarly, *TEA1*, a Ty enhancer activator, and *RGT1*, a glucose-responsive transcription factor, showed increased occupancy in *S. paradoxus*. Conversely, *CBF1*, a centromere binding factor also involved in stress response, showed higher FAIRE signals in *S. paradoxus* than *S. cerevisiae*, implying increased occupancy in *S. cerevisiae*.

4.3.5 Effects on gene expression

To examine the relationship between differences in chromatin accessibility and transcriptional divergence between *S. cerevisiae* and *S. paradoxus*, we performed RNA-Seq on the haploid parents and interspecific hybrid and tested for the *cis* and *trans* effects on gene expression values. Out of the 4,899 genes that could be aligned between species, 4,181 exhibited significant *cis* effects and 3,117 showed significant *trans* effects. Overall, *cis* and *trans* effects

on gene expression levels were smaller than those on chromatin accessibility, (Spearman rank-sum test, $p < 2.2 \times 10^{-16}$ for both *cis* and *trans* effects, Figure 4.5A).

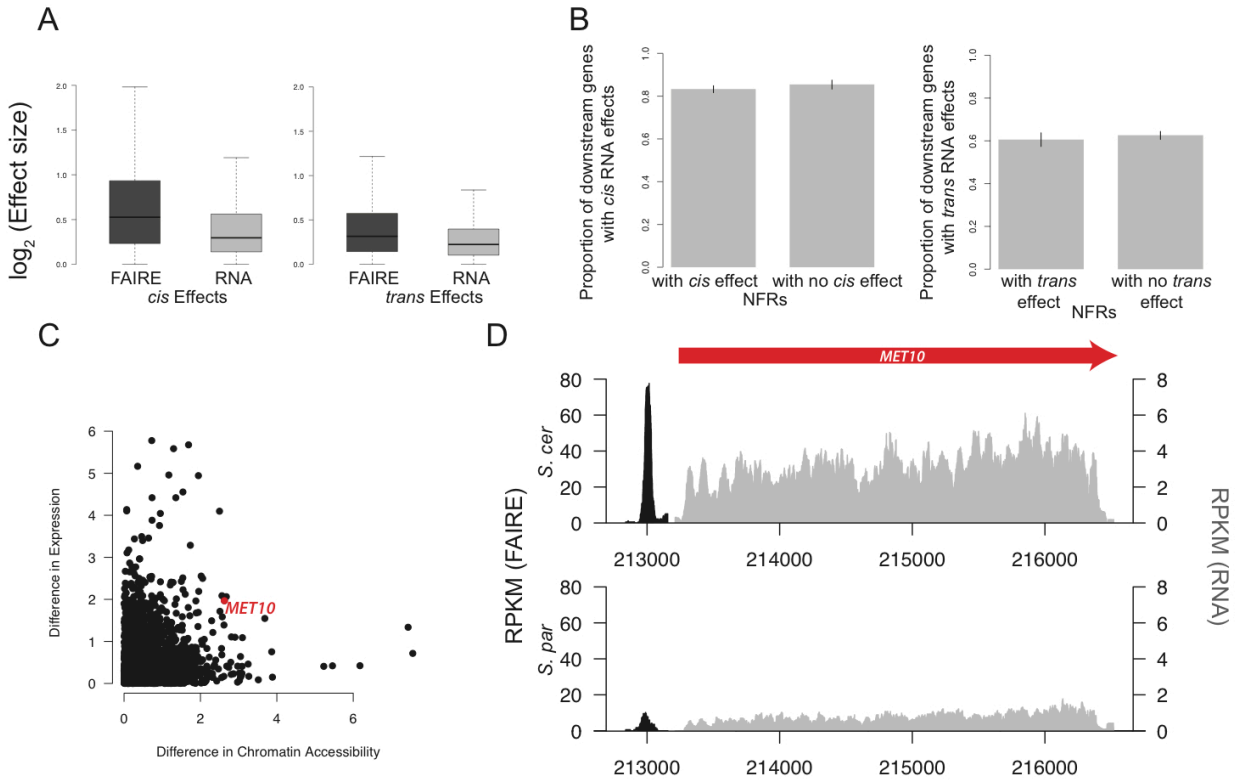

Figure 5. Gene expression and chromatin accessibility.

A. Boxplot of $\log_2$(effect size) of both *cis* and *trans* effects for FAIRE (dark grey) and RNA (light grey). B. Barplot of the percentage of genes with significant *cis* effects in RNA that are downstream of NFRs with and without *cis* effects (left). Barplot of the percentage of genes with significant *trans* effects in RNA that are downstream of NFRs with and without *trans* effects (right). C. Scatterplot of the $\log_2$(absolute value of the difference in chromatin accessibility between the two species) vs $\log_2$(absolute value of the difference in expression between the two species. The red dot indicates data from the *MET10* gene, whose FAIRE-Seq and RNA data are shown in panel D. For clarity, the FAIRE-Seq data is only shown in a 100bp window on either side of the NFR. FAIRE signal is shown in black, and RNA signal is shown in grey.

We next tested whether genes with a significant *cis* or *trans* effect in chromatin were more likely to have a significant *cis* or *trans* effect in transcript abundance. Specifically, we divided genes into categories of those downstream of an NFR with a *cis* effect on chromatin versus those downstream of an NFR without a *cis* effect on chromatin. We then compared the

percentage of genes showing *cis* effects on RNA in these two categories. Surprisingly, we did not find evidence that *cis* or *trans* effects in NFRs were more likely to be upstream of *cis* or *trans* effects on RNA, as would be expected if there was a simple correspondence between *cis* and *trans* effects in NFRs and RNA (see Figure 4.5B, Table C.2). This was true even when using varying cutoffs for the *cis* and *trans* effects, including ones that took into account the magnitude of effect sizes (Table C.2).

The relationship of *cis* and *trans* effects observed in gene expression and chromatin structure may be complicated by differences in statistical power. For example, 85% of all genes show significant *cis* effects on RNA. Thus, even if *cis* effects in NFRs are not more likely to be found upstream of *cis* effects on RNA, they could still contribute to gene expression variation between *S. cerevisiae* and *S. paradoxus*. To this end, we assessed whether expression differences between species could be modeled as a function of the *cis* and *trans* effects found upstream of each gene. Specifically, we fit the simple linear model: expression difference = Intercept + *cis* effect + *trans* effect + *cis* * *trans* effect + error, using the lm function in R. We found that both *cis* effects and *trans* effects on chromatin were significantly related to expression differences between species ($p = 0.002$, $p = 4.18 \times 10^{-5}$ respectively) though they explained a very small proportion of the total variance in expression between species (0.8% combined). The interaction term of *cis* and *trans* effects was not significant ($p > 0.05$). Interestingly, the motif for GZF3, which is significantly overrepresented in *cis* NFRs, was overrepresented in *cis* NFRs upstream of genes with *cis* effects on RNA.

Finally, we found no significant correlation between the magnitude of differences in chromatin accessibility and differences in gene expression between the parental species (Spearman rank-sum test, $p = 0.11$, Figure 4.5C). However, for a subset of NFRs, differences in chromatin

accessibility and gene expression do appear to be highly correlated. To identify these regions, we compared the $\log_2$(*S. paradoxus*/*S. cerevisiae*) for NFRs and gene expression at downstream genes and identified those whose absolute value of the difference between the two ratios was less than 0.25. We identified 701 such regions; one example is shown in Figure 4.5D.

4.4 Discussion

The ability to assay chromatin accessibility at high-resolution and on a genome-wide scale has enabled comprehensive insights into the structure and function of chromatin in many cell types, developmental stages, and organisms. Here, we were particularly interested in the evolutionary dynamics of changes in chromatin accessibility between two closely related yeast species. Broad-scale patterns of chromatin accessibility have been well conserved between *S. cerevisiae* and *S. paradoxus* (Figure 4.1), but superimposed on this background of conservation, we estimate that nearly 50% of NFRs exhibit differential accessibility.

To better understand the relative contributions of *cis* and *trans* effects on differences in chromatin accessibility observed between *S. cerevisiae* and *S. paradoxus*, we developed novel statistical methods to analyze FAIRE-Seq data from diploid hybrids. Similar to previous findings on RNA levels (Brem *et al.* 2002, Wittkopp *et al.* 2004, Wittkopp *et al.* 2008, Tirosh *et al.* 2009), differences in chromatin accessibility are caused by changes both in *cis* and in *trans.* In our data, *cis* effects were of greater magnitude and were more abundant. Recently, Lee *et al.* performed a study similar to ours and assessed *cis* and *trans* effects on chromatin structure in a cross between two strains of *S. cerevisiae* (Lee *et al.* 2013). In contrast to our observations, they found that *trans* QTL were more pervasive than *cis* QTL (92.1% of associations versus 7.9% of associations) (Lee *et al.* 2013). We hypothesize that these disparate observations are primarily

the consequence of differences in the evolutionary trajectory of chromatin accessibility QTL in within versus between species data. In particular, *trans*-acting chromatin QTL are likely to be subject to more intense purifying selection due to their potential pleiotropic effects, and tend to be eliminated over longer time periods (Ronald and Akey 2007). This hypothesis is consistent with findings for expression QTL studies, which showed that *trans*-eQTL were more common within species and *cis*-eQTL were more common between species (Tirosh *et al.* 2009, Emerson *et al.* 2010). Consistent with this hypothesis, we found that *cis* and *trans* effects were significantly negatively correlated, indicating that chromatin accessibility in each species is subject to stabilizing selection and perturbations of chromatin structure are, on average, deleterious.

We estimated *cis* and *trans* effects for both chromatin accessibility and gene expression levels. Unexpectedly, the presence of *cis* or *trans* effects on chromatin accessibility in NFRs was not significantly associated with *cis* or *trans* effects on RNA. In other words, gene expression levels with significant *cis* or *trans* effects were not more likely to have an NFR with significant *cis* or *trans* effects on chromatin accessibility. Thus, it appears that many of the changes in chromatin accessibility in NFRs between *S. cerevisiae* and *S. paradoxus* do not necessarily have transcriptional consequences. One factor that may contribute to this observation is that compensatory changes downstream of chromatin accessibility, such as mutations that influence mRNA stability, may evolve to maintain levels of gene expression. In addition, many changes in chromatin accessibility may simply be functionally benign.

Furthermore, an important caveat is that our data was obtained from a single environmental condition, and it is plausible that stronger correlations between chromatin and gene expression QTL may exist when analyzing data from either a different environment or

across multiple environments. Nonetheless, the lack of a clear relationship between chromatin and gene expression QTL in our data is interesting in light of recent observations from the ENCODE Project that have found a large proportion of the human genome has reproducible biochemical activity (Bernstein *et al.* 2012). Our results suggest caution in assuming all, or perhaps even most, of such sequences are functionally important.

4.5 Materials and Methods

Strain growth, FAIRE, and RNA-Seq

65 ml of each of 2 biological replicates of the *S. paradoxus* strain CBS432 and the two *S. cerevisiae* strains DBVPG1373 and UWOPS05_217_3 were grown to mid-log phase. 15 ml were used for RNA-seq and 50 ml were used for FAIRE. FAIRE was performed as described in Simon *et al.* 2012, with some modification. The cells were fixed with 1% formaldehyde for 35 minutes with mixing. Cells were sonicated using the Fisher Scientific Sonic Dismembrator Model 100 for three cycles of 15 one-second bursts with 1 second rest in between, keeping the cells on ice for at least 30 seconds between cycles. The remainder of the protocol was followed as in Simon *et al.* (2013). RNA isolation was done using the hot phenol protocol (Rose *et al.* 1990), and RNA was treated with Turbo DNAse before library construction.

Library construction and sequencing

Libraries were constructed for the FAIRE samples using the Illumina TruSeq DNA kit, starting with approximately 200 ng FAIRE DNA, following their standard kit protocol but omitting the fragmentation step. RNA libraries were prepared using the Illumina TruSeq RNA kit, following their standard protocol. Libraries were pooled into two lanes, one for the FAIRE samples and

one for the RNA samples, and were sequenced on the HiSeq 2000. Raw sequence data and processed files are available at the GEO database with accession number GSE55717.

Read mapping

Reads were mapped to genomes assembled in Skelly *et al.* 2013 for the *S. cerevisiae* haploid samples using bwa and samtools (Li and Durbin 2009, Li *et al.* 2009). For the *S. paradoxus* strain CBS432, we used the last updated reference version from the SGRP (Liti *et al.* 2009). For the diploid samples, we mapped to a combined FASTA containing both genomes. We tested whether mapping to each genome separately for the diploid samples resulted in increased mapping; it did not. For the diploid samples, we generated simulated reads and mapped to the combined FASTA. For all further analyses, we restricted analysis to NFRs for which greater than 90 percent of simulated reads mapped back to the correct region. We also sequenced a genomic DNA sample. We also filtered out NFRs where the absolute value of the $\log_2$(ratio of reads between the two species) for the genomic DNA was greater than 0.3.

Identifying NFRs

We identified NFRs as follows: specifically, starting at the beginning of the coding region of the gene, we looked for the peak of chromatin accessibility within 300bp upstream of the start codon. We then defined the edges of the NFR as the base-pair after which at least 3 bases had had a chromatin accessibility count of less than 10. We did this separately for each biological replicate and each species. For each gene separately, we then merged NFRs if they were within 200bp.

Filtering NFRs and genes

In order to convert between the two species coordinates, we created a multiple alignment between the two species using LASTZ and TBA (Harris 2004, Blanchette 2004). We inferred scoring parameters using the two species of interest. Using this multiple alignment, we then converted the NFRs called in CBS432 to *S. cerevisiae* coordinates, and found the union of all NFRs called across the samples. We used this union of NFRs for further tests. We also filtered the NFRs based on a reciprocal alignment filter, where we required that NFRs align to only one region in the other species, based on the multiple alignment. This allowed us to filter out regions with duplications or deletions between the two species.

Identifying differentially accessible NFRs

Using samtools, we summed the count of reads mapping in each species across each NFR or gene in both biological replicates. Note that we did this in the native coordinates for each species, filtering out sites which were called as indels in the multiple alignment. We then used the R package DESeq (Anders and Huber 2010) to assess differential FAIRE signal between species. This method takes into account biological replicates, and models the count distribution using a negative binomial distribution. We used the R package qvalue to estimate q-values (Storey and Tibshirani 2003). We used a significance threshold of FDR = 0.05 unless otherwise noted.

Statistical model to detect *trans* effects

If differences in chromatin accessibility between *S. cerevisiae* and *S. paradoxus* are due to *trans*-acting factors, the relative chromatin accessibility in the haploid parents will be different than the relative chromatin accessibility in the diploid hybrid (Fig. 4.2). We leveraged the FAIRE-Seq

data to detect differences in the relative levels of chromatin accessibility between F_1 hybrids and the parental species. Specifically, let N_c and N_p be the total number of reads across the genome mapping to polymorphic sites in the *S. cerevisiae* and *S. paradoxus* haploid parents, respectively. For a particular locus j, Y_c and Y_p denote the observed number of reads mapping to *S. cerevisiae* and *S. paradoxus*, respectively. Then assume: $Y_c|r_c \sim$ Binomial (N_c, r_c) and $Y_p|r_p \sim$ Binomial(N_p, r_p), where r_c and r_p denote the probabilities of observing a read mapping to *S. cerevisiae* or *S. paradoxus* for a particular locus, respectively. Since N_c and N_p are large, and r_c and r_p are small, we can approximate these binomials by Poissons to give: $Y_c|r_c \sim$ Poisson($N_c\, r_c$) and $Y_p|r_p \sim$ Poisson($N_p\, r_p$).

We define $\theta_P = r_c/r_p$ to be the ratio of these probabilities in the parents and $R = N_c/N_p$ to be the ratio of the total numbers of reads in each parent. Then, $Y_c|Y_c+Y_p, s_c \sim$ Binomial(Y_c+Y_p, s_c), where $s_c = N_c r_c/ (N_c r_c + N_p r_p) = R\theta_P /(R\theta_P +1)$ is the probability of observing a read map to *S. cerevisiae*, without adjusting for differences in the total number of reads mapping to each species. We can thus write $\log(S_c/1\text{-}S_c) = \log R + \log \theta_P$, such that θ_P is the odds of observing a read map to *S. cerevisiae* compared to *S. paradoxus* for a particular locus in the haploid parents, adjusted for differences in the total number of reads mapping to each species.

For the diploid hybrid, let Z_c and Z_p denote the number of reads mapping to *S. cerevisiae* and *S. paradoxus* SNPs within locus j, respectively. Thus, $Z_c|Z_c+ Z_p, p_c \sim$ Binomial(Z_c+Z_p, p_c), where p_c is the probability of observing a read map to the *S. cerevisiae* allele for a particular locus. The odds of observing a read map to *S. cerevisiae* in the hybrid for a particular gene is $\theta_H = p_c /(1-p_c)$. In the following, let Y_{cj}, Y_{pj}, Z_{cj}, and Z_{pj} represent the data as defined above, but with j = [1,2] indexing biological replicate.

Thus, the locus specific models are:

$Y_{cj}|Y_{cj}+Y_{pj}, s_{cj} \sim \text{Binomial}(Y_{cj} + Y_{pj}, s_{cj})$,

$Z_{cj}|Z_{cj}+Z_{pj}, p_{cj} \sim \text{Binomial}(Z_{cj} + Z_{pj}, p_{cj})$

$\text{logit}\ s_{cj} = \log R_j + \log \theta_P + \delta_j$

$\text{logit}\ p_{cj} = \log \theta_P + \Delta + \varepsilon_j$

where $R_j = N_{cj} / N_{pj}$, $\delta_j \sim N(0, \sigma^2)$ and $\varepsilon_j \sim N(0, \sigma^2)$ represent random effects that allow for excess-binomial variation. Here, Δ is the parameter of interest and provides an estimate of the difference between $\log(\theta_P)$ and $\log(\theta_H)$, as described above. The above framework is an example of a generalized linear mixed model (GLMM) and we used a Bayesian approach to inference with relatively flat hyperpriors. One computationally intensive method for summarizing the posterior would be Markov chain Monte Carlo (MCMC) but the integrated nested Laplace approximation (INLA) as described in Paul *et al.* 2010 provides an efficient alternative for GLMMs (Fong *et al.* 2010). We used the R implementation of INLA to estimate Δ. We examined a 95% posterior interval estimate for Δ and recorded whether this interval contained 0 or not. If the interval does not contain 0 it indicates that chromatin accessibility differs.

Statistical model to detect *cis* effects

To detect *cis* effects, we developed a model to test for differential accessibility between alleles within the diploid hybrid. Let Z_{cj}, and Z_{pj} represent the data as defined above. We can therefore write:

$Z_{cj}|Z_{cj} + Z_{pj}, p_{cj} \sim \text{Binomial}(Z_{cj} + Z_{pj}, p_{cj})$

$\text{logit}\ p_{cj} = \log \theta_H + \varepsilon_j$

with $\varepsilon_j \sim N(0,\sigma^2)$ representing random effects that allow for excess-binomial variation. In this model, θ_H is the parameter of interest and provides an estimate of the odds of a read mapping to

the *S. cerevisiae* allele compared to the *S. paradoxus* allele in the diploid hybrid for a particular gene. We again used the R program INLA to estimate the posterior for $\log(\theta_H)$ and in particular examine whether the 95% posterior interval estimate contains 0.

Simulations

We carried out extensive simulations to evaluate the operating characteristics of our model. Specifically, for the *trans* model, we set the total number of reads mapping to polymorphic sites for species 1 (N_{c1}) equal to 5×10^6, and drew the total number of reads mapping to polymorphic sites for the other species and replicate from a normal distribution with mean N_{c1} and standard deviation N_{c1}. We then drew the value for r_c, the probability of a read mapping to *S. cerevisiae* for a particular locus from an exponential distribution with rate 10,000. For $N_{c1} = 5 \times 10^6$, this results in a mean of 500 reads mapping to a locus, with most having less than 500 reads, consistent with the observed data. We drew the value for r_p, the probability of a read mapping to *S. paradoxus* for a particular locus, from a normal distribution with mean r_c and standard deviation r_c and took the absolute value to ensure r_p was greater than zero. Using these values, we derived Y_c and Y_p, the number of reads mapping to *S. cerevisiae* and *S. paradoxus,* respectively, for a particular locus, for two biological replicates as specified by the model. For Z_c and Z_p, the number of reads mapping to the *S. cerevisiae* and *S. paradoxus* alleles in the hybrid summed across polymorphic sites in a particular locus, we either derived these using the same r_c and r_p values as above, to simulate a locus which showed no *trans* effect, or we set the value of $\log_2(\theta_P) - \log_2(\theta_H)$ equal to 0.1, 0.5, or 0.8, to simulate a locus with a *trans* effect. Note, this spans the range of detected *trans* effects. For 100 replicates, we simulated a collection of 6000 loci, 5000 of which did not show a *trans* effect and 1000 or which did show a *trans* effect.

For each of the 100 replicates, we then used the method described above to test whether the 95% posterior interval estimate for Δ for each locus contained zero.

To evaluate the *cis* test, we again started with the same values for the total number of reads. To simulate a locus with no *cis* effect, we set the value of $\log_2(Z_c/Z_p)$ equal to zero, and to simulate a locus with a *cis* effect, we set the value of $\log_2(Z_c/Z_p)$ equal to 0.1, 0.5, or 0.8. Again, for 100 simulations, we simulated a collection of 6000 loci, 5000 showing no *cis* effect and 1000 showing a *cis* effect. For each simulated set of loci, we then used the statistical method above to test whether the 95% posterior interval estimate for $\log(\theta_H)$ for each locus contained zero to test for a significant *cis* effect.

We found that the false discovery rate for both *cis* and *trans* based on a test based on a 95% interval was 0.05. Moreover, we found that the *trans* test has reduced power compared to the *cis* test, as expected because there were more parameters that could vary across biological replicates. However, with an effect size=0.5 for both the *cis* and *trans* tests, there was significant power to detect the *cis* or *trans* effects (Table C.1).

Motif analysis

We called motifs separately in both species, using MEME, using their standard p value cutoff of $p < 10^{-4}$ (Bailey *et al.* 2009). This results in the same cutoffs used for both species. Motifs that were not called in both species were considered polymorphic. We filtered out motifs where the polymorphism was due to indels in order to mitigate alignment errors. The motif calls used for this analysis are available as supplementary data on our website (http://akeylab.gs.washington.edu/downloads.shtml).

We compared the proportion of disrupted motifs (those that were called in only one species) in *cis* NFRs to non-*cis* NFRs using the Fisher exact test. We determined significance by permutations; we permuted the assignment of *cis* or not *cis* NFRs 1000 times and obtained p values from the permutations. We then used the positive False Discovery Rate approach to determine significance (Storey and Tibshirani 2003).

Occupancy at *trans* NFRs

We obtained the RPKM in a 200bp window surrounding motifs that were conserved across species in *trans* NFR regions for each of the two species. We filtered out motifs that did not have at least five instances of conserved motifs. We fit a cubic smoothing spline to the mean coverage using the R function spline. We then bootstrapped the data 1000 times by resampling from the motifs for each species. At five bp intervals across the region, we then tested whether the coverage was significantly different between the species, using the confidence intervals obtained from the bootstrapping. We then manually inspected the significant motifs ($p < 0.05$) to identify those which appeared to affect the FAIRE signal at or the near the motif.

Chapter 5

Summary and future directions

5.1 Summary

In this dissertation, I have used yeast as a model organism to assess the effects of noncoding variants and chromatin architecture on gene expression. First, I asked how well association mapping works in yeast and found that complicated population structure makes it difficult to use yeast in association mapping without careful choice of which strains to use.

Secondly, I assessed the evolutionary pressures acting on noncoding regions in yeast, specifically the differences in pressures on potentially functional sites, motif binding sites, and other noncoding sites, and found that there was strong purifying pressure acting at motif binding sites. I then used data from yeast strains to associate noncoding variants with differences in gene expression.

Finally, I used methods to map chromatin accessibility to ask how differences in the accessibility of chromatin affect differences in gene expression between species. I found that differences in chromatin accessibility were numerous, that these differences are driven primarily by genetic changes in *cis*, and that the effects of chromatin accessibility on gene expression appear to be modest.

5.2 Alternative approaches for association mapping

In Chapter 2 of this dissertation, I found that the combinations of strains sequenced so far in yeast can lead to increased false positive rates when used in association tests. There has been a great deal of work recently to develop methods to control for complicated population structure in association mapping (Yang *et al.* 2014). However, in Chapter 2, we found that even using these methods did not fully reduce the rate of false positives. We suggested alternative approaches to

reduce the problem of population structure in yeast, including choosing intelligently which of the sequenced strains to use. Indeed, one advantage of working in a model organism is the ability to intelligently choose the starting population for association mapping. It has become clear that using global collections of strains leads to complicated patterns of population structure (Liti *et al.* 2006, Schacherer *et al.* 2009). From a population genetics perspective, strains from different parts of the world with complicated population structure patterns are not an idealized 'population'. In addition to choosing intelligently from the already sequenced strains, another approach to this could be to collect new groups of strains that would be more likely to fit the population genetics definitions of an idealized population, such as multiple strains from the same geographic location. This would presumably lead to less complicated population structure, although it would also lead to a less diverse group of strains.

5.3 Studying regulatory changes in different environments

In Chapters 3 and 4 of this dissertation, I looked at the effects of noncoding variation on gene expression in two environments, namely phosphate limitation and rapid growth. One limitation of much of the work done so far in yeast on gene regulation is that it has focused on a single environmental condition, namely, rapid growth. There are numerous reasons why this might not be the ideal condition under which to study gene regulation. One reason is that expression is not unchanging during the log phase of growth, and now that we are using more sensitive measurements we may want to find a way to get more consistent environmental responses (Hayes *et al.* 2002). Secondly, growth on a nutrition source with every amino acid required is unlikely to be representative of the conditions a wild or fermentation yeast would experience. Finally, studying a single condition will inevitably give an incomplete picture of regulatory networks and which variants are functional. For instance, a study of gene expression evolution in

three different environments found that *cis* effects were well correlated across conditions, while *trans* effects were more likely to be condition-specific (Tirosh *et al.* 2009). This means that we may be missing information about the role of *trans* effects by studying only a single environment. Finally, it's possible that more can be learned from integrating information about regulatory responses across multiple conditions.

5.4 Integrating more levels of information

Another approach that may yield more insights into gene regulation and its downstream affects is to integrate multiple levels of genome-wide datasets. For example, there are now many steps in of the process of gene expression which can be measured using genome-wide high-throughput methods (Figure 5.1)

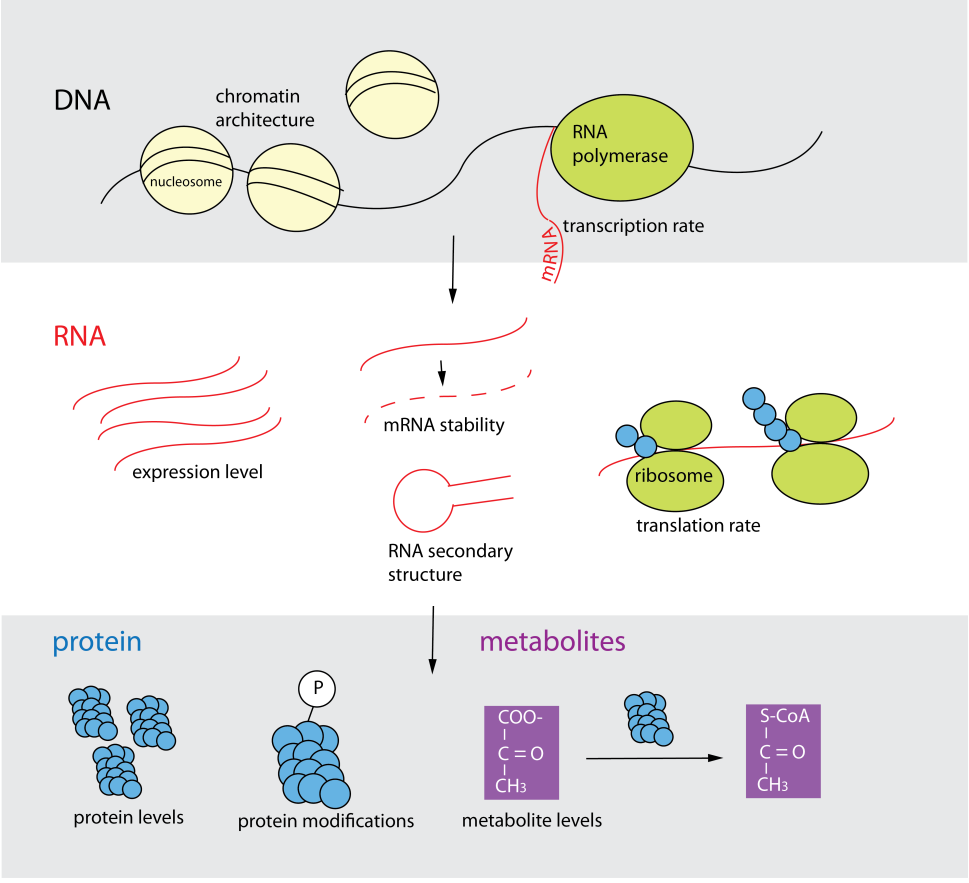

Figure 5.1 Molecular intermediates and rates which can be measured genome-wide.

Much has been learned in the past few years about how some of the later steps of this process are carried out. In particular, there has been work looking at the genetic basis of protein differences. Studies have attempted to address whether and how differences in gene expression are related to differences in protein levels. Overall, it appears that some of the variation in mRNA levels appears to be buffered in protein levels (Foss 2011, Ghazalpour 2011, Low *et al.* 2013). In yeast, researchers found that loci linked to variation in gene expression were also likely to be linked to variation in protein levels, and that previous studies which had tried to compare the two had been unable to find as high of a correspondence due to differences in power between the studies (Skelly *et al.* 2013, Albert *et al.* 2014). However, overall about 50% of eQTL were manifest as pQTL, emphasizing the importance of post-transcriptional processes to protein levels (Albert *et al.* 2014). In general, studies of protein levels are still relatively low-throughput compared to gene expression and other functional genomics phenotypes. For example, in the mice study they were able to assay 486 proteins, Albert *et al.* (2014) studied 174 abundantly expressed genes, with possibly the largest study to date assaying 2100 proteins (Picotti *et al.* 2013). There is still much to learn about how gene expression levels are related to protein expression and ultimately to phenotype.

Another approach taken in integrative studies is to attempt to identify regulatory modules or to construct regulatory networks. There has been a great deal of interest in using gene expression to predict phenotypes, particularly for clinical applications (Beer *et al.* 2002, Van de Vijver *et al.* 2002, Curtis *et al.* 2012). In model organisms, researchers have begun conducting large scale experiments where transcript abundance, genotype, and an array of phenotype data are collected at once in order, such as in Drosophila, Arabidopsis, and yeast (Ayroles *et al.* 2009, Fu *et al.* 2009, Skelly *et al.* 2013). Expression data can also be used to construct networks.

These networks can then be used to learn more about functional relationships among genes (Emilsson 2008, Chen *et al.* 2008, Zhu *et al.* 2008). Ultimately, genetic networks will be useful in that they can inform relationships among genes, and if we can understand the flow of information through a pathway, we may be better able to understand the effects of perturbations within a pathway.

5.5 Experimental testing of regulatory alleles

Identifying specific regulatory alleles after identifying linkage or association to a particular region remains challenging (Ehrenreich *et al.* 2009, Tirosh *et al.* 2009, Thompson and Regev 2010, Yvert *et al.* 2003). To illustrate this, some groups have gone on to map the specific differences causing expression differences at particular genes (Tao *et al.* 2006, Prud'homme *et al.* 2006). Tao *et al.* (2006) identified and experimentally validated five distinct *cis*-acting variants affecting gene expression at the *KRT1* gene in human cell lines. To fine-scale map causal SNPs, they predicted protein binding sites to identify candidate variants, tested for protein binding at those sites, and used reporter assays to confirm the functional effects of specific alleles. They found that of 5 SNPs they validated as having functional effects, 3 acted to increase expression and 2 to decreased expression (Tao *et al.* 2006). Finding multiple functional variants in a region, and ones which do not act in the same direction, appears to be a common outcome (Flint and Mackay 2009). Overall, *cis*-regulatory effects on gene expression can be complex and combinatorial in nature.

This type of experimental testing of regulatory alleles is still somewhat challenging and slow to carry out. A number of labs have recently worked on methods that attempt to scale up the number of alleles that can be tested at one time. One goal of these approaches is to learn more about how the spectrum of possible mutations might affect phenotypes, in order to be more

predictive when presented with a new mutation. Many of the methods developed so far involve generating libraries of mutations within a short stretch of DNA and measuring the effects (Fowler *et al.* 2010, Patwardhan *et al.* 2012, Bonde *et al.* 2014). These approaches will help generate a better null distribution of what the effects of different mutations might be.

5.6 Predictions of causative alleles

The goal of many studies of regulatory variation is to better inform our ability to make predictions about how genetic changes to regulatory regions may affect downstream phenotypes. Using the information we have available now, many papers have tried to make predictions about what types of amino acid or single nucleotide changes might be causative for a particular disease or phenotype (Lee *et al.* 2009). For example, for coding variants, PolyPhen uses information about potentially damaging amino acid changes to proteins to predict functional changes within coding regions (Cooper *et al.* 2010). Similarly, measurements of genetic constraint have been used in a similar manner (Siepel *et al.* 2005). Kircher *et al.* (2014) used machine learning to build predictions of potentially 'causal' genetic variants using data such as conservation, functional annotations such as DNaseI hypersensitivity, and measures of function such as PolyPhen. Overall, functional annotations improved predictions, although overall the strongest signal came from conservation information (Kircher *et al.* 2014). In many cases, we still do not have a good sense of what function different variants might have (i.e. a good 'training set'), making this type of prediction challenge harder, though the experimental approaches mentioned in the previous section may be one way to address this.

5.7 Concluding remarks

In conclusion, we have learned a great deal so far about the function of noncoding regions. However, overall, the questions which motivated some of the earliest researchers who

hypothesized the importance of regulatory changes to evolutionary change remain unanswered, and we are only beginning to learn about the importance of regulatory change to evolutionary processes. Methods to map functional noncoding regions have the potential to greatly increase our understanding of how DNA is translated into phenotypes, combined with experimental tests of regulatory function. We are also beginning to learn about what types of genetic changes have led to differences in chromatin architecture, gene expression, and protein levels between or within species. There is still much to learn about the multiple steps of producing cellular phenotypes from DNA from RNA to proteins and how genetic variation affects these different steps. Hopefully from the vast amount of functional genomics data currently being produced, we will soon be able to answer more definitely questions about how evolutionary change occurs, what contribution different types of regulatory changes have made to such changes, and the mechanisms by which these regulatory changes occur.

References

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ (2004) Bmp4 and morphological variation of beaks in Darwin's finches. *Science* 305:1462-5.

Akey J, Ruhe A, Akey D, Wong A, Connelly C, Madeoy J, Nicholas T, Neff M (2010) Tracking footprints of artificial selection in the dog genome. *Proc Natl Acad Sci USA* 107:1160-1165.

Albert FW, Treusch S, Shockley AH, Bloom JS, Kruglyak L (2014) Genetics of single-cell protein abundance variation in large yeast populations. *Nature* 506: 494-497.

Allison AC (1954) The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria. *Trans R Soc Trop Med Hyg* 48:312-8.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11:R106.

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

Asenjo AB, Rim J, Oprian DD (1994) Molecular determinants of human red/green color discrimination. *Neuron* 12:1131-8.

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

Ayroles J, Carbone M, Stone EA, Jordan K, Lyman R, Magwire M, Rollmann S, Duncan L, Lawrence F, Anholt R, et al. (2009) Systems genetics of complex traits in Drosophila melanogaster. *Nat Genet* 41:299-307.

Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* 3:21-29.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37:W202-8.

Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, et al. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 8:816-24.

Birney E, Lieb JD, Furey TS, Crawford GE, Iyer VR (2010) Allele-specific and heritable chromatin signatures in humans. *Hum Mol Genet* 19:R204-R209.

Blanchette M (2004) Aligning multiple genomic sequences with the threaded blockset al.igner. *Genome Res* 14:708-715.

Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, vonHoldt BM, et al. (2010) A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol*. 8:e1000451.

Bonde MT, Kosuri S, Genee HJ, Sarup-Lytzen K, Church GM, Sommer MO, Wang HH (2014) Direct Mutagenesis of Thousands of Genomic Targets Using Microarray-Derived Oligonucleotides. *ACS Synth Biol* ePub ahead of print.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M (2007) Divergence of transcription factor binding sites across related yeast species. *Science* 317:815-9.

Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB (2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol* 8: e1000343.

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752-5.

Breuker CJ, Debat V, Klingenberg CP (2006) Functional evo-devo. *Trends Ecol Evol* 21:488-92.

Britten RJ, Davidson EH (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* 46:111-38.

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* 36:D102-6.

Buenrostro J, Giresi PG, Zaba L, Chang H, Greenleaf W (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Meth* 10:1213-1218.

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

Chen K, van Nimwegen E, Rajewsky N, Siegal ML (2010) Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae. *Genome Biol Evol* 2:697-707.

Chen Y, Zhu J, Lum P, Yang X, Pinto S, Macneil D, Zhang C, Lamb J, Edwards S, Sieberts S, et al. (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature* 452:429-435.

Connelly CF, Akey JM (2012) On the prospects of whole-genome association mapping in Saccharomyces cerevisiae. *Genetics* 191:1345-53.

Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods* 7:250-251.

Curtis C, Shah SP, Chin S, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486:346-352.

Cusanovich D, Pavlovic B, Pritchard JK, Gilad Y (2014) The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genet* 10: e1004226.

Degner JF, Pai AA, Pique-Regi R, Veyrieras J, Gaffney DJ, Pickrell JK, Leon SD, Michelini K, Lewellen N, Crawford GE, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482:390-4.

Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* 19:1114-21.

Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* 3(5):e99.

Dorschner MO, Hawrylycz M, Humbert R, Wallace J, Shafer A, Kawamoto J, Mack J, Hall R, Goldy J, Sabo PJ, et al. (2004) High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods* 1:219-225.

Edwards A, Ayroles J, Stone E, Carbone M, Lyman R, Mackay T (2009) A transcriptional network associated with natural variation in Drosophila aggressive behavior. *Genome Biol* 10: R76.

Ehrenreich IM, Gerke JP, Kruglyak L (2009) Genetic dissection of complex traits in yeast: insights from studies of gene expression and other phenotypes in the BYxRM cross. *Cold Spring Harb Symp Quant Biol* 74:145-53.

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

Emilsson V, Thorleifsson G, Zhang B, Leonardson A, Zink F, Zhu J, Carlson S, Helgason A, Walters G, Gunnarsdottir S, et al. (2008) Genetics of gene expression and its effect on disease. *Nature* 452:423-428.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I (2002) Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 30:233–237.

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

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799-816.

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

Farnham PJ (2009) Insights from genomic profiling of transcription factors. *Nat Rev Genet* 10:605-616.

Fehrmann R, Jansen RC, Veldink J, Westra H, Arends D, Bonder M, Fu J, Deelen P, Groen H, Smolonska A, et al. (2011) Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA. PLoS Genet 7:e1002197.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164-166.

Flint J, Mackay T (2009) Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res* 19:723-733.

Fong Y, Rue H, Wakefield J (2010) Bayesian inference for generalized linear mixed models. *Biostatistics* 11:397-412.

Foss E, Radulovic D, Shaffer S, Goodlett D, Kruglyak L, Bedalov A (2011) Genetic Variation Shapes Protein Networks Mainly through Non-transcriptional Mechanisms. *PLoS Biol* 9:e1001144.

Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S (2010) High-resolution mapping of protein sequence-function relationships. *Nat Methods* 7:741-746.

Francesconi M, Jelier R, Lehner B (2011) Integrated Genome-Scale Prediction of Detrimental Mutations in Transcription Networks. *PLoS Genet* 7:e1002077.

Fu J, Keurentjes J, Bouwmeester H, America T, Verstappen F, Ward J, Beale M, De Vos R, Dijkstra M, Scheltema R, et al. (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat Genet* 41:166-167.

Galas DJ, Schmitz A (1978) DNAseI footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* 5:3157-3170.

Ghazalpour A, Bennett B, Petyuk V, Orozco L, Hagopian R, Mungrue I, Farber C, Sinsheimer J, Kang H, Furlotte N, et al. (2011) Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *PLoS Genet* 7:e1001393.

Giresi P, Kim J, Mcdaniell R, Iyer V, Lieb J (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 17:877-885.

Golding GB, Dean AM (1998) The Structural Basis of Molecular Adaptation. *Mol Biol Evol* 5:355–369.

Gossett AJ, Lieb JD (2012) In Vivo Effects of Histone H3 Depletion on Nucleosome Occupancy and Position in Saccharomyces cerevisiae. *PLoS Genet* 8:e1002771.

Gross DS, Adams CC, Lee S, Stentz B (1993) A critical role for heat shock transcription factor in establishing a nucleosome-free region over the TATA-initiation site of the yeast HSP82 heat shock gene. *EMBO* 12:3931-45.

Guenther CA, Tasic B, Luo L, Bedell MA, Kingsley DM (2014) A molecular basis for classic blond hair color in Europeans. Nature 46:748-752.

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 66:1669–1679.

Han M, Grunstein M (1988) Nucleosome loss activates yeast downstream promoters in vivo. *Cell* 55:1137-45.

Harms MJ, Thornton JW (2010) Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* 20:360-6.

Harris RS (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Hayes A, Zhang N, Wu J, Butler PR, Hauser NC, Hoheisel JD, Lim FL, Sharrocks AD, Oliver S (2002) Hybridization array technology coupled with chemostat culture: Tools to interrogate gene expression in Saccharomyces cerevisiae. *Methods* 26:281-90.

He B, Holloway A, Maerkl SJ, Kreitman M (2011) Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules. PLoS Genet 7:e1002053.

Heffer A, Pick L (2013) Conservation and Variation in HoxGenes: How Insect Models Pioneered the Evo-Devo Field. *Annu Rev Entomol* 58:161-179.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 6:283-9.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106:9362-7.

Hinnebusch A, Natarajan K (2002) Gcn4p, a Master Regulator of Gene Expression, Is Controlled at Multiple Levels by Diverse Signals of Starvation and Stress. *Eukaryot Cell* 1:22-32.

Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61:995-1016.

Hogan GJ, Lee CK, Lieb JD (2006) Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. PLoS Genet 2:e158.

Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. *PLoS One* 4:e7767.

Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 296:1205-1214.

Jansen RC, Nap NJ (2001) Genetical genomics: the added value from segregation. *Trends Genet* 17:388-91.

Jiang C, Pugh B (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* 10:161-172.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178:1709-23.

Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV, et al. (2013) Extensive Variation in Chromatin States Across Humans. *Science* 342:750-2.

Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucl Acids Res* 31:3576-3579.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241-54.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Pääbo S (2005) Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees. *Science* 309:1850-1854.

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* 188:107-16.

Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46:310-315.

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. *Science* 265:2037-48.

Lappalainen T, Sammeth M, Friedländer MR, Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Science* 501:506-511.

Lecrenier N, Foury F (2000) New features of mitochondrial DNA replication system in yeast and man. *Gene* 246:37-48.

Lee K, Kim SC, Jung I, Kim K, Seo J, Lee HS, Bogu GK, Kim D, Lee S, Lee B, et al. (2013) Genetic landscape of open chromatin in yeast. *PLoS Genet* 9:e1003229.

Lee S, Dudley A, Drubin D, Silver P, Krogan NJ, Pe'er D, Koller D (2009) Learning a Prior on Regulatory Potential from eQTL Data. *PLoS Genet* 5:e1000358.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298:799-804.

Lee TI, Young RA (2000) Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* 34:77-137.

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* 39:1235-44.

Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* 3:e161.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/Map format and SAMtools. *Bioinformatics*. 25:2078-9.

Ling F, Shibata T (2002) Recombination-dependent mtDNA partitioning: in vivo role of Mhr1p to promote pairing of homologous DNA. *EMBO J.* 21:4730-40.

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

Liu H, Styles CA, Fink GR (1996) Saccharomyces cerevisiae S288C has a mutation in FLO8, a gene required for filamentous growth. *Genetics* 144:967-78.

Lockshon D, Zweifel SG, Freeman-Cook LL, Lorimer HE, Brewer BJ, Fangman WL (1995) A role for recombination junctions in the segregation of mitochondrial DNA in yeast. *Cell* 81:947-55.

Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, Toonen P, Schafer S, Hübner N, van Breukelen B, Mohammed S, et al. (2013) Quantitative and Qualitative Proteome Characteristics Extracted from In-Depth Integrated Genomics and Proteomics Analysis. *CellReports* 5:1469-1478.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337:1190-5.

McDaniell R, Lee B, Song L, Liu Z, Boyle A, Erdos M, Scott L, Morken M, Kucera K, Battenhouse A, et al. (2010) Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. *Science* 328:235-239.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351:652-4.

Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tárraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, et al. (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res* 8:W210-3.

Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L (2011) Mining for genotype-phenotype relations in Saccharomyces using partial least squares. *BMC Bioinformatics* 12: 318.

Merker R, Klein H (2002) HPR1 affects ribosomal DNA recombination and cell life span in Saccharomyces cerevisiae. *Mol Cell Biol* 22:421-429.

Mizuno A, Tabei H, Iwahuti M (2006) Characterization of low-acetic-acid-producing yeast isolated from 2-deoxyglucose-resistant mutants and its application to high-gravity brewing. *J Biosci Bioeng* 101:31-7.

Miele V, Vaillant C, D'aubenton-Carafa Y, Thermes C, Grange T (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucl Acids Res* 36:3746-3756.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB (2006) Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila. *PLoS Comput Biol* 2:e130.

Mou C, Pitel F, Gourichon D, Vignoles F, Tzika A, Tato P, Yu L, Burt D, Bed'hom B, Tixier-Boichard M, et al. (2011) Cryptic Patterning of Avian Skin Confers a Developmental Facility for Loss of Neck Feathering. *PLoS Biol* 9:e1001028.

Muller L, Lucas J, Georgianna D, Mccusker J (2011) Genome-wide association analysis of clinical vs. nonclinical origin provides insights into Saccharomyces cerevisiae pathogenesis. *Mol Ecol* 20:4085–4097.

Mustonen V, Lässig M (2005) Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proc Natl Acad Sci USA* 102:15936-41.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489:83-90.

Olsson M, Meadows J, Truvé K, Rosengren Pielberg G, Puppo F, Mauceli E, Quilez J, Tonomura N, Zanna G, Docampo M, et al. (2011) A Novel Unstable Duplication Upstream of HAS2 Predisposes to a Breed-Defining Skin Phenotype and a Periodic Fever Syndrome in Chinese Shar-Pei Dogs. *PLoS Genet* 7:e1001332.

Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S, Cooper GM, et al. (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 30:265-70.

Paul M, Riebler A, Bachmann LM, Rue H, Held L (2010) Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Stat Med* 29:1325-39.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J, Stephens M, Gilad Y, Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768-772.

Picotti P, Clément-Ziza M, Lam H, Campbell DS, Schmidt A, Deutsch EW, Röst H, Sun Z, Rinner O, Reiter L, et al. (2013) A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* 494:266-270.

Pollard K, Hubisz M, Rosenbloom K, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20:110-121.

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904-909.

Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, Davis RW, Esposito RE (2000) The core meiotic transcriptome in budding yeasts. *Nat Genet* 26:415-23.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. *Am J Hum Genet* 67:170-81.

Prud'homme B, Gompel N, Rokas A, Kassner V, Williams T, Yeh S, True JR, Carroll SB (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* 440:1050-1053.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suites of utilities for comparing genomic features. *Bioinformatics* 26:841-2.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raijman D, Shamir R, Tanay A (2008) Evolution and selection in yeast promoters: analyzing the combined effect of diverse transcription factor binding sites. *PLoS Comput Biol* 4:e7.

Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Mol Biol Evol* 13:735-48.

Rando O, Chang H (2009) Genome-Wide Views of Chromatin Structure. *Annu Rev Biochem* 78:245-271.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516-7.

Rockman MV, Kruglyak L (2009) Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet.* ***5***:e1000419.

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

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

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

Rose MD, Winston F, Hieter P (1990) Methods in yeast genetics: A laboratory course manual. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. 198 p.

Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, et al. (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* 3:511-518.

Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, Mayford M, Lockhart DJ, Barlow C (2000) Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc Natl Acad Sci USA* 97:11038-43.

Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature* 458:342-345.

Schmidt D, Wilson M, Ballester B, Schwalie P, Brown G, Marshall A, Kutter C, Watt S, Martinez-Jimenez C, Mackay S, et al. (2010) Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science* 328:1036-40.

Schmitt ME, Brown TA, Trumpower BL (1990) A rapid and simple method for preparation of RNA from Saccharomyces cerevisiae. *Nucleic Acids Res* 18:3091-2.

Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore I, Wang J, Widom J (2006) A genomic code for nucleosome positioning. *Nature* 442:772-778.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, Schluter D, Kingsley DM (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 428:717-23.

Shibata Y, Sheffield N, Fedrigo O, Babbitt C, Wortham M, Tewari A, London D, Song L, Lee B, Iyer VR, et al. (2012) Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection. *PLoS Genet* 8:e1002789

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm and yeast genomes. *Genome Res* 15:1034-50.

Simon JM, Giresi PG, Davis IJ, Lieb JD (2013) A detailed protocol for formaldehyde-assisted isolation of regulatory elements (FAIRE). *Curr Protoc Mol Biol* Chapter 21:Unit21.26.

Skelly D, Johansson M, Madeoy J, Wakefield J, Akey J (2011) A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res* 21:1728-1737.

Skelly D, Ronald J, Akey J (2009) Inherited Variation in Gene Expression. *Annu Rev Genom Human Genet* 10:313-332.

Skelly D, Merrihew G, Riffle M, Connelly C, Kerr E, Johansson M, Jaschob D, Graczyk B, Shulman N, Wakefield J, et al. (2013) Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res* 23:1496-1504.

Stanbrough M, Rowen DW, Magasanik B (1995) Role of the GATA factors Gln3p and Nil1p of Saccharomyces cerevisiae in the expression of nitrogen-regulated genes. *Proc Natl Acad Sci USA* 92:9450-4.

Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB, Thurman RE, Sandstrom R, et al. (2013) Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* 154:888-903.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100:9440-5.

Stormo GD (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16:16-23.

Stuart GW, Moffett K, Baker S (2002) Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics* 18:100-8.

Tao H, Cox D, Frazer K (2006) Allele-Specific KRT1 Expression Is a Complex Trait. *PLoS Genet* 2:e93.

Thompson DA, Regev A (2009) Fungal regulatory evolution: cis and trans in the balance. *FEBS Letters* 583:3959-3965.

Threadgill D, Hunter K, Williams R (2002) Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm Genome* 13:175-178.

Tirosh I, Barkai N (2008) Two strategies for gene regulation by promoter nucleosomes. *Genome Res* 18:1084-91.

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

Tirosh I, Weinberger A, Carmi M, Barkai N (2006) A genetic signature of interspecies variations in gene expression. *Nat Genet* 38:830-4.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genet* 39:31–40.

Tournamille C, Colin Y, Cartron JP, Le Van Kim C (1995) Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nature Genet* 10:224–228.

Tsankov AM, Thompson DA, Socha A, Regev A, Rando O (2010) The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol* 8:e1000414.

Van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347:1999-2009.

Vavouri T, Elgar G (2005) Prediction of cis-regulatory elements using binding site matrices--the successes, the failures and the reasons for both. *Curr Opin Genet Dev* 15:395-402.

Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, Stamatoyannopoulos JA, Akey JM (2012) Personal and population genomics of human regulatory variation. *Genome Res* 22:1689-97.

Visel A, Rubin EM, Pennacchio LA (2009) Genomic views of distant-acting enhancers. *Nature* 461:199-205.

Wapinski I, Pfiffner J, French C, Socha A, Thompson DA, Regev A (2010) Gene duplication and the evolution of ribosomal protein gene regulation in yeast. *Proc Natl Acad Sci USA* 107:5505-10.

Ward LD, Kellis M (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* 30:1095-1106.

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, Fisher EM, Tavaré S, Odom DT (2008) Species-specific transcription in mice carrying human chromosome 21. *Science* 322:434-8.

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. *Nature* 430:85-8.

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

Wittkopp PJ, True JR, Carroll SB (2002) Reciprocal functions of the Drosophila yellow and ebony proteins in the development and evolution of pigment patterns. *Development* 129:1849-58.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 8:206-16.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434:338-345.

Yang J, Zaitlen NA, Goddard M, Visscher P, Price A (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* 46:100-106.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 29:75-8.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203-208.

Yuan GC, Liu JS (2008) Genomic Sequence Is Highly Predictive of Local Nucleosome Depletion. *PLoS Comput Biol* 4:e13.

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

Zelenaya-Troitskaya O, Newman SM, Okamoto K, Perlman PS, Butow RA (1998) Functions of the High Mobility Group Protein, Abf2p, in Mitochondrial DNA Segregation, Recombination and Copy Number in *Saccharomyces cerevisiae*. *Genetics* 148:1763–1776.

Zhang K, Lin L (2003) HaploBlockFinder: haplotype block analyses. *Bioinformatics* 19:1300-1301.

Zhang Z, Dietrich FS (2005) Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. *Nucleic Acids Res* 33:2838-51.

Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M (2010) Genetic analysis of variation in transcription factor binding in yeast. *Nature* 464:1187-1191.

Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner R, Schadt E (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* 40:854-861.

Appendix A- Supplementary Material for Chapter 2

A.1. Tables

Table A.1. mtDNA Copy Number in SGRP Strains

	mtDNA Relative Copy Number	
Strain	Replicate 1	Replicate 2
BY	1.00	1.00
RM	2.22	2.08
DBVPG6765	1.88	1.44
SK1	2.01	1.96
DBVPG6044	1.48	1.60
DBVPG1788	1.67	1.71
DBVPG1373	2.17	2.69
DBVPG1853	2.45	2.40
Y55	1.44	1.61
YPS128	1.18	1.38
DBVPG1106	1.64	1.59
DBVPG6040	1.32	1.44
YIIc17_E5	0.64	1.27
BC187	2.41	4.15
YPS606	1.39	1.57
L-1374	2.36	1.77
L-1528	1.99	2.11
NCYC110	1.89	2.65
NCYC361	1.71	1.56
K11	1.97	2.13
Y9	1.45	1.47
Y12	1.94	2.22
YS2	2.06	2.17
YS4	1.92	3.10
YS9	1.36	1.21
UWOPS83-787.3	1.32	1.57
UWOPS03-461.4	1.24	1.32
UWOPS05-217.3	1.58	3.19
UWOPS05-227.2	1.63	1.41
W303	1.40	1.51
322134S	2.78	2.41
378604X	2.10	1.51
273614N	2.38	2.22
YJM978	2.17	NA
YJM981	2.53	NA
YJM975	2.64	NA

Table A.2. mtDNA Copy Number at the Most Highly Associated SNP

Strain	mtDNA Copy Number	Allele at the Most Highly Associated SNP
273614N	2.3013	C
322134S	2.5985	C
378604X	1.8042	C
BC187	3.2815	C
DBVPG1106	1.612	C
DBVPG1373	2.428	C
DBVPG1788	1.6925	C
DBVPG1853	2.4258	C
DBVPG6040	1.3804	C
DBVPG6044	1.5387	C
DBVPG6765	1.6612	C
K11	2.0528	C
L_1374	2.0679	C
L_1528	2.0512	C
NCYC110	2.2716	C
RM11_1A	2.1507	C
SK1	1.9893	C
Y12	2.0785	C
Y55	1.5246	C
YJM975	2.4028	C
YJM978	2.1657	C
YJM981	2.5284	C
YS2	2.1183	C
YS4	2.5073	C
NCYC361	1.6346	T
REF	1	T
UWOPS03_461_4	1.2815	T
UWOPS05_217_3	2.3843	T
UWOPS05_227_2	1.5186	T
UWOPS83_787_3	1.4444	T
W303	1.4551	T
Y9	1.4579	T
YIIc17_E5	0.9506	T
YPS128	1.2793	T
YPS606	1.4761	T
YS9	1.2826	T

Appendix B- Supplementary Material for Chapter 3

B.1 Figures

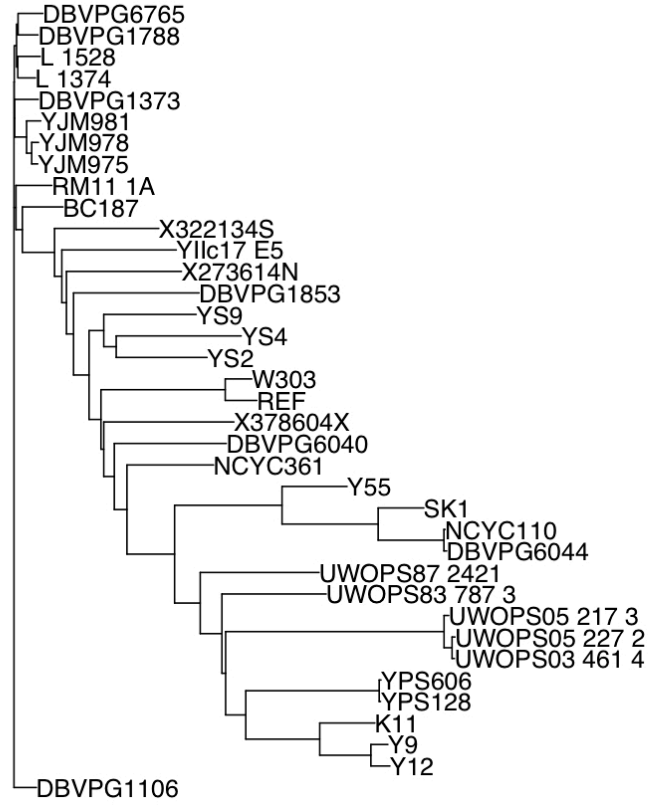

Figure B.1. Genome-wide phylogeny of the 37 strains. Neighbor-joining tree showing the 37 strains used in this paper.

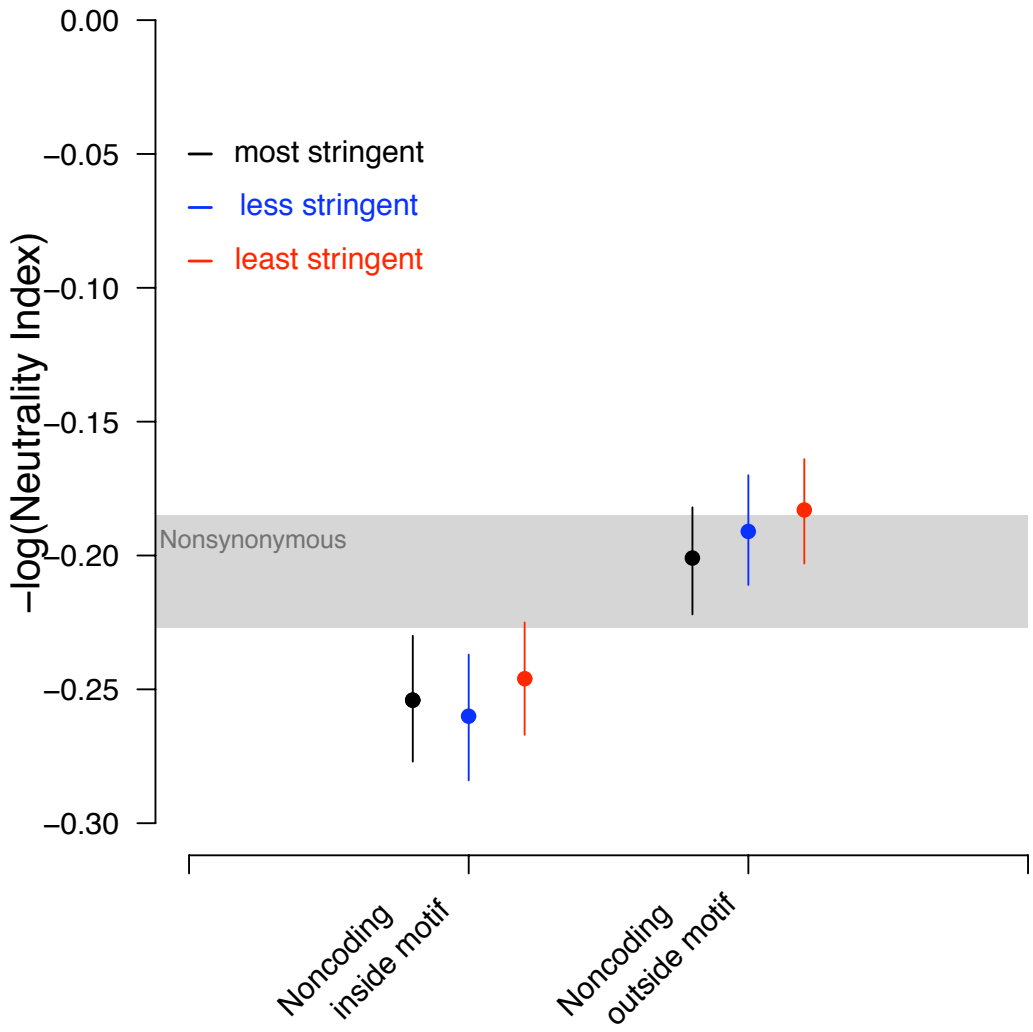

Figure B.2. Evolutionary pressures at noncoding sites using varying cutoffs for motif calling. –log(Neutrality Index) scores for two classes of sites (noncoding sites falling within predicted motifs and noncoding sites falling outside predicted motifs) are plotted. –log(NI) scores were obtained by summing information across all sites of a particular class and using synonymous sites within genes as putatively neutral sites. Confidence intervals were obtained by bootstrapping (see Methods). For noncoding sites, three different cutoffs were used for calling motifs. The most stringent cutoff results are colored black, a less stringent cutoff is colored blue, and the most stringent cutoff is colored red (see Methods).

B.2 Tables

Table B.1. Highly differentiated intergenic regions with more than 10 total motifs in the region.

Chr	Start	End	Flanking genes	Pairwise # of Motif Difs / Total motifs	Total motifs calls in region
15	516842	517642	YOR103C, YOR104W	0.298	27
6	68696	69112	-	0.268	22
7	11730	12480	-	0.253	19
4	524711	525437	YDR036C, YDR037W	0.252	26
16	752917	753298	YPR114W	0.223	19
4	1496541	1496783	YDR529C	0.222	13
4	371758	372244	YDL045W-A	0.221	15
3	303027	304357	YCR101C	0.219	48
12	243350	243886	-	0.219	18
15	24293	25271	YOL157C, YOL156W	0.219	24
15	266170	266264	-	0.218	11
7	323233	325333	YGL096W	0.212	43
15	160594	162355	YOL086C, YOL084W	0.211	11
10	67568	67848	-	0.210	13
2	398271	398607	YBR079C	0.209	13
15	388684	389212	YOR031W	0.208	22
5	560360	561699	-	0.207	43
16	26064	26610	-	0.206	18
2	366598	366967	YBR062C	0.205	14
11	218861	219967	YKL119C, YKL117W	0.205	13
3	266843	267430	YCR089W	0.198	13
5	562620	566224	YER186C, YER187W	0.198	58
12	830363	831114	YLR351C, YLR352W	0.197	26
4	292140	292780	YDL092W	0.197	18
16	225740	226167	YPL172C	0.194	14
5	313494	314529	YER076C	0.193	21
9	268472	268649	YIL046W	0.193	12
16	63860	64976	-	0.192	34
7	882236	882816	-	0.191	21

12	1060885	1062916	YLR460C, YLR461W	0.190	18
10	357922	358293	-	0.190	11
7	98589	98972	YGL207W	0.188	16
13	183363	190243	-	0.188	35
9	382379	382624	YIR016W	0.187	13
12	130612	131203	YLL010C	0.187	11
4	743872	744308	YDR143C	0.185	11
7	144054	144813	YGL191W	0.185	26
11	478475	478876	YKR021W	0.183	11
14	572317	573000	YNL033W	0.183	21
13	494494	494997	YMR112C, YMR113W	0.181	20
2	504281	504847	YBR133C, YBR135W	0.178	13
15	1070239	1071790	YOR387C	0.178	43
12	588920	589355	YLR225C, YLR226W	0.177	14
7	607147	607566	YGR059W	0.177	19
4	967819	968129	YDR255C	0.177	11
4	392055	392656	YDL035C	0.173	15
3	305464	307797	YCR102C, YCR104W	0.171	94
11	182430	182962	YKL139W	0.171	12
13	37647	38195	YML116W	0.170	19
12	817761	819311	YLR344W	0.169	17
14	743541	744360	YNR061C	0.169	23
9	131662	132240	YIL121W	0.168	17
15	277605	278056	YOL023W	0.167	13
16	76239	76668	YPL249C-A	0.167	12
14	573855	574507	YNL032W	0.166	17
1	196179	203393	YAR050W	0.166	134
15	845791	846267	YOR280C	0.165	15
4	979206	993130	YDR261C, YDR262W	0.165	37
7	996234	996873	YGR252W	0.165	15
5	14415	16354	YEL071W	0.165	44
14	591162	591428	YNL023C	0.165	11
5	68792	69756	YEL046C, YEL044W	0.164	30
12	490595	491867	-	0.164	34
1	57386	57518	YAL046C, YAL044W-A	0.162	11
2	90224	90738	YBL069W	0.162	16

11	489298	491006	YKR026C, YKR027W	0.162	42
4	512107	520513	YDR034C	0.161	58
13	86739	87122	YML092C	0.160	12
15	165331	165713	YOL083W	0.159	13
4	1058811	1059623	YDR298C, YDR299W	0.159	21
4	1196256	1196671	YDR361C	0.158	12
14	341970	342517	YNL156C, YNL155W	0.157	19
4	321552	322225	YDL076C, YDL075W	0.157	18
2	453218	453786	-	0.155	11
7	7090	8469	YGL259W	0.153	12
14	28346	28737	YNL326C	0.153	12
7	440430	441285	-	0.153	13
5	549512	549718	-	0.153	17
13	330230	330791	YMR029C, YMR030W	0.152	18
4	694173	694697	YDR122W	0.151	18
14	599232	599937	YNL019C	0.151	26
1	27969	31567	YAL063C, YAL062W	0.151	42
12	539591	540010	YLR192C	0.150	14
1	151168	152258	YAL001C, YAR002W	0.149	27
12	416659	417006	YLR136C, YLR137W	0.149	14
4	453648	454119	YDR003W	0.149	15
16	347388	348442	YPL109C, YPL108W	0.148	37
15	514279	515244	YOR100C, YOR101W	0.147	25
9	385698	389568	YIR018C-A	0.147	132
12	897672	898650	YLR387C, YLR388W	0.147	13
7	858609	859067	-	0.147	12
7	139967	140373	YGL194C-A	0.147	11
11	171134	171787	YKL148C, YKL146W	0.147	23
4	1224387	1224749	YDR374C, YDR374W-A	0.147	13
7	9395	11109	YGL258W	0.146	60
4	1256840	1257350	YDR390C	0.145	11
10	26086	26886	YJL216C,	0.144	29

			YJL214W		
14	745344	746943	YNR062C, YNR063W	0.144	42
16	308220	308826	YPL128C	0.143	19
4	1502153	1503305	YDR533C	0.143	31
15	1052928	1055542	YOR381W	0.143	62
10	638965	639931	YJR113C, YJR115W	0.143	34
16	359403	360205	YPL103C, YPL101W	0.143	15
13	208860	209524	YML035C, YML034W	0.143	28
14	405566	406359	YNL118C, YNL117W	0.143	13
4	892490	892872	YDR214W	0.142	12
7	598625	599420	YGR055W	0.142	20
12	36360	37331	YLL052C	0.141	22

Table B.2. Motifs under purifying selection.

Motif Name	$-\log_{10}$(Neutrality Index)
ABF1	-0.32 (-0.49 - -0.10)
ABF2	-0.29 (-0.42 - -0.14)
ACE2	-0.32 (-0.37 - -0.26)
ADR1	-0.45 (-0.52 - -0.37)
AFT2	-0.49 (-0.62 - -0.33)
ARG80	-0.34 (-0.40 - -0.27)
ARG81	-0.45 (-0.58 - -0.28)
ARR1	-0.27 (-0.37 - -0.17)
ASG1	-0.27 (-0.37 - -0.17)
AZF1	-0.40 (-0.48 - -0.32)
CAT8	-0.32 (-0.39 - -0.25)
CBF1	-0.69 (-0.96 - -0.35)
CEP3	-0.22 (-0.39 - -0.03)
CHA4	-0.40 (-0.55 - -0.19)
CRZ1	-0.32 (-0.42 - -0.20)
CST6	-0.32 (-0.47 - -0.15)
CUP9	-0.47 (-0.62 - -0.32)
DAL82	-0.49 (-0.64 - -0.33)
DOT6	-0.80 (-1.08 - -0.48)
ECM22	-0.17 (-0.31 - -0.02)
ECM23	-0.25 (-0.38 - -0.11)
EDS1	-0.44 (-0.58 - -0.28)
FHL1	-0.35 (-0.47 - -0.20)
FKH1	-0.73 (-1.13 - -0.26)
GAT1	-0.22 (-0.34 - -0.05)
GAT3	-0.38 (-0.48 - -0.27)
GAT4	-0.40 (-0.51 - -0.28)
GCR1	-0.40 (-0.56 - -0.20)
GCR2	-0.29 (-0.37 - -0.21)
GIS1	-0.47 (-0.56 - -0.38)
GLN3	-0.35 (-0.38 - -0.31)
GZF3	-0.25 (-0.34 - -0.15)
HAC1	-0.47 (-0.66 - -0.21)
HAL9	-0.36 (-0.40 - -0.32)
HAP1	-0.28 (-0.47 - -0.05)
HAP2	-0.36 (-0.39 - -0.32)
HCM1	-0.33 (-0.39 - -0.26)
HMRA2	-0.25 (-0.34 - -0.16)
HSF1	-0.44 (-0.57 - -0.29)
LEU3	-0.55 (-0.81 - -0.25)
LYS14	-0.48 (-0.68 - -0.23)
MAC1	-0.19 (-0.32 - -0.04)
MATA1	-0.43 (-0.50 - -0.35)

MATALPHA2	-0.46 (-0.60 - -0.28)
MBP1	-0.33 (-0.40 - -0.25)
MBP1::SWI6	-0.42 (-0.54 - -0.28)
MCM1	-0.35 (-0.55 - -0.10)
MET31	-0.51 (-0.69 - -0.27)
MET32	-0.38 (-0.48 - -0.27)
MET4	-0.27 (-0.43 - -0.08)
MIG1	-0.54 (-0.64 - -0.44)
MIG2	-0.50 (-0.64 - -0.35)
MIG3	-0.49 (-0.59 - -0.39)
MOT3	-0.35 (-0.38 - -0.31)
MSN2	-0.52 (-0.58 - -0.46)
NHP10	-0.54 (-0.80 - -0.23)
NHP6A	-0.31 (-0.54 - -0.02)
OPI1	-0.36 (-0.47 - -0.24)
PDR1	-0.33 (-0.49 - -0.14)
PHD1	-0.31 (-0.49 - -0.11)
PHO2	-0.38 (-0.42 - -0.35)
PUT3	-0.32 (-0.50 - -0.09)
RDR1	-0.40 (-0.55 - -0.24)
RDS1	-0.26 (-0.41 - -0.06)
RDS2	-0.36 (-0.52 - -0.16)
REB1	-0.60 (-0.81 - -0.36)
REI1	-0.43 (-0.60 - -0.23)
RFX1	-0.39 (-0.51 - -0.24)
RGT1	-0.36 (-0.57 - -0.12)
RIM101	-0.37 (-0.47 - -0.26)
RME1	-0.40 (-0.55 - -0.21)
RPH1	-0.46 (-0.63 - -0.25)
RSC3	-0.36 (-0.50 - -0.20)
RSC30	-0.76 (-0.98 - -0.51)
RTG3	-0.50 (-0.70 - -0.24)
SFP1	-0.88 (-1.11 - -0.48)
SIP4	-0.30 (-0.43 - -0.17)
SKN7	-0.32 (-0.38 - -0.26)
SPT2	-0.30 (-0.37 - -0.24)
SPT23	-0.25 (-0.32 - -0.18)
SRD1	-0.45 (-0.55 - -0.33)
STB4	-0.26 (-0.38 - -0.14)
STB5	-0.34 (-0.45 - -0.23)
STE12	-0.32 (-0.39 - -0.24)
STP1	-0.30 (-0.44 - -0.15)
STP2	-0.29 (-0.49 - -0.03)
STP3	-0.44 (-0.62 - -0.23)
STP4	-0.34 (-0.54 - -0.11)
SUM1	-0.50 (-0.55 - -0.45)

SWI5	-0.32 (-0.40 - -0.23)
TBF1	-0.48 (-0.60 - -0.37)
TOS8	-0.39 (-0.49 - -0.29)
UPC2	-0.28 (-0.33 - -0.23)
XBP1	-0.27 (-0.36 - -0.16)
YAP3	-0.27 (-0.38 - -0.15)
YAP5	-0.37 (-0.42 - -0.30)
YBR239C	-0.20 (-0.35 - -0.02)
YDR026C	-0.75 (-1.05 - -0.41)
YDR520C	-0.45 (-0.69 - -0.17)
YER130C	-0.42 (-0.53 - -0.30)
YER184C	-0.34 (-0.46 - -0.21)
YGR067C	-0.36 (-0.55 - -0.12)
YKL222C	-0.32 (-0.47 - -0.16)
YLL054C	-0.31 (-0.47 - -0.14)
YLR278C	-0.44 (-0.63 - -0.21)
YML081W	-0.49 (-0.62 - -0.35)
YNR063W	-0.24 (-0.41 - -0.03)
YOX1	-0.36 (-0.47 - -0.22)
YPR013C	-0.43 (-0.57 - -0.26)
YPR022C	-0.53 (-0.69 - -0.35)
YRM1	-0.45 (-0.53 - -0.35)
ZMS1	-0.55 (-0.74 - -0.32)

Table B.3. Regions significant for being under purifying selection by the MK test.

Chr	Start	Stop	Flanking gene(s)	D(motif)	P(motif)	D(synonymous)	P(synonymous)	P-value (Bonferonni corrected)
4	1278430	1279202	YDR406W	0	7	281	7	4.17×10^{-7}
4	720301	723001	YDR132C	56	13	471	7	1.16×10^{-4}
2	680557	682173	-	0	10	61	9	2.45×10^{-4}
12	390271	393484	YLR121C, YLR125W	66	51	144	25	3.48×10^{-4}
4	1164655	1167208	YDR345C	45	15	176	3	4.36×10^{-4}
4	892490	892872	YDR214W	8	8	264	10	1.68×10^{-3}
13	897602	898403	YMR311C, YMR312W	16	12	107	6	1.47×10^{-2}
15	1047801	1049508	YOR378W	0	8	114	27	2.06×10^{-2}
4	1055889	1056547	YDR297W	8	7	129	4	2.82×10^{-2}
4	1379589	1380047	-	0	3	595	10	3.35×10^{-2}
4	1509706	1510892	YDR538W	28	8	179	2	4.30×10^{-2}

Table B.4. Power to detect associations as a function of effect size in our data set.

Effect size (% variance explained)	Significance threshold (uncorrected *p*-value)	Power
	0.05	0.788
0.25	1×10^{-5}	0.022
	0.05	0.986
0.5	1×10^{-5}	0.389
	0.05	1
0.75	1×10^{-5}	0.925

Appendix C- Supplementary Material for Chapter 4

C.1 Figures

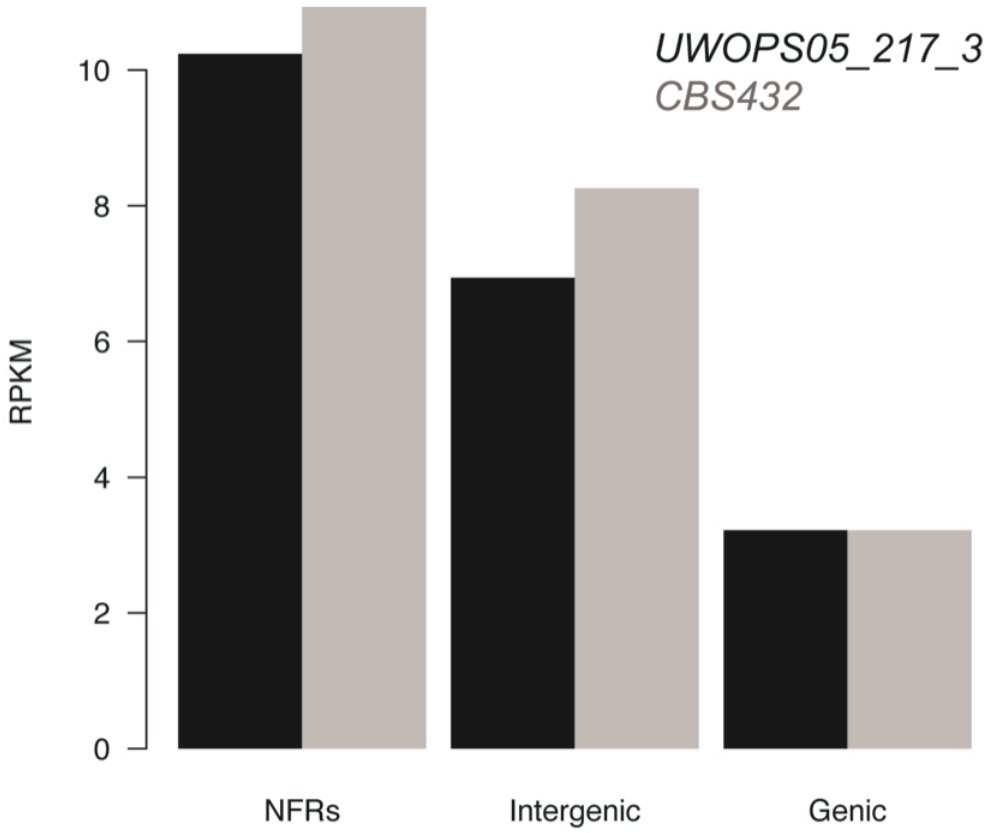

Figure C.1. Enrichment of FAIRE signal in NFRs and intergenic regions. RPKM for *the S. cerevisiae* strain UWOP05_217_3 and the *S. paradoxus* strains CBS432 is shown in three types of regions, nucleosome-free regions (NFRs), intergenic regions, and genic regions.

C.2 Tables

Table C.1. Power and false positive rate for *cis* and *trans* tests.

Type of test	Effect size	Power (at posterior probability=0.95)	False positive rate (at posterior probability=0.95)
cis	0.1	0.32	0.05
cis	0.5	0.92	0.05
cis	0.8	0.97	0.05
trans	0.1	0.18	0.05
trans	0.5	0.81	0.05
trans	0.8	0.91	0.05

Table C.2. Summary of different criteria used to investigate the relationship between chromatin and gene expression QTL.

Criteria for calling *cis* effects	***cis* effect on NFR**		**No *cis* effect on NFR**		**OR**	***p-value***
	cis effect on RNA	No *cis* effect on RNA	*cis* effect on RNA	No *cis* effect on RNA		
PP > 0.95	1790	358	971	165	0.85	0.12
PP > 0.95 & $\log_2$(magnitude of NFR effect) > 1	526	122	2235	401	0.77	0.03
PP > 0.95 & $\log_2$(magnitude of NFR and RNA effect) >1	185	1312	208	1579	1.07	0.55
PP > 0.998	1125	459	1244	459	0.90	0.17
PP > 0.998 & $\log_2$(magnitude of NFR effect) > 1	387	166	1982	749	0.88	0.23
PP > 0.998 & $\log_2$(magnitude of NFR and RNA effect) >1	136	1073	230	1875	1.02	0.91
PP > 0.95 & $\log_2$(*trans* effect) <0.10 for RNA	466	1682	219	917	1.16	0.11

Notes: PP and OR denote posterior probability and odds ratio, respectively.