Patterns and determinants of variation in functional genomics phenotypes in the yeast *Saccharomyces cerevisiae*

Daniel A. Skelly

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Joshua M. Akey, Chair

Maitreya J. Dunham

Philip Green

Program Authorized to Offer Degree:
Department of Genome Sciences

University of Washington

Abstract

Patterns and determinants of variation in functional genomics phenotypes in the yeast *Saccharomyces cerevisiae*

Daniel A. Skelly

Chair of the Supervisory Committee:
Associate Professor Joshua M. Akey
Department of Genome Sciences

Phenotypic variation among individuals within populations is ubiquitous in the natural world, and a preeminent challenge in biology is understanding the contribution of genetic variation to this phenotypic variation. Despite technological advances in the development of genome-scale methods for querying molecular phenotypes, our understanding of the molecular basis of morphological and physiological variation remains rudimentary. In this dissertation, I outline computational methods I have developed and analyses I have conducted in the yeast *Saccharomyces cerevisiae* to make inferences about the relationship between DNA sequences and the molecular phenotypes to which they give rise. First, I describe a population genomics study of a class of genomic elements, intron splice sequences, in a diverse set of complete *S. cerevisiae* genomes. I obtained quantitative estimates of the strength of purifying selection acting on these sequences, and present analyses suggesting that introns in some subsets of genes are actively maintained in natural populations of *S. cerevisiae*. Next, I shift my focus to the genetic basis of variation in a particular molecular phenotype, gene expression. I examine genes that show allele-specific expression (ASE) due to *cis*-regulatory variation, and present a Bayesian statistical model for quantifying ASE measured by RNA-Seq. A novel feature of this model is the ability to detect variable ASE, where the level of ASE differs across a transcript, as can occur in the case of variations in transcript structure. Finally, I explore molecular phenotypic variation more comprehen-

sively, presenting results of an analysis of deeply phenotyped *S. cerevisiae* strains. I analyze genome sequence, gene expression, protein abundance, metabolite abundance, and cellular morphological phenotypes in this phenomics study. I identify abundant natural variation across all phenotypic classes, pinpoint loci that act in *cis* to affect RNA and protein levels, and provide initial clues as to the predictability of phenotypic traits that vary between individuals within a species. I conclude by discussing the need for new statistical models to make use of the rich information contained in functional genomics datasets and the necessity of considering environmental context when disentangling the functional consequences of genetic variation.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ASE: allele-specific gene expression

BY: *S. cerevisiae* strain BY4716, a laboratory strain

FDR: false discovery rate

MCMC: Markov chain Monte Carlo

PSSM: position-specific scoring matrix

QTL: quantitative trait locus

RM: *S. cerevisiae* strain RM11-1a, a wild vineyard strain

SGD: *Saccharomyces* Genome Database

UTR: untranslated region

ACKNOWLEDGMENTS

I would like to thank Gwen Ebert and Leslie Kohlberg for helping me to recognize my strengths, and Maria Klapa for giving me my first exposure to scientific research. I thank Carol Eunmi Lee for teaching me a great deal about the practice of scientific research and inspiring me to attend graduate school. I am grateful to Michael Newton and David Schwartz for piquing my interest in computational and statistical genomics.

As a graduate student, I have been fortunate to have a thoughtful and helpful committee: Maitreya Dunham, Phil Green, Joseph Felsenstein, and Matt Kaeberlein. I thank the many colleagues with whom I have interacted as part of the Yeast Resource Center, especially Trisha Davis, Eric Muller, William Noble, Sara Cooper, Michael MacCoss, Gennifer Merrihew, Michael Riffle, and Daniel Jaschob. Chapter 4 of this thesis would not exist without their contributions, ideas, and encouragement. It has been a privilege to work with Jon Wakefield, who has been an ideal collaborator in all respects. His patience has been admirable and his humor has been a welcome companion to all of our interactions.

The Akey lab has been a fantastic place to grow as a graduate student. I especially thank Caitlin Connelly, Marnie Johansson, Jennifer Madeoy, and James Ronald for their assistance with experiments and analyses that contributed to Chapters 2, 3, and 4. I am incredibly grateful to Joshua Akey for his endless enthusiasm, patience, and committment to my personal development as a scientist, and for infusing his lab with a spirit of camaraderie, openness, and humor. He has taught me much about what it takes to be a great scientist.

I would like to thank my parents and my sister, Brenna. They have been incredibly supportive and have always believed in me. Finally, I wish to thank my wife Beth. I will always be grateful for her commendable patience during the most intense periods of my graduate work. She has given me unconditional love, and I look forward to continuing our journey through life together.

Chapter 1

INTRODUCTION

1.1 The genetic basis of phenotypic variation

Individual members of a species differ in myriad ways, including morphology, physiology, and disease susceptibility. A fundamental challenge in biology is to understand the contribution of DNA sequence variation to variation in the observable characteristics of an organism. In recent years, dramatic improvements in DNA sequencing technology and precipitous drops in sequencing costs have made possible detailed studies of the amount and distribution of polymorphism within species [1–5]. However, our understanding of the functional effects of sequence variation and the molecular basis of morphological and physiological variation remains rudimentary.

There are many examples of traits with well-understood genetic causes. For example, in humans, laborious gene mapping studies yielded the genetic loci responsible for cystic fibrosis [6] and Huntington's disease [7]. More recently, advances in DNA sequencing technology have facilitated the identification of genes underlying rare Mendelian syndromes of unknown cause [8]. Causal genes or strong candidate loci have also been identified for numerous interesting traits in model organisms, such as coat color in mice [9], skeletal morphology in stickleback fish [10], and inflorescence architecture in maize [11]. A unifying feature of these traits is their genetic simplicity; phenotypic variation within the population can be divided into qualitative categories rather than spanning a continuous range.

In contrast to these genetically simple traits, most commonly observable phenotypic traits are quantitative, and variation within the population is continuous and not easily divided into distinct categories [12]. Examples of such traits include human height [13], yeast sporulation efficiency [14], and *Drosophila* sensory bristle number [15]. For these genetically complex traits, multiple quantitative trait loci (QTL) act in concert with the environment to produce the trait. Although genetic linkage analysis has been a highly successful strategy

for discovering large-effect loci that contribute to phenotypic variation, it is less likely to be fruitful in the case of quantitative traits with genetically complex underpinnings [16]. Recently, genome-wide association studies (GWAS) have begun to make inroads on this front and have identified thousands of loci where common polymorphisms contribute to variation in quantitative traits [17]; however, these studies suffer from weaknesses of their own, with little mechanistic information accompanying loci identified and the potential for false positives due to population stratification.

1.2 Biomolecules as phenotypes

A complementary strategy to GWAS is to directly study the molecular phenotypes that mediate the conversion of heritable genetic variation into observable phenotypic differences. Such molecular phenotypes include transcript abundance and structure; protein abundance, isoforms, and modifications; metabolite presence and abundance; and the activity of these biomolecules in the context of the networks through which they interact. In this dissertation, I will refer to these measurable quantities as *molecular phenotypes* or *functional genomics phenotypes*. The term "functional genomics" is appropriate because it is often used to describe how a static genome gives rise to dynamic molecules that interact to determine genomic function. I contrast these molecular phenotypes with phenotypes that fall under the more conventional usage of the term: those that are directly visible without the use of technical procedures, which I term *organismal phenotypes*. I note that this is not an absolute distinction, as many phenotypes do not fall clearly within either category or may manifest only under specific environmental conditions. Nevertheless, these terms will serve as convenient shorthand for distinguishing between the two broadly different classes of phenotypes.

Can variation in functional genomics phenotypes be linked to variation in organismal phenotypes? An increasing number of empirical studies highlight the pervasive phenotypic effects of regulatory variation, such as skeletal morphology in stickleback fish [18], beak morphology in Darwin finches [19], cuticular pigmentation in Drosophila [20], and muscle growth in pigs [21]. In humans, regulatory alleles have been linked to infectious, autoimmune, psychiatric, neoplastic, and neurodegenerative disease susceptibility [22–27]. Even

relatively modest expression changes can have significant biological consequences, as seen for the tumor suppressor gene *APC*, in which a 50% change in gene expression can lead to development of familial neoplasia [28].

Since proteins and metabolites typically have a more direct functional role than messenger RNA, which serves to transmit information, it is not surprising that there are examples of variation in proteins or metabolites that affect organismal phenotypes. Perhaps the best known are defects in individual proteins that are responsible for many well-studied Mendelian diseases, such as cystic fibrosis [6], Huntington's disease [7], and phenylketonuria [29]. Similarly, metabolites can serve as signatures of future diabetes, with a role that is likely causal in nature [30]. Thus, as expected given their role in activating the static instructions provided by the genome, functional genomics phenotypes including RNA, protein, and metabolite levels can all affect organismal phenotypes.

1.3 The genetic basis of variation in functional genomics phenotypes

In order to understand in detail the steps through which biomolecules mediate the formation of organismal phenotypes encoded by the genome, it is necessary to consider the genetic basis of variation in functional genomics phenotypes. Although several groups have studied the genetic determinants of protein and/or metabolite levels [31–34], the genetic architecture of transcriptional variation is best understood, and I will focus on insights from this line of inquiry.

One of the more surprising observations gleaned from studies of expression QTL is that the genetic architecture of transcriptional variation is complex. In the first empirical study of the genetics of global gene expression, $\approx 20\%$ of significantly varying genes had no linkage to any genomic locus, although the study was well-powered to identify QTL explaining a high fraction of transcript variation [35]. This observation suggests that polygenic control of transcript levels is common, and has been supported by additional studies [36, 37]; it follows that only a small minority of transcripts appears to show heritable variation that is due to a single expression QTL [38].

The architecture of transcriptional variation shows many additional hallmarks of genetic complexity, including epistasis, genotype-environment interaction, and pleiotropy. A

significant portion of gene expression variation results from interacting loci [39, 40], and transcript levels appear to be influenced by gene-environment interaction, although only a few environments and model organisms have been examined [41–43]. A few characteristics of variation in transcript abundances provide clues to the types of alleles that segregate in populations and contribute to such variation. For example, gene expression levels frequently exhibit transgressive segregation [38], the emergence of a phenotype in offspring that is more extreme than that observed in either parental line. This phenomenon is often attributed to the presence of epistatic interactions between alleles or the segregation of alleles of opposing additive effects in parental populations [44]. In sum, the genetic basis of variation in functional genomics phenotypes is complex, and will in many cases be difficult to unravel completely. For some such phenotypes it may be more productive to attempt to predict the phenotypic value using other phenotypes that covary in a similar fashion, without a complete understanding of the underlying genetic causes.

1.4 Measuring molecular phenotypes

Advances in technology in the past two decades have led to a precipitous drop in the cost of conducting DNA sequencing [45]. This development has opened up previously inaccessible avenues of research and given us an unprecedented window onto levels of genetic variation present in natural populations [1–5]. Simultaneously, a confluence of advances in instrumentation, computational resources, and algorithmic development have led to a similar revolution in the measurement of other molecular phenotypes, albeit one occurring at a slower pace.

Since the first application of DNA microarrays to measuring gene expression levels in 1995 [46], the accurate quantification of transcript abundance at a genome-wide level has been of great interest. Recent advances in DNA sequencing have led to the conception of RNA-Seq [47, 48], where RNA fragments are converted to cDNA and directly sequenced on a high-throughput DNA sequencing machine. The millions of short reads that result can be used to quantify transcript abundance. Among the advantages of RNA-Seq compared to microarray-based methods of transcript quantification are greater dynamic range, ability to distinguish transcript isoforms, and a precision that is, in theory, limited only by coverage

depth [49].

In addition to advances in genome-wide transcript quantification, a variety of technological developments have facilitated more extensive measurements of other biomolecules of interest, such as protein and metabolite levels. In particular, improvements in the accuracy and robustness of chromatography and mass spectrometry, as well as enhancements in algorithms used for processing output, are enabling us to measure proteins and metabolites in a label-free manner. In the case of proteins, genome sequences have provided us with a complete catalog of coding sequences, making possible a shotgun proteomics approach where a mixture of proteins is separated via liquid chromatography and subjected to tandem mass spectrometry. Subsequently, the chromatographic peak intensity of single precursor peptide ions can be compared between samples or treatments [50], providing a reproducible surrogate for relative protein level. The measurement of small molecule metabolites on a large scale is a less mature field, but continuous improvements in analysis methods are allowing for the quantitation of ever larger and more chemically diverse sets of compounds [51, 52].

1.5 Yeast as a model system for studying variation in molecular phenotypes

The budding yeast *Saccharomyces cerevisiae* has a storied history as one of the premier model organisms for cellular and molecular biology, primarily due to its ease of culturing and genetic manipulation. Since becoming the first eukaryote with a completely sequenced genome in 1996 [53], *S. cerevisiae* has also proven a useful model for the study of eukaryotic genome structure and evolution. Thus, yeast is a useful model system with several advantages: (1) a small, well-annotated genome; (2) assignment of function to a significant fraction of its genes; (3) a well-studied collection of metabolic pathways and protein complexes; and (4) a rich literature from which to draw. These advantages are particularly useful for interpreting high-dimensional datasets that consist of simultaneous measurements of thousands of molecular phenotypes.

In addition to our understanding of yeast biology gleaned primarily from a small number of common laboratory strains, recent studies have described phenotypic variation found in wild or "domesticated" strains [54]. Surveys of genome-wide polymorphism have revealed strong population structure among wild strains, with several well-defined, geographically

isolated lineages [1, 55]. Differences in growth rate, coloration, and freeze tolerance under particular environmental conditions vary between wild isolates, and this variation has been correlated with variation in gene expression levels [56]. Similarly, a selection of wild isolates showed abundant variation in stress sensitivity and gene expression when subjected to a panel of 14 different environmental conditions [57].

Finally, I note that yeast has been a testing ground for many genomic technologies that have eventually found widespread usage in other organisms. Early studies using microarrays [46], RNA-Seq [47], ChIP-chip [58], chromosome conformation capture [59], proteomics [60], and metabolomics [61] were all conducted in *S. cerevisiae*. Thus, yeast serves as an ideal model system for measuring and analyzing variation in functional genomics phenotypes.

1.6 Objectives

My thesis work has focused on developing computational methods for making inferences about the relationship between DNA sequences and the molecular phenotypes to which they give rise. Specifically, my objectives were:

1. Analyze the evolutionary forces governing patterns of DNA sequence variation among a class of genomic elements in a natural population. As a case study, I conducted a population genomics analysis of intron splicing in *S. cerevisiae*.

2. Investigate the genetic basis of variation in transcript abundance. As a first step toward better understanding the characteristics of loci that harbor *cis*-regulatory variation, I developed a statistical model for measuring allele-specific gene expression via RNA-Seq.

3. Examine the quantitative characteristics of molecular phenotypic diversity. I analyzed data from an extensively phenotyped set of *S. cerevisiae* strains to provide insight into the patterns and determinants of variation in RNA, protein, metabolite and morphological traits.

Chapter 2

POPULATION GENOMICS OF INTRON SPLICING IN *SACCHAROMYCES CEREVISIAE*

This chapter contains material published in [62].

2.1 Summary

Introns are a ubiquitous feature of eukaryotic genomes, and the dynamics of intron evolution between species has been extensively studied. However, comparatively few analyses have focused on the evolutionary forces shaping patterns of intron variation within species. To better understand the population genetic characteristics of introns, I performed an extensive population genetics analysis on key intron splice sequences obtained from 38 strains of *Saccharomyces cerevisiae*. As expected, I found that purifying selection is the dominant force governing intron splice sequence evolution in yeast, formally confirming that intron-containing alleles are a mutational liability. In addition, through extensive coalescent simulations, I obtained quantitative estimates of the strength of purifying selection ($2N_e s \approx 19$) and used diffusion approximations to provide insights into the evolutionary dynamics and sojourn times of newly arising splice sequence mutations in natural yeast populations. In contrast to previous functional studies, my evolutionary analyses comparing the prevalence of introns in essential and non-essential genes suggest that introns in non-ribosomal protein genes are functionally important and tend to be actively maintained in natural populations of *S. cerevisiae*. Finally, I demonstrate that heritable variation in splicing efficiency is common in intron-containing genes with splice sequence polymorphisms.

2.2 Introduction

A distinguishing feature of eukaryotic genomes is the presence of intervening nucleotides that interrupt protein-coding sequences. The majority of these introns are removed by the spliceosome, an ancient molecular machine that was likely present in the most recent com-

mon ancestor of all living eukaryotes [63–65]. The abundance of spliceosomal introns varies widely between taxa, from just a handful of introns in some protists [66] to hundreds of thousands of introns in vertebrates and plants. Intron sizes, too, vary over several orders of magnitude between species [67, 68]. Despite these profound differences in intron characteristics between taxa, all introns are governed by the same evolutionary forces that regulate genetic elements in any genome [69].

The budding yeast *Saccharomyces cerevisiae* is an important model system for examining the evolution of eukaryotic genomes. The introns of *S. cerevisiae* are unusual among eukaryotes in several respects. Although introns are still being discovered and characterized in this well-annotated genome [70–73], less than 10% of yeast genes contain introns. *S. cerevisiae* introns are small (typically < 600 bp) [74], and only a few yeast genes have been reported to undergo alternative splicing [70, 73, 75]. *S. cerevisiae* introns are characterized by highly conserved $5'$, $3'$, and branch point sequences. The first yeast introns discovered possessed splicing sequences that fit a strict consensus motif [76–78]. More recently, a limited number of introns with splice motifs that match a more relaxed consensus have been identified [71–73,75], although the information content of short yeast intron splice sequences tends to exceed that present in the short introns of a typical multicellular eukaryote [79]. Molecular studies have revealed that all positions in the $5'$, $3'$, and branch point sequences of yeast introns are likely to play some role in determining splicing efficiency, with a few positions especially critical for pre-mRNA splicing [78, 80, 81]. In particular, the first and second positions of the intron, the adenosine in the penultimate position of the branch point, and the AG terminating the intron appear to be necessary to achieve any appreciable level of proper splicing [81–84]. In addition, interdependencies between bases, even outside the conserved splice sites, can render the effects of mutations at some positions unpredictable [85].

The functional significance of introns in *S. cerevisiae* is poorly understood. The ancestor of extant fungi was likely intron-rich, with an estimated density of roughly four introns per kilobase [86]. It has been hypothesized that introns are on their way out of the yeast genome, with intron loss mediated by homologous recombination of reverse-transcribed cDNAs [87]. An implication of this model is that yeast introns are largely genomic relics unlikely to have

functional significance. In support of this hypothesis, there are many examples of introns that can be deleted from the genome without obvious phenotypic consequences, at least under standard laboratory conditions [88–90]. In contrast, while most yeast introns have no known functional importance, it is clear that some encode functional elements, such as snoRNAs [91] or promoters [92]. Moreover, yeast introns appear to play a more general role in the regulation of gene expression and protein production [93,94]. Introns can be involved in splicing autoregulation [95], transcriptional and translational enhancement [93, 96], and transcriptional response to environmental stresses [94]. Notably, perturbation of subtle layers of transcript regulation or systems for responding to environmental stimuli may not be detectable under standard laboratory conditions, but might be critical to organismal fitness in more challenging wild environments.

The evolutionary forces governing intron dynamics have been subject to considerable debate [97]. Evolutionary analyses of introns have surveyed a variety of phylogenetic depths, from kingdom [86,98,99] to subphylum [100,101]. These comparisons have revealed that intron gains and losses are common over long evolutionary timescales. Notably, however, there have been few studies examining the evolutionary forces shaping patterns of intron variation over shorter timescales [69, 102, 103]. Population genetic analyses of intron polymorphism are a powerful approach for exploring the evolutionary trajectory of polymorphisms within introns and the importance of introns as genomic elements.

Here, I describe a systematic population genomics analysis of intron splicing in yeast. Specifically, I analyzed patterns of polymorphism in key intron splice sequences in 38 strains of *S. cerevisiae* with fully sequenced genomes [1]. As expected, polymorphisms are rare in sequences important for pre-mRNA splicing in *S. cerevisiae*, consistent with the elimination of deleterious mutations by purifying selection. I performed extensive simulations using the ancestral selection graph [104] to derive quantitative estimates of the strength of purifying selection acting upon these critical intron splice sequences. I compare these estimates to the strength of selection acting on non-synonymous sites, and apply diffusion approximations to explore the evolutionary dynamics of splice sequence polymorphisms. The strong purifying selection I observe acting on intron splice sequences formally confirms that intron-containing alleles are a mutational liability [69] and renews questions about why introns exist in the

yeast genome. Additional analyses suggest that extant introns in yeast are not merely genomic relics, but that introns tend to be actively maintained in natural populations of *S. cerevisiae*.

2.3 Methods

2.3.1 Sequence data

I used complete haploid genome sequences for 35 *S. cerevisiae* strains sequenced and assembled as part of the Saccharomyces Genome Resequencing Project (`http://www.sanger.ac.uk/Teams/Team71/durbin/sgrp/`), along with the reference *S. cerevisiae* genome (October 2007 sequence; `http://www.yeastgenome.org/`) and two previously sequenced genomes, RM11-1a (`http://www.broad.mit.edu/annotation/genome/saccharomyces_cerevisiae`) and YJM789 [105]. I examined sequences annotated as spliceosomal introns in *Saccharomyces* Genome Database [106], excluding introns in dubious genes and introns lacking evidence of splicing in previous experimental studies [74, 75]. I also included spliceosomal introns deposited in the Yeast Intron Database [74] and reported in the recent literature [71, 72], for a total of 292 introns in 276 genes. The majority of the introns that I studied have experimental support ($> 75\%$), with nine introns being initially discovered using unbiased experimental techniques for identifying spliceosomal introns [71,72]. The remaining introns were largely annotated by gene- and intron-finding programs based on the *S. cerevisiae* genome and characteristics of known experimentally verified introns [74, 106]. I did not include novel splice variants observed in recent large-scale studies of the yeast transcriptome [70, 73], as many of these observations lack additional experimental support and it is often unclear whether low-abundance sequences might reflect rare alternative splice variants or mis-splicing. After retrieval of the intron-containing gene sequences from the reference genome, I used `megablast` [107] to identify homologous sequences in the remaining 37 yeast strains. I aligned the 38 sequences for each gene using `MAFFT` (Katoh and Toh 2008). I obtained maximum-likelihood estimates of genome-wide levels of synonymous and non-synonymous site divergence for all pairwise comparisons between strains using `PAML` [108].

The strains I examined from the *Saccharomyces* Genome Resequencing Project include nucleotides that have been imputed by taking into account phylogenetic relationships between strains, to correct likely sequencing errors and fill in missing data [1]. I re-analyzed the data using only strains that had $< 3\%$ imputed data, and found qualitatively similar results (Appendix B.1). As such, I used complete assemblies (including imputed data) for all further analyses, and I expect my conclusions to be robust to the presence of imputed nucleotides.

2.3.2 Experimental determination of splicing efficiency

With assistance from Caitlin Connelly, I estimated intron splicing efficiency across seven introns in *S. cerevisiae* strains BY4716 (isogenic to S288C, the yeast reference genome strain), DBVPG1373, K11, UWOPS03-461.4, UWOPS83-787.3, YJM975, YS2, and YS4. Caitlin obtained four biological replicates per strain per intron, grew the strains to mid-log phase (OD_{660} 0.8–1.0) in rich medium (YPD), extracted RNA by the acid phenol method [109], and made cDNA by random priming using the Superscript III First-Strand Synthesis kit (Invitrogen Corp., Carlsbad, CA). I designed primers in `Primer3` [110] to amplify the products of genes YBL108C, YBR215W, YLR199C, YLR445W, YML025C, YNL004W, and YNL038W. Caitlin visualized the gene products on 2% agarose gels using ethidium bromide and used the program `ImageQuant` (Molecular Dynamics, Inc., Sunnyvale, CA) to quantify the amount of spliced and unspliced gene product. I quantified splicing efficiency as the fraction of spliced product to the sum of spliced and unspliced product. I analyzed differences in intron splicing efficiency using the nonparametric Kruskal-Wallis rank sum test in `R` [111].

2.3.3 Sequence analysis

S. cerevisiae introns are characterized by highly conserved $5'$, $3'$, and branch point sequences [76–78]. I examined six, three, and seven base pairs, respectively, of these sequences, which constitute the positions corresponding to the most highly conserved residues in consensus splice sequences [78, 112]. To summarize nucleotide variation between strains, I used the

formula $\hat{\pi} = \sum_{i=1}^{S} h_i w_i$, where h_i is an unbiased estimate of nucleotide diversity for the ith segregating site [113] and w_i is a weight for each site calculated by dividing the number of strains with non-gap nucleotides at site i by the total number of strains. By ignoring sequence gaps, I minimized the effect of alignment errors or incomplete sequences on our calculations. I plotted position-specific nucleotide diversities, and generated sequence logos, using the `R` software environment [111, 114].

2.3.4 Estimating the magnitude of purifying selection

To obtain quantitative estimates of the magnitude of selection acting on intron splice sequences, I conducted simulations using a computer implementation of the ancestral selection graph [104] provided by James Ronald. I simulated strains as sampled individuals from one common population (panmictic model) or from a model that included population structure (structure model). The complete demographic history of these strains is likely to be complex, but I sought to construct a simple structured model that recapitulated levels of synonymous site divergence observed between strains to gauge the robustness of my results to demographic uncertainty. I constructed a phylogenetic tree based on synonymous site divergences, and observed a topology very similar to a tree based on genome-wide pairwise SNP differences [1]. At a crude level, this tree can be subdivided into two divergent groups, loosely corresponding to strains involved in baking and wine production and those from Europe versus Asian, African, and several wild non-European strains. I estimated the time to most recent common ancestor (TMRCA) of these groups to be approximately $1.3N_e$ generations (roughly 11,500 years, adopting estimates of the mutation rate and generation time provided in [115]) using the method of Tang et al. [116].

As the "European" group exhibits markedly lower synonymous site divergence than the "Asian/African" group (0.0058 versus 0.0097), I chose to model a population bottleneck in the European group occurring after the estimated TMRCA of the two groups. A population bottleneck can be parameterized in terms of the increase in population homozygosity that results from a decrease in population size. In a randomly mating haploid population of finite size N, the inbreeding coefficient F reflects the chance that two randomly drawn copies of a

gene are identical by descent [117]. Working with $H = 1-F$, the probability of non-identity of two gene copies after t generations is $H_t = (1 - 1/N)H_{t-1} = (1 - 1/N)^t$ assuming the population is non-inbred at generation 0. Therefore, the increase in homozygosity caused by a bottleneck where the population is held at size N for t generations is $F_t = 1 - (1 - 1/N)^t$. Using the approximation $\log(1 - x) \approx -x$ (for small x) leads to the formula $F = t/N$, although this approximation breaks down at about $F > 0.2$, and it becomes more accurate to parameterize severe bottlenecks using the formula $\log(1 - F) = -t/N$. I modeled a bottleneck that began 0.5 N_e generations (roughly 5,000 years) ago, and searched a coarse grid of F =[0.1, 0.2, ..., 0.9] to determine that a relatively severe bottleneck was necessary to fit the observed data. I conducted a finer search across a range of bottlenecks where F = [0.675, 0.7, 0.725, ..., 0.975, 0.999, 0.9999]. I selected the best-fitting bottleneck by minimizing the sum of squared differences between the observed and simulated ratios of nucleotide diversity: (1) between-group to overall, (2) "European" group to "Asian/African" group, and (3) between-group to mean within-group, calculated at synonymous or simulated neutral sites. A severe bottleneck ($F = 0.875$) provided a very close fit to the observed data using these measures.

My implementation of the ancestral selection graph applies to evolution in haploid populations or diploid populations in which selection acts additively. I simulated using a four-allele model, where each simulation included a neutral site and a linked selected site at which the scaled selection coefficient was $\sigma = 2N_e s$ for the selected allele and $\sigma = 0$ for the remaining three alleles. Direct comparison of selected and linked neutral sites ensures that my estimates reflect the effect of selection acting directly on intron splice sequences rather than hitchhiking or background selection. Mutation occurred at the neutral and selected loci with rate $\theta/2$ along each branch with $\theta = 2N_e\mu = 0.0095$ (estimated from the synonymous site substitution rate). I ran simulations for $2N_e s = 0.0, 0.2, 0.4, \ldots, 13.0$ and $2N_e s = 14.0, 15.0, 16.0, \ldots, 50.0$. To increase the speed of simulations for strong selection, for $2N_e s > 13$ I sampled the remaining lineages from the stationary distribution at time $40N_e$ generations in the past as suggested by Pritchard [118]. I verified the results for strong selection using a theoretical formula for the distribution of gene frequencies in a panmictic population under a two allele-model with reversible asymmetric mutation and

one selectively favored allele (Figure B.1; Appendix B.3).

I simulated 1,000 replicates of 292 simulations for each selective coefficient. To assess the correspondence of simulations conducted for each selective class with the observed data, I calculated the reduction in nucleotide diversity at selected (intronic) sites, relative to neutral (synonymous) sites. To obtain confidence intervals for my model selection statistic, I calculated nucleotide diversity for each of the 1,000 replicates, and obtained the interval containing 95% of the realized nucleotide diversities. The number of simulations (unlinked sites) per replicate, 292, was chosen to match the size of the intron dataset, which consists of 292 introns.

2.3.5 Diffusion approximations for evolutionary dynamics of splice sequence polymorphisms

I used diffusion approximations derived by Kimura and Ohta [119] to explore the evolutionary dynamics of intron splice sequence polymorphisms. These formulas allow for the examination of the fixation probabilities and sojourn times of alleles subject to arbitrary selective advantage or disadvantage and present at arbitrary initial frequencies in the population. I calculated the mean sojourn time of an allele subject to selective disadvantage s using the formula $u(p)\bar{t}_1(p) + [1 - u(p)]\bar{t}_0(p)$, where $u(p)$ is the probability of ultimate fixation of an allele present at initial frequency p. $\bar{t}_1(p)$ and $\bar{t}_0(p)$ are the average number of generations until fixation conditional on ultimate fixation of the allele, and the average number of generations until loss conditional on ultimate loss of the allele, respectively. These formulas are [119]:

$$\bar{t}_1(p) = \int_p^1 \psi(\xi)u(\xi)\{1 - u(\xi)\}d\xi + \frac{1 - u(p)}{u(p)} \int_0^p \psi(\xi)u^2(\xi)d\xi$$

$$\bar{t}_0(p) = \frac{u(p)}{1 - u(p)} \int_p^1 \psi(\xi)\{1 - u(\xi)\}^2 d\xi + \int_0^p \psi(\xi)\{1 - u(\xi)\}u(\xi)d\xi$$

Assuming no recurrent mutation, for selection against an allele with disadvantage s in a haploid population of size N_e,

$$u(p) = \frac{1 - \exp(2n_e sp)}{1 - \exp(2n_e s)}$$

and

$$\psi(x) = 2N_e \frac{\int_0^1 \exp(2N_e sx)dx}{x(1 - x)\exp(2n_e sx)}.$$

Similarly, the probability that a mutant at frequency p rises to at least frequency p' is

$$u(p) = \frac{1 - \exp(2N_e sp)}{1 - \exp(2N_e sp')}.$$

The waiting time until the frequency of such events (which are rare under the conditions we discuss) is exponentially distributed with parameter λ equal to the frequency of the event, with an expected value of $1/\lambda$.

To calculate the number of new intron splice sequence mutations per day, I used the following estimates: (1) wild yeast are likely to reproduce at a rate of approximately eight generations per day [115], (2) the mutation rate at synonymous sites is approximately $\mu = 1.8 \times 10^{-10}$ [115], and (3) the effective population size of yeast is roughly $N_e = 26$ million (calculated using $\theta = 2N_e u = 0.0095$ estimated from synonymous sites).

2.3.6 Modeling intron presence/absence using genic characteristics

I used logistic regression to model the presence/absence of introns using genic characteristics as linear predictors. Specifically, I considered (1) the classification of each gene as essential or non-essential under standard laboratory conditions [120], (2) whether the gene encodes a ribosomal protein, (3) the codon adaptation index (CAI) [121] as an estimate of the relative expression level of each gene, (4) the genic GC content, and (5) dN/dS for sequences from all 38 strains as a proxy for the rate of protein evolution, calculated using `PAML` [108]. I implemented the model using the `glm` function in `R` [111]. Starting with all single predictors and second-order interaction terms, I used the `drop1` function [111] to remove predictors that did significantly improve the fit of the model. My final model consisted of predictors 1–4 above as well as the interaction between the first and second predictors.

2.4 Results

2.4.1 Polymorphisms are rare in key splice sequences

I compiled a list of 292 introns in 276 genes assembled from a variety of sources [71, 72, 74], `http://www.yeastgenome.org/`, excluding introns in dubious genes and introns lacking evidence of splicing in previous experimental studies [74, 75]. In the 38 yeast strains I

examined, I was able to identify sequences corresponding to the majority ($> 50\%$) of each intron for 286 introns. For the remaining six introns, there were a total of 28 instances where a strain was missing over half the intron sequence. These instances probably do not reflect true intron presence-absence polymorphisms. Rather, I attribute the missing bases to incomplete sequence coverage (24/28 instances exist in strains with < 1.5X genome sequence coverage; in all 20 instances where the complete intron is missing, a portion of the coding sequence is missing as well). However, it remains a formal possibility that these six introns represent true deletions of a large portion of the intron [102].

I focused on six, seven, and three base pairs of the $5'$, branch point, and $3'$ splice sequences, respectively (Figure 2.1). Among the 38 strains I examined, I identified 21 polymorphisms within 23 introns in 20 genes (two polymorphisms occur in splice sequences that are shared among multiple splice variants; Table 2.1). I found no polymorphisms in the remaining 269 introns in 256 genes. It is likely that many of the polymorphisms I identified are functionally neutral. For example, eight of the 21 polymorphisms involve the alleles [C/T]AG at the $3'$ splice site. Since 125 introns in my set use a CAG $3'$ splice site and 154 use a TAG $3'$ splice site, it is unlikely that a switch between the two has dramatic effects on splicing. However, as I demonstrate below, a subset of these polymorphisms do affect splicing efficiency.

I estimated position-specific nucleotide diversity for the conserved intron splice sequences (Figure 2.1). Residues previously identified as most critical for achieving splicing – the first two and last two positions of the intron, and the adenosine in the penultimate position of the branch point [81–84] – showed complete invariance, with no polymorphisms present in any strain for any intron. Interestingly, several positions located in the branch point sequence (2, 3, 4, 5, and 7) and the fifth position in the $5'$ splice site also showed either extremely low levels of polymorphism or complete invariance (Figure 2.1). Experimental studies of the effects of mutations at these positions have produced mixed results [76, 80, 81, 84, 122]; the low levels of variation I observed suggest that levels of functional constraint at these sites are of a similar order of magnitude to previously identified sites critical to splicing.

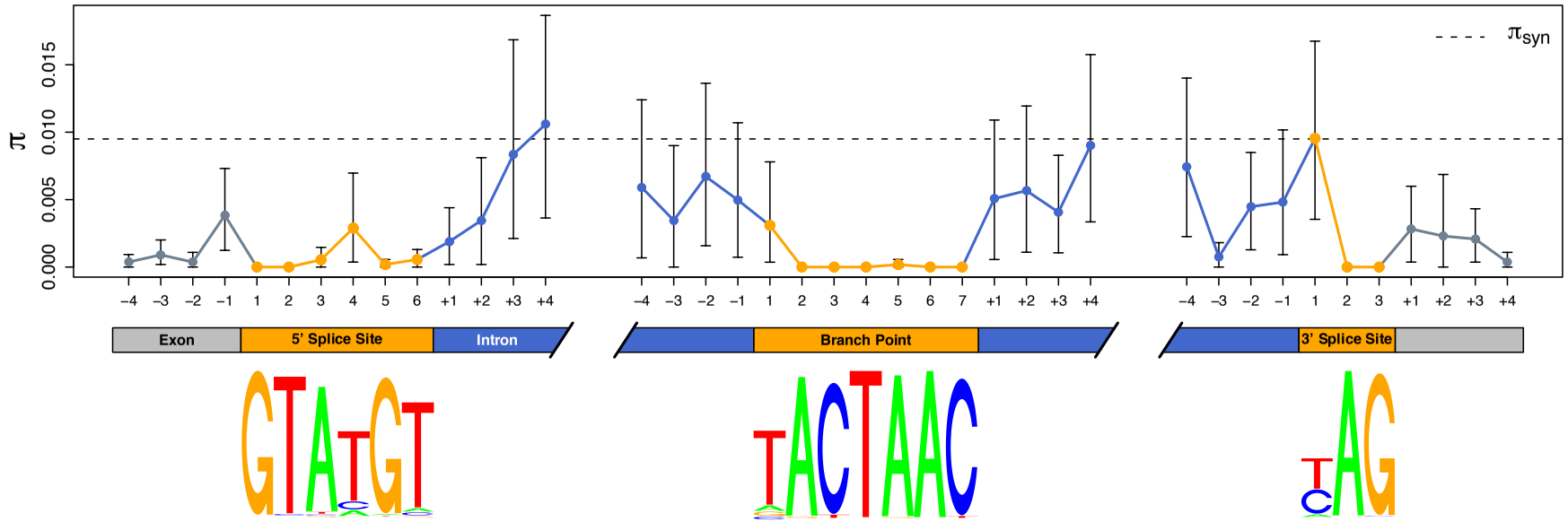


Figure 2.1: A window of four bases on each side of the splice site sequences is shown (colored gray for exonic sequences and blue for intronic sequences). Error bars show 95% confidence intervals based on 1000 resamplings. Sequence logos beneath splice site sequences were generated using our set of 292 introns, with the height of each position scaled according to information content.

Table 2.1: Summary of polymorphisms identified in yeast intron splice sequences

ORF	Gene name[a]	Location	Reference sequence	Alternate sequence	*S. paradoxus* sequence[b]
YBL018C	*POP8*	5′ splice site	GTA**T**GT	GTA**C**GT	GTACGT
YDR367W		5′ splice site	GT**ATG**T	GT**TGA**T	-
YGL033W	*HOP2*	5′ splice site	GT**T**AAG	GT**C**AAG	GTAAAG
YKL186C	*MTR2*	5′ splice site	GTATG**T**	GTATG**A**	ACATGA
YLR445W		5′ splice site	GTA**A**GT	GTA**G**GT	GTAAGT
YML025C	*YML6*	5′ splice site	GTA**C**GT	GTA**T**GT	GTACGT
YNL246W	*VPS75*	5′ splice site	GTA**T**GT	GTA**A**GT	GTAAGT
YBR215W	*HPC2*	branch point	**G**ATTAAC	**C**ATTAAC	TACTAAC
YCL002C		branch point	**G**ACTAAC	**A**ACTAAC	GACTAAC
YKL150W	*MCR1*	branch point	**T**ACTAAC	**A**ACTAAC	TACTAAC
YLR199C	*PBA1*	branch point	**G**ACTAAC	**A**ACTAAC	GACTAAC
YLR316C	*TAD3*	branch point	**A**ACTAAC	**G**ACTAAC	AACTAAC
YNL004W	*HRB1*	branch point	TACT**A**AT	TACT**G**AT	TACTAAT
YBR084C-A	*RPL19A*	3′ splice site	**C**AG	**T**AG	CAG
YBR089C-A	*NHP6B*	3′ splice site	**T**AG	**C**AG	TAG
YKL006C-A	*SFT1*	3′ splice site	**C**AG	**T**AG	CAG
YKL186C	*MTR2*	3′ splice site	**C**AG	**T**AG	-
YNL038W	*GPI15*	3′ splice site	**C**AG	**T**AG	CAG
YNL312W	*RFA2*	3′ splice site	**C**AG	**T**AG	TAG
YOR182C	*RPS30B*	3′ splice site	**T**AG	**C**AG	CAG
YOR234C	*RPL33B*	3′ splice site	**T**AG	**C**AG	TAG

[a] Blank gene names indicate uncharacterized ORFs.

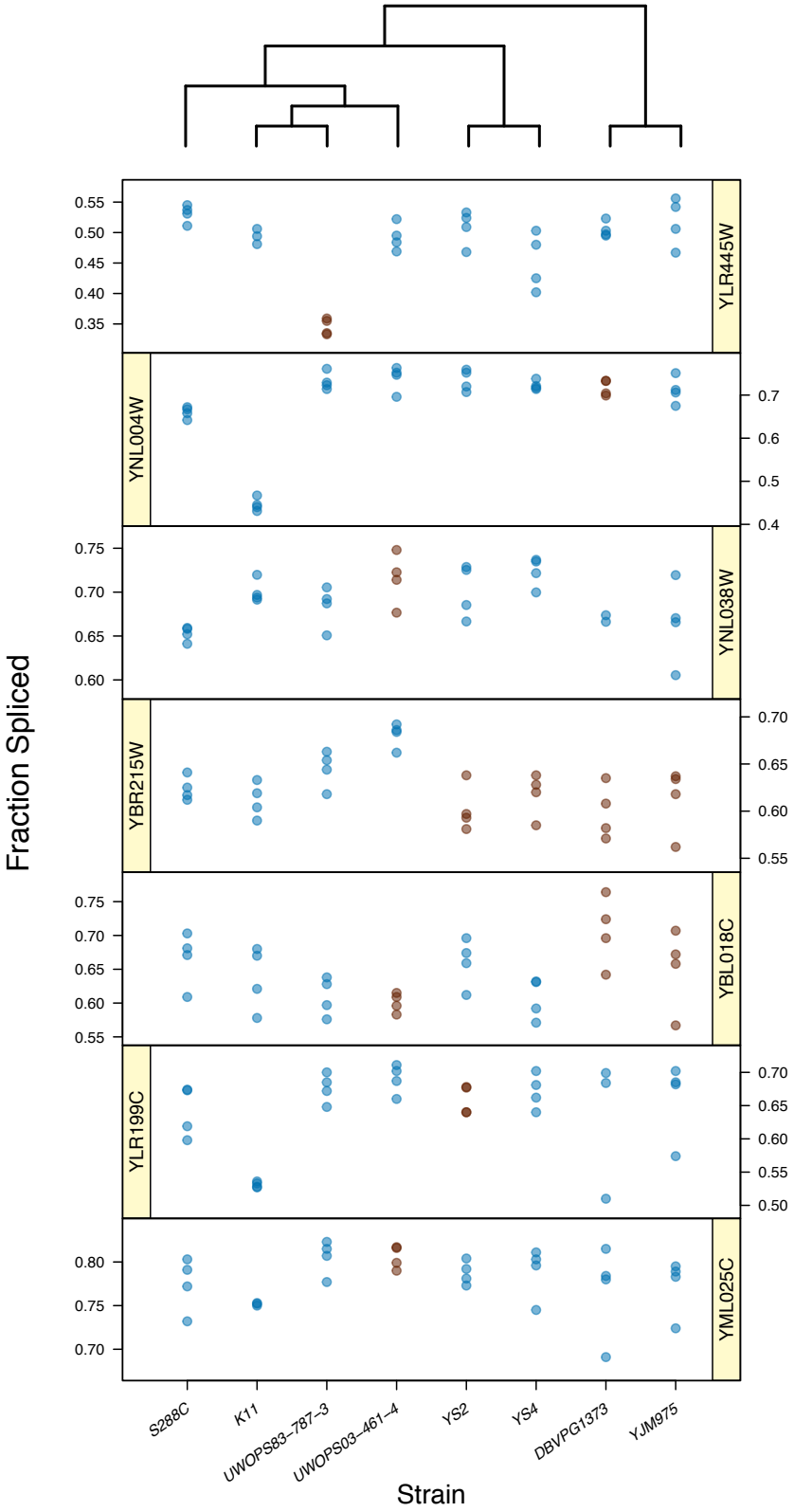
[b] Dash indicates that the *S. paradoxus* allele at this position could not be confidently identified.

2.4.2 Natural variation in intron splicing efficiency

With the assistance of Caitlin Connelly, I surveyed intron splicing efficiency across seven introns in eight strains in order to better understand natural variation in intron splicing. We examined introns that had at least one polymorphic splice sequence nucleotide, and focused on eight strains chosen to ensure that we captured variation present at both alleles for all seven introns (Figure 2.2). Caitlin estimated splicing efficiency by quantifying the amounts of spliced and unspliced gene product on electrophoretic gels, using four biological replicates per strain. Interestingly, I observed significant variation among strains in splicing efficiency for the majority of introns (4/7 introns; Kruskal-Wallis rank sum test, $p < 0.05$; an additional two introns are marginally significant; 2.2). In one case (YLR445W, a protein of unknown function), a 5′ splice site polymorphism from the common GTAAGT (used in 13 other introns and present at the orthologous 5′ splice site in *Saccharomyces paradoxus*, the closest known relative of *S. cerevisiae*) to the sequence GTAGGT (not found as a 5′ splice site in other introns) resulted in dramatically lower splicing efficiency in strain UWOPS83-787.3 (Figure 2.2, top panel). Thus, heritable variation in splicing efficiency is common, and polymorphisms in splice sequences can contribute to this variation.

In the remaining six introns, the polymorphisms I observed in splice sequences did not clearly coincide with increases or decreases in experimentally measured splicing efficiencies (Figure 2.2), suggesting that the genetic basis of splicing efficiency is complex. In addition to conserved splicing sequences, these results suggest that *trans*-acting factors and other *cis*-acting factors (such as additional sequence motifs or spurious splice sequences) also contribute to the efficiency of the splicing reaction [85, 123–125]. It is not entirely surprising that many of the splice sequence polymorphisms I examined did not track with our measured splicing efficiencies. Specifically, the test set of introns included three polymorphisms between alternate 5′ or 3′ splice site sequences that are common across the global set of introns, suggesting that they should not dramatically affect splicing. Moreover, polymorphisms present in the population may persist precisely because their functional effect on splicing is minimal. Finally, splicing efficiency in these strains was only measured under standard laboratory conditions, and some polymorphisms may have environment-specific

Figure 2.2: Splicing efficiency for seven introns measured across the eight strains shown at bottom. Panels indicate the fraction of spliced gene product measured for four biological replicates of each strain for the gene indicated along the axis. Blue dots indicate strains with the reference (strain S288C) splice sequence allele, and dark red dots indicate strains with the alternative allele. The top four panels show introns with significant variation in splicing efficiency among strains (Kruskal-Wallis rank sum test, $p < 0.05$). Results for introns in genes YBL018C and YLR199C were marginally significant ($p = 0.052$ and $p = 0.060$, respectively), and YML025C did not show significant variation in splicing efficiency among strains ($p = 0.13$). Note that scaling of the vertical axis differs between introns. The phylogeny above the figure depicts the approximate genealogical relationship between strains.


Fraction Spliced
YLR445W
YNL004W
YNL038W
YBR215W
YBL018C
YLR199C
YML025C
S288C
K11
UWOPS83−787−3
UWOPS03−461−4
YS2
YS4
DBVPG1373
YJM975
Strain

effects on splicing.

2.4.3 Quantitative estimates of the strength of selection acting on intron splice sequences

The significantly reduced levels of diversity within critical splicing sequences (Figure 2.1) suggest that most newly arisen mutations at these sites are deleterious and removed by purifying selection. To obtain quantitative estimates of the strength of purifying selection acting on intron splice sequences, I used a computer implementation of the ancestral selection graph [104] provided by James Ronald. The ancestral selection graph describes a genealogical process that extends the coalescent by properly taking into account the effect of natural selection [104]. I simulated strains as sampled individuals from one common population (panmictic model) or from a model that included population structure, with two subpopulations that split at some time in the past, one of which subsequently experienced a bottleneck (structure model). Obviously, both models are simplifications of the real demographic history of these 38 strains. However, the simple model of population structure I used recapitulates major patterns of synonymous site divergence within and between subpopulations (see Methods). In addition, it is useful to examine varying models to gauge the robustness of my results to demographic uncertainty [126].

I evaluated the fit of the panmictic and structure models to the observed data using the ratio of nucleotide diversity at selected (intronic) sites to diversity at neutral (synonymous) sites. Synonymous sites are subject to weak selective constraint in yeast [127], which suggests that normalization using synonymous sites will lead to slight underestimates of the magnitude of selection against splice sequence polymorphisms; nevertheless, this bias will not be present for relative comparisons between intronic and non-synonymous sites that are both normalized using synonymous sites (see below). The value of this summary statistic was broadly similar across selection coefficients for our two demographic models, with slightly lower values for the structure model (Figure 2.3). I first estimated the strength of purifying selection acting on intron splice sequences as a class, then considered the strength of selection acting on each intronic site. Given that the panmictic and structure models give very similar results across the range of selective classes I examined (Figure 2.3), I provide

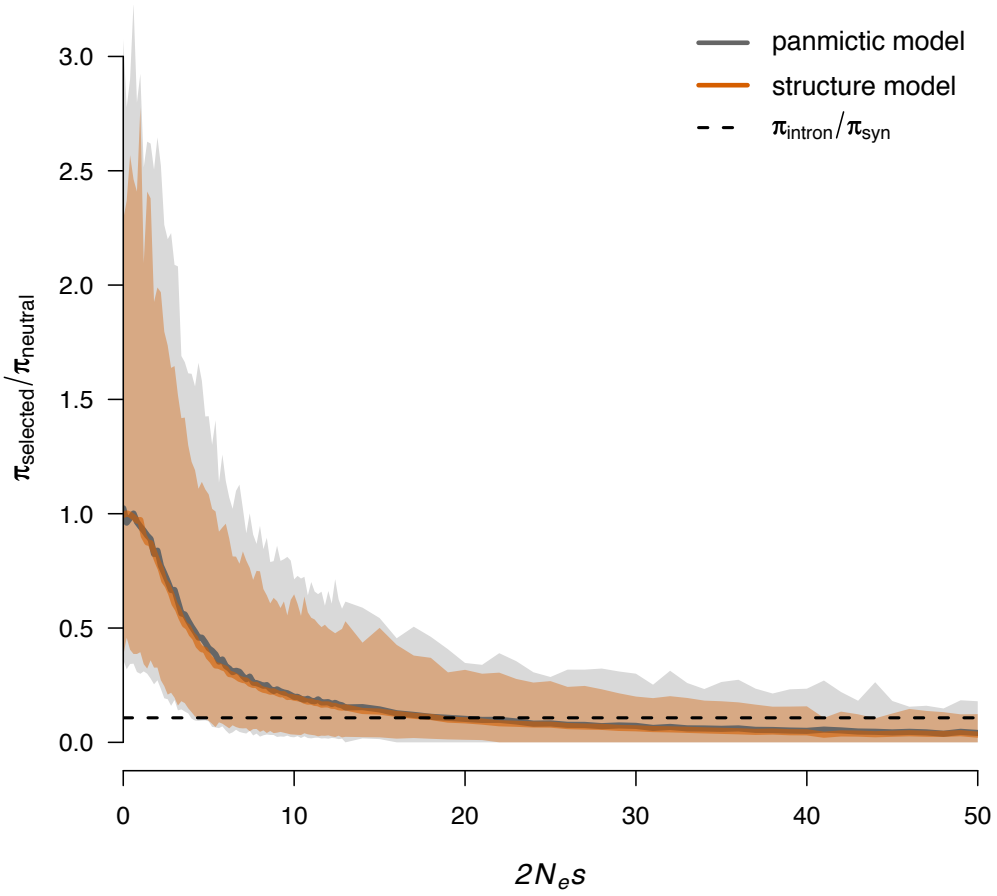


Figure 2.3: Figure shows the reduction in diversity at simulated selected site, relative to linked neutral site, as a function of the strength of selection against new mutations at the site. Solid lines show the mean of this ratio, across selection coefficients, for the two demographic models studied. Lighter shading gives 95% confidence intervals based on 1000 simulated replicates sized to match the intron dataset. The dashed line shows the reduction in diversity observed in the intron splice sequence dataset, relative to synonymous sites.

Table 2.2: Estimates of the strength of selection on intron splice sequences are specified in terms of the scaled selection coefficient, $2N_es$. First row indicates base present at each position in the $5'$ splice site, branch point, or $3'$ splice site as indicated. Second row indicates estimate of the strength of selection.

$5'$ splice site						Branch Point							$3'$ splice site		
G	T	A	T	G	T	T	A	C	T	A	A	C	C/T	A	G
13.0	13.0	6.0	1.6	11.0	6.0	0.8	13.0	13.0	13 .0	11.0	13.0	13.0	0.0	13.0	13.0

estimates of the strength of selection based on the results of simulations using the panmictic model. The best fit to the observed intron splice sequence data occurred at $2N_es \approx 19$ (Figure 2.3); 95% confidence intervals based on simulations suggest that the minimum strength of purifying selection is $2N_es > 4$. My simulation scheme ensures that these estimates reflect selection acting directly on intron splice sequences rather than hitchhiking or background selection (see Methods).

Next I investigated the strength of selection acting on individual splice sequence nucleotides. This analysis is complicated by the lack of variation at individual sites for sites subject to very strong selection. To this end, I focused on estimating the lower limits to selection at each site. Lower limits were obtained using the confidence intervals associated with my simulations (Figure 2.3), which properly account for the stochastic properties of coalescent genealogies as well as sampling variation present in the relatively small intron dataset. I observed considerable heterogeneity in the magnitude of selection across yeast intron splice sequence nucleotides (Table 2.2). One class of sites, as mentioned above, showed either very low levels of polymorphism or complete invariance. For these sites (first and last two positions in the intron, branch point positions 2-7, and fifth position in $5'$ splice site), I estimate the scaled selection coefficient to be at least $2N_es \approx 11$ (Table 2.2). The lower limit to the strength of selection was approximately one order of magnitude weaker at the fourth position in the $5'$ splice site and first position of the branch point (Table 2.2). Notably, levels of variation were indistinguishable from neutrality only at the first position in the $3'$ splice site.

Finally, I estimated the strength of selection on non-synonymous sites using the same demographic models described above. A minority of non-synonymous mutations may have been driven to high frequency by positive selection (which would downwardly bias my estimates), although positive selection acting on intron splice sequence mutations is also conceivable. I evaluated the fit of models with varying selection intensities using the same metric as above, the ratio of the nucleotide diversity at selected (non-synonymous) versus putatively neutral (synonymous) sites. My estimate of the magnitude of purifying selection acting on the average non-synonymous site is $2N_e s = 10.6$. Although this estimate is subject to uncertainty, these results suggest that the strength of purifying selection acting on intron splice sequences as a class is nearly double that acting on an average non-synonymous site.

2.4.4 Evolutionary dynamics of intron splice sequence polymorphisms

An advantage of obtaining quantitative estimates of the strength of selection acting on intron splice sequences (Figure 2.3, Table 2.2) is that these estimates can be used to better understand the evolutionary dynamics of existing genetic variation and newly arising mutations. I used diffusion approximations derived by Kimura and Ohta [119] to explore the fixation probabilities and sojourn times of alleles as a function of the estimated selection coefficients across intron splice sequence positions (Figure 2.4). Above, I estimate the average strength of selection against splice sequence mutations as a whole to be $2N_e s \approx 19$. With purifying selection of this magnitude, it is approximately nine million times more likely that a newly arising mutation at a neutral site will eventually rise to fixation than a newly arising mutation at a selectively constrained site. Nevertheless, the mean sojourn times of newly arising neutral and deleterious mutations are quite similar (roughly 20% longer for neutral than for deleterious mutations), since a large fraction of both classes of mutations are lost soon after arising (Figure 2.4, middle panel). Interestingly, newly arising strongly deleterious ($2N_e s \approx 19$) mutations that are destined for ultimate fixation arrive there nearly three times faster on average than new neutral mutations (Figure 2.4, left panel). This somewhat counterintuitive result arises from the fact that low and moderate frequency mutations that are strongly deleterious are overwhelmingly likely to be lost. Thus, the only new strongly

deleterious mutations that rise to ultimate fixation are those exceptionally rare mutants that rapidly and continually increase stochastically in frequency faster than the mutant alleles can be purged from the population by purifying selection.

Heterogeneous selective pressures across intron splice sequence nucleotides result in significantly different predicted evolutionary trajectories for polymorphisms that arise at different positions within splice sequences. Using estimates of the mutation rate, reproductive capacity, and effective population size of yeast (see Methods), I estimate that an average of eleven new mutations arise each day at any particular intronic splice sequence position in the global yeast population, scattered among the 292 introns I studied. The fate of these mutations varies widely, depending on the magnitude of selection against new mutations at the position where they occur. Even for selectively neutral mutations, the probability of ultimate fixation is only about one in 25 million for the large global yeast population. For the splice sequence positions where purifying selection is detectable but weak (Table 2.2), substitutions occur at about 60% the rate at neutral sites, while for the class of sites that shows about an order of magnitude stronger selection, the substitution rate is only roughly 0.02% that at neutral sites. The differences are less extreme when considering the probability that a newly arising mutation becomes common in the population, since the fate of new mutants is determined largely by drift while they remain rare. For example, the average waiting time until a new mutation at a particular splice sequence nucleotide (within one of the 292 introns we study) attains 10% frequency is roughly 650 years for a neutral mutation, 690 years for a weakly deleterious mutation ($2N_e s = 1$), and 1,120 years for a strongly deleterious mutation ($2N_e s = 10$) (see Figure 2.4, right panel, for waiting times scaled by N_e). Thus, although strongly deleterious new mutants are ultimately fixed at exceedingly low rates, their behavior at relatively low frequencies does not differ greatly from neutral variants. In some cases, changes in environmental conditions or compensatory evolution could lead initially deleterious mutations managing to drift to moderate frequency to become selectively favored and rise to ultimate fixation.

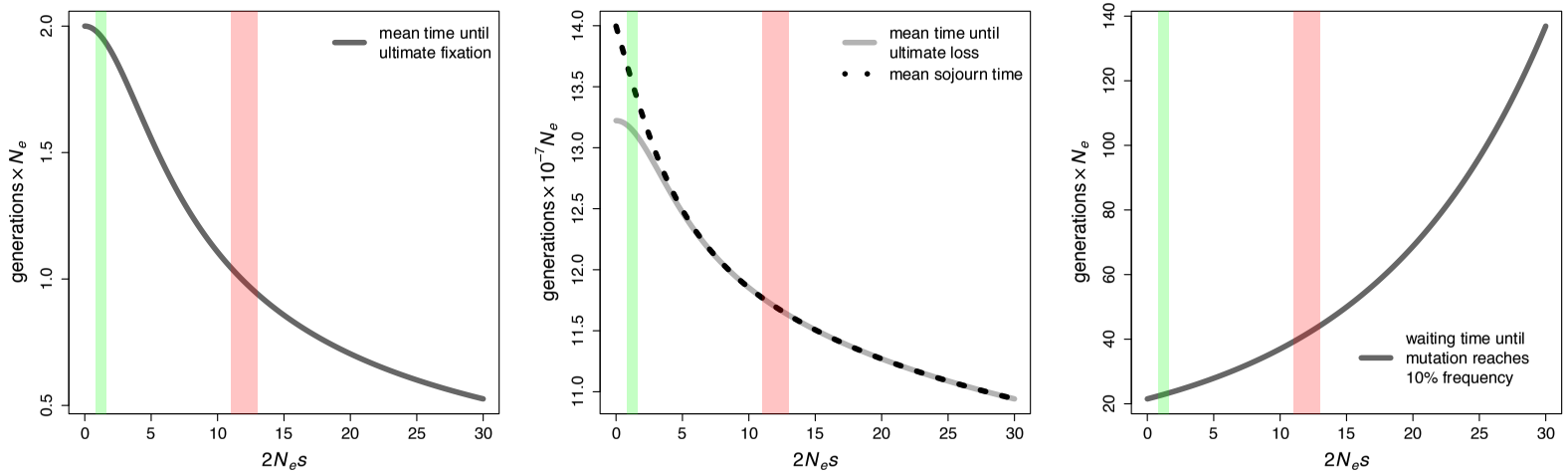


Figure 2.4: The three panels depict the fates of newly arising mutations as a function of the selection coefficient against new mutations. Times shown on the y-axes are scaled in terms of the effective population size as shown. Opaque boxes depict lower limits to the magnitude of selection at weakly constrained splice nucleotides (green) and strongly constrained splice nucleotides (red), as discussed in the text and Table 2.2. Left panel shows the mean time for newly arising mutations to achieve fixation, conditional upon ultimate fixation. Middle panel shows the mean time for newly arising mutations to be lost conditional upon ultimate loss (solid line) and the mean sojourn time of newly arising mutations (dotted line). The lines converge as $2N_es$ increases, since ultimate loss becomes increasingly more probable as selection against new mutants increases. Right panel shows the expected waiting time for a newly arising mutation to reach 10% frequency.

2.4.5 Why are extant yeast introns retained?

Although my analyses clearly demonstrate that purifying selection acts on intron splice sequences, it is less clear what specifically is deleterious about perturbing intron splicing in yeast. One possibility is that most introns are on their way out of the yeast genome [87] and are not functionally important. Under this scenario, the sole consequence of perturbing splicing arises when mutations in critical intronic splice sequences disrupt the splicing process, leading to the accumulation of splicing intermediates or improperly spliced transcripts [81, 122]. These defective precursor molecules either result in proteins likely to have impaired function or are eliminated by nonsense-mediated mRNA decay and alternative degradation pathways [128–130], which would effectively knock out a gene.

An alternative possibility is that yeast introns, in general, are functionally important and play a role in gene regulation. In yeast, the modulation of splicing efficiency can effect rapid response to environmental challenges [94] and contribute to important biological processes, such as the initiation of meiosis [131]. Moreover, the interactions between the spliceosome and proteins responsible for transcription, capping, polyadenylation, RNA export, and nonsense-mediated mRNA decay [132] support an integral role for introns in affecting transcriptional and translational yield [93]. Under this scenario, perturbing splicing could have the same severe effects on the splicing reaction itself as described above, but could also have deleterious consequences for normal gene regulation.

To examine the functional importance of yeast introns, I compared the prevalence of introns within genes classified as essential or non-essential under standard laboratory conditions [120]. The rationale of this analysis is that because mutations in critical splice sequences disrupt proper splicing and often lead to loss of gene function, introns create a sizeable target for mutation to null alleles [69]. My above estimates of strong purifying selection acting on intron splice sequence mutations formally justify this premise by confirming that intron-containing alleles are a mutational liability, since newly arising splice sequence mutations tend to be strongly deleterious. As such, functionless introns that reside passively within genes will tend to be lost (on an evolutionary timescale) more rapidly from essential genes than from non-essential genes. Conversely, functionally important introns

will tend to be preserved in all genes, perhaps even more so in essential genes in cases where the function of the intron involves regulation of the gene in which it resides. In the results described below, we analyzed ribosomal and non-ribosomal intron containing genes separately because ribosomal genes exhibit several features (such as high mean levels of expression and larger mean intron sizes) that distinguish them from non-ribosomal protein genes [74, 133].

Interestingly, introns are significantly over-represented in essential genes that code for non-ribosomal proteins (Figure 2.5). The prevalence of introns in this class of genes suggests that introns tend to have important functions that preserve their presence within non-ribosomal protein genes. Among ribosomal protein genes, there is no significant difference in the proportion of essential versus non-essential genes containing introns, within each of the duplicated and non-duplicated subclasses of ribosomal protein genes (Figure 2.5). These data do not support the hypothesis that introns in ribosomal protein genes are functionless genomic relics, since essential ribosomal protein genes are not significantly less likely to harbor introns. Even so, the lack of a clearer pattern may result from a combination of small sample sizes as well as the fact that many ribosomal protein genes classified as non-essential are nevertheless likely to impose severe growth defects in homozygous mutant form [120].

One proposed mechanism for intron loss in yeast involves homologous recombination of reverse-transcribed cDNAs [87]. Although other processes may also contribute to intron loss (e.g. simple genomic deletion) [134], the loss of intronic sequence by RNA-mediated recombination has been experimentally demonstrated in *S. cerevisiae* [135]. This mechanism predicts that highly expressed genes should lose their introns more rapidly than those expressed at lower levels. To assess whether the distribution of introns in essential and non-essential genes (Figure 2.5) might be driven by this neutral process rather than reflecting the preservation of functionally important introns, I used logistic regression to model the presence/absence of introns using genic characteristics as linear predictors (see Methods). I found that a gene's essentiality classification remained a significant predictor of intron presence ($p < 0.01$) even after accounting for transcript abundance (using codon bias as a surrogate measure of expression level), suggesting that my observation of an over-

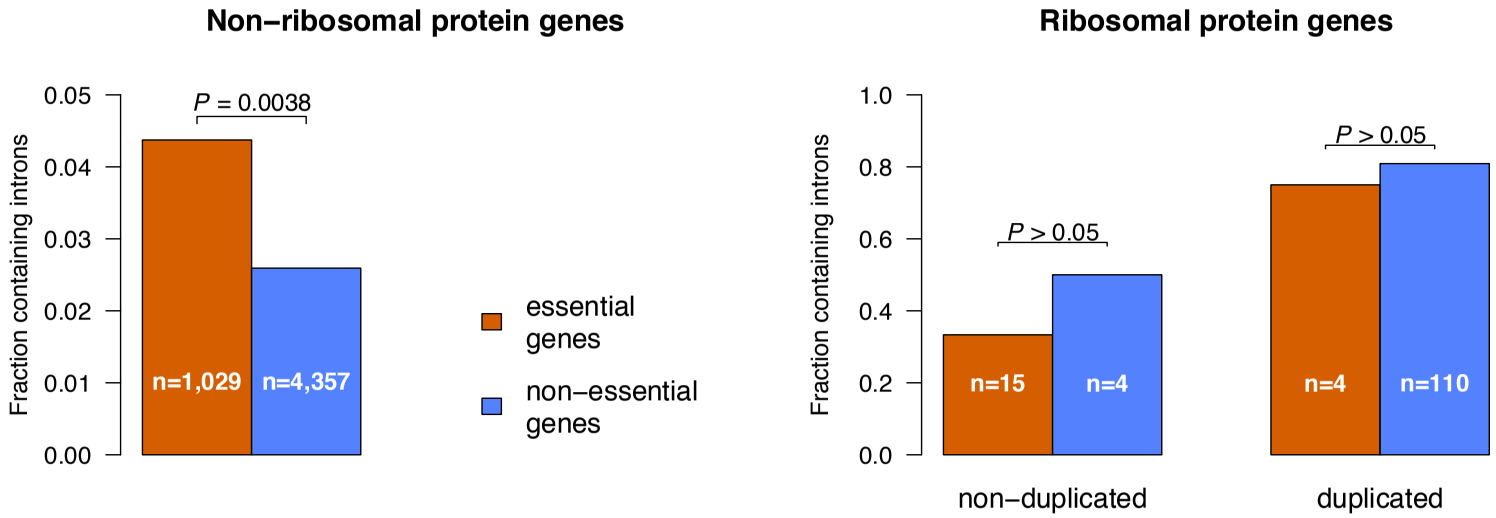


Figure 2.5: Figure shows the proportion of genes containing introns, divided according to whether the genes are classified as essential [120]. Top panel depicts the fraction of intron-containing genes among non-ribosomal protein genes. Bottom panel depicts the fraction of intron-containing genes among ribosomal protein genes, divided into ribosomal protein genes that are duplicated and those that are not. Note the difference in vertical scaling between top and bottom panels, as introns are much more common in ribosomal protein genes. p-values show results from Fisher's exact test of the null hypothesis that the proportion of genes containing introns does not differ between each pair of orange and blue bars. Sample sizes for each category are noted in white text within each bar. Genes that were not classified as essential or non-essential by Giaever et al. [120] are omitted from this figure.

representation of introns in essential non-ribosomal protein genes does reflect the functional importance of these introns.

2.5 Discussion

The molecular details of intron splicing have been studied in considerable detail. Experimental studies have been extraordinarily valuable for unraveling the molecular basis of the splicing process and for determining the molecular consequences of specific mutations in splice sequences [80, 81, 84]. Nevertheless, such approaches are inherently limited by the difficulty in accurately recapitulating the conditions experienced by wild yeast populations, as well as the incomplete detection of every phenotype that affects fitness. My analyses complement such studies by taking advantage of the characteristic signature imparted on DNA sequence variation by natural selection. My results are consistent with previous studies that have uniformly identified the first and last two bases of the intron and the sixth branch point position as critical for proper splicing [81–84]. For several other positions where the consensus is less clear (the fifth position of the 5′ splice site as well as branch point positions 2, 3, 4, 5, and 7) [76, 80, 81, 84, 122], I suggest that the level of selective constraint is comparable and the positions are similarly integral to the splicing process in wild environments.

I estimate the strength of purifying selection acting on yeast intron splice sequences as a class to be nearly double that governing the evolution of an average non-synonymous polymorphism or a *cis*-acting polymorphism influencing gene expression [136]. Thus, most newly arisen mutations in intron splice sequences are deleterious and are eliminated by purifying selection, although some mildly deleterious alleles may attain appreciable frequencies. This high level of selective constraint partially reflects the specific sequences I examined, which constitute the most critical residues for intron splicing. In contrast, collections of non-synonymous polymorphisms or *cis*-regulatory polymorphisms are likely to contain many positions at which mutations are functionally neutral, leading to an underestimate of the strength of selection at functionally critical sites. For example, it has been estimated that 36% of non-synonymous single nucleotide polymorphisms are deleterious in *S. cerevisiae* [137]. Similarly, estimates of the strength of selection acting on extant *cis*-

regulatory polymorphisms would be diluted by the presence of neutral polymorphisms in promoters and 3′ UTRs [136]. Some of the constraint I detect may also reflect the functional importance of regulated splicing (see below). Might the strength of purifying selection acting on yeast intron splice sequences vary systematically between genes? I examined a variety of genic features (gene ontology terms, GC content, dN/dS, and codon bias) separately in ribosomal protein encoding and non-ribosomal protein encoding genes with and without polymorphic intron splice sequences (Appendix B.2). I observed no detectable heterogeneity between genes with or without polymorphic intron splice sequences, suggesting that levels of functional constraint for intron splicing are broadly similar within ribosomal protein encoding genes and within non-ribosomal protein encoding genes.

In this chapter, I present evidence suggesting that introns tend to be actively maintained in *S. cerevisiae*. It is important to note that there are several possible mechanisms through which intronic sequences might contribute to organismal function. First, the modulation of splicing efficiency contributes to important biological processes, such as the initiation of meiosis [131], and can be an important mechanism for gene regulation [94]. Selective constraint attributable to this function would be reflected in the strong purifying selection we observe acting on key splice sequences, and would contribute to a greater retention of introns in essential genes (Figure 2.5). Second, the close association between the spliceosome and transcriptional machinery [132] points to a general role for introns in affecting transcriptional and translational yield of the genes in which they reside [93]. Since this intronic feature is independent of specific splice sequence nucleotides, it is not reflected in purifying selection on splice sequences, although it is likely to contribute to the retention of introns in essential genes. Finally, intronic bases not directly involved in the splicing reaction could encode functional elements such as promoters [92], snoRNAs [91], or binding sites for regulatory proteins. When the functional importance of such intron-encoded elements is unrelated to the importance of the gene in which the intron resides, this form of constraint is not detectable by my analyses.

Fink's proposal [87] that intron loss in yeast occurs largely through homologous recombination of reverse-transcribed cDNAs predicts the 5′ bias in intron location observed in the yeast genome. It might be expected that this bias would be absent in essential genes, where

I argue that introns tend to be preserved due to their functional importance (Figure 2.5). In fact, I observed a strong 5′ bias in location for introns in both essential and non-essential genes (Figure B.2). For both ribosomal and non-ribosomal protein genes, there is no significant difference in the distribution of intron locations between essential and non-essential genes (Wilcoxon-Mann-Whitney test; $p > 0.05$). However, this does not contradict my assertion that introns in essential non-ribosomal protein genes have been preserved due to functional importance. First, the sample sizes for the nonparametric test above are small, and the power to detect a subtle difference in locational bias is likely to be low. Second, an evolutionary process of intron gain occurring uniformly throughout genes and intron loss preferentially occurring at the 3′ end of genes predicts a steady-state distribution of introns that is 5′ biased. Thus, unless insertion of a new intron into a gene is immediately adaptive, most introns that filter through the sieve of natural selection will be positioned near the 5′ end of the gene.

These population genomics analyses allowed me to characterize the forces governing the evolutionary trajectory of polymorphisms in key splice sequences in *S. cerevisiae*. My analyses demonstrate that these sequences are subject to strong functional constraint. I propose that introns are not merely genomic relics on their way out of the yeast genome; patterns of intron prevalence in essential and non-essential genes suggest that, at least in non-ribosomal protein genes, introns appear to be actively maintained for their functional importance. My relatively high estimate of the magnitude of purifying selection governing the evolution of splice sequences reflects the need for intronic bases to be properly removed from transcripts, but is also likely to arise from the functional importance of regulated splicing. Ultimately, disentangling the possible contributions of yeast introns to organismal function will require detailed studies of splicing in different environments and at different points in the cell cycle. Obtaining a better understanding of the dynamics and functional importance of introns in yeast may inform our understanding of the prevalence of introns in more complex eukaryotic genomes.

Chapter 3

A STATISTICAL MODEL FOR ALLELE-SPECIFIC GENE EXPRESSION MEASURED BY RNA-SEQ

This chapter contains material published in [138]. I gratefully acknowledge Jon Wakefield, without whose assistance the statistical model described in this chapter could not have been conceived, specified, or implemented.

3.1 Summary

Variation in gene expression is thought to make a significant contribution to phenotypic diversity among individuals within populations. Although high-throughput cDNA sequencing offers a unique opportunity to delineate the genome-wide architecture of regulatory variation, new statistical methods need to be developed to capitalize on the wealth of information contained in RNA-Seq datasets. To this end, I developed a powerful and flexible hierarchical Bayesian model that combines information across loci to allow both global and locus-specific inferences about allele-specific expression (ASE). I applied this methodology to a large RNA-Seq dataset obtained in a diploid hybrid of two diverse *Saccharomyces cerevisiae* strains, as well as to RNA-Seq data from an individual human genome. My statistical framework accurately quantifies levels of ASE with specified false discovery rates, achieving high reproducibility between independent sequencing platforms. I pinpoint loci that show unusual and biologically interesting patterns of ASE, including allele-specific alternative splicing and transcription termination sites. This methodology provides a rigorous, quantitative, and high-resolution tool for profiling ASE across whole genomes.

3.2 Introduction

Gene expression is the fundamental initial step in the process by which static genomic information gives rise to dynamic organismal phenotypes. Variation in gene expression has the potential to contribute significantly to phenotypic diversity within species and diver-

gence between species [139, 140]. There is a diverse array of well-characterized examples of phenotypes influenced by regulatory polymorphisms, ranging from pelvic morphology in sticklebacks [10] to malaria susceptibility in humans [141]. Although heritable variation in gene expression levels appears to be ubiquitous among individuals within species, an understanding of the distribution of regulatory variation and the mechanisms by which regulatory polymorphisms act remains limited [142, 143].

Heritable differences in gene expression between individuals are ultimately caused by polymorphisms that affect the expression level of either one or both alleles (*cis*-acting or *trans*-acting polymorphisms, respectively) in a diploid. A powerful approach for identifying *cis*-acting regulatory variation is measuring allele-specific expression (ASE). An observation of differential allelic expression in a heterozygote indicates the presence of one or more variants that act in *cis* to affect the expression level of the gene. ASE has been studied by an assortment of methods, including variations of PCR [144, 145], pyrosequencing [146], array-based platforms [147–150], and chromatin immunoprecipitation [151]. A unifying theme of these studies has been that ASE is widespread both within and between a wide variety of species.

More recently, high-throughput sequencing has been used to assess ASE [152–159], which affords a number of advantages compared to previous approaches. An RNA-Seq approach, where cDNA is isolated and subjected to high-throughput sequencing, gives measurements of ASE genome-wide for both protein-coding genes and non-coding RNA, provided transcribed polymorphisms are present to distinguish between alleles. Sequencing also offers an improved dynamic range over microarrays, and results in digital allele counts with precision limited only by depth of coverage. Despite these advantages, inferences about ASE from high-throughput sequencing data have been made with simple statistical methods, which do not efficiently use all of the information contained in these large and complex datasets.

To address the lack of statistical methods for detecting ASE tailored to high-throughput sequencing data, I developed a Bayesian hierarchical model to analyze allelic read counts. I demonstrate that, compared to existing approaches, this model is more powerful, accurately quantifies false discovery rates (FDR), and facilitates more meaningful biological inferences. I use this method to characterize the landscape of ASE in a diploid hybrid of two diverse

strains of *S. cerevisiae* and find that my data are consistent with an overall proportion of nearly 80% of measured genes exhibiting ASE, 1,991 of which are significant at FDR = 5%. Using this statistical model, I also identified numerous genes with biologically interesting examples of ASE, including allele-specific alternative splicing and transcription termination sites. In addition, to highlight the advantages of this approach for more complex genomes, I applied my method to an RNA-Seq data set in humans. Overall, my analysis highlights the utility of using a carefully designed statistical framework to leverage the massive amount of information present in RNA-Seq datasets to reveal biological insights.

3.3 Methods

3.3.1 Experimental design

The experimental portions of this chapter were performed by Marnie Johansson and Jennifer Madeoy. Marnie and Jennifer mated strains BY4716 and RM11-1a and used auxotrophic deletions to select for the diploid hybrid during mating. These strains have been described in detail elsewhere [35]. They grew the strains to mid-log phase (OD_{600} 0.8-1.0) in rich media (YPD), extracted RNA by the acid phenol method [109], and confirmed RNA integrity using an Agilent 2100 Bioanalyzer (Agilent Technologies). They extracted genomic DNA using a modified version of the yeast smash and grab protocol [160].

I provide a brief overview of sequencing library preparation here, and give full details with kit numbers in Appendix C. Marnie and Jennifer prepared genomic DNA libraries according to manufacturer recommended protocols. For all RNA samples, they performed poly(A) enrichment and one round of ribosomal RNA depletion. For RNA samples submitted to the Illumina Genome Analyzer II, they fragmented RNA to 60-200 bp, made cDNA by random priming, and followed manufacturer recommended protocols for the remainder of sequencing library preparation. For RNA samples submitted to the Applied Biosystems Inc. (ABI) SOLiD machine, they prepared libraries according to manufacturer recommended protocols. All SOLiD samples were tagged with four barcodes per library.

3.3.2 Yeast allele-specific read mapping

I obtained complete genome sequences for BY from the *Saccharomyces* Genome Database (June 2008 sequence; `http://www.yeastgenome.org/`) and for RM from the Broad Institute (`http://www.broadinstitute.org/`). After repeat masking [161] the sequences, I used `LASTZ` (`http://www.bx.psu.edu/miller_lab`) to infer alignment scoring parameters appropriate for aligning the BY and RM genomes and to generate pairwise alignments between all chromosomes of the two strains. I then used `TBA` [162] to compute a whole-genome alignment that is not biased in favor of any particular reference genome. I masked any nucleotides that were ambiguous in either genome, projected this alignment to both BY and RM to construct reference genomes for the strains, and mapped all reads to both genomes. To align reads in colorspace or nucleotide space I used the program `BFAST` [163] (Section C.4).

Next, I examined the alignment of each read to the BY genome and to the RM genome in order to search for reads with distinguishable allelic origin. I analyzed only the highest-scoring alignment of each read to each genome. I required reads to map to approximately the same genomic location in BY and RM; specifically, I required each read to map within the same alignment block in each strain. I used a simple probabilistically-motivated, base quality-aware scoring scheme implemented in the program `cross_match` (`http://phrap.org/phredphrapconsed.html`) to score the alignment of the read to the genome of each strain (Section C.4), and considered a read to be a candidate BY read if the score was higher for the alignment to the BY genome, and vice versa. Any read with an alignment to one genome that scores higher must overlap a SNP, indel, or chromosomal breakpoint between the strains. At a small proportion of SNPs, read mapping is strongly biased toward one of the two alleles, as has been noted previously in humans [152]. To overcome this potential source of bias, I simulated 50-bp reads with sequencing errors overlapping every SNP and indel ascertained from our whole genome multiple alignment of BY and RM, and mapped the simulated reads using the same pipeline described above (Section C.4). For our experimentally acquired data, I then filtered out all allelically-mapped reads that overlapped a SNP showing deviation greater than 5% from equal mapping of alleles in my simulated

reads. To assign reads to genes, I used gene annotations from the *Saccharomyces* Genome Database, along with 5′ and 3′ UTRs predicted by RNA-Seq [47]. I ignored SNPs or indels that occurred within more than one overlapping genomic feature. For reads that overlapped multiple SNPs, I randomly assigned the read count to one of the SNPs. It has been noted by other investigators that base composition has a significant effect on the propensity of a molecule to be sequenced using high-throughput sequencing technologies [159,164,165]. This phenomenon could affect my results only when the BY and RM alleles at a particular locus differ greatly in base composition (which is rare), since my analysis only compares relative allelic expression. Nevertheless, I performed a correction for GC content by (1) calculating expected sequencing depth for windows of a given GC content using our genomic DNA data, and (2) adjusting relative RNA read counts based on the difference in predicted read depth between fragments of BY or RM allelic GC content (Section C.5).

Finally, I removed any reads marked as potential PCR duplicates to ensure that differential allelic expression was not due to differential allelic amplification. For our Illumina single-end and paired-end data, I used Picard's `MarkDuplicates` command-line tool (`http://picard.sourceforge.net/`). For our ABI SOLiD data, I took advantage of the four molecular barcodes tagging each sequencing library. Since the barcodes are embedded in bridge primers used for PCR amplification, reads possessing different barcodes must originate from distinct molecules. As such, for each genomic position I kept a maximum of one read per barcode, and marked the remaining reads as PCR duplicates.

3.3.3 Human allele-specific read mapping

I obtained four lanes of RNA-Seq data (two 35 bp and two 46 bp single-end datasets) generated by Pickrell et al. [159] for individual NA18498. This individual had the most RNA-Seq reads of any sample sequenced by Pickrell et al. [159]. I obtained phased genotype information from the International HapMap Project [166]. I mapped reads to the reference human genome (hg18/build 36) using the program `GSNAP` version 2011-03-11 [167], which features SNP-tolerant alignment. I also took advantage of `GSNAP`'s ability to detect splicing events using a database of known splice junctions compiled using Ensembl gene annotations

[168]. I ran `GSNAP` with the options `--use-snps`, `--splicesites`, `--max-mismatches=0.05`, `--npaths=1`, `--trim-mismatch-score=0`, and `--quiet-if-excessive` to obtain unique alignments of each read. To ensure that `GSNAP`'s SNP-tolerant alignment feature eliminated the mapping bias in favor of the reference allele [152], I simulated reads (35 and 46 bp in length, the lengths of the actual reads) of both alleles at every position overlapping the SNP. I mapped these reads to the human genome using the same commands used to map the real data. I found that mapping bias was completely eliminated for all but a small number of SNPs ($\approx 2,600$ SNPs, or $\approx 1.5\%$ of all SNPs), which I removed from further consideration.

In order to obtain allele-specific read counts, I grouped SNPs by Ensembl-annotated gene and examined any genic SNPs overlapping a mapped read. I assigned reads as originating from haplotypes A or B (as defined by the phased HapMap data; labels are arbitrary for my purposes). For reads overlapping multiple SNPs, I randomly chose a single SNP and incremented the read count for that SNP. This procedure results in allele-specific read counts for SNPs within each gene that are stratified as originating from either haplotype A or B, which served as input to my Bayesian model. As the RNA-Seq data from Pickrell et al. [159] was not accompanied by genomic DNA sequence data, I used the same estimates of the dispersion in read counts as my full analysis of the yeast data (i.e. the analysis of Illumina and ABI data using estimates derived from yeast genomic DNA sequencing data).

3.3.4 Statistical analysis

I used the `R` statistical environment for all statistical analyses [111]. For my initial analysis of allelic count data using the binomial exact test, I used the `binom.test` function. I provide a brief summary of my Bayesian hierarchical model here, and a detailed description in Section C.3. I construct a three-stage hierarchical model for allelic read counts. I denote the count of reads mapping to RM at SNP j in gene i and replicate r as Y_{ijr}, and in the first stage model these counts are binomially distributed with parameters N_{ijr} (coverage at the SNP) and p_{ij}. At the second stage, the p_{ij} arise from a gene-specific beta distribution with parameters α_i and β_i. The second stage allows for the possibility that p_i may not be constant across all SNPs within gene i. These steps can be collapsed to give a beta-binomial model. I

re-parameterize the beta distribution as $p_i = \alpha_i/(\alpha_i + \beta_i)$ and $e_i = 1/(1 + \alpha_i + \beta_i)$, which have straightforward interpretations as the mean amount of ASE (p_i) and the dispersion around the mean (e_i) for gene i. As the dispersion e_i approaches zero the counts converge to binomially distributed. Finally, I place a two-component mixture prior on p_i, e_i:

$$p_i, e_i|\hat{a}, \hat{d}, f, g, h, \pi_0 \sim \begin{cases} \text{Beta}(\hat{a}, \hat{a}) \times \text{Beta}(1, \hat{d}) & \text{with probability } \pi_0 \\ \text{Beta}(f, g) \times \text{Beta}(1, h) & \text{with probability } 1 - \pi_0 \end{cases}$$

The parameters $\hat{a}$ and $\hat{d}$ are estimated from genomic DNA data, and provide a measure of the "noise" in read counts due to technical variability. I estimated these parameters separately using genomic DNA data from each technology platform, and found that the estimates were similar (95% credible intervals overlapped), so in my analysis of data from both platforms I used the median of all posterior samples as our estimate for these parameters: $\hat{a} \approx 3600$ and $\hat{d} \approx 550$. I implement this model using MCMC, running multiple Markov chain simulations for at least 500,000 iterations and examining time series plots of model parameters to verify convergence. For any list of $i = 1, \ldots, n$ genes (out of m total genes) and $s = 1, \ldots, s$ draws from the posterior distribution of each parameter obtained via MCMC, if one lets $\boldsymbol{\theta} = (f, g, h, \pi_0, \hat{a}, \hat{d})$, the FDR achieved when calling those genes significant can be calculated using the formula

$$\text{FDR} = \sum_{i=1}^{n} 1 - p(\text{C2}|y), \text{where}$$

$$\Pr(\text{C2}|y) = \frac{1}{S} \sum_{s=1}^{S} p(\text{C2}|p_i^{(s)}, e_i^{(s)}, \boldsymbol{\theta}_i^{(s)}).$$

In this formula, C2 signifies component 2 of the two-component mixture prior described above and is calculated using Bayes' formula as detailed in Section C.3.

3.4 Results

3.4.1 Allele-specific expression in the yeast genome

With assistance in the wet lab from Marnie Johansson and Jennifer Madeoy, I measured ASE by RNA-Seq in a diploid hybrid of two diverse strains of *S. cerevisiae*, the laboratory strain BY4716 (BY, isogenic to S288C) and the wild vineyard strain RM11-1a (RM). I obtained

Table 3.1: Information on RNA-Seq datasets

Platform	Read length	Samples	Paired end?	Mapped reads (millions)	ASE reads (millions)[a]	Genes > 10X coverage[b]
ABI SOLiD	50 bp	2	No	51.78	1.19	4,483
Illumina GAII[c]	76 bp	1	Yes	21.89	2.16	3,899

[a] Reads assigned as originating from either the BY or RM allele, overlapping a BY/RM variant site that was not susceptible to biased read mapping, and not marked as a potential PCR duplicate.

[b] Coverage is defined here as the number of ASE reads that overlapped at least one base between the gene's annotated start and end coordinates.

[c] Reads obtained were 2×76 bp; each paired-end read is counted as a single read.

sequence data from two independent high-throughput sequencing platforms, the Applied Biosystems Inc. (ABI) SOLiD System and the Illumina Genome Analyzer II (GAII; Table 3.1). To eliminate any read-mapping bias [152], which could lead to erroneous inferences of allelic differences in transcript abundance, I developed strict criteria to call reads as matching the BY or RM allele (Section 3.3). Furthermore, an insidious possible source of bias is differential allelic amplification that would manifest as differential allelic expression. To correct for this, I removed any reads that were potential PCR duplicates. Although this is a conservative approach for single-end reads, I found that naively including duplicate reads induced both global and gene-specific artifacts (Section C.7). Overall, I obtained allele-specific read counts for approximately 4,500 protein-coding genes and non-coding RNAs in the yeast genome, each of which is expressed in rich media and contains at least one transcribed polymorphism.

As a first attempt at quantifying ASE in the diploid hybrid, I summed counts of reads from the BY and RM alleles across all SNPs in each gene and conducted a simple binomial exact test of the null hypothesis that each allele is equally expressed. This is the primary

test that has been employed in previous studies of ASE using RNA-Seq (Section C.8). This test assumes that read counts within each gene are binomially distributed, with a significant test result indicating evidence for ASE. I restricted my analysis to genes with coverage of at least 20 allele-distinguishing reads. 1208 and 1094 genes showed nominally significant ASE (binomial test $p < 0.05$) in our Illumina GAII and ABI SOLiD data sets, respectively. Although the simplicity of this test is appealing, it has several limitations. First, it is not clear how to allow for the possibility of extra-binomial variation in read counts caused by technical variability, as the binomial test cannot be tuned to the context of the experiment. Second, it is not straightforward to combine information from different experiments or replicates to obtain a composite measure of confidence in ASE. Third, it is difficult to calculate an accurate estimate of the FDR, the fraction of genes called as showing ASE that do not truly show ASE. Methods that make use of the complete distribution of p-values [169] to estimate this quantity are difficult to apply with the binomial exact test because the distribution of p-values under the null hypothesis is not uniformly distributed (Section C.2). Finally, summing counts of reads across SNPs to estimate ASE across a gene is undesirable as it masks heterogeneity in ASE at individual SNPs. When properly modeled, such information can provide insights regarding the mechanistic basis of ASE, as we demonstrate below.

3.4.2 A hierarchical Bayesian model for measuring genome-wide allele-specific expression

To address the limitations of standard binomial tests for ASE, I developed a powerful and flexible Bayesian hierarchical model for allelic read count data (Figure 3.1; see Section 3.3.4 for further details). First, I calibrate my model for genes without ASE (Figure 3.1A) using sequence data from genomic DNA, for which allele counts should vary only according to statistical sampling and technical variability. This enables me to allow for some "noise" in allele counts that does not have true biological relevance. Second, the model is motivated by my desire to classify genes according to whether they show ASE and whether patterns observed across SNPs are consistent with a constant level of ASE across the gene. I designed this model to partition genes into two broad categories: genes not showing ASE

(Figure 3.1A), and those showing ASE (Figure 3.1B). Thus, I can directly estimate the global fraction of genes that show ASE in an experiment; I use the notation π_0 to denote the fraction of genes that do not show ASE, with $1 - \pi_0$ representing the fraction that do show ASE. For genes showing ASE, I allow levels of ASE to vary across SNPs, as might be expected for genes with complicated patterns of ASE due to biological mechanisms such as allele-specific splicing, alternative polyadenylation site usage, or alternative transcription start sites (Figure 3.1C). I label genes with varying levels of ASE across SNPs as showing variable ASE. Notably, this model can accommodate multiple replicate datasets from different sequencing platforms in a statistically rigorous manner, while allowing for the possibility of platform-specific estimates of technical variability.

I performed inference with this model using Markov chain Monte Carlo (MCMC). I conducted simulations to explore the power and robustness of my approach compared to the binomial exact test. I simulated read counts with levels of overdispersion (which could be introduced due to technical variability) estimated from our data (Section C.3.4), and calculated the posterior probability of ASE for each simulated gene. I calculated the true positive and false positive rate across thresholds based on p-values for the binomial exact test and posterior probabilities of ASE for our model. Figure 3.2a shows that my model outperforms the binomial exact test across all thresholds. I also calculated a Bayesian analog to the FDR, which accurately represented the true false discovery rate (Figure 3.2b).

3.4.3 Consistent estimates of ASE across different sequencing platforms

As noted above, an important feature of my model is that it can combine replicate data from different sequencing platforms in a statistically rigorous framework. However, before combining all of our sequencing data, I first compared estimates of ASE derived independently from separate analyses of the Illumina GAII and ABI SOLiD datasets. Overall, I found high concordance between results from the two platforms. I obtained more usable allele-specific reads from our Illumina GAII data (Table 3.1), allowing me to call more genes as showing significant ASE than in our ABI SOLiD data. I examined lists of genes called as showing ASE at FDR = 5%, and found 453 genes called significant in both experiments

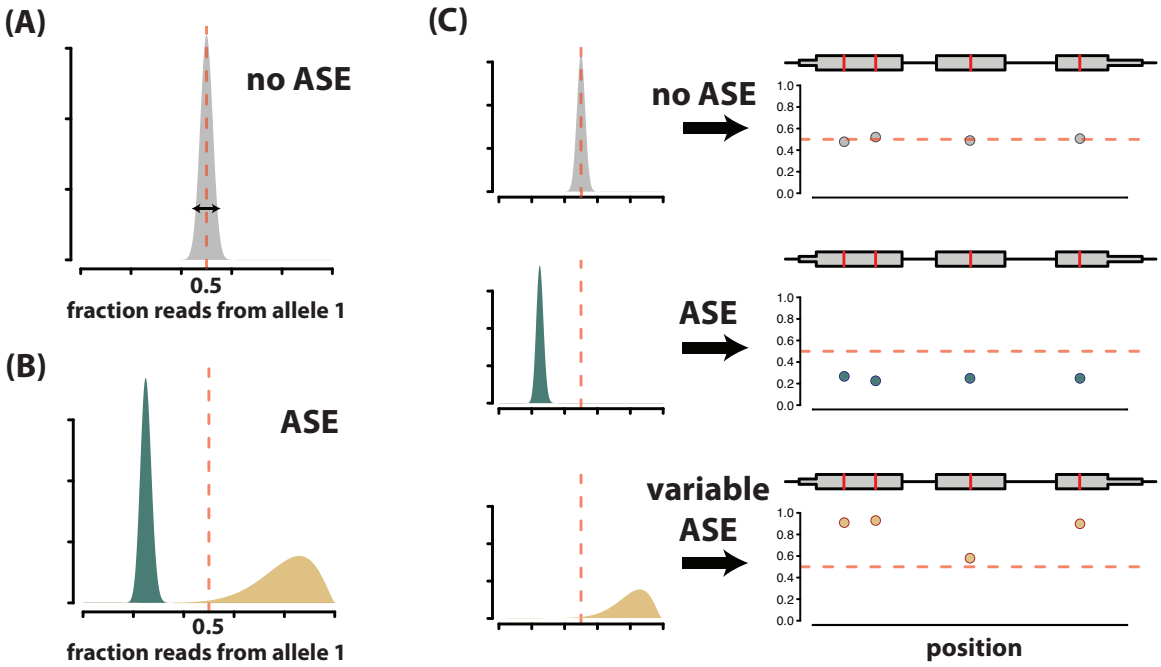


Figure 3.1: Schematic outline of Bayesian model for ASE. (A) The true fraction of reads from allele one should be exactly 0.5 in genomic DNA from a diploid. I use genomic DNA sequencing to calibrate my model in order to account for noise in read counts (arrow depicting width of distribution) at all SNPs as a result of technical variability inherent in the sequencing process. (B) Genes are partitioned into two categories: genes with ASE and those without ASE. For genes without ASE, the distribution of the fraction of reads from allele one is estimated as in (A). I borrow information across all genes to estimate the mean and variability of the corresponding distributions for genes with ASE, the second category. Some genes in this category have a mean different from 0.5 but low dispersion in read counts, like genes without ASE (blue distributions). Other genes in this category have greater dispersion in read counts (tan distributions). (C) Distributions for the fraction of reads from allele one are estimated for each gene. Differences in mean and variability of these gene-specific distributions allow for genes that do not show ASE (top), genes that show ASE that is constant across the transcript (middle) and genes that potentially show complex patterns of ASE (variable ASE), such as allele-specific alternative splicing (bottom). Panels on the left show the gene-specific distributions of the fraction of reads from allele one, while panels on the right show simulated allele-specific read counts for a three-exon gene (grey boxes) containing four SNPs (red lines). Dots below gene model indicate, at each SNP, the fraction of reads matching allele one.

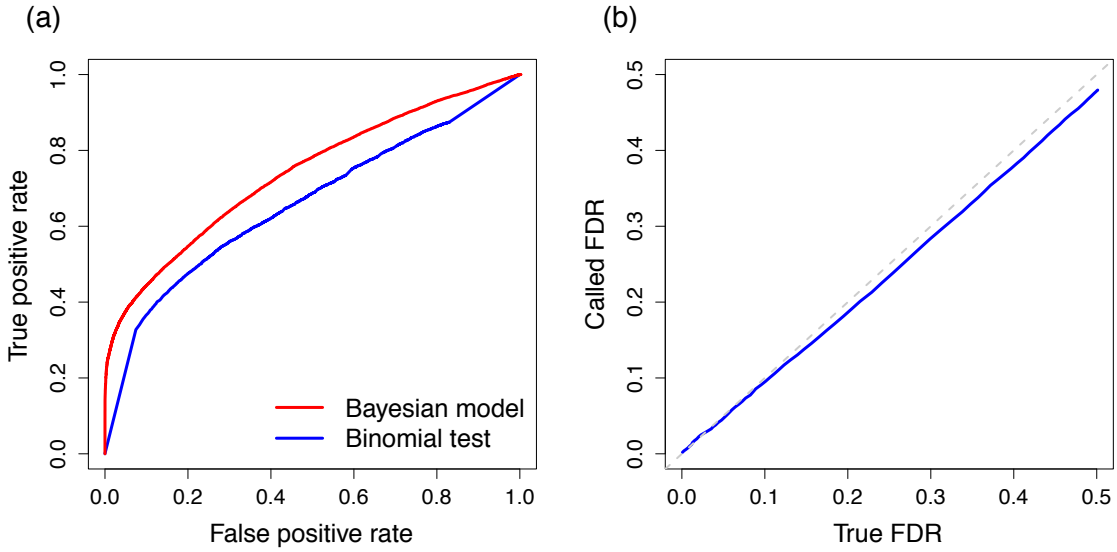


Figure 3.2: Performance of the Bayesian model for ASE. (a) Receiver operating characteristic (ROC) curve showing the performance of my model compared to the binomial exact test. Read counts were tabulated on simulated data with overdispersion as described in the Section C.3.4. ROC curve plots the number of true positives called correctly and the number of false positives called incorrectly using p-value thresholds from 0 to 1 for the binomial test and posterior probabilities of no ASE from 0 to 1 for my Bayesian model. (b) Observed FDR closely tracks the true FDR. Observed and true FDR were calculated for simulated data with overdispersion as described in the Supplemental Methods. The dotted light grey line shows $y = x$.

(Figure 3.3a). Given the incomplete power of each experiment, one would not expect perfect overlap between lists of significant genes. I used estimates of the power and false positive rate for each experiment along with simulations to demonstrate that the observed overlap between experiments is reasonably close to the expected level (Section C.3.5).

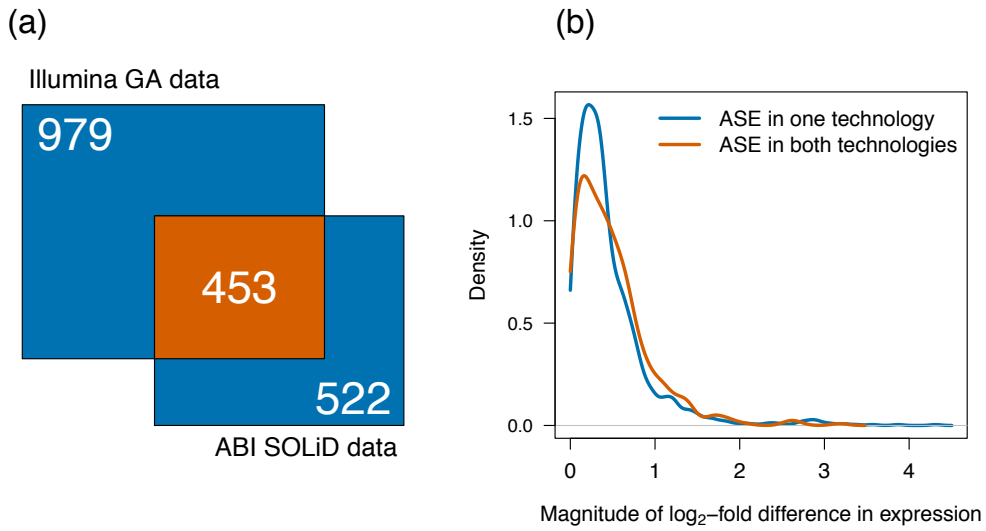


Figure 3.3: ASE on different sequencing platforms. (a) Overlap between genes showing significant ASE at FDR = 5% for two sequencing platforms. The orange square indicates genes called significant in data from both platforms, while blue squares show genes called significant on only one platform, indicated on the far side of the square. Numbers in white indicate the number of genes falling into each category; the area of each square is proportional to this number. (b) Magnitude of $\log_2$-fold difference in gene expression for genes called significant at FDR = 5%. $\log_2$-fold differences are computed with respect to the allele with lower expression, causing all values to be positive. Lines shown are continuous approximations to discrete densities. Blue line shows the density for genes called significant using only one sequencing platform, while orange line shows the density for genes called significant in data from both sequencing platforms.

Genes exhibiting significant ASE in both experiments might be expected to show, on average, more deviation from equal allelic expression than genes with significant ASE in only one experiment. Indeed, the median magnitude of $\log_2$-fold change in expression for genes showing ASE in both experiments was significantly higher than for genes showing ASE in only one experiment (Figure 3.3b; 0.393 vs. 0.337, permutation test $p = 0.0055$). Lastly, I examined the probability that each gene showed ASE in our Illumina GAII dataset and

compared this to the same probability calculated using our ABI SOLiD dataset. I observed a modest but highly significant correlation between these two measures of ASE (Spearmans $\rho = 0.09$; $p = 1.0 \times 10^{-8}$). Given the concordance between our measurements from the two technologies, I focus below on results inferred by simultaneously analyzing the data from both sequencing platforms.

3.4.4 Bayesian hierarchical model reveals features of allele-specific expression

An important advantage of my statistical model is that I am able to leverage the wealth of information contained in an RNA-Seq dataset to make precise inferences about global parameters, averaging over the conclusions drawn from any single gene. For example, one basic quantity of interest is the total fraction of genes that exhibit ASE. This can be computed easily from my model using my estimate of π_0, the fraction of genes that do not exhibit ASE. Combining all of my data, I estimate that approximately 79% of genes interrogated show ASE between BY and RM (95% credible interval 76–83%; Figure 3.4a). I also inferred the distribution of the magnitude of ASE for genes showing ASE, and used this to plot the distribution of fold-change in expression for genes with ASE (Figure 3.4b). In biological terms, this distribution supports the notion that most expression changes are relatively small ($> 90\%$ of genes with ASE show expression changes < 1.5 fold).

Next, I sought to identify which genes showed the strongest evidence for ASE overall, as well as which genes were good candidates for variable ASE. I identified 1,991 genes with significant evidence for ASE (5% FDR, corresponding to posterior probability of ASE > 0.82). Among these genes, 22 are known non-coding RNAs, and the remainder encode proteins. A previous study employed expression levels measured using microarrays and identified 1,428 genes with significant evidence for local eQTL [145]. I obtained allele-specific expression measurements for 1,198 of these genes, and detected significant evidence (5% FDR) for ASE in 637, supporting the assertion that these genes show *cis*-regulatory variation. Additionally, among 43 genes previously verified by quantitative PCR to show ASE [145], I called 30 as showing significant ASE.

I also compared inferences of ASE made using our Bayesian model to those made using

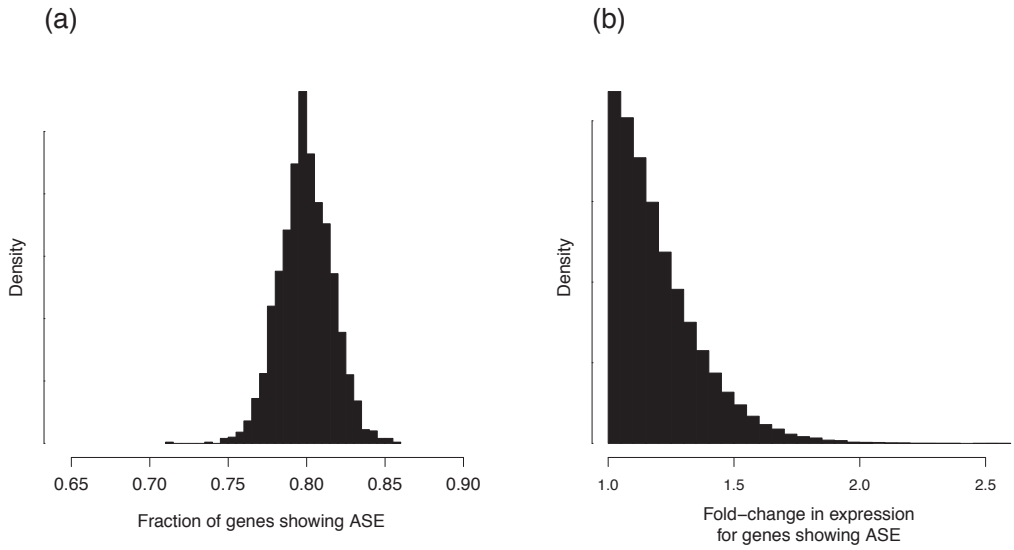


Figure 3.4: Global features of ASE in the yeast genome. (a) Posterior distribution of the fraction of genes showing ASE, $1 - \pi_0$. (b) Posterior distribution for the size of the fold-change in expression of genes showing ASE. Fold-change values are shown relative to the allele expressed at a lower level, meaning that all values are greater than one. Distribution depicts the estimated probability density from which the magnitude of allele-specific expression would be drawn for a new gene known to show allele-specific expression. Distribution shown was simulated by randomly drawing values by technology from posterior samples of a_t and b_t, which determine the shape of the probability density of the binomial parameter p for genes showing ASE.

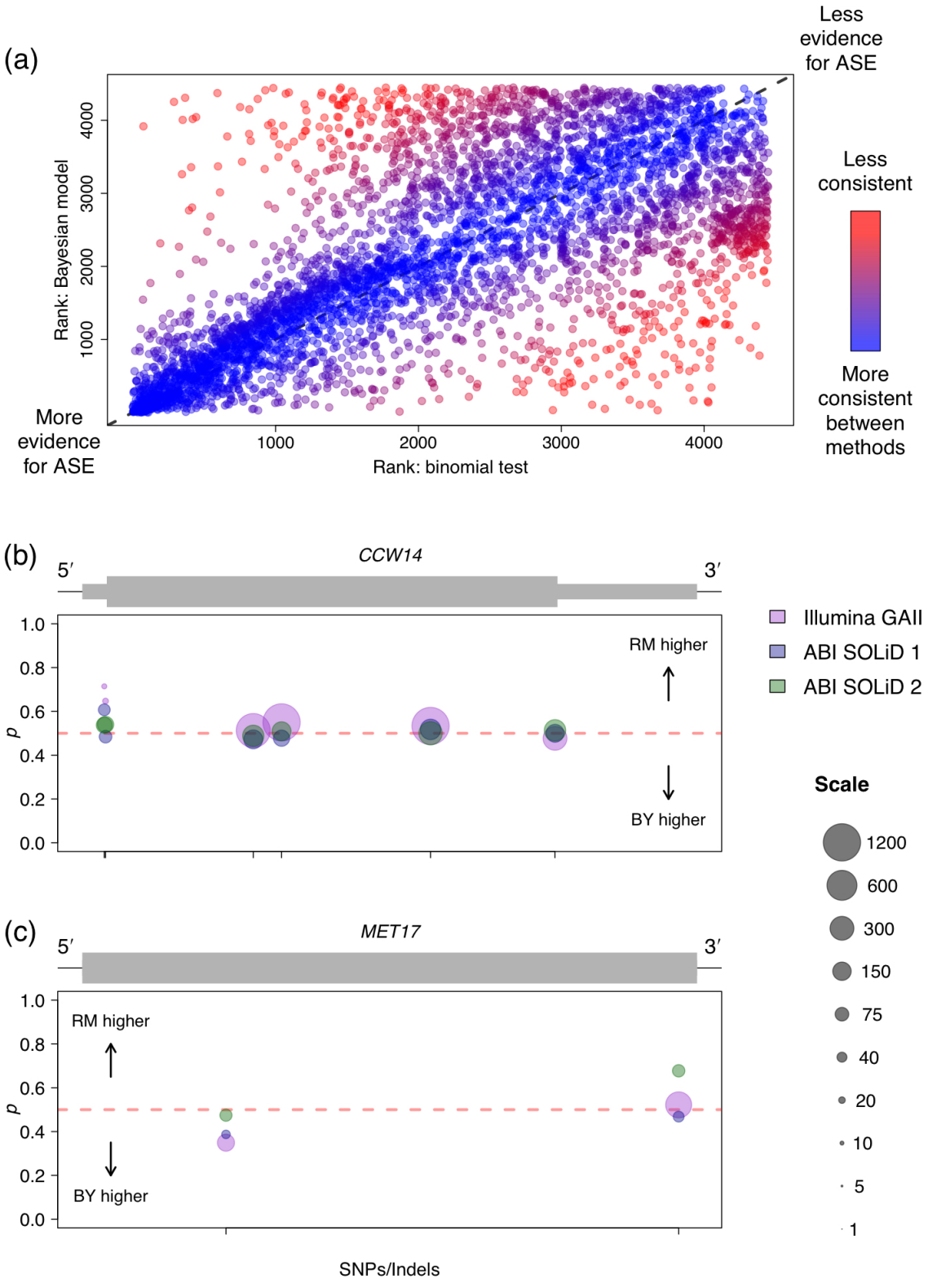
a binomial test of equal allelic expression for read counts summed across all datasets for all SNPs in each gene. As expected, measures of ASE were, on the whole, strongly correlated between methods (Spearmans $\rho = 0.67$; $p < 2.2 \times 10^{-16}$). However, there were many exceptions to this pattern. Figure 3.5a shows a plot of genes ranked by evidence for ASE in the two models, and demonstrates that while most genes have similar rankings (blue points), a non-trivial proportion of genes show highly discrepant results (red points; 1078 genes differ in rank by at least 25% of the total number of genes).

I further examined genes with highly discrepant measures of ASE between methods. There were 192 genes ranked among the top third most likely to show ASE by one method, but the bottom third least likely to show ASE by the other method. Figure 3.5b-c shows examples of two such genes, and illustrates some of the advantages of our method over the binomial test. Figure 3.5b shows the gene *CCW14* (YLR390W-A), a cell wall glycoprotein called as exhibiting significant ASE using the binomial exact test ($p = 1.2 \times 10^{-5}$). The sequence coverage at SNPs within this gene is high and allelic read counts occur at ratios close to 50:50, but there is enough departure from equal allelic expression for rejection of the binomial test. It seems likely that the slight variations from perfectly equal allelic expression are due to technical variability rather than some underlying biological mechanism, and reassuringly, this gene shows no evidence for ASE using our Bayesian model (posterior probability of ASE < 0.4). Figure 3.5c shows the gene *MET17* (YLR303W), a methionine and cysteine synthase called as significant using my Bayesian model (FDR = 5%) but not the binomial test ($p = 0.86$). The data shows a modest but reproducible change in ASE from read counts higher for the BY allele to read counts higher for the RM allele moving $5'$ to $3'$. This example emphasizes the fact that our model can detect variable ASE, while such genes are difficult to identify by the binomial test.

3.4.5 Variable ASE leads to mechanistic insights into allele-specific expression

There are a variety of possible mechanisms that might cause a gene to exhibit variable ASE, such as allele-specific alternative splicing, allele-specific polyadenylation site usage, allele-specific transcription start sites, or allele-specific antisense transcription encroaching

Figure 3.5: Comparison of results from binomial test versus Bayesian model of ASE. (a) Plot comparing ranks of genes in terms of evidence for ASE for the binomial test versus the Bayesian model. Ranks were determined using p-values for the binomial test, and posterior probabilities of ASE for the Bayesian model. Ties were broken by random assignment of ranks to genes with equal p-values/posterior probabilities of ASE. Points are colored according to consistency of ranks between methods. As shown in the color bar to the right, redder points represent genes with ranks that are less consistent between methods, while bluer points show genes ranked more consistently between methods. Dotted grey line in background follows $y = x$. (b) Allele-specific read counts for the gene *CCW14*, which is called as showing ASE using the binomial test, but not with the Bayesian model. Plot depicts the gene model (gray rectangles), with thick rectangles representing exons, thinner rectangles representing $5'$ and $3'$ UTRs, and the thin black line representing intergenic sequence. Circles plotted below the gene model show allele-specific read count data organized by SNPs/indels within the gene. Circles are centered on the point p =(BY count)/(BY count + RM count), and sized according to the total number of reads contributing to the observation. Scaling of circle size follows the scale given on the far right, with all observations with >1200 reads set to the largest size shown (1200). Circle colors indicate which experiment the observation is derived from, as shown in the legend on the far right. Ticks on the x-axis indicate the location of SNPs or indels used to distinguish between alleles. Sequence coverage is high and the slight departures from 50:50 allelic expression are likely due to technical variability rather than some underlying biological mechanism. (c) Allele-specific read counts for the gene *MET17*, which is called as showing ASE using the Bayesian model, but not with the binomial test. Plot is organized and colored identically to (b). The data shows a modest but reproducible change in ASE from read counts higher for the BY allele to read counts higher for the RM allele moving $5'$ to $3'$.


(a)
Less evidence for ASE
Less consistent
More consistent between methods
More evidence for ASE
Rank: Bayesian model
Rank: binomial test
1000
2000
3000
4000
(b)
CCW14
5′
3′
RM higher
BY higher
p
1.0
0.8
0.6
0.4
0.2
0.0
Illumina GAII
ABI SOLiD 1
ABI SOLiD 2
Scale
1200
600
300
150
75
40
20
10
5
1
(c)
MET17
5′
3′
RM higher
BY higher
SNPs/Indels

over a portion of a gene. I found candidate genes showing variable ASE by ranking genes by the parameter e_i, which measures dispersion around the mean level of ASE for all SNPs in a gene (Section C.3). I found examples of genes with variable ASE likely caused by some of the above mechanisms by visually examining read counts at loci that ranked among the 10% most variable (Figure 3.6). Figure 3.6a shows the gene *RPL25* (YOL127W), a protein component of the large ribosomal subunit with read counts consistent with allele-specific alternative splicing. In particular, I observed high read counts and reproducibly equal allelic expression in the second exon of the gene, but lower read counts and expression biased in favor of the BY allele at four SNPs in the intron. My observations are consistent with the sampling of a modest number of immature mRNA transcripts, with the BY allele present at a higher level, and a larger number of mature mRNA transcripts, with equal allelic expression. One possible mechanistic explanation for this observation is that splicing of the BY allele is inefficient, causing either a longer persistence time of immature mRNAs or a higher percentage of intron retention in mature mRNA than for the RM allele.

In addition to allele-specific alternative splicing, I found genes that appeared to demonstrate allele-specific variation in transcriptional start or stop sites. For example, Figure 3.6b shows the gene *SUP35* (YDR172W), which codes for a protein involved in translation termination. We observed nearly equal expression of both alleles within the coding sequence of the gene, but higher expression of the BY allele at a SNP in the $3'$ untranslated region (UTR) of the gene (Figure 3.6b). This observation suggests that the length of the $3'$ UTR may vary between alleles, with a shorter $3'$ UTR associated with the RM allele. Another example of variation in transcript structure is the gene *AFG3* (YER017C), a component of a mitochondrial inner membrane protease. For *AFG3* I observed equal allelic expression at the $5'$ end of the gene, and strong but reproducibly biased expression in favor of the RM allele near the $3'$ end of the gene (Figure 3.6c). This pattern is consistent with premature termination of transcription in the BY background, or a shorter $3'$ UTR associated with the BY allele. I note that, because our data derives from 50-76 bp reads (paired-end reads for Illumina GAII data), observations of ASE at particular SNPs could reflect variation in transcript structure located some distance from the SNP in question. The ability to identify these biologically complicated examples of ASE is an important strength that is unique to

the statistical approach that I developed.

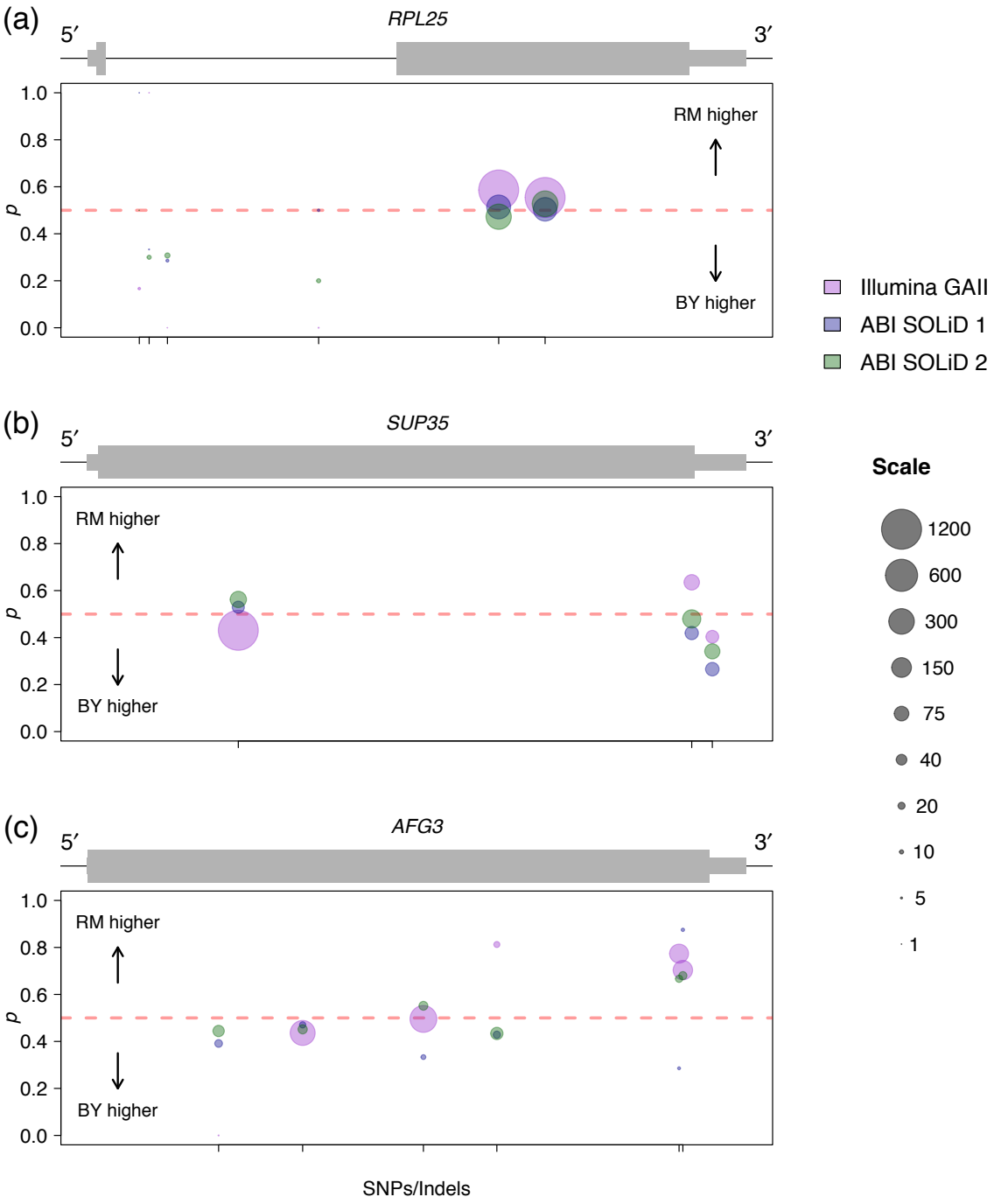
3.4.6 Application to measuring ASE in the human genome

To explore the utility of my method for characterizing ASE in a more complex mammalian genome, I obtained RNA-Seq reads from four lanes on the Illumina GAII generated by Pickrell et al. [159] from an individual of African descent, a member of the Yoruba in Ibadan, Nigeria with high-quality phased genotypes available from the International HapMap Project [166]. This individual is heterozygous at $\approx 164,000$ annotated transcribed sites, and I detected reads with distinguishable alleles mapping to 5,780 genes. This dataset has significantly lower sequencing depth than the yeast data described above, with only 2,082 genes containing 10 or more reads that overlap a transcribed polymorphism. Nevertheless, I conducted the analysis as a proof of principle that my model could be applicable to organisms with larger genomes.

Pickrell et al. [159] conducted a targeted test of 244 genes with significant evidence for local expression QTL to explore whether ASE contributed to expression variation among 69 individuals. In contrast, I carried out a genome-wide survey of ASE in this single individual (NA18498). By performing my analysis on a single individual, I avoided the possible complication of differences in genetic background confounding the relative expression levels of two alleles.

I identified 17 genes with evidence for significant ASE (5% FDR) in individual NA18498. These genes corresponded well to those identified by summing reads across SNPs and performing a binomial test. As it is difficult to calibrate the FDR for the binomial test, I chose a p-value threshold of 0.001 (corresponding to an expectation of approximately 5 expected false positives), which resulted a significant test result for 18 genes. Of these 18 genes, my Bayesian model identified 15 (FDR = 5%). The genes called as showing significant ASE by the binomial test but not using my model all had a high skew in allelic expression but few reads mapping (< 30), while the two genes called as significant by my model were marginally significant by the binomial test ($p < 0.05$). Although I pinpoint 17 genes as showing significant ASE, I estimate the fraction of the complete set of genes tested showing

Figure 3.6: Examples of genes showing variable ASE. Plots are organized and colored identically to Figure 3.5b-c. (a) Allele-specific read counts for the gene *RPL25*. Thin black line represents both intronic and intergenic sequence. Read counts indicate reproducibly equal expression in exon two of the gene, but expression biased in favor of the BY allele at four SNPs within the intron, consistent with allele-specific differences in splicing. (b) Allele-specific read counts for the gene *SUP35*. Higher expression of the BY allele at a SNP in the $3'$ UTR suggests allele-specific variation in UTR length. (c) Allele-specific read counts for the gene *AFG3*. Higher expression of the RM allele near the $3'$ end of the gene is consistent with allele-specific variation in transcript structure that could occur some distance away from the SNP tagging the ASE.


(a)
RPL25
5′
3′
1.0
0.8
0.6
0.4
0.2
0.0
p
RM higher
BY higher
(b)
SUP35
5′
3′
RM higher
BY higher
(c)
AFG3
5′
3′
RM higher
BY higher
SNPs/Indels
Illumina GAII
ABI SOLiD 1
ABI SOLiD 2
Scale
1200
600
300
150
75
40
20
10
5
1

ASE to be approximately 19% (Figure 3.7a; 95% credible interval 11-30%). Although it is difficult to obtain a precise figure for the fraction of genes showing ASE in an individual human due to differences in study design, power, and statistical methodology, this range is generally consistent with previous studies of ASE in humans [150, 170–172].

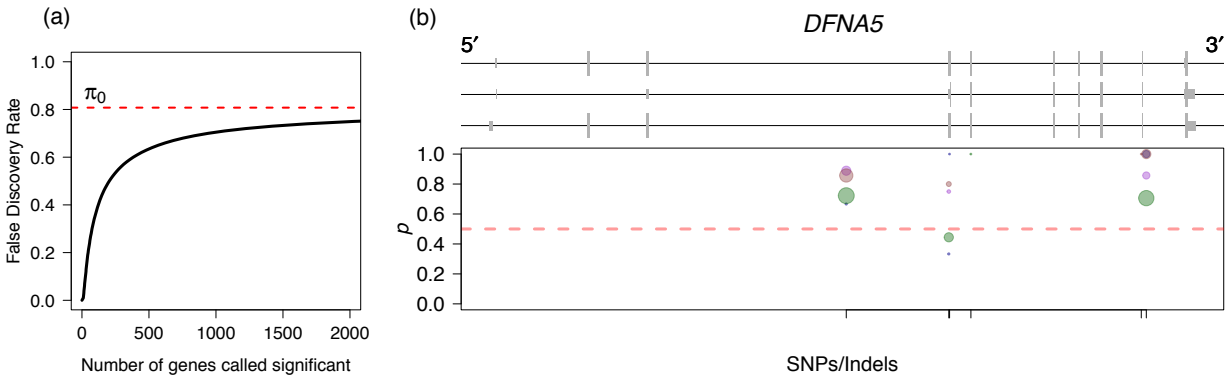


Figure 3.7: ASE in the human genome. (a) Plot of the false discovery rate as a function of the number of genes called significant. Since the human RNA-Seq dataset is low coverage for most genes, it is not possible to identify many genes showing significant ASE without risking a relatively large proportion of false discoveries. (b) Human gene *DFNA5*, which shows significant ASE in individual NA18498. Plot is organized identically to Figure 3.5b-c, with different colored dots representing measurements obtained from separate Illumina sequencing lanes. Although the number of reads is low for any given dot, the proportion of reads from allele one is consistently higher than that for allele two.

I also searched for genes showing complicated patterns of ASE that might inform our understanding of mechanisms of ASE at these loci. Given the low overall coverage of this dataset, I did not find any convincing examples of variable ASE. However, an examination of read counts at multiple SNPs within a gene can still be informative about potential mechanisms of ASE. For example, Figure 3.7b shows the gene *DFNA5* (ENSG00000105928), which has three transcript isoforms and is implicated in non-syndromic hearing impairment in humans [173]. Although read counts at SNPs within this gene are quite low, this gene was called as showing significant ASE (FDR = 5%). As is apparent from Figure 3.7b, the proportion of reads from allele one is consistently high across all SNPs in the gene, with the exception of two points with relatively few reads (green point just below 0.5 represents

a total of only 9 reads). Such read counts would be most consistent with a variant in the promoter affecting transcription initiation or a variant in the 3′ UTR affecting decay rates that acts uniformly across the transcript, rather than allele-specific variation in transcript structure. In the future, advances in sequencing technology and RNA-Seq read-mapping software are likely to lead to datasets with deeper coverage and more accurate reconstruction of transcript structure, which will allow a more complete picture of the landscape of ASE in humans.

3.5 Discussion

In this chapter, I have described a novel method for gaining insight on the genome-wide characteristics of *cis*-regulatory variation and discovering loci with complex patterns of ASE. I demonstrate that inferences of ASE made using different sequencing platforms are concordant, and identify approximately 2,000 genes showing ASE (FDR = 5%) between two diverse yeast strains. My model provides a framework for analyzing allele-specific read count data obtained at multiple SNPs within genes over multiple experimental replicates in a statistically rigorous manner. Combining information from SNPs across the length of a transcript, as well as allowing for technical variation in read counts, are key advantages that allow my model to outperform the binomial test. In addition, I demonstrate that explicitly allowing levels of ASE to vary across SNPs within genes can lead to the identification of genes showing biologically interesting patterns of ASE that may have remained invisible by other analysis methodologies (Figure 3.6). Modeling complicated mechanisms of ASE is likely to be even more critical as we move towards studying ASE in deeply sequenced mammalian transcriptomes, where phenomena such as alternative splicing are pervasive.

A unique strength of my approach is its ability to simultaneously make use of all of the sequence data to infer global parameters of interest. Of the genes that have transcribed polymorphisms, I estimate that nearly 80% exhibit ASE. This estimate is higher than a previous estimate for the same two yeast strains (approximately 20%) based on verification of *cis*-acting regulatory variation by allele-specific quantitative PCR for genes with local eQTL [145]. However, several details of differences in study design and methodology can account for this discrepancy. First, only genes with a transcribed polymorphism can be

assessed for ASE with RNA-Seq, while the estimate of Ronald et al. [145] relied on gene expression QTL that were detected without this requirement. Ronald et al. [145] showed that there is a higher rate of local regulatory variation (most of which acts in *cis* to produce ASE) in more polymorphic regions of the yeast genome. Thus, my estimate is likely higher in part due to measurements made on genes found in regions of the yeast genome ascertained to have a high occurrence of *cis*-acting regulatory variants. Second, microarray measurements of gene expression levels may miss some of the transcript variants that we detect and classify as variable ASE if probes are designed to regions of the gene with equal allelic expression. Finally, I note that RNA-Seq affords the opportunity to measure transcript levels with very high precision [49]. Given the large number of polymorphic noncoding sites found between BY and RM ($> 30,000$), it may be that nearly every gene in the genome shows some level of ASE when measured with sufficient precision, which raises a fundamental question: what level of ASE is biologically significant? In the future, it will be critical to move beyond describing and cataloging variation in transcript levels toward a more complete understanding of the functional relevance of expression variation.

Finally, although I applied our statistical methodology to study ASE, this framework is general and can be used to characterize allelic differences of any functional genomics phenotypes derived from sequence data, such as methylation [174] or protein-DNA interactions [175]. As new applications of high-throughput sequencing are conceived [176], it will become increasingly important to develop statistical methods tailored to these large and formidably complex data sets in order to maximize the biological insights derived from such experiments.

Chapter 4

PHENOMICS ANALYSIS OF GENETICALLY DIVERSE *SACCHAROMYCES CEREVISIAE* STRAINS

I gratefully acknowledge the contributions of colleagues from the Yeast Resource Center for their roles in the generation of the data described in this chapter and suggestions of interesting possible analyses.

4.1 Summary

In contrast to our understanding of genetic variation, our knowledge of the quantitative characteristics of phenotypic diversity remain poorly defined. Here, I describe a comprehensive phenomics dataset of over 15,000 gene expression, protein, metabolite, and cellular morphological phenotypes in 22 genetically diverse strains of *Saccharomyces cerevisiae*. Our deep phenotyping data expand the parts list of this well-studied model organism, as I identified novel or previously unconfirmed transcripts, introns, and peptides. Over 50% of all measured traits vary significantly across strains and $\approx$ 85% of all phenotypes are highly correlated ($\rho > 0.76$; FDR=5%) with at least one other trait (median = 6, maximum = 328). I identified 366 robust associations between genetic variants and phenotypes. Finally, I show how high-dimensional molecular phenomics datasets can be leveraged to accurately predict phenotypic variation between strains, often with greater precision than afforded by DNA sequence information alone. My results provide new insights into the spectrum of phenotypic diversity that exists in natural populations, the genetic basis of such diversity, and the characteristics influencing the ability to accurately predict particular phenotypes.

4.2 Introduction

Considerable progress has been made in characterizing genomes, allowing comprehensive insights into patterns of genetic diversity in many organisms [1–5]. Interpreting the functional and phenotypic consequences of genetic variation, however, remains challenging, and

is exacerbated by the paucity of data on the quantitative characteristics of phenotypes. One approach to bridge the gap between genetic variation and organismal phenotypes is the comprehensive and systematic collection of carefully measured phenotypes, an approach referred to as phenomics [177–179]. To date, phenomics studies have often been limited in either the number of phenotypes or individuals studied [180–183]. However, advances in functional genomics technology, instrumentation, and computational biology are providing the necessary tools to extensively phenotype large numbers of individuals.

In this chapter, I describe a comprehensive and carefully constructed phenomics dataset consisting of >15,000 molecular and morphological traits collected in 22 genetically diverse yeast strains. These strains represent a diverse collection of *S. cerevisiae* isolates sampled from six continents and a wide variety of microenvironments (Table D.1). This data reveals new insights into the patterns and determinants of phenotypic variation and will be a powerful community resource that fosters a deeper understanding of the principles governing the relationship between genotypes and phenotypes.

4.3 Methods

Due to the collaborative nature of this project, I was not directly involved in most aspects of the data generation. In this chapter, I have largely restricted the methods described to those I was directly involved in, and I include details of experimental and computational analyses conducted by others in Appendix D.

4.3.1 Whole-genome sequencing

See Section D.2.1 for a description of sample preparation and DNA sequencing methodology.

Genotyping

I used `BWA` version 0.5.9 [184] to map DNA sequence reads to the S288c reference genome (release 64/UCSC sacCer3), after substituting non-reference nucleotides at sites of known SNPs [1] for each strain where appropriate. Specifically, to construct these strain-specific reference sequences I employed sequence data generated by Liti et al. [1] and only considered

sites to be high quality non-reference alleles if the base quality exceeded 30 (error probability 0.001) and the site passed "neighbourhood quality standard" [1]. After mapping reads, I sorted BAM files and marked duplicate reads using `Picard` version 1.29 (`http://picard.sourceforge.net`).

I performed indel realignment and base quality recalibration, followed by SNP calling, using `GATK` version 1.0.5454 [185]. Since `GATK`'s UnifiedGenotyper tool is designed for calling SNPs in diploid genomes, I implemented a simple minimum overall depth filter of four reads (that passed `GATK`'s internal quality control metrics) and, at each passing site, manually extracted likelihoods for the two homozygous genotypes. I adopted a conservative approach and required the likelihood of the non-reference allele (i.e. homozygous non-reference) to be at least 1,000 times greater than the likelihood of the reference allele (i.e. homozygous reference) before calling the site as non-reference. Manual inspection of sites where the likelihood of the non-reference allele was only slightly higher than that of the reference allele did not provide convincing evidence for the true presence of a SNP. The precise threshold (of non-reference to reference likelihoods) used affects only a small minority of SNPs (<200 SNPs have a likelihood ratio between 1 and 1,000 in 22/23 strains). I did not attempt to call insertions, deletions, or large structural variants using our short-read data.

Preparing strain-specific reference genomes

In order to prepare strain-specific reference genomes for mapping RNA-Seq reads, I extracted my genotype calls at every base in each strain's genome. Starting with strain-specific reference genomes constructed by combining sequence data from Liti et al. [1] with the S288c reference genome (above), I replaced nucleotides with the correct allele at sites where I called a different allele. At sites where I had insufficient short-read data to produce a genotype, I kept the strain-specific reference nucleotide.

To prepare strain-specific databases for peptide matching, I extracted sequences from the strain-specific reference genomes using the coordinates of annotated genes in the reference S288c genome, then translated these sequences into protein sequences.

Phylogeny based on whole-genome sequence

I present a phylogeny of the strains in Figure 4.4B. To obtain this phylogeny, I first started with complete genomes of *S. cerevisiae* from *Saccharomyces* Genome Database [106] and *S. paradoxus* from Liti et al. [1]. I used the program `TBA` [162] to align these two genomes, and projected the resulting MAF file to the *S. cerevisiae* reference. Since the strain-specific reference genomes that I constructed (Section 4.3.1) were in the same genomic coordinates as the *S. cerevisiae* reference genome, I appended these genomes on to the MAF file produced by `TBA`. I used the `ape` version 3.0-5 package for phylogenetic manipulations [186] in `R` [187].

4.3.2 Normalization and data analysis

Overall, I implemented a two-stage model to test for differential abundance of RNA, protein, and metabolite levels. In the first stage, I used a linear model to remove effects due to batch and other factors not of primary interest, as detailed below for each data type. I then obtained normalized data values by extracting residuals from this model. In the second stage, I considered each gene separately, and tested for a strain effect using a random effects model with normalized data values from stage one used as input. I used a similar approach to test for differences in morphological traits, but modified the approach slightly because we had measurements from many cells for each trait. I used `R` [187] for all statistical analysis reported below.

RNA-Seq

I mapped RNA-Seq reads to strain-specific reference genomes (Section 4.3.1) using the program `BFAST` version 0.6.4e [163] with options `-K 100` and `-M 500` to `bfast match`. I aligned colorspace reads using a main index with mask `111111111111111111` (hash width 14) and secondary indexes with masks `1111101110111010100101011011111`, `10111101011010010-1100001101000111111`, and `1011100110100110010011111010100010111111` (all using hash width 14). I output the results in SAM format and converted to BAM format using `samtools` [188].

I computed strand-specific read depth across all annotated *S. cerevisiae* genes and non-

coding RNAs using `bedtools` version 2.15.0 [189], after filtering out reads with a mapping quality below 30. I normalized for RNA-Seq library size using normalization factors calculated using the trimmed mean of M-values method [190] as implemented in the `edgeR` package [191]. While examining potential batch effects, I noticed a particularly strong dependence on RNA-Seq flowcell – samples run on one flowcell consistently grouped separately from samples run on the other two flowcells (Figure 4.1). As such, I used the ComBat

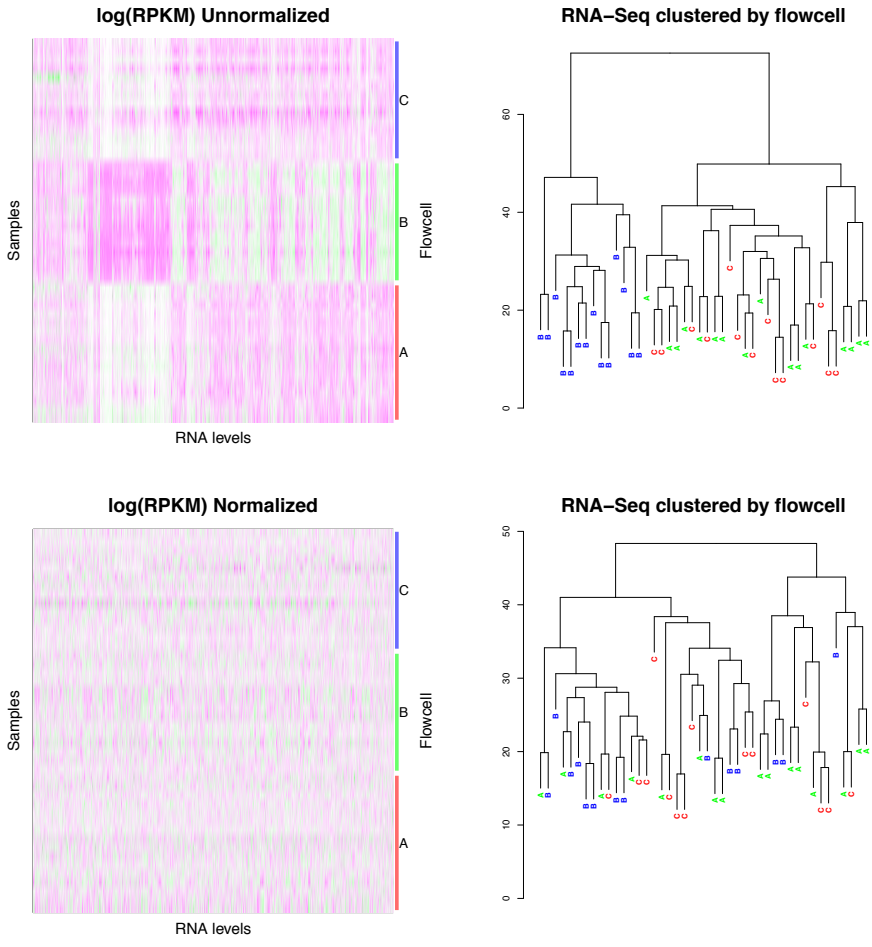


Figure 4.1: RNA-Seq flowcell batch effects. Top heatmap and dendrogram show RNA-Seq data plotted before flowcell normalization. In heatmap, magenta indicates lower expression, green higher expression, and white intermediate; samples are grouped by flowcell as shown on the right y-axis. It is evident from the top heatmap and dendrogram that samples are strongly clustered by flowcell; this effect is removed in the bottom heatmap and dendrogram.

function in the `sva` package [192] to explicitly correct for flowcell batch effects. Next, I performed quantile normalization [193], and log-transformed the counts to better approximate a normal distribution. In order to remove unknown or unmeasured sources of variation from the data, I ran `sva` [194]. In order to remove known potential batch effects, I used a fixed-effects linear model with the following covariates: RNA preparation batch, chemostat harvest round, chemostat in which sample was grown, and chemostat processing batch. I tested for differential expression by using gene-specific linear mixed models with a single covariate, strain, specified as a random effect and residuals from the fixed-effect normalization model as the response variable. To evaluate significance of gene expression differences, I calculated a likelihood ratio statistic for each gene, performed 10,000 permutations of the strain labels (refitting the model and recalculating the statistic for each permutation), and calculated a p-value based on this empirical null distribution.

Quantitative proteomics

Starting with the quantitative peptide abundance data output by `Topograph` (described in Section D.2.4), I divided by the total area of each sample to normalize for differences in total protein abundance. Since ionization efficiencies in the mass spectrometer are peptide-dependent, it would be difficult to compare abundances of peptides that differ in amino acid sequence between strains. As such, I focused only on peptides that matched uniquely to the same single protein in all strains, which constituted the majority of the dataset, 72% (6845/9497) of measured peptides. I also examined the amino acids immediately before and after each peptide sequence in all strains, discarding any peptides where a polymorphism could have affected a tryptic site.

I performed quantile normalization [193], and log-transformed the measurements to better approximate a normal distribution. In order to remove unknown or unmeasured sources of variation from the data, I ran `sva` [194]. In order to remove known potential batch effects, I used a fixed-effects linear model with the following covariates: protein lysis batch, protein digestion batch, mass spectrometer run order, chemostat in which sample was grown, chemostat harvest round, chemostat processing batch, and time post-quality control run

that the sample was run. I tested for differential peptide abundance by using peptide-specific random effects models with a single covariate, strain, specified as a random effect and residuals from the fixed-effect normalization model as the response variable. To evaluate significance of peptide abundance differences, I calculated a likelihood ratio statistic for each peptide, performed 10,000 permutations of the strain labels (refitting the model and recalculating the statistic for each permutation), and calculated a p-value based on this empirical null distribution.

Metabolomics

Starting with the metabolite abundance data output by `Guineu` (described in Section D.2.5) I first log-transformed the measurements to better approximate a normal distribution. I noted that the data showed a strong batch effect related to the instrument on which the samples had been run (with one instrument in Seattle, WA and a second in Huntsville, AL). I used the ComBat function in the `sva` package [192] to explicitly correct for this batch effect. There were metabolites missing from some samples so I used a nearest-neighbor averaging approach to impute the missing data [195]. In order to remove unknown or unmeasured sources of variation from the data, I ran `sva` [194]. In order to remove known potential batch effects, I used a fixed-effects linear model with chemostat in which sample was grown and chemostat harvest round as covariates. I tested for differential metabolite abundance by using metabolite-specific random effects models with a single covariate, strain, specified as a random effect and residuals from the fixed-effect normalization model as the response variable. To evaluate significance of metabolite abundance differences, I calculated a likelihood ratio statistic for each peptide, performed 10,000 permutations of the strain labels (refitting the model and recalculating the statistic for each permutation), and calculated a p-value based on this empirical null distribution.

Morphology

Since we measured hundreds or thousands of cells for each trait, I fit a separate linear mixed model for each trait to correct for potential batch/technical effects and test for

strain differences in traits. I tested for differences between strains in 199 directly measured traits and 199 traits consisting of the coefficient of variation (CV) of the directly measured traits [196], for a total of 398 traits.

For directly measured traits, I fit a mixed effects model to test for differences between strains; the response was the measured trait values and covariates included a fixed chemostat fermenter effect, a fixed image acquisition date, and a random strain effect. The null model consisted of the same setup with the random strain effect excluded. To evaluate significance of morphological differences, I calculated a likelihood ratio statistic for each trait, performed 10,000 permutations of the strain labels (refitting the model and recalculating the statistic for each permutation), and calculated a p-value based on this empirical null distribution.

For CV traits, I obtained residuals from a fixed effect model containing chemostat fermenter and image acquisition date as covariates (the null model above). I then calculated the coefficient of variation from these corrected values to obtain a single value (the CV) per sample. I tested for strain differences by using trait-specific random effects models with a single covariate, strain, specified as a random effect and the calculated coefficient of variation values for each sample as the response variable. To evaluate significance of differences in morphological trait coefficient of variation, I calculated a likelihood ratio statistic for each trait, performed 10,000 permutations of the strain labels (refitting the model and recalculating the statistic for each permutation), and calculated a p-value based on this empirical null distribution.

4.3.3 Network structure of phenotypic correlations

Hive plot

I collapsed peptide measurements into a single value for each protein by taking the average. I used nearest neighbor averaging to impute missing values (in the metabolite and morphological data) [195]. I combined the resulting 1,645 protein values with 6,207 gene expression values, 115 metabolite values, and 398 morphology trait values. I calculated Spearman correlations between each pair of phenotypes, resulting in >34,000,000 total correlations. I permuted the dataset and recalculated correlations, finding that a cutoff of $\rho = 0.7625$ corre-

sponded to a FDR of approximately 5%. I modified code from the `HiveR` package, available at `http://academic.depauw.edu/~hanson/HiveR/HiveR.html`, to create hive plot figures.

Subnetworks

I used the `MCODE` [197] plugin in `Cytoscape` [198], which detects subsets of densely interconnected nodes in a network, to identify highly connected subnetworks. I used the default settings: degree cutoff = 2, node score cutoff = 0.2, k-core = 2, max depth = 100.

4.3.4 Identification of novel and unconfirmed biomolecules

Estimating the transcribed fraction of the genome

I used `gem-mappability` [199], with read length 50 and max mismatches set to two, to compute the mappability at each base in the yeast genome. Here the mappability is defined as the number of 50 bp sequences in the genome that match the sequence starting at this base while allowing for 2 mismatches [199]. I divided the genome into 50 bp bins and classified each bin as mappable if all bases in the bin had a mappability of 1 (i.e. unique sequence). For the analyses below, I focused only on mappable bins ($\approx 92\%$ of the genome).

I identified a set of loci that are unlikely to be transcribed by the following steps, perfomed using `bedtools` version 2.15.0 [189]:

1. Start with a list of all annotated regions that could be transcribed (genes, non-coding RNAs, rRNA, snoRNA, snRNA, tRNA, transposable elements, and pseudogenes). Add annotations for $5'$ and $3'$ UTRs where possible [47].

2. Add 200 bp on either side of the features above.

3. Merge this set of features with the set of loci encompassing all bases with `phastcons` conservation scores >0.5 based on an alignment of 7 yeast species [200].

4. Take the complement of the above set of features to arrive at a set of loci that are unlikely to be transcribed. I subdivided each feature into nonoverlapping 50 bp bins,

which resulted in 11,430 mappable bins spread across all chromosomes. This corresponded to roughly 570 kb or $\approx 5\%$ of the yeast genome.

I combined reads from all our RNA-Seq samples, requiring each read to map with high quality and not be marked as a potential PCR duplicate, and worked from this set of reads. I obtained a "background" distribution for non-transcribed regions by calculating the read depth in each 50 bp bin I identified as unlikely to be transcribed. Although some of these bins may contain genomic sequences transcribed to produce a functional product, I assumed that for most bins the large majority of reads likely represent transcriptional noise. For the remaining genomic bins, I calculated a p-value by using the background distribution as an empirical null distribution. The presence of truly functionally transcribed bins in this background set makes our approach conservative.

Introns

I used `TopHat` version 2.0.3 [201] with options `--color`, `--quals`, `--bowtie1`, `--library-type=fr-secondstrand`, `--min-intron-length` 15, `--max-intron-length` 2000, `--min-coverage-intron` 15, `--max-coverage-intron` 2000, `--min-segment-intron` 15, `--max-segment-intron` 2000, `--transcriptome-max-hits` 10 to map all of our RNA-Seq reads to polymorphized reference genomes for each strain, providing a list of known splice junctions present in *Saccharomyces* Genome Database [106] using the `--raw-juncs` option. I combined predicted splice junctions for all strains and filtered out junctions corresponding to previously annotated introns. I excluded junctions supported by less than 5 reads or junctions resulting in putative introns less than 20 bp in length, leaving a total of 3,371 candidate novel splice junctions.

I used a list of 292 well-supported introns [62] to compile position-specific scoring matrices (PSSMs) for the 5′, 3′, and branch point sequences that have clear functional roles in pre-mRNA splicing [78]. The matrices I used to produce PSSMs for the 5′ and 3′ splice sequences were derived from the first six and last three bases, respectively, of each of the 292 introns. For the branch point sequence I used the sequence present in the yeast intron database [74]. A strong match identified using position-specific scoring matrices derived

from known functional sequences would suggest that the 5′, 3′, and branch point sequences in a candidate intron are likely to be functional.

First, I used `FIMO` version 4.8.1 [202] to obtain scores for each of the 292 true introns, using a first-order Markov model for background frequencies determined from the complete yeast reference genome. I restricted a match to my 5′ splice site PSSM at the beginning of each intron and a match to my 3′ splice site PSSM at the end of each intron, and allowed the branch point PSSM to match anywhere between the seventh and third from last bases of the intron. I summed log-odds scores output by `FIMO` for the three PSSM matches, taking only the best sequence match for each PSSM.

Second, I followed the same process described above to score each of the predicted 3,371 candidate novel introns. I found 50 candidate novel introns with total PSSM match scores higher than the lowest-scoring 25% of true introns. These candidate novel introns were supported by at least five `TopHat` read alignments (median 21, mean 103). I ignored introns that fell in repetitive regions of the genome (e.g. Y′ elements), leaving a total of 45 candidate introns. Of these 45 introns, 42 overlapped previously annotated introns. The three introns that did not overlap previously annotated introns included:

1. one intron supported by 249 reads and starting in the gene YMR147W that was previously thought to be spliced only under heat shock conditions (chrXIII:559782-560157) [73].

2. one intron, supported by 37 reads, near the 3′ end of the gene *HOP2* that would truncate the Hop2 protein by 7 amino acids and change 5 amino acids at the C-terminal end of the protein.

3. one intron, supported by 21 reads, in the 5′ UTR of the gene *BCK2* and that overlaps a spliced EST (with different splice junctions) detected by Grate et al. (unpublished; Genbank ID: EG999335.1) using RT-PCR.

The 42 remaining putative novel introns overlapped annotated introns. I observed polymorphisms in splice site sequences in five cases, but the polymorphisms did not severely

disrupt splice site consensus motifs and did not appear to correlate with frequency of either the previously annotated or novel splice forms.

RNA transcripts

I used `Tophat` version 2.0.3 [201] with the same options as above (Section 4.3.4) to map all of our RNA-Seq reads to strain-specific reference genomes for each strain. I ran `Cufflinks` version 2.0.0 [203] with the parameters `--min-intron-length` 15, `--max-intron-length` 2000, `--multi-read-correct`, `--3-overhang-tolerance` 300, `--intron-overhang-tolerance` 15, `--max-bundle-length` 15000 , and `--frag-bias-correct` to assemble transcripts using mapped reads. I focused on predicted intergenic transcripts or antisense transcripts >50 bp in size and at least 1 kb from any annotated gene or non-coding RNA on the same strand. I required 95% of the 50 bp bins tiling across the locus to be "uniquely mappable", as defined in Section 4.3.4, before considering a predicted transcript further. This resulted in 621 predicted transcripts: 383 intergenic transcripts and 238 antisense transcripts that overlapped a protein-coding gene or noncoding RNA on the sense strand.

Some antisense transcripts overlapped more than one gene on the opposite strand. When comparing correlations between the predicted antisense transcript and its corresponding sense gene, I used only the gene most overlapped by the predicted transcript. To analyze strain differences in expression of antisense and intergenic transcripts, I employed the same methods used to detect differential expression of annotated genes, as described in Section 4.3.2.

Proteins

See Section D.2.4 for a summary of the methods Gennifer Merrihew used to search for novel/unconfirmed proteins and protein modifications.

4.3.5 Association Mapping

Mapping cis*-regulatory gene expression QTL*

In order to map *cis*-regulatory gene expression QTL, for each gene I focused on SNPs located within the region 500 bp upstream of the annotated gene start to 500 bp downstream of the gene end. I focused on variants near the gene of interest, which presumably act primarily in *cis* to influence RNA and protein levels, because complex patterns of population structure in *S. cerevisiae* render genome-wide association studies susceptible to a high type I error rate [204]. I only considered variants where the minor allele was present in at least four strains. I used the program Haploview [205] to select tag SNPs with $r^2 \geq 0.6$. Next, I employed the program `EMMA` [206], which uses a mixed model approach, to control for population structure. Although the presence of population structure leads to an elevated type I error rate for association testing using these yeast strains, the corrections performed by `EMMA` partially mitigate these concerns and result in a type I error rate that is only slightly elevated above that expected in the absence of population structure [204].

To conduct association tests, I recorded the test statistic output by `EMMA` for each SNP-gene combination, and kept the maximum statistic for each gene. To determine the significance of this statistic, I performed 1,000 permutations where I replicated this strategy on data where the gene expression levels were shuffled randomly. After calculating p-values using the null distribution estimated from permuted data, I used standard methods [169] to calculate the false discovery rate, and for further analysis I focused on *cis*-regulatory gene expression QTL identified as significant at FDR=5%.

Mapping cis*-regulatory protein expression QTL*

In order to map *cis*-regulatory protein expression QTL, I performed exactly the same procedures as described above (Section 4.3.5), but used peptide levels in the association tests. I did permutations and calculated false discovery rates on a peptide-level basis.

To explore the concordance between gene expression QTL and protein expression QTL, I focused on the significant gene expression QTL and plotted the p-values for SNP-peptide associations for the same genes. The absence of a significant peptide association in 63/68

transcript associations might reflect a lack of power rather than a truly missing association. I analyzed the distribution of p-values for this subset of peptide associations using a conservative method for estimating the fraction of truly significant tests [169] and found that 53% of the peptides were estimated to have a true association, indicating that most large-effect variants affecting RNA levels affect peptide levels as well.

4.3.6 Predicting phenotypic variation

Unless otherwise noted, I constructed models using random forest regression, which allows for complex and nonlinear interactions among a heterogeneous set of predictor variables and has been shown to be effective across a wide range of modeling problems [207, 208]. I used the `R` package `randomForest` [209] for random forest analyses reported in this section.

Linear model for intrastrain phenotype prediction

I used a simple linear model in order to estimate the amount of variation explained between genes within a single individual at the RNA, protein, metabolite, or morphological level. The linear models I constructed were used to model, for each strain: 6,207 RNA levels, 1,643 protein levels, 115 metabolite levels, and 392 morphological traits. I obtained protein levels by averaging the peptide measurements for all peptides that mapped to each gene. In my models, the response variable was a vector of phenotype levels across strain i, and the covariates were vectors of (1) the phenotype level in a strain closely related to strain i and (2) the mean phenotypic level across all strains, excluding strain i. Closely related strains were assessed using a phylogeny constructed at the genome-wide level from all strains, and a single strain was selected randomly when multiple strains were equally closely related. I fit models in `R` [187], and quantified variance explained using adjusted R-squared.

Exploiting the phenotypic correlation structure for interstrain phenotype prediction

As described in the results section, I exploited the phenotypic correlation structure to make predictions of interstrain phenotypic variation. This approach was motivated by the observation that, given two phenotypes known to be highly correlated, each could serve as

a useful predictor of variation in the other. Beginning with the 8,365 phenotypes used to make Figure 4.6A (i.e. peptide measurements collapsed into a single mean number for each protein), I sequentially withheld each strain, recalculated pairwise correlations between all phenotypes, and recorded the phenotypes that were highly correlated (FDR = 5%) with each other phenotype.

I used a simple random forest model to make predictions for all 5,494 phenotypes that varied significantly between strains. To make a prediction for phenotype i in strain j, I first retrieved the list of all phenotypes highly correlated to phenotype i when strain j was withheld. I trained a random forest model using these phenotypes to predict phenotype i in the remaining strains, and used this model to make a prediction for strain j. I calculated the percent of variation explained by my models using the formula $1 - \mathrm{MSE}_{\mathrm{model}}/\mathrm{MSE}_{\mathrm{null}}$, where MSE indicates the mean squared error between true values and predicted values. Null predictions used to calculate $\mathrm{MSE}_{\mathrm{null}}$ consisted of the mean value of the phenotype across all strains other than the one whose phenotypic value was being predicted. I explored additional methods for prediction, including multiple linear regression and principal components analysis based methods, but found that differences in performance between different statistical methods for prediction were minimal.

I identified tag traits using a greedy algorithm where I first identified the phenotype correlated with the largest number of other phenotypes, then removed these phenotypes (as they are "tagged" by the phenotype selected) and repeated the process until acquiring the desired number of tag traits. I used tag traits to make predictions in an identical manner as above, but only training random forest models using highly correlated tag traits rather than all highly correlated phenotypes.

Addition of heterogeneous predictors for interstrain phenotype prediction

I used random forest models to make use of a range of heterogeneous sources of data to predict RNA and protein levels for genes that varied significantly between strains. I focused only on genes for which I obtained both RNA and protein data, which resulted in 1,303 RNA levels and 660 protein levels (these numbers differ because I focused only on RNA/protein

levels that differed significantly between strains). I obtained protein levels by averaging the peptide measurements for all peptides that mapped to each gene. My full model consisted of a large number of predictors with potential relevance for determining RNA and protein levels. Training a single-gene model with only 22 strains is difficult, so I constructed separate models for RNA and protein phenotypes and performed joint prediction of all phenotype $\times$ strain combinations (Figure 4.8). Thus, in the RNA prediction model the response vector consisted of $22 \times 1,303$ RNA levels, and in the protein model the response vector consisted of 22×660 protein levels.

Below, I refer to "globally" and "locally" closely related strains — both designations describe strains that are closely related, with globally related strains nearest each other in a matrix of genetic distances built from complete genome sequences, and locally related strains nearest each other in a matrix of genetic distances built from sequences at a particular locus. Thus, globally closely related strains are the same at all loci, while locally closely related strains can vary depending on locus. In cases where multiple strains were equally distant phylogenetically, I took the mean RNA/protein value across all equally closely related strains.

To avoid confusion, below I outline predictors used in the model of protein levels only (Figure 4.8). The model for RNA levels was identical, but with RNA in place of protein and vice versa. The matrix of predictor variables for the protein model included the following (for gene i, strain j):

1. Predictions obtained using other highly correlated traits (previous section)

2. Genic characteristics: gene essentiality [120]; fitness of gene deletant [210]; whether gene i encodes a ribosomal protein

3. DNA sequence characteristics: GC content of the gene; codon bias [121]; whether gene i contains a known intron; whether gene i's promoter is predicted to contain a TATA box [211] or a GA element [212]; divergence (rate of evolution) from *S. paradoxus*, in the promoter, coding sequence, and a 10 kb bin centered around gene

i; presence/absence of predicted instances of 153 transcription factor binding motifs upstream of gene i

4. RNA levels: RNA level of gene i in strain j, RNA level of gene i in globally closely related strain; RNA level of gene i in locally closely related strain

5. Protein levels; protein level of gene i in globally closely related strain; protein level of gene i in locally closely related strain

6. Strain j

7. Clustering, pathways, and functional annotation: protein levels for top 3 proteins predicted to interact with gene i's product (M. Riffle et al., unpublished data); protein and RNA levels for the first 10 genes grouped with gene i according to co-expressed gene clusters identified by Yvert et al. [213], biochemical pathways curated in the `yeastCyc` database [214], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [215], and Gene Ontology SLIM categories available from *Saccharomyces* Genome Database [106]; and genic membership in these clusters, pathways, and Gene Ontology categories

I filled in missing values by taking the mean across all non-missing observations. I centered RNA and protein values for each gene to have mean zero and variance one. I calculated the percent of variation explained by our models on a gene-by-gene basis using the formula $1 - \mathrm{MSE}_{\mathrm{model}}/\mathrm{MSE}_{\mathrm{null}}$, where MSE indicates the mean squared error between true values and predicted values. Null predictions used to calculate $\mathrm{MSE}_{\mathrm{null}}$ consisted of the mean value of the phenotype across all strains other than the one whose phenotypic value was being predicted. I obtained model predictions using out-of-bag data from the random forest model. Out-of-bag predictions were similar to those obtained by withholding data from a gene and performing predictions after model fitting.

Models for RNA and protein in specific pathways

I constructed targeted models to predict RNA and protein levels for specific pathways. Here, I considered each pathway separately. I used annotations from `yeastCyc` [214] and *Saccharomyces* Genome Database [106] to group genes into pathways. I built simple random forest models using only RNA levels for genes in the pathway to predict protein levels for genes in the pathway, and vice versa. I explored whether the inclusion of additional information (genic characteristics and DNA sequence characteristics listed in the previous section) might improve predictive accuracy, and found that the improvement gained by including these factors was generally minimal.

Models for predicting metabolite levels

I constructed models to predict individual metabolite levels using either a random forest or linear regression approach. I chose genes in pathways that produced or consumed the metabolite, or where the metabolite is an intermediate, and used RNA and protein levels for these genes as predictors in my models. I obtained information on the pathways in which each metabolite is involved using `yeastCyc` [214] and *Saccharomyces* Genome Database [106].

4.4 Results and discussion

4.4.1 High-dimensional phenotyping and genome sequencing

To comprehensively measure molecular and morphological phenotypes, while mitigating confounding variables, my colleagues and I used a randomized study design and obtained biological replicates for each measured trait (Figure 4.2). Specifically, we grew strains to steady state in chemostats under phosphate-limited conditions (Section D.2.2), which allowed us to control for growth rate variation among strains and maintain a constant external cellular milieu. We selected phosphate-limited media as preliminary work showed that this condition resulted in widespread transcriptional differences among *S. cerevisiae* strains. For each sample, we performed RNA-Seq to characterize gene expression and transcript structure; chromatography and mass spectrometry to measure protein and metabolite

abundances; and quantitative microscopy to measure morphological phenotypes (Figure 4.2; Section D.2).

In addition, we resequenced the genome of each strain to high coverage (mean $\approx$ 30X; Section D.2.1). This data supplemented existing low-coverage Sanger sequence data [1] and allowed me to call additional SNPs, map RNA-Seq reads in an unbiased fashion, and identify peptides expected to differ in amino acid sequence between strains. Concordance with previously published Sanger-based sequences [1] was over 99.5% in all strains. Previous low-coverage sequencing [1] reported approximately 230,000 SNPs (20,000-80,000 SNPs per strain) in these strains relative to the *S. cerevisiae* reference sequence (S288c); our high-coverage short-read sequencing yielded an additional 50,000 SNPs (1,500-44,000 per strain). Liti et al. [1] used a phylogenetically-motivated imputation procedure to fill in missing sequence in their low-coverage genomes. For bases where there was sufficient coverage to call genotypes, my genotype call was discordant with 2.5% to 37.5% of imputed sites across strains. Using Sanger sequencing, Marnie Johansson obtained 5.9 kb in two strains overlapping 86 imputed SNPs, where my short-read genotyping calls disagreed with the imputed genotype in 41 cases. I found 100% agreement between the Sanger sequence and our calls (Section D.2.1), highlighting the difficulty of accurately imputing sequence in a model organism with complex and heterogeneous patterns of population structure.

4.4.2 Expanding and refining the parts list of a model organism

The deep molecular phenotyping and genome sequencing we performed provides a unique opportunity to further expand the biological "parts list" of *S. cerevisiae*. Although this organism is perhaps the best-studied eukaryote, I found evidence for novel or previously unconfirmed transcripts, introns, and proteins. Across all samples combined, we obtained approximately 13.2 Gb of uniquely mappable genome sequence (equivalent to $>$ 1000X coverage of the genome) and 38.6 Gb of uniquely mappable transcript sequence. To estimate the fraction of the genome transcribed into RNA, I first constructed a background distribution of expression from genomic regions that were not conserved among other yeasts, a minimum of 200 bp from annotated genomic features, and non-repetitive. I tested for transcription

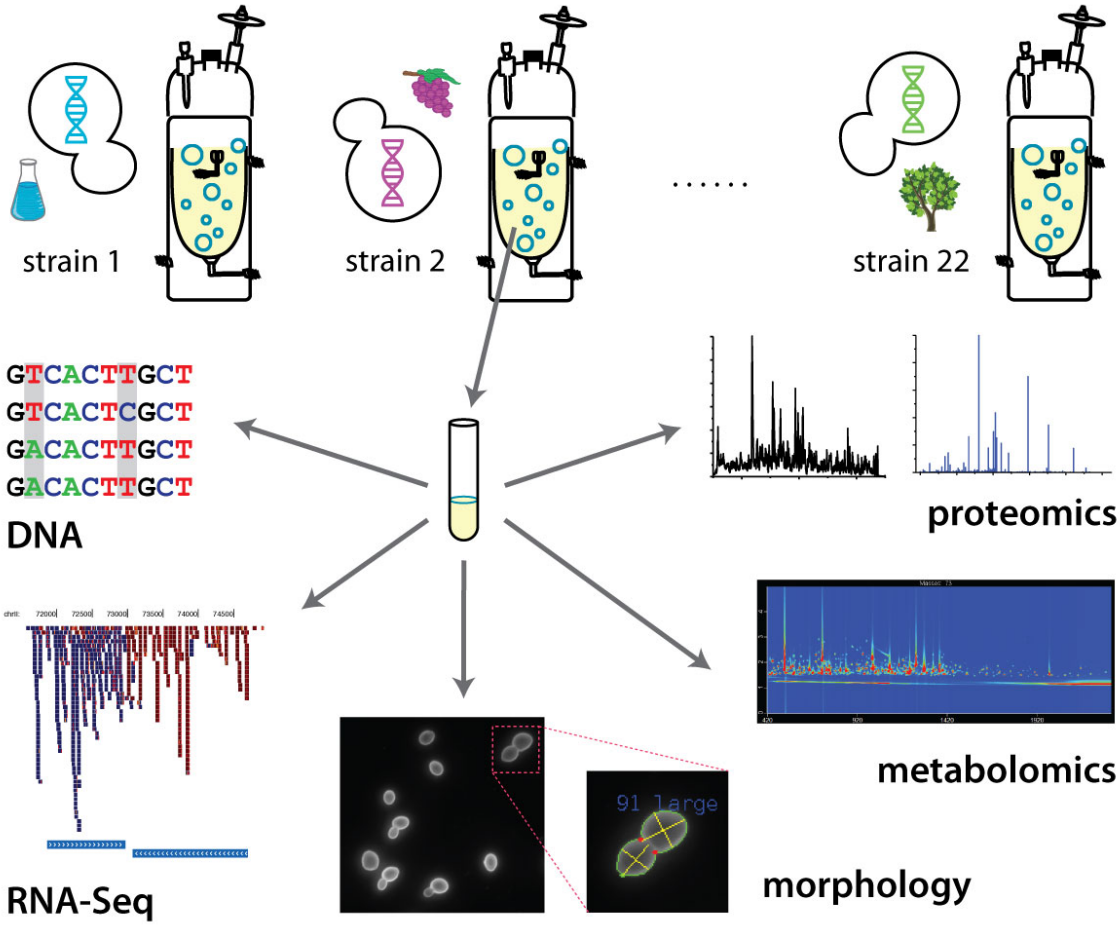


Figure 4.2: Experimental design facilitates high-dimensional phenotyping. A schematic of the experimental design used to obtain phenotypic data for each of 22 strains. Icons next to strains represent the variety of sources from which strains were originally isolated (Table D.1). Schematic outlines the process of obtaining phenotype data for a single strain. In total (aggregated across all strains), we obtained 16 Gb of DNA sequence; 820 million RNA-Seq reads; 912,000 mass spectra that we used to infer peptide levels; metabolic measurements of molecules in 40 different biochemical pathways; and 2,000 images that together captured 21,000 cells whose morphological characteristics we tabulated.

above this background level in 50-bp bins along the genome, and found that $\approx 80\%$ of the uniquely mappable fraction of the genome is transcribed under phosphate-limited growth (Section 4.3.4).

To identify novel antisense and intergenic transcripts, I focused on transcripts predicted by Cufflinks [203] that were at least 50 bases in size and a minimum of 1 kb from any other annotated transcribed feature on the same strand. I found 621 transcripts that met these criteria: 383 intergenic transcripts and 238 antisense transcripts that overlapped a protein-coding gene or noncoding RNA on the sense strand (Section 4.3.4). Overall, intergenic and antisense transcripts were expressed at low levels (median 3.0 RPKM for intergenic transcripts and 4.5 RPKM for antisense transcripts, as opposed to 40.5 RPKM across all annotated genes). I observed significant differences in expression (FDR = 5%) between strains for 155 (40%) intergenic transcripts and 176 (74%) antisense transcripts. Antisense transcripts can regulate gene expression via mechanisms that repress sense transcription, such as transcriptional interference [216, 217] or epigenetic modifications [218]. I tested whether strain differences in expression of antisense transcripts were correlated with expression of the closest sense strand gene, and found significant anticorrelations for 52 transcript/gene pairs (FDR=5%; one example is shown in Figure 4.3A).

In addition, I used our RNA-Seq data to identify strain differences in untranslated regions (UTRs). Focusing on 4,882 verified genes with previously annotated UTR boundaries [47], I found 1,188 that were expressed in all strains (>1 RPKM) and had >10-fold differences in UTR read coverage between strains. 72% of these candidate differences in UTR length occurred in $5'$ UTRs; an example in the pyridoxine transporter *TPN1* is shown in Figure 4.3B. There was a marginally significant enrichment for genes associated with cellular bud ($p = 0.036$, Bonferroni corrected) among genes with UTR differences.

I also identified 45 candidate novel splice junctions. One confirmed an unannotated but previously reported intron thought to be spliced only under heat shock conditions [73]. Two additional predictions result in introns at the $3'$ end of the gene *HOP2* (causing minor alterations to the Hop2 protein) and in the $5'$ UTR of the gene *BCK2*. The remaining 42 novel splice junctions shared a splice donor (41/42) or acceptor (1/42) site with an existing annotated intron. The median ratio of reads mapping to previously annotated

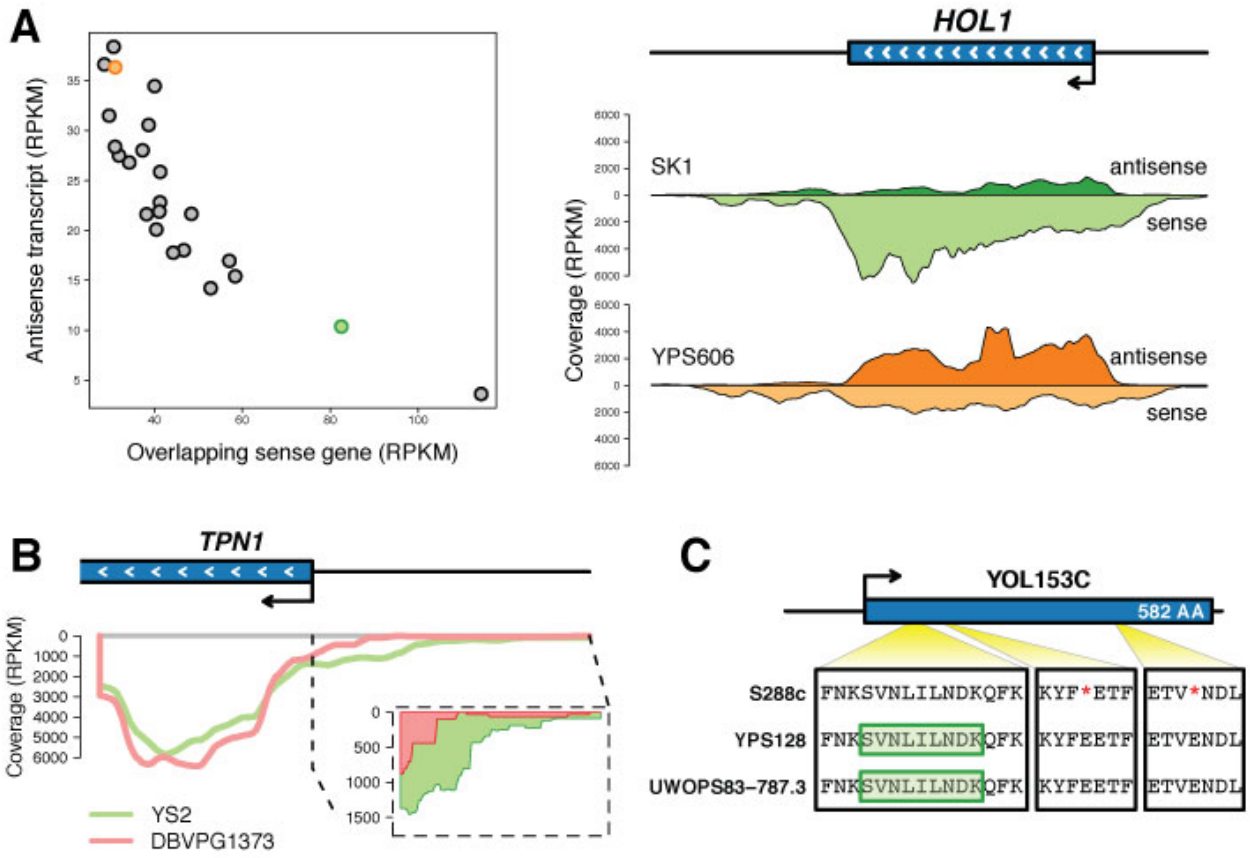


Figure 4.3: Exploring the parts list of *S. cerevisiae*. (A) Example of negative correlation ($\rho = -0.835$) between sense transcription of the gene *HOL1* and antisense transcription at the locus. Left panel shows relationship between sense and antisense transcription; each point represents data from a single strain. Orange and green points are shown in more detail in right panel. Right panel depicts the gene model and smoothed RNA-Seq read coverage from two strains highlighted in left panel; sense coverage is shown below the zero-line in lighter color, and antisense coverage is shown above the zero-line in darker color. (B) Difference in UTR usage in the gene *TPN1*. Smoothed RNA-Seq read coverage from two strains is depicted below gene model. Inset shows detail in the 5′ UTR. (C) Evidence for translation of the gene YOL153C. YOL153C is categorized as a hypothetical protein in *Saccharomyces* Genome Database, and is likely a pseudogene in S288c as evidenced by two in-frame stop codons (red asterisks). We detected peptide evidence (green boxes) of this protein in strains YPS128 and UWOPS83-787.3, which do not contain the two stop codons.

versus novel predicted splice junctions at the same locus was 118:1, suggesting that many of these predicted novel junctions may be low-frequency alternative splice forms or splicing errors. In many cases, the novel splice forms would cause profound changes in the translated protein (changing at least ten amino acids in 15 cases and truncating at least one-fifth of the protein in 23 cases), suggesting a possible role for these novel splice variants in gene regulation.

To identify novel peptides using our quantitative shotgun mass spectrometry data, I first searched for matches to annotated *S. cerevisiae* genes that either lack firm experimental evidence ("uncharacterized" genes) or are thought unlikely to encode an expressed protein ("dubious" genes). I found uniquely matching peptides for 3 dubious and 63 uncharacterized genes. Next, I searched for evidence of translation of pseudogenes or hypothetical proteins identified in previous genome-sequencing projects (Section D.2.4). In two strains, UWOPS83-787.3 and YPS128, I found evidence for translation of YOL153C (Figure 4.3C), which is is a presumed pseudogene with two in-frame stop codons in the reference S288c genome. In our strain-specific genome sequences, these two strains (along with 17 others) do not possess the two stop codons present in S288c. In addition, I detected peptides matching predicted/hypothetical genes from other genome sequencing projects, supporting predictions for two genes in strain AWRI796 [219], five genes in strain EC1118 [220], three genes in strain Jay291 [221], and six genes predicted by Liti et al. [1]. Finally, I used our proteomics data to search for several post-translational modifications including oxidized methionine and phosphorylated serine and threonine (Section D.2.4). I found modifications (often present in only a subset of strains) in 122 peptides comprising 84 unique proteins (FDR = 5%). Overall, these results add to the extensively curated list of biomolecules and catalog of variation present in *S. cerevisiae* and suggest that deep molecular phenotyping will be useful for refining the parts lists of more complex organisms.

4.4.3 Pervasive phenotypic diversity

I found widespread heritable variation within every class of phenotypic data measured, with over 50% of all measured traits varying between strains. Specifically, 74% (4,565) of tran-

script levels, 23% (1553) of peptides, 10% (12) of metabolites, and 64% (255) morphological traits significantly varied (FDR = 5%) across strains. For the morphological phenotypes, most traits were correlated with a small number of broad characteristics of cellular morphology, including 28% (110) of traits that were significantly correlated with whole cell size and 31% (121) that were significantly correlated with cell shape (order of the elliptical approximation of the cell). Following Nogami et al. [196], the traits that I tabulated included both directly measured traits and their coefficient of variation (CV). I found more directly measured traits differed between strains (151/199) than CV traits (104/199).

For both transcript and protein levels, genes that varied most across strains were involved in aerobic respiration and the electron transport chain, with highly significant gene ontology enrichment ($p < 10^{-5}$) for cellular respiration, ATP synthesis coupled proton transport, and mitochondrial respiratory chain complexes. Indeed, I found consistent differences between strains in the overall activity of central carbon metabolic pathways, reflecting contrasting strategies for energy generation (Figure 4.4A). The strong anticorrelation between the activity of genes involved in fermentation versus aerobic respiration was largely (though not entirely) associated with the major phylogenetic division between the strains (Figure 4.4B); strains involved in the production of alcoholic beverages as well as their close relatives tended to be more active fermenters.

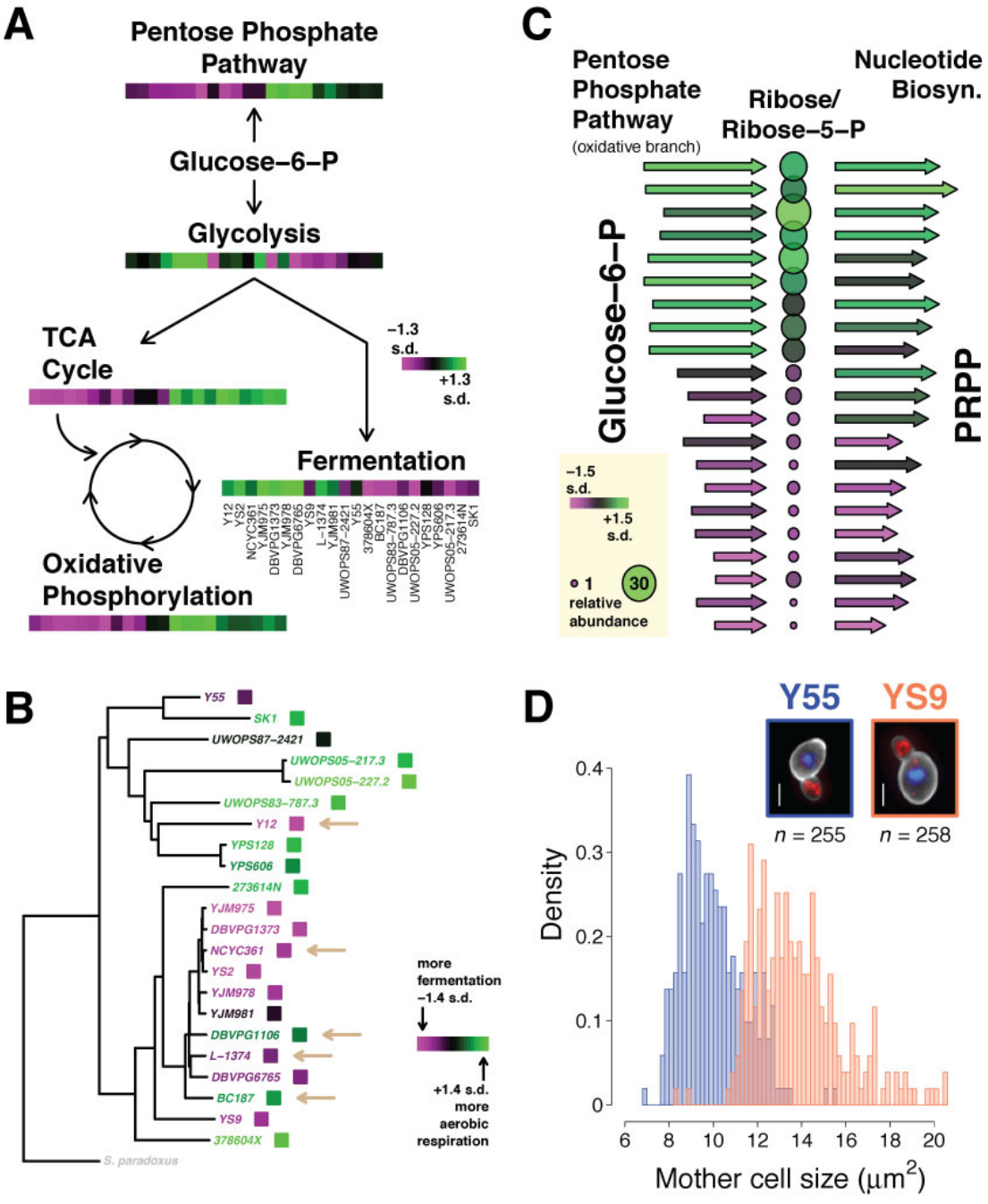
I examined each differentially abundant metabolite in the context of 162 well-annotated biochemical pathways. Overall, metabolites tended to correlate significantly (FDR = 5%) with a large number of pathways (mean = 58, standard deviation = 30), consistent with the highly interconnected nature of metabolism. The metabolite ribose (due to the derivatization process, this measurement included both free ribose and ribose-5-phosphate) was significantly correlated (FDR = 5%; $|\rho| > 0.43$) with the largest number of pathways, 96. Ribose-5-phosphate is produced by the pentose phosphate pathway and required for nucleotide biosynthesis, and I observed varying levels of activity of these pathways and corresponding ribose/ribose-5-phosphate levels across strains (Figure 4.4C). At the morphological level, I observed consistent, heritable differences in cell size (Figure 4.4D).

4.4.4 Identifying large-effect cis*-regulatory transcript and protein QTL*

To search for genomic variants underlying variation in functional genomics phenotypes, I performed association tests between variants within 500 bp of each gene and its corresponding RNA and peptide levels. Although the number of individuals we sampled is small ($N = 22$), simulations indicate that I have moderate to high power to detect large-effect variants (Section D.3.1). I focused on common variants near the gene of interest, which presumably act primarily in *cis* to influence RNA and protein levels, because complex patterns of population structure in *S. cerevisiae* render genome-wide association studies susceptible to a high type I error rate [204]. Before conducting association tests, I controlled for population structure using mixed models and selected tag SNPs with $r^2 > 0.6$. I found 64 significant peptide-SNP associations (from 42 distinct proteins) and 302 significant RNA-SNP associations (Figure 4.5A; FDR = 5%). The genetic variants underlying these associations have large effects, explaining on average nearly 53% of the variation in peptide or RNA level.

Of the 69 significant transcript associations that also had peptide data, six had at least one significant peptide association. Of these six, five were associated with the same SNP, consistent with variants that affect transcript level and thus indirectly affect protein level. Figure 4.5B shows one such example in the gene *PPN1*, an endopolyphosphatase involved in phosphate metabolism. The absence of a significant peptide association in 63/69 transcript associations suggests that the heritable basis of regulatory variants influencing RNA and protein levels is largely distinct [33, 222]. However, it also might reflect a lack of statistical power. To investigate these two hypotheses, I estimated the fraction of truly significant peptide associations among the set of genes with significant RNA associations using a conservative method based on the distribution of p-values (Figure 4.5C) [169]. I estimate that 53% of the peptides have a true association, indicating that a substantial fraction of large-effect variants that influence RNA levels also affect peptide levels.

Figure 4.4: Pervasive heritable phenotypic variation. (A) Overview of central carbon metabolism. Heatmaps indicate pathway activity in each strain: RNA and protein data for all genes in each pathway was combined, the first principal component extracted, and the numerical sign adjusted to ensure higher numbers corresponded to higher average RNA and protein abundance across the pathway and vice versa. Strains are shown in the same order (listed under the fermentation heatmap) for all heatmaps. (B) Phylogeny based on complete genome sequences, with strain names colored according to the key shown. Tan arrows indicate strains used in the fermentation of alcoholic beverages. Pathway activity for each strain was calculated as in (A) using genes in the TCA cycle and genes involved in fermentation. (C) Pathways leading to the production of phosphoribosyl pyrophosphate (PRPP) from glucose-6-phosphate, with pathway activity displayed vertically for 21 strains. Arrows represent pathway activity, with longer arrows/brighter green indicating higher activity and shorter arrows/brighter magenta indicating lower activity; a single measure of pathway activity was calculated as in (A). Arrows on left indicate pathway activity of the oxidative branch of the pentose phosphate pathway. Circles indicate measured levels of ribose/ribose-5-phosphate and are colored and sized accordingly. Arrows on right indicate activity of 5-phospho-ribosyl-1(alpha)-pyrophosphate synthetase, the heteromultimeric complex that synthesizes PRPP. (D) Differences in mother cell size between a small and large strain. Histograms are composed of measurements made on individual cells. Inset photos show a typical cell from each strain (with size near the strain mean). White scale bars show ≈ 2 μm. Actin stain is shown in red, DNA stain in blue, and cell wall stain in greyscale in the merged images.


A
Pentose Phosphate Pathway
Glucose–6–P
Glycolysis
TCA Cycle
Oxidative Phosphorylation
Fermentation
–1.3 s.d.
+1.3 s.d.
Y12
YS2
NCYC361
YJM975
DBVPG1373
YJM978
DBVPG6765
YS9
L–1374
YJM981
UWOPS87–2421
Y55
378604X
BC187
UWOPS83–787.3
DBVPG1106
UWOPS05–227.2
YPS128
YPS606
UWOPS05–217.3
273614N
SK1
B
Y55
SK1
UWOPS87–2421
UWOPS05–217.3
UWOPS05–227.2
UWOPS83–787.3
Y12
YPS128
YPS606
273614N
YJM975
DBVPG1373
NCYC361
YS2
YJM978
YJM981
DBVPG1106
L–1374
DBVPG6765
BC187
YS9
378604X
S. paradoxus
more fermentation –1.4 s.d.
+1.4 s.d. more aerobic respiration
C
Pentose Phosphate Pathway (oxidative branch)
Ribose/ Ribose–5–P
Nucleotide Biosyn.
Glucose–6–P
PRPP
–1.5 s.d.
+1.5 s.d.
1 relative abundance
30
D
Y55
YS9
n = 255
n = 258
Density
0.4
0.3
0.2
0.1
0
6
8
10
12
14
16
18
20
Mother cell size (µm2)

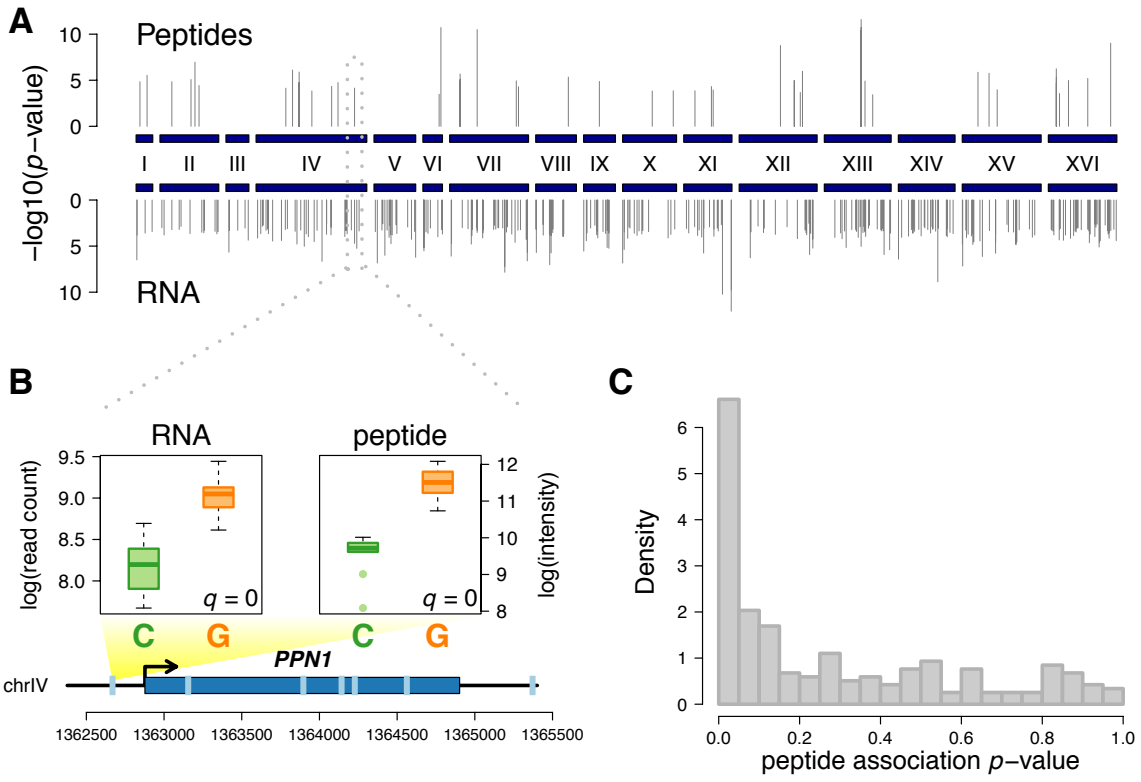


Figure 4.5: *Cis*-regulatory RNA and peptide QTL. (A) Manhattan plot showing results for significant RNA and peptide *cis*-association tests. Grey vertical lines indicate individual tests. Blue boxes and associated Roman numerals across the middle of the panel indicate the 16 chromosomes of yeast. (B) Example, in the gene *PPN1*, where RNA and peptide levels show association with the same polymorphism. Light blue ticks along gene model indicate locations of tag SNPs tested for association. SNP that was significantly associated with RNA and peptide levels shown is the first SNP. Yellow gradient originating at this SNP expands to boxplots of RNA and peptide levels separated by allele; boxes indicate lower quartile, median, and upper quartile, and whiskers extend to half the interquartile range. (C) Histogram of p-values for 236 peptide associations in 69 genes with significant RNA associations.

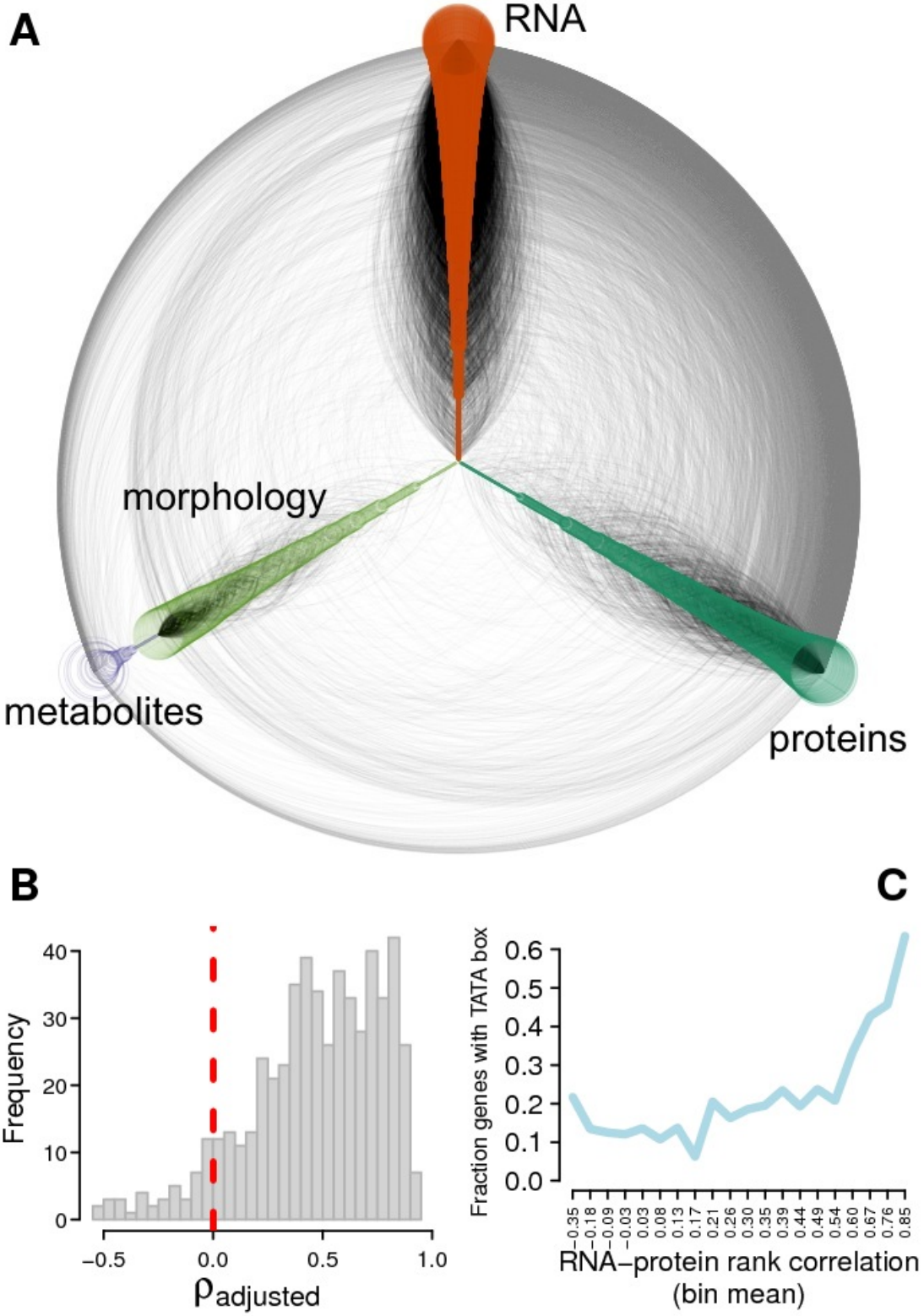
4.4.5 Densely connected network structure of phenotypic correlations

To explore the correlation structure among traits, I calculated pairwise correlation coefficients among 8,365 phenotypes (collapsing all peptide measurements into a single mean number for each protein) and identified 68,558 correlations, involving a total of 7,078 phenotypes, which were significant at a FDR of 5%. Approximately 60% (41,649) of the trait comparisons were positively correlated. The excess of positive correlations is partially attributable to the fact that levels of RNA and proteins in the same pathway or protein complex tend to be positively correlated (mean $\rho = 0.12$ across $n = 427$ pathways and protein complexes) but those from different pathways are equally likely to be negatively as positively correlated (mean $\rho = 0.008$). Overall, there were strong correlations both within (79%) and between (21%) data types, with a particularly dense set of connections within and between highly correlated RNA and protein phenotypes (Figure 4.6A).

Of the 7,078 phenotypes correlated to at least one other trait, the mean number of significant correlations to other traits was 19.4 (bootstrap 95% confidence interval 18.6–20.2). RNA levels were correlated with the largest number of other traits on average (20.4) and metabolite levels the fewest (12.0). The single most highly correlated phenotype was the RNA level of the histidine tRNA synthetase *HTS1*, which was correlated with 328 other phenotypes from all four data types but consisting largely of other RNA levels ($n = 293$). Among the 50 most highly correlated RNA and 50 most highly correlated protein levels, I observed strong enrichment for genes involved in energy generation, the mitochondrial respiratory chain, and ATP synthesis. I found 162 densely connected subnetworks ranging in size from 3 to 129 phenotypes, with 91 (56%) containing phenotypes of only one data type and 10 (6%) containing phenotypes from at least three separate data types. The highest scoring subnetwork (based on size and density) [197] consisted largely of RNA and/or protein levels for genes in the cytochrome c oxidase, cytochrome bc1, and ATP synthase complexes.

Previous studies in a diverse complement of organisms have reported widely varying levels of RNA-protein correlation [223–225]. These studies have largely measured RNA-protein correlation between different genes within an individual, whereas I sought to measure RNA-protein correlation between individuals (strains) on a gene-by-gene basis [222].

Figure 4.6: Dense network structure of phenotypic correlations. (A) Hive plot showing network composed of highly correlated phenotypic traits. Nodes arrayed along the three axes respresent individual phenotypes. Nodes are colored by data type, with orange nodes showing RNA levels, lime green nodes showing metabolites, purple nodes showing morphological traits, and sea green nodes showing protein levels. Lines drawn between nodes indicate that the two phenotypes are highly correlated. Black lines indicate connections (i.e. high correlation between phenotypes) within the same data type, and gray lines indicate connections between data types. (B) RNA-protein correlations for 542 genes called significantly differentially expressed at the RNA level and with at least one significantly differentially abundant peptide level. ρ_{adjusted} indicates a correlation calculated by subtracting the mean of correlations taken after randomizing the data 1,000 times from the true correlation. Vertical red dotted line is drawn at 0. (C) Fraction of genes containing a TATA box as a function of RNA-protein correlation. Each point plotted shows the fraction of genes with a TATA box among a bin of approximately 80 genes with similar RNA-protein correlations (whose mean is shown on the x-axis).


A
RNA
morphology
metabolites
proteins
B
Frequency
40
30
20
10
0
−0.5
0.0
0.5
1.0
ρadjusted
C
Fraction genes with TATA box
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.35
−0.18
−0.09
−0.03
0.03
0.08
0.13
0.17
0.21
0.26
0.30
0.35
0.39
0.44
0.49
0.54
0.60
0.67
0.76
0.85
RNA–protein rank correlation
(bin mean)

We measured RNA and protein levels using aliquots of cells taken from the same chemostat sample, minimizing environmental/batch effects that could lower correlations. I found a modest correlation (median 0.33; Spearman's ρ), with 44% (728 out of 1,636 genes with RNA and protein data) of genes having a significant correlation (FDR = 5%). However, restricting the analysis to genes called significantly differentially expressed (at the RNA level) and with at least one differentially abundant peptide (see below) increased the median Spearman correlation to 0.50 and resulted in $\approx$ 85% of genes having a significant RNA-protein correlation (FDR = 5%; Figure 4.6B). Genes with the highest RNA-protein correlations were strongly enriched for TATA box containing genes (Figure 4.6C). TATA-containing genes show greater variability in RNA and protein abundance between strains compared to TATA-less genes (t-test, $p < 1 \times 10^{-5}$), as has been shown for gene expression levels between different yeast species [226]. Thus, the larger variation among strains in these genes likely dominates measurement variation, resulting in stronger RNA-protein correlations. Alternatively, TATA-containing genes may be subject to less post-transcriptional regulation than TATA-less genes. Nevertheless, the relatively modest correlations between RNA and protein levels suggest a substantial role for post-translational modifications and protein degradation in the control of steady-state protein abundances [227].

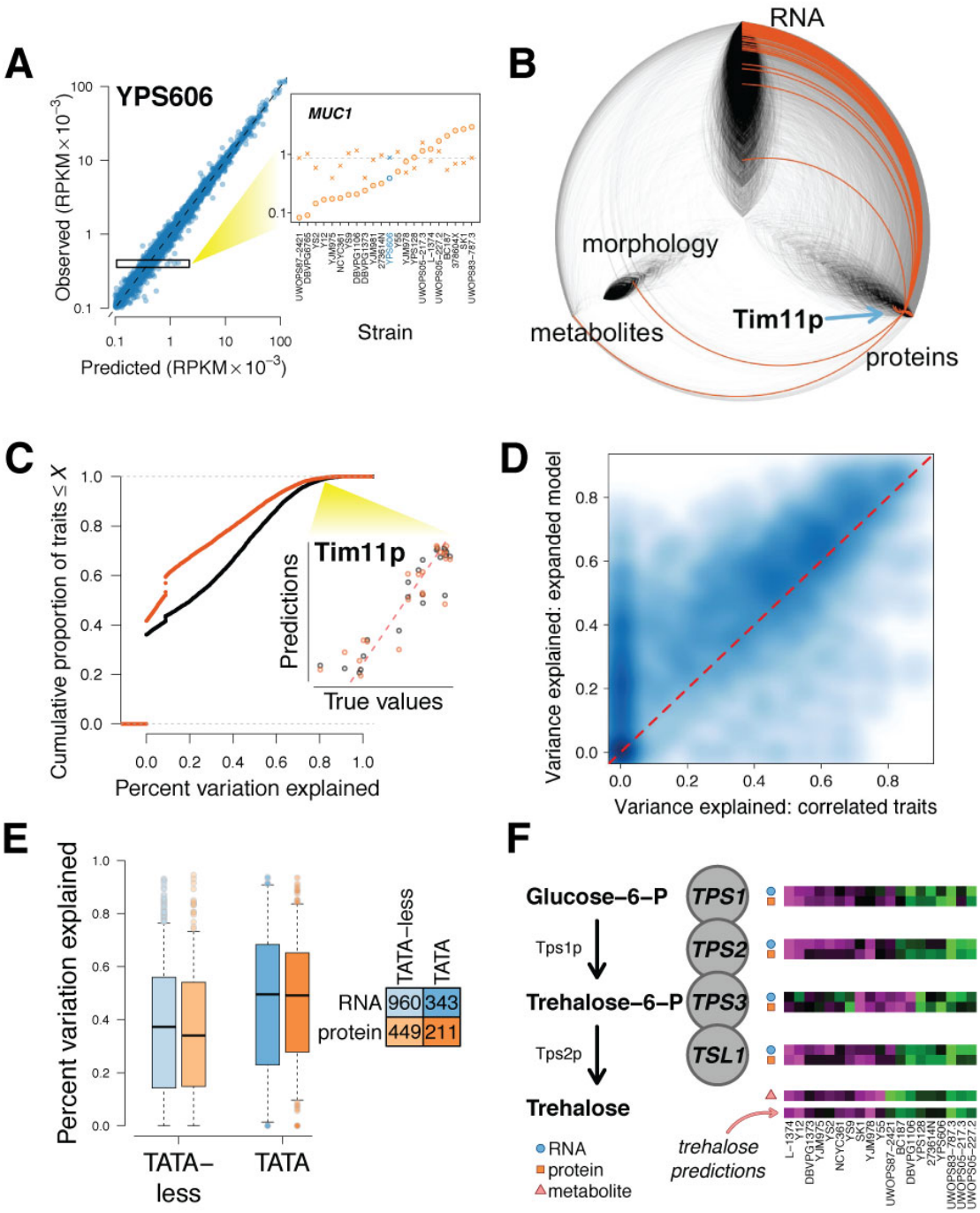
4.4.6 Integrative phenomics facilitates the prediction of phenotypes

The ability to accurately predict phenotypes would have profound consequences for basic and biomedical science [228–230], yet remains a challenging problem. I first predicted all RNA, protein, metabolite, and morphological traits in each single strain using simple models in which predicted phenotypes in the i^{th} strain were a linear function of the phenotype in its closest relative and the mean across other strains. These models accurately predict RNA, protein, and metabolite levels (median R^2_{adj} across strains = 0.97, 0.88, and 0.89, respectively) as well as morphological traits (median R^2_{adj} across strains > 0.99). However, while this analysis captures relative differences in abundance between genes within individuals, it does not robustly predict variation in abundance between individuals for a particular trait (Figure 4.7A shows this in the context of gene expression levels).

To address this problem, I leveraged the complex correlation structure of our dataset (Figure 4.6A) to predict interstrain variation for all 5,494 phenotypes that vary significantly between strains (with peptide measurements collapsed into a single mean number for each protein). To perform prediction I employed random forest regression, a statistical technique that allows for complex and nonlinear interactions among predictor variables [207]. Specifically, I sequentially withheld each strain, recalculated phenotypic correlations, and used only highly correlated (FDR = 5%) phenotypes as predictor variables in separate random forest regression models for each phenotype (for example, orange lines in Figure 4.7B show phenotypes highly correlated with the abundance of the Tim11 protein). Across all phenotypes, my predictions can account for a median of 30% of variation (Figure 4.7C, black line), and for 28% of phenotypes (1,545) my predictions explain at least 50% of variation. For example, I can account for 86% of variation in abundance of the Tim11 protein, which is a subunit of the mitochondrial F1F0 ATPase required for ATP synthesis (Figure 4.7C inset, black points). Other well-predicted phenotypes were highly enriched for mitochondrial and ribosomal functions. However, for 1,984 (36%) of phenotypes, my predictions failed to explain more than 10% of variation, indicating that information beyond values of correlated traits is necessary for robust predictions.

To explore how informative additional predictors could be, I incorporated functional annotation data available for a subset of the phenotypes I measured into my model. Specifically, I considered variation in 1,303 RNA and 660 protein levels that differed significantly between strains, using an approach similar to above with the addition of $\approx 1,000$ heterogeneous predictor variables. Additional predictors included RNA and protein levels of other genes with similar functions, genic characteristics, sequence features, and pathway annotations (Figure 4.8). I trained the model on data from all genes simultaneously in order to ensure that predictors were weighted accurately. Overall, the predictions explain $\approx 45\%$ of the variation in both RNA (median 46.8%) and protein (median 44.8%) levels, significantly better than the $\approx 36\%$ (median 37.0% and 35.6%, respectively) of variation explained using correlated traits alone (Figure 4.7D). In some cases, the difference in prediction accuracy was dramatic; for 98 (6%) phenotypes, predictions using correlated traits alone explained less than 10% of variation, but the inclusion of additional covariates increased accuracy to

Figure 4.7: Integrating data to predict phenotypes. (A) Simple models can accurately predict gene expression levels when compared between genes (left, predictions for $n = 5,385$ genes in strain YPS606 are shown), but do not fully capture variation between strains at a specific gene (inset right, gene expression levels for *MUC1* are shown for all 22 strains, with YPS606 highlighted in blue; units on y-axis are same as at left). At the *MUC1* locus, predicted values (Xs) are clustered around the mean expression across strains (gray dotted line), but observed values (circles) diverge substantially. Observed RPKM [48] values have been normalized. (B) Hive plot arranged identically to Figure 4.6A, with orange edges indicating connections to the node representing abundance of the Tim11 protein (blue arrow). (C) Empirical cumulative distribution function (CDF) displaying predictive accuracy for correlation-based phenotype predictions. Black line indicates CDF for predictions made using all phenotypes, and orange line for predictions made using only 1,000 tag traits. Inset shows predictions for abundance of the Tim11 protein made using all traits (black dots) and tag traits only (orange dots). (D) Smooth scatter plot comparing performance of prediction models discussed in the text. Darker blue indicates higher density of points, and lighter blue lower density. Dotted red line is drawn at $y = x$. (E) Boxplot indicating percent variation explained for models of RNA and protein levels. Boxes indicate lower quartile, median, and upper quartile, and whiskers extend to half the interquartile range. (F) Model for predicting levels of the metabolite trehalose, which is synthesized by the trehalose-phosphate synthase complex. Heatmaps show relative levels of RNA, protein, or metabolites. Blue circles, orange squares, and pink triangles distinguish between RNA, protein, and metabolites. Each heatmap is arranged with the strains ordered left to right in the order shown at bottom.


A
YPS606
MUC1
Observed (RPKM×10^−3)
Predicted (RPKM×10^−3)
Strain
B
RNA
morphology
metabolites
Tim11p
proteins
C
Cumulative proportion of traits ≤ X
Tim11p
Predictions
True values
Percent variation explained
D
Variance explained: expanded model
Variance explained: correlated traits
E
Percent variation explained
TATA–less
TATA
RNA 960 343
protein 449 211
F
Glucose–6–P
TPS1
Tps1p
Trehalose–6–P
TPS3
TPS2
Tps2p
TSL1
Trehalose
RNA
protein
metabolite
trehalose predictions

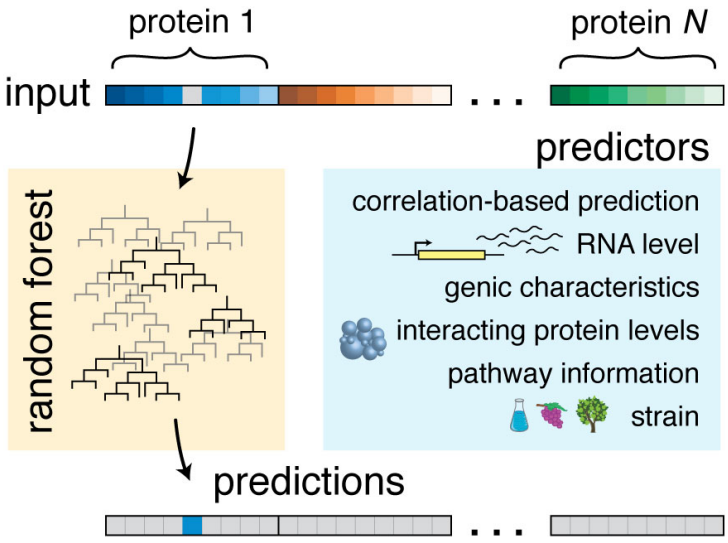


Figure 4.8: Schematic of random forest setup. Schematic outlines my approach to predicting abundance of one protein in one strain. Across top, a vector of phenotype × strain measurements is used to train the random forest model. On right, predictor variables included in the model. Along bottom, a prediction for the missing square at top (gray) is shown in blue.

greater than 40% of variation explained.

For both RNA and protein predictions in my expanded model, the most informative predictors were the correlation-based predictions (above) and abundance of the opposite data type (RNA or protein) for the gene in question, followed by strain and RNA/protein levels in other closely related strains. The presence of a TATA box was associated with better-predicted genes; I could explain over half of the variation in RNA (median 54.6%; $n = 343$) and protein levels (median 55.1%; $n = 211$) for TATA box-containing genes (Figure 4.7E). Predictions for the same TATA box-containing genes using the method above (without additional predictor variables) explained a median 41.7% of the variation, reinforcing the suggestion that these predictors can substantially improve prediction accuracy, at least for some subsets of phenotypes. Next, to explore the predictive power of DNA sequence alone, I predicted variation in RNA and protein levels using only sequence and annotation information. Specifically, I used genic characteristics, sequence features, and

pathway annotations as predictor variables, and found that I was able to explain far less variation: a median 24.6% across all RNA levels, and 21.7% across all protein levels.

Moreover, I also implemented targeted models for specific pathways and protein complexes whose steps and constituents are well understood. I was able to make highly accurate predictions in some cases, explaining at least 75% of the variation for half or more of measured RNA and protein levels in pathways including ATP synthesis and the electron transport chain, trehalose biosynthesis, glycogen catabolism, and protein levels of the RNA polymerase I complex. I also constructed models to predict metabolites that differed in abundance between strains using genes in biochemical pathways known to involve the metabolite. For some metabolites (e.g. ribose/ribose-5-phosphate, trehalose) I achieved high predictive accuracy, explaining over 50% of the variation in metabolite levels (Figure 4.7F), but other metabolite levels were poorly predicted, probably due to the influence of numerous pathways on metabolic flux.

Just as tag SNPs can be used to capture a large fraction of genetic variation with a small number of SNPs, a relatively small number of phenotypes (which I term tag traits) can capture a significant fraction of phenotypic variability. I implemented a simple greedy algorithm to identify tag traits highly correlated to many other phenotypes. Using only the top 1,000 tag traits (12% of the data), I was able to explain at least 50% of variation in 975 (18%) phenotypes that differed significantly between strains (Figure 4.7C, orange line). Abundance of the Tim11 protein was well-predicted using tag traits (Figure 4.7C, inset, orange points), and well-predicted phenotypes were enriched for largely identical functions as above. Among the 2,569 phenotypes for which the full models explained at least 1/3 of the variation among strains, tag trait models explained a median 79.1% of variation explained using all traits, indicating that tag traits can make use of a relatively small portion of the data to capture a significant fraction of variability between strains.

4.5 Conclusions

The catalog of high-dimensional quantitative traits that we measured in a genetically diverse set of yeast strains provides new insights into the structure and characteristics of phenotypic diversity and the determinants of trait predictability. My results provide a

framework for understanding the structure of molecular and morphological phenotypic diversity in budding yeast, and have implications for the successful implementation of personal genomics [183]. Moreover, this data will serve as a useful starting point for transitioning from models of bacterial cells [231] to models of eukaryotes. As technology development, advances in instrumentation, and algorithmic improvements allow for increasingly comprehensive phenomics studies, a promising future direction will be to extend this approach to multiple environments where organisms are naturally found. Finally, the data discussed in this chapter will be a useful community resource, and is available in multiple forms at `http://www.yeastrc.org/g2p/`.

Chapter 5

CONCLUDING REMARKS

5.1 Summary

In this thesis, I have presented a variety of analyses aimed at the development of methods for making inferences about the relationship between DNA sequences and the molecular phenotypes to which they give rise.

First, I presented a detailed population genetics study of introns in *S. cerevisiae*. Strictly speaking, introns are genomic elements rather than what I have termed a molecular phenotype; nevertheless, introns affect transcript structure and may play a role in the regulation of RNA and protein abundances [93]. More generally, this study presented a framework for making inferences about the evolutionary forces governing a class of genomic elements influencing genome function.

Next, I described a Bayesian model for allele-specific gene expression measured by RNA-Seq. A novel feature of this model is the ability to detect variable ASE, where the level of ASE differs across a transcript, as can occur in the case of variations in transcript structure. In the future, it will be useful to more directly investigate the specific *cis*-regulatory polymorphisms that affect gene expression levels, perhaps in concert with data on sites of protein-DNA binding (e.g. via ChIP-seq) and DNase I hypersensitive sites.

Finally, I detailed my analyses of a diverse set of *S. cerevisiae* strains that have been deeply phenotyped using RNA-Seq, proteomics, metabolomics, and quantitative microscopy. Broadly, my analyses quantified abundant variation at each level in the phenotypic hierarchy from genome sequence to morphology, and provided initial clues as to the predictability of phenotypic traits that vary between individuals within a species. This rich dataset should provide many additional insights, especially with the application of more principled models of variation affecting specific traits of interest. There are abundant opportunities to further the phenomics paradigm, whether by increasing the number of individuals sampled,

obtaining a broader phylogenetic sample, examining additional environments, or simply phenotyping more deeply.

5.2 New statistical models to utilize genome-scale data

The "big data" revolution in biology precipitated by advances in genomics and computation offers abundant opportunities for the development of new statistical models to perform inference on questions of interest. Basic models for performing pairwise tests of gene expression differences between two RNA-Seq samples have been developed [191,232], although dedicated models for more sophisticated scenarios are currently lacking. Methods for estimating genetic networks are not well developed, although large gene expression and protein abundance datasets offer a potentially powerful source of information for this task [233]. Similarly, methods for modeling causal relationships, although promising, have met limited application [234, 235]. Although there is a need for better statistical methods to interrogate high-throughput datasets, the complexity of the data generated via new instruments warrants caution, as biases are still poorly understood (e.g. biases due to transcript length, GC content, or stochastic variation in library preparation for high-throughput sequencing data) [165, 236, 237]. Nevertheless, the successful construction of models of biomolecular interactions at the whole cell level [231], even for just a subset of pathways or organelles, will require sophisticated inferences to be made from genome-scale datasets.

5.3 Placing variation in an environmental context

In natural environments, organisms are not found in sterile petri dishes filled with rich media. Undoubtedly, an organism's response to stressful conditions is an important component of fitness. Many studies have documented environmental influences on phenotypic variation, although the study of gene × environment interactions in functional genomics phenotypes is in its infancy [41–43]. Interestingly, these studies hint that *trans*-acting regulatory variation may be particularly likely to be condition-dependent [42, 43]. In Chapter 4, I studied phenotypic variation in a single environment, phosphate limitation. In order to obtain a realistic view of the functional effect of genetic variation on natural populations, we must expand our studies to additional environments, particularly those that have relevance to

the studied organism's ecological niche.

5.4 Conclusions

In 1911 Wilhelm Johanssen proposed the term "phenotype" and distinguished between genotype and phenotype [238, 239]. Over a century later, we finally have the tools to investigate the molecular basis by which genetic variation engenders phenotypic variation. Nevertheless, the molecular details underlying variation in most quantitative traits remain elusive, although progress has been made in a few examples [14, 240]. The next century will surely see remarkable progress along this front.

BIBLIOGRAPHY

[1] Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJT, van Oudenaarden A, Barton DBH, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ: **Population genomics of domestic and wild yeasts**. *Nature* 2009, 458:337–41.

[2] Consortium GP: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, 467:1061–73.

[3] Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P, Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX, Clark RM, Rätsch G, Mott R: **Multiple reference genomes and transcriptomes for Arabidopsis thaliana**. *Nature* 2011, .

[4] Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellåker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhomra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Bala S, Stalker J, Mott R, Durbin R, Jackson IJ, Czechanski A, Guerra-Assunção JA, Donahue LR, Reinholdt LG, Payseur BA, Ponting CP, Birney E, Flint J, Adams DJ: **Mouse genomic variation and its effect on phenotypes and gene regulation**. *Nature* 2011, 477:289–94.

[5] Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM, GO B, GO S, Project NES: **Evolution and functional impact of rare coding variation from deep sequencing of human exomes**. *Science* 2012, 337:64–9.

[6] O'Sullivan BP, Freedman SD: **Cystic fibrosis**. *Lancet* 2009, 373:1891–904.

[7] Walker FO: **Huntington's disease**. *Lancet* 2007, 369:218–28.

[8] Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery**. *Nat Rev Genet* 2011, 12:745–55.

[9] Manceau M, Domingues VS, Mallarino R, Hoekstra HE: **The developmental role of Agouti in color pattern evolution**. *Science* 2011, 331:1062–5.

[10] Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM, Petrov D, Jónsson B, Schluter D, Bell MA, Kingsley DM: **Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer**. *Science* 2010, 327:302–5.

[11] Doebley J, Stec A, Gustus C: **teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance**. *Genetics* 1995, 141:333–46.

[12] Mackay TF: **The genetic architecture of quantitative traits**. *Annu Rev Genet* 2001, 35:303–39.

[13] Visscher PM: **Sizing up human height variation**. *Nat Genet* 2008, 40:489–90.

[14] Deutschbauer AM, Davis RW: **Quantitative trait loci mapped to single-nucleotide resolution in yeast**. *Nat Genet* 2005, 37:1333–40.

[15] Mackay TF, Lyman RF: **Drosophila bristles and the nature of quantitative genetic variation**. *Philos Trans R Soc Lond, B, Biol Sci* 2005, 360:1513–27.

[16] Risch N, Merikangas K: **The future of genetic studies of complex human diseases**. *Science* 1996, 273:1516–7.

[17] Witte JS: **Genome-wide association studies and beyond**. *Annu Rev Public Health* 2010, 31:9–20 4 p following 20.

[18] Shapiro MD, Bell MA, Kingsley DM: **Parallel genetic origins of pelvic reduction in vertebrates**. *Proc Natl Acad Sci USA* 2006, 103:13753–8.

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

[20] Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB: **Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila**. *Nature* 2005, 433:481–7.

[21] Laere ASV, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, Archibald AL, Haley CS, Buys N, Tally M, Andersson G, Georges M, Andersson L: **A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig**. *Nature* 2003, 425:832–6.

[22] Bray NJ, Buckland PR, Williams NM, Williams HJ, Norton N, Owen MJ, O'Donovan MC: **A haplotype implicated in schizophrenia susceptibility is associated with reduced COMT expression in human brain**. *Am J Hum Genet* 2003, 73:152–61.

[23] Grady WM, Willis J, Guilford PJ, Dunbier AK, Toro TT, Lynch H, Wiesner G, Ferguson K, Eng C, Park JG, Kim SJ, Markowitz S: **Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer**. *Nat Genet* 2000, 26:16–7.

[24] Kochi Y, Yamada R, Suzuki A, Harley JB, Shirasawa S, Sawada T, Bae SC, Tokuhiro S, Chang X, Sekine A, Takahashi A, Tsunoda T, Ohnishi Y, Kaufman KM, Kang CP, Kang C, Otsubo S, Yumura W, Mimori A, Koike T, Nakamura Y, Sasazuki T, Yamamoto K: **A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities**. *Nat Genet* 2005, 37:478–85.

[25] Shin HD, Winkler C, Stephens JC, Bream J, Young H, Goedert JJ, O'Brien TR, Vlahov D, Buchbinder S, Giorgi J, Rinaldo C, Donfield S, Willoughby A, O'Brien SJ, Smith MW: **Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10**. *Proc Natl Acad Sci USA* 2000, 97:14467–72.

[26] van der Zee J, Ber IL, Maurer-Stroh S, Engelborghs S, Gijselinck I, Camuzat A, Brouwers N, Vandenberghe R, Sleegers K, Hannequin D, Dermaut B, Schymkowitz J, Campion D, Santens P, Martin JJ, Lacomblez L, Pooter TD, Peeters K, Mattheijssens M, Vercelletto M, den Broeck MV, Cruts M, Deyn PPD, Rousseau F, Brice A, Broeckhoven CV: **Mutations other than null mutations producing a pathogenic loss of progranulin in frontotemporal dementia**. *Hum Mutat* 2007, 28:416.

[27] Watts JA, Morley M, Burdick JT, Fiori JL, Ewens WJ, Spielman RS, Cheung VG: **Gene expression phenotype in heterozygous carriers of ataxia telangiectasia**. *Am J Hum Genet* 2002, 71:791–800.

[28] Yan H, Dobbie Z, Gruber SB, Markowitz S, Romans K, Giardiello FM, Kinzler KW, Vogelstein B: **Small changes in expression affect predisposition to tumorigenesis**. *Nat Genet* 2002, 30:25–6.

[29] Eisensmith RC, Woo SL: **Molecular basis of phenylketonuria and related hyperphenylalaninemias: mutations and polymorphisms in the human phenylalanine hydroxylase gene**. *Hum Mutat* 1992, 1:13–23.

[30] Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, Lewis GD, Fox CS, Jacques PF, Fernandez C, O'Donnell CJ, Carr SA, Mootha VK, Florez JC, Souza A, Melander O, Clish CB, Gerszten RE: **Metabolite profiles and the risk of developing diabetes**. *Nat Med* 2011, 17:448–53.

[31] Chan EKF, Rowe HC, Hansen BG, Kliebenstein DJ: **The complex genetic architecture of the metabolome**. *PLoS Genet* 2010, 6:e1001198.

[32] Foss EJ, Radulovic D, Shaffer SA, Ruderfer DM, Bedalov A, Goodlett DR, Kruglyak L: **Genetic basis of proteome variation in yeast**. *Nat Genet* 2007, 39:1369–75.

[33] Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, Mungrue IN, Farber CR, Sinsheimer J, Kang HM, Furlotte N, Park CC, Wen PZ, Brewer H, Weitz K, Camp DG, Pan C, Yordanova R, Neuhaus I, Tilford C, Siemers N, Gargalovic P, Eskin E, Kirchgessner T, Smith DJ, Smith RD, Lusis AJ: **Comparative analysis of proteome and transcriptome variation in mouse**. *PLoS Genet* 2011, 7:e1001393.

[34] Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikäinen LP, Kangas AJ, Soininen P, Würtz P, Silander K, Dick DM, Rose RJ, Savolainen MJ, Viikari J, Kähönen M, Lehtimäki T, Pietiläinen KH, Inouye M, McCarthy MI, Jula A, Eriksson J, Raitakari OT, Salomaa V, Kaprio J, Järvelin MR, Peltonen L, Perola M, Freimer NB, Ala-Korpela M, Palotie A, Ripatti S: **Genome-wide association study identifies multiple loci influencing human serum metabolite levels**. *Nat Genet* 2012, 44:269–76.

[35] Brem RB, Yvert G, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast**. *Science* 2002, 296:752–5.

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

[37] Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M, Hubner N, Aitman TJ: **Heritability and tissue specificity of expression quantitative trait loci**. *PLoS Genet* 2006, 2:e172.

[38] Brem RB, Kruglyak L: **The landscape of genetic complexity across 5,700 gene expression traits in yeast**. *Proc Natl Acad Sci USA* 2005, 102:1572–7.

[39] Brem RB, Storey JD, Whittle J, Kruglyak L: **Genetic interactions between polymorphisms that affect gene expression in yeast**. *Nature* 2005, 436:701–3.

[40] Storey JD, Akey JM, Kruglyak L: **Multiple locus linkage analysis of genomewide expression in yeast**. *PLoS Biol* 2005, 3:e267.

[41] Landry CR, Oh J, Hartl DL, Cavalieri D: **Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes**. *Gene* 2006, 366:343–51.

[42] Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JAG, Hazendonk E, Prins P, Plasterk RHA, Jansen RC, Breitling R, Kammenga JE: **Mapping determinants of gene expression plasticity by genetical genomics in C. elegans**. *PLoS Genet* 2006, 2:e222.

[43] Smith EN, Kruglyak L: **Gene-environment interaction in yeast gene expression**. *PLoS Biol* 2008, 6:e83.

[44] Rieseberg LH, Archer MA, Wayne RK: **Transgressive segregation, adaptation and speciation**. *Heredity* 1999, 83 (Pt 4):363–72.

[45] Mardis ER: **A decade's perspective on DNA sequencing technology**. *Nature* 2011, 470:198–203.

[46] Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, 270:467–70.

[47] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing**. *Science* 2008, 320:1344–9.

[48] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, 5:621–8.

[49] Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, 10:57–63.

[50] Wang M, You J, Bemis KG, Tegeler TJ, Brown DPG: **Label-free mass spectrometry-based protein quantification technologies in proteomic analysis**. *Briefings in functional genomics & proteomics* 2008, 7:329–39.

[51] Cooper SJ, Finney GL, Brown SL, Nelson SK, Hesselberth J, MacCoss MJ, Fields S: **High-throughput profiling of amino acids in strains of the Saccharomyces cerevisiae deletion collection**. *Genome Res* 2010, 20:1288–96.

[52] Almstetter MF, Oefner PJ, Dettmer K: **Comprehensive two-dimensional gas chromatography in metabolomics**. *Anal Bioanal Chem* 2012, 402:1993–2013.

[53] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes**. *Science* 1996, 274:546, 563–7.

[54] Fay JC, Benavides JA: **Evidence for domesticated and wild populations of Saccharomyces cerevisiae**. *PLoS Genet* 2005, 1:66–71.

[55] Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L: **Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae**. *Nature* 2009, 458:342–5.

[56] Fay JC, McCullough HL, Sniegowski PD, Eisen MB: **Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae**. *Genome Biol* 2004, 5:R26.

[57] Kvitek DJ, Will JL, Gasch AP: **Variations in Stress Sensitivity and Genomic Expression in Diverse S. cerevisiae Isolates**. *PLoS Genet* 2008, 4:e1000223.

[58] Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins**. *Science* 2000, 290:2306–9.

[59] Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation**. *Science* 2002, 295:1306–11.

[60] Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Shevchenko A, Boucherie H, Mann M: **Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels**. *Proc Natl Acad Sci USA* 1996, 93:14440–5.

[61] Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG: **A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations**. *Nat Biotechnol* 2001, 19:45–50.

[62] Skelly DA, Ronald J, Connelly CF, Akey JM: **Population genomics of intron splicing in 38 Saccharomyces cerevisiae genome sequences**. *Genome Biol Evol* 2009, 2009:466–78.

[63] Nixon JEJ, Wang A, Morrison HG, McArthur AG, Sogin ML, Loftus BJ, Samuelson J: **A spliceosomal intron in Giardia lamblia**. *Proc Natl Acad Sci USA* 2002, 99:3701–5.

[64] Simpson AGB, MacQuarrie EK, Roger AJ: **Eukaryotic evolution: early origin of canonical introns**. *Nature* 2002, 419:270.

[65] Vanácová S, Yan W, Carlton JM, Johnson PJ: **Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis**. *Proc Natl Acad Sci USA* 2005, 102:4430–5.

[66] Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best AA, Cande WZ, Chen F, Cipriano MJ, Davids BJ, Dawson SC, Elmendorf HG, Hehl AB, Holder ME, Huse SM, Kim UU, Lasek-Nesselquist E, Manning G, Nigam A, Nixon JEJ, Palm D, Passamaneck NE, Prabhu A, Reich CI, Reiner DS, Samuelson J, Svard SG, Sogin ML: **Genomic minimalism in the early diverging intestinal parasite Giardia lamblia**. *Science* 2007, 317:1921–6.

[67] Russell CB, Fraga D, Hinrichsen RD: **Extremely short 20-33 nucleotide introns are the standard length in Paramecium tetraurelia**. *Nucleic Acids Res* 1994, 22:1221–5.

[68] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf

YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J, Consortium IHGS: **Initial sequencing and analysis of the human genome**. *Nature* 2001, 409:860–921.

[69] Lynch M: **Intron evolution as a population-genetic process**. *Proc Natl Acad Sci USA* 2002, 99:6118–23.

[70] Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T: **A large-scale full-length cDNA analysis to explore the budding yeast transcriptome**. *Proc Natl Acad Sci USA* 2006, 103:17846–51.

[71] Juneau K, Palm C, Miranda M, Davis RW: **High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing**. *Proc Natl Acad Sci USA* 2007, 104:1522–7.

[72] Zhang Z, Hesselberth JR, Fields S: **Genome-wide identification of spliced introns using a tiling microarray**. *Genome Res* 2007, 17:503–9.

[73] Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtukova I, Gnirke A, Nusbaum C, Thompson DA, Friedman N, Regev A: **Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing**. *Proc Natl Acad Sci USA* 2009, 106:3264–9.

[74] Spingola M, Grate L, Haussler D, Ares M: **Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae**. *RNA* 1999, 5:221–34.

[75] Davis CA, Grate L, Spingola M, Ares M: **Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast**. *Nucleic Acids Res* 2000, 28:1700–6.

[76] Langford CJ, Klinz FJ, Donath C, Gallwitz D: **Point mutations identify the conserved, intron-contained TACTAAC box as an essential splicing signal sequence in yeast**. *Cell* 1984, 36:645–53.

[77] Teem JL, ABOVICH N, Kaufer NF, Schwindinger WF, Warner JR, LEVY A, WOOLFORD J, Leer RJ, van Raamsdonk-Duin MM, Mager WH: **A comparison of yeast ribosomal protein gene DNA sequences**. *Nucleic Acids Res* 1984, 12:8295–312.

[78] Woolford JL: **Nuclear pre-mRNA splicing in yeast**. *Yeast* 1989, 5:439–57.

[79] Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns**. *Proc Natl Acad Sci USA* 2001, 98:11193–8.

[80] Jacquier A, Rodriguez JR, Rosbash M: **A quantitative analysis of the effects of 5' junction and TACTAAC box mutants and mutant combinations on yeast mRNA splicing**. *Cell* 1985, 43:423–30.

[81] Fouser LA, Friesen JD: **Mutations in a yeast intron demonstrate the importance of specific conserved nucleotides for the two stages of nuclear mRNA splicing**. *Cell* 1986, 45:81–93.

[82] Newman AJ, Lin RJ, Cheng SC, Abelson J: **Molecular consequences of specific intron mutations on yeast mRNA splicing in vivo and in vitro**. *Cell* 1985, 42:335–44.

[83] Jacquier A, Rosbash M: **RNA splicing and intron turnover are greatly diminished by a mutant yeast branch point**. *Proc Natl Acad Sci USA* 1986, 83:5835–9.

[84] Vijayraghavan U, Parker R, Tamm J, Iimura Y, Rossi J, Abelson J, Guthrie C: **Mutations in conserved intron sequences affect multiple steps in the yeast splicing pathway, particularly assembly of the spliceosome**. *EMBO J* 1986, 5:1683–95.

[85] Castanotto D, Rossi JJ: **Cooperative interaction of branch signals in the actin intron of Saccharomyces cerevisiae**. *Nucleic Acids Res* 1998, 26:4137–45.

[86] Stajich JE, Dietrich FS, Roy SW: **Comparative genomic analysis of fungal genomes reveals intron-rich ancestors**. *Genome Biol* 2007, 8:R223.

[87] Fink GR: **Pseudogenes in yeast?** *Cell* 1987, 49:5–6.

[88] Ng R, Domdey H, Larson G, Rossi J, Abelson J: **A test for intron function in the yeast actin gene**. *Nature* 1985, .

[89] Ho CK, Abelson J: **Testing for intron function in the essential Saccharomyces cerevisiae tRNA(SerUCG) gene**. *J Mol Biol* 1988, 202:667–72.

[90] Parenteau J, Durand M, Véronneau S, Lacombe AA, Morin G, Guérin V, Cecez B, Gervais-Bird J, Koh CS, Brunelle D, Wellinger RJ, Chabot B, Elela SA: **Deletion of Many Yeast Introns Reveals a Minority of Genes that Require Splicing for Function**. *Mol Biol Cell* 2008, 19:1932–41.

[91] Maxwell ES, Fournier MJ: **The small nucleolar RNAs**. *Annu Rev Biochem* 1995, 64:897–934.

[92] Thompson-Jäger S, Domdey H: **The intron of the yeast actin gene contains the promoter for an antisense RNA**. *Curr Genet* 1990, 17:269–73.

[93] Juneau K, Miranda M, Hillenmeyer ME, Nislow C, Davis RW: **Introns regulate RNA and protein abundance in yeast**. *Genetics* 2006, 174:511–8.

[94] Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C: **Rapid, transcript-specific changes in splicing in response to environmental stress**. *Mol Cell* 2007, 27:928–37.

[95] Li B, Vilardell J, Warner JR: **An RNA structure involved in feedback regulation of splicing and of translation is critical for biological fitness**. *Proc Natl Acad Sci USA* 1996, 93:1596–600.

[96] Furger A, O'Sullivan JM, Binnie A, Lee BA, Proudfoot NJ: **Promoter proximal splice sites enhance transcription**. *Genes Dev* 2002, 16:2792–9.

[97] Belshaw R, Bensasson D: **The rise and falls of introns**. *Heredity* 2006, 96:208–13.

[98] Fedorov A, Merican AF, Gilbert W: **Large-scale comparison of intron positions among animal, plant, and fungal genes**. *Proc Natl Acad Sci USA* 2002, 99:16128–33.

[99] Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: **Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution**. *Curr Biol* 2003, 13:1512–7.

[100] Bon E, Casaregola S, Blandin G, Llorente B, Neuvéglise C, Munsterkotter M, Guldener U, Mewes HW, Helden JV, Dujon B, Gaillardin C: **Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns**. *Nucleic Acids Res* 2003, 31:1121–35.

[101] Sharpton T, Neafsey D, Galagan J, Taylor J: **Mechanisms of intron gain and loss in Cryptococcus**. *Genome Biol* 2008, 9:R24.

[102] Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M: **Intron presence-absence polymorphism in Drosophila driven by positive Darwinian selection**. *Proc Natl Acad Sci USA* 2002, 99:8121–6.

[103] Omilian AR, Scofield DG, Lynch M: **Intron presence-absence polymorphisms in Daphnia**. *Mol Biol Evol* 2008, 25:2129–39.

[104] Neuhauser C, Krone SM: **The genealogy of samples in models with selection**. *Genetics* 1997, 145:519–34.

[105] Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M, Wilhelmy J, Komp C, Tamse R, Wang X, Jia P, Luedi P, Oefner PJ, David L, Dietrich FS, Li Y, Davis RW, Steinmetz LM: **Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789**. *Proc Natl Acad Sci USA* 2007, 104:12825–30.

[106] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED: **Saccharomyces Genome Database: the genomics resource of budding yeast**. *Nucleic Acids Res* 2012, 40:D700–5.

[107] Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences**. *J Comput Biol* 2000, 7:203–14.

[108] Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. *Mol Biol Evol* 2007, 24:1586–91.

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

[110] Rozen S, Skaletsky HJ: **Primer3 on the www for general users and for biologist programmers**. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, edited by Krawetz S, Misener S. Totowa, NJ: Humana Press, 2000, 365–386.

[111] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010.

[112] Lopez PJ, Séraphin B: **Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition**. *RNA* 1999, 5:1135–7.

[113] Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism**. *Genetics* 1989, 123:585–95.

[114] Bembom O: *seqLogo: Sequence logos for DNA sequence alignments*, 2007.

[115] Fay JC, Benavides JA: **Hypervariable noncoding sequences in Saccharomyces cerevisiae**. *Genetics* 2005, 170:1575–87.

[116] Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW: **Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition**. *Genetics* 2002, 161:447–59.

[117] Crow JF, Kimura M: *An introduction to population genetics theory*. New York: Harper and Row, 1970.

[118] Pritchard JK: **Are rare variants responsible for susceptibility to complex diseases?** *Am J Hum Genet* 2001, 69:124–37.

[119] Kimura M, Ohta T: **The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population**. *Genetics* 1969, 61:763–771.

[120] Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Güldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kötter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, yun Wang C, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M: **Functional profiling of the Saccharomyces cerevisiae genome**. *Nature* 2002, 418:387–91.

[121] Sharp PM, Li WH: **The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications**. *Nucleic Acids Res* 1987, 15:1281–95.

[122] Parker R, Guthrie C: **A point mutation in the conserved hexanucleotide at a yeast 5' splice junction uncouples recognition, cleavage, and ligation**. *Cell* 1985, 41:107–18.

[123] Couto JR, Tamm J, Parker R, Guthrie C: **A trans-acting suppressor restores splicing of a yeast intron with a branch point mutation**. *Genes Dev* 1987, 1:445–55.

[124] Kivens W, Siliciano PG: **RNA sequences upstream of the 3' splice site repress splicing of mutantyeast ACT1 introns**. *RNA* 1996, 2:492–505.

[125] Spingola M, Ares M: **A yeast intronic splicing enhancer and Nam8p are required for Mer1p-activated splicing**. *Mol Cell* 2000, 6:329–38.

[126] Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L: **Population history and natural selection shape patterns of genetic variation in 132 genes**. *PLoS Biol* 2004, 2:e286.

[127] Akashi H: **Gene expression and molecular evolution**. *Curr Opin Genet Dev* 2001, 11:660–6.

[128] Danin-Kreiselman M, Lee CY, Chanfreau G: **RNAse III-mediated degradation of unspliced pre-mRNAs and lariat introns**. *Mol Cell* 2003, 11:1279–89.

[129] Hilleren PJ, Parker R: **Cytoplasmic degradation of splice-defective pre-mRNAs and intermediates**. *Mol Cell* 2003, 12:1453–65.

[130] Sayani S, Janis M, Lee CY, Toesca I, Chanfreau GF: **Widespread impact of nonsense-mediated mRNA decay on the yeast intronome**. *Mol Cell* 2008, 31:360–70.

[131] Engebrecht JA, Voelkel-Meiman K, Roeder GS: **Meiosis-specific RNA splicing in yeast**. *Cell* 1991, 66:1257–68.

[132] Maniatis T, Reed R: **An extensive network of coupling among gene expression machines**. *Nature* 2002, 416:499–506.

[133] Ares M, Grate L, Pauling MH: **A handful of intron-containing genes produces the lion's share of yeast mRNA**. *RNA* 1999, 5:1138–9.

[134] Lynch M, Richardson AO: **The evolution of spliceosomal introns**. *Curr Opin Genet Dev* 2002, 12:701–10.

[135] Derr LK, Strathern JN, Garfinkel DJ: **RNA-mediated recombination in S. cerevisiae**. *Cell* 1991, 67:355–64.

[136] Ronald J, Akey JM: **The evolution of gene expression QTL in Saccharomyces cerevisiae**. *PLoS ONE* 2007, 2:e678.

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

[138] Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM: **A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data**. *Genome Res* 2011, 21:1728–37.

[139] Britten RJ, Davidson EH: **Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty**. *The Quarterly review of biology* 1971, 46:111–38.

[140] King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees**. *Science* 1975, 188:107–16.

[141] Tournamille C, Colin Y, Cartron JP, Kim CLV: **Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals**. *Nat Genet* 1995, 10:224–8.

[142] Rockman MV, Kruglyak L: **Genetics of global gene expression**. *Nat Rev Genet* 2006, 7:862–72.

[143] Skelly DA, Ronald J, Akey JM: **Inherited variation in gene expression**. *Annual review of genomics and human genetics* 2009, 10:313–32.

[144] Cowles CR, Hirschhorn JN, Altshuler D, Lander ES: **Detection of regulatory variation in mouse genes**. *Nat Genet* 2002, 32:432–7.

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

[146] Wittkopp PJ, Haerum BK, Clark AG: **Evolutionary changes in cis and trans gene regulation**. *Nature* 2004, 430:85–8.

[147] Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP: **Allelic variation in gene expression is common in the human genome**. *Genome Res* 2003, 13:1855–62.

[148] Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L: **Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays**. *Genome Res* 2005, 15:284–91.

[149] Pant PVK, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA: **Analysis of allelic differential expression in human white blood cells**. *Genome Res* 2006, 16:331–9.

[150] Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan JB, Hudson TJ: **Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression**. *PLoS Genet* 2008, 4:e1000006.

[151] Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP: **In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading**. *Nat Genet* 2003, 33:469–75.

[152] Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK: **Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data**. *Bioinformatics* 2009, 25:3207–12.

[153] Main BJ, Bickel RD, McIntyre LM, Graze RM, Calabrese PP, Nuzhdin SV: **Allele-specific expression assays using Solexa**. *BMC Genomics* 2009, 10:422.

[154] Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, LeProust EM, Eggan K, Church GM: **Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human**. *Nat Methods* 2009, 6:613–8.

[155] Emerson JJ, Hsieh LC, Sung HM, Wang TY, Huang CJ, Lu HHS, Lu MYJ, Wu SH, Li WH: **Natural selection on cis and trans regulation in yeasts**. *Genome Res* 2010, 20:826–36.

[156] Heap GA, Yang JHM, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ, Albert TJ, Rodesch MJ, Clayton DG, Todd JA, van Heel DA, Plagnol V: **Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing**. *Hum Mol Genet* 2010, 19:122–34.

[157] McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ: **Regulatory divergence in Drosophila revealed by mRNA-seq**. *Genome Res* 2010, 20:816–25.

[158] Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population**. *Nature* 2010, 464:773–7.

[159] Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing**. *Nature* 2010, 464:768–72.

[160] Hoffman CS, Winston F: **A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of Escherichia coli**. *Gene* 1987, 57:267–72.

[161] Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0**, 2010. URL `http://www.repeatmasker.org`.

[162] Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W: **Aligning multiple genomic sequences with the threaded blockset aligner**. *Genome Res* 2004, 14:708–15.

[163] Homer N, Merriman B, Nelson SF: **BFAST: an alignment tool for large scale genome resequencing**. *PLoS ONE* 2009, 4:e7767.

[164] Bullard JH, Purdom EA, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments**. *BMC Bioinformatics* 2010, 94.

[165] Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing**. *Nucleic Acids Res* 2008, 36:e105.

[166] Consortium IH, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PIW, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJR, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE: **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, 467:52–8.

[167] Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads**. *Bioinformatics* 2010, 26:873–81.

[168] Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: **The Ensembl genome database project**. *Nucleic Acids Res* 2002, 30:38–41.

[169] Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci USA* 2003, 100:9440–5.

[170] Bray NJ, Buckland PR, Owen MJ, O'Donovan MC: **Cis-acting variation in the expression of a high proportion of genes in human brain**. *Hum Genet* 2003, 113:149–53.

[171] Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, Lavergne K, Villeneuve A, Gaudin T, Brändström H, Beck A, Verner A, Kingsley J, Harmsen E, Labuda D, Morgan K, Vohl MC, Naumova AK, Sinnett D, Hudson TJ: **A survey of genetic and epigenetic variation affecting human gene expression**. *Physiol Genomics* 2004, 16:184–93.

[172] Ge B, Pokholok D, Kwan T, Grundberg E, Morcos L, Verlaan D, Le J, Koka V, Lam K, Gagné V, Dias J, Hoberman R, Montpetit A, Joly M, Harvey E, Sinnett D, Beaulieu P, Hamon R, Graziani A, Dewar K, Harmsen E, Majewski J, Göring H, Naumova A, Blanchette M, Gunderson K, Pastinen T: **Global patterns of cis variation in human cells revealed by high-density allelic expression analysis**. *Nat Genet* 2009, .

[173] van Camp G, Coucke P, Balemans W, van Velzen D, van de Bilt C, van Laer L, Smith RJ, Fukushima K, Padberg GW, Frants RR: **Localization of a gene for non-syndromic hearing loss (DFNA5) to chromosome 7p15**. *Hum Mol Genet* 1995, 4:2159–63.

[174] Shoemaker R, Deng J, Wang W, Zhang K: **Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome**. *Genome Res* 2010, 20:883–9.

[175] Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA: **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting**. *Nat Methods* 2009, 6:283–9.

[176] Morozova O, Hirst M, Marra MA: **Applications of new sequencing technologies for transcriptome analysis**. *Annual review of genomics and human genetics* 2009, 10:135–51.

[177] Schork NJ: **Genetics of complex disease: approaches, problems, and solutions**. *Am J Respir Crit Care Med* 1997, 156:S103–9.

[178] Freimer N, Sabatti C: **The human phenome project**. *Nat Genet* 2003, 34:15–21.

[179] Houle D, Govindaraju DR, Omholt S: **Phenomics: the next challenge**. *Nat Rev Genet* 2010, 11:855–66.

[180] Warringer J, Ericson E, Fernandez L, Nerman O, Blomberg A: **High-resolution yeast phenomics resolves different physiological features in the saline response**. *Proc Natl Acad Sci USA* 2003, 100:15724–9.

[181] Ratnakumar S, Hesketh A, Gkargkas K, Wilson M, Rash BM, Hayes A, Tunnacliffe A, Oliver SG: **Phenomic and transcriptomic analyses reveal that autophagy plays a major role in desiccation tolerance in Saccharomyces cerevisiae**. *Mol Biosyst* 2011, 7:139–49.

[182] Warringer J, Zörgö E, Cubillos FA, Zia A, Gjuvsland A, Simpson JT, Forsmark A, Durbin R, Omholt SW, Louis EJ, Liti G, Moses A, Blomberg A: **Trait variation in yeast is defined by population history**. *PLoS Genet* 2011, 7:e1002111.

[183] Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M: **Personal omics profiling reveals dynamic molecular and medical phenotypes**. *Cell* 2012, 148:1293–307.

[184] Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, 25:1754–60.

[185] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nat Genet* 2011, 43:491–8.

[186] Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language**. *Bioinformatics* 2004, 20:289–90.

[187] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[188] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, 25:2078–9.

[189] Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, 26:841–2.

[190] Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biol* 2010, 11:R25.

[191] Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, 26:139–40.

[192] Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments**. *Bioinformatics* 2012, 28:882–3.

[193] Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, 19:185–93.

[194] Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis**. *PLoS Genet* 2007, 3:1724–35.

[195] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays**. *Bioinformatics* 2001, 17:520–5.

[196] Nogami S, Ohya Y, Yvert G: **Genetic complexity and quantitative trait loci mapping of yeast morphological traits**. *PLoS Genet* 2007, 3:e31.

[197] Bader GD, Hogue CWV: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, 4:2.

[198] Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, 27:431–2.

[199] Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, Ribeca P: **Fast computation and applications of genome mappability**. *PLoS ONE* 2012, 7:e30377.

[200] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, 15:1034–50.

[201] Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, 25:1105–11.

[202] Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif**. *Bioinformatics* 2011, 27:1017–8.

[203] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, 28:511–5.

[204] Connelly CF, Akey JM: **On the Prospects of Whole-Genome Association Mapping in Saccharomyces cerevisiae**. *Genetics* 2012, 191:1345–53.

[205] Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, 21:263–5.

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

[207] Breiman L: **Random forests**. *Machine Learning* 2001, 45:5–32.

[208] Segal MR: **Machine learning benchmarks and random forest regression**. Technical report, University of California, 2004.

[209] Liaw A, Wiener M: **Classification and regression by randomforest**. *R News* 2002, 2:18–22.

[210] Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW: **Systematic screen for human disease genes in yeast**. *Nat Genet* 2002, 31:400–4.

[211] Basehoar AD, Zanton SJ, Pugh BF: **Identification and distinct regulation of yeast TATA box-containing genes**. *Cell* 2004, 116:699–709.

[212] Seizl M, Hartmann H, Hoeg F, Kurth F, Martin DE, Söding J, Cramer P: **A conserved GA element in TATA-less RNA polymerase II promoters**. *PLoS ONE* 2011, 6:e27595.

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

[214] Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res* 2011, .

[215] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic Acids Res* 2012, 40:D109–14.

[216] Martens JA, Laprade L, Winston F: **Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene**. *Nature* 2004, 429:571–4.

[217] Hongay CF, Grisafi PL, Galitski T, Fink GR: **Antisense transcription controls cell fate in Saccharomyces cerevisiae**. *Cell* 2006, 127:735–45.

[218] Camblong J, Iglesias N, Fickentscher C, Dieppois G, Stutz F: **Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in S. cerevisiae**. *Cell* 2007, 131:706–17.

[219] Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, Egholm M, Chambers PJ: **Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of Saccharomyces cerevisiae**. *PLoS Genet* 2011, 7:e1001287.

[220] Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, Cambon B, Legras JL, Wincker P, Casaregola S, Dequin S: **Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118**. *Proc Natl Acad Sci USA* 2009, 106:16333–8.

[221] Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OVC, Missawa SK, Galzerani F, Costa GGL, Vidal RO, Noronha MF, Dominska M, Andrietta MGS, Andrietta SR, Cunha AF, Gomes LH, Tavares FCA, Alcarde AR, Dietrich FS, McCusker JH, Petes TD, Pereira GAG: **Genome structure of a Saccharomyces cerevisiae strain widely used in bioethanol production**. *Genome Res* 2009, 19:2258–70.

[222] Foss EJ, Radulovic D, Shaffer SA, Goodlett DR, Kruglyak L, Bedalov A: **Genetic Variation Shapes Protein Networks Mainly through Non-transcriptional Mechanisms**. *PLoS Biol* 2011, 9:e1001144.

[223] Greenbaum D, Colangelo C, Williams K, Gerstein M: **Comparing protein abundance and mRNA expression levels on a genomic scale**. *Genome Biol* 2003, 4:117.

[224] de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C: **Global signatures of protein and mRNA expression levels**. *Mol Biosyst* 2009, 5:1512–26.

[225] Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control**. *Nature* 2011, 473:337–42.

[226] Tirosh I, Weinberger A, Carmi M, Barkai N: **A genetic signature of interspecies variations in gene expression**. *Nat Genet* 2006, 38:830–4.

[227] Vogel C, Marcotte EM: **Insights into the regulation of protein abundance from proteomic and transcriptomic analyses**. *Nat Rev Genet* 2012, 13:227–32.

[228] Zbuk KM, Eng C: **Cancer phenomics: RET and PTEN as illustrative models**. *Nat Rev Cancer* 2007, 7:35–45.

[229] Ng PC, Murray SS, Levy S, Venter JC: **An agenda for personalized medicine**. *Nature* 2009, 461:724–6.

[230] Gonzaga-Jauregui C, Lupski JR, Gibbs RA: **Human genome sequencing in health and disease**. *Annu Rev Med* 2012, 63:35–61.

[231] Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW: **A whole-cell computational model predicts phenotype from genotype**. *Cell* 2012, 150:389–401.

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

[233] Markowetz F, Spang R: **Inferring cellular networks–a review**. *BMC Bioinformatics* 2007, 8 Suppl 6:S5.

[234] Chen LS, Emmert-Streib F, Storey JD: **Harnessing naturally randomized transcription to infer regulatory relationships among genes**. *Genome Biol* 2007, 8:R219.

[235] Kang EY, Ye C, Shpitser I, Eskin E: **Detecting the presence and absence of causal relationships between expression of yeast genes with very few samples**. *J Comput Biol* 2010, 17:533–46.

[236] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays**. *Genome Res* 2008, 18:1509–17.

[237] Oshlack A, Wakefield MJ: **Transcript length bias in RNA-seq data confounds systems biology**. *Biol Direct* 2009, 4:14.

[238] Churchill FB: **William Johannsen and the genotype concept**. *J Hist Biol* 1974, 7:5–30.

[239] Johannsen W: **The genotype conception of heredity**. *The American Naturalist* 1911, 45:129–159.

[240] Ehrenreich IM, Bloom J, Torabi N, Wang X, Jia Y, Kruglyak L: **Genetic architecture of highly complex chemical resistance traits across four yeast strains**. *PLoS Genet* 2012, 8:e1002570.

[241] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, 25:25–9.

[242] Wigginton JE, Cutler DJ, Abecasis GR: **A note on exact tests of Hardy-Weinberg equilibrium**. *Am J Hum Genet* 2005, 76:887–93.

[243] MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ: **Skyline: an open source document editor for creating and analyzing targeted proteomics experiments**. *Bioinformatics* 2010, 26:966–8.

[244] Hsieh EJ, Hoopmann MR, MacLean B, MacCoss MJ: **Comparison of database search strategies for high precursor mass accuracy MS/MS data**. *J Proteome Res* 2010, 9:1138–43.

[245] Eng JK, McCormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database**. *Journal of the American Society of Mass Spectrometry* 1994, 5:976–989.

[246] Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ: **Semi-supervised learning for peptide identification from shotgun proteomics datasets**. *Nat Methods* 2007, 4:923–5.

[247] Zhang B, Chambers MC, Tabb DL: **Proteomic parsimony through bipartite graph analysis improves accuracy and transparency**. *J Proteome Res* 2007, 6:3549–57.

[248] Hsieh EJ, Shulman NJ, Dai DF, Vincow ES, Karunadharma PP, Pallanck L, Rabinovitch PS, MacCoss MJ: **Topograph, a software platform for precursor enrichment corrected global protein turnover measurements**. *Mol Cell Proteomics* 2012, 11:1468–74.

[249] Fowler DM, Cooper SJ, Stephany JJ, Hendon N, Nelson S, Fields S: **Suppression of statin effectiveness by copper and zinc in yeast and human cells**. *Mol Biosyst* 2011, 7:533–44.

[250] Castillo S, Mattila I, Miettinen J, Orešič M, Hyötyläinen T: **Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry**. *Anal Chem* 2011, 83:3058–67.

[251] Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S, Suzuki G, Watanabe M, Hirata A, Ohtani M, Sawai H, Fraysse N, Latgé JP, François JM, Aebi M, Tanaka S, Muramatsu S, Araki H, Sonoike K, Nogami S, Morishita S: **High-dimensional and large-scale phenotyping of yeast mutants**. *Proc Natl Acad Sci USA* 2005, 102:19015–20.

[252] Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness**. *Nat Genet* 2006, 38:203–8.

Appendix A

INHERITED VARIATION IN GENE EXPRESSION

This appendix contains material from [143], with permission.

Inherited Variation in Gene Expression

Daniel A. Skelly, James Ronald, and Joshua M. Akey

Department of Genome Sciences, University of Washington, Seattle, Washington, 98195; email: daskelly@u.washington.edu, jscr@u.washington.edu, akeyj@u.washington.edu

Annu. Rev. Genomics Hum. Genet. 2009. 10:313–32

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev-genom-082908-150121

1527-8204/09/0922-0313$20.00

Key Words

gene expression variation, eQTL, allele-specific expression, genetical genomics, microarrays, expression heterogeneity

Abstract

Variation in gene expression constitutes an important source of biological variability within and between populations that is likely to contribute significantly to phenotypic diversity. Recent conceptual, technical, and methodological advances have enabled the genome-scale dissection of transcriptional variation. Here, we outline common approaches for detecting gene expression quantitative trait loci, and summarize the insights gleaned from these studies regarding the genetic architecture of transcriptional variation and the nature of regulatory alleles. Particular emphasis is placed on human studies, and we discuss experimental designs that ensure that increasingly large and complex studies continue to advance our understanding of gene expression variation. We conclude by discussing the evolution of gene expression levels, and we explore prospects for leveraging new technological developments to investigate inherited variation in gene expression in even greater depth.

INTRODUCTION

Heritability: the proportion of total phenotypic variation that is attributable to genetic variation

Gene expression is an important molecular phenotype, providing the initial step in bridging the divide between static genomic information and dynamic organismal phenotypes. Thus, variation in gene expression levels is thought to constitute a significant source of phenotypic diversity among individuals within populations and to contribute to the evolutionary divergence between species (12, 55). An increasing number of empirical studies highlight the pervasive phenotypic effects of regulatory variation, such as skeletal morphology in stickleback fish (96), beak morphology in Darwin finches (1), cuticular pigmentation in *Drosophila* (39), and muscle growth in pigs (111). In humans, regulatory alleles have been linked to susceptibility to infectious, autoimmune, psychiatric, neoplastic, and neurodegenerative disease (7, 41, 61, 99, 110, 118). Even relatively modest expression changes can have significant biological consequences, as seen for the tumor suppressor gene *APC*, in which a 50% change in gene expression can lead to development of familial neoplasia (123). Additional examples of phenotypes influenced by gene expression have been extensively reviewed elsewhere (57, 59, 122).

More recently, considerable interest has focused on inferring general principles of naturally occurring gene expression variation, such as the amount of gene expression variation governed by genetic variation, the nature of specific regulatory alleles, and mechanisms of transcriptional variation. Here, we review recent developments in the burgeoning field of gene expression genetics. We focus on several important themes that have emerged from the analysis of heritable variation of gene expression levels across a wide spectrum of species, and how such studies have illuminated principles of regulatory evolution. In addition, we provide a more detailed synopsis of recent work in humans and suggest future avenues of research to better delimit the relationship between genetic and gene expression variation.

A SIGNIFICANT PROPORTION OF EXPRESSION VARIATION IS HERITABLE

The advent of microarrays in the mid 1990s heralded a new era wherein it became possible to measure the abundances of a large number of transcripts simultaneously. Experiments using microarray technology have established that there is widespread variation in gene expression levels between individuals within natural populations (14, 75, 77, 103, 109). For such variation to contribute to evolutionary change, it must have a heritable component. The heritability of gene expression variation on a genome-wide scale was first estimated in a cross between a laboratory and a wild strain of *Saccharomyces cerevisiae* (11). Strikingly, the median heritability of transcript levels among segregants for genes that showed parental differences in expression was 84%, indicating a substantial genetic component to transcriptional variation in yeast.

Early studies in humans demonstrated familial aggregation of expression levels suggestive of a genetic component to gene expression variation (14), and found that approximately one third of genes differentially expressed between members of CEPH families had significant heritability (68, 93). A more recent study of blood and adipose tissues in Icelanders detected significant heritability in ~55%–75% of transcripts, with the genetic component explaining 30% of transcriptional variation on average (30). Because the proportion of genes detected as heritable depends on the statistical power of the study, and heritability estimates for particular transcripts are specific to the population and environment in which they are measured (35, 113), estimates from different studies are not directly comparable. Nevertheless, it is clear that a significant proportion of variation in gene expression is heritable in all organisms studied to date.

APPROACHES FOR IDENTIFYING GENE EXPRESSION QUANTITATIVE TRAIT LOCI

Once the heritable nature of gene expression variation was documented, attention shifted

toward identifying genomic loci that harbor regulatory variation. In the following, we summarize the two basic strategies that are most commonly used to map gene expression quantitative trait loci (QTL).

Genetical Genomics

A powerful paradigm to emerge in the past decade has been the combination of traditional genetic mapping methods with microarray technology in an approach coined "genetical genomics" by Jansen & Nap (47). The basic idea is to obtain marker genotypes and measure gene expression levels in either related or unrelated individuals, treat each gene expression level as a quantitative trait, and correlate patterns of genetic variation with expression variation (**Figure 1**). The resulting set of expression QTL comprises genomic locations that are defined by statistically significant correlation to one or more transcript levels. This approach has been used to dissect the genetic basis of global gene expression in a diverse cadre of organisms, including humans (see references in **Table 1**) and other mammals (45, 93), plants (23, 56), flies (33), worms (65), and yeast (11).

The sheer number of traits measured in such genome-wide analyses affords the opportunity to make general inferences about the genetic architecture of quantitative traits. As eloquently articulated by Rockman & Kruglyak (89), the

Quantitative trait loci (QTL): a position in the genome where genetic variation is correlated with variation of a quantitative trait of interest

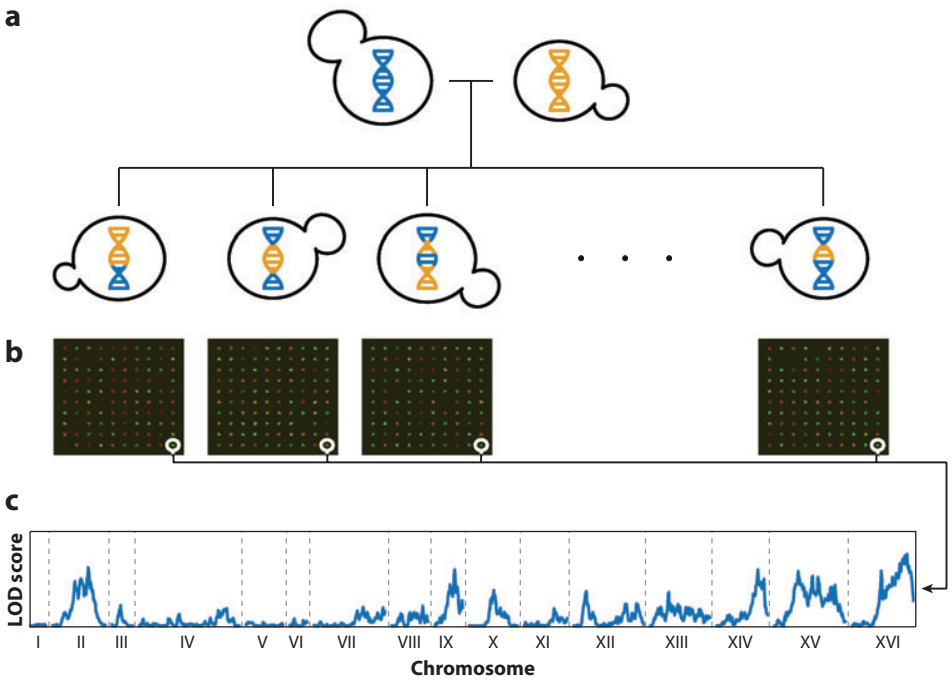


Figure 1

A typical genetical genomics experiment. This illustration involves a cross between two genetically distinct yeast strains. Segregants are depicted below the parents (*a*), with genomes that consist of randomly assorted contributions from both parental genomes. Gene expression levels of all segregants are measured using microarrays (*b*). Traditional linkage analysis is used to map QTL for each of the gene expression traits (i.e., the thousands of spots measured on each microarray). The result of a linkage analysis for one transcript level (*c*). Roman numerals designate the 16 chromosomes of yeast, and the height of the peaks depicts the LOD (logarithm of odds) score, representing the strength of evidence in favor of genetic linkage for each position in the genome.

Linkage disequilibrium (LD): the nonrandom association of alleles between two or more loci. In essence, if knowing the allele at locus one provides information about the allele present at locus two, then LD is said to exist

simultaneous study of thousands of traits provides the opportunity to explore in detail the landscape of all possible genetic architectures. Furthermore, mapping the genetic determinants of transcript levels in an unbiased fashion provides information about loci that can affect gene expression by any of a multitude of possible mechanisms. We discuss insights into the architecture of transcriptional variation and mechanisms of gene expression QTL action below.

Detecting Allele-Specific Expression

An alternative avenue for identifying regulatory variation is to specifically test for allelic expression differences at individual genes (19, 124). The most common approach for detecting allele-specific expression is to compare the expression level between two alleles distinguishable by a transcribed polymorphism; the observation of differential expression between alleles suggests that either the polymorphism studied directly affects expression or is in linkage disequilibrium (LD) with the actual causal polymorphism (**Figure 2**). Measurements of allelic expression are typically made within heterozygous individuals, allowing for the internal control of *trans*-acting and environmental factors.

A variety of methods have been used to quantify allele-specific differences in gene expression. If a transcribed polymorphism is used to distinguish alleles, transcript levels can be determined by allele-specific quantitative PCR (66, 95) or PCR and single-base extension followed by fluorescence- or mass spectrometry-based detection (6, 19, 27, 80). Alternatively, methods based on hybridization differences between allelic transcripts may be used. These methods appear to be generally less precise than the aforementioned options, although they have shown impressive potential to dramatically increase the number of genes that can be queried in a single experiment (5, 66, 79, 91). An entirely different approach involves performing chromatin immunoprecipitation using antibodies to phosphorylation sites specific to transcription initiation by RNA polymerase II (58). This method allows the quantification of allele-specific expression for genes with polymorphisms that are nearby though not transcribed.

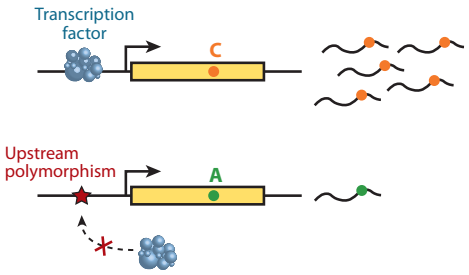


Figure 2

Detecting allele-specific gene expression. Two alleles whose expression level differs due to an upstream polymorphism are shown. They are distinguishable by a transcribed polymorphism. In the C-containing allele (*top*), a transcription factor promotes a high level of transcription under the conditions depicted, resulting in many RNA molecules containing G at the polymorphic position (the complementary base to C; *orange dots*). In the A-containing allele (*bottom*), an upstream polymorphism (*star*) prevents the transcription factor from binding, keeping transcription at a basal level and resulting in few RNA molecules containing U at the polymorphic position (the complementary base to A; *green dot*). Here, it is assumed that the two polymorphisms are in LD. A typical assay for allele-specific expression could identify this gene as showing differential allelic expression, but would not isolate the specific causative polymorphism.

THE NOMENCLATURE OF GENE EXPRESSION QTL

Using the techniques detailed above, expression QTL can be described in positional and mechanistic terms. That is, where are the loci that contribute to variation in transcript abundance, and how do those loci act to affect expression? To distinguish between these two fundamentally distinct ways of classifying expression QTL, a new nomenclature has emerged (89, 92): positional information is described by the terms local and distant, and mechanistic information is denoted by the terms *cis* and *trans*.

The positional classification of expression QTL follows naturally from the relationship between the genomic origin of a particular transcript and the genomic location of its expression QTL. If the expression QTL is close to the gene encoding the transcript under inspection, it is classified as local; otherwise, it is classified as distant. The precise physical distance used to classify an expression QTL as local or distant is arbitrary, although it can be framed to ensure that the probability of an expression trait correlating with a marker close to the gene simply by chance is small (11).

The specific mechanisms through which an expression QTL mediates transcriptional variation are numerous, but can be classified broadly into two categories. *Cis*-acting regulatory QTL affect transcript levels in an allele-specific manner, an influence that is usually exerted from a position coincident with the gene being regulated. In contrast, *trans*-acting regulatory QTL modify the expression of both alleles of a given transcript. It is typically assumed that a large proportion of local regulatory variation acts in *cis*, whereas most distant regulatory variation acts in *trans*. In general, a detailed nucleotide-level description of the mechanistic basis of inherited variation in gene expression is lacking for all but a few transcripts that are related to disease (see Reference 57 for a review) or to morphological traits (85), or that were selected for further characterization from a large group of expression QTL (92, 108, 127).

THE ARCHITECTURE OF TRANSCRIPTIONAL VARIATION IS COMPLEX

A surprising observation gleaned from studies of expression QTL is that the genetic architecture of transcriptional variation is complex. This is perhaps most clearly seen in the first empirical study of the genetics of global gene expression (11). In this study, 1528 genes showed differential expression with high confidence between two strains of *S. cerevisiae*, but only 308 of these genes showed statistically significant linkage to at least one genomic locus. Because the study design was well powered to identify expression QTL explaining a high fraction of transcript variation, the failure to detect loci underlying ~80% of expression differences suggests that polygenic control of transcript levels is common (11).

The polygenic basis of transcriptional variation has since become abundantly clear in a variety of other organisms (e.g., 84, 93). Correspondingly, only a small minority of transcripts appears to show heritable variation that is due to a single expression QTL (9). Even transcripts governed by a single expression QTL might have a complex genetic basis. For example, in one human expression QTL dissected to the nucleotide level, differences in allelic expression of the transcript result from the cumulative effects of at least five separate *cis*-regulatory polymorphisms (108).

The architecture of transcriptional variation shows many additional hallmarks of genetic complexity, including epistasis, genotype-environment interaction, and pleiotropy. Although the identification of specific pairs of loci that interact epistatically can be challenging, it is clear that a significant portion of gene expression variation results from interacting loci (10, 102). Similarly, a substantial proportion of transcript levels appears to be influenced by gene-environment interaction (where the effect of a regulatory QTL depends on the environment), although only a few environments and model organisms have been examined (62, 65, 101). For those environments that have been studied, distant QTL appear more likely to show condition-specific effects than local QTL, possibly because *trans*-acting factors are often proteins that can interact with environment-specific cellular constituents (65, 101).

The existence of multiple traits that show linkage to the same genomic region (11, 28, 29, 69, 93, 94) suggests that pleiotropy is widespread, although this observation is complicated by two factors. First, hotspots containing transcriptional regulators that affect many gene expression traits may be attributable to statistical artifacts or artificial selective forces (see section below on The Evolution

Epistasis: a statistical interaction between genotypes at two or more loci

Pleiotropy: the phenomenon in which one gene affects multiple phenotypes

Allelic heterogeneity: the presence of different alleles at a locus that result in similar phenotypic effects. Regulatory variation showing allelic heterogeneity is detectable using linkage methods, but problematic for association studies that correlate specific alleles with transcript levels

of Gene Expression). Second, the imprecision with which QTL are localized may mask the presence of multiple QTL or quantitative trait nucleotides that each affect distinct expression traits. The recent dissection of a QTL hotspot in mice highlights the fact that a single QTL affecting multiple expression levels may, upon closer observation, actually consist of multiple linked QTL that are not necessarily functionally related (71).

Finally, a few characteristics of variation in transcript abundances provide clues to the types of alleles that segregate in populations and contribute to such variation. For example, gene expression levels frequently exhibit transgressive segregation (9), the emergence of a phenotype in offspring that is more extreme than that observed in either parental line. This phenomenon is often attributed to the presence of epistatic interactions between alleles or the segregation of alleles of opposing additive effects in parental populations (86). The presence of different alleles with similar phenotypic effects at a particular locus indicates allelic heterogeneity, a phenomenon that might explain some instances where expression QTL identified by linkage are not replicated by association study designs (15).

A MULTITUDE OF MECHANISMS CONTRIBUTE TO EXPRESSION VARIATION

Local Regulatory Variation

The majority of local regulatory variation appears to act in *cis* to affect expression in an allele-specific manner (92). The process by which DNA sequence information is converted into mRNA provides numerous opportunities for genetic perturbations to alter allele-specific expression (**Figure 3*a***). Perhaps the most straightforward way in which allelic expression can be changed is by the presence of a polymorphism in a regulatory sequence, such as a transcription factor binding site. Such variation would likely effect differential allelic expression by influencing the rate of transcription initiation. Consistent with this mechanism, genes with local regulatory variation in yeast show enrichment for polymorphisms in the promoter region, some of which fall within predicted transcription factor binding sites (92). A similar increased concentration of expression QTL around the transcription start site has been observed in humans (112).

In a complementary fashion, transcript levels can be determined by precise, transcript-specific decay rates (116). The presence of regulatory polymorphisms near the 3′ end of the transcript (92, 112) might be indicative of variants affecting mRNA stability (e.g., 49, 72, 78, 97). Polymorphisms in coding or intronic sequences could affect the splicing process and lead to intron retention, exon skipping, or production of alternative transcripts. At a broader scale, DNA sequence variation could produce allele-specific expression by exerting effects on chromatin structure (29, 50).

Finally, although it is often assumed that local regulatory variation acts in *cis*, it can also act in *trans* and affect both alleles. In one well-characterized example, a coding polymorphism in the gene *AMN1* in *S. cerevisiae* results in a missense amino acid change that impairs the capacity of the Amn1 protein for negative self-regulation, leading to increased expression of both alleles (92). Similarly, the presence of a closely linked gene that regulates a particular transcript can also lead to local regulatory variation that acts in *trans*.

Distant Regulatory Variation

Distant regulatory variation is generally assumed to act in *trans* through mechanisms that affect both alleles at a locus. Presumably, most distant regulatory QTL assert their effects by perturbing links in the transcriptional regulatory network. These QTL occur in coding or *cis*-regulatory sequences within genes that either directly or indirectly affect the expression of one or more other genes (**Figure 3*b***, ***ii*** and ***iii***). Distant regulatory variation can also occur in regulatory sequences, such as enhancers

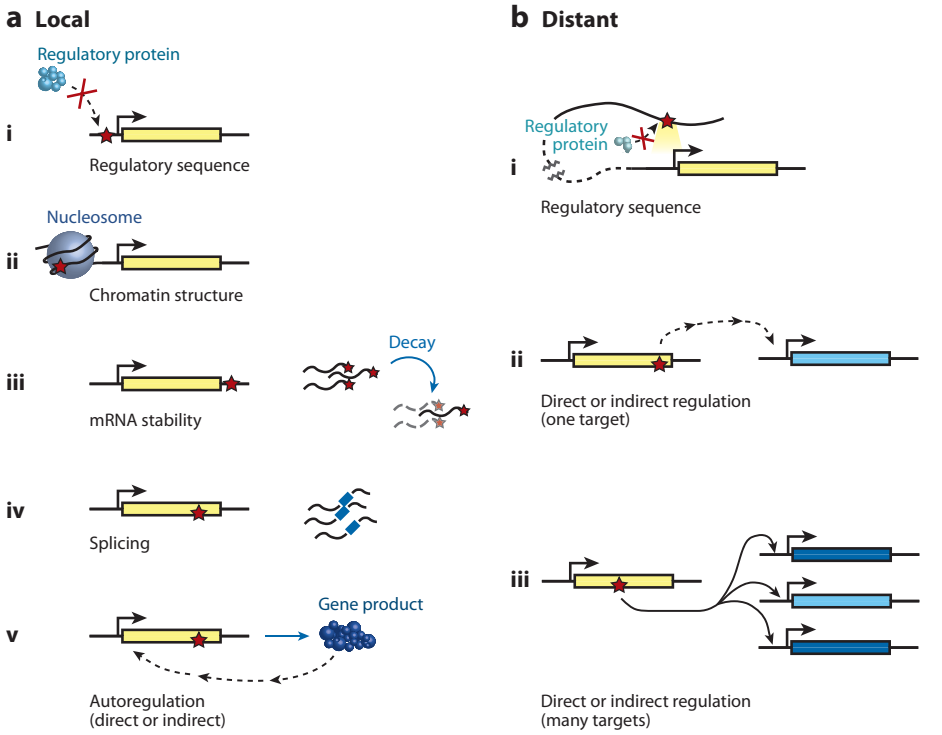


Figure 3

Molecular mechanisms of local and distant regulatory variation. (*a*) Local regulatory variation acts from a position near the gene of interest. This type of variation can impact gene expression levels by affecting (*i*) the binding of regulatory proteins to regulatory sequences, (*ii*) nucleosome binding or chromatin remodeling to influence chromatin structure, (*iii*) sequences that contribute to transcript-specific decay rates to determine mRNA stability, (*iv*) transcript structure as determined by the fidelity of intron splicing, and (*v*) regulation of the gene by its own product or the product of a gene downstream in the transcriptional regulatory network. (*b*) Distant regulatory variation acts from a position far from the gene of interest. This type of variation can impact gene expression levels by affecting (*i*) the binding of regulatory proteins to distant regulatory sequences or (*ii* and *iii*) regulation of one or more genes directly or at some point downstream in the transcriptional regulatory network.

(**Figure 3*b*,*i***), which could act in *cis* or in *trans* to effect differential expression (24).

One plausible theory is that distant regulatory QTL represent polymorphisms in transcription factors that influence the expression of genes serving as their targets. Although this scenario may explain some distant regulatory QTL (e.g., 11), overall there appears to be no enrichment for transcription factors among genes containing distant regulatory polymorphisms (29, 127). The presence of regulatory hotspots, loci that affect the expression levels of a large number of distant genes, has been noted in many studies (11, 28, 29, 69, 93, 94; for a discussion of these observations, see the section below on The Evolution of Gene Expression, below). Finally, it has been suggested that distant regulatory QTL have smaller effect sizes and are less replicable than local regulatory QTL (81, 84), although it remains unclear whether their replicability is a function only of effect size, or whether it is due to greater environmental/tissue specificity of distant regulatory QTL (84, 101).

Copy Number Variation

As copy number variants (CNVs) are simply genomic lesions of larger size than more commonly explored sources of DNA sequence variation, their mechanisms of action can be as diverse as those described above. Due to their size, CNVs are perhaps uniquely able to simultaneously contribute to regulatory variation across genes that are not necessarily functionally related but are found within the same genomic region. In the single study that has addressed the importance of CNVs to expression variation on a genomic scale, large-scale (>100 kb) CNVs were associated with 10%–25% as many gene expression traits as were mapped using SNPs (105). However, it is difficult to directly compare the relative contributions of SNPs and CNVs to transcriptional variation, as the impact of smaller CNVs has yet to be rigorously addressed. Another interesting observation to emerge from Stranger et al. (105) is that CNVs associated with transcript levels tend to lie close to the corresponding gene, suggesting that CNVs often affect transcript abundance by disrupting the gene itself or nearby regulatory regions (105). In one example, a CNV containing a duplication of the gene *Fgfbp3* in the mouse strain C57BL6/J alters expression of the gene in the spleen but not in the brain, suggesting the disruption of a brain-specific regulatory sequence in the duplicated copy (117).

HUMAN EXPRESSION GENETICS

Although large-scale studies of inherited gene expression variation have been performed in many species, the majority of analyses have focused on humans (15, 16, 28–30, 40, 68, 69, 73, 94, 100, 104–106, 112). **Table 1** summarizes the salient details of the 15 genetical genomics studies performed in humans to date. While the number of independent studies of inherited expression variation in humans seems to present an obvious opportunity for comparison, in practice this is exceedingly difficult. For example, study-specific differences exist in the tissue that expression levels were derived from, the technology platform used to measure transcript abundance, the statistical methods used to map expression QTL, the threshold for declaring a linkage or association test significant, and the operational definitions of local and distant expression QTL (**Table 1**). In addition, each study likely suffers from low statistical power, as the sample sizes are relatively modest given the large number of phenotypes analyzed; thus, it is not surprising that the overlap among significant expression QTL is generally low across studies (21, 38, 40). Despite these difficulties, several important issues emerge from the details of **Table 1**.

First, the majority of human analyses have been done on either the same set, or subset, of transformed B lymphoblastoid cell lines (B-LCLs) derived from individuals studied in the International HapMap Project (46). This somewhat curious detail is principally due to the ready availability of the samples and the pre-existing dense SNP genotype data for these individuals obtained in the course of the HapMap project. However, the extent to which patterns of gene expression in the B-LCLs recapitulate that of untransformed cells and other tissues is not known, and some observations, such as aberrant patterns of methylation induced by transformation (42), warrant caution. Ideally, future studies will focus on identifying expression QTL in nontransformed cells from a more diverse spectrum of tissues, building upon recent studies performed in blood (30, 40), liver (94), adipose (30), and the cerebral cortex (73).

Second, the fraction of expression QTL that are local or distant varies considerably across studies (**Table 1**). For instance, Goring et al. (40) estimate that 99% of their expression QTL are local, whereas Duan et al. (29) found that only ~5% of their expression QTL were local. These discrepancies also exist among the set of HapMap B-LCL studies, where the fraction of local expression QTL ranges from ~5% to nearly 70% or greater (**Table 1**). These large inconsistencies are difficult to explain, and the fact that they are observed in studies using the same samples and genotype data argues against

Table 1 Summary of genetical genomics studies in humans[a]

Reference	Samples	Tissue	Study design	Technology platform	Sample size	Genes examined	Number of gene expression QTL	Fraction of local QTL	Number of hotspots
Monks et al. (68)	CEPH	B-LCLs	Linkage	Agilent 60-mer	167	2430	55	24%	0
Morley et al. (69)	CEPH	B-LCLs	Linkage	Affymetrix Human Genome Focus	94	3554	147	20%	2
Cheung et al. (15)	CEU	B-LCLs	Association	Affymetrix Human Genome Focus	57	374[b]	65	-	-
Stranger et al. (104)	CEU	B-LCLs	Association	Illumina 50-mer	60	374	10-43	77-100%	-
Spielman et al. (100)	ASN + CEU	B-LCLs	Association	Affymetrix Human Genome Focus	142	1097	~90–100 per pop.	10-26%	-
Stranger et al. (105)	ASN + CEU + YRI	B-LCLs	Association	Illumina Human WG-6	210	14,072	~300–400 per pop. (SNP) ~50–100 per pop. (CNV)	-	-
Stranger et al. (106)	ASN + CEU + YRI	B-LCLs	Association	Illumina Human WG-6	270	13,643	~300–400 per pop.	-	-
Dixon et al. (28)	Britons	B-LCLs	Association	Affymetrix U133 Plus 2.0	308	14,819	2546	69%	3
Goring et al. (40)	Mexican Americans	Blood (lymphocyte)	Linkage	Illumina Human WG-6	1240	18,519	~1360	99%	0
Myers et al. (73)	Europeans	Cerebral cortex	Association	Illumina HumanRefseq-8	193	14,078	373	27%	-
Duan et al. (29)	CEU + YRI	B-LCLs	Association	Affymetrix Human Exon 1.0 ST	176	12,747	~1100–2800 per pop.	~5%	14-38
Veyrieras et al. (112)	ASN + CEU + YRI	B-LCLs	Association	Reanalysis of Stranger et al. 2007	210	11,446	1586	-	-
Emilsson et al. (30)	Icelanders	Blood + adipose	Linkage/ Association	Agilent 60-mer	670-1000 (linkage) 150 (association)	20,877	~1500–2600 (linkage) ~2800–3400 (association)	98%	0-3
Schadt et al. (94)	Caucasians	Liver	Association	Agilent 60-mer	427	34,266	3517	87%	23
Choy et al. (16)	ASN + CEU + YRI	B-LCLs	Association	Affymetrix U133A/ Illumina WG-6	198	1,000	~100–200	-	-

[a]Technology platform column lists the platform used to obtain gene expression measurements. Genes examined column indicates the number of expression traits to which correlation of molecular markers was tested. Number of expression QTL and number of hotspots are listed as reported in the paper, at the primary significance level that the authors used for carrying out analyses and interpreting results. Ranges provide approximate results for cases where authors analyzed populations separately or presented results using several methods of analysis. For studies where no fraction of local expression QTL is indicated, only local associations were tested. Studies missing a value in the final column either tested only local associations or did not mention hotspots/master regulators in the text. For studies that reported only the number of transcript-SNP associations, without combining SNPs in close proximity or SNPs showing high LD, we combined SNPs significantly associated with the same transcript within 4 Mb of each other into a single QTL. For one case where we had to retrieve microarray probe information to classify QTL location, local expression QTL were defined as QTL falling within 5 Mb of the location of a probeset. If results were not explicitly stated in the paper, we calculated table values from the authors' supplementary data following criteria specified in their paper. Abbreviations are as follows: CEPH = Centre d'Etude du Polymorphisme Humain; ASN = HapMap samples from Beijing, China + Tokyo, Japan; CEU = HapMap samples from US residents with northern and western European ancestry; YRI = HapMap samples from Yoruba people of Ibadan, Nigeria; B-LCLs = B-lymphoblastoid cell lines; CNV = copy number variant.
[b]Transcripts tested for local association had prior evidence for local linkage.

Confounding variables: two or more variables whose effects cannot be distinguished from one another. The unrecognized presence of confounding variables can lead to erroneous inferences about cause and effect

Expression heterogeneity: patterns of gene expression variation due to variables that are unknown, unmeasured, or too complicated to model

explanations related simply to statistical power. Given that the studies with the largest sample sizes (30, 40, 94) would have relatively high statistical power to detect distant expression QTL, yet predominantly find local expression QTL (**Table 1**), we suggest that the evidence for pervasive distant expression QTL of moderate to large effect size in humans remains tenuous.

Third, and related to the point above, the evidence for distant regulatory hotspots in humans is weak. The majority of studies have found either a very small number (<3) of hotspots with robust statistical support or none at all (**Table 1**). As we discuss below, there are both technical and evolutionary reasons to suspect that the number of regulatory hotspots due to common genetic variation in natural populations will be small.

In summary, the large number of expression QTL studies in humans has shown that regulatory loci are abundant and can be mapped to specific positions in the genome. A general pattern is beginning to emerge such that the majority of heritable transcriptional variation appears to be due to the combination of a relatively modest number of distant loci, perhaps a few with widespread transcriptional effects, and a much larger number of local regulatory alleles (30, 94). However, the quantitative details vary considerably across studies, with the fraction of expression QTL that are local or distant and the existence of regulatory hotspots being the most conspicuous differences. Although biological factors, such as distinct genetic architectures of transcriptional variation between tissues, may partly account for these discrepancies, additional nonbiological factors related to study design might also contribute to the heterogeneous results across studies, as we discuss in the following section.

EXPERIMENTAL DESIGN AND EXPRESSION HETEROGENEITY

The goal of all gene expression studies is to make biologically meaningful inferences about transcriptional variation, which requires careful attention to experimental study design. However, many factors not of primary interest can contribute to transcriptional variation (17, 52). For instance, there are many steps in a typical microarray experiment, such as isolation and labeling of RNA, hybridization conditions, and time of sample processing, which are well-known sources of technical variation (2, 17, 52, 125). In addition to technical variation, there are countless additional variables that also induce expression variation, such as sex, age, genetic and environmental heterogeneity (34, 63), the time of day a sample is collected (120), and passage number or other confounding variables present in cell lines (16, 126).

There is a substantial risk of drawing incorrect inferences about patterns of gene expression variation if confounding variables and expression heterogeneity (63) are not properly taken into account in the design and analysis of microarray experiments. The most recognizable problem results from batch effects, in which samples are processed in different groups during one or more steps between sample collection and data acquisition. Obviously, if the primary variable of interest (e.g., presence or absence of disease) is confounded with batch (i.e., all affected individuals are processed in one batch and all unaffected individuals are processed in a second batch), differences in gene expression levels may be observable that have nothing to do with the variable of interest. Even more insidious are cases where potential confounding variables that influence gene expression are either unknown or not measured.

To illustrate these problems, consider the hypothetical example in **Figure 4**, which demonstrates how batch effects and confounding variables can complicate inferences of expression variation. Specifically, in this example, gene expression levels are compared between five individuals with a disease and five individuals without a disease. In addition to disease status, several additional covariates that are known (sex and age) and unknown (use of a nonprescription drug) affect expression levels. If these samples are processed in two batches, each consisting of only affected or unaffected individuals, technical variation among

batches can lead to spurious inferences of differential expression (**Figure 4*a***). Furthermore, erroneous inferences of differential expression can also occur if the effects of confounding variables are not properly taken into account (**Figure 4*a***).

The most straightforward way to mitigate batch effects and confounding variables is through randomized study designs, which have been discussed in detail elsewhere (2, 17, 52, 125). In our hypothetical example, we could randomly allocate affected and unaffected individuals between the two batches, such that perhaps three randomly selected affected individuals and two randomly selected unaffected individuals are processed in batch one and the remaining samples are processed in batch two. Even better, the randomization could be performed so that known covariates are balanced among batches (**Figure 4*b***), allowing meaningful estimates of their contribution and subsequent adjustment to expression variation. Note, however, that expression heterogeneity can persist even in the most carefully designed studies that employ randomization (63). For example, randomization would not ameliorate the expression variation mediated by the unknown drug by sex interaction in **Figure 4**.

In short, it is impractical, if not impossible, to design a "perfect" large-scale study of gene expression, or any other high-dimensional molecular phenotype, that explicitly controls for all potential confounding variables. In studies of hundreds or thousands of individuals, it is not feasible to collect, culture, and process all samples simultaneously, or measure every potential confounding variable contributing to expression heterogeneity. Nonetheless, careful attention to study design, including randomization in all experimental aspects, is critical (2, 125). Even so, expression heterogeneity from unmeasured confounding factors will likely persist (63). Fortunately, new statistical methods have been developed to identify and correct for the unwanted effects of unmeasured confounding variables in studies of gene expression (51, 63), which should become integral components in analysis pipelines.

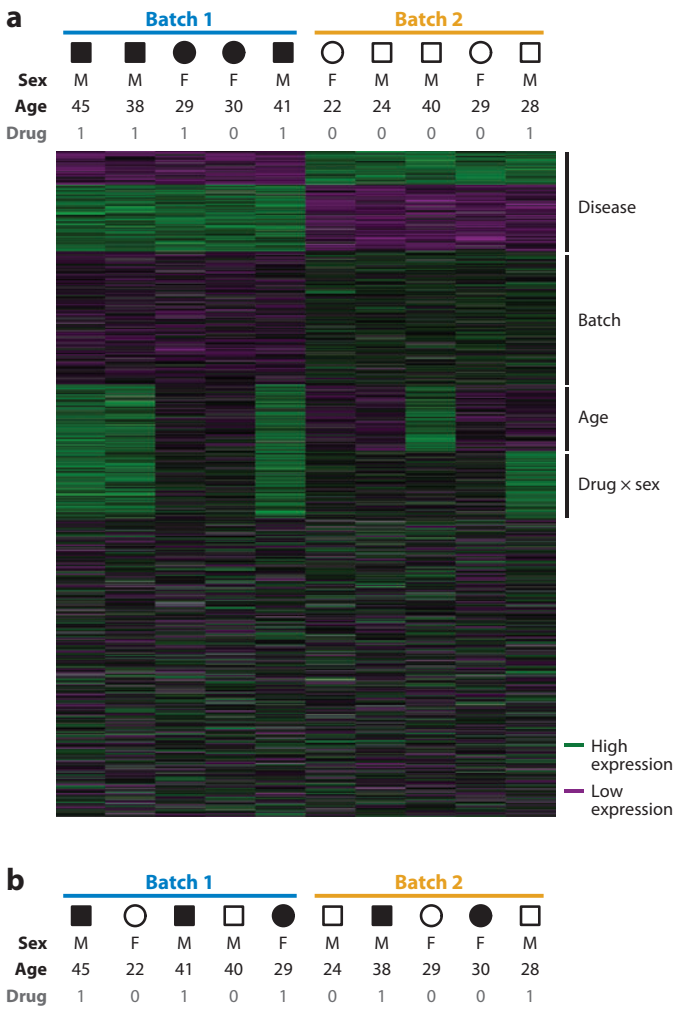


Figure 4

Study design is critical for making meaningful inferences of gene expression variation. (*a*) Hypothetical example of how batch effects and confounding variables can lead to incorrect inferences about patterns of gene expression variation. Expression levels are shown for 1000 genes (*rows*) in ten individuals (*columns*). Cases and controls are indicated by black and white boxes or circles, respectively. Colors indicate the relative transcript abundance ranging from low (*magenta*) to high (*green*). Known confounding variables for each individual (sex and age) are indicated at the top of the figure, as well as an unknown confounding variable (whether the individual is taking an over the counter drug) shown in gray. Note that because of the study design, in which all affected individuals are processed in a single batch, 500 differentially expressed genes are identified, only 100 of which are actually related to the disease (the source of each differentially expressed gene is indicated by the vertical bars along the right of the figure). (*b*) A randomized study design, in which affected and unaffected individuals are not processed simultaneously and known covariates are balanced among batches, which can help ameliorate expression heterogeneity and erroneous inferences of gene expression variation (see text for details).

THE EVOLUTION OF GENE EXPRESSION

An important consequence of the recent interest in heritable gene expression variation is that it has paved the way for systematic analyses into the tempo and mode of gene expression evolution. The impetus for such studies can be traced back to at least three decades ago, with the provocative assertion that gene expression changes might underlie many of the phenotypic differences between humans and chimpanzees (55), an argument primarily based on the smaller than expected divergence in protein sequences between the species. Although the relative contribution of changes in gene expression levels versus changes in protein sequence to phenotypic change remains controversial (43, 122), it is clear that regulatory variation with pleiotropic effects provides a putative mechanism through which phenotypic diversity might be rapidly generated and tested by selection with minimal genetic change (121). Striking empirical evidence for this hypothesis is evident in induced regulatory mutations in the *Hox* gene cluster in *Drosophila* that have effects mimicking arthropod body plan diversity (13), and selected regulatory mutations fixed during the domestication of maize (18, 115).

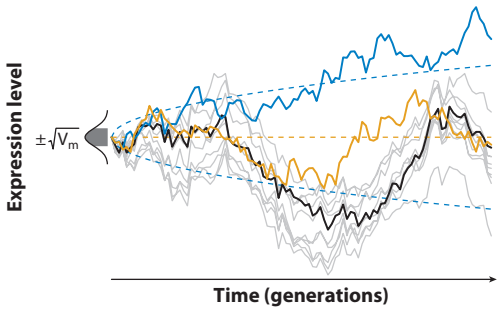


Figure 5

Modeling gene expression evolution by diffusion processes. Horizontal and vertical axes represent time in generations and gene expression level, respectively. The solid blue curve represents a neutral model of gene expression evolution in which transcript levels change according to a Brownian motion model (31). At each generation, the expression level changes randomly from its current value by an amount drawn from a normal distribution with mean 0 and variance V_m, where V_m is mutational variance. The dashed blue curves show the expected magnitude of the deviation as a function of time under Brownian motion. The orange curve is a realization of the Ornstein-Uhlenbeck process, a model in which natural selection constrains gene expression levels to fluctuate around an optimum (31). The black and gray curves represent realizations of Brownian motion for 10 traits with underlying genetic correlation. Note that correlation among the trajectories could lead to apparent evidence in favor of a neutral model or positive selection model with accelerated change (near the time midpoint) or apparent evidence in favor of purifying selection to maintain stability (near the time endpoint) if significance is assessed across genes.

In general, two study designs have been used to make inferences about gene expression evolution. In the first approach, gene expression levels are measured in related species spanning a range of divergence times, and the observed expression differences are modeled as a function of time using diffusion processes (37). Gene expression levels that appear to diverge either more slowly or rapidly relative to neutral expectations are inferred to be targets of purifying or positive selection, respectively (**Figure 5**). Such analyses have led to adaptive (87), neutral (53, 54), and selective constraint-based (25, 44, 88) models to explain the evolution of transcript levels between species.

Though theoretically appealing, several technical difficulties have become apparent with this approach. For example, as can be seen in **Figure 5**, the variance among neutral and selectively constrained trajectories can be substantial, making it difficult to reject either underlying model. A typical solution to this problem is to average expression levels across genes in an attempt to better infer the dynamics of the underlying process (25, 53, 88). However, genetic correlation among gene expression levels, due for example to a shared transcriptional regulator, can complicate evolutionary analyses of quantitative characters (31). In the presence of strong underlying genetic correlations, averaging levels of expression divergence across genes only provides a more precise picture of a single realization of the evolutionary process, not a more precise picture of the evolutionary process (see **Figure 5**). Rather than averaging information over genes, it would be preferable to average information over more independent evolutionary processes for each gene by sampling more species. However, as increasingly divergent species are sampled, DNA sequence

differences interfere with hybridization on typical microarrays, leading to potential biases in gene expression measurements (36). This hybridization artifact causes gene expression trajectories to appear to diverge rapidly, with interspecies differences due to both expression changes and sequence differences.

Despite these difficulties, a consensus has emerged that, on average, gene expression profiles do not diverge rapidly enough to follow a neutral model (37). However, because the Brownian motion–based neutral expectation for gene expression change is an unbounded process, it may be impossible for cellular phenotypes to meet this expectation even in the absence of selection. Thus, it remains unclear whether the underlying basis for slow gene expression divergence is natural selection or simply biochemical limitations on the rate of mRNA transcription or stability that can be achieved within the cell (88). Even if purifying selection is invoked to explain deviations from Brownian motion, little is known about the way in which selection might act. For example, it has been suggested that evolution might proceed mainly by strong constraint on major regulatory genes (88). Alternatively, regulatory mutations in most genes may be subject to persistent weak selection (76, 114).

The second, complementary, study design for studying gene expression evolution leverages the information contained in expression QTL studies (90). Specifically, in cases where expression QTL are localized with high precision, regions harboring regulatory polymorphisms can be identified and studied. This allows the large set of well-developed tools for studying DNA sequence evolution to be applied to the problem of gene expression evolution. Furthermore, because the underlying expression QTL represent evolutionarily independent events, analysis of regulatory evolution at the level of DNA sequence variation may circumvent difficulties associated with genetic correlations among expression levels.

We recently applied this approach to investigate the evolutionary forces shaping patterns of genetic variation for 1206 *cis*-regulatory QTL identified in a cross between two divergent strains of *S. cerevisiae* (90). This analysis revealed that purifying selection against deleterious alleles is the dominant force governing *cis*-regulatory evolution in *S. cerevisiae*, and approximately 24% fewer genes show *cis*-acting expression variation relative to what would be expected if these expression changes were selectively neutral. In addition, we found that the average strength of selection acting on a typical regulatory mutation is rather weak (scaled selection coefficient of $\sim$2), implying that stochastic forces play a significant role in shaping patterns of *cis*-regulatory diversity. Under such a nearly neutral regime, deleterious *cis*-regulatory alleles can be present at appreciable frequencies in a population, and thus an interplay of forces, including changes in population size, LD among selected alleles, and epistatic selection (76), may have prominent roles in *cis*-regulatory evolution.

One strength of studying the evolution of gene expression levels at the level of DNA sequence variation is that it allows more detailed and mechanistic hypotheses of regulatory evolution to be explored. For example, the existence of major *trans*-regulatory hotspots in expression QTL studies is controversial. Recently, Breitling et al. (8) suggested that many apparent *trans*-regulatory hotspots could largely be explained as statistical artifacts. The observation in yeast that most individual expression changes are mildly deleterious suggests that the cumulative selective effects against major *trans*-acting QTL would be so strong that they would be rapidly eliminated from the population. Thus, from an evolutionary perspective, major *trans*-acting QTL should be rare. Indeed, closer inspection of the major *trans*-acting QTL hotspots in yeast suggests caution in interpreting their prevalence in natural populations. Specifically, many of the yeast *trans*-acting QTL hotspots are likely attributable to auxotrophic markers or alleles selected during the domestication of yeast to the lab (90). These loci were subjected to exceptionally strong positive selection, the magnitude of

which is unlikely to be commonly experienced in natural populations.

LOOKING AHEAD: THE ERA OF RNA SEQUENCING HAS BEGUN

Although microarrays have proven tremendously useful for the measurement of transcript levels on a genome-wide basis, it appears increasingly probable that this task will soon be subsumed by technologies that directly sequence entire transcriptomes. Such an approach, dubbed RNA-Seq (74), makes use of massively parallel DNA sequencing technologies to determine the sequences of cDNA fragments, which are then mapped back to a reference genome. Although this approach is still in its infancy, it has a number of attractive features that obviate some of the limitations of microarray methods. For example, microarrays rely on nucleic acid hybridization and measurement of the emission intensity of a fluorophore-labeled target, an inherently continuous signal. This constrains the dynamic range interrogated by microarrays, and makes accurate measurement of low-abundance transcripts particularly problematic. These limitations are ameliorated by RNA-Seq, which provides digital quantitation of transcript levels in the form of read counts mapping to the reference genome. In theory, precision of a digital expression readout using RNA-Seq should be limited only by coverage depth; indeed, it has been shown that cDNA standards spiked at known concentration show linearity across a dynamic range of five orders of magnitude (70). RNA-Seq does not require probe design steps or interrogate only a subset of the genome, although it does require a complete reference sequence to which reads can be mapped.

Furthermore, concerns about sequence polymorphism affecting hybridization and mimicking local expression QTL (3) are largely avoided by RNA-Seq. Finally, it has been speculated that the interplatform reproducibility of RNA-Seq will be superior to microarrays (98). This question has not been directly addressed, although there appears to be high technical reproducibility between RNA-Seq experiments within and between laboratories (67, 107) using the Illumina Genome Analyzer. However, it is important to emphasize that the issues of study design described above transcend technology platform and apply equally to additional methods for measuring gene expression levels such as RNA-Seq.

Although promising, a number of challenges exist in RNA-Seq experiments. For instance, the short read lengths and high error rates of base calls may make it difficult to map sequence reads back to a genome, especially for large and complex mammalian genomes (70). Methods for analysis are still evolving, although there has been considerable progress toward developing fully probabilistic methods for mapping sequence reads to a genome (64). Finally, RNA-Seq suffers from one of the same limitations as microarrays: Highly similar sequences usually cannot be distinguished. Nevertheless, as next-generation sequencing technology matures, the compound effects of increasing read lengths and decreasing error rates will diminish the difficulty of mapping reads correctly and facilitate analysis of the deluge of data.

CONCLUSIONS

Genome-wide technologies for measuring gene expression levels and interrogating DNA sequence variation have allowed the genetic underpinnings of transcriptional variation to begin to be elucidated in a wide variety of organisms. These studies have already provided novel insights into the genetic architecture of gene expression variation and the nature of regulatory alleles, but many important questions remain unanswered. Among the issues that remain enigmatic are the contribution of rare alleles to transcriptional variation, the effects of environmental variation on gene expression, and the nature and prevalence of tissue-specific regulatory variation. A deeper understanding of gene expression evolution may come from studies that investigate the evolution of transcriptional variation at the DNA sequence level. Such studies can serve as a foundation

for models of "functional evolution," which describe the evolutionary trajectories of alleles that impact other functional genomics phenotypes, such as gene repertoire and copy number (4, 26, 48, 119), protein levels (20, 22, 32, 60), and metabolic profiles (26, 82, 83), that vary in natural populations.

Ultimately, the purpose of discovering gene expression QTL is to better understand the mechanistic details of how DNA sequence variation perturbs dynamic transcriptional networks, and how such regulatory networks contribute to disease susceptibility, phenotypic diversity, and evolutionary change. To this end, it will be critical to move beyond gene expression QTL, which are descriptions of regulatory variants defined in purely statistical terms, to the identification of the actual regulatory alleles. This is clearly a daunting challenge, even in model organisms, and will require the synthesis of experimental approaches for characterizing gene regulation and methodological and technical advances that will allow the extraction of meaningful information from increasingly complex and heterogeneous genome-wide datasets.

SUMMARY POINTS

1. A significant fraction of transcriptional variation is heritable.
2. The identification of gene expression QTL has provided detailed insights into the genetic architecture of transcriptional variation, often revealing unexpected complexity, and the nature of regulatory alleles.
3. Study design is critical for making biologically meaningful inferences of gene expression variation.
4. Transcriptome sequencing (RNA-Seq) is poised to transform studies of gene expression variation.

FUTURE ISSUES

1. How much of gene expression variation is functionally neutral and how much contributes to inherited variation in protein levels and organismal phenotypes?
2. What are the best strategies for identifying specific regulatory alleles underlying gene expression QTL?
3. How stable are patterns and characteristics of heritable gene expression variation across tissue types and environmental perturbations?
4. What is the relative contribution of rare and common alleles to transcriptional variation?
5. New methods are needed to fully exploit increasingly complex and heterogeneous sources of data, such as complete genome sequences and additional functional genomics phenotypes, in order to understand the topology, function, and evolution of regulatory networks.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported in part by a Sloan Fellowship in Computational Biology and NIH grant 5R01GM078105 to J.M.A., as well as an NIH training grant to the University of Washington (D.A.S.). We apologize to all of our colleagues whose work could not be cited because of constraints on space.

LITERATURE CITED

1. Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ. 2004. *Bmp4* and morphological variation of beaks in Darwin's finches. *Science* 305:1462–65
2. Akey JM, Biswas S, Leek JT, Storey JD. 2007. On the design and analysis of gene expression studies in human populations. *Nat. Genet.* 39:807–8
3. Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC. 2007. Sequence polymorphisms cause many false *cis* eQTLs. *PLoS ONE* 2:e622
4. Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF. 2007. Genome dynamics in a natural archaeal population. *Proc. Natl. Acad. Sci. USA* 104:1883–88
5. Bjornsson HT, Albert TJ, Ladd-Acosta CM, Green RD, Rongione MA, et al. 2008. SNP-specific array-based allele-specific expression analysis. *Genome Res.* 18:771–79
6. Bray NJ, Buckland PR, Owen MJ, O'Donovan MC. 2003. *Cis*-acting variation in the expression of a high proportion of genes in human brain. *Hum. Genet.* 113:149–53
7. Bray NJ, Buckland PR, Williams NM, Williams HJ, Norton N, et al. 2003. A haplotype implicated in schizophrenia susceptibility is associated with reduced *COMT* expression in human brain. *Am. J. Hum. Genet.* 73:152–61
8. Breitling R, Li Y, Tesson BM, Fu J, Wu C, et al. 2008. Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* 4:e1000232
9. Brem RB, Kruglyak L. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA* 102:1572–77
10. Brem RB, Storey JD, Whittle J, Kruglyak L. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436:701–3
11. **Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752–55**
12. 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
13. Carroll SB. 1995. Homeotic genes and the evolution of arthropods and chordates. *Nature* 376:479–85
14. **Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, et al. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33:422–25**
15. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437:1365–69
16. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, et al. 2008. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* 4:e1000287
17. Churchill GA. 2002. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32:490–95
18. Clark RM, Wagler TN, Quijada P, Doebley J. 2006. A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. *Nat. Genet.* 38:594–97
19. Cowles CR, Hirschhorn JN, Altshuler D, Lander ES. 2002. Detection of regulatory variation in mouse genomes. *Nat. Genet.* 32:432–37
20. Damerval C, Maurice A, Josse JM, de Vienne D. 1994. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. *Genetics* 137:289–301
21. de Koning DJ, Haley CS. 2005. Genetical genomics in humans and model organisms. *Trends Genet.* 21:377–81
22. de Vienne D, Maurice A, Josse JM, Leonardi A, Damerval C. 1994. Mapping factors controlling genetic expression. *Cell. Mol. Biol.* 40:29–39

The first study to map the genetic determinants of global gene expression variation.

The first study to evaluate natural variation in gene expression in humans.

23. DeCook R, Lall S, Nettleton D, Howell SH. 2006. Genetic regulation of gene expression during shoot development in *Arabidopsis*. *Genetics* 172:1155–64
24. Dekker J. 2008. Gene regulation in the third dimension. *Science* 319:1793–94
25. Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat. Genet.* 37:544–48
26. Díaz-Mejía JJ, Pérez-Rueda E, Segovia L. 2007. A network perspective on the evolution of metabolism by gene duplication. *Genome Biol.* 8:R26
27. Ding C, Cantor CR. 2003. A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. *Proc. Natl. Acad. Sci. USA* 100:3059–64
28. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. 2007. A genome-wide association study of global gene expression. *Nat. Genet.* 39:1202–7
29. Duan S, Huang RS, Zhang W, Bleibel WK, Roe CA, et al. 2008. Genetic architecture of transcript-level variation in humans. *Am. J. Hum. Genet.* 82:1101–13
30. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. 2008. Genetics of gene expression and its effect on disease. *Nature* 452:423–28
31. Felsenstein J. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer
32. Fu N, Drinnenberg I, Kelso J, Wu JR, Paabo S, et al. 2007. Comparison of protein and mRNA expression evolution in humans and chimpanzees. *PLoS ONE* 14:e216
33. Genissel A, McIntyre LM, Wayne ML, Nuzhdin SV. 2008. *Cis* and *trans* regulatory effects contribute to natural variation in transcriptome of *Drosophila melanogaster*. *Mol. Biol. Evol.* 25:101–10
34. Gibson G. 2008. The environmental contribution to gene expression profiles. *Nat. Rev. Genet.* 9:575–81
35. Gibson G, Weir B. 2005. The quantitative genetics of transcription. *Trends Genet.* 21:616–23
36. Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res.* 15:674–80
37. Gilad Y, Oshlack A, Rifkin SA. 2006. Natural selection on gene expression. *Trends Genet.* 22:456–61
38. Gilad Y, Rifkin SA, Pritchard JK. 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* 24:408–15
39. Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. 2005. Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* 433:481–87
40. Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. 2007. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* 39:1208–16
41. Grady WM, Willis J, Guilford PJ, Dunbier AK, Toro TT, et al. 2000. Methylation of the *CDH1* promoter as the second genetic hit in hereditary diffuse gastric cancer. *Nat. Genet.* 26:16–17
42. Hannula K, Lipsanen-Nyman M, Scherer SW, Holmberg C, Höglund P, Kere J. 2001. Maternal and paternal chromosomes 7 show differential methylation of many genes in lymphoblast DNA. *Genomics* 73:1–9
43. Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61:995–1016
44. Hsieh WP, Chu TM, Wolfinger RD, Gibson G. 2003. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* 165:747–57
45. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, et al. 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* 37:243–53
46. The International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426:789–96
47. Jansen RC, Nap JP. 2001. Genetical genomics: the added value from segregation. *Trends Genet.* 17:388–91
48. Johnson PL, Slatkin M. 2006. Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res.* 16:1320–27
49. Jones TR, Cole MD. 1987. Rapid cytoplasmic turnover of c-*myc* mRNA: requirement of the 3′ untranslated sequences. *Mol. Cell. Biol.* 7:4513–21
50. Kadota M, Yang HH, Hu N, Wang C, Hu Y, et al. 2007. Allele-specific chromatin immunoprecipitation studies show genetic influence on chromatin state in human genome. *PLoS Genet.* 3:e81
51. Kang HM, Ye C, Eskin E. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180:1909–25

An impressive large-scale study of gene expression QTL in human blood and adipose tissues.

Outlined the genetical genomics study design.

52. Kerr MK, Churchill GA. 2001. Statistical design and the analysis of gene expression microarray data. *Genet. Res.* 77:123–28
53. Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, et al. 2004. A neutral model of transcriptome evolution. *PLoS Biol.* 2:e132
54. Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, et al. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309:1850–54
55. King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–16
56. Kirst M, Myburg AA, De León JP, Kirst ME, Scott J, Sederoff R. 2004. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of *Eucalyptus*. *Plant Physiol.* 135:2368–78
57. Knight JC. 2005. Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* 83:97–109
58. Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP. 2003. In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat. Genet.* 33:469–75
59. Kleinjan DA, van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* 76:8–32
60. Klose J, Nock C, Herrmann M, Stühler K, Marcus K, et al. 2002. Genetic analysis of the mouse brain proteome. *Nat. Genet.* 30:385–93
61. Kochi Y, Yamada R, Suzuki A, Harley JB, Shirasawa S, et al. 2005. A functional variant in *FCRL3*, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nat. Genet.* 37:478–85
62. Landry CR, Oh J, Hartl DL, Cavalieri D. 2006. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene* 366:343–51

Describes a method for identifying and correcting the effects of confounding variables contributing to expression heterogeneity.

63. Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3:1724–35

64. Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18:1851–58
65. Li Y, Álvarez OA, Gutteling EW, Tijsterman M, Fu J, et al. 2006. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet.* 2:e222
66. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, et al. 2003. Allelic variation in gene expression is common in the human genome. *Genome Res.* 13:1855–62
67. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18:1509–17
68. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, et al. 2004. Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* 75:1094–105
69. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, et al. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature* 430:743–47

One of the first attempts to extend the RNA-Seq technique to a mammalian genome.

70. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5:621–28

71. Mozhui K, Ciobanu DC, Schikorski T, Wang X, Lu L, Williams RW. 2008. Dissection of a QTL hotspot on mouse distal chromosome 1 that modulates neurobehavioral phenotypes and gene expression. *PLoS Genet.* 4:e1000260
72. Muhlrad D, Parker R. 1992. Mutations affecting stability and deadenylation of the yeast *MFA2* transcript. *Genes Dev.* 6:2100–11
73. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, et al. 2007. A survey of genetic human cortical gene expression. *Nat. Genet.* 39:1494–99
74. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–49
75. Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol. Biol. Evol.* 21:1308–17
76. Ohta T. 2002. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. USA* 99:16134–37
77. Oleksiak MF, Churchill GA, Crawford DL. 2002. Variation in gene expression within and among natural populations. *Nat. Genet.* 32:261–66

78. Pandey NB, Marzluff WF. 1987. The stem-loop structure at the 3′ end of histone mRNA is necessary and sufficient for regulation of histone mRNA stability. *Mol. Cell. Biol.* 7:4557–59
79. Pant PV, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. 2006. Analysis of allelic differential expression in human white blood cells. *Genome Res.* 16:331–39
80. Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, et al. 2004. A survey of genetic and epigenetic variation affecting human gene expression. *Physiol. Genomics* 16:184–93
81. Peirce JL, Li H, Wang J, Manly KF, Hitzemann RJ, et al. 2006. How replicable are mRNA expression QTL? *Mamm. Genome* 17:643–56
82. Perlstein EO, Ruderfer DM, Ramachandran G, Haggarty SJ, Kruglyak L, Schreiber SL. 2006. Revealing complex traits with small molecules and naturally recombinant yeast strains. *Chem. Biol.* 13:319–27
83. Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L. 2007. Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nat. Genet.* 39:496–502
84. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, et al. 2006. Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet.* 2:e172
85. Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, et al. 2006. Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. *Nature* 440:1050–53
86. Rieseberg LH, Archer MA, Wayne RK. 1999. Transgressive segregation, adaptation and speciation. *Heredity* 83:363–72
87. Rifkin SA, Kim J, White KP. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* 33:138–44
88. Rifkin SA, Houle D, Kim J, White KP. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* 438:220–23
89. Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. *Nat. Rev. Genet.* 7:862–72
90. Ronald J, Akey JM. 2007. The evolution of gene expression QTL in *Saccharomyces cerevisiae*. *PLoS ONE* 2:e678
91. Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L. 2005. Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res.* 15:284–91
92. Ronald J, Brem RB, Whittle J, Kruglyak L. 2005. Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet.* 1:e25
93. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, et al. 2003. Genetics of gene expression surveyed in maize, mouse, and man. *Nature* 422:297–302
94. Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6:e107
95. Serre D, Gurd S, Ge B, Sladek R, Sinnett D, et al. 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. *PLoS Genet.* 4:e1000006
96. Shapiro MD, Bell MA, Kingsley DM. 2006. Parallel genetic origins of pelvic reduction in vertebrates. *Proc. Natl. Acad. Sci. USA* 103:13753–58
97. Shaw G, Kamen R. 1986. A conserved AU sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell* 46:659–67
98. Shendure J. 2008. The beginning of the end for microarrays? *Nat. Methods* 5:585–87
99. Shin HD, Winkler C, Stephens JC, Bream J, Young H, et al. 2000. Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. *Proc. Natl. Acad. Sci. USA* 97:14467–72
100. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. 2007. Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* 39:226–31
101. Smith EN, Kruglyak L. 2008. Gene-environment interaction in yeast gene expression. *PLoS Biol.* 6:e83
102. Storey JD, Akey JM, Kruglyak L. 2005. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.* 3:e267
103. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. 2007. Gene-expression variation within and among human populations. *Am. J. Hum. Genet.* 80:502–9
104. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. 2005. Genome-wide associations of gene expression variation in humans. *PLoS Genet.* 1:e78

One of the largest genome-wide analyses of expression QTL in over 400 human liver samples.

105. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315:848–53
106. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. 2007. Population genomics of human gene expression. *Nat. Genet.* 39:1217–24
107. 't Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, et al. 2008. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.* 36:e141
108. Tao H, Cox DR, Frazer KA. 2006. Allele-specific *KRT1* expression is a complex trait. *PLoS Genet.* 2:e93
109. Townsend JP, Cavalieri D, Hartl DL. 2003. Population genetic variation in genome-wide gene expression. *Mol. Biol. Evol.* 20:955–63
110. van der Zee J, Le Ber I, Maurer-Stroh S, Engelborghs S, Gijselinck I, et al. 2007. Mutations other than null mutations producing a pathogenic loss of progranulin in frontotemporal dementia. *Hum. Mutat.* 28:416
111. Van Laere AS, Nguyen M, Braunschweig M, Nezer C, Collette C, et al. 2003. A regulatory mutation in *IGF2* causes a major QTL effect on muscle growth in the pig. *Nature* 425:832–36
112. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* 4:e1000214
113. Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* 9:255–66
114. Wagner A. 2005. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* 22:1365–74
115. Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, et al. 2005. The origin of the naked grains of maize. *Nature* 436:714–19
116. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. 2002. Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. USA* 99:5860–65
117. Watkins-Chow DE, Pavan WJ. 2008. Genomic copy number and expression variation within the C57BL/6J inbred mouse strain. *Genome Res.* 18:60–66
118. Watts JA, Morley M, Burdick JT, Fiori JL, Ewens WJ, et al. 2002. Gene expression phenotype in heterozygous carriers of ataxia telangiectasia. *Am. J. Hum. Genet.* 71:791–800
119. Whitaker RJ, Banfield JF. 2006. Population genomics in natural microbial communities. *Trends Ecol. Evol.* 21:508–16
120. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, et al. 2003. Individuality and variation in gene expression patterns in human blood. *Proc. Natl. Acad. Sci. USA* 100:1896–901
121. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20:1377–419
122. Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat. Rev. Genet.* 8:206–16
123. Yan H, Dobbie Z, Gruber SB, Markowitz S, Romans K, et al. 2002. Small changes in expression affect predisposition to tumorigenesis. *Nat. Genet.* 30:25–26

The first study to systematically evaluate allelic variation in gene expression in humans.

124. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. 2002. Allelic variation in human gene expression. *Science* 297:1143

125. Yang H, Harrington CA, Vartanian K, Coldren CD, Hall R, Churchill GA. 2008. Randomization in laboratory procedure is key to obtaining reproducible microarray results. *PLoS ONE* 3:e3724
126. Yu H, Cook TJ, Sinko PJ. 1997. Evidence for diminished functional expression of intestinal transporters in Caco-2 cell monolayers at high passages. *Pharm. Res.* 14:757–62
127. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, et al. 2003. *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.* 35:57–64

Appendix B

SUPPLEMENT TO CHAPTER 2

This appendix contains material published in [62].

B.1 Robustness of observations to imputed data

The genome sequences I used varied in sequence coverage, with most between 1-4X coverage [1]. The complete assemblies of the *Saccharomyces* Genome Resequencing Project strains include nucleotides that have been imputed by taking into account phylogenetic relationships between strains, to correct likely sequencing errors and fill in missing data [1]. In my intron splice sequence dataset, an average of 32% of the sequence is imputed per site. There was considerable variation between strains in the amount of imputed sequence in the dataset, from $< 3\%$ (the reference genome, RM11-1A, SK1, W303, Y55, and YJM789) to 67% (YJM981). Nucleotide diversity for the reduced set of genomes with $< 3\%$ imputed data was not significantly different compared to the complete set of 38 strains (0.00092 versus 0.00101; $P > 0.05$). Furthermore, in the reduced dataset with $< 3\%$ imputed sequence, we observed nine of the 21 splice sequence polymorphisms found in the complete dataset, with 10 of the 12 missing polymorphisms present at low frequencies ($< 10\%$) in the complete dataset. The dataset of genomes with high sequence coverage would be expected to lack a significant fraction of the polymorphisms present in the complete dataset, as the relatively divergent sake strains, Malaysian strains, and North American oak tree isolates [1] are not represented in the smaller dataset. I suggest that the imputation process had little bearing on my general observations for two reasons: (1) the majority of the polymorphisms not detected in the smaller dataset are likely missing because of the much smaller number of strains sampled rather than an incorrectly imputed base, and (2) the measure of nucleotide variation that I used for summarizing polymorphism (nucleotide diversity) is only marginally affected by low frequency polymorphisms [113]. As such, I used complete assemblies (including imputed

data) for all further analyses, and I expect my conclusions to be robust to the presence of imputed nucleotides in the genome sequences.

B.2 Genic features of introns with polymorphic splice sequences

Since genes encoding ribosomal proteins display features (such as high mean levels of expression) differentiating them from non-ribosomal protein genes [133], I separated ribosomal and non-ribosomal intron-containing genes for all analyses. There was no significant difference in the proportion of genes with polymorphic intron splice sequences among ribosomal and non-ribosomal genes (Fisher's exact test; $p > 0.05$). I performed gene ontology searches [241] using the GO Term Finder in *Saccharomyces* Genome Database (`http://yeastgenome.org/`), specifying the set of all (non-)ribosomal intron-containing genes as the background set, and found no significant function, location, or biological process terms for genes containing polymorphic intron splice sequences. Genic GC content was not significantly different for genes with polymorphic splice sequences (t-test; $P > 0.05$). I calculated dN/dS for sequences from all 38 strains using `PAML` [108] as a proxy for the rate of protein evolution and the codon adaptation index (CAI) [121] as an estimate of the relative expression level of each gene, and found no significant difference in dN/dS or mean CAI between genes with/without polymorphic intron splice sequences (t-test; $p > 0.05$). Thus, the lack of heterogeneity in genic features associated with polymorphic splice sequences suggests that levels of functional constraint on intron splicing are similar across ribosomal genes or non-ribosomal genes.

B.3 Verification of simulations using theoretical model

I verified the results of my simulations for strong selection coefficients ($2N_e s > 10$) using a theoretical formula, originally derived by Sewall Wright, for the distribution of gene frequencies in a panmictic population under a two allele-model with reversible asymmetric mutation and one selectively favored allele:

$$f(x) = C \exp^{2N_e sx} x^{2N_e v-1}(1-x)^{2N_e u-1},$$

where x is the frequency of the allele which has selective advantage s, v is the mutation rate to the preferred allele, u is the mutation rate to the unpreferred allele, and C is a normalizing

constant adjusted so that the distribution sums to unity [117]. I used asymmetric mutation rates, with the mutation rate to the preferred allele one-third the mutation rate to the unpreferred allele. I obtained nucleotide diversity by calculating the binomial probability of selecting two individuals with different alleles integrated over the complete distribution of x:

$$\binom{2}{1} \frac{\int_0^1 \exp^{2N_e s x} x^{2N_e v}(1-x)^{2N_e u} dx}{\int_0^1 \exp^{2N_e s x} x^{2N_e v-1}(1-x)^{2N_e u-1} dx}$$

using the `R` software environment [111]. For purifying selection with the advantageous allele being favored at a level stronger than $2N_e s \approx 5$, the theoretical result calculated using this two-allele model provides excellent correspondence with results from simulations using the four-allele model (Figure B.1). The accuracy of the approximation derives from the fact that, in the four-allele model, the proportion of polymorphisms between two alleles that are both not selectively favored becomes vanishingly small as the selection coefficient rises above $2N_e s \approx 5$ and most individuals in the population possess the favored allele.

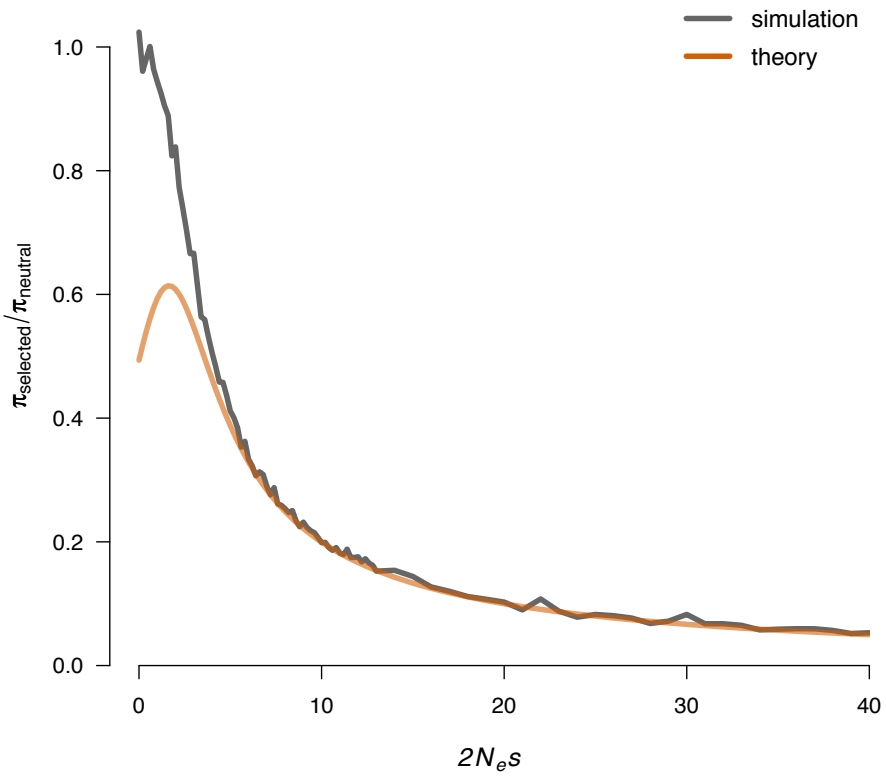


Figure B.1: Figure shows reduction in diversity at simulated selected site, relative to linked neutral site, as a function of the strength of selection. Both the two-allele and four-allele model are specified to include one selectively favored allele. Values plotted were obtained by simulation for the four-allele model, and calculated analytically for the two-allele model (see Section B.3 for details). The two-allele theoretical model provides an excellent approximation to the simulated four-allele model for selection coefficients larger than $2N_es \approx 5$. This occurs because, in the four-allele model, the proportion of polymorphisms between two alleles that are both not selectively favored becomes vanishingly small as the selection coefficient rises above $2N_es \approx 5$ and most individuals in the population possess the favored allele.

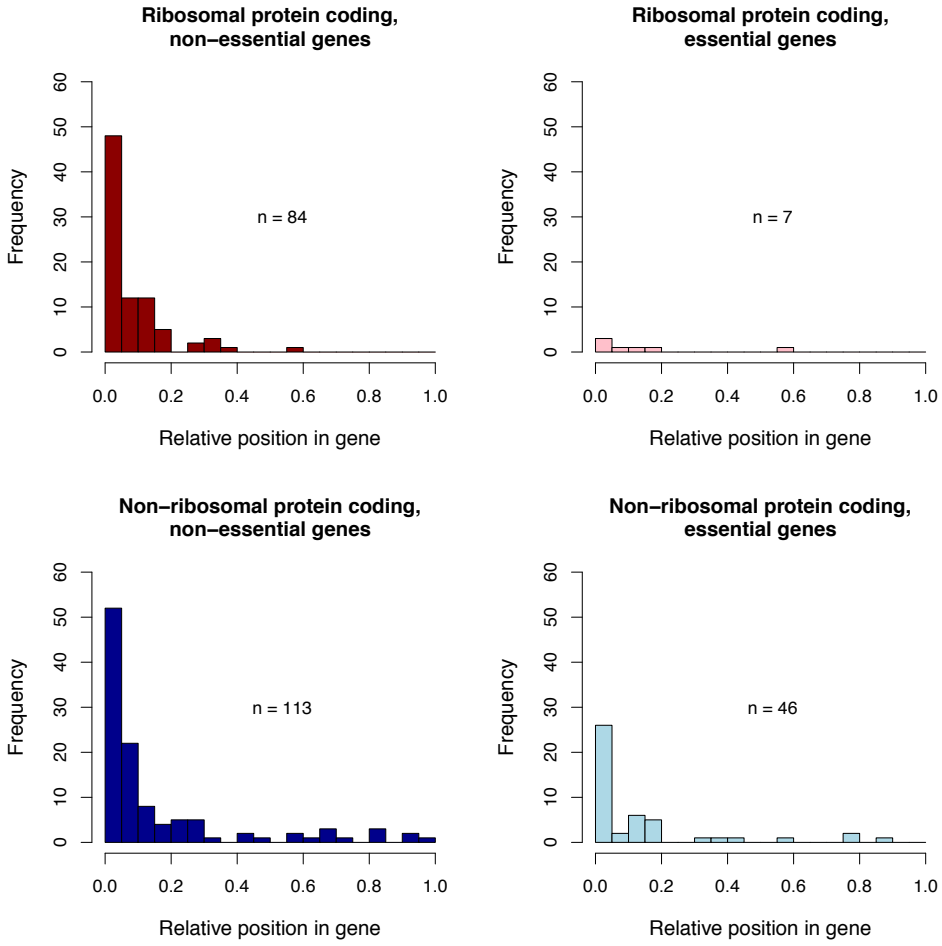


Figure B.2: Histograms showing the relative positions of introns within genes, divided according to whether genes are classified as essential or non-essential [120], and whether they code for ribosomal proteins. 5′ UTR introns and introns in genes not classified as essential/non-essential are omitted for clarity. Sample sizes of each class are shown in the center of each histogram.

Appendix C

SUPPLEMENT TO CHAPTER 3

This appendix contains material published in [138]. I gratefully acknowledge Jon Wakefield, without whose assistance the statistical model described in this chapter could not have been conceived or implemented.

C.1 Experimental protocols

C.1.1 RNA samples

For samples submitted to the Illumina Genome Analyzer II, Marnie and Jennifer performed poly(A) enrichment using the Dynabeads mRNA Purification Kit (Invitrogen #610-06) and used the RiboMinus Eukaryote Kit for ribosomal depletion (Invitrogen #A10837-08). They fragmented RNA by metal-ion catalysis (Ambion #AM8740). They made cDNA by random priming using the Superscript III First-Strand Synthesis Kit (Invitrogen #18080-093). They performed end repair, A-tailing, adaptor ligation, and gel purification according to recommended protocols (Illumina #1004898A). For samples submitted to the ABI SOLiD System, they performed poly(A) enrichment using the MicroPoly(A)Purist Kit (Ambion #AM1919) and used the RiboMinus Eukaryote Kit for ribosomal depletion. They prepared sequencing libraries using the SOLiD Total RNA-Seq Kit (Applied Biosystems #4452437A). All SOLiD samples were tagged with four barcodes per library using the SOLiD Fragment Library Barcoding Kit (Applied Biosystems #4443045B).

C.1.2 Genomic DNA samples

For genomic DNA, Marnie and Jennifer performed end repair, A-tailing, adaptor ligation, and gel purification according to recommended Illumina protocols (Illumina #1003806B) or used the SOLiD Fragment Library Construction Kit (Applied Biosystems #4443473). SOLiD samples were tagged with barcodes as above.

C.2 Testing for ASE using the binomial exact test

To test for ASE using the binomial exact test, I summed counts of reads called as BY and RM across all SNPs in each gene. When I had multiple samples sequenced using the same technology platform, I simply added read counts between replicates. I performed a two-sided binomial exact test of the null hypothesis that the BY and RM read counts are equal (i.e. binomial success probability = 0.5). I used the `binom.test` function in `R` to carry out this test for each gene [111].

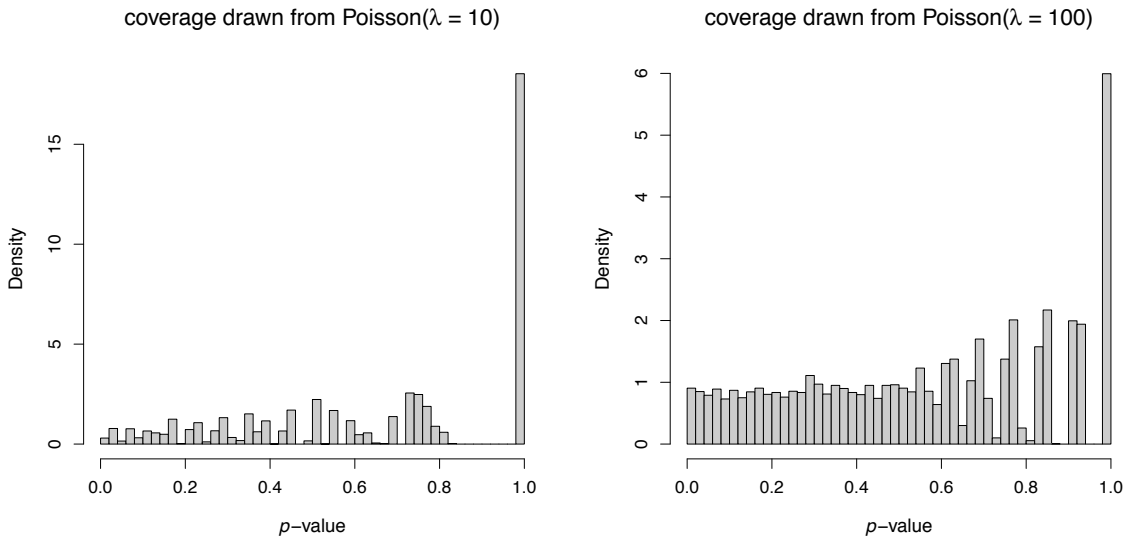


Figure C.1: Distributions of p-values under the null hypothesis for the binomial exact test

I simulated binomially distributed read counts under the null hypothesis of equal expression of both alleles to examine the null distribution of binomial test p-values. For a well-behaved test, the distribution of p-values under the null distribution should be approximately uniformly distributed between 0 and 1. As is apparent from Figure C.1, this assumption is grossly violated for the binomial exact test for these data. This problem is well-recognized in the context of testing for Hardy-Weinberg equilibrium [242].

I did not examine the accuracy of calling allele-specific expression using the binomial test at various false discovery rate thresholds as it is unclear how best to accurately calibrate

false discovery rates for a statistical test with such an irregular distribution of p-values under the null hypothesis.

C.3 Statistical model for ASE

I adopt a sequential approach where I conduct the following three steps:

1. Run a simple model on genomic DNA data to identify biased SNPs. Filter out these SNPs for all subsequent analyses.

2. Run a model for genomic DNA read counts that estimates overdispersion in this "null" data where no genes should show ASE. Use the parameters estimated in this step for step 3.

3. Run a model to detect ASE in read counts derived from RNA.

I describe these steps in more detail below.

C.3.1 Identifying biased SNPs

We sequenced genomic DNA from the BY/RM diploid hybrid in order to obtain "null" read counts that should be 50% BY and 50% RM for each SNP. I expected that allelic read counts would show some variability beyond that expected from statistical sampling, due to factors such as the many steps involved in preparation of sequencing libraries. I also noticed that read counts at a minority of SNPs in our genomic DNA data appeared highly biased, even after removing SNPs where reads generated *in silico* from the BY and RM genomes did not show 50/50 mapping of alleles (see section C.4.2). I constructed a simple model to identify SNPs that are highly biased towards the BY or RM allele so they could be filtered out in our RNA-Seq data.

My model for read counts Y_j is:

$$\begin{aligned} Y_j|\pi_0,\alpha,\delta,\epsilon \sim \pi_0 \times \text{Beta-Binomial}(N_j,\alpha,\alpha) \\ + (1-\pi_0) \times \text{Beta-Binomial}(N_j,\delta,\epsilon) \end{aligned} \tag{C.1}$$

where j indexes SNPs. The beta-binomial distribution arises when the probability p_j of success for each trial is not fixed but rather drawn from a beta distribution. This allows for the possibility of read counts that are overdispersed with respect to strictly binomially distributed read counts. A Beta-Binomial(N, α, α) distribution approaches a Binomial$(N, 0.5)$ distribution as $\alpha \to \infty$. I note that the beta-binomial distribution is different from the *negative binomial* distribution, which has been used by some investigators (e.g. [191, 232]) to model gene expression variation in the context of RNA-Seq data. The beta-binomial distribution is useful for modeling overdispersed binomially distributed counts, while the negative binomial distribution is useful for modeling overdispersed Poisson distributed counts.

Model (C.1) consists of two components. The first component, representing the majority of SNPs/genes (a fraction π_0), models read counts with $p_j \approx 0.5$. The second component is used to capture outlier loci as described below ($p_j \gg 0.5$ or $p_j \ll 0.5$). I constrain $\delta < 1$ and $\epsilon < 1$ to ensure that the distribution from which p_j's are drawn for the second component is U-shaped. The latter is desired because I expect highly biased SNPs/outlier genes to have probabilities close to 0 or 1. As priors on $\pi_0, \alpha, \delta, \epsilon$ we use

$$\begin{aligned}
\pi_0 &\sim \text{Unif}(0, 1) \\
\alpha &\sim \text{Log-Normal}(4.3, 1.8) \\
\delta &\sim \text{Unif}(0, 1) \\
\epsilon &\sim \text{Unif}(0, 1)
\end{aligned}$$

I chose the prior on α by performing a coarse grid search across a range of parameter values, simulating data, and verifying that simulated read counts were reasonable. I ran this model on all of our genomic DNA data, performing inference using Markov chain Monte Carlo (MCMC). I calculated the posterior probability that a particular SNP was highly biased using

$$\begin{aligned}
\Pr(\text{biased}|y_j) &= \Pr(\text{non-null}|y_j) \\
&= \frac{\Pr(y_j|\text{non-null})\Pr(\text{non-null})}{\Pr(\text{non-null})\Pr(y_j|\text{non-null}) + \Pr(y_j|\text{null})\Pr(\text{null})} \\
&= \frac{\texttt{dbetabin}(y_j, N_j, \hat{\delta}, \hat{\epsilon}) \times (1 - \hat{\pi_0})}{\texttt{dbetabin}(y_j, N_j, \hat{\delta}, \hat{\epsilon}) \times (1 - \hat{\pi_0}) + \texttt{dbetabin}(y_j, N_j, \hat{\alpha}, \hat{\alpha}) \times \hat{\pi_0}}
\end{aligned}$$

where `dbetabin` represents the beta-binomial density function. I removed SNPs with $P(\text{biased}|y_j) > 0.5$ from all further analysis. I found that there was a clear separation between well-behaved SNPs and those that appeared highly biased, with the set of SNPs called as biased remaining relatively invariant to changes in the threshold $P(\text{biased}|y_j)$ across a range from ≈ 0.1 to 0.9.

C.3.2 Model for genomic DNA data

As explained above, we sequenced genomic DNA from the BY/RM diploid hybrid in order to obtain "null" read counts that should be 50% BY and 50% RM for each SNP, with the expectation that allelic read counts would show some variability beyond that expected from statistical sampling, due to factors such as the many steps involved in preparation of sequencing libraries. I constructed a model for $i = 1, \ldots, m$ genes, with $j = 1, \ldots, m_i$ SNPs in gene i. I used the model to estimate the amount of overdispersion in read counts in our null data, which I then used to calibrate our model for RNA-Seq data (see C.3.3). For read counts Y_{ij} the model is

$$Y_{ij}|\alpha_i, \beta_i \sim \text{Beta-Binomial}(N_{ij}, \alpha_i, \beta_i)$$

I discuss the beta-binomial distribution and its relationship to the negative binomial distribution above (see C.3.1). I reparameterize α_i and β_i in terms of the mean p_i and dispersion e_i as

$$p_i = \alpha_i/(\alpha_i + \beta_i) \qquad \alpha_i = p_i(1 - e_i)/e_i$$
$$e_i = 1/(1 + \alpha_i + \beta_i) \qquad \beta_i = (1 - e_i)(1 - p_i)/e_i$$

As $e_i \to 0$ we approach the binomial model. The priors on p_i, e_i are:

$$p_i \sim \text{Beta}(a, a)$$
$$e_i \sim \text{Beta}(1, d)$$

I will use the notation Beta(α, β) to indicate the beta distribution, which has density

$$p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1 - \theta)^{\beta-1}$$

As priors on a and d I use

$$a \sim \text{Log-Normal}(4.3, 1.8)$$
$$d \sim \text{Exp}(0.0001)$$

These priors were chosen using a coarse grid search of the parameter space to obtain distributions that placed substantial probability across all realistic ranges for the parameters a and d, which I assessed by simulating data from these priors and verifying that the resulting read counts appeared reasonable. Thus, the posterior is given by

$$\begin{aligned} p(\boldsymbol{p}, \boldsymbol{e}, a, d|\boldsymbol{Y}) &\propto p(\boldsymbol{Y}|\boldsymbol{p}, \boldsymbol{e}) \times p(\boldsymbol{p}, \boldsymbol{e}|a, d) \times p(a, d) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{m_i} p(Y_{ij}|p_i, e_i) \times p(p_i, e_i|a, d) \times p(a, d) \end{aligned}$$

where

$$p(Y_{ij}|p_i, e_i) = p(Y_{ij}|\alpha_i, \beta_i) = \texttt{dbetabin}(Y_{ij}, N_{ij}, \alpha_i, \beta_i)$$
$$p(p_i, e_i|a, d) = \texttt{dbeta}(p_i, a, a) \times \texttt{dbeta}(e_i, 1, d)$$

with `dbeta` signifying the beta density and `dbetabin` the beta-binomial density. The model I have developed is not amenable to analytic methods of implementation, and so I implement using MCMC. I construct relatively straightforward independent Metropolis-Hastings random walk steps for each parameter, parameterizing on the log scale for a and d and the logistic scale for p_i and e_i, and using a normal proposal. I make use of estimates of a and d in our model for RNA-Seq data described below (C.3.3). I estimated $\hat{a}$ and $\hat{d}$ as the medians of the posterior densities of a and d, respectively.

C.3.3 Model for RNA-Seq data

For our RNA-Seq data, I construct a three-stage hierarchical model for allelic read counts for $i = 1, \ldots, m$ genes with $j = 1, \ldots, m_j$ SNPs in each gene. I denote the count of reads mapping to BY at SNP j in gene i as Y_{ij}, and in the first stage these counts are binomially distributed with parameters N_{ij} (coverage at the SNP) and p_{ij} (amount of ASE; $p_{ij} \approx 0.5$

signifies no ASE, $p_{ij} \neq 0.5$ indicates ASE). At the second stage, the p_{ij} arise from a gene-specific beta distribution with parameters α_i and β_i. This second stage allows for the possibility that p_{ij} may not be constant across all SNPs within gene i. The two stages of this model can be collapsed to give a beta-binomial model. I reparameterize the Beta(α_i, β_i) distribution in a similar manner as above (C.3.2):

$$p_i = \alpha_i/(\alpha_i + \beta_i) \qquad \alpha_i = p_i(1 - e_i)/e_i$$
$$e_i = 1/(1 + \alpha_i + \beta_i) \qquad \beta_i = (1 - e_i)(1 - p_i)/e_i$$

Let $\boldsymbol{\theta} = (f, g, h, \pi_0)$. I use a two component mixture prior on p_i, e_i:

$$p_i, e_i|\hat{a}, \hat{d}, \boldsymbol{\theta} \sim \begin{cases} \text{Beta}(\hat{a}, \hat{a}) \times \text{Beta}(1, \hat{d}) & \text{with probability } \pi_0 \\ \text{Beta}(f, g) \times \text{Beta}(1, h) & \text{with probability } 1 - \pi_0 \end{cases}$$

where $\hat{a}$ and $\hat{d}$ are obtained from the genomic DNA model (C.3.2). These two components naturally correspond to two classes of genes:

1. Class 1: Those showing no ASE, whose read counts are distributed according to

$$Y_{ij}|\hat{a}, \hat{d} \sim \text{Beta-Binomial}(N_{ij}, p_i, e_i)$$
$$p_i \sim \text{Beta}(\hat{a}, \hat{a})$$
$$e_i \sim \text{Beta}(1, \hat{d})$$

   The values of $\hat{a}$ are large ($\approx$ 3000 to 6500 for our data), which means that the p_i drawn from the Beta($\hat{a}, \hat{a}$) distribution are tightly concentrated around 0.5. Likewise, the values of $\hat{d}$ are also large ($\approx$ 550 for our data), which means that the dispersion e_i for each gene is small, as one would expect for null genes.

2. Class 2: Those showing ASE that is either constant at some level $p_i \neq 0.5$ across all SNPs in the gene, or that varies among SNPs in the gene.

The posterior is given by

$$\begin{aligned} p(\boldsymbol{p}, \boldsymbol{e}, \boldsymbol{\theta}|\boldsymbol{Y}, \hat{a}, \hat{d}) &\propto p(\boldsymbol{Y}|\boldsymbol{p}, \boldsymbol{e}) \times p(\boldsymbol{p}, \boldsymbol{e}|\hat{a}, \hat{d}, \boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{m_i} \prod_{r=1}^{R_t} p(Y_{ijr}|p_i, e_i) \times p(p_i, e_i|\hat{a}, \hat{d}, \boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \end{aligned}$$

where

$$p(Y_{ijr}|p_i, e_i) = p(Y_{ijr}|\alpha_i, \beta_i) = \texttt{dbetabin}(Y_{ijr}, N_{ijr}, \alpha_i, \beta_i)$$
$$p(p_i, e_i|\hat{a}, \hat{d}, \boldsymbol{\theta}) = \pi_0 \times \texttt{dbeta}(p_i, \hat{a}, \hat{a}) \times \texttt{dbeta}(e_i, 1, \hat{d}) + (1 - \pi_0) \times \texttt{dbeta}(p_i, f, g) \times \texttt{dbeta}(e_i, 1, h)$$

where `dbetabin` is the beta-binomial density function, and `dbeta` is the beta density function. The model is completed by placing priors on f, g, and h. To specify priors I parameterize f and g as $q = f/(f+g)$ and $r = 1/(1+f+g)$ which means that $f = q(1-r)/r$ and $g = (1-q)(1-r)/r$. Then

$$q \sim \text{Beta}(\alpha_q, \beta_q)$$
$$r \sim \text{Beta}(\alpha_r, \beta_r)$$

I performed a coarse grid search across a range of parameter values for α_q, β_q, α_r, and β_r, simulating samples of a and b for each combination of parameters and drawing values of p_{ij}. I found that $\alpha_q = \beta_q = 100$ and $\alpha_r = 1$, $\beta_r = 20$ resulted in reasonable values of a and b that produced realizations of p_{ij} close to those observed in the real data. I use

$$h \sim \text{Exp}(0.03)$$
$$\pi_0 \sim \text{Unif}(0, 1)$$

for the remaining priors. The prior on h was chosen to place substantial probability across all realistic ranges for that parameter. I simulated data using the above priors and verified that the resulting read counts were reasonable. As before, I implement this model using MCMC, with straightforward independent random walk Metropolis-Hastings steps for each parameter. I parameterize on the log scale for f, g, h and on the logistic scale for π_0, p_i, e_i, and use a normal proposal.

With $s = 1, \ldots, S$ draws from the posterior distribution of each parameter obtained via MCMC, the posterior probability that each gene shows ASE is just the posterior probability that gene i falls in component 2 (i.e. genes in Class 2 above). Let $\boldsymbol{\theta} = (f, g, h, \pi_0)$ and

abbreviate component 2 as C2. Then we have

$$\Pr(\text{C2}|y) = \frac{1}{S}\sum_{s=1}^{S} p(\text{C2}|p_i^{(s)}, e_i^{(s)}, \boldsymbol{\theta}^{(s)})$$

where

$$\begin{aligned} p(\text{C2}|p, e, \boldsymbol{\theta}) &= \frac{p(p, e|\text{C2}, \boldsymbol{\theta})p(\text{C2})}{p(p, e|\text{C2}, \boldsymbol{\theta})p(\text{C2}) + p(p, e|\text{C1}, \boldsymbol{\theta})p(\text{C1})} \\ &= \frac{\texttt{dbeta}(p|f, g)\texttt{dbeta}(e|1, h)(1-\pi_0)}{\texttt{dbeta}(p|f, g)\texttt{dbeta}(e|1, h)(1-\pi_0) + \texttt{dbeta}(p|\hat{a})\texttt{dbeta}(e|1, \hat{d})\pi_0} \end{aligned}$$

where `dbeta` is the beta density function.

C.3.4 Simulations to evaluate statistical model

In order to compare the power and robustness of the statistical approach described above to the binomial exact test, I simulated allele-specific read counts by mimicking the characteristics of our experimental data where possible. I chose to base these simulations on our Illumina GAII data because this was the largest of the datasets. Specifically, I simulated datasets that consisted of $n = 2000$ genes, with each gene containing j SNPs (where j is drawn from the distribution of the number of SNPs per gene present in the real data). Coverage levels of SNPs within genes were randomly drawn from the true coverage levels of SNPs within the dataset. I simulated 20 independent datasets, and present results below averaged across these simulations.

I first simulated genomic DNA count data, which I analyzed in the same manner as our real DNA count data. For each gene i, I drew a value $x_i \sim \text{N}(0, 0.188)$ representing the log fold change in expression between the two alleles of that gene. I estimated this standard deviation (0.188) using our real DNA count data. I then converted this to a binomial success probability for gene i using the formula $p_i = \exp(x_i)/[1 + \exp(x_i)]$. Each of the j SNPs within gene i had counts distributed binomially with parameter p_{ij}, and I added a small amount of "noise" to each mean binomial proportion p_i via $p_{ij} = \text{logit}^{-1}\left[\text{logit}\left(p_i\right) + \varepsilon_j\right]$, where $\varepsilon_j \sim \text{N}(0, 0.1)$. For $p_i = 0.5$ this corresponds to approximately 95% of the p_{ij}'s falling within the interval $(0.45, 0.55)$, as observed in the real data.

For simulated RNA count data, I simulated 50% of the genes as showing no ASE, and 50% showing ASE (i.e. $\pi_0 = 0.5$). For genes showing ASE, I drew a value $x_i \sim$

$\mathrm{N}(0, 0.658)$ representing the log fold change in expression between the two alleles of that gene and converted this to a binomial success probability for gene i using the formula $p_i = \exp(x_i)/[1 + \exp(x_i)]$. The standard deviation of the x_i's (0.658) was estimated from our real RNA count data. As with the DNA count data, each of the j SNPs within gene i had counts distributed binomially with parameter p_{ij}. To obtain p_{ij}'s, I added "noise" to each mean binomial proportion p_i as before: $p_{ij} = \mathrm{logit}^{-1}\left[\mathrm{logit}\ (p_i) + \varepsilon_j\right]$, where $\varepsilon_j \sim \mathrm{N}(0, 0.1)$.

Note that these simulations required that I assume a distributional form for the fold change in expression between the two alleles of each gene, and the distributional form for noise in read counts added to the simulations. I modeled these processes using normal distributions rather than taking the forms used in my statistical model in order to avoid biasing the results in favor of my model.

I computed a receiver operating characteristic (ROC) curve by tabulating the number of true positives called correctly and the number of false positives called incorrectly using p-value thresholds from 0 to 1 for the binomial exact test and posterior probabilities of ASE from 1 to 0 for our Bayesian statistical model. A plot of the ROC curve demonstrates that my model outperforms the binomial exact test as a classifier of ASE (Figure 3.2a).

In order to better understand the properties of our model I also examined whether the false discovery rate (FDR) was calibrated accurately. I use an indicator variable Z_i to label whether gene i shows ASE ($Z_i = 1$) or does not show ASE ($Z_i = 0$), which is known since the data was generated via simulation. I can compute the posterior probability that each gene shows ASE as described in section C.3.3, and I can compute the FDR for any chosen list of $i = 1, \ldots, n$ genes (out of m total genes) called as showing ASE using the formula

$$\mathrm{FDR} = \frac{1}{n} \sum_{i=1}^{n} \Pr(Z_i = 0 | \boldsymbol{y})$$

My simulations demonstrated that the FDR is calibrated accurately (Figure 3.2b; dotted red line follows $y = x$), with perhaps a very slight conservative bias in the reported FDR at relatively high rates (FDR > 0.2).

C.3.5 Expected level of overlap for measurements from different sequencing platforms

For sequencing platform t (with $t = \text{GA}, \text{ABI}$), I first constructed a list of genes called as showing significant ASE at a known FDR (typically 5%). I wish to estimate the fraction of genes expected to be called significant in both experiments. To this end, let $G_{ti} = 0/1$ if gene i in experiment t shows no ASE/ASE. Then I want to estimate

$$\frac{1}{m}\sum_{i=1}^{m}\Pr(G_{1i} = 1 \cap G_{2i} = 1). \tag{C.2}$$

The power to detect a signal varies across genes, and so G_{1i} and G_{2i} are not independent. To obtain an estimate of (C.2) requires a model, and so I perform a simulation.

I simulated read counts with coverage levels identical to the observed data. I used my model for RNA-Seq data (C.3.3) to generate read counts at each SNP, where I substituted for the parameters a, d, f, g, and h the posterior medians of these parameters generated by running MCMC on our observed data. I subsequently analyzed the simulated data using the same methods as our observed data.

In order to compare some of the characteristics of our observed data with the simulated data, I used estimates of the FDR and False Non-Discovery Rate (FNDR) to calculate the expected number of false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN), as well as the probability of type I error (T1) and probability of type II error (T2). I can calculate these values using

$$\begin{aligned}
m &= \text{total number of genes} \\
n_t &= \text{number of genes called significant in experiment } t \\
\text{FP}_t &= \text{FDR}_t \times n_t \\
\text{FN}_t &= \text{FNDR}_t \times (m - n_t) \\
\text{TP}_t &= n_t - \text{FP}_t \\
\text{TN}_t &= (m - n_t) - \text{FN}_t \\
\text{T1}_t &= \text{FP}_t/(\text{TN}_t + \text{FP}_t) \\
\text{T2}_t &= \text{FN}_t/(\text{TP}_t + \text{FN}_t)
\end{aligned}$$

For the two experiments reported in the main text, I calculated

$$\mathrm{T1}_{\mathrm{GA}} = 0.067$$

$$\mathrm{T1}_{\mathrm{ABI}} = 0.029$$

$$\mathrm{T2}_{\mathrm{GA}} = 0.60$$

$$\mathrm{T2}_{\mathrm{ABI}} = 0.67$$

I found that my simulated data had reasonably similar type I and type II error rates:

$$\mathrm{T1}_{\text{simulated GA}} = 0.10$$

$$\mathrm{T1}_{\text{simulated ABI}} = 0.11$$

$$\mathrm{T2}_{\text{simulated GA}} = 0.61$$

$$\mathrm{T2}_{\text{simulated ABI}} = 0.59$$

I tabulated the number of genes called significant at FDR = 5% in both my observed and simulated data. As shown in Figure C.2, the level of overlap for our observed data is

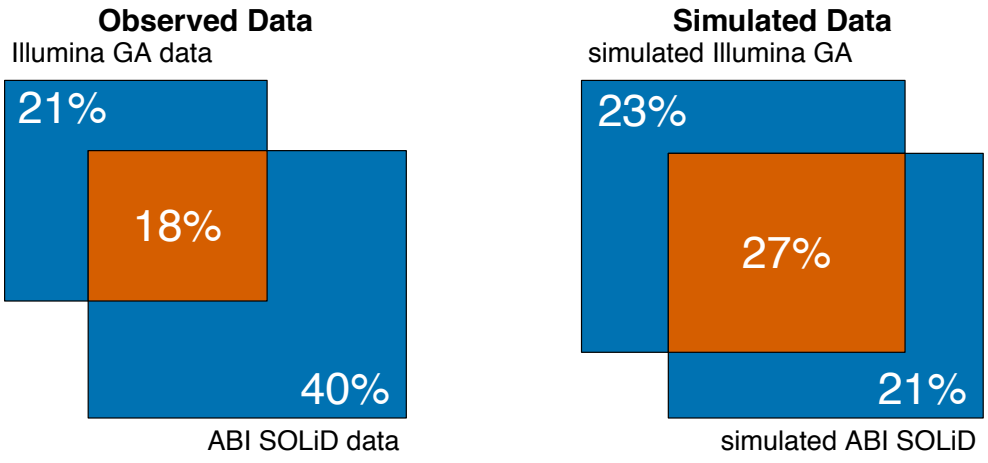


Figure C.2: Numbers in each square represent the number of genes falling in that grouping (significant only in GA data, significant only in ABI data, or significant in both), divided by the sum of the number of genes called significant in each of the two datasets.

reasonably similar to what I calculated in the simulated data.

The higher overlap in my simulated data likely reflects at least two factors:

1. I generated simulated data using our model for RNA-Seq data (C.3.3) in order to produce datasets with similar type I and type II error rates as seen in the real data. However, as shown above, the type I and type II error rates were not identical between the simulated and real data. For example, the type I error rate is significantly higher in the simulated ABI data.

2. My simulated data was generated via a known model and analyzed using a statistical framework employing the same model. For real data the read counts are generated by an unknown mechanism. I believe our model captures much of the underlying complexity of this mechanism, but any model is necessarily a simplification of reality.

C.4 Read mapping

C.4.1 Details of read mapping procedure

I obtained complete genome sequences for BY from the Saccharomyces Genome Database (June 2008 sequence; `http://www.yeastgenome.org/`) and for RM from the Broad Institute (`http://www.broadinstitute.org/`). After repeat masking the sequences [161], I used `LASTZ` (`http://www.bx.psu.edu/miller_lab`) to infer alignment scoring parameters appropriate for aligning the BY and RM genomes and to generate pairwise alignments between all chromosomes of the two strains. I then used `TBA` [162] to compute a whole-genome alignment that is not biased in favor of any particular reference genome. I masked any nucleotides that were ambiguous in either genome and projected this alignment to both BY and RM to construct reference genomes for the strains. To the genome of each strain I added all unaligned segments from the opposing strain as a single contig. I mapped all reads to the BY and RM genomes using the program `BFAST` [163]. I aligned reads using primary and secondary indices (Table C.1) suggested in Homer et al. [163]. I aligned reads using options `-K 100` and `-M 500` for `bfast match`, and examined only reads that had a single highest-scoring alignment to each genome. I output the results in SAM format and

converted to BAM format using `samtools` [188].

Table C.1: Indices used for `BFAST` mapping

data type	mask	hash width	index type
nucleotide	`1111111111111111`	14	1°
nucleotide	`111101001101111010101011111`	14	2°
nucleotide	`1111111111111001111`	14	2°
nucleotide	`111101110110010100111111`	14	2°
colorspace	`1111111111111111`	14	1°
colorspace	`11111011101110101001010111011111`	14	2°
colorspace	`10111101011010010110000110100011111111`	14	2°
colorspace	`10111001101001100100111101010001011111`	14	2°

I examined the alignment of each read to the BY genome and to the RM genome in order to search for reads with distinguishable allelic origin. I required reads to map to approximately the same genomic location in BY and RM; specifically, I required each read to map within the same alignment block in each strain. I used the CIGAR string present in the BAM output file to reconstruct the alignment between each read and the BY and RM genomes. I used a simple probabilistically-motivated, base quality-aware scoring scheme implemented in the program `cross_match` (`http://phrap.org/phredphrapconsed.html`) to score the alignment of the read to the genome of each strain. This scheme awards matching bases a score of 6 and mismatching bases a score of $6 - (q + 5)$ where q is the base quality. I considered a read to be a candidate BY read if the score was higher for the alignment to the BY genome, and vice versa.

C.4.2 Read mapping simulations

In our data and in previous datasets examined [152], a small proportion of SNPs is biased toward one of the two alleles. Degner et al. [152] identified the presence of flanking sequences sharing identity with another region of the genome as one factor contributing to this read-mapping bias at some SNPs in humans. To overcome this potential source of bias, I simulated 50 bp reads overlapping every SNP and indel ascertained using our whole-genome alignment of BY and RM. I mapped these reads using the same methods as for our real data. For our experimentally acquired data, I then filtered out any SNP showing either a bias of at least 5% greater than equal mapping of alleles among our simulated reads or less than 95% of my simulated reads mapping to the correct genomic location and allelic background.

C.5 Correction for GC content

It has been noted by other investigators that base composition has a significant effect on the propensity of a molecule to be sequenced using high-throughput sequencing technologies [159,164,165]. My overall strategy was to correct for GC content by using our genomic DNA data to measure how often sequences of varying GC content were sequenced, and correcting read counts derived from our RNA data to reflect the fact that alleles of particular GC content might be over- or under-sequenced solely due to their base composition. I followed many of the methods suggested by Pickrell et al. [159] to perform this correction. More specifically, my correction for GC content consisted of the following steps:

1. Divide the yeast genome into b 200 base pair bins and calculate the read depth N_b in each of the bins from our genomic DNA data

2. Calculate the mean read depth $\bar{N}_i$ across all b_i bins of GC content i

3. Calculate the $\log_2$ relative enrichment of reads f_i of each GC content i: $r_i = \log_2(\bar{N}_i/\bar{N})$

4. Fit a spline to the relationship between GC content i and $\log_2$ relative enrichment of reads r_i in bins of GC content i. I performed this analysis in `R` [111] using the function

`smooth.spline` with `spar` = 0.8.

5. Calculate a "correction factor" f_i for each bin corresponding to the amount that reads of GC content i have been over- or under-sequenced. I used the `predict` function in `R` [111] to obtain a predicted $\log_2$ relative enrichment $\hat{r}_i$ given the spline estimated above. This correction factor could be used to correct absolute RNA read counts in bin b with GC content i for over- or under-sequencing using the formula $N_b^{\text{corrected}} = N_b \times 2^{-f_i}$

6. Use the correction factors calculated for BY and RM at bins overlapping each SNP to correct the *relative* read counts in our RNA data. Rather than altering both the BY and the RM read counts, I alter only one read count, by only the amount the more extreme correction factor f_i differs from the less extreme correction factor. For example, if both the BY and the RM alleles are predicted to be over-sequenced, with $f_i^{BY} = 1.10$ and $f_i^{RM} = 1.12$, I keep the BY read count unchanged and multiply the RM read count by $f_i^{RM} - f_i^{BY} = 1.02$. I round altered counts to the nearest integer.

This scheme allows me to correct for possible differences in sequencing depth driven by differences in base composition between the BY and RM alleles, as I am only interested in comparing read counts derived from the BY and RM alleles within each locus. Since the BY and RM alleles are usually very similar in base composition, my scheme has the desired effect of keeping most counts largely unaltered, and only altering counts where differences in base composition leading to different probabilities of sequencing the BY and RM allele could plausibly affect conclusions about ASE. Estimates of global parameters were nearly identical regardless of whether I performed the GC content correction described above.

C.6 Comparing absolute expression levels between technologies

I performed a comparison of absolute transcript abundances for each of the two RNA-Seq platforms from which I obtained data. For this analysis I focused only on absolute expression levels, ignoring all information about allele-specific expression and averaging the measurements for each gene for the ABI platform (since I obtained RNA-Seq data for two samples for this platform). First, I compared absolute expression levels measured

on each platform and, as expected, found a highly significant correlation between these measurements (Figure C.3; Pearson's $\rho = 0.78$, $p < 2.2 \times 10^{-16}$).

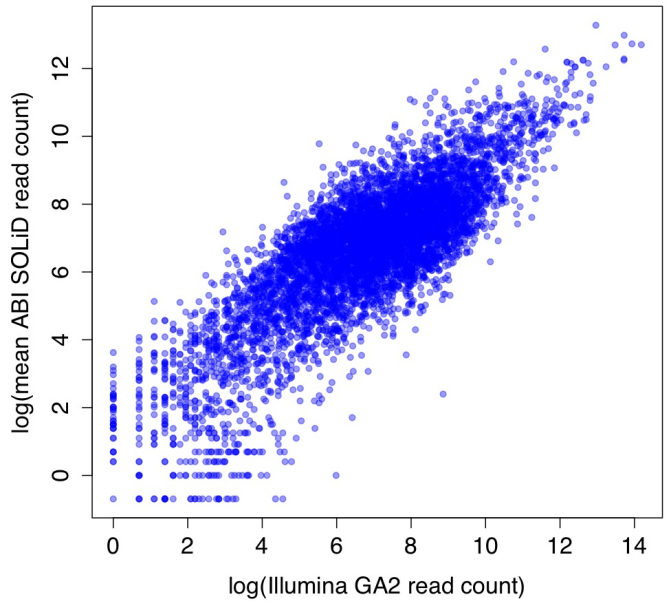


Figure C.3: Comparison of absolute expression levels measured using the ABI SOLiD and Illumina GA2 sequencing platforms

Next I explored whether differences in absolute transcript abundance between the two technology platforms could explain some of the differences in calling ASE observed for the two platforms. Consider a single gene which truly shows ASE. If this gene is sequenced to equal depth by both platforms, the test for ASE will be similarly powered in both analyses. However, if stochastic variations in sequencing depth result in the gene being sequenced to a higher level in one RNA-Seq dataset and a lower level in another, an analysis of the gene in the former dataset will have higher power to detect ASE than the latter. I binned genes by the absolute value of the difference in coverage between the Illumina and ABI datasets, and found a negative correlation between this difference in coverage and the probability that the ASE calls made on each platform agreed (Figure C.4). To confirm that the coverage difference between technologies is negatively correlated with the fraction of ASE calls where

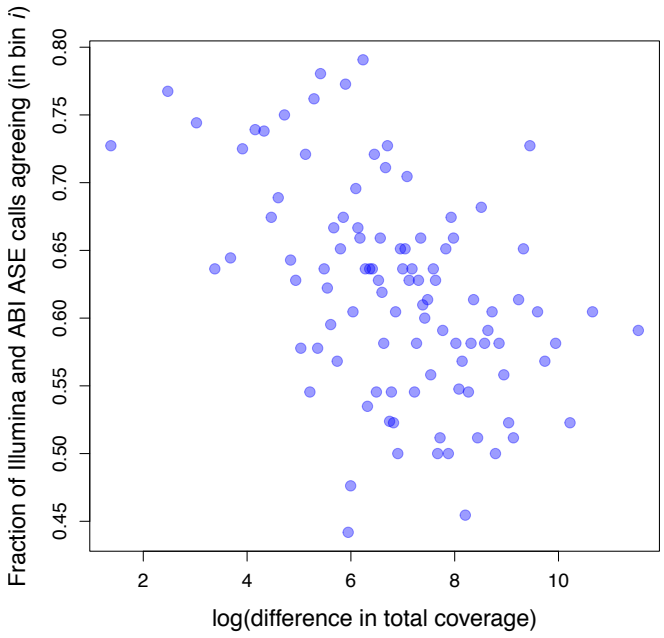


Figure C.4: Probability is plotted as a function of the difference in total coverage between the platforms (binned and shown on log scale)

both platforms agree, I divided the data into 100 bins based on this coverage difference and calculated the fraction of ASE calls in concordance between the two platforms for each bin. I then fit the linear model

$$\text{ASE concordance} = \log(\text{total coverage}) + \log(\text{coverage difference}) + \epsilon$$

This resulted in a weakly significant estimate for the contribution of coverage difference to the model ($p = 0.0249$) with the expected negative effect on concordance of ASE calls between platforms. This indicates that some portion of the discrepancy in genes called as showing ASE between platforms is due to stochastic variation in sequencing coverage of those genes.

C.7 Analysis of data without removing PCR duplicates

Although I removed PCR duplicates for our primary results as described in the main text, I explored the effect of not removing these reads. Our Illumina data was paired-end while

our ABI data was single-end. As a result, I removed more potential duplicates from our ABI data since I was unable to distinguish between reads that were true duplicates and those that mapped to the same $5'$ location by chance. Using all of our data with duplicate reads included, I ran the same analyses reported in the main text.

I found that including duplicate reads resulted in increased noise in read counts. Specifically, the posterior distributions of a (which affects the tightness of the p_i around 0.5 for null genes - see C.3.2) overlapped between the analyses with and without including PCR duplicates, for both technology platforms. However, estimates of d (which governs the amount of dispersion within each null gene - see C.3.2) were significantly lower when including PCR duplicates, indicating greater within-gene noise in read counts. The 95% credible intervals for each scenario are shown in Table C.2. A greater value of within-gene dispersion for

Table C.2: Confidence intervals for d both with and without including PCR duplicates

platform	PCR duplicates included?	2.5% quantile	97.5% quantile
ABI	yes	192	226
ABI	no	441	617
Illumina	yes	464	543
Illumina	no	526	628

"null" read counts means that it is more difficult to distinguish genes showing ASE from those not showing ASE.

I also found that my estimate of π_0, the global fraction of genes not showing ASE, was lower if I included duplicate reads in my analyses (i.e. more genes were identified as showing ASE when duplicate reads were included). I analyzed the Illumina Genome Analyzer II and ABI SOLiD System data separately to determine how the estimates changed depending on whether duplicate reads were included. Interestingly, estimates of π_0 decreased much more dramatically for the ABI data than for the Illumina data (for which there are far fewer reads marked as duplicates), indicating that including duplicate reads strongly affects estimates

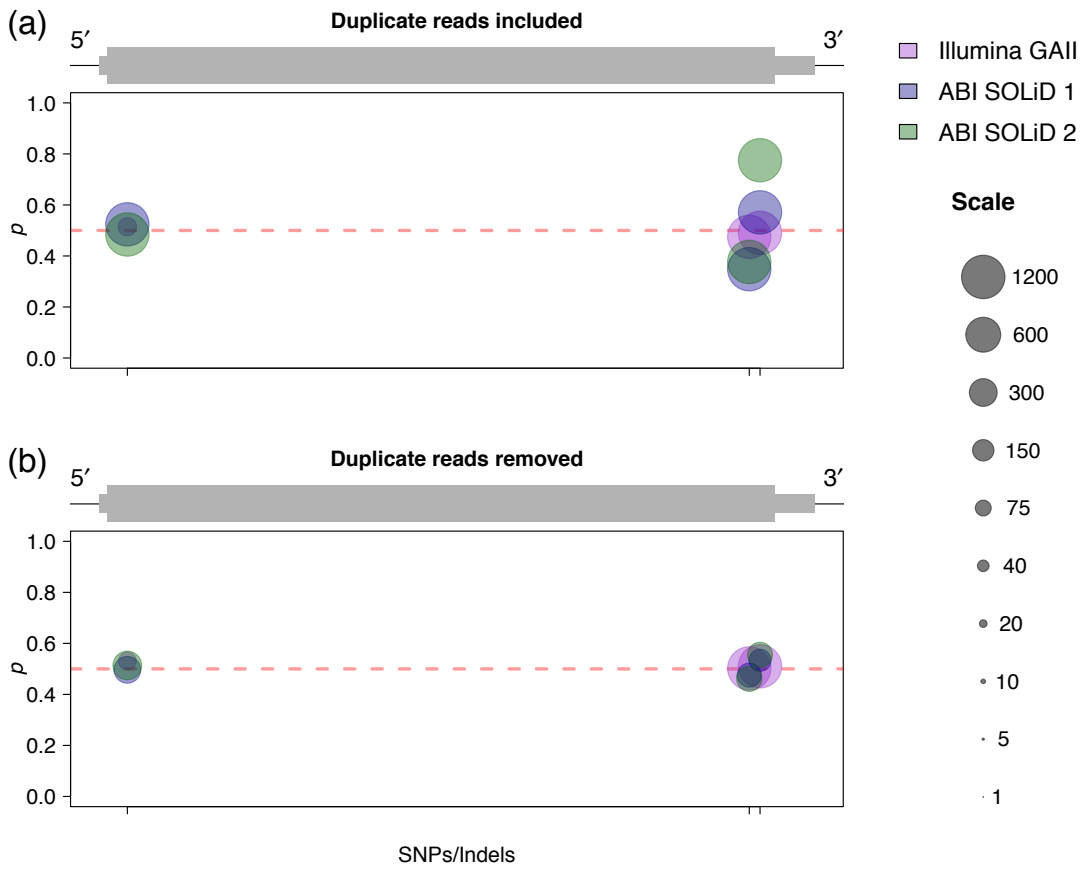


Figure C.5: Data for gene *SSA2*. Organization and coloring of plots is identical to that in Fig. 4b-c and Fig. 5a-c in the main text. (a) Data with duplicate reads included. (b) Data with duplicate reads removed.

of global parameters.

Finally, I found genes that showed highly discrepant results between analyses of the data with and without duplicate reads included. These genes tended to have low posterior probabilities of ASE when duplicate reads were removed, but higher posterior probabilities of ASE when including duplicate reads. One example is the gene *SSA2* (YLL024C). I found a very large number of reads mapped to the first two SNPs of this gene in all datasets. Despite a large number of reads mapping, the level of ASE at these SNPs is quite inconsistent between datasets if duplicate reads are included (Figure C.5a). Furthermore, at least for some datasets there is suggestive evidence that this gene may show ASE. However, if I remove duplicate reads from the analysis the datasets give far more consistent results, and any evidence for ASE disappears (Figure C.5b).

C.8 Unifying previous statistical tests for allele-specific expression

In this section, I note previously published tests for detecting allele-specific expression using RNA-Seq. For tests that are applicable to experimental designs different from ours, I show how the test relates to our situation (detecting allele-specific expression using transcribed heterozygous sites in a single individual). Previous studies of allele-specific expression using RNA-Seq, and the statistical method employed by each, are shown in Table C.3. I also note that Main et al. [153] present a method for analyzing allele-specific read counts using a linear model in a scenario where a small number of genes has been sequenced to very high coverage, but that it is unclear how to adapt their test to our situation.

C.8.1 Test of Pickrell et al. [159] applied to our data

In Pickrell et al. [159], the authors sequenced RNA from $n = 70$ individuals and looked for allele-specific expression at a subset of genes (those genes showing significant evidence for putative *cis*-acting gene expression QTL) among all individuals in the population. To conduct their test, the authors sum up the read counts at all SNPs in a gene. For a single

Table C.3: Statistical tests employed by previous studies of allele-specific expression by RNA-Seq

Reference	Binomial exact test	χ^2 test	Notes
Degner et al. [152]	✓		
Emerson et al. [155]	✓		Reduces to asymptotic equivalent of binomial exact test for our experimental design (see C.8.2)
Heap et al. [156]		✓	
McManus et al. [157]	✓		
Montgomery et al. [158]	✓		Customized to individual sequencing lanes
Pickrell et al. [159]	✓		Reduces to asymptotic equivalent of binomial exact test for our experimental design (see C.8.1)
Zhang et al. [154]		✓	

gene, Pickrell et al. [159] model read counts as beta-binomially distributed, i.e.

$$Y = \text{Number of reads from allele 1}$$
$$N = \text{Number of reads from allele 1 + allele 2}$$
$$Y \sim \text{Beta-binomial}(N, \alpha, \beta)$$

where alleles 1 and 2 are defined as the haplotype that tends to cause higher expression of the gene and the haplotype that tends to cause lower expression of the gene. The authors use a beta-binomial model to allow for overdispersion in the data *between individuals*, which they note could arise due to differences in genetic background [159]. Then

$$p(Y|N, \alpha, \beta) \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(Y+\alpha)\Gamma(N-Y+\beta)}{\Gamma(N+\alpha+\beta)}$$

Pickrell et al. [159] maximize the overall log-likelihood of the data

$$\log P(\text{data}) = \sum_{i=1}^{n} \log p(y_i | N_i, \alpha, \beta)$$

for the n individuals and use a likelihood ratio test of $\alpha = \beta$ vs. $\alpha \neq \beta$ to obtain a p-value.

For this study, I examined only one individual rather than 70. Since I restrict my study to a single individual, I do not need to consider overdispersion in read counts between individuals. If I parameterize the beta-binomial distribution as

$$\mu = \frac{\alpha}{\alpha + \beta}$$
$$\rho = \frac{1}{1 + \alpha + \beta}$$

one can think of ρ as the dispersion between individuals and μ as the mean level of ASE for a gene. By the logic above, $\rho = 0$ since my study consists of only one individual. Thus $\alpha + \beta \to \infty$, and the likelihood reduces to the binomial. Thus the test of Pickrell et al. [159], adapted to my experimental design, results in a likelihood ratio test for $p = 0.5$ vs. $p \neq 0.5$ with a binomial likelihood where $p \equiv \mu$. This test is asymptotically equivalent to the binomial exact test.

C.8.2 Test of Emerson et al. [155] applied to our data

In Emerson et al. [155], the authors examined regulatory variation in the same diploid *S. cerevisiae* hybrid that I studied. They implement a binomial model that is designed to detect both *cis* and *trans*-regulatory variation in their experimental design. My experimental design focused on detecting *cis*-regulatory variation, but the model of Emerson et al. [155] can be adapted for this situation. The authors defined

$$d = \frac{\text{number of cells containing allele 1}}{\text{number of cells containing allele 2}}$$
$$e = \frac{\text{expression level per cell for allele 1}}{\text{expression level per cell for allele 2}}$$
$$p = \frac{de}{de + 1}$$

If we let

$$Y = \text{Number of reads from allele 1}$$

$$N = \text{Number of reads from allele 1 + allele 2}$$

then the authors estimate $\hat{d}$ over all n genes, using genomic DNA sequence data, with the maximum likelihood estimator

$$\hat{d} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} N_i - Y_i}$$

In my case, $d \approx 1$, as Emerson et al. [155] report for sequencing of diploid hybrid genomic DNA. To test for allele-specific expression due to *cis*-regulatory variation Emerson et al. [155] maximize the binomial likelihood

$$L(p|Y, N) = \binom{N}{Y} p^Y (1-p)^{N-Y} \tag{C.3}$$

where

$$p = \frac{de}{de+1} = \frac{e}{e+1}$$

and conduct a likelihood ratio test with the null hypothesis $e = 1$ which implies $p = 0.5$. Thus, for my situation this amounts to a likelihood ratio test of $p = 0.5$ vs. $p \neq 0.5$ with a binomial likelihood, which is asymptotically equivalent to the binomial exact test.

Appendix D

SUPPLEMENT TO CHAPTER 4

D.1 Yeast strains

For this study, we purchased strains used in Liti et al. [1] from the National Collection of Yeast Cultures (`http://www.ncyc.co.uk/sgrp.html`). Caitlin Connelly confirmed strains as haploid through mating tests or made them haploid by integrating a *KanMX* cassette at the *HO* locus, sporulating the transformants, and isolating haploid spores. Caitlin confirmed that the cassette had integrated at the *HO* locus through PCR using one primer flanking the cassette and one inside the cassette.

Table D.1 provides a brief overview of the strains studied, and more detailed information about each strain is available in Liti et al. [1].

D.2 Phenotyping

D.2.1 DNA sequencing

Sample preparation

Marnie Johansson grew strains to mid-log phase (OD_{660} 0.8-1.0) in yeast extract peptone dextrose and extracted DNA by the phenol:chloroform:IAA method. She performed sequencing library preparation as previously described [5] and obtained sequences using the Illumina Hi-Seq platform (50 bp paired-end reads), with individual samples barcoded to enable multiple libraries per lane.

Validation of SNP calls by Sanger sequencing

Marnie sequenced nine randomly chosen regions across seven chromosomes in two strains (NCYC361 and UWOPS83-787.3) in order to validate my short-read genotype calls in regions of imputed sequence [1]. These two strains are among the most genetically dis-

tinct of all the strains we studied and have among the most imputed sequence of any strain. The regions I chose totaled 5.9 kb in size and overlapped 86 imputed SNPs, where our short-read genotyping calls disagreed with the imputed genotype in 41 cases. Marnie grew strains to mid-log phase (OD_{660} 0.8-1.0) in yeast extract peptone dextrose and extracted DNA using the MasterPure Yeast DNA purification kit (Epicentre). I designed primers using the program Primer3 [110] and Marnie sequenced the regions using Sanger sequencing on an ABI 3130xl machine. The regions from which we obtained high-quality sequence were: chrV:240884-241529, chrIX:399255-400030, chrX:48040-48719, chrX:678485-679128, chrXI:588177-588789, chrXIII:332837-333434, chrXIII:900894-901537, chrXV:642438-643100, and chrXVI:840028-840694. I aligned each Sanger sequence read to strain-specific reference genomes using the program `SMALT` (`http://www.sanger.ac.uk/resources/software/smalt/`) and found 100% concordance with genotyping calls we made using short-read data. There were small insertions/deletions apparent in the Sanger sequence reads (relative to the strain-specific reference genome), but I did not attempt to call these variants using short-read data.

D.2.2 Chemostat culture

Chemostat culturing was performed by Emily Mitchell. Emily streaked strains from -80°C freezer stocks in randomized batches of 6-10 at a time onto YPD plates. She put the plates at 30°C for 2 days before picking single colonies to grow overnight in 3 mL YPD. She inoculated chemostats with 1 mL YPD culture, and allowed growth for 24 hours before beginning continuous culture. Emily maintained a dilution rate of 0.17 +/- .01 volumes per hour for 3-4 days, until cultures were deemed to have reached steady state. She defined steady state as stabilizing to within 10% of the previous days density measurements. In order to quantify steady state she used measurements by Klett colorimeter and by OD_{600}. To avoid any perturbation, Emily took sample culture passively at the effluent port. If the culture density stabilized, she harvested the chemostat and used the samples for RNA, protein, metabolite, and microscopy studies. She repeated the above schedule for all strains, such that replicates were harvested, in most cases, in less than 25 generations. Some strains

Figure D.1: Brief information on the properties of each strain studied. For more detailed information on the strains, see Liti et al. [1]. [1]We obtained partial data for strains DBVPG6040 and DBVPG6044. Due to difficulties with chemostat growth, we obtained only one successful sample of each strain. We gathered quantitative proteomics and cellular morphology for both strains, and metabolomic data for DBVPG6044, but did not obtain gene expression or genome sequence data. [2]Strain K11 was determined not to be phosphate limited, due to a much larger number of differentially expressed genes than other samples and a great enrichment among these genes for phosphate transporters (with K11 having lower expression than all other strains).

Strain name	Isolated	Source	Strain type
273614N	Newcastle, UK	fecal	clinical
378604X	Newcastle, UK	sputum	clinical
BC187	Napa Valley, CA	barrel	fermentation
DBVPG1106	Australia	grapes	fermentation
DBVPG1373	Netherlands	soil	wild
DBVPG6040[1]	Netherlands	juice	fermentation
DBVPG6044[1]	West Africa	bili wine	fermentation
DBVPG6765	unknown	-	-
K11[2]	Japan	sake	fermentation
L-1374	Chile	must	fermentation
NCYC361	Ireland	wort	fermentation
SK1	USA	soil	lab
UWOPS05-217.3	Malaysia	plant	wild
UWOPS05-227.2	Malaysia	bee	wild
UWOPS83-787.3	Bahamas	fruit	wild
UWOPS87-2421	Hawaii	plant	wild
Y12	Africa	palm wine	fermentation
Y55	France	grape	lab
YJM975	Bergamo, Italy	vaginal	clinical
YJM978	Bergamo, Italy	vaginal	clinical
YJM981	Bergamo, Italy	vaginal	clinical
YPS128	Pennsylvania, USA	oak tree	wild
YPS606	Pennsylvania, USA	oak tree	wild
YS2	Australia	-	baking
YS9	Singapore	-	baking

required increased inoculum (2 ml) and a slightly extended timeline (up to 32 generations) to reach steady state.

Chemostat media was phosphate limited, and contained the following per liter: 100 mg calcium chloride, 100 mg sodium chloride, 500 mg magnesium sulfate, 5 g ammonium sulfate, 1 g potassium chloride, 10 mg potassium phosphate, 500 μg boric acid, 40 μg copper sulfate, 100 μg potassium iodide, 200 μg ferric chloride, 400 μg manganese sulfate, 200 μg sodium molybdate, 400 μg zinc sulfate, 2 μg biotin, 400 μg calcium pantothenate, 2 μg folic acid, 2 mg inositol, 400 μg niacin, 200 μg p-aminobenzoic acid, 400 μg pyridoxine, 200 μg riboflavin, 400 μg thiamine, and 5 g glucose.

At harvest, Emily took the following samples (in the order listed):

1. 3 mL of effluent (passively collected) for microscopy (Section D.2.6)

2. 50 mL culture for RNA-Seq. Emily immediately (within one minute of initial culture perturbation) filtered the sample, froze cells using liquid nitrogen, and stored at -80°C until processing (Section D.2.3)

3. 50 mL culture for protein analysis. Emily pelleted at 4°C, washed in cold 50mM ammonium bicarbonate pH 7.8, re-pelleted, and frozen the sample using liquid nitrogen at -80 °C until processing (Section D.2.4

4. 50 mL for metabolite analysis, which was processed immediately (Section D.2.5).

D.2.3 RNA-Seq

Beginning with aliquots from samples taken from the chemostats, Marnie Johansson extracted RNA by the acid phenol method. She performed poly(A) enrichment (MicroPoly(A) Purist Kit, Ambion) followed by ribosomal depletion (RiboMinus Kit, Invitrogen). She prepared RNA-Seq libraries, barcoded samples, and performed sequencing (50-bp single-end reads) according to the manufacturers recommendations using the ABI SOLiD v4 (SOLiD Whole Transcriptome Analysis Kit, ABI). Marnie randomly allocated our samples across

three flowcells. We obtained 5-50 million reads per sample (equivalent to $\approx 20 - 200$X coverage of the genome).

D.2.4 Quantitative proteomics

Lysis

Beginning with aliquots from samples taken from our chemostats, Beth Graczyk lysed samples in batches of 5-10 samples. She pelleted aliquots of 40 mL of cells for each sample by centrifugation for 3 minutes at 2,000 rpm. She resuspended cells in 0.7 mL of 50 mM ammonium bicarbonate pH 7.8 and lysed with 0.7 mL glass beads by vortexing at speed 8-9 for 20-30 minutes in cold room until complete lysis. Beth first centrifuged lysate at $5,000 \times g$ for 5 minutes at 4°C to clear debris and centrifuged again at $100,000 \times g$ for one hour at 4°C to separate soluble and insoluble fractions.

Digestion

Gennifer Merrihew digested samples in three different randomized batches. She resolubilized the insoluble pellet using 0.1% RapiGest (Waters Corporation) in 50 mM ammonium bicarbonate pH 7.8 and sonicated for 20 seconds at speed #3 followed by 5 minutes of heat at 100°C. After the samples were cooled, she used a BCA protein assay (Pierce) to measure the protein concentration. She then reduced samples with DTT (dithiothreitol), alkylated with IAA (iodoacetic acid) and digested with sequence modified trypsin (Promega) at a 1:50 trypsin:protein ratio for one hour at 37°C. Gennifer added 200 mM Hcl to cleave RapiGest and halt digestion, and centrifuged to separate peptides from debris. She cleaned samples with mixed-mode cation exchange columns (MCX) (Waters Corporation).

LC-MS/MS

Gennifer Merrihew performed the steps detailed in this section. She used fused silica microcapillary columns of 75 μm inner diameter (Polymicro Technologies, Phoenix, AZ) packed in-house by pressure loading 40 cm of Jupiter 90 Å C12 material (Phenomenex, Torrance, CA). In order to assess quality of the column before and during analysis, she used an

equimolar mix of a six protein bovine digest (Michrom Bioresources, Inc., Auburn, CA). She analyzed three of these quality control runs prior to any sample analysis and after every six sample runs analyzed another quality control run. She randomized samples and ran them in replicate. She loaded three μg of each sample digest and 200 femtomole of the six protein bovine digest onto the column by the NanoACQUITY UPLC (Waters Corporation, Milford, MA) system. For buffer solutions she used: were water, 0.1% formic acid (buffer A) and acetonitrile, 0.1% formic acid (buffer B). The 100 minute gradient of the six protein bovine digest quality control consisted of 69 minutes of 93% buffer A and 7% buffer B, 1 minute of 65% buffer A and 35% buffer B, 10 minutes of 20% buffer A and 80% buffer B and 20 minutes of 93% buffer A and 7% buffer B at a flow rate of 0.25 μl/min. The 180 minute gradient for the sample digest consisted of 140 minutes of 91% buffer A and 9% buffer B, 20 minutes of 80% buffer A and 20% buffer B, 6 minutes of 20% buffer A and 80% buffer B and 14 minutes of 91% buffer A and 9% buffer B at a flow rate of 0.25 μl/min. Peptides were eluted form the column and electrosprayed directly into an LTQ-FT mass spectrometer (ThermoFisher, San Jose, CA) with the application of a distal 3 kV spray voltage. For the six protein bovine digest quality control analysis, a cycle of one 25,000 resolution full-scan mass spectrum (400-1400 m/z) followed by five selected reaction monitoring (SRM) spectra analyzing five peptides and 4-5 fragment ions per peptide at 35% normalized collision energy with a 2 m/z isolation window. For the sample digests, she used a cycle of one 50,000 resolution full-scan mass spectrum (400-1400 m/z) followed by five data-dependent MS/MS spectra at 35% normalized collision energy with a 2 m/z isolation window. She used the ThermoFisher XCalibur data system to control application of the mass spectrometer and UPLC solvent gradients.

Data Analysis

Gennifer Merrihew and Nick Shulman performed the data analysis steps outlined in this section. They analyzed the SRM six protein bovine digest using `Skyline` [243]. They processed high resolution MS data using `Bullseye` [244] to optimize precursor mass information. To identify peptides, they searched the MS/MS data using `SEQUEST` [245] against

a fasta database containing all the protein sequences from all the strains. They determined peptide spectrum match false discovery rates using `Percolator` [246] at a q-value threshold of 0.01 and a posterior error probability threshold of 1. They assembled the peptides into protein identifications using an in-house implementation of `IDPicker` [247].

In order to obtain a quantitative measure of peptide abundance, Gennifer used the program `Topograph` [248]. `Topograph` searches for a chromatographic peak for each identified peptide in each sample, and integrates the intensity over the retention time of the peak. In samples where `SEQUEST` had identified a particular peptide in an MS/MS scan, `Topograph` searched for a peak in the chromatogram that overlapped the range of times where the peptide was identified. For samples without an MS/MS identification for a particular peptide, `Topograph` performed a pairwise retention time alignment against those samples which did have an MS/MS identification for that peptide. The retention time alignment used a loess regression to find the best path which came nearest to the peptide identifications that were in common between the two samples. In samples requiring alignment, `Topograph` restricted its search for peaks to the range of times of the aligned identifications.

Gennifer also performed searches to find novel peptides using a six-frame translated fasta database created for each strain except for DBVPG6040 and DBVPG6044 which were searched against databases for NCYC361 and SK1, respectively (their closest relatives, since we did not obtain genome sequence for these two strains). The peptides were then assembled into protein identifications using an in-house implementation of `IDPicker`. Gennifer performed a similar search for translation of pseudogenes. Finally, she performed a third search for protein modifications, specifically oxidized methionine and phosphorylation of serine, threonine, and tyrosine. We did not expect to find many phosphorylation sites in these modification searches because we did not use phosphatase inhibitors during lysis of the samples.

D.2.5 Metabolite profiling

The steps detailed in this section were performed by Sara Cooper.

Metabolite extraction

Beginning with aliquots from samples taken from our chemostats, for each biological replicate Sara divided 20 mLs of culture into two 10 mL aliquots. Each was prepared separately as a technical replicate yielding a total of four replicates for each strain. Sara centrifuged at 4°C for 3 minutes at 3000g, washed the pellet with 10 mL water, and spun again. She immediately resuspended the pellet in 500 μl water and added 500 μl of cold methanol. After mixing, she incubated the mixture on a dry ice-ethanol bath at -40°C for 30 minutes. Then she thawed the frozen mixture on ice for 10 minutes and spun at 4 °C for 5 minutes at 3000g. The supernatant was frozen at -80°C until the time of analysis.

Derivatization

Sara performed derivatizations as previously described [249]. Briefly, 100 μL of each sample was dried down under vacuum. Methylene chloride was added to dried sample and removed under vacuum to remove residual water. To the dry sample, she added 30 μL of 20 mg/ml methoxyamine in pyridine. She incubated the samples at 30°C for 60 minutes. Then we added 70 μL of MSTFA+1% TCMS (Thermo catalog number 48915) and incubated at 60°C for 60 minutes.

Two-dimensional gas chromatography with TOF mass spectrometry

All metabolite analysis was done on a Leco Pegasus 4D system (GC×GC-TOFMS). Sara acquired and processed data using the `ChromaTOF` software. She capped and injected derivatized samples using a CTC Analytics autosampler (Gerstel). The GC columns were as follows: primary column–20 m × 250 μm × 0.4 μm RTX-5MS (Restek), secondary column–2 m × 180 μm id × 0.2 μm RTX-200 (Restek). Sara injected 1 μL of each sample using a split ratio of 1:5. The initial GC oven temperature was 60°C, the modulator temperature was 30°C above the primary oven temperature and the initial secondary oven temperature was 75°C. The inlet temperature was 280°C and the transfer line was 280°C. The flow rate of the helium was 1 ml/min. The oven temperatures increased at a constant rate of 7°C/min to a final oven temperatures of 310°C and 325°C, respectively. Modulation to achieve two-

dimensional separation was 5 seconds with a 0.4 s hot pulse and 2.1 s cold pulse. The ion source was 250°C and the data were acquired at a rate of 100 Hz from 70-600 m/z. Sara ran the majority of samples (78/91) on a Pegasus 4D system located in Seattle, WA and the remainder (13/91) on a second Pegasus 4D system located in Huntsville, AL. All methods were identical. Some data processing was different due to differences in average signal. We used location of sample processing as a covariate in normalizations (see below). Unless stated, the samples were otherwise identical.

Data processing and analysis

Sara used the software `ChromaTOF` (Leco) for peak calling and deconvolution. The following key parameters were set: peak width 1st dimension: 10 s, peak width 2nd dimension 0.1 s, match required to combine: 750, signal to noise ratio: 10 (for the subset of samples acquired on machine 2, overall signals were higher, so we required a signal to noise ratio of 25).

To facilitate comparison of samples, Sara created a reference sample. A single sample was used as a template and was edited manually to remove duplicate peaks and assign quant masses to maximize accuracy of quantification. For each sample, the `ChromaTOF` software determines whether there is a peak in the reference that matches. If so, the the QuantMass of the matching peak is assigned to the same unique mass as the reference. The requirement for achieving a match between an unknown peak and the reference were: spectral similarity – 500, 1st dimension retention time – within 10 s, 2nd dimension retention time – within 0.2 s. This processing step improves our ability to compare metabolite levels between samples. Peak areas were normalized by dividing the area under the curve by the median area of all peaks for each sample. This metric normalizes for differences in sample concentration and injection volume. For the subset of samples run on the second Leco Pegasus 4D system, the secondary retention times were adjusted by a linear equation determined by mapping known metabolites between samples run on each of the machines. The equation used to adjust the secondary retention time was $y = (x - 0.3521)/0.6848$.

Sara eliminated from analysis all samples with an insufficient average signal to noise ratio. Ultimately this resulted in 91 samples, representing 23 different yeast strains. Sara

used the software package `Guineu` v1.0 [250] to align the common metabolites among 91 individual sample files. `Guineu` parameters were set at RT1 deviation: 10 s, RT1 penalty = 25, RT2 deviation 0.2 s RT2 penalty = 25, minimum spectra match: 500, name bonus = 50.

The `Guineu` output generated a list of 419 metabolites that appeared in at least 10 of the 91 samples assayed. Of those, 83 were manually annotated as background peaks because they appeared in equal or higher concentrations in a blank sample. Of the remaining metabolites, we used `ChromaTOF`s library matching algorithm (an implementation of AMDIS) to identify the best metabolite identifications available in NIST libraries as well as a custom library including our own standards and the Fiehn Library (Leco). We conducted additional manual annotation to validate metabolite identities based on retention times when available. Finally, we generated a list of 117 metabolites that were present in at least half of the samples and for 93 of those we have potential, though not validated, identifications. The list of metabolites generated using these methods was used as input for all other analysis described.

D.2.6 Cellular morphology

The steps detailed in this section were performed by Eric Muller.

Cell fixation and staining

Immediately following the harvest from chemostats, Eric fixed 1.2 mL cells in media in 3% formaldehyde, 0.1 M potassium phosphate buffer, pH 6.7. The formadehyde was deactivated in 10 mM ethanolamine in 0.1 M potassium phosphate buffer, pH 6.7. Following centrifugation, Eric suspended cells in 0.1 M potassium phosphate, pH 6.7, then put on ice, sonicated with 12 pulses, 0.5 sec/pulse, pelleted at 2000 rpm for 7 minutes in a microfuge and resuspended in 10-20 μl PBS. Eric simultaneously stained DNA, cell wall, and actin with 1 μg/ml DAPI, 200 μg/ml ConA-Alexa Fluor 488 and 0.33 μM Phalloidin-Alexa Fluor 546 (Invitrogen) in PBS with 0.1% Triton X-100.

Fluorescence microscopy

Eric mounted cells on an agarose pad as described (`http://www.youtube.com/watch?v=ZrZVbFg9NE8`) except the pad was not dried before adding cells. He acquired images on a DeltaVision Core using a 100X UPlanApo NA 1.35 objective and the Photometrics CoolSnapHQ camera. The `Calmorph` software [251] was developed to process images acquired from the same camera, so pixel dimensions in the acquired images were the same in both studies. Eric acquired images with 1×1 binning with a 1024×1024 image size in a 20 section Z-series, 0.2 μm/section. He then converted the Z-stack to a single image using the `Softworx` software quick projection-max intensity protocol. He binned images 2×2 and converted to scaled 8-bit tiffs . He processed tiff images with `Calmorph` [251] to generate ≈ 100 quantitative morphological traits for each strain. Daniel Jaschob modified `Calmorph` to accept the dimensions of our images and the tiff format, but otherwise Eric used the program unchanged. Eric collected data for a total of ≈ 800 cells per strain, in two replicates. `Calmorph` calculates cell dimensions in pixels, which in this study correspond to 0.1290 μm/pixel. `Calmorph` outputs measurements that can be used to construct a total of 501 different traits [196]. I discarded any `Calmorph` measurements related to image brightness as Eric considered this unreliable to measure. I also did not use any "total stage" traits calculated at the population level by `Calmorph` since cells were selected to include some from each cell cycle stage rather than being selected completely randomly, leaving a total of 398 traits.

D.3 Association Mapping

D.3.1 Power and false positive rates

Caitlin Connelly conducted simulations to determine the power and false positive rate for our tests of association. For the simulations, she picked 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. Caitlin 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

Effect size	α Level	Power
0.25	0.05	0.788
	1×10^{-5}	0.022
0.5	0.05	0.986
	1×10^{-5}	0.389
0.75	0.05	1
	1×10^{-5}	0.925

Table D.1: Power to detect associations of a variety of effect sizes. α levels shown are not corrected for multiple testing. Effect size is specified in terms of percent variance explained.

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 $= p(1-p)k^2/(p(1-p)k^2 + 1 - 1/n) \approx 1/(1 + 1/(p(1-p)k^2)$ 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 [252]. To assess power, Caitlin tested for association between the simulated data and the genotype at the causal variant for each of the 1000 simulations using EMMA [206]. To assess the type I error rate, she picked 1000 random SNPs and asked how often they showed association in any of the 1000 simulated datasets. As Table D.1 shows, we have reasonable power to detect associations of large effect. Moreover, the type I error rate is only slightly elevated above that found in an idealized scenario in the absence of population structure (Figure D.2).

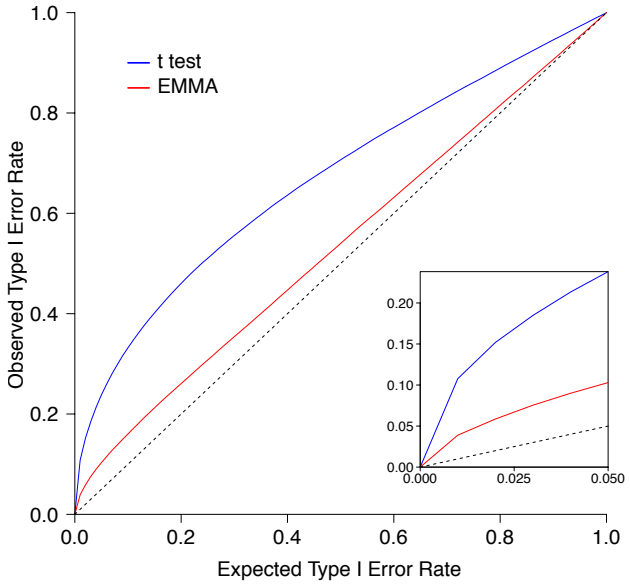


Figure D.2: Type I error rate in simulated data. The mean observed type I error rate from 1000 simulations is plotted versus the expected type I error rate for association tests done using a simple t test (blue) and EMMA (red). The theoretical expectation in the absence of population structure is shown as a dashed line. Inset, detail of the observed vs. expected type I error rates at low expected error rates.

VITA

Daniel Skelly was born in Milwaukee, Wisconsin on December 17, 1982. He earned a Bachelor of Science degree in Zoology from the University of Wisconsin - Madison in 2005. He joined the University of Washington graduate program in Genome Sciences in 2007 and completed his thesis work in the laboratory of Dr. Joshua M. Akey.