A cradle-to-grave analysis of *cis*-regulatory variation in yeast

Jennifer Margaret Andrie

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Joshua M. Akey, Chair

Stanley Fields

Jay A. Shendure

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

A cradle-to-grave analysis of *cis*-regulatory variation in yeast

Jennifer Margaret Andrie

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

Cis-regulatory variation is an important source of phenotypic variation within populations and a major target of adaptive divergence between species. However, the molecular processes that are influenced by *cis*-regulatory variation remain poorly understood. To this end, we crossed two genetically diverse wild-derived strains of *Saccharomyces cerevisiae* and studied allele-specific differences in six molecular phenotypes, including chromatin structure, rates of RNA transcription, decay and translation, RNA secondary structure, and binding of proteins to RNA. Furthermore, we performed high-coverage sequencing of both the genome, using PacBio, and the transcriptome, and *de novo* assembled each parental strain to mitigate read mapping biases and ensure accurate estimates of allele-specific phenotypes. We show that *cis*-regulatory variation has pervasive influence on the conversion of genotype into

phenotype, and that pleiotropy is a predominant feature in the architectural landscape of *cis*-regulatory mutations. Our comprehensive data also provide novel mechanistic insights into *cis*-regulatory variation, revealing an important role for RNA secondary structure and RNA binding protein-RNA interactions in determining RNA decay rate and translation efficiency. Overall, we find that relationships between allelic differences in the measured molecular phenotypes are highly complex. Our data represent the most comprehensive analysis conducted to date of how genetic variation influences gene expression and, therefore, provide, an important advancement toward the understanding of how genetic variation produces phenotypic variation.

INTRODUCTION:

The awesome power of yeast genetics for understanding *cis*-regulatory variation

A fundamental objective in modern biology is to determine how genotype produces phenotype. Over the last decade, rapid advancements in the field of genomics have led to extraordinary reductions in the cost of sequencing and, concomitantly, a wealth of genome sequence data (reviewed in Goodwin *et al.* 2016). To date, over 21,000 genomes from over 1,000 organisms have been deposited in the National Center for Biotechnology Information (NCBI) Genome database (https://www.ncbi.nlm.nih.gov/genome/browse/). Furthermore, several large projects, including the 1000 Genomes Project in humans, the 1001 Genomes Project in *Arabidopsis thaliana*, the 1002 Yeast Genomes Project in yeast, as well as smaller projects, like the 100-genomes strains sequencing in yeast, are creating of vast catalogs of the natural genetic variation present across populations (Strope *et al.* 2015; Sudmant *et al.* 2015; The 1000 Genomes Project Consortium 2015; The 1001 Genomes Consortium 2016; http://1002genomes.u-strasbg.fr). However, understanding how natural genetic variation among individuals generates differences in functional genomics traits, including gene and protein expression, function, and interaction, and, consequently, in organismal phenotypes, remains a formidable challenge.

To study the relationship between genotypic variation and phenotypic variation in eukaryotic cells, perhaps the most powerful experimental model organism is the budding yeast *Saccharomyces cerevisiae* (Nieduszynski and Liti 2011; Botstein and Fink 2011). Yeast first emerged as a model organism for eukaryotic cell biology in the 1960's and

1970's, primarily through the genetic and biochemical experiments of Fred Sherman (reviewed in Liebman and Haber 2013). However, it was the successful transformation in 1978 of yeast with a plasmid that had been replicated in *Escherichia coli* and the subsequent development of the "awesome power of yeast genetics," comprising of a myriad of recombinant DNA technologies that allowed researchers to easily move genes in and out of yeast cells, that ultimately distinguished yeast as the premier eukaryotic model organism for establishing relationships between gene structure and function (Duina *et al.* 2014). The utility of *S. cerevisiae* for uncovering genotype-phenotype relationships further increased when it became the first eukaryotic genome to be fully sequenced and annotated (Goffeau *et al.* 1996); and later, the first organism with a deletion collection of nearly every non-essential open reading frame (Winzeler *et al.* 1999; Giaever *et al.* 2002). More recently, yeast has become the pioneer organism in the emerging field of functional genomics (Botstein and Fink 2011; Skelly and Magwene 2015). Specifically, yeast was the first eukaryote to be subjected to genome-wide analyses of gene expression levels (DeRisi *et al.* 1997), mRNA stability (Wang *et al.* 2002), and translation rates (Ingolia *et al.* 2009), and, additionally, was one of the first eukaryotes to have protein-protein interactions characterized (Uetz *et al.* 2000; Ito *et al.* 2000), the global proteome analyzed (Washburn *et al.* 2001; Ghaemmaghami *et al.* 2003), and the global metabolome profiled (Raamsdonk *et al.* 2001). As new functional genomics assays develop, they facilitate our ability to measure population-level variation in functional genomics phenotypes and to delineate the genetic architecture underlying this variation; yeast has also been at the forefront of these population-level studies (Skelly and Magwene 2015). In one of the earliest such studies, Brem *et al.* developed a

method to map quantitative trait loci (QTL) underlying inheritance of gene expression in segregants of a cross between two genetically divergent yeast strains (2002). They showed that the variation underlying most gene expression is highly complex, with many QTLs exhibiting weak effects, and with pervasive transgressive segregation and epistasis (Brem *et al.* 2002). Currently, the features that originally made it a premier model organism, including its experimental tractability and extremely well characterized genome, combined with the growing availability of whole-genome sequences for hundreds of *Saccharomyces cerevisiae* strains (Gallone *et al.* 2016; Strope *et al.* 2016; Zhu *et al.* 2016) as well as the high level of genetic diversity among strains ($\pi = 4 \text{ X } 10^{-3}$) (Peter and Schacherer 2016), continue to make yeast an attractive and powerful experimental system for dissecting the connection of genotype to phenotype.

A critical component of mapping genetic variation to phenotypic variation will be determining how *cis*-regulatory variation influences functional genomic phenotypes. Specifically, heritable regulatory variation can broadly be classified as either acting in *cis*, meaning the variation resides within or around the gene being regulated, or in *trans*, meaning the variation resides at a location distant from the gene being regulated (Figure 1, adapted from Skelly *et al*. 2009). In humans, *cis*-regulatory variants have been identified that affect susceptibility to autoimmune, infectious, neoplastic, neurodegenerative, and psychiatric diseases (Skelly *et al*. 2009). In addition, *cis*-regulatory variation also plays a prominent role in evolutionary diversification, and has been shown to influence phenotypes such as beak morphology in Darwin's finches (Abzhanov *et al.* 2004). To ascertain the mechanisms by which *cis*-regulatory variation generates phenotypic variation, a commonly used approach, first developed by Brem *et*

al., as described above, is to map QTL for a functional genomic trait of interest, such as gene expression levels, translation efficiency, or protein abundance from segregants of a cross or individuals in a population (2002; for another example, see Battle *et al.* 2015). QTL identified in or near the gene they affect are considered *cis*, while those distantly located are labeled as *trans*. A more direct approach for determining the impact of *cis*-regulatory variation is to measure allelic differences in the functional genomic trait of interest in a diploid hybrid (for example, see Skelly *et al.* 2011). A powerful advantage of this approach over QTL methods is that it internally controls for *trans*-acting regulatory variation as well as environmental factors. In yeast, allele-specific methods have already been used successfully to examine the *cis*-regulatory genetic variation underlying individual functional genomic phenotypes, including chromatin accessibility (Connelly *et al.* 2014), mRNA decay rate (Andrie *et al.* 2014), and translation efficiency (Albert *et al.* 2014). However, these studies have only begun to scratch the surface in terms of exploring how allele-specific variation influences functional genomic phenotypes.

Specifically, the number of functional genomics technologies that measure the interactions between DNA, RNA, and protein, as well as properties of these biomolecules, like structure or dynamic behavior is rapidly expanding, allowing the consequences of *cis*-regulatory variation to be more broadly studied than previously possible. Importantly, not only have these new technologies have increased the breadth of molecular phenotypes we can assay genome-wide, but also the ease with which researchers can conduct assays, due to shorter, simpler protocols more readily applied to a broader array of organisms. For example, the recently developed Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq)

method is a fast and sensitive alternative to the DNase-seq or FAIRE-seq assays that measure chromatin accessibility genome-wide, and the MNase-seq assay that identifies nucleosome positions in accessible regions of the genome (Buenrostro *et al.* 2015). The ATAC-seq method requires 1000-fold less starting sample than the alternative methods listed above, and sequencing-ready libraries are produced in a mere few hours (Buenrostro *et al.* 2015). Researchers have successfully applied this method to a diversity of organisms, including humans (Buenrostro *et al.* 2013), *Arabidopsis thaliana* (Lu *et al.* 2016), and yeast (Schep *et al.* 2015). Likewise, the newly developed RNA proximity ligation (RPL) method (Ramani *et al.* 2015) provides quicker, simpler alternative to DMS-seq (Rouskin *et al.* 2013) for the high-throughput determination of secondary and tertiary RNA structures. Additionally, methods such as Protein Interaction Profile sequencing (PIP-seq) (Silverman *et al.* 2014), which assesses global RNA-binding protein-RNA (RBP-RNA) interactions, as well as methods that examine RNA modifications like Pseudo-seq (Carlile *et al.* 2014) and Ψ-seq (Schwartz *et al.* 2014), which both identify pseudouridine modification sites transcriptome-wide, provide researchers with tools to globally assess molecular phenotypes that were previously unfeasible to study on a genome- or transcriptome-wide scale.

While each of the many functional genomic phenotypes is interesting in isolation, an integrated analysis of how genetic variation affects all of these phenotypes will be essential for improving our understanding of how genotype affects phenotype. Recently, the first two such an analyses of this kind were conducted in human lymphoblastoid cell lines (LCLs) (Battle *et al.* 2015; Cenik *et al.* 2015). Motivated by the observation that variation in mRNA and protein expression levels are often uncorrelated (Foss *et al.* 2007;

Ghazalpour *et al.* 2011; Picotti 2013; Albert *et al.* 2014), these studies attempted a comprehensive analysis of the functional consequences of genetic variation among individuals by coordinately examining steady-state RNA abundance via RNA-seq, translation efficiency via ribosome profiling, and steady-state protein abundance via quantitative mass spectrometry (Battle *et al.* 2015; Cenik *et al.* 2015). Interestingly, they came to conflicting conclusions (Battle *et al.* 2015; Cenik *et al.* 2015). Battle *et al.* detected a scarcity of translation-specific QTL, and therefore, infer that attenuation of mRNA expression levels at the protein level is likely due to post-translational processes, as opposed to differences in translation (2015). Conversely, Cenik *et al.* find that ribosome occupancy correlated better with protein levels than RNA abundance correlated with protein levels, and therefore, argue that genetic variation may penetrate to phenotype through changes in translation (2015). As evidenced by the disagreement between these two studies, further investigation of the relationships between RNA abundance, translation efficiency, and protein abundance is needed. However, to accurately map the connections between genotype and phenotype, studies must also to expand to include synchronous measurements of a broader array of functional genomic phenotypes; future studies should strive to incorporate structural properties of DNA and RNA like chromatin accessibility and RNA secondary structure, as well as the dynamics underlying RNA and protein abundance through measurement of transcription rates, mRNA decay rates, and protein decay rates. Additionally, mapping the influence of genetic variation on phenotypic variation will also require researchers to distinguish the effects *cis*- and *trans*-regulatory variation. Indeed, coordinated measurement of several functional genomic phenotypes remains ripe for future analysis. Given its many advantages described above,

and, importantly, its amenability to allele-specific approaches that powerfully differentiate *cis*- from *trans*-regulatory variation, yeast offers an exceptional model system for beginning to combine population genomic approaches with synchronous measurement of several functional genomic phenotypes in order to map the influence of genetic variation, and in particular, *cis*-regulatory variation, on phenotypic variation.

FIGURES

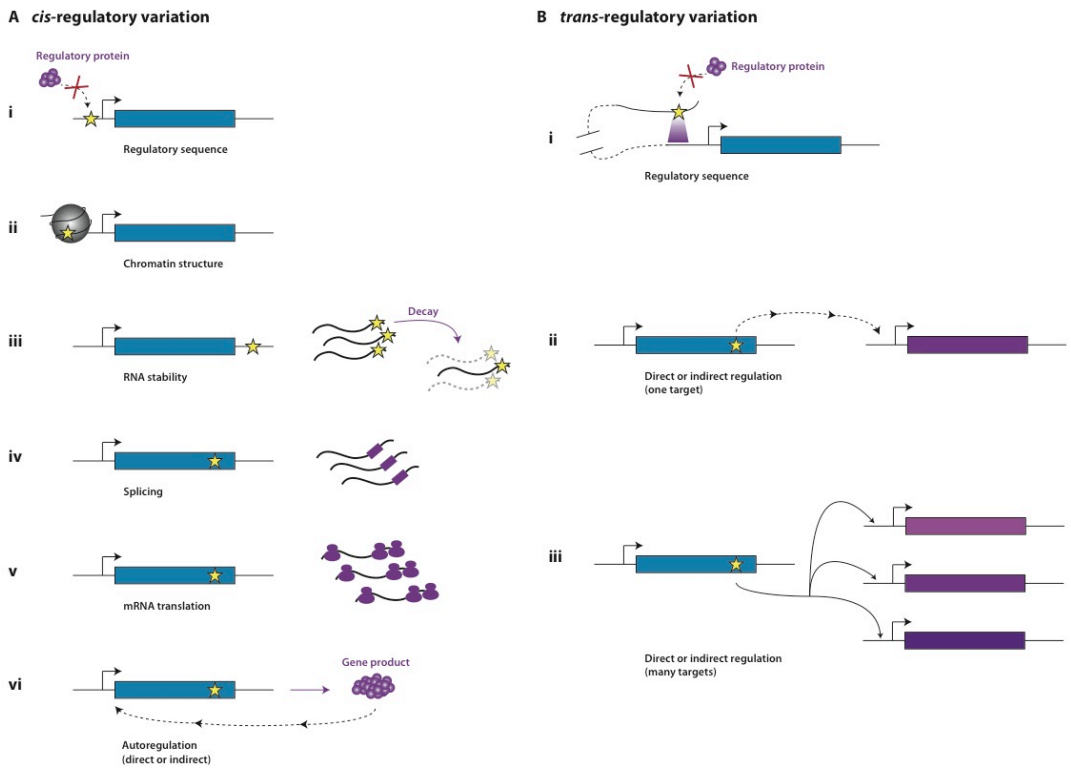


Figure 1. Molecular mechanisms of *cis*- and *trans*-regulatory variation (adapted from Skelly *et al.* 2009). A) *Cis*-regulatory variation acts from a position near the gene of interest. This type of variation can impact gene expression levels by affecting i) the binding of regulatory proteins to regulatory 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, v) sequences that contribute to transcript-specific translation rates to determine protein production and vi) regulation of the gene by its own product or the product of a gene downstream in the transcriptional regulatory network. B) *Trans*-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.

REFERENCES

1000 Genomes Project Consortium, 2015 A global reference for human genetic variation. Nature 526: 68-74.

1001 Genomes Consortium, 2016 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166: 481-491.

Botstein, D., and G. R. Fink, 2011 Yeast: an experimental organism for 21st Century biology. Genetics 189: 695-704.

Abzhanov, A., M. Protas, G. R. Grant, P. R. Grant, and C. J. Tabin, 2004 Bmp4 and morphological variation of beaks in Darwin's finches. Science 305: 1462-1465.

Albert, F. W., D. Muzzey, J. S. Weissman, and L. Kruglyak, 2014 Genetic influences on translation in yeast. PLoS Genet. 10: e1004692.

Andrie, J. M., J. Wakefield, and J. M. Akey, 2014 Heritable variation of mRNA decay rates in yeast. Genome Res. 24: 2000-2010.

Battle, A., Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford *et al.*, 2015 Impact of regulatory variation from RNA to protein. Science 347: 664-667.

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

Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, 2013 Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10: 1213-1218.

Buenrostro, J. D., B. Wu, H. Y. Chang, and W. J. Greenleaf, 2015 ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. 109: 21.29.1-21.29.9.

Carlile, T. M., M. F. Rojas-Duran, B. Zinshteyn, H. Shin, K. M. Bartoli *et al.*, 2014 Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature 515: 143-146.

Cenik, C., E. S. Cenik, G. W. Byeon, F. Grubert, S. I. Candille *et al.*, 2015 Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. Genome Res. 25: 1610-1621.

Connelly, C. F., J. Wakefield, and J. M. Akey, 2014 Evolution and genetic architecture of chromatin accessibility and function in yeast. PLoS Genet. 10: e1004427.

DeRisi, J. L., V. R. Iyer, and P. O. Brown, 1997 Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686.

Duina, A. A., M. E. Miller, and J. B. Keeney, 2014 Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. Genetics 197: 33-48.

Foss, E. J., D. Radulovic, S. A. Shaffer, D. M. Ruderfer, A. Bedalov *et al.*, 2007 Genetic basis of proteome variation in yeast. Nat Genet. 39: 1369-1375.

Gallone, B., J. Steensels, T. Prahl, L. Soriaga, V. Saels *et al.*, 2016 Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. Cell 166: 1397-1410.

Ghaemmaghami, S., W. K. Huh, K. Bower, R. W. Howson, A. Belle *et al.*, 2003 Global analysis of protein expression in yeast. Nature 425: 737-741.

Ghazalpour, A., B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian *et al.*, 2011 Comparative analysis of proteome and transcriptome variation in mouse. PLoS

Genet. 7: e1001393.
Giaever, G., A. M. Chu, L. Ni, C. Connelly, L. Riles *et al.*, 2002 Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387-391.
Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.*, 1996 Life with 6000 genes. Science 274: 546, 563-567.
Goodwin, S., J. D. McPherson, and W. R. McCombie, 2016 Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17: 333-351.
Ingolia, N. T., S. Ghaemmaghami, J. R. Newman, and J. S. Weissman, 2009 Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218-223.
Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori *et al.*, 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. U. S. A. 98: 4569-4574.
Liebman, S. W., and J. E. Haber, 2013 Retrospective. Fred Sherman (1932-2013). Science 342: 1059.
Lu, Z., B. T. Hofmeister, C. Vollmers, R. M. DuBois, and R. J. Schmitz, 2016 Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. Nucleic Acids Res. pii: gkw1179.
Nieduszynski, C. A., and G. Liti, 2011 From sequence to function: Insights from natural variation in budding yeasts. Biochim. Biophys. Acta. 1810: 959-966.
Peter, J., and J. Schacherer, 2016 Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. Yeast 33: 73-81.
Picotti, P., M. Clément-Ziza, H. Lam, D. S. Campbell, A. Schmidt *et al.*, 2013 A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. Nature 494: 266-270.
Raamsdonk, L. M., B. Teusink, D. Broadhurst, N. Zhang, A. Hayes *et al.*, 2001 A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat. Biotechnol. 19: 45-50.
Ramani, V., R. Qiu, and J. Shendure, 2015 High-throughput determination of RNA structure by proximity ligation. Nat. Biotechnol. 33: 980-984.
Rouskin, S., M. Zubradt, S. Washietl, M. Kellis, and J. S. Weissman, 2014 Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature 505: 701-705.
Schep, A. N., J. D. Buenrostro, S. K. Denny, K. Schwartz, G. Sherlock *et al.*, 2015 Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res. 25: 1757-1770.
Schwartz, S., D. A. Bernstein, M. R. Mumbach, M. Jovanovic, R. H. Herbst *et al.*, 2014 Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell 159: 148-162.
Silverman, I. M., F. Li, A. Alexander, L. Goff, C. Trapnell *et al.*, 2014 RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biol. 15: R3.
Skelly, D. A., J. Ronald, and J. M. Akey, 2009 Inherited variation in gene expression. Annu. Rev. Genomics Hum. Genet. 10: 313-332.
Skelly, D. A., M. Johansson, J. Madeoy, J. Wakefield, and J. M. Akey, 2011 A powerful and flexible statistical framework for testing hypotheses of allele-specific gene

expression from RNA-seq data. Genome Res. 21: 1728-1737.
Skelly, D. A., and P. M. Magwene, 2016 Population perspectives on functional genomic variation in yeast. Brief Funct. Genomics 15: 138-146.
Strope, P. K., D. A. Skelly, S. G. Kozmin, G. Mahadevan, E. A. Stone *et al.*, 2015 The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome Res. 25: 762-774.
Sudmant, P. H., T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov *et al.*, 2015 An integrated map of structural variation in 2,504 human genomes. Nature 526: 75-81.
Uetz, P., L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson *et al.*, 2000 A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623-627.
Washburn, M. P., D. Wolters, and J. R. Yates 3rd, 2001 Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19: 242-247.
Wang, Y., C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag *et al.*, 2002 Precision and functional specificity in mRNA decay. Proc. Natl. Acad. Sci. U. S. A. 99: 5860-5865.
Winzeler, E. A., D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson *et al.*, 1999 Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285: 901-906.
Zhu, Y. O., G. Sherlock, and D. A. Petrov, 2016 Whole genome analysis of 132 clinical Saccharomyces cerevisiae strains reveals extensive ploidy variation. G3 6: 2421-2434.

Heritable variation of mRNA decay rates in yeast

Jennifer M. Andrie,[1] Jon Wakefield,[2] and Joshua M. Akey[1]

[1]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA;* [2]*Department of Statistics, University of Washington, Seattle, Washington 98195, USA*

Gene expression levels are determined by the balance between rates of mRNA transcription and decay, and genetic variation in either of these processes can result in heritable differences in transcript abundance. Although the genetics of gene expression has been a subject of intense interest, the contribution of heritable variation in mRNA decay rates to gene expression variation has received far less attention. To this end, we developed a novel statistical framework and measured allele-specific differences in mRNA decay rates in a diploid yeast hybrid created by mating two genetically diverse parental strains. We estimate that 31% of genes exhibit allelic differences in mRNA decay rates, of which 350 can be identified at a false discovery rate of 10%. Genes with significant allele-specific differences in mRNA decay rates have higher levels of polymorphism compared to other genes, with all gene regions contributing to allelic differences in mRNA decay rates. Strikingly, we find widespread evidence for compensatory evolution, such that variants influencing transcriptional initiation and decay have opposite effects, suggesting that steady-state gene expression levels are subject to pervasive stabilizing selection. Our results demonstrate that heritable differences in mRNA decay rates are widespread and are an important target for natural selection to maintain or fine-tune steady-state gene expression levels.

[Supplemental material is available for this article.]

Variation in gene expression levels constitutes a significant source of phenotypic diversity among individuals within populations and contributes to the evolutionary divergence between species (Skelly et al. 2009; Jones et al. 2012). In humans, regulatory variants affecting gene expression influence susceptibility to autoimmune, infectious, neoplastic, neurodegenerative, and psychiatric diseases (Skelly et al. 2009). In Darwin's finches, regulatory variation affecting beak morphology likely played a role in their speciation (Abzhanov et al. 2004). Likewise, gene expression variation underlies the skeletal morphology differences in stickleback fish that distinguish saltwater from freshwater species (Jones et al. 2012).

Heritable regulatory variation can broadly be classified as either acting in *cis* or *trans* (Skelly et al. 2009). While *trans*-regulatory effects on gene expression are undoubtedly important, studies in several eukaryotic organisms, including yeast, fruit flies, mice, rats, and humans, suggest that *cis*-regulatory effects constitute a substantially higher proportion of the genetic variance in gene expression within species than do *trans* effects (Schadt et al. 2003; Hughes et al. 2006; Petretto et al. 2006; Emilsson et al. 2008; Pickrell et al. 2010; Skelly et al. 2011). In the budding yeast *Saccharomyces cerevisiae*, for example, nearly 80% of the genes that have transcribed polymorphisms between two diverse strains exhibit allele-specific expression differences (Skelly et al. 2011). In humans, it has been estimated that ~90% of single nucleotide polymorphisms influencing gene expression levels are due to *cis*-regulatory mechanisms (Pickrell et al. 2010). Furthermore, *cis*-regulatory differences accumulate at a faster rate than *trans*-regulatory differences between closely related species (Wittkopp et al. 2008; Tirosh et al. 2009; Romero et al. 2012).

The balance between mRNA synthesis and decay determines steady-state levels of transcript abundance, and genetic variation affecting either of these processes can contribute to heritable differences in transcript abundance. However, to date, most research has concentrated on genetic variation affecting steady-state mRNA levels, failing to distinguish regulatory variation affecting transcription from that affecting decay (Skelly et al. 2009). Studies that have explored different classes of heritable variation underlying differences in steady-state gene expression focus primarily on transcription initiation, cataloging variation both within and between species in transcription factor binding sites, chromatin structure, and DNA methylation sites (Gerstein et al. 2010; The modENCODE Consortium et al. 2010; The ENCODE Project Consortium 2012; Connelly et al. 2014). In contrast, regulatory variants underlying differences in mRNA decay rate have received considerably less attention (Dori-Bachash et al. 2011; Pai et al. 2012).

To better delimit the contribution of *cis*-regulatory variation to heritable differences in mRNA decay rates, we developed a novel statistical framework and measured allele-specific differences in decay in a diploid hybrid created from two genetically diverse strains of the budding yeast, *S. cerevisiae*. We demonstrate that allelic differences in mRNA decay rates are widespread, affecting the expression levels of nearly 31% of measurable genes. Interestingly, we observe that a significant proportion of changes in decay rate are coupled to opposing changes in transcriptional initiation, suggesting pervasive compensatory evolution to stabilize or fine-tune steady-state gene expression levels. Our results also provide insights into the mechanisms through which *cis*-regulatory variation acts to influence mRNA decay rates, highlighting an important role for variants that affect mRNA secondary structure.

Results

Overview of experimental design

We measured rates of allele-specific mRNA decay (ASD) in a diploid yeast produced by mating two genetically diverse haploid

Corresponding author: akeyj@uw.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.175802.114.

24:2000–2010 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/14; www.genome.org

Saccharomyces cerevisiae strains: the laboratory strain BY4716 (BY), which is isogenic to the reference sequence strain S288C, and the wild Californian vineyard strain RM11-1a (RM) (Liti et al. 2009). A schematic of the experimental design is shown in Figure 1. Briefly, we introduced *rpb1-1*, a temperature-sensitive mutation in an RNA polymerase II subunit encoded by the gene *RPO21*, to each of the haploid yeast strains, mated the strains, and grew the resulting hybrid diploid to mid-log phase at 24°C, before rapidly shifting the culture to 37°C to inhibit transcription (Fig. 1A; Nonet et al. 1987). RNA-seq was performed on culture samples taken at 0, 6, 12, 18, 24, and 42 min subsequent to the temperature shift (Fig. 1A). To identify ASD, we used transcribed polymorphisms to distinguish between parental transcripts and compared the relative levels of transcript abundance over the time course (Fig. 1B). Note, this experimental design internally controls for *trans*-acting regulatory variation as well as environmental factors. Under the null hypothesis of no ASD, the proportion of reads from the BY transcript ($p_{BY} = \frac{N_{BY}}{N_{BY}+N_{RM}}$) observed over the time course remains unchanged (Fig. 1B). However, genes with ASD will exhibit an increasing or decreasing proportion of BY reads as a function of time (Fig. 1B). In total, we measured ASD from three independent biological replicates.

Statistical modeling of allele-specific mRNA decay

We developed a novel statistical framework to identify ASD. In brief, we use a linear logistic model to measure the change in the proportion, p_{BY}, of reads derived from the BY transcript as a function of time. To assess statistical significance, we use a Bayesian hierarchical Markov chain Monte Carlo model (see Methods). In this model, the prior probability of the alternative hypothesis (i.e., that a gene exhibits ASD) is determined from the totality of data. We also estimate the mean and variance of the decay rate differences under the alternative hypothesis from the data (see Methods). The primary motivation for developing this more sophisticated framework is that it accounts for genes that exhibit small departures from nonconstancy due to high read counts in a more principled manner than alternative approaches (Dori-Bachash et al. 2011). We evaluated the power and operating characteristics of our statistical framework through extensive simulations and found that it generally has higher power and more accurately estimates π_0 (the proportion of genes consistent with the null hypothesis of no allelic differences in mRNA decay) compared to alternative approaches under a wide variety of parameters (see Supplemental Methods and Results; Supplemental Fig. S1; Supplemental Table S1).

Pervasive influence of *cis*-regulatory variation on mRNA decay rates

Through a careful filtering pipeline, which included whole-genome sequencing of RM to mitigate read mapping bias (Degner et al. 2009; see Methods), we identified 27,569 transcribed single nucleotide variants (SNVs) in 4381 genes that could be used to assign whether individual RNA-seq reads derived from the BY or the RM allele of each gene. Of the ~222 million RNA-seq reads we obtained across all replicates and all time points in our study, 13.57 million reads, averaging 2.26 ± 0.65 million reads per time point, were informative, such that they mapped to a variant site and could unambiguously be assigned as originating from BY or RM.

We applied the statistical inference framework described above to 3544 genes that passed our filtering criteria (see Methods). From the Bayesian hierarchical MCMC model, we estimated $1 - \pi_0$, the proportion of genes that exhibit ASD, to be 0.31. Thus, ~31% of all measured genes are inferred to be inconsistent with the null hypothesis and exhibit allelic differences in decay rates. Of these, we can identify 350 genes at a false discovery rate of 10% (Fig. 2A). Note, this corresponds to a false nondiscovery rate of 24%. The set of genes called significant agrees well with a simpler approach of correcting for multiple testing by the QVALUE software (Storey 2002; Storey and Tibshirani 2003; Storey et al. 2004) and imposing a threshold on the magnitude of effect sizes needed to be called as significant (see Supplemental Methods and Results; Supplemental Fig. S2). We note that, in order to inhibit transcription, we subjected the yeast to mild heat shock. The decay rates observed in our data set are specific to the environmental condition that they were measured in and, thus, may be different in other states, such as log phase growth (Sun et al. 2012). Additionally, in theory, some of the differences in decay rate that we measured could be due to allele-specific transcriptional responses to heat shock, since there can be small amounts of leaky transcription in the first 5–15 min following the temperature shift (Nonet et al. 1987); however, such heat shock-induced differences

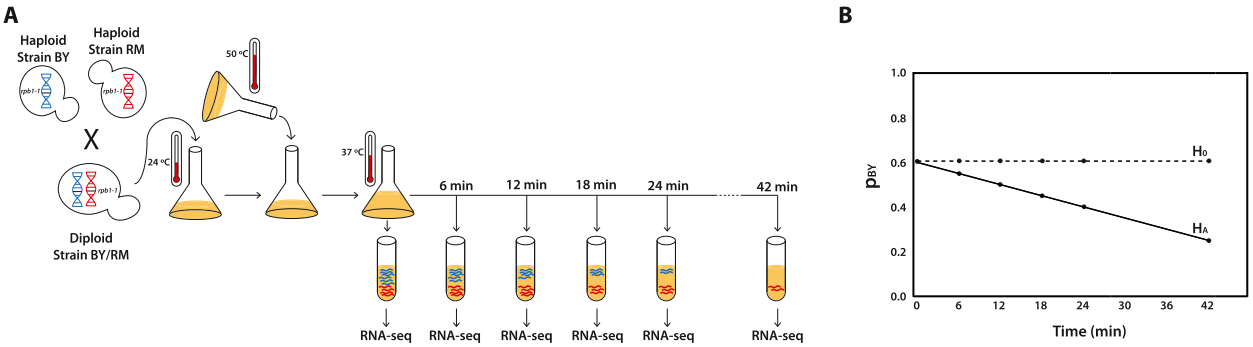


Figure 1. Overview of experimental design. (*A*) We replaced the wild-type allele of the *RPO21* (also known as *RPB1*) gene with the temperature-sensitive *rpb1-1* allele in both the BY4716 (BY) and the RM11-1a (RM) strains of *S. cerevisiae* (Nonet et al. 1987). We mated these two haploid strains to produce a diploid hybrid and grew the diploid to mid-log phase at the permissive temperature of 24°C. We rapidly shifted the temperature of the culture to 37°C, halting transcription. Immediately following the temperature shift, and at 6, 12, 18, 24, and 42 min after the shift, we isolated mRNA and performed RNA-seq. (*B*) By quantifying the relative levels of the BY and RM alleles for each gene, we estimated p_{BY}, the proportion of transcripts derived from BY, at each time point. Under the null hypothesis (H_0; dashed line) of no allele-specific differences in mRNA decay rates, p_{BY} remains constant. Under the alternative hypothesis (H_A; solid line) of allelic differences in mRNA decay, we expect p_{BY} to change as a function of time. For the gene represented by the solid line in the example pictured, p_{BY} decreases significantly over time, indicating that the BY allele is decaying more quickly than the RM allele of this gene.

in allele-specific transcription are unlikely to comprise most of the differences we observe.

The exponential of the slope of the linear logistic model fit to each gene is a direct estimate of the difference in mRNA decay rate between the BY and RM alleles of that gene (see Methods). Among the genes with significant ASD, the effect size of the decay rate difference ranges from a 1.81×10^{-3} to a 5.62×10^{-2} change in the odds of observing an mRNA allele of the BY strain given a 1-min increase in time, with a median difference of 1.01×10^{-2} (Fig. 2B). This median difference corresponds to an ~83% increase over 1 hr in the odds of observing an mRNA allele of the BY strain. The BY allele decays more quickly than the RM allele in 161 genes, while the RM allele decays more quickly than the BY allele in 189 genes. Genes with allelic differences in mRNA decay rates spanned a broad range of gene ontology terms, and we did not detect significant enrichment for particular functions or biological processes after correcting for multiple comparisons.

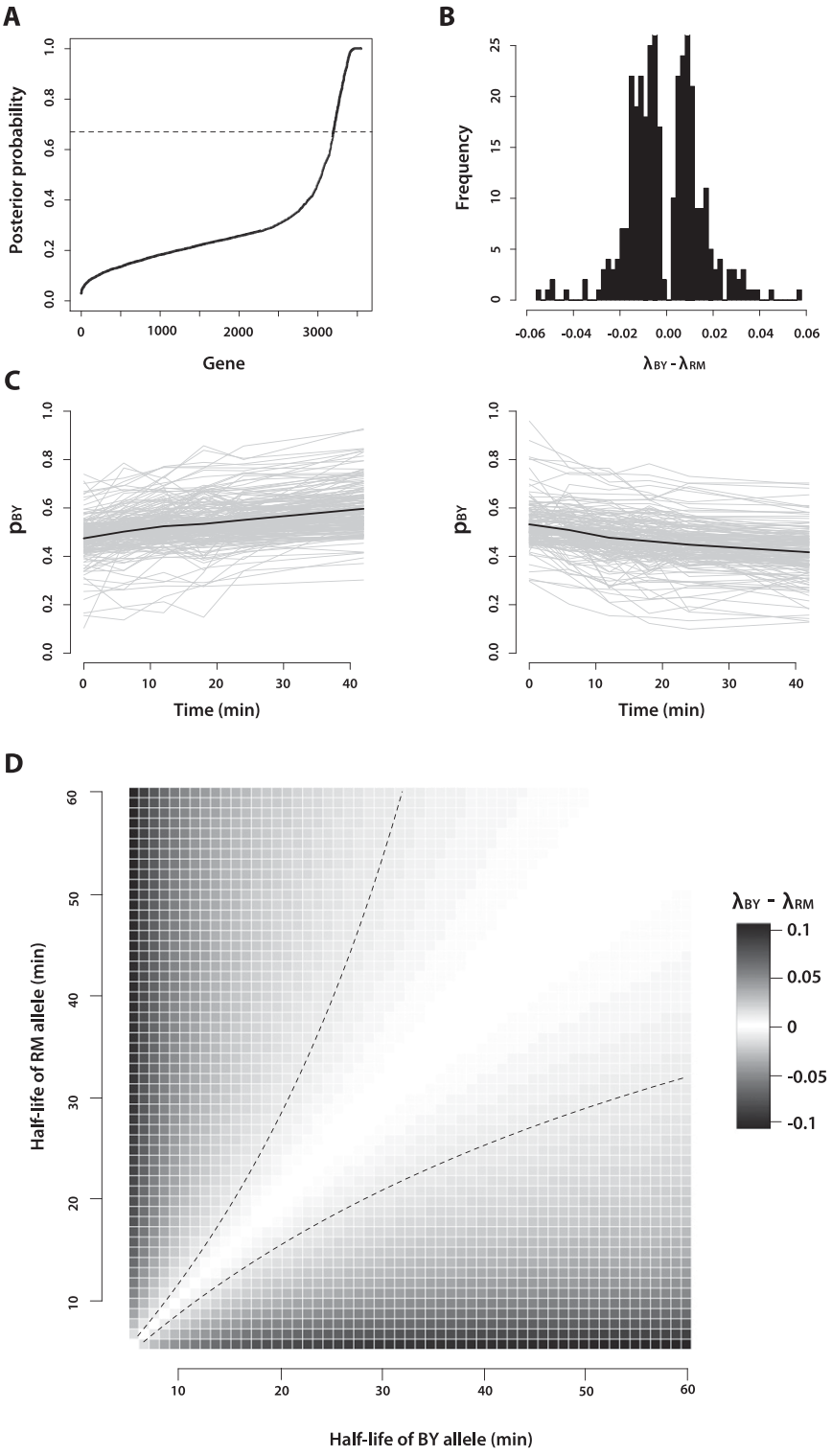


Figure 2. (Legend on next page)

Allelic differences in mRNA decay reveal widespread compensatory evolution

To investigate the relationship between ASD and steady-state gene expression levels, we first inferred allele-specific expression (ASE) at the 0-min time point in our time course, which is a reasonable proxy for steady-state levels of transcript abundance. Using the method developed by Connelly et al. (2014), we find that 1137 genes exhibit ASE (posterior probability > 0.95) (Fig. 3). Five hundred and ninety-five of the 1137 genes (52.3%) that exhibit steady-state ASE have higher levels of the RM transcript, and 542 genes (47.7%) have higher levels of the BY transcript. The median $\log_2$-fold change for all genes with allele-specific steady-state expression differences is 0.43.

Of the 350 genes with significant ASD, 182 (52.0%) also exhibit ASE (Fig. 3). Strikingly, of the 182 genes with both ASD and ASE, 129 (70.9%) have increased decay rates in the allele with higher levels of steady-state expression, suggesting that there are variants influencing rates of transcriptional initiation with opposite effects to those influencing decay (Fig. 3). Similarly, the 168 genes that exhibit ASD but not ASE (Fig. 3) are also likely enriched for variants with opposing effects on transcriptional initiation and decay, since the difference in decay rate does not produce a corresponding difference in steady state levels. Thus, these data suggest that changes in mRNA decay rates in yeast are often coupled with opposite changes in transcription, consistent with pervasive compensatory evolution to stabilize or fine-tune steady-state gene expression levels.

Patterns of genetic diversity across transcripts with allelic differences in mRNA decay

Previous studies in yeast have found that genes with ASE exhibited higher levels of genetic diversity compared to those without

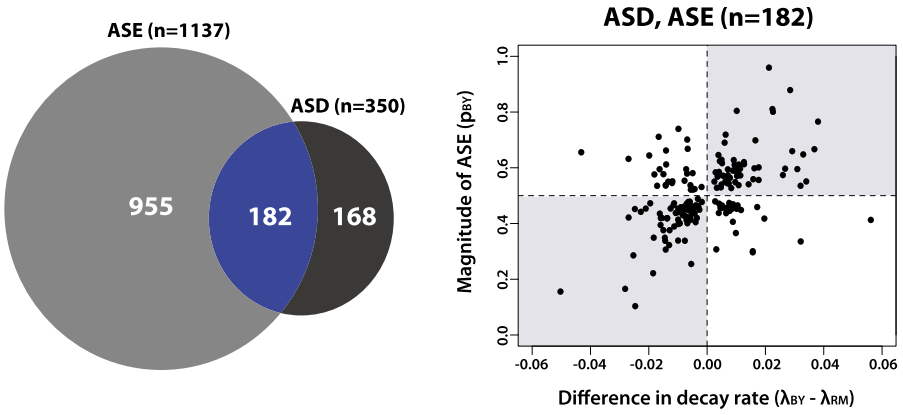


Figure 3. mRNA decay rates in yeast are often coupled to opposite changes in transcription initiation. (*Left*) Venn diagram showing the overlap of genes with significant allelic differences in steady-state gene expression (ASE) and decay (ASD). (*Right*) Scatter plot showing estimates of differences in decay rates between BY and RM (*x*-axis) versus the proportion of transcripts from the BY allele (p_{BY}) inferred from the 0-min time point for genes with both ASE and ASD. The shaded gray rectangles represent quadrants where magnitudes of ASD and ASE are discordant, suggesting compensatory evolution.

such differences (Ronald et al. 2005). To explore patterns of diversity in our data set, we first compared overall levels of variation among four classes of genes: those with only ASD, those with only ASE in steady-state levels, those with both ASD and ASE, and those with no allele-specific differences. Limiting our analysis to the 2954 genes with reliable UTR annotations (Nagalakshmi et al. 2008), we found that genes with any type of allele-specific difference have 1.4-fold more variants than genes without ASD or ASE (4.62 and 3.32 variants/kb, respectively; Mann-Whitney *U* test, *P*-value < 2.20×10^{-16}) (Fig. 4). Moreover, genes with ASD have 1.3-fold higher levels of variation compared to genes with only ASE (5.48 variants/kb compared to 4.31 variants/kb, respectively; Mann-Whitney *U* test, *P*-value = 2.26×10^{-10}) (Fig. 4). One complication in interpreting these findings is that genes with larger numbers of variants tend to have more informative reads, and, therefore, there is greater power to detect allelic differences in expression and decay. Indeed, variant density is significantly correlated with the number of informative reads ($r^2 = 0.158$, *P*-value < 2.20×10^{-16}). To more formally explore whether differences in SNV density are simply related to power, we performed logistic regression where the predictor variables were coded as zero if a gene did not show ASD and one if it did. We found that a model that included both the number of variants/kb and the number of informative reads as covariates fits the data significantly better than a model using just the number of informative reads alone (ANOVA *P*-value < 2.20×10^{-16}), suggesting that the increased levels of variation in genes with ASD and ASE are not simply a consequence of discovery bias.

To identify regions that may be enriched for variants that influence decay rates, we compared levels of genetic variation in the 5′ UTR, coding region, and 3′ UTR among the four classes of genes described above. Levels of variation are significantly elevated across all genic regions for genes with ASD compared to genes with no allelic differences (Fig. 4). Genes with both ASD and ASE have 1.46-fold, 1.68-fold, and 1.74-fold higher levels of variation than genes with no allelic differences in the 5′ UTR, coding region, and 3′ UTR, respectively (Mann-Whitney *U* test, *P*-value = 1.43×10^{-2}, < 2.20×10^{-16}, and 1.47×10^{-4}). Genes with ASD only have 1.88-fold, 1.64-fold, and 2.12-fold higher levels of variation than genes with no allelic differences in the 5′ UTR, coding region, and 3′ UTR, respectively (Mann-Whitney *U* test, *P*-value = 1.58×10^{-7}, 1.12×10^{-15}, and 6.77×10^{-4}). Thus, allele-specific differences in mRNA decay rate are likely driven by variants positioned throughout the transcript. Consistent with this hypothesis, genes that only contain SNVs in either their coding region or UTR are less likely to exhibit ASD compared to genes with variants in both their coding region and UTR (Fisher's exact test, *P*-value = 1.35×10^{-10} and 1.55×10^{-4}, respectively).

Genes with ASD are enriched for variants that influence mRNA structure

To test the hypothesis that variation in mRNA secondary structure contributes to allelic differences in mRNA decay rates, we compared the minimum Gibb's free energy (ΔG) associated with the predicted secondary structures of the BY and RM alleles for each mRNA transcript. Specifically, following standard practices (Tuller et al. 2010), we computed the ΔG of the predicted secondary structures for each of the 27,569 variants that we identified between BY and RM (Fig. 5A). We then calculated the absolute value of the difference in free energy between alleles, $|\Delta\Delta G| = |\Delta G_{BY} - \Delta G_{RM}|$, and for each gene, we recorded the maximum |ΔΔG| of all its variants. Variants with larger values of |ΔΔG| are predicted to have more severe structural consequences (Fig. 5A). Genes that exhibit ASD are enriched for variants with larger predicted effects on mRNA secondary structure as compared to genes without any allelic differences in decay or steady-state expression (1.32-fold increase in the maximum |ΔΔG| observed in genes with ASD; Mann-Whitney *U* test, *P*-value = 5.48×10^{-15}) (Fig. 5B). Although in silico predictions of differences in mRNA secondary structure are not perfect proxies for structures that occur in vivo, our observations are consistent with the hypothesis that allelic variation in mRNA secondary structure contributes to heritable variation in mRNA decay rates. Interestingly, the gene *HSP78*, which encodes a mitochondrial matrix chaperone, exhibits ASD and only contains a single variant (Supplemen-

Figure 2. Characteristics of genes that exhibit allele-specific mRNA decay. (*A*) Posterior probability that a gene exhibits allele-specific mRNA decay rates, as calculated from our Bayesian hierarchical Markov chain Monte Carlo model. The dashed line at posterior probability = 0.67 corresponds to the threshold we used to call genes as exhibiting significant allele-specific mRNA decay rates. (*B*) Histogram of the slope calculated from the linear logistic model for the 350 genes with significant (FDR = 10%) allele-specific mRNA decay rates. The exponential of the slope, which estimates $\lambda_{BY} - \lambda_{RM}$, is the change in the odds of observing a BY mRNA allele given a 1-min increase in time. (*C*) Decay rate time courses of all genes in which the RM allele decays significantly faster than the BY allele (*left*) and the BY allele decays significantly faster than the RM allele (*right*). The gray lines represent the decay rate time courses of each of the individual genes. The black lines represent the mean decay rate time courses for all of the genes included in each plot. (*D*) Correspondence of $\lambda_{BY} - \lambda_{RM}$ to half-life differences between the BY and RM alleles of a gene. The dashed black lines represent the positive and negative of the median effect size, where effect size is defined as $|\lambda_{BY} - \lambda_{RM}|$, observed among the genes we identified with significant allele-specific differences in decay rates.

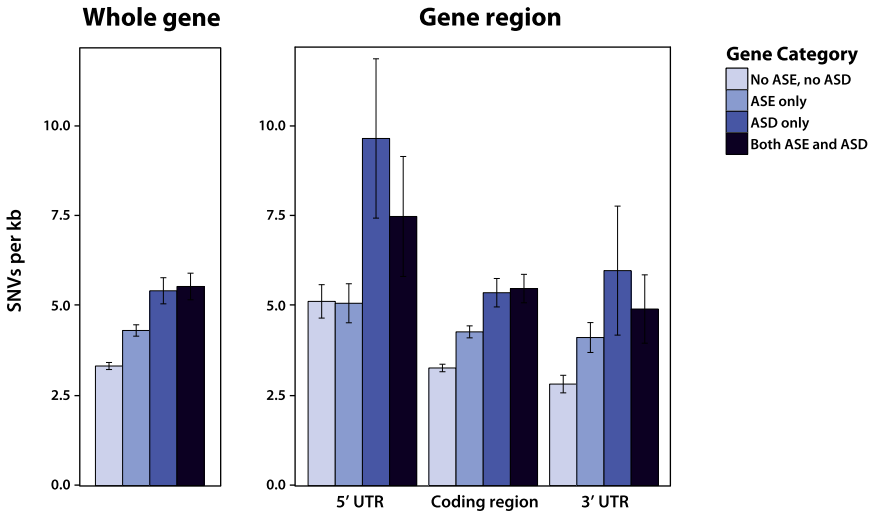


Figure 4. Levels of genetic diversity in genes with and without allele-specific differences in mRNA decay rates and steady-state expression levels. Bar plots show the mean number of single nucleotide variants (SNVs) between BY and RM per kilobase across the whole gene (*left*) or across each gene region separately (*right*). Error bars correspond to the 95% confidence interval of the mean. ASE and ASD denote allele-specific expression and allele-specific decay, respectively.

tal Table S2), which has a large predicted effect on mRNA secondary structure (Fig. 5C), suggesting that allelic differences in decay of this gene are likely mediated by structural differences.

No widespread evidence for coupling between decay rates and translational efficiency

Previous reports measuring mRNA decay rate in one or a few genes have suggested that the translation efficiency of an mRNA might be directly coupled to mRNA decay rate (for review, see Garneau et al. 2007 and Schoenberg and Maquat 2012). To evaluate whether such coupling is common, we compared the types of codon changes occurring between BY and RM in genes with and without decay rate differences. If mRNA decay rate is coupled to translation efficiency, we would expect that genes with ASD would have higher proportions of codon changes that impact translation efficiency, as compared to genes without such differences. To this end, we compared the proportion of preferred to unpreferred synonymous codon changes in genes with and without ASD and found that genes with ASD have a slightly lower proportion of codon preference changes, although this is not statistically significant (68.1% versus 68.7%; Mann-Whitney *U* test, *P*-value = 8.60×10^{-2}). Thus, we do not find widespread evidence that mRNA translation efficiency is directly coupled to mRNA decay rate. A caveat of this analysis is that more sensitive measures of translational efficiency may be needed to detect coupling. Moreover, these findings do not preclude the possibility that coupling exists for a subset of genes, which we may not have power to detect.

Discussion

We developed a novel statistical framework to measure allele-specific differences in mRNA decay rate in a diploid yeast hybrid created by mating two genetically diverse parental strains. A particular strength of our statistical approach is its ability to deal with genes that exhibit small departures from nonconstancy due to high read counts in a more principled manner than alternative approaches (Dori-Bachash et al. 2011). Using our statistical framework, we demonstrate the pervasive influence of *cis*-regulatory variation on mRNA decay rates, estimating that >30% of measurable genes exhibit ASD. Our results suggest that variation in mRNA decay rate is widespread across the genome, as well as among individuals within a species. Overall, our study provides further evidence of the importance of post-transcriptional processes in determining heritable differences in gene expression levels, which, in turn, impact phenotypic diversity among individuals within populations. Additionally, the novel statistical framework we developed has broad applications for future work in testing hypotheses of differential expression.

A striking feature of the data is that differences in mRNA decay rates are often coupled with opposite changes in transcription. It is difficult to precisely estimate the proportion of genes with significant ASD that is in the opposite direction of steady-state expression levels because of differences in the statistical power of detecting ASE and ASD. However, a naïve estimate suggests that up to 85% of genes with significant ASD are coupled with opposing effects on transcription (Fig. 3). These findings agree with previous studies, which observed that roughly 80% of differences between yeast species and 50% of differences among humans in mRNA decay rate are coupled to opposing differences in transcription (Dori-Bachash et al. 2011; Pai et al. 2012). Interestingly, in the remaining 15% of genes with significant ASD, there is no significant correlation ($r^2 = 0.255$, *P*-value = 6.53×10^{-2}) between the magnitude of the decay rate difference and the magnitude of the gene expression difference between the alleles. Collectively, these results suggest that steady-state gene expression levels are subject to strong stabilizing selection, and that heritable differences in mRNA decay rates are an important target for natural selection to maintain or fine-tune steady-state gene expression levels.

To explore which regions of the mRNA transcript are most important in determining mRNA decay rate differences, we compared the levels of genetic variation in the 5′ UTR, coding region, and 3′ UTR in genes with and without ASD. We hypothesized that the 3′ UTR would be the most important region governing ASD, and therefore, that genes exhibiting ASD would be especially enriched for polymorphisms between BY and RM in the 3′ UTR compared to genes without ASD or ASE. Instead, we observed that all three regions exhibited significantly more variation in genes with ASD compared to genes without ASD or ASE (Fig. 4). One explanation for these results is that the 3′ UTR contains the lowest overall amount of variation, suggesting that it is under significant functional constraint. If the 3′ UTR contains the highest density of *cis*-elements affecting mRNA decay rate, then changes to this region perhaps have a larger effect on mRNA decay rate, and therefore, are more likely to be removed by purifying selection. Conversely, changes in the 5′ UTR or coding region may cause differences in mRNA decay rates of a smaller effect size, and therefore be subject to less intense purifying selection. Thus, all three gene regions may be important determinants to within-species differences in mRNA decay rate. Our observation that genes with SNVs only in the coding region

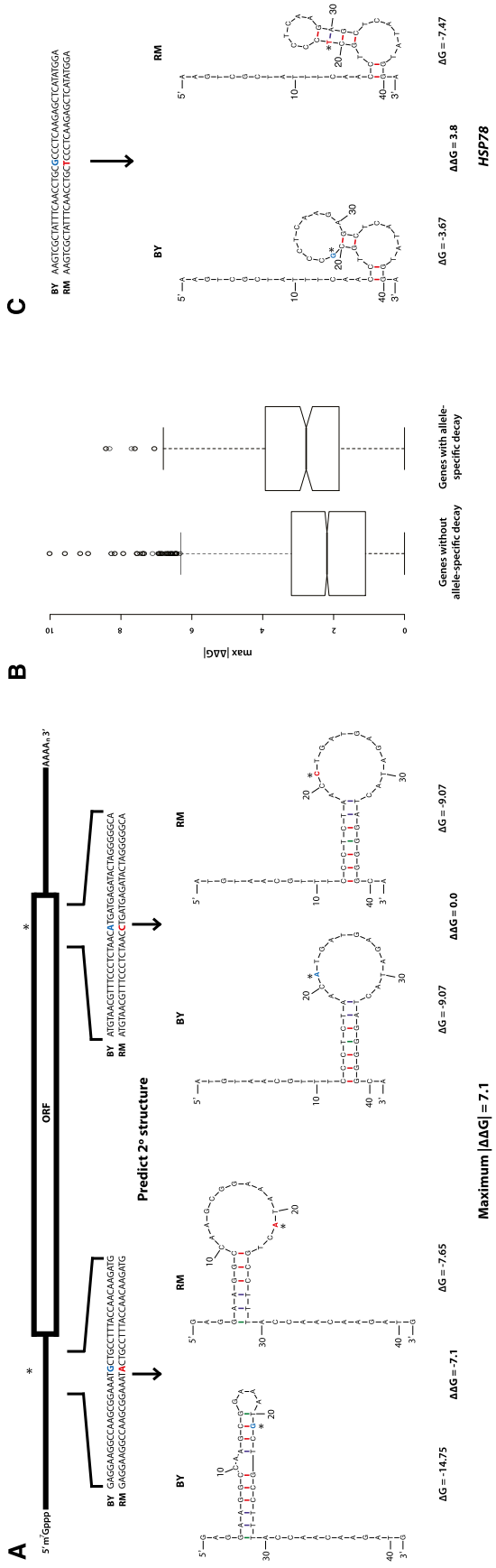


Figure 5. Genes with significant allele-specific decay are enriched for predicted differences in mRNA structure. (*A*) Calculation of ΔΔG for SNVs across a transcript. We determined the mRNA sequence for both the BY and RM alleles of each gene. For each SNV (denoted by *), we calculated the predicted minimum Gibb's free energy (ΔG) associated with the mRNA secondary structures of each allele using 41-bp regions of the mRNA transcript centered on each of the SNVs. The difference in ΔG (ΔΔG) between the BY and RM alleles for each 41-bp window was measured by subtracting the ΔG calculated for the RM allele from the ΔG calculated for the BY allele. We then calculated the absolute value of the difference in free energy between alleles, $|\Delta\Delta G| = |\Delta G_{BY} - \Delta G_{RM}|$, for each variant. For each gene, we recorded the maximum $|\Delta\Delta G|$ observed among all its variants. (*B*) The maximum $|\Delta\Delta G|$ of genes with and without allele-specific decay rate differences are significantly different (Mann-Whitney *U* test, *P*-value = 4.47×10^{-7}), suggesting that genes with allele-specific decay rates are significantly enriched for variants with larger predicted effects on mRNA secondary structure compared to genes without allelic differences in decay. (*C*) Predicted mRNA secondary structures surrounding an SNV in the gene *HSP78*, which encodes a mitochondrial matrix chaperone. The location of the SNV in each transcript allele is denoted by *; the BY and RM alleles at the variant site are colored blue and red, respectively. Note, *HSP78* only contains one SNV.

or, alternatively, only in the UTRs are less likely to exhibit ASD, also suggests that variation in all parts of the transcript can potentially impact decay rate.

Identifying allelic differences in mRNA decay rate is only the first step in the ultimate goal of identifying causal regulatory alleles and the mechanisms that they act through. To this end, it is interesting to note that 13 genes with ASD possess a single variant between the BY and RM alleles of the transcript (11 in coding regions and two in UTRs) (see Supplemental Table S2). These variants are strong candidate causal alleles, and as shown for *HSP78* (Fig. 5C), enable mechanistic hypotheses to be formulated and ultimately tested. Furthermore, it will be important to consider additional processes that could influence allele-specific decay. For example, nonsense-mediated decay (NMD) can be triggered by AUG codons in the 5′ UTR, and if a SNV introduced or disrupted a 5′ UTR AUG, it could influence mRNA decay rates between the two alleles. In the 2954 genes that have reliable UTR annotations, 34 contain SNVs in the 5′ UTR that introduce or disrupt an AUG codon. Of these 34 genes, nine exhibit significant ASD, which is significantly more than we would expect by chance (Fisher's exact test, *P*-value = 5.16×10^{-3}); however, only three of the nine genes show decay rate differences in the direction expected if allelic differences in mRNA decay rate were mediated by nonsense-mediated decay. Thus, this process likely makes a minor contribution to patterns of ASD in our data. More generally, dissecting the mechanistic basis of allelic variation in mRNA decay rates will facilitate the robust prediction of causal regulatory alleles from sequence data.

Another critical area of research will be exploring the interactions of genetic variation with the environment. Our study was conducted in yeast undergoing exponential growth in a rich medium that underwent mild heat shock at the time of transcriptional shut-off; however, we would expect that patterns of ASD would differ markedly under differing growth conditions, such as nutrient-limited media or the presence of high concentrations of chemicals like ethanol or the various heavy metals. Additionally, the effects of *cis*-regulatory variation on mRNA decay are also likely to vary between different stages of the yeast life cycle, including during meiosis and during vegetative growth. Another important limitation of our study is that it only examines allele-specific differences in mRNA decay rate between two diverse yeast strains. Nonetheless, our results highlight the important contribution that heritable variation in mRNA decay rates make to buffer steady-state differences in gene expression and suggest that additional post-transcriptional processes should be studied in greater detail for a more comprehensive understanding of mechanisms contributing to transcriptional diversity within and between species.

Methods

Yeast strains

For the purposes of this study, we replaced the wild-type copy of the *RPO21* (also known as *RPB1*) gene in the haploid *S. cerevisiae* strains BY4716 (BY) and RM11-1a (RM) (for detailed descriptions of these two strains, see Brem et al. 2002) with the *rpb1-1* mutant allele (Nonet et al. 1987). We began by identifying the mutations that make *rpb1-1* differ from wild-type *RPO21* via standard Sanger sequencing of the *RP021* locus in the strain Y262 (described in Herrick et al. 1990). We identified two mutations: a C to T substitution 206 bp after the translation start site, and a G to A substitution 4310 bp after the translation start site; both mutations are nonsynonymous. To substitute *RPO21* with *rpb1-1* in RM, we used a "pop-in, pop-out" strategy (Rothstein 1995; Duff and Huxley 1996). Specifically, in the "pop-in" step, we linearized a plasmid containing the *URA3* selectable marker and the *rpb1-1* mutant allele with a restriction enzyme that cut in the *rpb1-1* sequence segment. We then transformed the linearized plasmid into RM cells and selected for cells in which the plasmid had recombined into the genome using uracil prototrophy. At the completion of the "pop-in" step, RM carried a duplication of the target genomic sequence segment, in which one duplicate contained the *RPO21* wild-type allele and one duplicate contained the *rpb1-1* mutant allele; the plasmid sequences and *URA3* lay between the two duplicates. In the "pop-out" step, we added uracil back to the medium so that the *URA3* gene was no longer required for viability, allowing spontaneous recombination events to occur between the duplicated target sequences. To select for recombination events, we used 5-fluororotic acid (5-FOA), which is metabolized by URA3 into a toxic compound. A recombination event will result in either retention of the mutant *rpb1-1* allele or reversion to the wild-type *RPO21* allele. Using this strategy, we first replaced the C located 206 bp after the translation start site in *RPO21* with a T, and then, subsequently, we replaced the G located 4310 bp after the translation start site in *RPO21* with an A. To confirm successful substitution of the wild-type allele with the mutant allele at both sites, we used standard Sanger sequencing.

To replace *RPO21* with *rpb1-1* in BY, we employed a backcrossing strategy. We could not use the "pop-in, pop-out" strategy because BY already contained the *URA3* selectable marker. More specifically, we crossed BY to Y262, sporulated the hybrid diploid, and screened the resultant offspring for inability to grow at 37°C (Note: *rpb1-1* mutants do not replicate at this temperature). We then performed four more rounds of backcrossing between the hybrid offspring and BY, such that the resulting yeast strain carried the *rpb1-1* allele on an ~97% BY genetic background. We confirmed that the final product of our backcross carried the two single nucleotide variants that make *rpb1-1* mutants different from the wild-type *RPO21* by standard Sanger sequencing. We mated the BY and RM *rpb1-1* temperature-sensitive mutant strains and selected for the diploid hybrid by visually screening for BY and RM cells that had mated. We confirmed that our candidate diploid hybrids identified in our screen were, in fact, diploid using a standard Halo Mating Type Assay.

Measuring mRNA decay rates

mRNA decay rate time course sample collection

The BY × RM hybrid diploid we generated was grown at 24°C to mid-log phase (OD_{600} 0.8–1.0) in 60 mL of yeast extract peptone dextrose (YEPD). We abruptly shifted the culture to 37°C via addition of 60 mL of 50°C YEPD. Immediately following, and at 6, 12, 18, 24, and 42 min after addition of the 50°C medium, we collected 20-mL aliquots of the culture using vacuum filtration. To maintain the increased temperature of the culture, we housed it in a 37°C shaking incubator between collection time points. The collected yeast cells from each time point were flash-frozen in liquid nitrogen and then stored at −80°C for no more than 2 d before we extracted total RNA from the cells using a standard phenol-chloroform preparation. In total, we collected three replicates of our time course.

RNA sequencing

We used a TruSeq RNA Sample Prep v2 Kit (Illumina) to create a sequencing library from the total RNA collected for each decay rate time course time point from each replicate. Per the protocol

for the kit, we isolated mRNA from the total RNA using two rounds of poly(A) selection, then fragmented the isolated mRNA into ~150 base pair (bp) fragments, and finally, used random hexamer primers to produce cDNA. Poly(A) selection, by definition, retains only mRNA with intact poly(A) tails in the resultant RNA pool. The most commonly used pathway of mRNA decay, referred to as deadenylation-dependent decapping, begins with shortening of the poly(A) tail by deadenylases, followed by removal of the 5′ cap structure by decapping enzymes, and, finally, 5′ to 3′ exonucleolytic degradation of the decapped intermediate (for review, see Wilusz et al. 2001 and Garneau et al. 2007). Thus, our experimental design is unable to detect allelic differences that affect the decapping rate or rate of exonucleolytic degradation. We chose to use poly(A) selection to isolate mRNA despite its inability to detect differences in the later stages of mRNA decay because previous studies of deadenylation-dependent decapping have demonstrated that mRNA decay proceeds very rapidly following deadenylation, and that deadenylation, as opposed to decapping or exonucleolytic degradation, is the rate-limiting step in the mRNA decay process (for review, see Wilusz et al. 2001 and Garneau et al. 2007). We created barcoded sequencing libraries from the cDNA from each sample and, in an effort to minimize technical variation between the data acquired from different decay rate time points, all samples from all replicates were sequenced in the same lane on an Illumina HiSeq 2000 (50-bp paired-end reads).

Whole-genome sequencing of RM

For whole genome sequencing of the *S. cerevisiae* strain RM, we inoculated the strain from −80°C freezer stock into 5 mL YEPD and grew the culture at 30°C to saturation. We pelleted the cells from the culture by centrifugation, decanted the supernatant, and froze the cells at −80°C. We extracted DNA using a Genomic-tip 100/G Kit (Qiagen) and then concentrated the sample using a standard ethanol precipitation. We prepared a DNA sequencing library using a TruSeq DNA Sample Prep v2 Kit (Illumina). Per the protocol for the kit, we used a Covaris sonicator to shear the DNA into ~300- to 400-bp fragments, and, after ligating adaptors onto the DNA fragments, we additionally size-selected for 300- to 400-bp fragments by running the ligation products out on an agarose gel and gel-extracting the appropriate band. We performed whole-genome sequencing using an Illumina MiSeq (151-bp paired-end reads).

Read mapping

We obtained complete genome sequences for BY from the *Saccharomyces* Genome Database (version R64-1-1, released February 3, 2011; http://www.yeastgenome.org) (Engel et al. 2013) and for RM from the Broad Institute (http://www.broadinstitute.org). We used BWA version 0.5.9 (Li and Durbin 2009) to map both the DNA and RNA sequence reads to the BY genome and, separately, the RM genome. After mapping reads, we sorted BAM files and marked duplicate reads using Picard version 1.43 (http://picard.sourceforge.net).

Identification of variant sites for assigning the allele of individual RNA-seq reads

To obtain a set of variants for allele-specific read calling in the BY × RM diploid, we 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. We then used threaded blockset aligner (TBA) (Blanchette et al. 2004) to compute a whole-genome alignment that is not biased in favor of a particular reference genome. We cataloged all SNVs, as well as all indels, identified in the alignment. As we were only interested in transcribed differences between the BY and RM genomes, we removed from our variant list all sites not within annotated BY open reading frames (obtained from the *Saccharomyces* Genome Database; http://www.yeastgenome.org) and their corresponding untranslated regions (UTRs) (UTR lengths were determined from Nagalakshmi et al. 2008). Manual review of the remaining variant sites using the program Integrated Genome Viewer (http://www.broadinstitute.org/igv/) revealed that many of the indels identified from the BY and RM alignment produced by TBA, as well as the SNVs closely flanking these indels, were miscalled. Therefore, we removed all indels and all SNVs within 10 bp of an indel from our variant list. Likewise, because we suspected that most or all of the SNVs we identified in genes with unusually high numbers of variants per kb were artifacts of alignment errors, we discarded all SNVs located in genes that exhibited greater than fivefold the average variant density of all genes across the genome. Due to the difficulty in distinguishing which gene an RNA-sequencing read that aligns to a location in which two yeast genes overlap derives from, we also threw out any variants that overlapped more than one annotated yeast gene. Finally, we removed any variants to which reads obtained from whole genome sequencing of RM were assigned more often to the BY allele of the variant than to the RM allele by the method developed by Skelly et al. (2011) (briefly described below) to assign whether individual RNA-seq reads derived from the BY or the RM allele of each gene.

Assignment of the allele of individual RNA sequencing reads

We performed assignment of individual RNA-seq reads as originating from either the BY or the RM allele of each gene as described in Skelly et al. (2011), with the following two changes. First, any read with an alignment to one genome that scores higher had to overlap one of the SNVs between BY and RM that were identified as described above. Second, we did not perform a correction for GC content.

Measuring allele-specific differences in mRNA decay rate

To determine whether a gene exhibited allele-specific mRNA decay rate differences, we developed a novel linear logistic model that we applied in conjunction with a quasi-likelihood ratio test to measure the change across our time course in the calculated proportion of reads deriving from the BY allele as compared to the total number of informative reads at each gene. Specifically, in our model, we let $N_j(t)$ be the number of mRNA transcripts for strain j, $j = 1, 2$ (representing BY and RM) at time t. We then assumed that the rate of decay is $dN_j/dt = -\lambda_j t$, with $\lambda_j > 0$, so that $N_j(t) = N_{0j} \exp(-\lambda_j t)$, where N_{0j} is the count at time 0 for strain j. For each time point, t, the number of RNA-seq reads that we can assign to a strain, $n_j(t)$, is a fraction, f_t, of the total number of mRNA transcripts for that strain, such that $n_j(t) = f_t N_j(t)$.

We then assumed the model

$$n_j(t) \sim \text{Poisson}\left[f_t N_{0j} \exp\left(-\lambda_j t\right)\right].$$

Under this model, the distribution of the counts for strain 1 (BY) given the total is binomial [we could make a binomial approximation since $n(t) \ll N(t)$] with denominator $n_1(t) + n_2(t)$ and probability (of strain 1):

$$p(t) = \frac{f_t N_{01}\exp(-\lambda_1 t)}{f_t N_{01}\exp(-\lambda_1 t) + f_t N_{02}\exp(-\lambda_2 t)} = \frac{\frac{N_{01}}{N_{02}}\exp(-[\lambda_1 - \lambda_2]t)}{\frac{N_{01}}{N_{02}}\exp(-[\lambda_1 - \lambda_2]t) + 1}.$$

Taking the logit gives:

$$\log\left(\frac{p(t)}{1-p(t)}\right) = \log\left(\frac{N_{01}}{N_{02}}\right) - [\lambda_1 - \lambda_2]t = \alpha + \beta t.$$

Although different fractions of the total mRNA transcript pool are sampled at each time point, these fractions cancel in the above calculation, so that we compute the relative proportion of strain 1 (BY) alleles in the pool from which we have sampled at each time point. The $\exp(\alpha)$ is the odds that we observe an mRNA allele of strain 1 at time $t = 0$. In our linear logistic model, we estimate α; however, from the above derivation we know that these odds are N_{01}/N_{02}, but this proportion is unobserved. The parameter $\exp(\beta)$ is the change in the odds of observing an mRNA allele of the strain 1 type given a 1-min increase in time. Thus, $\exp(60 \times \beta)$ is the change in the odds of observing an mRNA allele of strain 1 given a 1-hr increase in time. For example, if $\exp(60 \times \beta) = 2$, then the odds of observing an mRNA allele of strain 1, when compared to the odds of observing an mRNA allele of strain 2 (RM), doubles over 1 hr. If decay rates are the same in both strains, then $\lambda_1 = \lambda_2$, which is equivalent to $\beta = 0$ in the logistic model. The parameter estimate $\widehat{\beta}_i$, along with the associated standard error, are subsequently used within a hierarchical model, as detailed shortly.

The null can be rejected with small departures from non-constancy due to high counts, if a frequentist test (such as a quasi-likelihood test) is used. This is a recognized problem with frequentist testing in which power is not accounted for in the setting of significance thresholds. Hence, to determine if β was significantly different from 0, and therefore, whether a gene exhibited allele specific differences in mRNA decay rate, we used a Bayesian hierarchical model. In our model, we let Y_i be the estimate of the slope $\widehat{\beta}_i$ for the ith gene, and σ_i^2 be the variance of this estimate. We then assumed $Y_i | \mu_i \sim_{ind} N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, m$, where m is the number of genes. We specified a mixture model for the collection $[\mu_1, \ldots, \mu_m]$, with

$$\mu_i = \begin{cases} 0 & \text{with probability } \pi_0 \\ \sim N(\delta, \tau^2) & \text{with probability } \pi_1 = 1 - \pi_0 \end{cases}.$$

The second mixture component contains the non-null genes. We integrated out over μ_i to obtain a three-stage model, and we use mixture component indicators $H_i = 0/1$ to denote the zero/normal membership model for transcript i. The model is:

Stage 1:

$$Y_i | H_i, \delta, \tau, \pi_0 \sim_{ind} \begin{cases} N(0, \sigma_i^2) & \text{if } H_i = 0 \\ N(\delta, \sigma_i^2 + \tau^2) & \text{if } H_i = 1 \end{cases}.$$

Stage 2:

$$H_i | \pi_1 \sim_{iid} \text{Bernoulli}(\pi_1), i = 1, \ldots, m.$$

Stage 3:

Independent priors on the common parameters :

$$p(\delta, \tau, \pi_0) = p(\delta)p(\tau)p(\pi_0)$$

with

$$p(\delta) \propto 1,$$

$$p(\tau) \propto 1/\tau,$$

$$p(\pi_0) = 1,$$

so that we had improper priors for δ and τ^2. This model is appealing since we deal with overdispersion in the data using a reliable and distribution-free frequentist method and then take the information on the parameter of interest only (the differences), namely the estimate and its associated standard error, to model within the hierarchy. By only concentrating on the key parameters, we avoid having to make model assumptions concerning parameters of no interest.

We implemented this model via a Markov chain Monte Carlo algorithm in which we introduced indicator variable ω_i to denote the mixture component of gene i. For our analysis, we only evaluated the 3544 genes that had at least 10 informative reads at each of the six time points, as well as less than a 50-fold difference in expression of the two alleles at the 0-min time point. General background of this testing framework can be found in Wakefield (2013).

To formally determine whether gene i exhibited allele-specific mRNA decay rate differences, we placed a threshold of 0.67 on the posterior probability $r_i = \Pr(H_i = 1 | \text{data})$ of being non-null. At this threshold, the false discovery rate (FDR) is 0.099 and the false nondiscovery rate (FNDR) is 0.244. The FDR and FNDR are model-based estimates and are calculated as follows. For the list of R (say) genes i that pass the threshold, we calculate the sum of $1 - r_i$ (i.e., the posterior probability of no difference in mRNA decay rate) and divide by the total number of "noteworthy" genes, R, to give the FDR. For all the remaining (3544 – R) non-noteworthy genes, we sum the r_i (i.e., the posterior probability of a difference in mRNA decay rate) and then divide by (3544 – R) to give the FNDR.

Gene Ontology analysis

To assess whether there was any significant enrichment for genes involved in a particular molecular function, cellular component, or biological process in the set of genes we identified with allelic differences in mRNA decay, or in the two smaller subsets of genes in which one allele or the other decayed more quickly, we submitted each set of genes to AmiGO's GO Term Enrichment Tool (http://amigo1.geneontology.org/cgi-bin/amigo/term_enrichment). We co-submitted all 3544 genes we analyzed for allele-specific mRNA decay rate differences as the input background set and selected SGD as the database filter. We chose 0.01 as our maximum P-value threshold and two as the minimum number of gene products.

Measuring allele-specific differences in mRNA steady-state levels

Using the numbers of mRNA transcripts from BY and RM for each gene at time point $t = 0$ min as a proxy for steady-state expression levels, we determined whether a gene exhibited allele-specific steady-state expression differences by performing the *cis* test exactly as described in Connelly et al. (2014), with the following modification: The test was performed with three, rather than two, replicates. Our primary motivation for choosing this method, as opposed to alternative approaches (Skelly et al. 2011), is that its statistical framework is most closely related to the framework we implemented for identifying allele-specific mRNA decay rate differences.

Classification of genes by type of allele-specific differences

For comparison between genes with allele-specific mRNA decay rate differences, allele-specific steady-state expression differences, and no allele-specific differences, we divided the genes into four classes: those with allele-specific differences in mRNA decay rate only, those with allele-specific differences in steady-state levels only, those with allele-specific differences in both mRNA decay rate and steady-state levels, and those with no allele-specific differences. Specifically, we categorized genes with a posterior probability greater than 0.67 in our Bayesian hierarchical Markov chain Monte Carlo model, but with a posterior probability less than 0.95 in our test for allele-specific steady-state expression differences as only having allele-specific differences in mRNA decay rate. We considered genes with a posterior probability greater than 0.67 in our Bayesian hierarchical Markov chain Monte Carlo model and a posterior probability greater than 0.95 in our test for allele-specific steady-state expression differences as exhibiting both allele-specific differences in mRNA decay rate and allele-specific differences in steady-state expression levels. Genes with a posterior probability greater than 0.95 in our test for allele-specific steady-state expression differences, but which did not have a posterior probability greater than 0.67 in our Bayesian hierarchical Markov chain Monte Carlo model, were classified as only having allele-specific differences in steady-state expression levels. For our final category of genes with no allelic differences, we grouped together genes with a posterior probability less than 0.30 in our Bayesian hierarchical Markov chain Monte Carlo model and a posterior probability less than 0.95 in our test for allele-specific steady-state expression differences. We choose 0.30 rather than 0.67 as the cut-off for the posterior probability in our Bayesian hierarchical Markov chain Monte Carlo model for this group of genes in an effort to minimize the number of false negatives (for allele-specific differences in mRNA decay rate) in this group.

Secondary structure analysis

To evaluate the differences in mRNA secondary structure between the BY and RM alleles of each gene, we began by determining the mRNA sequence for both the BY and RM alleles of each gene using the set of 27,569 variants we identified between BY and RM, the BY genome sequence (from the *Saccharomyces* Genome Database, version R64-1-1, released February 3, 2011; http://www.yeastgenome.org) (Engel et al. 2013), annotations of the BY open reading frames (from the *Saccharomyces* Genome Database; http://www.yeastgenome.org), and the predicted untranslated region lengths (UTRs) for the BY open reading frames (Nagalakshmi et al. 2008). We then used the UNAFold software package's hybrid-ss-min tool to compute the predicted minimum Gibb's free energy (ΔG) associated with the mRNA secondary structures of each allele of each transcript at 30°C (we chose to use 30°C because this is the standard temperature at which yeast are grown in the laboratory) (Markham and Zuker 2005, 2008). More specifically, following standard practices (Tuller et al. 2010), we calculated the ΔG of a 41-bp mRNA region surrounding each of the 27,569 SNVs in our data set (Markham and Zuker 2005, 2008). The variant of interest was placed at the center of each 41-bp window; however, if the variant was <20 bp from the end of the mRNA transcript, the 41-bp window was shifted such that the beginning or the end coincided with the beginning or the end of the mRNA transcript, as appropriate. The difference in ΔG (ΔΔG) between the BY and RM alleles for each 41-bp window was measured by simply subtracting the ΔG calculated for the RM allele from the ΔG calculated for the BY allele. We then calculated the absolute value of the difference in free energy between alleles, $|\Delta\Delta G| = |\Delta G_{BY} - \Delta G_{RM}|$, for each variant. For each gene, we recorded the maximum $|\Delta\Delta G|$ we observed among all its variants.

Data access

Sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE60617.

Acknowledgments

We thank Jenny Madeoy and Marnie Johansson for expert technical assistance in strain construction, the Northwest Genomics Center for sequencing, and members of the Akey laboratory for helpful discussions and feedback on the manuscript. This work was supported by NIH grants 5R01GM094810 and 1R01GM098360 to J.M.A.

References

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ. 2004. *Bmp4* and morphological variation of beaks in Darwin's finches. *Science* **305:** 1462–1465.
Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14:** 708–715.
Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **436:** 701–703.
Connelly CF, Wakefield J, Akey JM. 2014. Evolution and genetic architecture of chromatin accessibility and function in yeast. *PLoS Genet* **10:** e1004427.
Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25:** 3207–3212.
Dori-Bachash M, Shema E, Tirosh I. 2011. Coupled evolution of transcription and mRNA degradation. *PLoS Biol* **9:** e1001106.
Duff K, Huxley C. 1996. Targeting mutations to YACs by homologous recombination. *Methods Mol Biol* **54:** 187–198.
Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al. 2008. Genetics of gene expression and its effect on disease. *Nature* **452:** 423–428.
The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.
Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, *et al*. 2013. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3* **4:** 389–398.
Garneau NL, Wilusz J, Wilusz CJ. 2007. The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* **8:** 113–126.
Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330:** 1775–1787.
Herrick D, Parker R, Jacobson A. 1990. Identification and comparison of stable and unstable mRNAs in *Saccharomyces cerevisiae*. *Mol Cell Biol* **10:** 2269–2284.
Hughes KA, Ayroles JF, Reedy MM, Drnevich JM, Rowe KC, Ruedi EA, Cáceres CE, Paige KN. 2006. Segregating variation in the transcriptome: *cis* regulation and additivity of effects. *Genetics* **173:** 1347–1355.
Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484:** 55–61.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25:** 1754–1760.
Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458:** 337–341.
Markham NR, Zuker M. 2005. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res* **33:** W577–W581.
Markham NR, Zuker M. 2008. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol* **453:** 3–31.
The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330:** 1787–1797.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.
Nonet M, Scafe C, Sexton J, Young R. 1987. Eucaryotic RNA polymerase conditional mutant that rapidly ceases mRNA synthesis. *Mol Cell Biol* **7:** 1602–1611.
Pai AA, Cain CE, Mizrahi-Man O, De Leon S, Lewellen N, Veyrieras JB, Degner JF, Gaffney DJ, Pickrell JK, Stephens M, et al. 2012. The

contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet* **8:** e1003000.
Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M, et al. 2006. Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet* **2:** e172.
Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464:** 768–772.
Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet* **13:** 505–516.
Ronald J, Brem RB, Whittle J, Kruglyak L. 2005. Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet* **1:** e25.
Rothstein R. 1995. Targeting, disruption, replacement, and allele rescue: integrative DNA transformation in yeast. In *Guide to yeast genetics and molecular biology* (ed. Guthrie C, Fink GR), pp. 281–301. Academic Press, San Diego.
Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422:** 297–302.
Schoenberg DR, Maquat LE. 2012. Regulation of cytoplasmic mRNA decay. *Nat Rev Genet* **13:** 246–259.
Skelly DA, Ronald J, Akey JM. 2009. Inherited variation in gene expression. *Annu Rev Genomics Hum Genet* **10:** 313–332.
Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. 2011. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res* **21:** 1728–1737.
Storey JD. 2002. A direct approach to false discovery rates. *J R Stat Soc Ser B Methodol* **64:** 479–498.
Storey JD, Tibshirani R. 2003. Statistical significance for genome-wide experiments. *Proc Natl Acad Sci* **100:** 9440–9445.
Storey JD, Taylor JE, Siegmund D. 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc Ser B Methodol* **66:** 187–205.
Sun M, Schwalb B, Schulz D, Pirkl N, Etzold S, Larivière L, Maier KC, Seizl M, Tresch A, Cramer P. 2012. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res* **22:** 1350–1359.
Tirosh I, Reikhav S, Levy AA, Barkai N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324:** 659–662.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci* **107:** 3645–3650.
Wakefield J. 2013. Hypothesis testing and variable selection. In *Bayesian and frequentist regression methods*, pp. 157–200. Springer, New York.
Wilusz CJ, Wormington M, Peltz SW. 2001. The cap-to-tail guide to mRNA turnover. *Nat Rev Mol Cell Biol* **2:** 237–246.
Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat Genet* **40:** 346–350.

Received March 18, 2014; accepted in revised form September 17, 2014.

Heritable variation of mRNA decay rates in yeast

Jennifer M. Andrie, Jon Wakefield and Joshua M. Akey

Genome Res. 2014 24: 2000-2010 originally published online September 25, 2014
Access the most recent version at doi:10.1101/gr.175802.114

SUPPLEMENTARY MATERIAL

Supplementary Materials and Methods

Determining the appropriate statistical framework for our data set through simulations

In our statistical framework, one of the key inputs for determining whether a gene exhibits allele-specific mRNA decay rate is the variance in the estimates of the proportion, p_{BY}, of reads derived from the BY transcript across time points (see Methods). It has been suggested in the past that the variances in RNA-seq data sets might be better analyzed if it is assumed they are overdispersed (Anders and Huber 2010; Marioni *et al*. 2008). In our statistical framework, rather than a binomial model, we could allow for excess-binomial variation by taking a quasi-likelihood approach (McCullagh and Nelder 1989) with

$$\mathrm{E}[N_1(t)] = [N_1(t) + N_2(t)]\, p(t)$$

$$\mathrm{var}(N_1(t)) = k[N_1(t) + N_2(t)]\, p(t)(1 - p(t))$$

where k is a parameter that is estimated and allows for overdispersion.

Though we did not expect to find a high degree of overdispersion in our data set because we were estimating variance in a proportion of one allele over a time course rather than the variance in raw read counts between technical replicates (and, indeed, the median estimate of overdispersion for all genes in our data set is 0.927, while the mean is 1.247), we decided to further explore the effect of assuming overdispersion on our data analysis framework. Therefore, we simulated data sets in which the read counts from the genes derived from either a standard binomial distribution or, to allow for overdispersion, a beta-binomial distribution. More specifically, under the beta-binomial distribution, we

performed three sets of simulations: one in which the beta-binomial parameters *a* and *b* corresponded to an overdispersion parameter, *k*, of 1.25; one in which *a* and *b* corresponded to a *k* of 10; and one in which *a* and *b* corresponded to a *k* of 25. The beta-binomial parameters *a* and *b* can be determined from the desired amount of overdispersion *k*, as well as p_{BY} and the coverage at the simulated gene, *N*, by the following equations:

$$a = \frac{p_{BY}\,(N-k)}{(k-1)}$$

$$b = \frac{(1-p_{BY})\,(N-k)}{(k-1)}$$

Under all distributions, the mean $p_{BY} = 0.5$ for all time points for genes simulated under the null hypothesis. For genes simulated under the alternative hypothesis, the mean $p_{BY} = 0.5$ at the 0 minute time point, and the mean p_{BY} at subsequent time points was calculated to correspond to the median effect size observed for genes identified to exhibit allele-specific decay in our real data set. Finally, in our simulated data sets, the coverage for each gene was sampled from a Poisson distribution with $\lambda = 331.5$, where 331.5 is the median coverage per gene in our real data set. Using the statistical framework we developed, we analyzed each simulated data set either with or without assuming the data were overdispersed, by calculating the variance either under a standard binomial distribution or under a quasibinomial distribution, respectively. For comparison with our Bayesian method, we also assessed statistical significance using a likelihood ratio test and corrected for multiple testing with the QVALUE software (Storey 2002; Storey and Tibshirani 2003; Storey *et al*. 2004).

Identifying allele-specific differences in mRNA decay rate via a frequentist test

In addition to using a Bayesian hierarchical model (see Methods) to identify genes with significant allele-specific differences in mRNA decay rate, we also tested whether the parameter estimate $\widehat{\beta}_t$, which we obtained from our linear logistic model (see Methods), differed from zero (i.e. the gene exhibited allele-specific mRNA decay rate) using a likelihood ratio test. We corrected for multiple testing with the QVALUE software (Storey 2002; Storey and Tibshirani 2003; Storey *et al*. 2004). Using this approach, the null can be rejected with small departures from non-constancy due to high read counts. Therefore, we imposed a threshold of 0.004 on the magnitude of the change in the odds of observing an mRNA allele of the BY strain given a one minute increase in time; all genes with an effect size lower than 0.004 were discarded from the set of genes we identified at FDR = 10% as exhibiting allele-specific mRNA decay rate. We chose this threshold by comparing the distributions of the magnitude of the change in the odds of observing an mRNA allele of the BY strain given a one minute increase in time between genes with $q\text{-}value < 0.10$ and $q\text{-}value > 0.10$ (Figure S2A).

Supplementary Results

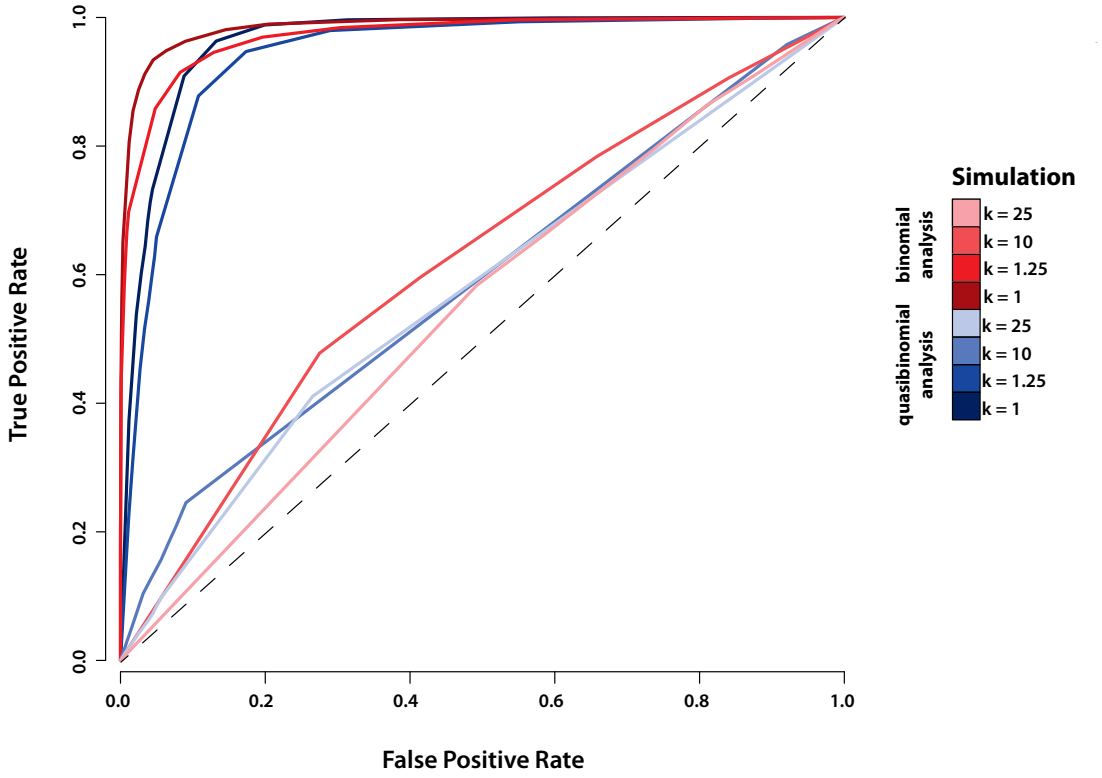
Our novel statistical framework with no overdispersion outperforms other methods of assessing statistical significance

From a standard binomial distribution, as well as from beta-binomial distributions corresponding to k = 1.25, 10, and 25, we simulated ten data set replicates, each of ten-thousand genes, for each π_0 between 0.1 and 0.9, incrementing by 0.2 (see Table S1). We found that if binomially-distributed data are analyzed under the assumption of no overdispersion, our newly-developed statistical framework more accurately calculates π_0 and better identifies individual genes as exhibiting allele-specific mRNA decay rate than the likelihood ratio test corrected for multiple testing with the QVALUE software (Figure S1 and Table S1). However, if binomially-distributed data are analyzed under an assumption of overdispersion, both methods of assessing statistical significance underestimate π_0 and over-identify individual genes as exhibiting allele-specific mRNA decay rate (Figure S1 and Table S1). Interestingly, for moderately overdispersed data (k = 1.25), it is also more accurate not to assume overdispersion when using our model (Figure S1 and Table S1). When using the likelihood ratio test corrected for multiple testing with the QVALUE software on moderately overdispersed data, it is only marginally better to assume overdispersion (Figure S1 and Table S1). If the data is more severely overdispersed (k = 10 or 25), both our newly-developed statistical framework and the likelihood ratio test corrected for multiple testing with the QVALUE software perform very poorly regardless of whether or not the method assumes overdispersion (Figure S1 and Table S1).

Gene sets identified as exhibiting allele-specific differences in mRNA decay rate via two different methods show a high degree of overlap

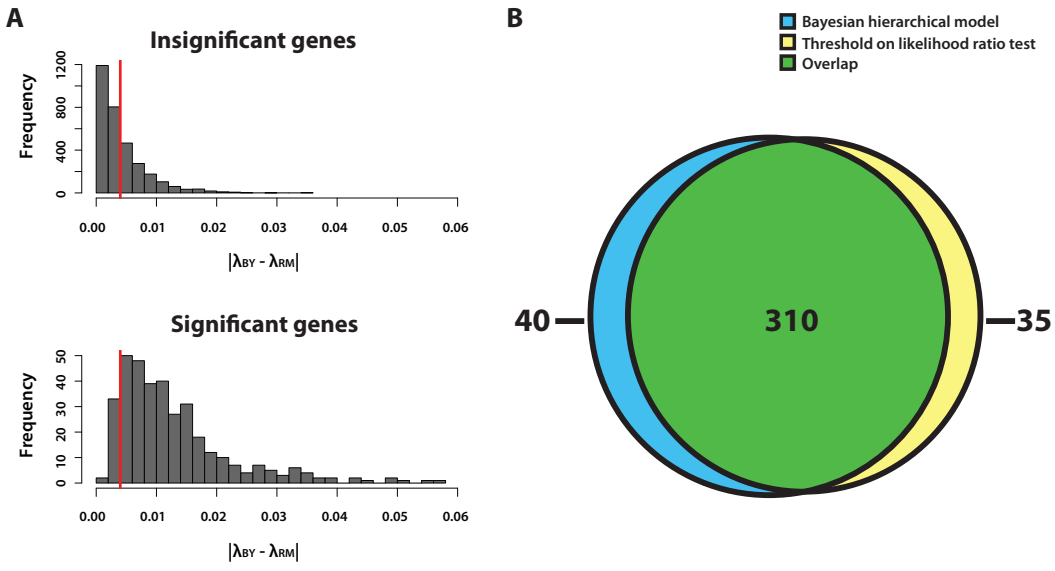
To identify genes exhibiting allele-specific differences in mRNA decay rate, we used a linear logistic model to measure the change in the proportion, p_{BY}, of reads derived from the BY transcript as a function of time. We then assessed statistical significance with two different methods for dealing with genes that have high read counts, but only small (and, therefore, likely non-biologically significant) departures from non-constancy: a Bayesian hierarchical Markov chain Monte Carlo model (see Methods), or a likelihood ratio test in which we imposed a threshold on the effect size a gene needed to exhibit in order to be called as significant (see Supplementary Methods). Using the Bayesian hierarchical model, we identify 350 genes as significant (see Results). Before imposing a threshold on the results of the likelihood ratio test, we identify 358 genes as significant at a FDR = 10%; however, only 323 genes of these genes meet our threshold of a greater than 0.004 change in the odds of observing an mRNA allele of the BY strain given a one minute increase in time. 310 (96.0%) of the 323 genes we identify with our thresholding method overlap with the set of 350 genes we identify using the Bayesian hierarchical Markov chain Monte Carlo model (Figure S1B). Thus, the two approaches agree well with one another.

Supplementary Figures



Supplementary Figure 1. ROC curves for analyses of simulated data sets using our novel statistical framework. From a standard binomial distribution, as well as from beta-binomial distributions corresponding to an overdispersion parameter, k, = 1.25, 10, and 25, we simulated ten data set replicates, each of ten-thousand genes, for a $\pi_0 = 0.7$ (i.e. 30% of genes exhibited allele-specific mRNA decay rate). We then analyzed each of these data sets in two ways using our newly developed statistical framework: first, we calculated the variance for each gene assuming the data set came from a standard binomial (i.e., not overdispersed) distribution; second, we calculated the variance for each gene assuming the data came from a quasibinomial (i.e. overdispersed) distribution. To make each ROC curve, we varied the threshold on the posterior probability of no difference in mRNA decay rate that we used to identify genes with significant differences

in allele-specific mRNA decay rate. The ROC curves calculated for data analyzed under the assumption of no overdispersion are shown in shades of red, while the ROC curves calculated for data analyzed under the assumption of overdispersion are shown in shades of blue. The black dotted line corresponds to the ROC curve we would expect to see if our analysis method were performing no better than random chance. As shown by the curves in the figure, it is more accurate to analyze binomially-distributed data under the assumption of no overdispersion than under the assumption of overdispersion when using our newly developed statistical framework. It is also more accurate to analyze modestly overdispersed (k = 1.25) data under the assumption of no overdispersion. For data with more severe overdispersion (k = 10 or k = 25), our newly developed statistical framework performs poorly, regardless of the assumptions made about overdispersion.



Supplementary Figure 2. A) Choice of a threshold for identifying genes with allele-specific mRNA decay rate differences based on comparison of the distributions of the parameter estimate $\widehat{\beta}_\iota$. We compared the distributions of the magnitude of the change in

the odds of observing an mRNA allele of the BY strain given a one minute increase in time ($|\lambda_{BY} - \lambda_{RM}|$) between the set of genes we identified from a likelihood ratio test at FDR = 10% as exhibiting allele-specific mRNA decay rate (bottom histogram) and all other genes (top histogram). Based on these two distributions, we chose a threshold of $|\lambda_{BY} - \lambda_{RM}| = 0.004$ (red vertical line) and discarded genes with a $|\lambda_{BY} - \lambda_{RM}|$ below this threshold from our set of significant genes. B) Overlap between gene sets identified as exhibiting allele-specific differences in mRNA decay rate via two different methods. Using a Bayesian hierarchical Markov chain Monte Carlo model to determine whether the parameter estimate $\widehat{\beta}_1$, which we obtained from linear logistic model, differed from zero, we identified 350 genes with allele-specific mRNA decay rate differences. Using a likelihood ratio test in which we imposed a threshold on the effect size a gene needed to exhibit in order to be called as significant, we identified 323 genes with allele-specific mRNA decay rates. 310 (96%) of the 323 genes we identify with our thresholding method overlap with the set of 350 genes we identify using the Bayesian hierarchical Markov chain Monte Carlo model. Thus, the two approaches agree well with one another.

Supplementary Tables

Supplementary Table 1. Estimates of π_0 and the number of significant genes from a likelihood ratio test corrected for multiple testing with the QVALUE software and from our novel statistical framework for eight sets of simulated data. From a standard binomial distribution, as well as from beta-binomial distributions in which the parameters *a* and *b* corresponded to an overdispersion parameter, *k*, of 1.25, 10, and 25, we simulated ten data set replicates, each of ten-thousand genes, for each π_0 between 0.1 and 0.9, incrementing by 0.2. In the table, each row corresponds to one group of ten replicates (each replicate consists of a set of ten-thousand genes), while column 1 identifies the distribution from which we sampled the data, column 2 gives the *k* for that distribution, and column 3 gives the proportion of genes that we simulated under the null hypothesis for each data set in the group of ten replicates. Column 4 identifies whether we analyzed the simulated data sets with or without assuming overdispersion, i.e. whether we calculated the variance under a standard binomial distribution or under a quasibinomial distribution, respectively. The mean and standard deviation of the π_0 we calculated for the ten replicates using likelihood ratio test corrected for multiple testing with the QVALUE software (Storey 2002; Storey and Tibshirani 2003; Storey *et al*. 2004), or using our newly-developed statistical framework involving a Bayesian hierarchical Markov chain Monte Carlo model are shown in columns 5 and 6, respectively. The mean and standard deviation of the number of genes we identified as exhibiting allele-specific mRNA decay rate across the ten replicates using likelihood ratio test corrected for multiple testing with the QVALUE software (Storey 2002; Storey and Tibshirani 2003; Storey *et al*. 2004), or

using our newly-developed statistical framework involving a Bayesian hierarchical Markov chain Monte Carlo model are shown in columns 7 and 8, respectively.

Simulated distribution	***k***	**Simulated π_0**	**Assumed analysis distribution**	**Likelihood ratio test and qvalue software π_0**	**Bayesian MCMC π_0**	**Likelihood ratio test and qvalue software number of genes (q < 0.10)**	**Bayesian MCMC number of genes (post. prob. > 0.65)**
Binomial	1	0.10	Binomial	0.104 ± 0.0136	0.100 ± 0.0030	9949 ± 122	8968 ± 34
Binomial	1	0.30	Binomial	0.305 ± 0.0170	0.300 ± 0.0045	7566 ± 109	6842 ± 60
Binomial	1	0.50	Binomial	0.495 ± 0.0330	0.499 ± 0.0050	5201 ± 98	4722 ± 53
Binomial	1	0.70	Binomial	0.692 ± 0.0423	0.699 ± 0.0036	2892 ± 51	2746 ± 36
Binomial	1	0.90	Binomial	0.898 ± 0.0434	0.898 ± 0.0040	769 ± 43	812 ± 32
Beta-binomial	1.25	0.10	Binomial	0.097 ± 0.0155	0.010 ± 0.0040	9975 ± 72	8957 ± 39
Beta-binomial	1.25	0.30	Binomial	0.268 ± 0.0211	0.291 ± 0.0057	7836 ± 115	6871 ± 56
Beta-binomial	1.25	0.50	Binomial	0.458 ± 0.0347	0.581 ± 0.0106	5402 ± 118	4827 ± 94
Beta-binomial	1.25	0.70	Binomial	0.617 ± 0.0392	0.588 ± 0.2215	3132 ± 112	3078 ± 670
Beta-binomial	1.25	0.90	Binomial	0.796 ± 0.0362	0.711 ± 0.0304	911 ± 68	1087 ± 92
Beta-binomial	10	0.10	Binomial	0.192 ± 0.0248	0.004 ± 0.0022	8831 ± 257	10000 ± 0
Beta-binomial	10	0.30	Binomial	0.216 ± 0.0256	0.0048 ± 0.0031	8453 ± 250	10000 ± 0
Beta-binomial	10	0.50	Binomial	0.247 ± 0.0253	0.010 ± 0.0068	8035 ± 226	10000 ± 0
Beta-binomial	10	0.70	Binomial	0.266 ± 0.0290	0.012 ± 0.0113	7707 ± 228	10000 ± 0
Beta-binomial	10	0.90	Binomial	0.287 ± 0.0231	0.013 ± 0.0095	7343 ± 209	10000 ± 0
Beta-binomial	25	0.10	Binomial	0.154 ± 0.0180	0.0070 ± 0.0053	9336 ± 178	10000 ± 0
Beta-binomial	25	0.30	Binomial	0.164 ± 0.0297	0.008 ± 0.0067	9227 ± 285	10000 ± 0
Beta-binomial	25	0.50	Binomial	0.170 ± 0.0172	0.010 ± 0.0095	9117 ± 174	10000 ± 0
Beta-binomial	25	0.70	Binomial	0.180 ± 0.0165	0.007 ± 0.0057	9028 ± 156	10000 ± 0
Beta-binomial	25	0.90	Binomial	0.185 ± 0.0149	0.010 ± 0.0074	8933 ± 131	10000 ± 0

Binomial	1	0.10	Quasibinomial	0.096 ± 0.0124	0.073 ± 0.0045	9983 ± 78	9233 ± 41
Binomial	1	0.30	Quasibinomial	0.282 ± 0.0269	0.238 ± 0.0205	7852 ± 168	7298 ± 168
Binomial	1	0.50	Quasibinomial	0.472 ± 0.0292	0.210 ± 0.0315	5602 ± 123	6704 ± 528
Binomial	1	0.70	Quasibinomial	0.663 ± 0.0307	0.368 ± 0.0202	3443 ± 71	4128 ± 168
Binomial	1	0.90	Quasibinomial	0.854 ± 0.0454	0.549 ± 0.0240	1422 ± 106	1842 ± 136
Beta-binomial	1.25	0.10	Quasibinomial	0.101 ± 0.0152	0.070 ± 0.0065	9951 ± 98	9289 ± 66
Beta-binomial	1.25	0.30	Quasibinomial	0.279 ± 0.0090	0.099 ± 0.1243	7666 ± 83	9162 ± 1907
Beta-binomial	1.25	0.50	Quasibinomial	0.469 ± 0.0232	0.155 ± 0.0190	5337 ± 82	8352 ± 853
Beta-binomial	1.25	0.70	Quasibinomial	0.663 ± 0.0375	0.325 ± 0.0284	3175 ± 122	4535 ± 392
Beta-binomial	1.25	0.90	Quasibinomial	0.848 ± 0.0235	0.485 ± 0.0369	1305 ± 93	2064 ± 251
Beta-binomial	10	0.10	Quasibinomial	0.607 ± 0.0382	0.006 ± 0.0014	1914 ± 125	10000 ± 0
Beta-binomial	10	0.30	Quasibinomial	0.678 ± 0.0434	0.012 ± 0.0077	1493 ± 106	10000 ± 0
Beta-binomial	10	0.50	Quasibinomial	0.753 ± 0.0348	0.035 ± 0.0303	1148 ± 66	10000 ± 0
Beta-binomial	10	0.70	Quasibinomial	0.830 ± 0.0328	0.122 ± 0.0532	834 ± 34	9992 ± 28
Beta-binomial	10	0.90	Quasibinomial	0.896 ± 0.0377	0.183 ± 0.0741	573 ± 84	9938 ± 118
Beta-binomial	25	0.10	Quasibinomial	0.773 ± 0.0418	0.024 ± 0.0165	828 ± 63	10000 ± 0
Beta-binomial	25	0.30	Quasibinomial	0.817 ± 0.0214	0.051 ± 0.0289	749 ± 86	10000 ± 0
Beta-binomial	25	0.50	Quasibinomial	0.844 ± 0.0438	0.117 ± 0.0676	651 ± 60	9994 ± 12
Beta-binomial	25	0.70	Quasibinomial	0.878 ± 0.0536	0.178 ± 0.0829	552 ± 76	9891 ± 497
Beta-binomial	25	0.90	Quasibinomial	0.919 ± 0.0493	0.197 ± 0.0624	468 ± 48	9936 ± 132

Supplementary Table 2. Genes that exhibit allele-specific mRNA decay and contain a lone single nucleotide variant. We identified 13 genes with allele-specific mRNA decay (ASD) that possess a single variant between the BY and RM alleles of the transcript. In the table, column 1 identifies the gene in which the lone single variant resides, while column 4 provides the location of the variant in the BY genome and column 6 provides the location of the variant in the RM genome. Column 2 shows the posterior probability of ASD we calculated for the gene from our statistical model (see Methods) and column 3 shows our estimate of the slope calculated from the linear logistic model for the gene. The exponential of the slope is the change in the odds of observing an mRNA allele of the BY strain given a one minute increase in time for the gene. Columns 5 and 7 provide the BY and RM alleles of the variant, respectively; column 8 lists the type of change the variant affects in the transcript.

GeneName	Posterior probability of allele-specific decay rate	$\lambda_{BY} - \lambda_{RM}$	Position in BY genome	BY allele	Position in RM genome	RM allele	Type of change
YDR162C	0.864	0.0173	Chr IV: 781102	A	supercontig 1.1: 719468	G	5' UTR
YDR258C	0.694	-0.0050	Chr IV: 974041	G	supercontig 1.1: 538400	T	Coding, nonsynonymous codon change (A -> S)
YIL106W	0.829	-0.0314	Chr IX: 166531	C	supercontig 1.14: 263381	T	Coding, nonsynonymous codon change (P -> L)
YIL053W	1.000	-0.0262	Chr IX: 255653	A	supercontig 1.14: 180392	C	Coding, synonymous codon change without preference change
YER070W	0.704	-0.0168	Chr V: 300737	G	supercontig 1.11: 293068	A	Coding, nonsynonymous codon change (T -> I)
YGR214W	1.000	0.0062	Chr VII: 921307	C	supercontig 1.2: 167564	T	Coding, nonsynonymous codon change (V -> I)
YLR200W	0.761	-0.0160	Chr XII: 549296	A	supercontig 1.10: 85244	G	Coding, nonsynonymous codon change (M -> V)
YMR314W	0.838	-0.0128	Chr XIII: 902531	G	supercontig 1.5: 880665	C	3' UTR
YNL134C	0.934	-0.0102	Chr XIV: 373133	T	supercontig 1.7: 396984	C	Coding, synonymous codon change with preference change
YOR367W	0.700	0.0010	Chr XV: 1026426	A	supercontig 1.3: 994675	C	Coding, nonsynonymous codon change (T -> P)
YPL212C	0.844	0.0227	Chr XVI: 152471	T	supercontig 1.4: 145874	C	Coding, synonymous codon change with preference change
YPL183W-A	0.875	-0.0155	Chr XVI: 199137	T	supercontig 1.4: 192474	C	Coding, nonsynonymous codon change (L -> S)
YPL117C	0.882	-0.0085	Chr XVI: 328316	C	supercontig 1.4: 321315	T	Coding, synonymous codon change with preference change

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106. DOI: 10.1186/gb-2010-11-10-r106.

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.

McCullagh P, Nelder JA. 1989. *Generalized Linear Models,* 2nd ed. Chapman and Hall/CRC Press, London.

Storey JD. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64:** 479-98.

Storey JD and Tibshirani R. 2003. Statistical significance for genome-wide experiments. *Proc Natl Acad Sci USA* **100:** 9440-5.

Storey JD, Taylor JE, and Siegmund D. 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B* **66:** 187-205.

Tirosh I, Reikhav S, Levy AA, Barkai N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324:** 659-62.

High-quality *de novo* genome and transcriptome assembly of two wild-derived *Saccharomyces cerevisiae* strains

Jennifer M. Andrie[*], David Gordon[*†], Evan E. Eichler[*†] and Joshua M. Akey[*]

[*]Department of Genome Sciences, University of Washington, Seattle, Washington, 98195.

[†]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, 98195.

Correspondence to:

Joshua M. Akey, PhD
Department of Genome Sciences
University of Washington School of Medicine
Box 355065
1705 NE Pacific Street
Seattle, WA 98195
(206) 543-7254
akeyj@uw.edu

Reference Numbers for Data Available in Public Repositories: SRR5168994, SRR5168993, SRR5170142, SRR5170141, SRR5171570, SRR5171571, SRR5176941, SRS1915938, SRS1915939

Running title: De novo yeast genome assembly

Key words: Yeast; genomics; genome annotation; genome assembly; RNA-Seq

Submitted to G3 on January 19, 2017 (MS ID#: G3/2017/039719)

ABSTRACT

Massively parallel short read sequencing has revolutionized genomics, enabling large-scale catalogs of genetic variation to be compiled within and between populations. Nonetheless, high-quality *de novo* genome assemblies are necessary to fully capture the entire spectrum of genetic variation and resolve structurally complex genomic regions. Here, we leverage PacBio's long read sequencing technology to generate *de novo* assemblies of two wild *Saccharomyces cerevisiae* strains (Y12 and DBVPG6044) as well as deep RNA-Seq data (>1000X) to facilitate unbiased genome annotation and identification of genomic variation. Moreover, our RNA-Seq data suggests that 96% of the genome is transcribed, anti-sense transcription is widespread, and unexpressed genomic regions are enriched for origins of replication. Finally, we evaluate how *de novo* genome assemblies mitigate read mapping bias in estimates of allele specific expression measured in Y12 x DBVPG6044 diploids.

INTRODUCTION

Compared to assembly methods that use a reference genome, *de novo* whole genome assembly enables more comprehensive characterization of genetic variation, including better annotation of functional elements, improved resolution of repetitive regions, and reduced levels of missing sequence (Chaisson *et al.* 2015). Consequently, *de novo* genome assemblies also improve the accuracy of read mapping for the rapidly expanding repertoire of sequencing-based assays to assess various molecular phenotypes, such as chromatin state, protein-DNA interactions, protein-RNA interactions, RNA expression levels, translation rates, and RNA secondary structure (Shendure and Lieberman Aiden 2012). In particular, when reads from these assays are mapped to more accurate reference genomes, biases that influence the accuracy of functional genomics phenotypes (Degner *et al.* 2009) can be ameliorated.

Currently, *de novo* assembly from short read next generation sequencing data remains a difficult problem, despite advances in sequencing technologies and computational algorithms (Sohn and Nam 2016). Specifically, uneven read depth arising from variation in PCR, cloning, GC bias, sequencing errors, copy number variants, and the topological complexity of repetitive elements often result in gaps and/or misassembly (Sohn and Nam 2016). To overcome these challenges and achieve chromosome-scale scaffolding, physical mapping methods, such as optical mapping or chromatin-interaction mapping have been used (Burton *et al.* 2013; Dong *et al.* 2013). Alternatively, *de novo* whole genome assembly with long read sequencing, such as with PacBio's Single Molecule, Real-Time (SMRT) Sequencing platform, generates highly contiguous assemblies, easily achieving chromosome-length contigs for bacterial and simple eukaryotic genomes given sufficient read-depth (Zowawi *et al.* 2015; Frank *et al.* 2016; McIlwain *et al.* 2016).

Although whole-genome sequencing has been performed on hundreds of *Saccharomyces cerevisiae* strains (Kim *et al*. 2014; Gallone *et al.* 2016; McIlwain *et al.* 2016; Strope *et al.* 2016; Zhu *et al.* 2016), these data were largely obtained with short read sequencing, thus precluding high-quality *de novo* assembly. Here, we describe high quality *de novo* genome assemblies obtained from long read sequencing data in two wild-derived haploid strains of *S. cerevisiae*, as well as *de novo* transcriptome assemblies from ultra high-coverage RNA-Seq data. We show these data provide novel insights into the structure and function of yeast genomes.

MATERIALS AND METHODS

Yeast strains

We obtained the wild-derived haploid *S. cerevisiae* strains Y12 (*MATa*) and DBVPG6044 (*MATα*) from our strain collection stored in glycerol at -80°. We inoculated cells from the glycerol stocks into tubes of 5 mL of liquid yeast extract peptone dextrose (YEPD), incubated the tubes in a rotating tube rack at 30° overnight, and then streaked 50 µL of culture onto YEPD plates. We incubated the plates at 30° for 24 hours, until single colonies were visible, and then wrapped the plates in parafilm and stored them at 4°C.

We mated Y12 (*MATa*) and DBVPG6044 (*MATα*) to create a hybrid diploid by mixing cells from single colonies of each strain together on a new YEPD plate and incubating the plate at 30° for 6 hours. We scraped some of the mixed cells from the plate, diluted those cells into ultrapure H_2O to create a cell slurry, plated the cell slurry onto a YEPD plate, and incubated the plate at 30°C for 24 hours. The resulting cell colonies exhibited clear division into two size groups; we selected three of the larger colonies as candidate diploid hybrids and created glycerol stocks for these three colonies following standard procedures. The stocks were added to our strain collection stored at -80°C.

We confirmed that our candidate diploid hybrids identified in our screen were, in fact, diploid using a standard Halo Mating Type Assay as well as by a Polymerase Chain Reaction (PCR) assay. For the PCR-based assay, we began by inoculating cells from each of the three colonies present on the YEPD plate into tubes of 5 mL liquid YEPD and incubated the tubes in a rotating tube rack at 30°C overnight. We then extracted the DNA from these cultures as follows: We transferred 1.5 mL of culture to a screw top tube, centrifuged the tube at room temperature for 3 min at 14,000 rpm, and poured off the supernatant. We added 200 µL of lysis buffer (10

mM Tris-HCl, pH=8.0; 1 mM EDTA; 100 mM NaCl; 1% SDS; 2% Triton X-100), 0.3 mL of glass beads (425-600 μm, Sigma-Aldrich, G-9268), and 200 μL of phenol:chloroform:isoamyl alcohol to the cell pellet, vortexed at top speed for 2 min, added 200 μL of TE buffer (10 mM Tris-HCl, pH=8.0; 1 mM EDTA), and briefly vortexed again. We centrifuged the tube at 4°C for 5 min at 14,000 rpm, transferred the aqueous phase to a new screw top tube, and added 400 μL of phenol:chloroform:isoamyl alcohol. After mixing by inverting the tube several times, we again centrifuged the tube at 4°C for 5 min at 14,000 rpm and transferred the aqueous phase to a 1.7 mL Eppendorf tube. We added two volumes of room temperature 100% ethanol, centrifuged the tube at 4°C for 2 min at 14,000 rpm, and discarded the supernatant. We washed the nucleic acid pellet twice with 500 μL of ice-cold 70% ethanol, centrifuged the tube at 4°C for 2 min at 14,000 rpm, discarded the supernatant, and placed the open tube in a tube rack at room temperature for approximately 10 min to dry the pellet. We resuspended the pellet in 200 μL TE buffer. To perform each PCR reaction, we combined 1 μL of the template DNA solution, 1 μL of 10X *Taq* (Mg-free) Reaction Buffer (M0320S New England BioLabs, Inc., Ipswich, MA), 0.5 μL of 2 mM dNTPs, 0.5 μL of 50 mM $MgCl_2$, 0.1 μL of 10 μM Primer mix, 0.1 μL of *Taq* (Mg-free) (M0320S New England BioLabs, Inc., Ipswich, MA), and 6.8 μL of ultrapure H_2O, to a final volume of 10 μL. We performed two PCR reactions per colony: one that contained the forward primer for the *MATα* locus (5'-TTACTCACAGTTTGGCTCCGGTGT-3') and the reverse primer for both *MAT* loci (5'-GAACCGCATGGGCAGTTTACCTTT-3'); and one that contained the forward primer for the *MATa* locus (5'-CTCCACTTCAAGTAAGAGTTTGGG-3') and the same reverse primer as above. We heated the PCR reactions to 95° for 3 min, performed 30 cycles of 45 seconds at 95°C, 30 seconds at 62°C, and 45 seconds at 72°C, and

ended with an incubation 72°C for 5 min before holding the reactions at 4°C. We checked the products of the PCR reactions using a standard gel electrophoresis assay.

PacBio whole-genome sequencing

To isolate DNA from the Y12 (*MATa*) and DBVPG6044 (*MATα*) strains, we began by inoculating single colonies of each strain into tubes of 5 mL of liquid YEPD, incubating the tubes in a rotating tube rack at 30°C overnight, and then inoculating 2 mL of the resultant cultures into 250 mL of liquid YEPD in 2000 mL Erlenmeyer flasks and incubating the cultures in a shaking incubator at 30°C for 4.5 hours at 200 rpm. Using a hemacytometer to estimate the cell density in the cultures, we aliquoted approximately 3 x 10^9 cells per strain into new tubes. We then extracted the DNA from the cells using the Qiagen Genomic Tips 100/G kit (Qiagen, Valencia, CA) following the manufacturer's instructions with the following modifications: We pre-warmed the zymolase reactions to 30°C in a water bath and then incubated them for 45 minutes at 30°C with shaking at 200 rpm; we pre-warmed the RNase/Proteinase reactions to 48°C in a water bath, and then incubated the reactions for 40 minutes at 50°C with shaking; we spun down the cell debris for 10 minutes at 5,000 rpm. We assessed the concentration and purity of the resultant genomic DNA using a Qubit® dsDNA BR Assay (Life Technologies, Grand Island, NY) and a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE). We further purified the DNA with a MoBio PowerClean Pro DNA Clean-up kit, eluting the DNA into 100 µL of MoBio reagent DC5 (MoBio Laboratories, Inc., Carlsbad, CA). We reassessed the concentration of the DNA with a Qubit® dsDNA BR Assay. To evaluate the size distribution of the extracted DNA, we used standard Pulse Field Gel Electrophoresis with 200 – 300 ng of DNA loaded per sample. We ran the gel at 1 second/6 second intervals for 16 hours at 160 V and 14°C.

The isolated DNA was sequenced on the PacBio RSII instrument at the University of Washington. Size distribution of the DNA was further analyzed by an Agilent DNA 1200 assay on the 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA). For each library, 7.5 µg of DNA was diluted to 60 ng/µL and sheared using a Covaris gTube (Covaris, Inc., Woburn, MA) with centrifuge settings of 3200 rpm for 2 x 1 minute, followed by 3400 rpm for 2 x 30 sec. The tube was then flipped and the same settings used for a second shear pass. After cleanup and concentration using AMPure PB beads (Pacific Biosciences, Menlo Park, CA) at a 0.45x ratio, a 20 kb SMRTbell library was constructed using the SMRTbell Template Prep Kit 1.0 according to the protocol "Procedure & Checklist - Greater Than 10 kb Template Preparation Using AMPure® PB Beads." Each library was annealed and bound to polymerase using the P6v2 kit, bound to MagBeads for 20 min, and sequenced on 4 SMRT cells at an on-plate concentration of 0.12 to 0.15 nM.

Illumina whole-genome sequencing

We streaked glycerol stocks of haploid Y12 (*MATa*) and haploid DBVPG6044 (*MATα*) onto YEPD plates and then incubated the plates at 30°C for 48 hours. We inoculated single colonies of each strain into 5 mL of liquid YEPD and then incubated in a rotating shaker at 30°C for 21 hours. To isolate DNA, we followed the protocol described above for the PCR-based assay to verify mating type, except that we resuspended in 100 µL TE. We quantified the DNA in 1:100 dilutions with the Qubit dsDNA HS Assay Kit (Life Technologies, Grand Island, NY). We prepared sequencing libraries from 1 ng of DNA using Nextera XT DNA Library Preparation Kit (FC-131-1024, Illumina, San Diego, CA). We sent the DNA libraries to the Northwest Genomics Center for sequencing. Cluster generation was performed on an Illumina

cBot using standard HiSeq 4000 chemistry cluster kits (Illumina, San Diego, CA). Paired-end 75-bp sequencing was performed on an Illumina HiSeq 4000 using standard SBS chemistry and HiSeq Control Software version 3.3.52 (Illumina, San Diego, CA).

De novo genome assembly and polishing

We performed *de novo* assembly of the Y12 and DBVPG6044 strain genomes using our long-read sequencing data generated on a PacBio RSII instrument and the string graph assembly algorithm, FALCON, version June 30, 2015 hash: cee6a58 (https://github.com/PacificBiosciences/FALCON). We polished our assemblies with the consensus algorithm, Quiver, version 1.1.0 (https://github.com/PacificBiosciences/GenomicConsensus). To assign names to the resulting contigs, we aligned each assembly to the S288c genome (version R64-2-1, released November 18, 2014; http://www.yeastgenome.org; Engel *et al.* 2013) with BLASR, version July 1, 2015 (https://github.com/mchaisso/blasr). As we were only interested in analyzing the nuclear genome of each strain, we removed any contigs aligning to the S288c mitochondrial genome from our assemblies.

Identifying structural variation in *de novo* genomes

To compare the structures of the Y12 and DBVPG6044 strain genomes we assembled *de novo* from long-read sequencing data generated on a PacBio RSII instrument, we visualized an alignment of each pair of orthologous chromsomes from the two strains using the Genome Pair – Rapid Dotter (Gepard) program (Krumsiek *et al.* 2007). For each pair of chromosomes, we uploaded the chromosomal sequence of both strains into the Gepard GUI, and directed the

program to calculate a dot plot using default parameters. We then manually inspected the resulting dot plots for regions containing insertions/deletions (represented as gaps in the dot plot) or inversions. We also assessed the Ty element and simple repeat content of each strain's genome with the RepeatMasker, version 4.0.1 program (http://www.repeatmasker.org). We used default settings for the program, with a custom Ty element library obtained from Carr *et al.* (2012).

Creating polymorphised genomes from the S288c reference genome

We obtained a complete genome sequence for S288c from the Saccharomyces Genome Database (version R64-2-1, released November 18, 2014; http://www.yeastgenome.org; Engel *et al.* 2013). We trimmed adapters from the reads using Trim Galore, version 0.4.1 (https://github.com/FelixKrueger/TrimGalore) with *--illumina* and *--paired* and then used BWA, version 0.7.13 (Li and Durbin 2009a) *mem* with *–M -R '@RG\tID:group1\tSM:sample1\tPL:illumina\tLB:lib1\tPU:unit1'* to map both the Y12 and DBVPG6044 DNA sequence reads generated on the Illumina HiSeq 2500 to the S288c genome. After mapping reads, we used the SortSam tool in Picard, version 1.111 (http://picard.sourceforge.net) to convert the SAM-formatted files output by BWA to BAM format and sort the BAM files by coordinate. We then marked duplicate reads using the Picard *MarkDuplicates* tool and indexed the resulting BAM files with the *BuildBamIndex* tool.

To identify single nucleotide variant (SNV) sites between the S288c reference genome and our two strains, we used GATK, version 3.5 (McKenna *et al.* 2010). We began by running the GATK tool *HaplotypeCaller* with options *--genotyping_mode DISCOVERY -stand_call_conf 30.0 -stand_emit_conf 10.0 --sample_ploidy 1* to identify variant sites. We used the GATK tool

BaseRecalibrator with the output of *HaplotypeCaller* as the *–knownSites* and the GATK *PrintReads* tool to recalibrate the sequencing data we initially provided to the *HaplotypeCaller* tool. We then ran the *HaplotypeCaller* tool as before on the recalibrated reads and extracted SNVs from the variant call set using the GATK tool *SelectVariants* with the option *-selectType SNP*. We further filtered the call set using the GATK tool *VariantFiltration* using the standard hard filtering parameters according to GATK Best Practices recommendations, including the options *--filterExpression "QD < 10.0 || FS > 60.0 || MQ < 50.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"* (DePristo *et al.* 2011; Van der Auwera *et al.* 2013). We created polymorphised versions of the S288c reference genome for each of the Y12 and DBVPG6044 strains by substituting the appropriate base at each of the SNV sites identified with GATK from our Illumina sequencing data. We did not include indels predicted by GATK from our Illumina sequencing data in the polymorphised genomes due to the increased incidence of errors by GATK in the indel calls versus the SNV calls.

RNA-Seq

To isolate RNA from the Y12 (*MATa*) and DBVPG6044 (*MATα*) strains, we began by inoculating single colonies of each strain into tubes containing 5 mL of liquid YEPD, incubating the tubes in a rotating tube rack at 30°C overnight, and then inoculating the resultant cultures into 50 mL of liquid YEPD in a 250 mL Erlenmeyer flasks to achieve a starting culture density of OD_{660} = 0.17 – 0.20. We incubated the 50 mL cultures in a shaking incubator at 30°C for 5.75 hours at 225 rpm and then collected the cells from the culture, now at OD_{660} = 0.9 – 1.1, using vacuum filtration onto a 0.45 micron nylon filter (R04SP04700, GE Osmonics). The filter was placed in a 15 mL Falcon™ tube and flash frozen in liquid nitrogen. We proceeded immediately

with a standard acid phenol-chloroform preparation to extract total RNA from the cells: We added 1 mL of ice-cold TES buffer (10 mM Tris-HCl, pH=7.5; 10 mM EDTA; 0.5% SDS) to each 15 mL tube, and then alternated vortexing at top speed for 15 seconds with resting on ice for 15 seconds until all the cells were washed off the filter. To pull the cell slurry to the bottom of the tube, we spun the tube in a tabletop centrifuge until the centrifuge reached 1,000 rpm. We transferred the cell slurry to a 1.5 mL screw top tube, added 400 µL of acid phenol pre-warmed to 65°C, and then incubated the sample at 65°C for one hour, vortexing every 10 minutes during the incubation. We then placed the tube on ice for 5 min before centrifuging at 4°C for 5 min at 14,000 rpm, transferring the aqueous phase to a new screw top tube, adding 400 µL of pre-warmed acid phenol, and briefly vortexing. We again centrifuged at 4°C for 5 min at 14,000 rpm, transferred the aqueous phase to a new screw top tube, added 400 µL of pre-warmed acid phenol, and briefly vortexed. After centrifuging at 4°C for 5 min at 14,000 rpm and transferring the aqueous phase to a new screw top tube a third time, we added 400 µL of room temperature chloroform, and briefly vortexed. We centrifuged the tube at 4°C for 5 min at 14,000 rpm and transferred the aqueous phase to a 1.7 mL Eppendorf tube. We added 0.1 volume of 3M Sodium Acetate and 2.5 volumes of cold 100% ethanol, and froze the sample at -80°C overnight. We centrifuged the tube at 4°C for 30 min at 14,000 rpm, and discarded the supernatant. We washed the nucleic acid pellet twice with 500 µL of ice-cold 100% ethanol, centrifuged the tube at 4°C for 2 min at 14,000 rpm, discarded the supernatant, and placed the open tube in a tube rack at room temperature for approximately 10 min to dry the pellet. We resuspended the pellet in 200 µL diethylpyrocarbonate-treated (DEPC-treated) H_2O. We quantified the samples using a Qubit® RNA BR Assay (Life Technologies, Grand Island, NY).

To remove DNA from the samples, we combined 200 µg of RNA in 172 µL DEPC-treated H_2O, 20 µL Turbo DNase Buffer, and 8 µL of Turbo DNase (2 U/µL) (Life Technologies, Grand Island, NY), and then incubated the mixture at 37° for 30 minutes. We re-isolated the RNA with a phenol:chloroform:isoamyl alcohol extraction: We placed the sample in a screw top tube, added an equal volume of phenol:chloroform:isoamyl alcohol cooled to 4°C, briefly vortexed, and centrifuged at 4°C for 5 min at 14,000 rpm. We transferred the aqueous phase to a new screw top tube, added an equal volume of room temperature chloroform, and briefly vortexed. We centrifuged the tube at 4°C for 5 min at 14,000 rpm and transferred the aqueous phase to a 1.7 mL Eppendorf tube. We added 0.1 volume of 3M Sodium Acetate and 2.5 volumes of cold 100% ethanol, froze the sample at -80°C for one hour, and then centrifuged the tube at 4°C for 30 min at 14,000 rpm. We discarded the supernatant and washed the nucleic acid pellet twice with 500 µL of ice-cold 70% ethanol, centrifuged the tube at 4°C for 2 min at 14,000 rpm, discarded the supernatant, and placed the open tube in a tube rack at room temperature for approximately 25 min to dry the pellet. We resuspended the pellet in 150 µL DEPC-treated H_2O. We quantified the samples using a Qubit® RNA BR Assay (Life Technologies, Grand Island, NY). We assessed the quality of the samples with an Agilent RNA 6000 Nano assay on the 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA).

To collect RNA from the Y12 (*MATa*) x DBVPG6044 (*MATα*) hybrid diploid we generated, we inoculated cells from our -80°C stock into a tube containing 5 mL of liquid YEPD, incubated the tube in a rotating tube rack at 30°C overnight, and then inoculated 1 mL the resultant culture into 50 mL of liquid YEPD in 250 mL Erlenmeyer flask to achieve a starting culture density of OD_{660} = 0.06. We incubated the 50 mL cultures in a shaking incubator at 30°C for 6.5 hours at 225 rpm and then used a hemacytometer to estimate the cell density in the

culture. We inoculated 2.22 x 10^5 cells from the 50 mL culture into 490 mL of liquid YEPD in a 2,000 mL Erlenmeyer flask and incubated the 490 mL cultures in a shaking incubator at 30°C for 15.5 hours at 225 rpm, to a final culture density of OD_{660} = 0.800. We collected the cells from 25 mL of culture using vacuum filtration, as previously described. We extracted total RNA from the cells and removed DNA using a standard acid phenol-chloroform preparation and DNase treatment, also previously described. In total, we collected three replicates of RNA from the Y12 (*MATa*) x DBVPG6044 (*MATα*) hybrid diploid.

To prepare the RNA samples for sequencing, we first depleted them of rRNA using the Illumina Ribo-Zero Gold rRNA Removal Kit (Yeast) (Illumina, San Diego, CA). We purified the product using the manufacturer's suggested ethanol precipitation protocol, but omitting the addition of glycerol to our samples. We quantified and assessed the quality of the product of the Illumina Ribo-Zero Gold rRNA Removal Kit (Yeast) with an Agilent RNA 6000 Pico assay on the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). We prepared strand-specific sequencing libraries for each of our samples using the Clontech SMARTer® Stranded RNA-Seq Kit (Takara Bio USA, Inc., Mountain View, CA) and 50 ng of input RNA. We fragmented the RNA for 5 minutes and used 10 cycles of PCR. We assessed the quality of our libraries with an Agilent High Sensitivity DNA assay on the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). We quantified the libraries with a Qubit® dsDNA HS Assay (Life Technologies, Grand Island, NY).

We sent the RNA libraries to the Northwest Genomics Center for sequencing. For the libraries generated from the haploid Y12 (*MATa*) and DBVPG6044 (*MATα*) strains, cluster generation was performed using an Illumina cBot using standard HiSeq 4000 cluster kits (Illumina, San Diego, CA). Paired-end 75-bp sequencing was performed on an Illumina HiSeq

4000 using standard SBS chemistry and HiSeq Control Software version 3.3.20 (Illumina, San Diego, CA). For the libraries generated from the Y12 (*MATa*) x DBVPG6044 (*MATα*) diploid hybrid strain, cluster generation was performed using an Illumina cBot using standard HiSeq 2500 high output version 4 chemistry cluster kits (Illumina, San Diego, CA). Paired-end 100-bp sequencing was performed on an Illumina HiSeq 2500 in high output mode using SBS version 4 chemistry and HiSeq Control Software version 2.2.58 (Illumina, San Diego, CA).

Transcriptome assembly and annotation

We assembled a transcriptome for each of the Y12 (*MATa*) and DBVPG6044 (*MATα*) strains using the software programs Trinity, version 2.2.0 (Grabherr *et al.* 2011) and PASA, version 2.0.2 (Hass *et al.* 2003). We first performed *de novo* transcriptome assembly for each strain using fastq files of the Illumina RNA-seq reads and Trinity with the options *--min_kmer_cov 2 --min_glue 20 --min_per_id_same_path 99 --max_diffs_same_path 1 --SS_lib_type FR --jaccard_clip --trimmomatic --quality_trimming_params "HEADCROP:3" --normalize_reads*. In addition to generating a fasta file containing all *de novo* predicted transcripts for each strain, this Trinity command also produced two sets of fastq files: one set in which the 5' three base pairs of each read was trimmed; and a second set produced through an *in silico* normalization of the first set. We used this second set of reads as input for running Trinity again to perform genome-guided *de novo* transcriptome assembly. For the genome-guided run of Trinity, we used the same options as before, omitting the options *trimmomatic --quality_trimming_params "HEADCROP:3" --normalize_reads*, and adding the option *--genome_guided_bam,* to which we supplied a bam file of the normalized reads aligned to the PacBio-based genome assembly for the appropriate strain. We generated the bam file for each

strain using TopHat, version 2.0.12 (Trapnell *et al.* 2009) with the options *-r 100 -i 40 --no-discordant --no-mixed --library-type fr-secondstrand*. After completing both runs of Trinity for each strain, we polished the resulting transcriptome assemblies with the *seqclean* module included in the PASA pipeline package. Before running PASA, we concatenated the polished full *de novo* and genome-guided *de novo* assemblies into one large fasta-formatted file for each strain. We also created a file containing the list of transcript accessions that corresponded to the full *de novo* assembly transcripts with the *accession_extractor.pl* script included in the PASA pipeline package. We then ran PASA using *Launch_PASA_pipeline.pl* with the options *-C -R --ALIGNERS blat,gmap --transcribed_is_aligned_orient --stringent_alignment_overlap 30.0* and using the concatenated transcript assemblies, the list of transcript accessions that corresponded to the full *de novo* assembly transcripts, and the PacBio-based genome assembly as inputs.

To predict the proteins encoded by the transcripts in our assemblies, we used the program TransDecoder, version 2.1.0 (https://transdecoder.github.io) We extracted long open reading frames with the *LongOrfs* tool using the options *–S –m 100*, which indicate the program should only search for open reading frames (ORFs) on the sense strand that are greater than 100 basepairs in length. We predicted coding regions from the identified long ORFs with the *Predict* tool. To further refine both the transcriptome assembly output by PASA for each strain and the predicted ORFs for each strain, we merged predicted transcripts that contained overlapping predicted ORFs. We also merged neighboring predicted transcripts that both contained an ORF that either did not start on a start codon or stop on a stop codon. We similarly merged such transcripts with non-coding transcripts that overlapped abrupt ORF beginnings and ends. We then repeated ORF prediction with TransDecoder, as previously described, in the merged transcripts, mapped the newly-predicted ORFs back to the genome with GMAP, version 2014-

08-20 (Wu and Watanabe 2005), converted the gmap-formatted file to bed format with the *gmap_to_bed.pl* provided in the PASA pipeline package, and sorted the bed format file with BEDTools, version 2.25.0 (http://bedtools.readthedocs.io/en/latest/). To obtain our final set of predicted transcripts for each strain, we partitioned transcripts with multiple ORFs into transcriptional units with one ORF per transcript; to avoid ambiguity, sequence in between ORFs on such transcripts was not included in the newly annotated transcriptional units. To obtain our final set of predicted proteins for each strain, we again used TransDecoder, as previously described, to all predict ORFs in our final set of predicted transcripts for each strain.

To determine which non-coding transcripts were antisense to an ORF, we began by mapping the all predicted ORFs back to the genome, converting the gmap-formatted file to bed format, and sorting the bed format file, as described above. We then used the *intersect* tool in BEDTools, version 2.25.0 (http://bedtools.readthedocs.io/en/latest/), to identify the non-coding transcripts with at least 5% overlap on the opposite strand of a coding transcript using the *–f 0.05 –S* options.

Transcriptome and proteome comparison

We compared the transcriptomes and the proteomes between Y12 and DBVPG6044 using BLAST+, version 2.2.29 (Camacho *et al.* 2009) to search each predicted transcript and protein against the predicted transcripts or proteins in the other strain. Specifically, to compare the each predicted transcript in DBVPG6044, we used the *blastn* tool to identify to which predicted transcript in Y12 it aligned best. We then repeated the analysis for each predicted transcript in Y12, using the *blastn* tool to identify to which predicted transcript in DBVPG6044 it aligned best. If the predicted transcripts in each strain were reciprocal best hits, we categorized

them as orthologous to one another. If a predicted transcript did not have a reciprocal best hit, but did align with greater than 90% identity at the nucleotide level across at least 80% of the transcript with at least one predicted transcript in the other strain, we categorized the transcript as paralogous. We performed an analogous analysis for the predicted proteomes of Y12 and DBVPG6044 using the *blastp* tool. Finally, we identified which predicted coding transcripts in each strain did not align anywhere in the genome of the other strain by using the *blastn* tool to attempt to align each predicted coding transcript to the genome we assembled *de novo* from long-read sequencing data generated on a PacBio RSII instrument for the other strain.

RNA expression analysis

To examine the levels of transcription across the genome in each strain, we began by aligning the strand-specific RNA-Seq reads we generated with Illumina for each strain to the strain's genome we assembled *de novo* from long-read sequencing data generated on a PacBio RSII instrument. Specifically, for each strain, we aligned the RNA-Seq fastq files, previously generated by Trinity, in which the 5' three base pairs of each read was trimmed, to our *de novo* genome assemblies using STAR, version 2.5.2a (https://github.com/alexdobin/STAR) in two-pass mode with *--outFilterMultimapNmax 1 --alignIntronMax 20000 --alignMatesGapMax 20000 --outMultimapperOrder Random --outSAMmultNmax 1 --outSAMtype BAM SortedByCoordinate --bamRemoveDuplicatesType UniqueIdentical --twopassMode Basic*. We then removed all duplicate reads from the output bam files with the Picard, version 2.0.6 (http://broadinstitute.github.io/picard/) *MarkDuplicates* tool and the options *VALIDATION_STRINGENCY=SILENT REMOVE_DUPLICATES=true*. We then used Samtools, version 1.3 (Li *et al.* 2009b) *mpileup* tool with the options *–q 1 –B* to generate a pileup

for each of the alignments. To establish whether a particular position in the genome of each strain was expressed from one or both strands, we determined whether the read base column at that position in the pileup we generated for the strain contained a dot, which represents a match of a read to the reference base on the forward strand; a comma, which represents a match of a read to the reference base on the reverse strand; or there were both. If both a dot and a comma were present, we classified that position of the genome as expressed from both strands; if only a dot or a comma were present, we classified that position as expressed from only one strand; if neither was present, we classified that position as not expressed.

We also determined which regions of each genome contained greater than 50 bp of contiguously unexpressed sequence. For this analysis, we simulated 75-bp single-end reads from across each *de novo* genome as in Connelly *et al.* (2014), and then mapped the simulated reads back to the genome and removed duplicates, as described above. We next used Samtools *depth* tool to calculate the depth at each position in the genome for each of the real RNA-Seq data alignments as well as the simulated data alignments. We searched each strain's genome for windows of 50 bp or longer in which the simulated data aligned to the genome, but there were no reads aligning in the real data. For each unexpressed region, we used the BEDTools, version 2.25.0 *closest* tool with the option *–t all* to find the closest predicted transcript. We assigned names to each of the predicted transcripts that had a reciprocal best BLAST match, as previously described, with an annotated S288c gene (annotations obtained from the *Saccharomyces* Genome Database; http://www.yeastgenome.org; Engel *et al.* 2013). Then, using these gene names for the predicted transcripts, we performed a Gene Ontology (GO analysis) for the set of predicted transcripts nearest to unexpressed regions of each strain's genome with the WEB-based Gene SeTAnaLysisToolkit (WebGestalt) GO Analysis tool (Zhang *et al.* 2005; Wang *et al.* 2013). We

selected the *scerevisiae__genome* as a Reference Set for Enrichment Analysis, *Hypergeometric* as the Statistical Method, *BH* as the Multiple Test Adjustment, *Top10* as the Significance Level, and *2* as the Minimum Number of Genes for a Category. Additionally, for each strain, we examined the overlap of our set of unexpressed regions with a set of approximately 400 experimentally determined Autonomously Replicating Sequences (ARSs; (Siow *et al.* 2012; Liachko *et al.* 2013; McGuffee *et. al* 2013), also using the BLAST+, version 2.2.29 *blastn* tool.

Measuring allele-specific expression in a diploid hybrid

We mapped the RNA-seq reads from the Y12 (*MATa*) x DBVPG6044 (*MATα*) diploid hybrid to the PacBio-based Y12 (*MATa*) assembly, the PacBio-based DBVPG6044 (*MATα*) assembly, the S288c genome polymorphised with Y12 (*MATa*), and the S288c genome polymorphised with DBVPG6044 (*MATα*). We first trimmed the reads with Trim Galore!, version 0.4.1 (https://github.com/FelixKrueger/TrimGalore) and the options *--clip_R1 3 --three_prime_clip_R1 1 --three_prime_clip_R2 4*. We then aligned the reads to each genome using STAR, version 2.5.2a (https://github.com/alexdobin/STAR) in two-pass mode with *--genomeSAindexNbases 11 --outFilterMultimapNmax 1 --alignIntronMax 20000 --alignMatesGapMax 20000 --outSAMprimaryFlag AllBestScore --outMultimapperOrder Random --outSAMmultNmax 1 --outSAMtype BAM SortedByCoordinate --outSAMunmapped Within KeepPairs --bamRemoveDuplicatesType UniqueIdentical --twopassMode Basic*.

To obtain a set of variants for allele-specific read calling in the Y12 (*MATa*) x DBVPG6044 (*MATα*) diploid, we used MUGSY (Angiuoli *et al.* 2010) to compute whole-genome alignments that are not biased in favor of a particular reference genome between the two PacBio-based genomes and, separately, the two polymorphised genomes. We cataloged all SNVs

identified in each of the alignments. As we were only interested in transcribed differences between the Y12 and DBVPG6044 genomes, we removed from our variant lists all sites not within annotated transcripts. For the PacBio-based genomes, our set of annotated transcripts consisted of all predicted DBVPG6044 transcripts (based on our RNA-seq data, as previously described) that had a reciprocal best BLAST hit in the set of predicted Y12 transcripts (note: variants that resided in a transcript annotation in one strain, but not the other, were not included). For the polymorphised genomes, we used the annotated S288c open reading frames (obtained from the *Saccharomyces* Genome Database; http://www.yeastgenome.org; Engel *et al.* 2013) and their corresponding untranslated regions (UTRs) (UTR lengths were determined from Nagalakshmi *et al.* 2008). We also removed any variants to which reads obtained from whole genome sequencing of the Y12 (*MATa*) x DBVPG6044 (*MATα*) diploid had a calculated ratio of reads matching the DBVPG6044 allele to total reads at that site that was less than 0.25 or greater than 0.75.

We determined whether a transcript exhibited allele-specific steady-state expression differences following the method exactly as described in Andrie *et al.* (2014). For comparison between levels of allele-specific steady-state expression in the PacBio-based genome assemblies and the polymorphised genome assemblies, we considered only the transcripts for which we were able to identify a reciprocal best BLAST hit between the annotated S288c transcripts and the predicted DBVPG6044 transcripts that also had a reciprocal best BLAST hit with a predicted Y12 transcripts.

Data Availability

All raw sequencing data is available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession numbers SRR5168994, SRR5168993, SRR5170142, SRR5170141, SRR5171570, SRR5171571, and SRR5176941. The PacBio-based assemblies of the nuclear genomes for the Y12 and DBVPG6044 strains are available from the NCBI Whole Genome Shotgun (WGS) database under accession numbers SRS1915938, and SRS1915939.

RESULTS AND DISCUSSION

Whole-genome sequencing and *de novo* assembly of two *Saccharomyces cerevisiae* strains

Using the PacBio single-molecule, real-time (SMRT) sequencing platform, we generated 145–fold and 315-fold coverage, respectively, for two wild-derived haploid strains of *Saccharomyces cerevisiae*: the Sake strain Y12 and the West-African strain DBVPG6044. From this data, we constructed 11.69 and 11.78 Mb assemblies of the nuclear genomes for the Y12 and DBVPG6044 strains, respectively (Table 1). Our assemblies exhibited high contiguity with an N50 of 787 kb for Y12 and 792 kb for DBVPG6044; most chromosomes assembled into single contigs. The most notable exception was Chromosome XII, which assembled into two contigs, averaging 440 kb and 573 kb that included the unique regions surrounding the rDNA region, as well as 2 – 4 additional contigs, averaging 23 kb, that spanned the rDNA region. Additionally, Chromosome XIV in Y12 assembled into two contigs, which were 693 kb and 105 kb in length.

We compared the assemblies of DBVPG6044 and Y12 visually to identify all insertions, deletions, and structural rearrangements greater than 1.5 kb in length (Figure 1). We found 50 such events across the genome with Chromosome VII harboring the most differences (10 insertions and deletions). Notably, this analysis also revealed that just over 40 kb of the 89.4 kb difference in length between our two assemblies is caused by longer telomere regions in DBVPG6044 as compared Y12. Since we obtained almost twice as much coverage of DBVPG6044 as compared to Y12, these telomeric length differences are most likely a consequence of assembly quality, as opposed to true biological differences.

We additionally characterized the Ty elements and simple repeats in our genome assemblies. We found that Ty elements comprise 266 kb of each genome assembly. By comparison, the S288c reference genome contains 403 kb of Ty element sequence. Thus, roughly

half of the length difference between our assembly of DBVPG6044 and the 12.07 Mb nuclear genome of S288c, can be accounted for simply by differences in the amount of Ty element. Our genome assemblies, as well as S288c, contain approximately 120 kb of simple repeats and 25 kb of low complexity sequence.

Transcriptome sequencing and assembly of two *Saccharomyces cerevisiae* strains for genome annotation

To annotate genes in each of our assemblies, we began by performing strand-specific Illumina-based RNA-Seq at 1300-fold coverage of both the Y12 and the DBVPG6044 strains during mid-log phase growth. Using this data, we created transcriptome assemblies for each strain, comprising of 9,328 and 9,301 transcripts in Y12 and DBVPG6044, respectively (Table 1). We computationally predicted open reading frames (ORFs) in 4,821 and 4,758 transcripts, respectively. Of the remaining non-coding transcripts, 2,360 and 2,307 transcripts, respectively, are on the antisense strand of an ORF. Overall, each transcriptome covered 86% of the nuclear genome sequence of each strain, with 11% of the genome annotated as transcribed from both strands.

Using our transcriptome annotations, we compared the gene content between Y12 and DBVPG6044. We found that 5,779 of the transcripts and 4,201 of the ORFs we annotated in each strain are one-to-one orthologs. Of the remaining transcripts in each strain, approximately 3,000 have greater than 90% identity at the nucleotide level across at least 80% of the transcript with at least one transcript in the other strain, suggesting that the transcripts are paralogous. Interestingly, we found that two ORFs in Y12 and four ORFs in DBVPG6044 were not present

in the genomic sequence of the other strain, though they do align to known ORFs in S288c or to predicted proteins in other wild-derived strains.

In addition to assembling and annotating the transcriptomes of our two strains, we also used our deep RNA-seq coverage to evaluate how much of the nuclear genome is transcribed. We observed expression from both strands in approximately 91.5% of genome in both strains. An additional 5.0% of each genome exhibited expression on one strand only, while just over 3% of each genome was not expressed from either stand (Figure 2).

Given the pervasive expression of most of the yeast genome, we investigated the unexpressed regions in more detail. Across each genome, there were approximately 1,400 regions with greater than 50 bp contiguously unexpressed. Interestingly, we find a significant enrichment of tRNA genes in unexpressed regions as well as genes involved in viral capsid assembly and DNA integration (BH correction for multiple testing; p-value = 0.003, 0.007, and 0.008, respectively) (Zhang *et al.* 2005; Wang *et al.* 2013). Genomic rearrangements are frequently bounded by tRNAs and transposable elements and, additionally, colocalize with origins of replication, particularly origins firing earlier in S-phase (Di Rienzi *et al.* 2009). Using BLAST to align the unexpressed regions in our data set to origins of replication previously shown to be active in S288c (Siow *et al.* 2012; Liachko *et al.* 2013; McGuffee *et. al* 2013), we find that approximately 90 of the unexpressed regions in each strain are one-to-one reciprocal best matches with an origin, and an additional 350 unexpressed regions in each strain exhibit greater than 90% identity to at least one origin. These data suggest mechanisms to minimize the co-occurrence of transcriptional and replication machinery contribute to the landscape of gene expression in yeast (Helmrich *et al.* 2013).

Evaluating *de novo* assembly for mitigating read mapping bias

In theory, aligning RNA-seq reads from non-reference strains to their own *de novo* assembled genome would reduce allele-specific read mapping bias (Degner et al. 2009) compared to approaches that map reads to a reference genome. For example, a common approach for mitigating read mapping bias in functional genomics experiments is to edit the reference genome with polymorphisms identified from short-read sequencing data (referred to as "polymorphising"; Skelly et al. 2009; Connelly et al. 2014). To directly compare read mapping biases in the context of calling allele-specific expression (ASE), we first generated ~600-fold coverage Illumina WGS data for Y12 and DBVPG6044, and polymorphised the S288C reference genome with 61,084 Y12 and 75,234 DBVPG6044 single nucleotide variants (SNVs). Next, we mated the Y12 and DBVPG6044 strains to create a diploid hybrid, and then performed strand-specific Illumina-based RNA-Seq at 125-fold coverage on hybrid yeast grown to mid-log phase in rich medium. Finally, we compared estimates of ASE in the diploid between RNA-Seq reads aligned to our PacBio *de novo* assemblies versus the polymorphised genomes in the set of 4,103 one-to-one orthologous transcripts shared between S288c, Y12, and DBVPG6044.

The estimated proportion of genes showing ASE (posterior probability > 0.95) was not significantly different between the *de novo* and polymorphised genomes (~44% of genes with ASE in both approaches; $\chi^2 = 0.96$, *p-value* = 0.33), and 89% of significant genes overlapped between the two data sets. Thus, these results demonstrate that broad scale patterns of ASE are concordant when using the *de novo* versus polymorphised genomes in this data set. However, it is important to point out that Y12 and DBVPG6044 are equally distantly related to the S288c strain (Bergström *et al.* 2014), and diploids constructed from parental strains that differ in levels

of divergence to a reference strain would exhibit more recalcitrant read mapping bias in polymorphised genomes.

Conclusions

We generated high-quality *de novo* genome and transcriptome assemblies from two strains of budding yeast, and leveraged these data to develop comprehensive catalogs of sequence and structural variation and test hypotheses of genome structure and function. Our methods are directly applicable to other yeast strains, and as the price of long-range sequencing data continues to decline, we anticipate that *de novo* genome assemblies will become standard. Ultimately, high-quality *de novo* genome assemblies will allow a more complete collection of the molecular parts list to be developed, which will empower genotype-phenotype inferences and testing hypotheses about the evolution of molecular and organismal traits.

ACKNOWLEDGMENTS

We thank Riza Daza and Jay Shendure for supplying the equipment necessary for Pulse Field Gel Electrophoresis and for assistance in running the gels. We also acknowledge Jennifer Madeoy for preparing the DNA libraries for Illumina whole-genome sequencing of DBVPG6044 and Y12 and Kelsey Lynch and Elizabeth Kwan for providing a curated list of experimentally derived ARSs. This work was supported, in part, by US National Institutes of Health (NIH) grants R01HG002385 and U24HG009081 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute.

FIGURES

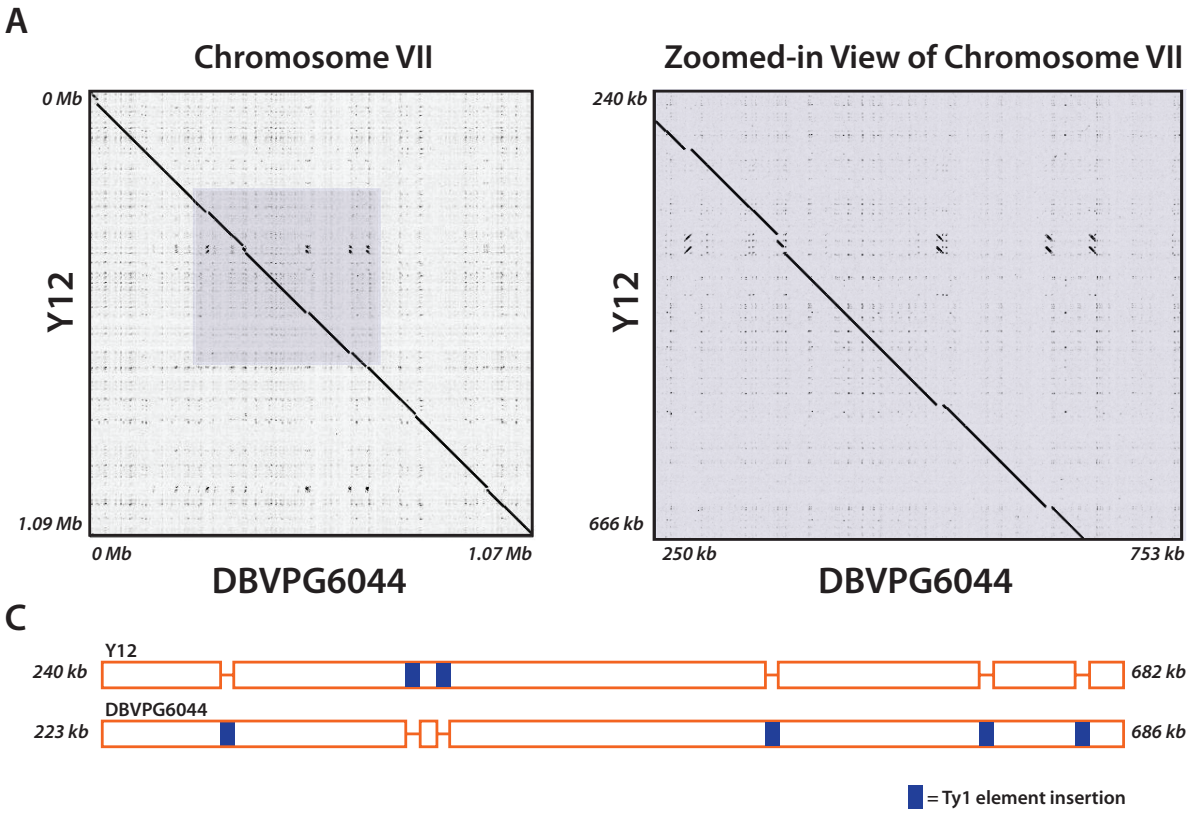
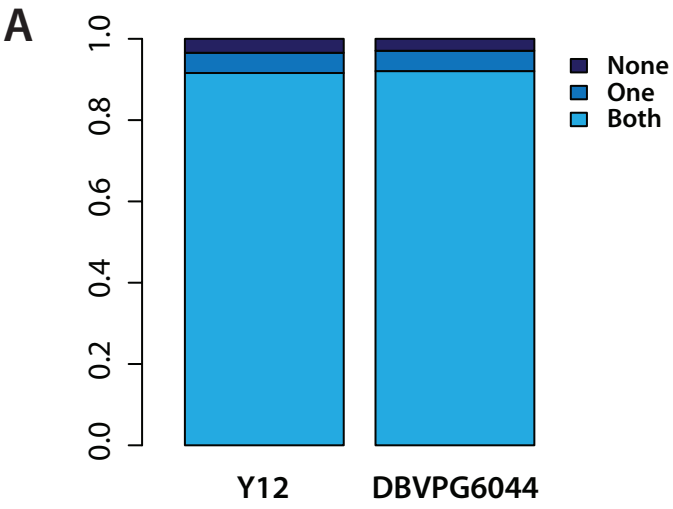


Figure 1. Structural variation between DBVPG6044 and Y12 on Chromosome VII due to differential Ty element insertion. A) Gepard dot plot showing the alignment between Chromosome VII in DBVPG6044 (horizontal axis) and Chromosome VII in Y12 (vertical axis). The diagonal black line represents matches between the Y12 and DBVPG6044 sequences. The shaded (dark gray) region in the center of the plot is enlarged in (B). B) Gepard dot plot showing a region of Chromosome VII in which there are six insertions/deletions, represented as gaps in the diagonal line representing the sequence alignment. C) Diagram of the region shown in (B). The Y12 sequence, top, and DBVPG6044 sequence, bottom, are represented by orange boxes. Orange lines represent gaps in the alignment of that sequence with the other strain's sequence. Insertions of the Ty1 repetitive element in each sequence are represented by blue rectangles.


A
1.0
0.8
0.6
0.4
0.2
0.0
None
One
Both
Y12
DBVPG6044

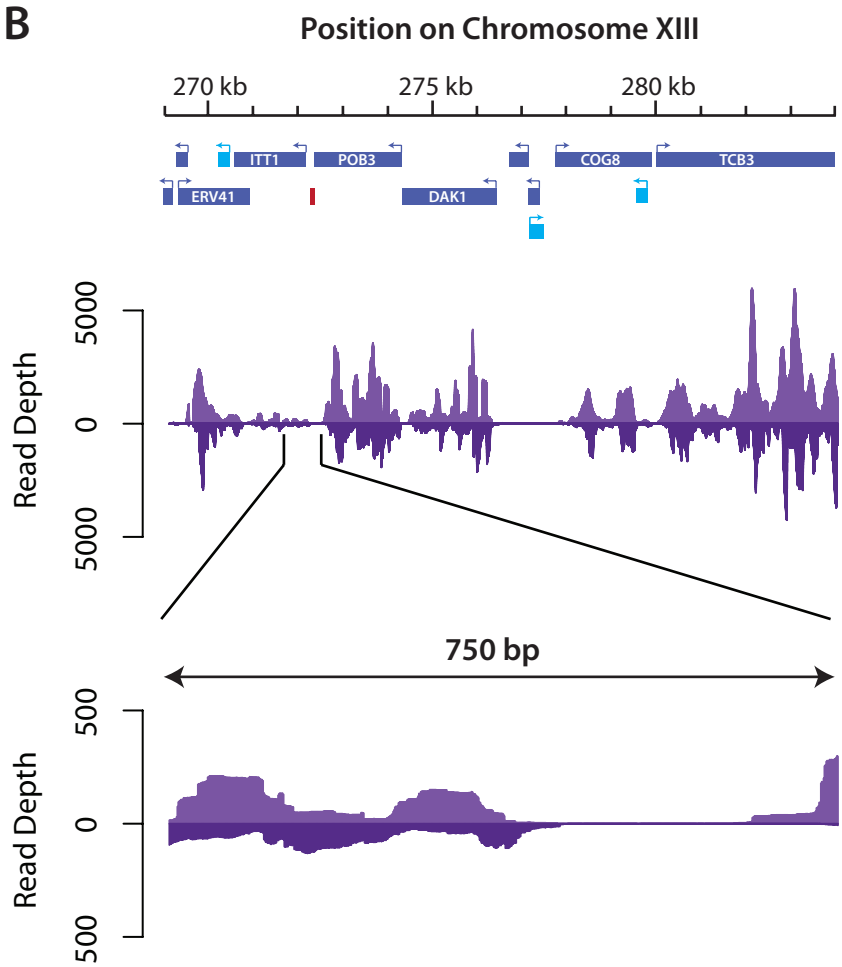

B
Position on Chromosome XIII
270 kb
275 kb
280 kb
ITT1
POB3
COG8
TCB3
ERV41
DAK1
Read Depth
5000
0
5000
750 bp
Read Depth
500
0
500

Figure 2. Pervasive expression of the yeast genome. A) Stacked bar plot showing the proportion of the genome expressed from both strands (light blue), one strand only (medium blue), or neither strand (dark blue) in the Y12 and DBVPG6044 strains, as determined from 1300X RNA-Seq coverage. B) RNA-Seq coverage for a region of Chromosome XIII in DBVPG6044. The top diagram shows the position on Chromosome XIII of the transcripts we annotated in the DBVPG6044 transcriptome for this region. The arrow on each transcript marks the direction of transcription. Transcripts shown in medium blue have a reciprocal best match by BLAST with a transcript we predicted in the Y12 strain; transcripts shown in light blue do not. Additionally, transcripts labeled in white with a gene name are orthologous (i.e. they have a reciprocal best match by BLAST) to the annotated S288c gene of the same name. The red rectangle at approximately 272 kb marks a region of >50 bp of contiguously unexpressed sequence. This particular unexpressed region has a reciprocal best match by BLAST with an unexpressed region we identified in the Y12 strain, and, additionally, has a reciprocal best match by BLAST with an active origin of replication in S288c. Below this diagram, the top graph shows the RNA-Seq read depth we measured on the Watson (light purple) and Crick (dark purple) strands for the region. The bottom graph shows a close-up of the read depth in a 750 bp region surrounding the contiguously unexpressed sequence located at 272 kb.

TABLES

Table 1. Genome and transcriptome assembly statistics.

Strain	Y12	DBVPG6044
Nuclear genome assembly size (bp)	11,688,012	11,777,415
Fold-coverage for PacBio whole genome sequencing	145	315
Fold-coverage at read length >15 kb for PacBio whole genome sequencing	45	63
Nuclear genome assembly N50 (bp)	786,744	791,957
Nuclear genome assembly N90 (bp)	404,883	431,505
Number of contigs in nuclear genome assembly	23	23
Number of rDNA region contigs in nuclear genome assembly	4	2
Ty Element content of nuclear genome assembly (bp)	266,369	265,954
Simple repeat content of nuclear genome assembly (bp)	119,651	124,063
Low complexity sequence content of nuclear genome assembly (bp)	25,740	23,528
Number of transcripts in transcriptome assembly	9,328	9,301
Transcriptome assembly size (bp)	11,467,627	11,240,433
Transcriptome assembly N50 (bp)	1,973	1,941
Predicted number of ORFs in transcriptome	4,821	4,758
Predicted number of antisense transcripts in transcriptome	2,360	2,307

REFERENCES

Andrie, J. M., J. Wakefield, and J. M. Akey, 2014 Heritable variation of mRNA decay rates in yeast. Genome Res. 24: 2000-2010.

Angiuoli, S. V., and S. L. Salzberg, 2011 Mugsy: Fast multiple alignment of closely related whole genomes. Bioinformatics 27: 334-342.

Bergström, A., J. T. Simpson, F. Salinas, B. Barré, L. Parts *et al.*, 2014 A high-definition view of functional genetic variation from natural yeast genomes. Mol. Biol. Evol. 31: 872-888.

Burton, J. N., A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman *et al.*, 2013 Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. 31: 1119-1125.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Carr, M., D. Bensasson, and C. M. Bergman, 2012 Evolutionary genomics of transposable elements in Saccharomyces cerevisiae. PLoS One 7: e50978.

Chaisson, M. J., R. K. Wilson, and E. E. Eichler, 2015 Genetic variation and the de novo assembly of human genomes. Nat. Rev. Genet. 16: 627-640.

Connelly, C. F., J. Wakefield, and J. M. Akey, 2014 Evolution and genetic architecture of chromatin accessibility and function in yeast. PLoS Genet. 10: e1004427.

Degner, J. F., J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori *et al.*, 2009 Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25: 3207-3212.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491-498.

Di Rienzi, S. C., D. Collingwood, M. K. Raghuraman, and B. J. Brewer, 2009 Fragile genomic sites are associated with origins of replication. Genome Biol Evol. 1: 350-63.

Dong, Y., M. Xie, Y. Jiang, N. Xiao, X. Du *et al.*, 2013 Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). Nat. Biotechnol. 31: 135–41.

Engel, S. R., and J. M. Cherry, 2013 The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple Saccharomyces cerevisiae strains at the Saccharomyces Genome Database. Database 2013:bat012.

Frank, J., C. Dingemanse, A. M. Schmitz, R. H. Vossen, G. J. van Ommen *et al.*, 2016 The Complete Genome Sequence of the Murine Pathobiont Helicobacter typhlonius. Front. Microbiol. 6: 1549.

Gallone, B., J. Steensels, T. Prahl, L. Soriaga, V. Saels *et al.*, 2016 Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. Cell 166: 1397-1410.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29: 644-652.

Haas, B. J., A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr. *et al.*, 2003 Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31: 5654-5666.

Helmrich, A., M. Ballarino, E. Nudler, and L. Tora, 2013 Transcription-replication encounters, consequences and genomic instability. Nat. Struct. Mol. Biol. 20: 412-418.

Kim, K. E., P. Peluso, P. Babayan, P. J. Yeadon, C. Yu *et al.*, 2014 Long-read, whole-genome shotgun sequence data for five model organisms. Sci. Data. 1: 140045.
Krumsiek, J., R. Arnold, and T. Rattei, 2007 Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23: 1026-1028.
Li, H., and R. Durbin, 2009a Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-60.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009b The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078-2079.
Liachko, I., R. A. Youngblood, U. Keich, and M. J. Dunham, 2013 High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast. Genome Res. 23: 698-704.
McGuffee, S. R., D. J. Smith, and I. Whitehouse, 2013 Quantitative, genome-wide analysis of eukaryotic replication initiation and termination. Mol. Cell. 50: 123-135.
McIlwain, S. J., D. Peris, M. Sardi, O. V. Moskvin, F. Zhan *et al.*, 2016 Genome Sequence and Analysis of a Stress-Tolerant, Wild-Derived Strain of Saccharomyces cerevisiae Used in Biofuels Research. G3 6: 1757-1766.
McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297-303.
Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha *et al.*, 2008 The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-1349.
Shendure, J., and E. Lieberman Aiden, 2012 The expanding scope of DNA sequencing. Nat. Biotechnol. 30: 1084-1094.
Siow, C. C., S. R. Nieduszynska, C. A. Müller, and C. A. Nieduszynski, 2012 OriDB, the DNA replication origin database updated and extended. Nucleic Acids Res. 40: D682-D686.
Sohn, J. I., and J. W. Nam, 2016 The present and future of de novo whole-genome assembly. Brief Bioinform. pii: bbw096.
Strope, P. K., D. A. Skelly, S. G. Kozmin, G. Mahadevan, E. A. Stone *et al.*, 2015 The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome Res. 25: 762-74.
Trapnell, C., L. Pachter, and S. L. Salzberg, 2009 TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111.
Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel *et al.*, 2013 From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43: 11.10.1-11.10.33.
Wang, J., D. Duncan, Z. Shi, and B. Zhang, 2013 WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 41: W77-W83.
Wu, T. D., and C. K. Watanabe, 2005 GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859-1875.
Zhang, B., S. Kirov, and J. Snoddy, 2005 WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. 33: W741-W748.
Zhu, Y. O., G. Sherlock, and D. A. Petrov, 2016 Whole genome analysis of 132 clinical Saccharomyces cerevisiae strains reveals extensive ploidy variation. G3 6: 2421-2434.
Zowawi, H. M., B. M. Forde, M. Alfaresi, A. Alzarouni, Y. Farahat *et al.*, 2015 Stepwise evolution of pandrug-resistance in Klebsiella pneumoniae. Sci. Rep. 5: 15082.

A cradle-to-grave analysis of *cis*-regulatory variation in yeast

Jennifer M. Andrie[1], David R. Morris[2], Jon Wakefield[3], and Joshua M. Akey[1]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA.

[2]Department of Biochemistry, University of Washington, Seattle, Washington, USA.

[3]Department of Statistics, University of Washington, Seattle, Washington, USA.

INTRODUCTION

Cis-regulatory variation is an important source of phenotypic variation within populations and a major target of adaptive divergence between species. In humans, *cis*-regulatory variation impacts susceptibility to autoimmune, infectious, neoplastic, neurodegenerative, and psychiatric diseases (Skelly *et al.* 2009). In Darwin's finches, cis-regulatory variation affecting beak morphology contributed their speciation (Abzhanov *et al.* 2004). However, despite the critical role *cis*-regulatory variation plays in producing phenotypic variation, our understanding of how *cis*-regulatory variation affects the multi-stage process that converts the information contained in the genome into observable organismal phenotypes remains poor.

Enabled by the recent, rapid expansion in functional genomics technologies that measure the interactions between DNA, RNA, and protein, as well as properties of these biomolecules, like structure or dynamic behavior, several studies have begun to address how *cis*-regulatory variation affects the individual molecular phenotypes that connect genotype to phenotype. For example, in their study comparing chromatin accessibility and function between the two yeast strains *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*, Connelly *et al.* showed that *cis*-regulatory effects on chromatin accessibility were not significantly associated with *cis*-regulatory effects on RNA expression levels, and suggest that the chromatin structure changes caused by *cis*-regulatory variation may be counter-balanced by compensatory changes in mRNA stability, or may simply be functionally benign (2014). In our own study of the effects of heritable variation on mRNA decay rates in two divergent strains of *S. cerevisiae*, we also found extensive evidence of compensatory changes in the *cis*-regulatory variation influencing RNA

transcription rates and decay rates (Andrie *et al.* 2014). Albert *et al.* examined the influence of genetic variation on translation the same two *S. cerevisiae* strains, and found that *cis*-acting effects on translation were of small magnitude and only subtly modulated RNA expression differences.

Also using functional genomics technologies, two studies recently moved beyond individual phenotypes and coordinately investigated how genetic variation, including *cis*-regulatory variation, influences steady-state RNA, translation efficiency, and steady-state protein abundance in human lymphoblastoid cell lines (LCLs) (Battle *et al.* 2015; Cenik *et al.* 2015). Motivated by the observation that variation in mRNA and protein expression levels are often uncorrelated (Foss *et al.* 2007; Ghazalpour *et al.* 2011; Picotti 2013; Albert *et al.* 2014), these studies sought to determine whether genetic variation affecting translation could account for the lack of correlation between RNA and protein levels. Battle *et al.* detected a scarcity of translation-specific QTL, and therefore, infer that attenuation of mRNA expression levels at the protein level is likely due to post-translational processes, as opposed to differences in translation (2015). Conversely, Cenik *et al.* find that ribosome occupancy correlated better with protein levels than RNA abundance correlated with protein levels, and therefore, argue that genetic variation may penetrate to phenotype through changes in translation (2015).

To more accurately delineate the effects of *cis*-regulatory variation on the conversion of genotype into phenotype, we performed synchronous measurement of allele-specific differences in six functional genomics phenotypes, spanning the cradle to grave of gene expression, including chromatin accessibility, transcription rate, RNA binding protein (RBP) binding, RNA secondary structure, RNA stability, and translation

rates in two genetically diverse, wild-derived strains of *S. cerevisiae*. We demonstrate that *cis*-regulatory variation has pervasive effects on high-dimensional molecular phenotypes, and pleiotropy is a predominant feature in the architectural landscape of *cis*-regulatory mutations. Our comprehensive data also provides novel mechanistic insights into *cis*-regulatory variation.

MATERIALS AND METHODS

Yeast strains and diploid hybrid construction

We obtained the wild-derived haploid *S. cerevisiae* strains Y12 (*MATa*) and DBVPG6044 (*MATα*) from our strain collection stored in glycerol at -80°. We inoculated cells from the glycerol stocks into tubes of 5 mL of liquid yeast extract peptone dextrose (YEPD), incubated the tubes in a rotating tube rack at 30° overnight, and then streaked 50 μL of culture onto YEPD plates. We incubated the plates at 30° for 24 hours, until single colonies were visible, and then wrapped the plates in parafilm and stored them at 4°.

We mated Y12 (*MATa*) and DBVPG6044 (*MATα*) to create a hybrid diploid by mixing cells from single colonies of each strain together on a new YEPD plate and incubating the plate at 30° for 6 hours. We scraped some of the mixed cells from the plate, diluted those cells into ultrapure H_2O to create a cell slurry, plated the cell slurry onto a YEPD plate, and incubated the plate at 30° for 24 hours. The resulting cell colonies exhibited clear division into two size groups; we selected three of the larger colonies as candidate diploid hybrids and created glycerol stocks for these three colonies following standard procedures. The stocks were added to our strain collection stored at -80°.

We confirmed that our candidate diploid hybrids identified in our screen were, in fact, diploid using a standard Halo Mating Type Assay as well as by a Polymerase Chain Reaction (PCR) assay. For the PCR-based assay, we began by inoculating cells from each of the three colonies present on the YEPD plate into tubes of 5 mL of liquid YEPD and incubated the tubes in a rotating tube rack at 30° overnight. We then extracted the DNA from these cultures as follows: We transferred 1.5 mL of culture to a screw top tube,

centrifuged the tube at room temperature for 3 min at 14,000 rpm, and poured off the supernatant. We added 200 µL of lysis buffer (10 mM Tris-HCl, pH=8.0; 1 mM EDTA; 100 mM NaCl; 1% SDS; 2% Triton X-100), 0.3 mL of glass beads (425-600 µm, Sigma-Aldrich, G-9268), and 200 µL of phenol:chloroform:isoamyl alcohol to the cell pellet, vortexed at top speed for 2 min, added 200 µL of TE buffer (10 mM Tris-HCl, pH=8.0; 1 mM EDTA), and briefly vortexed again. We centrifuged the tube at 4° for 5 min at 14,000 rpm, transferred the aqueous phase to a new screw top tube, and added 400 µL of phenol:chloroform:isoamyl alcohol. After mixing by inverting the tube several times, we again centrifuged the tube at 4° for 5 min at 14,000 rpm and transferred the aqueous phase to a 1.7 mL Eppendorf tube. We added two volumes of room temperature 100% ethanol, centrifuged the tube at 4° for 2 min at 14,000 rpm, and discarded the supernatant. We washed the nucleic acid pellet twice with 500 µL of ice-cold 70% ethanol, centrifuged the tube at 4° for 2 min at 14,000 rpm, discarded the supernatant, and placed the open tube in a tube rack at room temperature for approximately 10 min to dry the pellet. We resuspended the pellet in 200 µL TE buffer. To perform each PCR reaction, we combined 1 µL of the template DNA solution, 1 µL of 10X *Taq* (Mg-free) Reaction Buffer (M0320S New England BioLabs, Inc., Ipswich, MA), 0.5 µL of 2 mM dNTPs, 0.5 µL of 50 mM $MgCl_2$, 0.1 µL of 10 µM Primer mix, 0.1 µL of *Taq* (Mg-free) (M0320S New England BioLabs, Inc., Ipswich, MA), and 6.8 µL of ultrapure H_2O, to a final volume of 10 µL. We performed two PCR reactions per colony: one that contained the forward primer for the *MATα* locus (5'-TTACTCACAGTTTGGCTCCGGTGT-3') and the reverse primer for both *MAT* loci (5'-GAACCGCATGGGCAGTTTACCTTT-3'); and one that contained the forward primer for the *MATa* locus (5'-

CTCCACTTCAAGTAAGAGTTTGGG-3') and the same reverse primer as above. We heated the PCR reactions to 95° for 3 min, performed 30 cycles of 45 seconds at 95°, 30 seconds at 62°, and 45 seconds at 72°, and ended with an incubation 72° for 5 min before holding the reactions at 4°. We checked the products of the PCR reactions using a standard gel electrophoresis assay.

Yeast cell culture and sample collection

We inoculated cells from the glycerol stock of one of the Y12xDBVPG6044 diploid hybrid colonies, which we generated as described above, into a tube of 5 mL of liquid YEPD. We incubated the tube in a rotating tube rack at 30° overnight, inoculated 0.5 mL of the resultant culture into 50 mL of liquid YEPD in a 250 mL Erlenmeyer flask, and then incubated the flask at 30° for 5 – 7 hours at 225 rpm. We inoculated 2.22×10^5 cells of the resulting 50 mL culture into 500 mL of liquid YEPD in a 2000 mL Erlenmeyer flask. To calculate the volume of culture required for 2.22×10^5 cells, we measured the OD_{660} of the 50 mL culture using a spectrophotometer, and then calculated the cell density of the culture with a chart (obtained from http://www.pangloss.com/seidel/Protocols/ODvsCells.html and originally created by a researcher at UC Boulder) that correlates OD_{660} with the number of yeast cells per mL of culture. We incubated the 500 mL culture at 30° for exactly 15.5 hours at 225 rpm, until it reached mid-log phase ($OD_{660} = 0.8 – 0.9$).

To collect our samples, we removed, in rapid succession, the appropriate cell culture volume from the large culture for each assay that we planned to conduct. First, two aliquots of 50 mL of culture were removed for use in the polysome fractionation

assay. Next, two aliquots of 25 mL of culture were removed for use in the Nuclear Run-On (NRO) assay. We then removed four aliquots of five million cells (we determined the volume of culture needed for five million cells using the same chart as used above, which correlates OD_{660} with the number of yeast cells per mL of culture): we used two of the aliquots for the Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) and two of the aliquots for the RNA proximity ligation (RPL) assay. For each of the assays listed above, we collected two aliquots so that if an error occurred during preparation of the first aliquot, we would still have a second aliquot we could use for that particular assay in that particular replicate. Then, we removed an aliquot of 110 mL of culture into a 1000 mL Erlenmeyer flasks to be used for the mRNA decay rate time course. Finally, we removed 40 mL of culture for use in the Protein Interaction Profile sequencing (PIP-seq) assay. From the time of beginning our collection of the cells used in the polysome fractionation assay to beginning our collection of the cells used in the PIP-seq assay, no more than twenty minutes elapsed. Each of the individual assays we conducted on the collected cell samples is described below. In total, we collected three biological replicates in the manner described above.

For use as a control, we streaked glycerol stock of one of the Y12xDBVPG6044 diploid hybrid colonies onto a YEPD plate and then incubated the plates at 30°C for 48 hours. We inoculated a single colony of the diploid hybrid into 5 mL of liquid YEPD and then incubated in a rotating shaker at 30°C for 21 hours.

Hybrid diploid DNA library construction

We isolated DNA from the 5 mL culture of the Y12xDBVPG6044 diploid hybrid

following the protocol described above for the PCR-based assay to verify mating type, except that we resuspended in 100 µL TE. We quantified the DNA in 1:100 dilutions with the Qubit dsDNA HS Assay Kit (Life Technologies, Grand Island, NY). We prepared sequencing libraries from 1 ng of DNA using Nextera XT DNA Library Preparation Kit (FC-131-1024, Illumina, San Diego, CA).

ATAC-seq library construction

We collected two aliquots of five million cells into 1.7 mL microcentrifuge tubes, centrifuged the tubes at room temperature for 1 min at 14,000 rpm, and decanted the supernatants. We gently resuspended the cells in 200 µL SB buffer (1.4M Sorbitol, 4mM HEPES-KOH, pH 7.5, 0.5mM $MgCl_2$), centrifuged the tubes at room temperature for 1 min at 14,000 rpm, and decanted the supernatant. To spheroblast the cells, we resuspended the cell pellets in a mix of 400 µL SB buffer and 25 µL of a 40 U/mL dilution of 100T zymolase solution, and, after pre-warming the tubes containing the reaction mixture in 30° water bath for 1-2 min, incubated the tubes for 30 min at 30° with shaking at 225 rpm. We centrifuged the tubes at 4° for 5 min at 5,000 rpm, removed the supernatant, and added 50 µL of cold transposition mix (25 µL of TD buffer (Illumina Cat #FC-121-1030), 16 µL of Nextera enzyme, and 15 µL of ultrapure water; we previously determined the appropriate amount of Nextera enzyme using a dilution series). To transpose sequencing adaptors into native chromatin, we incubated the tubes for 30 min at 37°. We immediately purified the reaction product using a Qiagen MinElute Reaction Clean-up Kit, following the manufacturer's instructions with elution into 10 µL of buffer EB (Qiagen, Valencia, CA).

To amplify the transposed DNA fragments, we combined the 10 μL of transposed DNA product, 7.5 μL of nuclease-free water, 2.5 μL of a Nextera N7 Index, 2.5 μL of a Nextera N5 Index, 2.5 μL of Nextera PCR Primer Cocktail, and 25 μL of NEBNext High-Fidelity 2x PCR Master Mix (M0541, New England BioLabs, Inc., Ipswich, MA) in a PCR tube, and then performed PCR following the conditions specified in Buenrostro *et al.* 2013, except that we used ten thermocycles and did not monitor the PCR reaction using qPCR. We purified the PCR products using a Qiagen MinElute PCR Purification kit, following the manufacturer's instructions with elution into 20 μL of buffer EB (Qiagen, Valencia, CA).

We verified the quality of the library using an Agilent DNA 1000 kit on the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) as well by examining it with a standard gel electrophoresis assay. We quantified the concentration of the library with Qubit dsDNA BR Assay (Life Technologies, Grand Island, NY).

NRO assay

To analyze transcription on a genome-wide level, we performed an NRO assay as described in McKinlay *et al.* 2011, using all cells collected in the 25 mL aliquots of culture we reserved for the NRO assay. After precipitating the RNA labeled with Biotin-16-UTP and resuspending it in nuclease-free water, we quantified the concentration of the RNA and assessed its quality with an Agilent RNA Pico kit on the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA), before proceeding with strand-specific RNA-Seq library construction using the SMARTer® Stranded RNA-Seq kit following the manufacturer's instructions, as described below (Takara Bio USA, Inc., Mountain View,

CA). For the first and third replicates, we built one "labeled" RNA library; for the second replicate, we built two libraries. For the aliquot of total RNA we sequenced from each replicate for this assay, we performed quality assessments and quantification, removed rRNA, and constructed strand-specific RNA-Seq libraries exactly as described below (see RNA-Sequencing library construction).

RNA decay rate time course

To examine genome-wide rates of RNA decay, we added 110 μL of 100 mg/mL 1,10-phenanthroline to our 110 mL aliquot of yeast cell culture, to achieve a final concentration of 100 μg/mL 1,10-phenanthroline. Immediately following, and at 15, 20, and 30 min after addition of the drug, we collected 25 mL aliquots of the culture using vacuum filtration onto a 0.45 micron nylon filter (R04SP04700, Fisher Scientific, Waltham, MA). The filter was placed in a 15 mL Falcon™ tube and flash frozen in liquid nitrogen. To maintain the temperature of the culture, we housed it in a 30° incubator at 225 rpm between collection time points. The collected yeast cells from each time point were stored at -80° for no more than 2 days before we extracted total RNA from the cells using a standard phenol-chloroform preparation: We added 1 mL of ice-cold TES buffer (10 mM Tris-HCl, pH=7.5; 10 mM EDTA; 0.5% SDS) to each 15 mL tube, and then alternated vortexing at top speed for 15 seconds with resting on ice for 15 seconds until all the cells were washed off the filter. To pull the cell slurry to the bottom of the tube, we spun the tube in a tabletop centrifuge until the centrifuge reached 1,000 rpm. We transferred the cell slurry to a 1.5 mL screw top tube, added 400 μL of acid phenol pre-warmed to 65°, and then incubated the sample at 65° for one hour, vortexing every 10

minutes during the incubation. We then placed the tube on ice for 5 min before centrifuging at 4° for 5 min at 14,000 rpm, transferring the aqueous phase to a new screw top tube, adding 400 µL of pre-warmed acid phenol, and briefly vortexing. We again centrifuged at 4° for 5 min at 14,000 rpm, transferred the aqueous phase to a new screw top tube, added 400 µL of pre-warmed acid phenol, and briefly vortexed. After centrifuging at 4° for 5 min at 14,000 rpm and transferring the aqueous phase to a new screw top tube a third time, we added 400 µL of room temperature chloroform, and briefly vortexed. We centrifuged the tube at 4° for 5 min at 14,000 rpm and transferred the aqueous phase to a 1.7 mL Eppendorf tube. We added 0.1 volume of 3 M sodium acetate and 2.5 volumes of cold 100% ethanol, and froze the sample at -80° overnight. We centrifuged the tube at 4° for 30 min at 14,000 rpm, and discarded the supernatant. We washed the nucleic acid pellet twice with 500 µL of ice-cold 100% ethanol, centrifuged the tube at 4° for 2 min at 14,000 rpm, discarded the supernatant, and placed the open tube in a tube rack at room temperature for approximately 10 min to dry the pellet. We resuspended the pellet in 200 µL diethylpyrocarbonate-treated (DEPC-treated) H_2O. We quantified the samples using a Qubit® RNA BR Assay (Life Technologies, Grand Island, NY).

To remove DNA from the samples, we combined each 200 µL RNA sample with 230 µL DEPC-treated H_2O, 50 µL of 10X Turbo DNase I Buffer, and 20 µL of Turbo DNase I (2 U/µL) (Life Technologies, Grand Island, NY), and then incubated the mixture at 37° for 30 minutes. We re-isolated the RNA with a phenol:chloroform:isoamyl alcohol extraction: We placed the sample in a screw top tube, added an equal volume of phenol:chloroform:isoamyl alcohol pre-cooled to 4°, briefly vortexed, and centrifuged at

4° for 5 min at 14,000 rpm. We transferred the aqueous phase to a new screw top tube, added an equal volume of room temperature chloroform, and briefly vortexed. We centrifuged the tube at 4° for 5 min at 14,000 rpm and transferred the aqueous phase to a 1.7 mL Eppendorf tube. We added 0.1 volume of 3 M sodium acetate and 2.5 volumes of cold 100% ethanol, froze the sample at -80° for one hour, and then centrifuged the tube at 4° for 30 min at 14,000 rpm. We discarded the supernatant and washed the nucleic acid pellet twice with 500 µL of ice-cold 70% ethanol, centrifuged the tube at 4° for 2 min at 14,000 rpm, discarded the supernatant, and placed the open tube in a tube rack at room temperature for approximately 25 min to dry the pellet. We resuspended the pellet in 200 µL DEPC-treated H_2O and stored the pellet at -80°. We performed quality assessments and quantification, removed rRNA, and constructed strand-specific RNA-Seq libraries exactly as described below (see RNA-Sequencing library construction).

Polysome profiling

To prepare and fractionate polysomes, we modified a previously described method (MacKay *et al.* 2004). Specifically, working in an ice-water bath, we poured each 25 mL aliquot of culture we collected into a 250 mL flask containing 50 µL of 50 mg/mL cyclohexamide stock and 19 mL of crushed frozen YEPD containing 100 µg/mL cyclohexamide. We swirled the flasks for 1 – 2 min to bring culture rapidly down to 0° – 4°. We then transferred each sample to a 50 mL conical tube and centrifuged the tubes in Beckman table top centrifuge at 4° until the rotor speed reached 4,750 rpm. We removed the supernatant and resuspended the cells from each sample in 5 mL of ice-cold Martin lysis buffer (25 mM Tris-HCl, pH 7.5, 7.5 mM $MgCl_2$, 40 mM KCl, 1 mM DTT, 0.5

mg/mL heparin, and 100 μg/mL cyclohexamide). We transfered the cell slurries to new 50 mL tubes, pooling two samples into one tube, and centrifuged in a table top centrifuge as described above. After pouring off as much of the supernatant as possible, we flash froze the cell pellets in liquid nitrogen and stored them at -80°.

Working in an ice-water bath in a 4° cold room with all reagents ice-cold, we resuspended each cell pellet in 1 mL of Martin lysis buffer and then transferred the cell slurries to a 15 mL tube containing 2 g RNase-free glass beads. To lyse the cells, we vortexed each sample eight times for 30 sec each, cooling on ice for 30 sec between pulses. We added 100 μL of 10% Triton X-100 and 100 μL of 10% NaDOC with vortexing, and then held the samples on ice for 5 min before vortexing again. We transferred all liquid from the sample tubes into new microfuge tubes, added 200 μL Martin lysis buffer to the tube containing the glass beads, vortexed again for 15 sec, and then combined this with the rest of the lysate. We centrifuged the samples at 4° for 1 min at 13,000 rpm, and then transferred the supernatants to fresh microfuge tubes.

To perform polysome fractionation, we loaded 20 A_{260} units in 1 mL of Martin lysis buffer onto 10.8 mL linear 7%-47% sucrose gradients in 50 mM Tris-HCl, pH 7.5, 15 mM $MgCl_2$, 700 mM NaCl, 0.5 mg/mL heparin, and 100 μg/mL cyclohexamide. We centrifuged at 39,000 rpm in an SW40 Ti swinging bucket rotor (Beckman) for 2.25 hours at 4°. We collected twenty-four fractions from the top of each gradient into microfuge tubes containing 5 μL of 10% SDS, for a final concentration of 0.1% SDS. During fractionation, we monitored the polysome profiles at 254 nm.

To extract RNA from the polysome fractions, we began by adding SDS to a final concentration of 0.5% SDS and 20 mg/mL Proteinase K stock to a final concentration of

100 μg/mL Proteinase K to each 0.5 mL fraction, as well as a sample of 5 A_{260} units of the total lysate that we loaded onto each gradient. We incubated at 37° for 30 min, and then added an equal volume of acid phenol/chloroform to each sample. We centrifuged at 4° for 5 min at 13,000 rpm and then removed the aqueous phase to a clean tube. We next precipitated with ethanol by adding 10 M ammonium acetate to a final concentration of 2 – 2.5 M and 4 volumes of 100% ethanol. We chilled the samples at –80° for 30 min, and then centrifuged at 4° for 30 min at 13,000 rpm to pellet the RNA. We washed the pellet twice with cold 70% ethanol, and then air-dried the pellet by leaving the tube upside down on the bench for 15 min. We dissolved each RNA pellet in 25 μL of DEPC H_2O and pooled the gradient fractions as follows:

Pool	Fractions
A	1 – 6
B	9 – 11
C	13 – 14
D	15 – 17
E	18 – 19
F	20 – 22

Fraction pool A corresponds to transcripts to which no ribosomes are bound; fraction pool B corresponds to transcripts to which one ribosome is bound; fraction pool C corresponds to transcripts to which two ribosomes are bound; fraction pool D corresponds to transcripts to which three ribosomes are bound; fraction pool E corresponds to transcripts to which four to approximately seven ribosomes are bound; and fraction pool F corresponds to transcripts to which approximately eight or more

ribosomes are bound (Figure 4A). We removed DNA from the samples as described above (see RNA decay rate time course); however, we resuspended the samples in 100 µL of DEPC-treated H_2O before storing them at -80°. To remove heparin from our samples, which can inhibit with reverse transcriptase, we added 1/5 volume of 10 M LiCl to a final concentration of 2 M LiCl to each sample and incubated at -80° overnight. We centrifuged the tubes at 4° for 25 min at 13,000 rpm, removed the supernatant, and then added 1 mL of 70% ethanol, pre-cooled to -20°. We centrifuged at 4° for 5 min at 13,000 rpm and decanted the supernatant. We briefly spun once more, removed the residual ethanol, air-dried the tube for 5 min at room temperature, and then resuspended the pellets in 30 µL of DEPC-treated H_2O. We performed quality assessments and quantification, removed rRNA, and constructed strand-specific RNA-Seq libraries exactly as described below (see RNA-Sequencing library construction).

RPL assay

To examine RNA secondary structure, we performed RPL based loosely on the method described in Ramani *et al.* 2015. Specifically, we collected two aliquots of five million cells into 1.7 mL microcentrifuge tubes and spheroblasted the cells as described above (see ATAC-seq library construction). After spheroblasting the cells, we centrifuged the tubes at 4° for 5 min at 5,000 rpm and removed the supernatant. We then repaired RNA ends that had sheared during spheroblasting by adding 50 µL of T4 PNK mix (5 µL 10X T4 DNA Ligase Reaction Buffer (New England BioLabs, Inc., Ipswich, MA), 0.5 µL SUPERase-In^{TM} RNase Inhibitor (20 U/ µL) (Thermo Fisher Scientific, Waltham, MA), 2.5 µL T4 PNK (to a final concentration of 0.5 U/µL) (New England

BioLabs, Inc., Ipswich, MA), 42 µL 1X PBS + 0.2% NP-40) to each tube and immediately incubating at 37° for 30 min. We transferred the tubes to ice immediately, and then performed intramolecular RNA ligation by adding 450 µL ice-cold ligation mix (50 µL 10X T4 DNA Ligase Reaction Buffer (New England BioLabs, Inc., Ipswich, MA), 5 µL SUPERase-In^{TM} RNase Inhibitor (20 U/ µL) (Thermo Fisher Scientific, Waltham, MA), 12.5 µL T4 RNA Ligase 1 (to a final concentration of 0.25 U/ µL) (New England BioLabs, Inc., Ipswich, MA), 382.5 µL 1X PBS + 0.2% NP-40) to each tube and incubating at 16° for 6 hours.

We extracted total RNA from the cells using a standard phenol-chloroform preparation, as described above (see RNA decay rate time course), except that in the first step, we added 500 µL of acid phenol pre-warmed to 65° directly to the ligation complex mixes, and then incubated the samples at 65° for one hour, vortexing every 10 minutes during the incubation. We removed DNA from the samples as described above (see RNA decay rate time course); however, we resuspended the samples in 20 µL DEPC-treated H_2O before storing them at -80°. We performed quality assessments and quantification, removed rRNA, and constructed strand-specific RNA-Seq libraries exactly as described below (see RNA-Sequencing library construction).

PIP-seq assay

We adapted the PIP-seq assay developed by Silverman *et al.* 2014 for use in yeast cells. We began by placing a 40 mL aliquot of culture on rotating platform (150 rpm) at room temperature and adding 1080 µL of 37% formaldehyde solution dropwise, to a final concentration of 1% formaldehyde to the culture. We continued to rotate the culture for

15 minutes and then added 2 mL of 2.5 M glycine to achieve a final concentration of 125 mM glycine in the culture. We rotated for an additional 5 minutes at room temperature before dividing the culture into four 50 mL conical tubes of 10 mL of culture each, and briefly centrifuging at 4° to pellet the cells. We washed the cell pellets twice in 15 mL ice-cold PBS without resuspending the pellets into solution. We resuspended each aliquot in 1 mL PBS, transferred the cell slurry to a microcentrifuge tube, and centrifuged the tubes at 4° for 2 min at 13,000 rpm. We removed the supernatant and flash froze the pellets in liquid nitrogen. Pellets were stored at -80° until we were ready to proceed with the rest of the protocol.

To lyse the cells, we resuspended each cell pellet in 500 µL of cold freshly-prepared RIP lysis buffer (25 mM Tris-HCl, pH=7.4, 0.5% NP-40, 40 U/mL RNaseOUT™ Ribonuclease Inhibitor (Invitrogen, Carlsbad, CA), 10 µM dithiothreitol (DTT), and 50 µL protease inhibitors) and then added 300 – 400 mg RNase-free glass beads. Working in a 4° cold room, we vortexed each tube ten times for 30 sec each, cooling on ice for 30 sec between pulses. We transferred all liquid from the sample tubes into new microfuge tubes, added 200 µL RIP lysis buffer to the tube containing the glass beads, vortexed again for 15 sec, and then combined this with the rest of the lysate.

For the control samples, we combined 475 µL of lysate with 37.5 µL of RNase Stop/Proteinase K solution (5% SDS, 5 mM EDTA, pH = 8.0, 2 µg/ml Proteinase K) and incubated at room temperature for 15 min to inhibit RNase activity and denature/degrade proteins. We then added of one-tenth of a volume of 3 M sodium acetate and 2.5 volumes ice-cold ethanol to each proteinase-treated lysate sample, stored the samples at -80° for 30 min, and then centrifuged at 4° for 30 min at 14,000 rpm. We washed the pellets twice

with ice-cold 70% ethanol, centrifuged at 4° for an additional 1 min at 14,000 rpm, decanted the supernatant, air dried the pellet at room temperature for 15 min, and, finally, resuspended the sample in 475 µL RIP buffer.

Then, for both the control and experimental samples, we combined 475 µL of each lysate sample with 50 µL of RNase ONE™ buffer and 5 µL of RNase ONE™ Ribonuclease (Promega, Madison, WI). We incubated at 37° for 1 hr, inverting every 15 minute to ensure even digestion, and then added 37.5 µL of RNase Stop/Proteinase K solution (5% SDS, 5 mM EDTA, pH = 8.0, 2 µg/ml Proteinase K) to the samples and incubated at room temperature for 15 min to inhibit RNase activity and denature/degrade proteins. To reverse any remaining crosslinks, we incubated the samples at 65° for 2 hours, mixing every 15 min. We flash froze the samples in liquid nitrogen and stored them at -80° until we were ready to perform RNA extraction.

We extracted total RNA from the cells using a standard phenol-chloroform preparation, as described above (see RPL assay). We removed DNA from the samples as described above (see RNA decay rate time course); however, we resuspended the samples in 100 µL of DEPC-treated H_2O before storing them at -80°. We performed quality assessments and quantification, removed rRNA, and constructed strand-specific RNA-Seq libraries exactly as described below (see RNA-Sequencing library construction).

RNA-Sequencing library construction

We quantified the samples using a Qubit® RNA BR Assay (Life Technologies, Grand Island, NY). We assessed the quality of the samples with an Agilent RNA 6000

Nano assay on the 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA). To prepare the RNA samples for sequencing, we first depleted them of rRNA using the Illumina Ribo-Zero Gold rRNA Removal Kit (Yeast) (Illumina, San Diego, CA). We purified the product using the manufacturer's suggested ethanol precipitation protocol, but omitting the addition of glycerol to our samples. We quantified and assessed the quality of the product of the Illumina Ribo-Zero Gold rRNA Removal Kit (Yeast) with an Agilent RNA 6000 Pico assay on the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). We prepared strand-specific sequencing libraries for each of our samples using the Clontech SMARTer® Stranded RNA-Seq Kit per the manufacturer's instructions (Takara Bio USA, Inc., Mountain View, CA). For each sample, we used 50 ng of input RNA; or, if we did not have 50 ng, we used the entire RNA product we had left after depleting rRNA. We assessed the quality of our libraries with an Agilent High Sensitivity DNA assay on the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). We quantified the libraries with a Qubit® dsDNA HS Assay (Life Technologies, Grand Island, NY).

Sequencing

We sent the hybrid diploid DNA library to the Northwest Genomics Center for sequencing. Cluster generation was performed on an Illumina cBot using standard HiSeq 4000 chemistry cluster kits (Illumina, San Diego, CA). Paired-end 75-bp sequencing was performed on an Illumina HiSeq 4000 using standard SBS chemistry and HiSeq Control Software version 3.3.52 (Illumina, San Diego, CA).

We also sent the ATAC-Seq and all RNA-Seq libraries to the Northwest Genomics Center for sequencing. All sequencing libraries from the same replicate were pooled, for a total of three pools. Cluster generation was performed using an Illumina cBot using standard HiSeq 2500 high output version 4 chemistry cluster kits (Illumina, San Diego, CA). Paired-end 100-bp sequencing was performed on an Illumina HiSeq 2500 in high output mode using SBS version 4 chemistry and HiSeq Control Software version 2.2.58 (Illumina, San Diego, CA). We performed one lane of sequencing for each replicate, with all replicates run simultaneously on the same flow cell. We performed one additional lane of sequencing on a different flow cell for a pool of libraries that included the second "labeled" RNA library we built for the NRO assay from replicate 2, as well as the libraries we previously sequenced from replicate 2 for the NRO "total" RNA, decay rate time course, PIP-Seq, and polysome profiling samples.

Read mapping

For both the Y12 and the DBVPG6044 strain, we obtained nearly complete nuclear genome sequence through assembly of greater than 145-fold PacBio-based long-read sequencing coverage of the genome, as previously described in Andrie *et al.* 2017.

ATAC-Seq

We trimmed the ATAC-seq sequencing reads from each replicate with Trim Galore!, version 0.4.1 (https://github.com/FelixKrueger/TrimGalore) and the options --*nextera –paired* and then aligned the reads separately to both the Y12 and DBVPG6044 genomes using bowtie2, version 2.2.3 (Langmead and Salzberg 2012) and the options -*X*

2000 --no-discordant --no-mixed --sensitive. We converted the SAM-formatted files output by bowtie2 to BAM format with the *view* tool in samtools, version 1.3 (Li *et al.* 2009) using the options *-bhS*. We sorted and indexed the resulting BAM-formatted files with the *sort* and *index* tools in samtools, version 1.3 (Li *et al.* 2009), and then marked duplicate reads with the *MarkDuplicates* tool in Picard, version 2.6.0 (http://broadinstitute.github.io/picard) using the options *VALIDATION_STRINGENCY=SILENT CREATE_INDEX=true*. For use as a control, we mapped the reads we obtained from the hybrid diploid DNA samples following the exact same steps as described above for the ATAC-seq samples.

RNA-based assays

We trimmed reads from all RNA-based assays with Trim Galore!, version 0.4.1 (https://github.com/FelixKrueger/TrimGalore) and the options *--clip_R1 3 --three_prime_clip_R1 1 --three_prime_clip_R2 4 --paired*. For libraries sequenced twice on different lanes, we concatenated the FASTQ files we obtained from the Northwest Genomics Center before trimming the sequencing reads. We aligned the RNA-Seq reads from all assays separately to both the Y12 and DBVPG6044 genomes using STAR, version 2.5.2a (Dobin *et al.* 2013) in two-pass mode with *--genomeSAindexNbases 11 --outFilterMultimapNmax 1 --alignIntronMax 20000 --alignMatesGapMax 20000 --outSAMprimaryFlag AllBestScore --outMultimapperOrder Random --outSAMmultNmax 1 --outSAMtype BAM SortedByCoordinate --outSAMunmapped Within KeepPairs --bamRemoveDuplicatesType UniqueIdentical --twopassMode Basic*. For use as a control, we used Trim Galore!, version 0.4.1 (https://github.com/FelixKrueger/TrimGalore) and

the options --*nextera –paired* to trim the reads we obtained from the hybrid diploid DNA samples. We then aligned these reads separately to both the Y12 and DBVPG6044 genomes with STAR, version 2.5.2a (Dobin *et al.* 2013), exactly as described above.

Allele-specific read calling

Also as previously described in Andrie *et al.* 2017, we obtained transcript annotations for the Y12 and DBVPG6044 genomes by using Trinity (Grabherr *et al.* 2011) and PASA (Haas *et al.* 2003) to assemble the transcriptome of each strain from ultra-deep strand-specific Illumina-based RNA-Seq of each strain.

ATAC-seq

We merged the BAM-formatted alignment files from the three replicates of ATAC-seq with the *merge* tool in samtools, version 1.3 (Li *et al.* 2009). We called peaks of open chromatin in each strain's genome using the *callpeak* tool in MACS, version 2.1.0 (Zhang *et al.* 2008) with the options --*format BAMPE --gsize 11.78e6* and the concatenated BAM-formatted files as the --*treatment* file. Focusing on only the "narrow" peaks annotated by MACS, we filtered out any peaks for which MACS calculated a fold-change less than 1.110. We assigned each remaining peak to the closest transcript using the *closest* tool in BEDTools, version 2.25.0 (http://bedtools.readthedocs.io/en/latest/) to identify the transcript with the nearest transcription start site to the summit that MACS determined for the peak.

To obtain a set of variants for allele-specific read calling for the ATAC-seq assay, we used MUGSY (Angiuoli *et al.* 2010) to compute whole-genome alignments that were not biased in favor of a particular reference genome between the two strain genomes and then cataloged all SNVs identified in the alignment. As we were only interested in differences in chromatin accessibility between the Y12 and DBVPG6044 genomes, we refined our variant list to include only sites within peaks that were shared between the two strains. Specifically, we used the *blastn* tool in BLAST+, version 2.2.29 (Camacho *et al.* 2009) to search each predicted transcript in DBVPG6044 against all predicted transcripts in Y12, and determine to which Y12 transcript it aligned best. We then repeated the analysis for each predicted transcript in Y12, using the *blastn* tool to identify to which predicted transcript in DBVPG6044 it aligned best. For each SNV we identified, we then asked if it resided in an annotated peak in both strains and if that annotated peak was nearest to the same transcript, as defined by reciprocal BLAST best hit, in both strains. In addition to removing SNVs not in shared peak regions, we used our Y12xDBVPG6044 hybrid diploid whole genome sequencing data to identify and, subsequently, remove any variants with mapping biases. More precisely, we subjected the reads obtained from whole genome sequencing of the Y12xDBVPG6044 hybrid diploid to the allele-specific read-calling pipeline we used for our ATAC-seq assays, including alignment with bowtie2, as previously described, and assignment of individual reads to an allele, as we describe below. We then threw out any variants that had either no reads mapping or greater than two standard deviations above the mean number of reads mapping. As we expect that the proportion of reads deriving from a particular allele

should be 50% in the diploid, we also removed any variant with less than 30% or greater than 70% of reads called as matching the DBVPG6044 allele.

We performed assignment of individual ATAC-seq reads and DNA diploid hybrid reads as originating from either the Y12 or the DBVPG6044 allele of each peak as described in Andrie *et al.* 2014, and originally developed by Skelly *et al.* 2011.

RNA-based assays

To obtain a set of variants for allele-specific read calling for all of our RNA-based assays, we used MUGSY (Angiuoli *et al.* 2010) to compute whole-genome alignments between the two strain genomes and then cataloged all SNVs identified in the alignment, as described for the ATAC-seq assay. However, as we were only interested in transcribed differences between the Y12 and DBVPG6044 genomes for the RNA-based assays, we refined our variant list to include only sites within annotated transcripts. More specifically, we retained a variant only if the variant resided in the same transcript, as defined by reciprocal BLAST best hit (described above, see Allele-specific read calling for ATAC-seq), in both strains. Additionally, we used our Y12xDBVPG6044 hybrid diploid whole genome sequencing data to identify and, subsequently, remove any variants with mapping biases. More precisely, we subjected the reads obtained from whole genome sequencing of the Y12xDBVPG6044 hybrid diploid to the allele-specific read-calling pipeline we used for all of our RNA-based assays, except RPL, including alignment with STAR, as previously described, and assignment of individual reads to an allele, as we describe below. We then threw out any variants that had either no reads mapping or greater than two standard deviations above the mean number of reads

mapping. As we expect that the proportion of reads deriving from a particular allele should be 50% in the diploid, we also removed any variant with less than 30% or greater than 70% of reads called as matching the DBVPG6044 allele.

For all RNA-based assays except RPL, we performed assignment of individual RNA-Seq reads and DNA diploid hybrid reads as originating from either the Y12 or the DBVPG6044 allele of each transcript as described in Andrie *et al.* 2014, and originally developed by Skelly *et al.* 2011, with the following modification: Each read was assigned to a parental allele (i.e. Y12 or DBVPG6044), and subsequently, to a transcript overlapping that variant site in a strand-specific manner. If a read overlapped a variant site, but was not on the same strand as an annotated transcript, it was thrown out.

For allele-specific read calling for the RPL assay, we assigned each read to a parental allele (i.e. Y12 or DBVPG6044) as described above, but further processing followed the pipeline developed by Ramani *et al.* in their paper describing the RPL method (2015). We first converted the BAM-formatted files containing the reads that were assigned, through the pipeline developed by Skelly *et al.*, as deriving from either parent to FASTQ format with the *fastq* tool in samtools, version 1.3 (Li *et al.* 2009). Keeping the reads assigned to the Y12 allele separate from the reads assigned to the DBVPG6044 allele, we merged read pairs that contained redundant (i.e., sequence) content using SeqPrep, version 1.2 (https://github.com/jstjohn/SeqPrep) with the options *–g –o 10*, as in Ramani *et al.* (2015). Again keeping the reads assigned to the Y12 allele separate from the reads assigned to the DBVPG6044 allele, we concatenated the resulting fused reads and the remaining "singleton" forward and reverse reads to create one large FASTQ file per strain. Also as in Ramani *et al.*, we aligned the reads in the resulting

FASTQ file to the appropriate strain's transcriptome using STAR, version 2.5.2a (Dobin *et al.* 2013) in two-pass mode with *--genomeSAindexNbases 11 --genomeChrBinNbits=10 --outFilterMultimapNmax 1 --alignIntronMax 20000 --outSAMprimaryFlag AllBestScore --outMultimapperOrder Random --outSAMmultNmax 1 --outSAMunmapped Within --twopassMode Basic --outSJfilterOverhangMin 6 6 6 6 --outSJfilterCountTotalMin 1 1 1 1 --outSJfilterDistToOtherSJmin 0 0 0 0 --alignIntronMin 10 --chimSegmentMin 15 --chimScoreJunctionNonGTAG 0 --chimJunctionOverhangMin 6* (2015). We used the bash command *awk* to filter the aligned reads output by STAR for intronic reads, and then deduplicated those reads using a script written by Ramani *et al.* that collapses all alignments with identical start coordinates and CIGAR strings (2015). Using a custom *awk* script written by A. Dobin (available at https://github.com/alexdobin/STAR/blob/master/extras/scripts/sjFromSAMcollapseUandM.awk), we converted the deduplicated alignments to "splice junction" files that provided the coordinates of each ligation event and the read counts supporting that ligation event.

Statistical methods to measure allele-specific differences

All statistical analyses were performed in the R statistical programming language and environment (R Core Team 2013).

Chromatin accessibility and transcription rate

We used our ATAC-seq data set to identify transcripts with allelic differences in chromatin accessibility. Specifically, we first summed the number of reads deriving from

the Y12 allele and the number of reads deriving from the DBVPG6044 allele across all SNVs in peak regions associated with the same transcript in both strains. Thus, for each transcript shared between the two strains, we obtained total counts of reads deriving from the DBVPG6044 allele and total counts of reads deriving from the Y12 allele in regions of open chromatin associated with the transcript. To determine whether a transcript exhibited allele-specific chromatin accessibility differences, we performed the *cis* test exactly as described by Connelly *et al.*, with the following modification: the test was performed with three, rather than two, replicates (2014). Our primary motivation for choosing this method, as opposed to alternative approaches for measuring allele-specific expression differences (Skelly *et al.* 2011), is that its statistical framework is most closely related to the framework we implemented for identifying allele-specific mRNA decay rate differences and allele-specific differences in translation efficiency (described below).

We likewise used our "labeled" NRO data set to identify transcripts with allelic differences in transcription rate. For each replicate, we summed the number of reads deriving from the Y12 allele and the number of reads deriving from the DBVPG6044 allele across all SNVs in each transcript shared between the two strains. After filtering out transcripts with no reads mapping to one or both alleles in more than one replicate, we determined whether each transcript exhibited allele-specific transcription rate differences by performing the *cis* test exactly as described by Connelly *et al.*, but using three, instead of two replicates (2014).

RNA decay rate

To determine whether a transcript exhibited allele-specific RNA decay rate differences, we implemented the statistical framework we previously developed (Andrie *et al.* 2014) for measuring allelic differences in RNA decay rates. Briefly, for each time point in each replicate, we summed the number of reads deriving from the Y12 allele and the number of reads deriving from the DBVPG6044 allele across all SNVs in each transcript shared between the two strains. We then removed transcripts that did not have at least one read mapped to at least one allele in all time points in all replicates. For the transcripts remaining in our data set, we next summed the number of reads deriving from each allele across all replicates of each time point. As described by Andrie *et al.*, we measured the change in the proportion, p_{DB}, of reads derived from the DBVPG6044 transcript as a function of time using a linear logistic model (2014). To assess statistical significance, we applied a Bayesian hierarchical Markov chain Monte Carlo model.

Steady-state RNA expression levels

We used the time point $t = 0$ min from our RNA decay rate time course as a proxy for steady-state RNA expression levels. For each replicate, we summed the number of reads deriving from the Y12 allele and the number of reads deriving from the DBVPG6044 allele across all SNVs in each transcript shared between the two strains. After filtering out transcripts with no reads mapping to one or both alleles in more than one replicate, we determined whether each transcript exhibited allele-specific steady-state RNA expression level differences by performing the *cis* test exactly as described by Connelly *et al.*, but using three, instead of two replicates (2014).

Translation efficiency

To determine whether a transcript exhibited allele-specific translation efficiency differences, we developed a proportional odds logistic regression model to measure how the distribution of the DBVPG6044 allele and the Y12 allele differ across the polysome fractions that we collected. Specifically, for each fraction, we summed across all replicates the number of reads deriving from the Y12 allele and the number of reads deriving from the DBVPG6044 allele across all SNVs in each transcript shared between the two strains. We then removed transcripts that did not have at least one read mapped in all fractions. We also removed transcripts that did not have at least one read mapped to each allele in at least one of the fractions. We then fit our proportional odds logistic regression model to each of the remaining transcripts. In our model, the dependent variable, y^*, is the location of a transcript within the polysome gradient, and this is a continuous version of the discrete variable y that we observe. We cannot directly observe y^*, but rather can observe which fraction pool the transcript fell into, such that:

$$y = \begin{cases} 1 & \text{if } y^* \le \mu_1, \\ 2 & \text{if } \mu_1 < y^* \le \mu_2, \\ 3 & \text{if } \mu_2 < y^* \le \mu_3, \\ 4 & \text{if } \mu_3 < y^* \le \mu_4, \\ 5 & \text{if } \mu_4 < y^* \le \mu_5, \\ 6 & \text{if } \mu_5 < y^* \le \mu_6 \end{cases}$$

where the parameters μ_i are the endpoints in the polysome gradient of the fraction pools, A – F, that we collected. The probabilities associated with the fraction pools are $\{\pi_1, \pi_2, \ldots, \pi_i\}$, and the cumulative probability of a response less than or equal to i is given by:

$$P(y \le i) = \pi_1 + \ldots + \pi_i$$

The cumulative logit, which describes log-odds of a transcript being in fraction i or below versus in a fraction higher than i, is then defined as:

$$\log\left(\frac{P(y \le i)}{P(y > i)}\right) = \log\left(\frac{P(y \le i)}{1 - P(y \le i)}\right) = \log\left(\frac{\pi_1 + \ldots + \pi_i}{\pi_{i+1} + \ldots + \pi_I}\right)$$

The sequence of cumulative logits may be defined as:

$$L_1 = \log\left(\frac{\pi_1}{\pi_2 + \pi_3 + \ldots + \pi_r}\right)$$

$$L_2 = \log\left(\frac{\pi_1 + \pi_2}{\pi_3 + \pi_4 + \ldots + \pi_r}\right)$$

…

$$L_{r-1} = \log\left(\frac{\pi_1 + \pi_2 + \ldots + \pi_{r-1}}{\pi_r}\right)$$

where L_i is the log-odds of a transcript being in fraction i or below versus in a fraction higher than i. We incorporate the allele, j, j =0,1, of the transcript as a predictor variable, x, such that:

$$L_1 = \alpha_1 + \beta x$$

$$L_2 = \alpha_2 + \beta x$$

...

$$L_r = \alpha_{r-1} + \beta x$$

In this model, α_i is the log-odds of falling into or below fraction i when $x = 0$, and a single parameter β describes the change in log-odds of falling into or below any fraction

when $x = 1$. A positive coefficient β indicates, therefore, the increased tendency for the probability of a transcript of allele $j = 1$ to fall into lower fractions than a transcript of allele $j = 0$. To fit this model in R, we used the *polr* function from the *MASS* package with the read counts for each allele in each fraction input as *weights* in the function.

For each transcript, β is the summary statistic that is of primary interest. Asymptotically (i.e. in large samples) β has a normal distribution, centered on the true value and with variance equal to the standard error squared. Following Wakefield, we take this normal distribution as our likelihood and then add a normal prior centered at 0 and with variance W (2007). W is chosen to reflect how large we believe the log odds ratio will be. A Bayes factor can then be calculated which summarizes the evidence in favor of the null (i.e. $\beta = 0$) versus the alternative (i.e. β != 0). Further, a posterior probability on the null can be calculated if we add a prior probability on the null, π_0. We obtain the latter by examining the distribution of p-values over all transcripts and using the QVALUE methodology (Storey 2002; Storey and Tibshirani 2003; Storey *et al.* 2004). Using this estimate of π_0, we obtain posterior probability on all transcripts. We used a threshold of posterior probability > 0.5 to decide whether a transcript exhibits significant differences in translation efficiency between alleles.

RNA secondary structure

Using the "splice junction" files we generated as described above, we summed the number of ligation events in each transcript for each strain and each replicate. After filtering out transcripts with no ligation events for one or both alleles in more than one

replicate, we determined whether each transcript exhibited allele-specific secondary structure differences by performing the *cis* test exactly as described by Connelly *et al.*, but using three, instead of two replicates (2014).

RNA binding protein binding

We used our PIP-seq data set to identify SNVs in transcripts with allelic differences in RNA binding protein (RBP) binding to the variant site. To first identify variants located within RBP footprints, we used the DESeq2 (Love *et al.* 2014) package to determine which SNVs in our data set exhibited significant enrichment of sequencing reads in our footprinting samples as compared to our control samples. We separately created DESeq data set matrices for un-normalized counts of reads mapping to the Y12 allele and to the DBVPG6044 allele in our footprinting and control samples with the *DESeqDataSetFromMatrix* tool, such that we had one data set matrix for each strain. We pre-filtered each data set to remove rows with only 0 or 1 reads. We then ran a differential expression analysis with the *DESeq* tool to look for differences in expression due to treatment (i.e. footprint versus control), while using replicate as an additional factor in the experimental design. We corrected the p-values output by DESeq2 for multiple testing with the QVALUE software (Storey 2002; Storey and Tibshirani 2003; Storey *et al.* 2004). We retained all SNVs in our data set that had a q-value < 0.10 and were enriched for reads in the footprinting sample as compared to the control sample in either strain. We determined if these SNVs exhibited allelic differences in levels of RBP binding between the two alleles by applying the *cis* test to our footprinting samples exactly as described by Connelly *et al.*, but using three, instead of two replicates (2014).

Gene Ontology analysis

We assessed whether each set of transcripts we identified with significant allele-specific differences in a particular molecular phenotype, or combination of molecular phenotypes, was significantly enriched for genes involved in a specific molecular function, biological process, or cellular component using the Gene List Analysis tool from AmiGO 2 (Carbon *et al.* 2009). Specifically, we performed the PANTHER Overrepresentation Test (release 20160715) with the GO Ontology database (released 2017-01-26) (Ashburner *et al.* 2000; The Gene Ontology Consortium 2015) using default settings and "Saccharomyces cerevisiae" as the organism. We submitted all genes we tested for allele-specific differences in the particular molecular phenotype(s) as the "Reference List" and the subset of genes that we found to have significant allele-specific differences as the "Analyzed List." We note that any transcripts that did not have a reciprocal BLAST best hit (described above, see Allele-specific read calling for ATAC-seq) with an annotated S288c gene (annotations obtained from the *Saccharomyces* Genome Database and based on the S288C_reference_genome_R64-2-1_20150113 build; http://www.yeastgenome.org) were excluded from our lists. We performed the enrichment analysis three times, once each with "GO molecular function complete," "GO biological process complete," and "GO cellular component complete" as the "Annotation Data Set." We used the Bonferroni correction for multiple testing.

RESULTS

Overview of experimental design

We measured allele-specific differences in six molecular phenotypes, including chromatin accessibility, RNA secondary structure, binding of proteins to RNA, and rates of RNA transcription, decay and translation in a diploid yeast produced by mating two genetically diverse wild-derived haploid strains of *Saccharomyces cerevisiae*: the Sake strain Y12 and the West-African strain DBVPG6044 (Liti *et al.* 2009). A schematic of the experimental design is shown in Figure 1. Briefly, we grew the hybrid diploid yeast to mid-log phase at 30° in a large (500 mL) culture, and then, as simultaneously as possible, collected culture samples for each of the six molecular phenotypes we examined. We applied ATAC-seq to one culture sample to measure chromatin accessibility; a NRO assay coupled to RNA-Sequencing to another culture sample measure transcription rates; a time course with sampling at 0, 15, 20, and 30 minutes following treatment with the transcriptional-inhibitor 1,10-phenanthroline to a third culture sample to measure RNA decay rates; polysome profiling coupled to RNA-Sequencing of six distinct fractions across the polysome gradient (Figure 4A) to a fourth culture sample to measure mRNA translation efficiency; RPL to a fifth culture sample to measure RNA secondary structure; and, finally, PIP-seq to a sixth culture sample to measure binding of proteins to RNA (Figure 1). To identify allele-specific differences in each molecular phenotype, we used polymorphisms to distinguish between parental alleles of transcripts or regions of open chromatin, and compared the relative levels of each allele in each assay. In total, we measured allelic differences from three independent biological replicates.

Statistical modeling of allele-specific differences in phenotypes

To identify allelic differences in chromatin accessibility, RNA secondary structure, binding of proteins to RNA, and rates of RNA transcription and decay, we employed statistical methods previously developed by our lab (Connelly *et al.* 2014; Andrie *et al.* 2014). These methods use a Bayesian framework to assess the significance of the parameters of interest we calculate from our models, and therefore, account for genes that exhibit small departures from non-constancy due to high read counts in a more principled manner than alternative approaches (Connelly *et al.* 2014; Andrie *et al.* 2014). To identify allelic differences in translation efficiency, we developed an analogous statistical framework. Specifically, for each transcript, we use a proportional odds logistic regression model to quantify how the distribution of the DBVPG6044 allele and the Y12 allele differ across six polysome fractions corresponding to 1) no ribosomes, 2) one ribosome, 3) two ribosomes, 4) three ribosomes, 5) approximately four to seven ribosomes, and 6) approximately eight or more ribosomes bound to each transcript. To assess statistical significance, we use the Bayesian false-discovery probability method developed by Wakefield (see Materials and Methods) (2007).

Pervasive influence of *cis*-regulatory variation on molecular phenotypes

Chromatin accessibility

Utilizing the combined ATAC-seq data from all three of our replicates, we detected over 4,500 distinct regions of open chromatin associated with 2,467 transcripts per strain. Through a careful filtering pipeline, which included whole genome sequencing of the Y12xDBVPG6044 diploid hybrid as a control, we identified

13,884 single nucleotide variants (SNVs) in these regions that could be used to assign whether individual ATAC-seq reads derived from the Y12 or the DBVPG6044 allele (see Materials and Methods; Table 1). Of the approximately 217 million ATAC-seq reads we obtained across all replicates, 15.81 million reads were informative, meaning that they mapped to a variant site within a chromatin accessibility region and could unambiguously be assigned as originating from either DBVPG6044 or Y12 (Table 1). Applying the method developed by Connelly *et al.* (2014), we found that the chromatin accessibility regions associated with 1,220 transcripts have allele-specific differences in levels of accessibility (posterior probability > 0.95) (Table 1). 627 of the 1,220 transcripts (51.4%) that exhibit associated allele-specific differences in levels of accessibility are more open at the Y12 allele and 594 (48.6%) are more open at the DBVPG6044 allele. The median $\log_2$-fold change in chromatin openness for all transcripts with allele-specific chromatin accessibility differences is 0.15 (Figure 2). We found that transcripts associated with regions containing allele-specific chromatin accessibility differences span a broad range of gene ontology terms, and we did not detect significant enrichment for particular molecular functions, cellular components, or biological processes after correcting for multiple comparisons. However, we note that the allele-specific differences we identified do not include presence/absence differences between alleles, due to our filtering pipeline. Therefore, our findings do not preclude the possibility that allele-specific differences that lead to open versus closed chromatin state differences play a critical role for certain gene classes.

Transcription rates

Using an analogous filtering pipeline to the one that we employed for our ATAC-seq data (see Materials and Methods), we identified 44,758 transcribed SNVs in 5,424 transcripts that could be used to assign whether individual RNA-Seq reads from our NRO, RNA decay rate, and mRNA translation efficiency assays derived from the Y12 or the DBVPG6044 allele (see Materials and Methods). For our NRO assay, 1.14 million of the approximately 46 million RNA-Seq reads we obtained across all replicates from newly transcribed RNA molecules were informative, meaning that they mapped to a transcribed variant site and could unambiguously be assigned as originating from either DBVPG6044 or Y12 (Table 1). Applying the method developed by Connelly *et al.* (2014) to 3,303 transcripts that passed our filtering criteria for the NRO assay (see Materials and Methods), we found that 1,259 transcripts have allele-specific differences in transcription rate (posterior probability > 0.95). 624 of the 1,259 transcripts (49.6%) that exhibit allele-specific transcription rate differences have more transcription of the Y12 allele, while 635 transcripts (50.4%) have more transcription of the DBVPG6044 allele. The median $\log_2$-fold change in amount of transcription for all transcripts with allele-specific transcription rate differences is 1.18 (Figure 2). We found that transcripts with allele-specific differences in transcription rate span a broad range of gene ontology terms, and we did not detect significant enrichment for particular molecular functions, cellular components, or biological processes after correcting for multiple comparisons.

RNA decay rates

For our RNA decay rate assay, 53.22 million of the approximately 226 million RNA-Seq reads we obtained across all time points and all replicates were informative

(Table 1). We examined the influence of *cis*-regulatory variation on RNA decay rate by applying the statistical inference framework we previously developed (Andrie *et al.* 2014) to 4,861 transcripts that passed our decay rate assay filtering criteria (see Materials and Methods). From the Bayesian hierarchical MCMC model we use in this framework, we estimated $1 - \pi_0$, the proportion of transcripts that exhibit allele-specific decay to be 0.47. Thus, approximately 47% of all measured transcripts are inferred to be inconsistent with the null hypothesis and exhibit allelic differences in decay rates. Of these, we can identify 980 transcripts at a false discovery rate of 10% (Figure 3A). Note, this corresponds to a false non-discovery rate of 36%. The exponential of the slope of the linear logistic model fit to each transcript is a direct estimate of the difference in RNA decay rate between the Y12 and DBVPG6044 alleles. Among the genes with significant allelic differences in RNA decay rate, the effect size of the decay rate difference ranges from a 4.65×10^{-4} to a 6.74×10^{-2} change in the odds of observing an RNA allele of the DBVPG6044 strain given a one minute increase in time, with a median difference of 8.78×10^{-3} (Figure 3B). This median difference corresponds to an approximately 69% increase over one hour in the odds of observing an RNA allele of the DBVPG6044 strain. The Y12 allele decays more quickly than the DBVPG6044 allele in 502 transcripts (51.2%), while the DBVPG6044 allele decays more quickly than the Y12 allele in 478 transcripts (48.8%) (Figure 3C). We found that transcripts with allele-specific differences in RNA decay rate are significantly enriched for genes involved in cofactor binding (1.88 fold-enrichment, p-value = 2.29×10^{-2}), as well as carboxylic acid biosynthesis (1.85 fold-enrichment, p-value = 5.86×10^{-3}) and the alpha-amino acid metabolic process (1.77 fold-enrichment, p-value = 3.16×10^{-2}). *Cis*-regulatory variation in these genes classes may

have arisen as a result of the differences in nutrient availability in the environments where Y12 and DBVPG6044 reside.

Translation efficiency

For our translation efficiency assay, 68.48 million of the approximately 356 million RNA-Seq reads we obtained across all collected fractions and all replicates were informative (Table 1). We examined the influence of *cis*-regulatory variation on mRNA translation efficiency by applying the proportional odds logistic regression model framework described above to 5,014 transcripts that passed our translation efficiency assay filtering criteria (see Materials and Methods). From the Bayesian model we use to assess significance in this framework, we identify 1,279 transcripts as exhibiting allele-specific translation efficiency at a false discovery rate of 16% (Figure 4B). Note, this corresponds to a false non-discovery rate of 18%. The coefficient of the proportional odds logistic regression model fit to each transcript can be used to approximate the difference between the DBVPG6044 allele and the Y12 allele the number of ribosomes bound to each allele; however, due to complex the relationship between fraction and number of ribosomes bound to a transcript (Figure 4A), as well as being unable to accurately normalize read counts across fractions, we restricted our interpretation of this coefficient to its directionality only, and did not consider its magnitude. The Y12 allele is more efficiently translated than the DBVPG6044 allele in 640 transcripts (50.0%), while the DBVPG6044 allele is more efficiently translated than the Y12 allele in 639 transcripts (50.0%) (Figure 4C). We found that transcripts with allele-specific differences in translation efficiency are significantly enriched for genes involved in the cytosolic

ribosome (1.65 fold-enrichment, p-value = 4.91 x 10^{-2}), and are significantly depleted for genes involved in the phospholipid biosynthetic process (<0.2 fold-enrichment, p-value = 1.14 x 10^{-2}). These findings suggest that *cis*-regulatory variation affecting translation of phospholipid biosynthetic genes is under strong purifying selection.

Cis-regulatory variation exhibits widespread pleiotropic effects on molecular phenotypes

Previously, it has been hypothesized that *cis*-regulatory mutations are less pleiotropic than protein-coding mutations because they only affect a subset of the expression domains of the gene, as opposed to affecting the protein everywhere it is expressed (Paaby and Rockman 2013). To explore the extent to which *cis*-regulatory variation affects expression, we examined the overlap in transcripts exhibiting statistically significant allelic differences in chromatin accessibility, transcription rates, RNA decay rates, and translation efficiency. We found that 49.2% of transcripts with allele-specific differences in chromatin accessibility, 55.4% of transcripts with allele-specific differences in transcription rate, 64.1% of transcripts with allele-specific differences in RNA decay rates, and 55.9% of transcripts with allele-specific differences in translation efficiency differences also had allele-specific differences in at least one other phenotype (Figure 5). Of the 1,526 transcripts we tested in all four assays, 170 transcripts exhibit allele-specific differences in three of four molecular phenotypes, while 31 transcripts have allele-specific differences in all four molecular phenotypes (Figure 5). As we would expect that only variation present in the transcript would affect allele-specific differences in RNA decay and translation efficiency, we identified transcripts

with only one variant site between Y12 and DBVPG6044 that exhibited both allele-specific RNA decay and translation efficiency. Interestingly, we found six such transcripts, in which a single *cis*-regulatory variant affects both RNA stability and translation rate. Taken together, these results suggest that pleiotropy is a common feature of *cis*-regulatory variation, with *cis*-regulatory variants frequently influencing multiple steps in the conversion of DNA sequence information into organismal phenotype.

Tracking the potential chain of causality for allele-specific expression differences

A unique feature of our data set is its coordinated measurement of allelic differences in several molecular phenotypes. We leveraged this feature to better understand the impact of *cis*-regulatory variation on the conversion of genotype into phenotype.

Chromatin accessibility and transcription rates

We began by investigating the relationship between *cis*-regulatory variation affecting chromatin accessibility and transcription rates in the 1,527 transcripts for which we collected both ATAC-seq and NRO data. Interestingly, the majority of transcripts (61.7%) with allele-specific differences in chromatin accessibility in our data set do not exhibit significant allele-specific differences in transcription rates, indicating that many of the chromatin structure changes caused by *cis*-regulatory variation may be simply be functionally benign, as Connelly *et al.* previously suggested (2014). Among the 278 transcripts we identified as exhibiting allele-specific differences in both molecular

phenotypes, there is a weak, but positive correlation between which allele has a more open chromatin state and which allele is more highly transcribed (Pearson correlation coefficient = 0.16, p-value = 9.2×10^{-3}) (Figure 6A). Thus, these data additionally suggest that the relationship between the two molecular phenotypes may be not be straightforward, perhaps because the consequences of chromatin structure changes are also not straightforward: changes in chromatin accessibility could affect recruitment of both transcriptional activators and transcriptional repressors, which would have opposite effects on transcription rates.

Transcription rates, RNA decay rates, and steady-state RNA expression levels

We next examined the effects of *cis*-regulatory variation influencing transcription and decay rates by comparing allele-specific differences in these two rates to allele-specific differences in steady-state expression levels. Specifically, we used the 0 minute time point of our RNA decay rate time course as a proxy for steady-state expression and applied the method developed by Connelly *et al.* (2014) to identify which of the 4,938 transcripts that passed our filtering criteria have significant allele-specific steady-state expression (see Materials and Methods). We found 2,186 (44.3%) transcripts with allele-specific differences in steady-state expression levels (posterior probability > 0.95). 1,075 of the 2,186 transcripts (49.2%) have more of the Y12 allele, while 1,111 transcripts (50.8%) have more of the DBVPG6044 allele. The median $\log_2$-fold change in levels of expression for all transcripts with allele-specific steady-state expression level differences is 0.37 (Figure 2).

To investigate how transcription rate differences caused by *cis*-regulatory variation affect steady-state expression levels, we identified the 684 (20.7%) transcripts exhibiting allele-specific differences in both steady-state expression and transcription rate from the 3,303 transcripts for which we had steady-state expression, transcription rate, and decay rate data. Among these 684 transcripts, 472 (68.9%) show concordant differences, such that the allele that is more highly transcribed also has higher levels of steady-state expression (Figure 6B). Additionally, among the same 684 transcripts, we observed that the ratio at which the two alleles were being transcribed is significantly positively correlated with the steady-state expression ratio of the two alleles (Pearson correlation coefficient = 0.41, p-value < 2.2×10^{-16}) (Figure 6B). These findings suggest that *cis*-regulatory changes in transcription commonly drive corresponding changes in steady-state expression. However, the 212 transcripts that showed allele-specific steady-state expression differences opposite of what we would predict based on their allele-specific transcription rate differences, as well as the 575 transcripts that exhibit allele-specific transcription rate differences, but no steady-state expression differences, suggest that most commonly (62.5% of transcripts with allele-specific transcription rate differences), *cis*-regulatory variants influencing rates of transcription are coupled to variants with opposite effects on RNA decay rate or, perhaps, other post-transcriptional processes that regulate RNA levels.

To further explore the relationship between *cis*-regulatory variation affecting transcription and RNA decay, we next investigated how RNA decay rate differences caused by *cis*-regulatory variation affect steady-state expression levels. Of the 533 transcripts that show both types of allele-specific differences, 378 (70.9%) have increased

decay rates in the allele with higher levels of steady-state expression. Among the same 533 genes, we likewise observed that the difference we estimated in decay rate between the two alleles is significantly negatively correlated with the steady-state expression ratio of the two alleles (Pearson correlation coefficient = -0.46, p-value < 2.2×10^{-16}) (Figure 6C). As with our transcription rate data, these data suggest that there are *cis* variants influencing rates of transcription with opposite effects to those influencing decay.

We further tested this hypothesis by directly investigating the relationship between transcription rates and RNA decay rates. Interestingly, while 1,259 of the 3,303 transcripts in our data set exhibit allele-specific transcription rate differences, and 814 transcripts exhibit allele-specific RNA decay rate differences, only 334 transcripts (27% of transcripts with significant transcription rate differences and 41% of transcripts with significant decay rate differences) exhibit both types of allele-specific differences. Moreover, among these 334 transcripts, we found only a modest negative correlation between the ratio at which the two alleles were being transcribed and the difference we estimated in decay rate between the two alleles (Pearson correlation coefficient = -0.16, p-value < 4.4×10^{-3}) (Figure 6D). Thus, direct comparison of allele-specific differences in transcription rates and RNA decay rates only weakly supports our hypothesis that changes in transcription rates are often coupled to opposing changes in RNA stability. As our hypothesis also predicts that transcripts that exhibit allele-specific differences in RNA decay rates, but not in steady-state expression levels would be enriched for *cis* variants with opposing effects on transcriptional initiation and decay, since the difference in decay rate does not produce a corresponding difference in steady-state expression levels, we additionally examined the subset of 281 transcripts in our data set with allele-

specific differences in RNA decay rates, but not steady-state expression levels. We found that only 98 (34.9%) of these transcripts have allele-specific differences in transcription, and thus, that this subset of transcripts was modestly depleted for transcripts with allele-specific differences in transcription, as compared to transcripts with allele-specific differences in both RNA stability and steady-state expression levels (34.9% versus 44.3%; Chi-square test statistic = 6.34, p-value = 1.18×10^{-2}). Additionally, among these 98 transcripts, there is no correlation between which allele is more highly transcribed and which allele is more quickly degraded (Pearson correlation coefficient = 0.07, p-value = 0.51) (Figure 6D). Overall, while our comparisons of transcription and decay rates to steady-state RNA expression levels suggest that there is widespread stabilizing selection to maintain or fine-tune steady-state gene expression levels, our data also suggests that there is a complex relationship between gene expression levels and the variation underlying transcription and RNA stability. Specifically, direct comparison of allele-specific differences in transcription rates and RNA decay indicates that merely using the directionality and estimates of the magnitudes of transcription and decay rates will not be sufficient for accurately predicting gene expression levels, and that more sophisticated models will be needed. Such models may require incorporating the contributions of other post-transcriptional processes. For example, heat stress has been shown in yeast to result in the sequestration of mRNA and proteins into insoluble deposits (Cherkasov *et al.* 2015). Indeed, the drug treatment (1,10-phenanthroline) that we used to inhibit transcription and measure decay rates has been previously shown to mimic heat shock in yeast, and so such sequestration could be affecting our observations in our data set (Adams and Gross 1991).

Translation efficiency and steady-state RNA expression levels

In addition to investigating how gene expression levels are affected by *cis*-regulatory variation governing chromatin accessibility, transcription rates, and RNA decay rates, we also were interested in determining the connection between steady-state RNA expression differences and translation efficiency differences. In particular, previous studies have found very low correlations between RNA expression levels and protein abundance, suggesting that translation might be important than RNA levels for determining protein levels (Foss *et al.* 2007; Ghazalpour *et al.* 2011). We measured allele-specific differences in both steady-state RNA expression levels and translation efficiency in 4,862 transcripts. 1,530 of these transcripts show significant allele-specific differences in RNA expression levels, but not in translation efficiency. For such transcripts, we predict that allele-specific differences in RNA expression levels would correlate strongly to allele-specific differences protein levels, and therefore, that *cis* variation in these transcripts exerts its effects on phenotype mainly by influencing RNA levels. Contrastingly, 602 transcripts in our data set show allele-specific differences in translation efficiency, but not in steady-state RNA expression levels. For these transcripts, we predict that *cis* variation would produce phenotypic variation mainly by influencing translation rates. In addition to investigating transcripts with allele-specific differences in either steady-state RNA expression or translation, we also examined the 647 transcripts that exhibited significant allele-specific differences in both molecular phenotypes. Interestingly, we observed that 368 (56.9%) of these transcripts show concordant allele-specific differences in RNA expression levels and translation

efficiency, such that the more highly expressed allele is also more efficiently translated, while 279 (43.1%) transcripts show discordant differences. Thus, while *cis* variation affecting translation efficiency more often reinforces the *cis* variation affecting RNA expression differences, it also commonly serves to buffer those differences.

Mechanisms through which *cis*-variation affects RNA decay rates and translation efficiency

RNA stability versus translation efficiency

Previous reports measuring mRNA decay rates in one or a few genes have suggested that the translation efficiency of an mRNA might be directly coupled to mRNA decay rate (reviewed in Garneau *et al.* 2007 and Schoenberg and Maquat 2012). To evaluate the role of *cis* variation in such coupling, we compared allele-specific differences in RNA decay rate and translation efficiency in our data. If *cis* variation does govern coupling, we would expect that transcripts with allele-specific differences in RNA decay rate would have a higher proportion of significant allele-specific differences in translation efficiency than transcripts without RNA decay rate differences. Indeed, we observe that significantly more transcripts with allele-specific RNA decay rate also have allele-specific differences in translation efficiency as compared to transcripts without allele-specific RNA decay rate (33% versus 24%; Chi-square statistic = 33.4, p-value = 7.60×10^{-9}). However, the proportion of transcripts with allele-specific differences in both phenotypes is still relatively low. Additionally, we observed that there is no correlation between which allele is more quickly degraded and which allele is more highly translated: 171 transcripts exhibit faster decay of the allele that is less efficiently

translated, while 152 transcripts exhibit faster decay of the allele that is more efficiently translated. Therefore, we do not find widespread evidence that *cis*-regulatory variation is involved in the direct coupling of RNA decay rate to translation efficiency.

RNA secondary structure

We have shown previously that genes with allele-specific differences in RNA decay rate are enriched for variants that influence RNA structure (Andrie *et al.* 2014). Using our RPL data set, we further explored how *cis* variation producing secondary structure differences affects RNA decay rates, and, additionally, investigated the role of RNA structure in translation efficiency. For our RPL assay, we could measure the number of ligation events, and therefore, the amount of secondary structure in 86 transcripts. Both the inefficiency of ligation and the low probability of obtaining a read that overlapped a variant site limited our coverage to a mere 32,336 informative reads across all of our RPL replicates. 23 (26.7%) of the transcripts in which we measured RNA secondary structure via RPL exhibit significant allele-specific differences, as determined using the *cis* test developed by Connelly *et al.* (2014). Of these 23 transcripts with allele-specific secondary structure differences, we could measure allele-specific differences in RNA decay rates and translation efficiency in 21 transcripts. We found that 7 (33.3%) of transcripts with allele-specific secondary structure differences also have RNA decay rate differences; while 10 (47.6%) of transcripts also have translation efficiency differences. 4 (19.0%) transcripts with allele-specific secondary structure differences have both RNA decay rate and translation efficiency differences. In these 13 transcripts, secondary structure differences may be directly responsible for the RNA

stability and/or translation rate differences we observed. In the transcripts with allele-specific secondary structure differences that did not affect allele-specific RNA decay rate or translation efficiency, perhaps such differences are important to other processes in the cell, such as RNA localization. Alternatively, such differences may be functionally unimportant.

RNA binding protein binding

Using our PIP-seq assay, we also examined how *cis* variation producing differences in RNA binding protein (RBP) binding affects RNA decay rate and translation efficiency. From our PIP-seq data, we could only reliably identify 413 variant sites in 275 transcripts with evidence of RNA binding protein (RBP) binding. In this assay, we were most restricted by high levels of noise in the data. Using the *cis* test developed by Connelly *et al.* (2014), we identified 270 of these sites in 196 transcripts as exhibiting allele-specific differences in levels of RBP binding. We could measure allele-specific differences in RNA decay rate and translation efficiency in all 196 transcripts we identified with RBP binding differences. We observed that 76 (38.8%) of these transcripts also have allele-specific RNA decay rate differences, while 79 (40.3%) also have allele-specific translation efficiency differences. 34 (17.3%) transcripts with allele-specific RBP binding differences have both RNA decay rate and translation efficiency differences. In these 121 transcripts, RBP binding differences may directly lead to the RNA stability and translation rate differences we observed. In the transcripts with allele-specific RBP differences that did not affect allele-specific RNA decay rates or translation

efficiency, perhaps such differences are important to other processes in the cell, such as RNA localization, or perhaps, they are simply functionally benign.

Many RBPs recognize their binding sites based on RNA secondary structure, and even those RBPs that bind to specific sequence motifs may be sensitive to secondary structure context (Li *et al.* 2014). Therefore, we investigated whether allele-specific differences in RNA secondary structure are connected to allele-specific differences in RBP binding. Of the 37 transcripts for which we have both RPL and PIP-seq data, 2 transcripts show allele-specific differences in RNA secondary structure only, while 7 show allele-specific differences in both molecular phenotypes. These findings are consistent with *cis* variation acting through RNA secondary structure to influence patterns of RBP binding. Moreover, of the 7 transcripts with both secondary structure and RBP binding differences, 5 of these transcripts also have allele-specific differences in RNA decay rate differences and/or translation efficiency (1 transcript has only allele-specific RNA decay rate differences, 2 transcripts have only allele-specific translation efficiency differences, and 2 transcripts have both RNA decay rate and translation efficiency differences). These data suggest that, for this subset of transcripts, *cis* variation influences RNA decay rates and translation efficiency through changes to RNA secondary structure, which in turn affect recruitment of RBPs to the transcript.

DISCUSSION

We undertake a cradle-to-grave analysis of the *cis*-regulatory variation affecting gene and protein expression through measurement of allele-specific differences in six molecular phenotypes in a diploid yeast hybrid created by mating two genetically diverse parental strains. Using statistical frameworks that employ Bayesian methods to assess the significance of the allele-specific differences we quantified for the various molecular phenotypes, including a novel statistical framework we developed for identifying allelic differences in translation efficiency, we demonstrate the pervasive influence of *cis*-regulatory variation on the conversion of genotype into phenotype, estimating that between one quarter and one half of all measureable transcripts exhibit allelic differences in each of the individual molecular phenotypes we examined. Additionally, we uncover extensive pleiotropy of *cis*-regulatory variation, calculating that over half of transcripts with allelic differences in one molecular phenotype also exhibit allelic differences in at least one other phenotype. Overall, our results suggest that intraspecific *cis*-regulatory variation is prevalent across the genome and affects a broad array of processes involved in determining gene expression levels. Importantly, our results also provide evidence of the critical importance of transcriptional as well as post-transcriptional processes to determining heritable gene and protein expression levels, which, in turn, impact phenotypic diversity among individuals within populations.

An especially novel aspect of our experimental design is our synchronous measurement of molecular phenotypes in cells taken from the same culture, which allows us to powerfully track the potential chain of causality of allele-specific expression differences. Somewhat surprisingly, we find little evidence that allelic differences in

chromatin accessibility can be used to predict allelic differences in transcription rate, suggesting that understanding the effects of *cis* variation on the binding of specific transcription factors, rather than on more general patterns of chromatin structure, will be most important for predicting gene expression levels. Another unexpected finding of our study was that, while comparison of allelic differences in steady-state expression levels to allelic differences in transcription rate and decay rate indicated that the *cis*-variation affecting these two processes is under stabilizing selection, such that variation causing changes in transcription rate is buffered by variation causing changes in decay rate, we did not see evidence of such buffering when directly comparing allelic transcription rate and decay rate. Our comparison of transcription rate and decay rate to steady-state expression levels recapitulates our and others' previous findings (Dori-Bachash *et al.* 2011; Pai *et al.* 2012; Andrie *et al.* 2014); however, this study is the first, to our knowledge, to directly compare inter-individual variation in transcription rate and decay rate. The complex relationship between transcription rate and decay rate that we uncovered highlights the need for additional, integrated studies of the various processes affecting gene expression levels in order to correctly create predictive models of the influence of genotypic variation on phenotypic variation.

The importance of variation in RNA expression levels as compared to variation in translation rate for producing variation in protein levels, and, ultimately, phenotype has recently become a subject of debate (Battle *et al.* 2015; Cenik *et al.* 2015). While we did not measure protein levels in our study, we can nonetheless use the patterns of allele-specific translation efficiency differences we observed to track the potential chain causality of allele-specific differences in protein levels. For the 25% of transcripts we

identified in our data set with allele-specific translation efficiency differences, comparison with steady-state RNA expression levels suggests that translation either buffers or reinforces expression differences in 30% of the transcripts we identified with allele-specific RNA expression levels. Additionally, 22% of transcripts without allele-specific RNA expression differences have allelic differences in protein production due solely to allele-specific translational differences. Thus, we find extensive evidence that *cis* variation affecting translation plays a crucial role in determining inter-individual variation in protein production. Our results argue that robust predictive modeling of phenotype from genotype will critically depend on understanding the effects of genetic variation on translation.

To achieve the goal of accurately predicting phenotype from sequence data, we will also need to dissect the mechanistic basis of allelic variation. Due to problems with sequencing coverage and data noise, we had very low power to detect allelic differences in RNA secondary structure or RBP binding to RNA in our data set. However, our naive estimates suggest that of 62% of allelic differences in both RNA secondary structure and RBP binding cause corresponding allelic differences in RNA decay rate and/or translation efficiency. In addition to further examining RNA secondary structure and RBP-RNA interactions, future research aimed at understanding the precise mechanisms underlying allelic expression differences, and thereby, phenotypic variation, should also investigate how genetic variation affects RNA localization as well as post-transcriptional RNA modifications, such as pseudouridylation. Likewise, at the protein level, such studies should attempt to quantify allelic differences in localization as well as post-

translational modifications to proteins. As functional genomics technologies continue to improve, such studies will become increasingly feasible.

An additional crucial area of future study will be exploring the interactions of genetic variation with the environment. Our study was conducted in yeast undergoing exponential growth in a rich medium; however, we would expect that the effects of *cis*-regulatory variation would differ markedly under differing growth conditions, such nutrient-limited media, or the presence of high concentrations of chemicals like ethanol or the various heavy metals. Another limitation of our study is that it only investigates allele-specific differences between two diverse yeast strains; future research will need to more broadly explore variation within and between species. Nevertheless, by demonstrating the pervasive influence of *cis*-regulatory variation on molecular phenotypes spanning the cradle to grave of gene expression, our study provides a good starting point for beginning to build predictive models of how genetic variation ultimately produces phenotypic variation.

ACKNOWLEDGEMENTS

We thank Jennifer Madeoy and Dayna Akey for their help with sample collection and preparation. We also acknowledge Nastya Gridasova, Vijay Ramani, and Ian Silverman for providing their protocols and custom analysis scripts for the NRO, RPL, and PIP-seq assays, respectively.

FIGURES

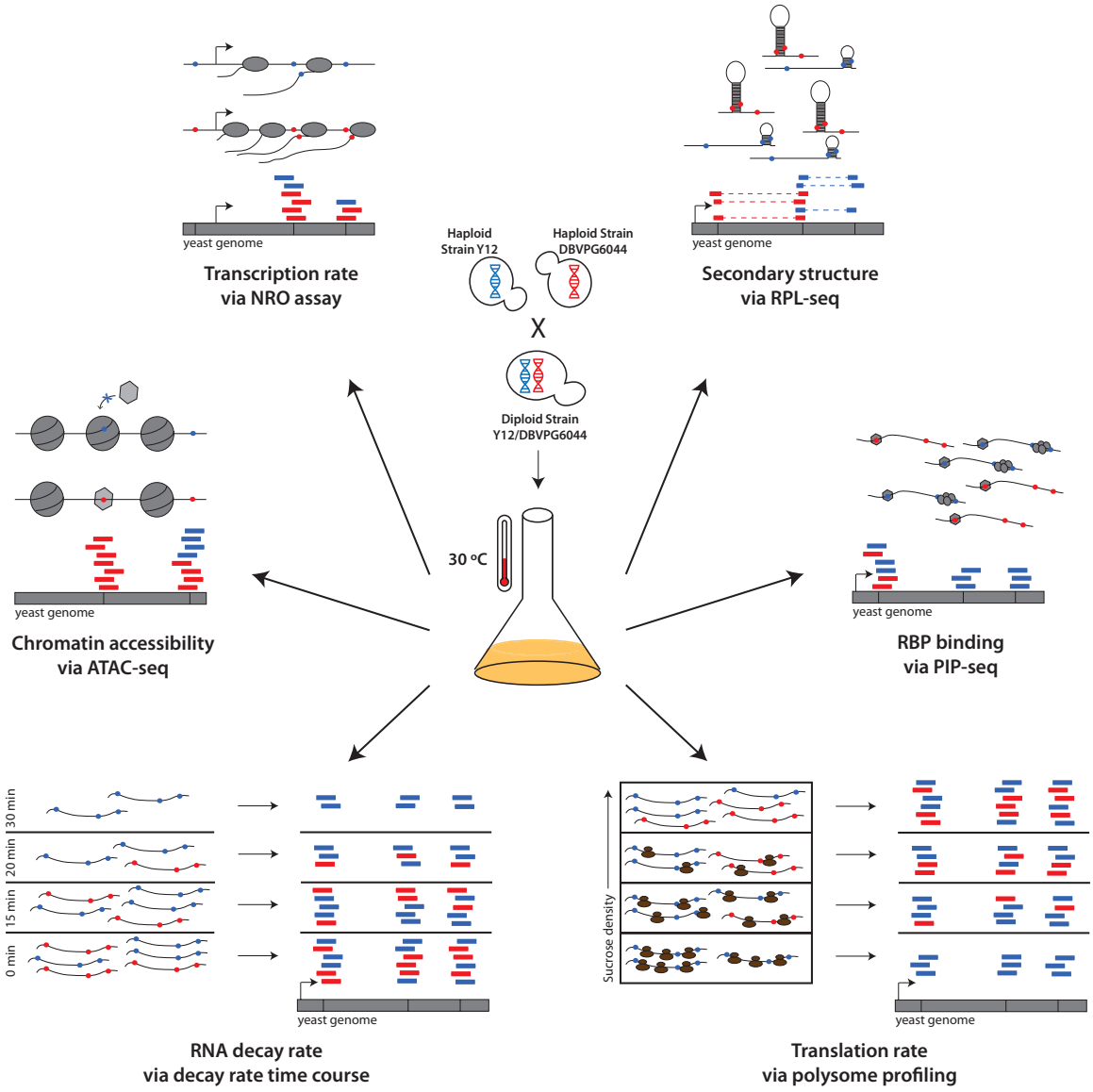


Figure 1. Overview of experimental design. We mated the haploid Y12 and DBVPG6044 strains of *Saccharomyces cerevisiae* to produce a diploid hybrid. The diploid was grown up at 30° to mid-log phase in a 500 mL culture of YEPD. As simultaneously as possible, aliquots of the culture were removed to perform six different assays, which measured six different molecular phenotypes: ATAC-seq was used to measure chromatin accessibility; a NRO assay coupled to RNA-Sequencing was used to

measure transcription rates; RPL was used to measure RNA secondary structure; PIP-seq was used to measure binding of proteins to RNA; polysome profiling coupled to RNA-Sequencing of distinct fractions across the polysome gradient was used to measure mRNA translation efficiency; and, a time course with sampling at 0, 15, 20, and 30 minutes following treatment with the transcriptional-inhibitor 1,10-phenanthroline was used to measure RNA decay rates. The sequencing reads from each assay were mapped back the parental Y12 and DBVPG6044 genomes. Any read overlapping a variant site between the two parental genomes was assigned as deriving from either the Y12 (red) or DBVPG6044 (blue) allele. In the schematic, gray bars represent the *S. cerevisiae* genome and vertical black lines represent variant sites between Y12 and DBVPG6044. Bent arrows indicate transcriptional start sites. Blue and red bars represent reads. In the ATAC-seq assay, reads are obtained from nucleosome-free regions. For the NRO assay, reads are obtained from newly-transcribed RNA. In the RPL assay, neighboring bases in RNA secondary structures are ligated together and the presence of unexpected "introns" in the RNA-Seq reads is used to identify where secondary structure exists. In the PIP-seq assay, RNA binding proteins (RBPs) are cross-linked to RNA via formaldehyde treatment and the RNA is digested with an RNase. The RNA-Seq reads obtained from the PIP-seq assay correspond to regions of the RNA protected from RNase digestion by bound RBPs. In the polysome profiling assay, RNAs are separated on a sucrose gradient by the number of ribosomes bound. We sequenced several fractions across the polysome gradient and measured the number of reads matching each allele in each fraction. For the RNA rate time course, we measured the number of reads matching each allele at four time points following transcriptional inhibition with the drug 1,10-phenanthroline.

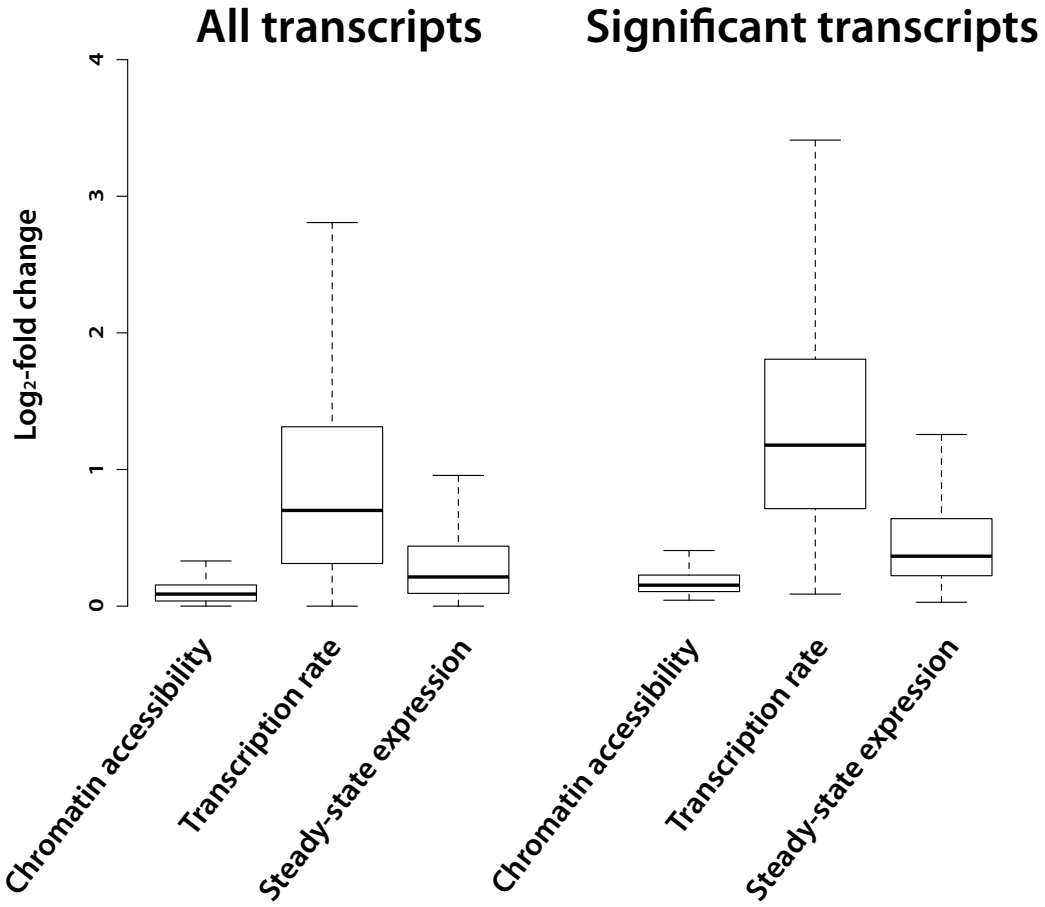


Figure 2. Effect sizes of allele-specific differences in chromatin accessibility, transcription rates, and steady-state RNA expression levels. The left boxplot shows the $\log_2$ of the fold change between the number of reads obtained for the Y12 allele as compared to the DBVPG6044 allele for all transcripts in which we measured chromatin accessibility via ATAC-seq, transcription rate via a NRO assay, and steady-state expression levels via the 0 minute time point of our RNA decay rate time course. The right boxplot shows the $\log_2$ of the fold change between the number of reads obtained for

the Y12 allele as compared to the DBVPG6044 allele for only those transcripts in which we identified significant allele-specific differences in the three molecular phenotypes. Note: To obtain effect sizes, all fold changes were calculated with the allele having more reads as the numerator and the allele having fewer reads as the denominator, such that all fold changes would be >1.

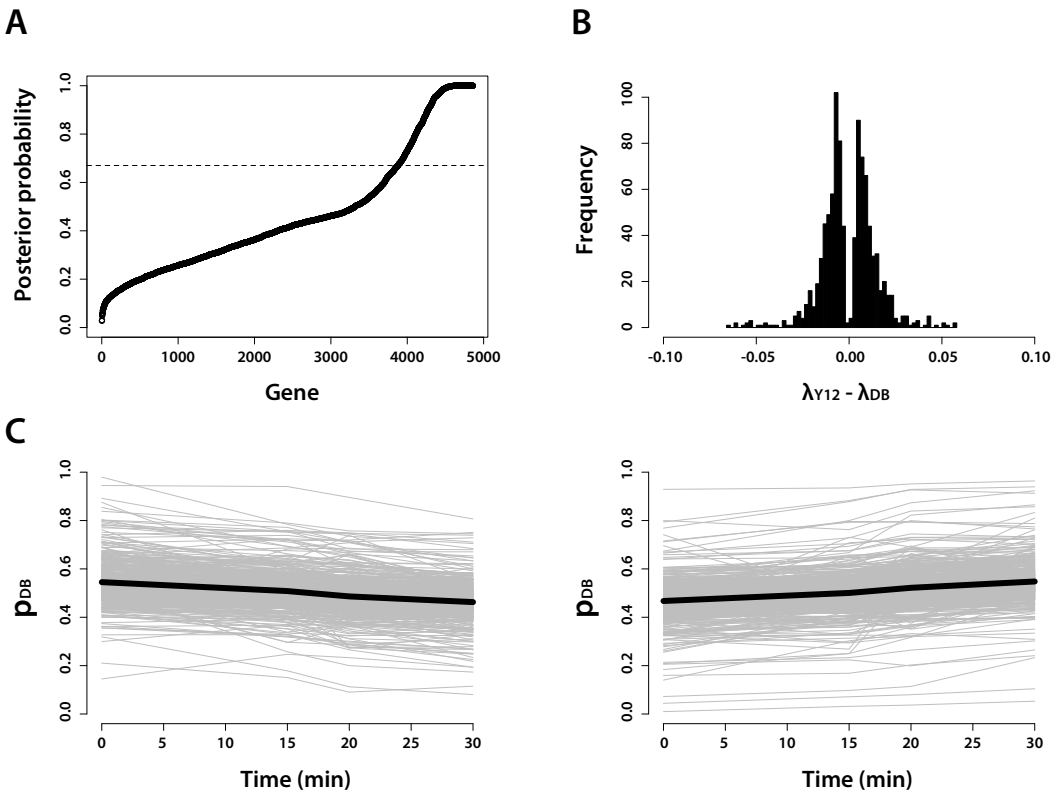


Figure 3. Characteristics of transcripts that exhibit allele-specific RNA decay rate. A) Posterior probability that a transcript exhibits allele-specific RNA decay rate, as calculated from our Bayesian hierarchical Markov chain Monte Carlo model. The dashed line at posterior probability = 0.67 corresponds to the threshold we used to call transcripts as exhibiting significant allele-specific RNA decay rate. B) Histogram of the slope

calculated from the linear logistic model for the 980 transcripts with significant (FDR = 10%) allele-specific RNA decay rate. The exponential of the slope, which estimates $\lambda_{Y12} - \lambda_{DBVPG6044}$, is the change in the odds of observing a Y12 RNA allele given a 1 minute increase in time. C) Decay rate time courses of all transcripts in which the DBVPG6044 allele decays significantly faster than the Y12 allele (left) and the Y12 allele decays significantly faster than the DBVPG6044 allele (right). The proportion of reads deriving from the DBVPG6044 allele as compared to all allelic reads (p_{DB}) is plotted as a function of time in minutes. The gray lines represent the decay rate time courses of each of the individual transcripts. The black lines represent the mean decay rate time courses for all of the transcripts included in each plot.

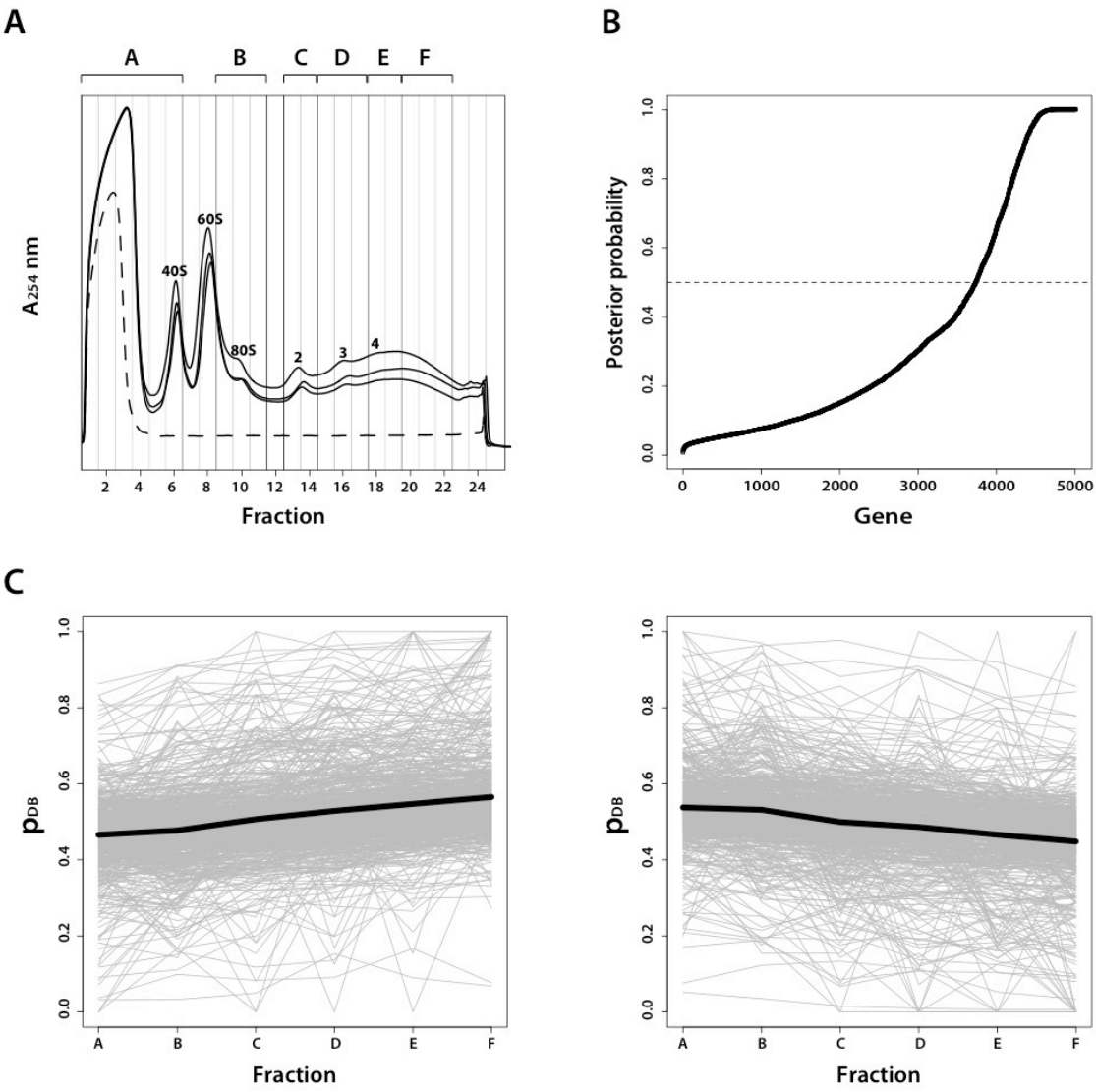


Figure 4. Characteristics of transcripts that exhibit allele-specific translation efficiency. A) Polysome profiles for each of our three replicates (black solid lines) and a buffer control (black dashed lines). The A_{254} nm is used to approximate the amount of biological material at each point in the sucrose gradient; however, the buffer also provides some signal, as shown. Peaks corresponding to the 40S ribosomal subunit, the 60S ribosomal subunit, one, two, three, and four ribosomes are marked on the profiles from the three replicates. We collected 24 fractions from each polysome gradient (x-

axis). The thin vertical lines in the plot show the portion of the polysome gradient collected into each fraction. For sequencing, the 24 fractions were pooled into six groups, demarcated as thicker black lines, and shown along the top of the plot as A (fractions 1-6), B (fractions 9-11), C (fractions 13-14), D (fractions 15-17), E (fractions 18-19), and F (fractions 20-22). B) Posterior probability that a transcript exhibits allele-specific translation efficiency, as calculated from our Bayesian model. The dashed line at posterior probability = 0.50 corresponds to the threshold we used to call transcripts as exhibiting significant allele-specific translation efficiency. C) The proportion of reads deriving from the DBVPG6044 allele as compared to all allelic reads (p_{DB}) is plotted as a function of fraction pool for all transcripts in which the DBVPG6044 allele is translated significantly more efficiently than the Y12 allele (left) and the Y12 allele is translated significantly more efficiently than the DBVPG6044 allele (right). The gray lines represent individual transcripts, while the black lines represent the mean pattern across fraction pools for all of the transcripts included in each plot.

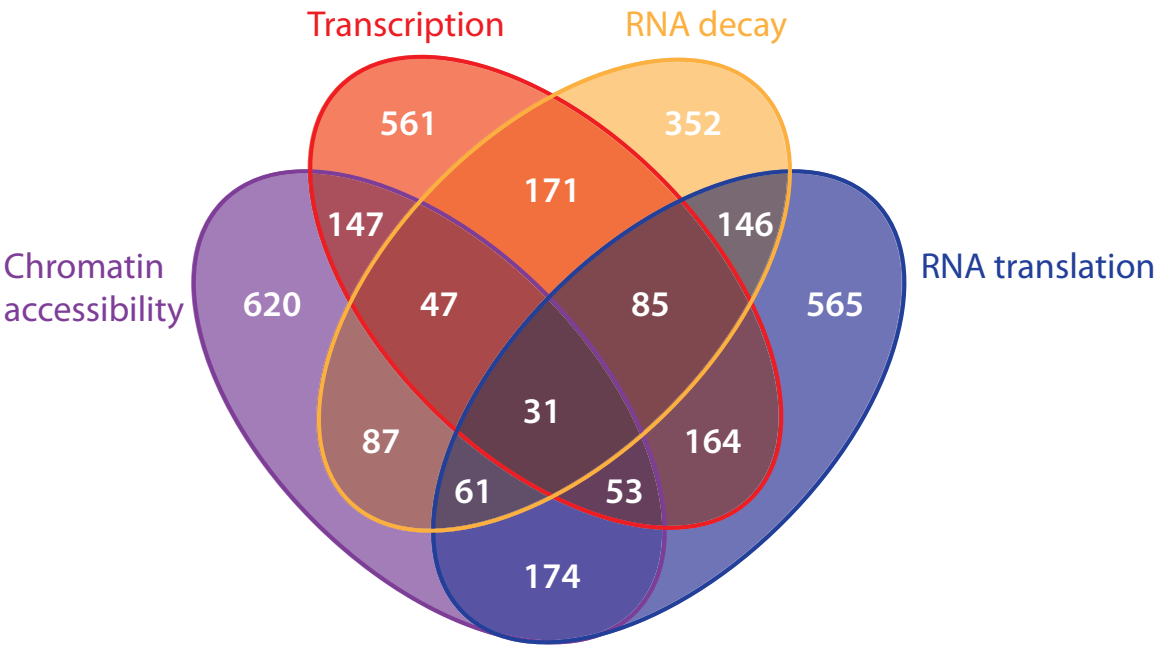


Figure 5. Overlap of transcripts exhibiting allele-specific differences in chromatin accessibility, transcription rates, RNA decay rates, and translation efficiency. The Venn diagram shows the number of transcripts we identified with significant allele-specific differences in each molecular phenotype or combination of molecular phenotypes. All transcripts assayed for each molecular phenotype are included.

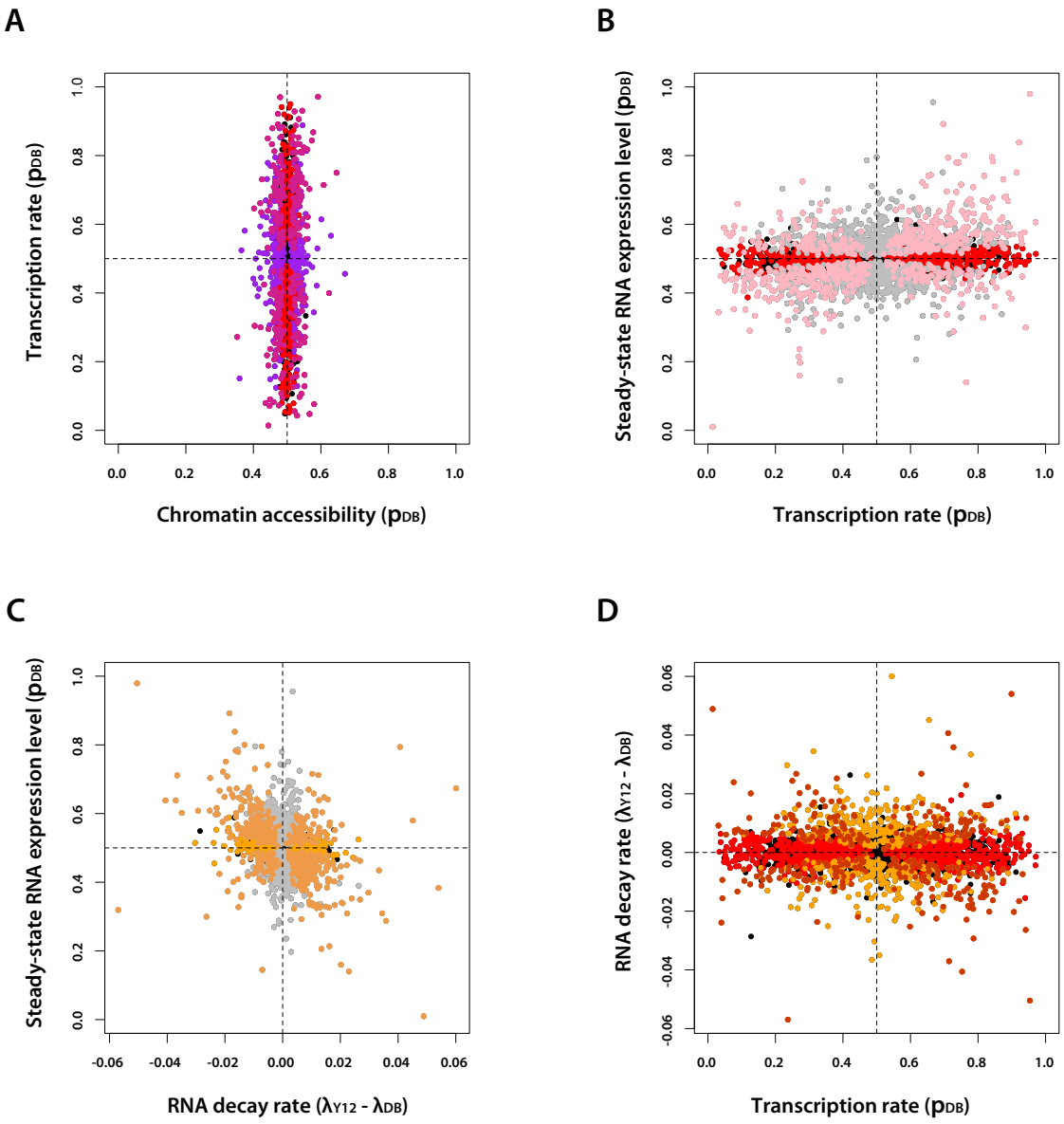


Figure 6. Correlations among allele-specific differences in chromatin accessibility, transcription rates, RNA decay rates, and steady-state RNA expression levels. A) Comparison of allele-specific chromatin accessibility and transcription rate. For each transcript, the proportion of reads deriving from the DBVPG6044 allele as compared to all allelic reads (p_{DB}) that we measured in our ATAC-seq assay is plotted on the x-axis and the p_{DB} we measured in our NRO assay is plotted on the y-axis. Transcripts that did

not have significant allele-specific differences in either molecular phenotype are represented by gray dots; transcripts with significant allele-specific differences in chromatin accessibility, but not transcription rate are represented by purple dots; transcripts with significant allele-specific differences in transcription rate, but not chromatin accessibility are represented by red dots; and transcripts with significant allele-specific differences in both transcription rate and chromatin accessibility are represented by magenta dots. B) Comparison of allele-specific transcription rate and steady-state RNA expression levels. For each transcript, the p_{DB} that we measured in our NRO assay is plotted on the x-axis and the p_{DB} we measured using the 0 minute time point of our decay rate time course as a proxy for steady-state is plotted on the y-axis. Transcripts that did not have significant allele-specific differences in either molecular phenotype are represented by gray dots; transcripts with significant allele-specific differences in transcription rate, but not steady-state RNA levels are represented by red dots; transcripts with significant allele-specific differences in steady-state RNA levels, but not transcription rate are represented by black dots; and transcripts with significant allele-specific differences in both transcription rate and steady-state RNA levels are represented by pink dots. C) Comparison of allele-specific RNA decay rate and steady-state RNA expression levels. For each transcript, the slope of the linear logistic model we fit to the decay rate time course for that transcript, $\lambda_{Y12} - \lambda_{DBVPG6044}$, which represents the difference in decay rate between the Y12 and DBVPG6044 alleles, is plotted on the x-axis and the p_{DB} we measured using the 0 minute time point of our decay rate time course as a proxy for steady-state is plotted on the y-axis. Transcripts that did not have significant allele-specific differences in either molecular phenotype are represented by

gray dots; transcripts with significant allele-specific differences in RNA decay rate, but not steady-state RNA levels are represented by yellow dots; transcripts with significant allele-specific differences in steady-state RNA levels, but not RNA decay rate are represented by black dots; and transcripts with significant allele-specific differences in both RNA decay rate and steady-state RNA levels are represented by gold dots. D) Comparison of allele-specific transcription rate and RNA decay rate. For each transcript, the p_{DB} that we measured in our NRO assay is plotted on the x-axis and the slope of the linear logistic model we fit to the decay rate time course for that transcript, $\lambda_{Y12} - \lambda_{DBVPG6044}$, which represents the difference in decay rate between the Y12 and DBVPG6044 alleles, is plotted on the y-axis. Transcripts that did not have significant allele-specific differences in either molecular phenotype are represented by gray dots; transcripts with significant allele-specific differences in transcription rate, but not RNA decay rate are represented by red dots; transcripts with significant allele-specific differences in RNA decay rate, but not transcription rate are represented by yellow dots; and transcripts with significant allele-specific differences in both transcription rate and RNA decay rate are represented by orange dots.

TABLES

Table 1. Sequencing read coverage, number of transcripts tested for allele-specific differences, and number of transcripts identified with significant allele-specific differences for our chromatin accessibility, transcription rate, RNA decay rate, and translation efficiency assays. Number of sequencing reads corresponds to the raw total number of sequencing reads obtained across all replicates, before any processing. Number of informative reads corresponds the number of filtered sequencing reads that map uniquely to a variant site in an accessible region (ATAC-seq) or transcript (all RNA-based assays) and could be assigned to a parental allele (i.e. Y12 or DBVPG6044).

Molecular phenotype	Assay	Number of sequencing reads	Number of informative sequencing reads	Number of transcripts tested	Number of significant transcripts identified
Chromatin accessibility	ATAC-seq	217,250,917	15,813,630	2,467	1,220
Transcription rate	Nuclear Run-On (NRO)	46,298,182	1,141,660	3,303	1,259
RNA decay rate	Decay time course with drug treatment	226,409,102	53,220,979	4,861	980
Translation efficiency	Polysome fractionation	356,057,087	68,482,137	5,014	1,279

REFERENCES

Abzhanov, A., M. Protas, G. R. Grant, P. R. Grant, and C. J. Tabin, 2004 Bmp4 and morphological variation of beaks in Darwin's finches. Science 305: 1462-1465.

Adams, C. C., and D. S. Gross, 1991 The yeast heat shock response is induced by conversion of cells to spheroplasts and by potent transcriptional inhibitors. J. Bacteriol. 173: 7429-7435.

Albert, F. W., D. Muzzey, J. S. Weissman, and L. Kruglyak, 2014 Genetic influences on translation in yeast. PLoS Genet. 10: e1004692.

Andrie, J. M., J. Wakefield, and J. M. Akey, 2014 Heritable variation of mRNA decay rates in yeast. Genome Res. 24: 2000-2010.

Andrie, J. M., D. Gordon, E. E. Eichler, and J. M. Akey, 2017 High-quality *de novo* genome and transcriptome assembly of two wild-derived *Saccharomyces cerevisiae* strains. Submitted to G3.

Angiuoli, S. V., and S. L. Salzberg, 2011 Mugsy: Fast multiple alignment of closely related whole genomes. Bioinformatics 27: 334-342.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25: 25-29.

Battle, A., Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford *et al.*, 2015 Impact of regulatory variation from RNA to protein. Science 347: 664-667.

Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, 2013 Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10: 1213-1218.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Carbon, S., A. Ireland, C. J. Mungall, S. Shu, B. Marshall, *et al.*, 2009 AmiGO Hub, Web Presence Working Group. AmiGO: online access to ontology and annotation data. Bioinformatics 25: 288-289.

Cenik, C., E. S. Cenik, G. W. Byeon, F. Grubert, S. I. Candille *et al.*, 2015 Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. Genome Res. 25: 1610-1621.

Cherkasov, V., T. Grousl, P. Theer, Y. Vainshtein, C. Glässer, *et al.*, 2015 Systemic control of protein synthesis through sequestration of translation and ribosome biogenesis factors during severe heat stress. FEBS Lett. 589: 3654-3664.

Connelly, C. F., J. Wakefield, and J. M. Akey, 2014 Evolution and genetic architecture of chromatin accessibility and function in yeast. PLoS Genet. 10: e1004427.

Dori-Bachash, M., E. Shema, and I. Tirosh, 2011 Coupled evolution of transcription and mRNA degradation. PLoS Biol. 9: e1001106.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, *et al.*, 2013 STAR: ultrafast universal RNA-Seq aligner. Bioinformatics 29: 15-21.

Foss, E. J., D. Radulovic, S. A. Shaffer, D. M. Ruderfer, A. Bedalov et al., 2007 Genetic basis of proteome variation in yeast. Nat Genet. 39: 1369-1375.

Garneau, N. L., J. Wilusz, and C. J. Wilusz, 2007 The highways and byways of mRNA

decay. Nat. Rev. Mol. Cell Biol. 8: 113-26.
The Gene Ontology Consortium, 2015 Gene Ontology Consortium: going forward. Nucl. Acids Res. 43 Database issue: D1049–D1056.
Ghazalpour, A., B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian *et al.*, 2011 Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. 7: e1001393.
Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644-652.
Haas, B. J., A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr. *et al.*, 2003 Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31: 5654-5666.
Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357-359.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078-2079.
Li, X., H. Kazan, H. D. Lipshitz, and Q. D. Morris, 2014 Finding the target sites of RNA-binding proteins. Wiley Interdiscip. Rev. RNA. 5: 111-130.
Liti, G., D. M. Carter, A. M. Moses, J. Warringer, L. Parts, *et al.*, 2009 Population genomics of domestic and wild yeasts. Nature 458: 337-341.
Love, M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. 15: 550.
MacKay, V. L., X. Li, M. R. Flory, E. Turcott, G. L. Law, *et al.*, 2004 Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. Mol. Cell Proteomics 3: 478-489.
McKinlay, A., C. L. Araya, and S. Fields, 2011 Genome-Wide Analysis of Nascent Transcription in *Saccharomyces cerevisiae*. G3 1: 549-558.
Paaby, A. B., and M. V. Rockman, 2013 The many faces of pleiotropy. Trends Genet. 29: 66-73.
Pai, A. A., C. E. Cain, O. Mizrahi-Man, S. De Leon, N. Lewellen, *et al.*, 2012 The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. PLoS Genet. 8: e1003000.
Picotti, P., M. Clément-Ziza, H. Lam, D. S. Campbell, A. Schmidt *et al.*, 2013 A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. Nature 494: 266-270.
R Core Team, 2013 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Ramani, V., R. Qiu, and J. Shendure, 2015 High-throughput determination of RNA structure by proximity ligation. Nat. Biotechnol. 33: 980-984.
Ronald, J., J. M. Akey, J. Whittle, E. N. Smith, G. Yvert, *et al.*, 2005 Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. Genome Res. 15: 284-291.
Schoenberg, D. R., and L. E. Maquat, 2012 Regulation of cytoplasmic mRNA decay. Nat. Rev. Genet. 13: 246-59.
Silverman, I. M., F. Li, A. Alexander, L. Goff, C. Trapnell et al., 2014 RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human

transcriptome. Genome Biol. 15: R3.
Skelly, D. A., J. Ronald, and J. M. Akey, 2009 Inherited variation in gene expression. Annu. Rev. Genomics Hum. Genet. 10: 313-332.
Skelly, D. A., M. Johansson, J. Madeoy, J. Wakefield, and J. M. Akey, 2011 A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-Seq data. Genome Res. 21: 1728-1737.
Storey, J. D., 2002 A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B 64: 479-498.
Storey, J. D. and R. Tibshirani, 2003 Statistical significance for genome-wide experiments. Proc. Natl. Acad. Sci. USA 100: 9440-9445.
Storey, J. D., J. E. Taylor, and D. Siegmund, 2004 Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. Journal of the Royal Statistical Society, Series B 66: 187-205.
Wakefield, J., 2007 A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am. J. Hum. Genet. 81: 208-227.
Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, *et al.*, 2008 Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9: R137.

CONCLUSIONS AND FUTURE DIRECTIONS

A fundamental objective in modern biology is to determine how genotype produces phenotype. However, despite rapid advancements in the field of genomics, which have led to extraordinary reductions in the cost of sequencing and, concomitantly, a wealth of genome sequence data (reviewed in Goodwin *et al.* 2016), understanding how natural genetic variation among individuals generates differences in gene and protein expression, function, and interaction, and, consequently, in organismal phenotypes, remains a formidable challenge. Through my doctoral research, which culminated in a comprehensive analysis of *cis*-regulatory variation across many molecular phenotypes, I have demonstrated the pervasive influence of *cis*-regulatory variation on the conversion of genotype into phenotype, better delineated the genomic architecture of *cis*-regulatory mutations, and enhanced our understanding of the mechanisms through which *cis*-regulatory variation produces expression variation, and ultimately, phenotypic variation.

When I began my doctoral research, the genetics of gene expression had been the subject of intense interest; however, the contribution of heritable variation in post-transcriptional processes including RNA decay to gene expression variation had received far less attention. To this end, we measured allele-specific differences in RNA decay rates in a diploid yeast hybrid created by mating two genetically diverse *Saccharomyces cerevisiae* strains: the laboratory strain BY4716 (BY) and the wild Californian vineyard strain RM11-1a (RM). In this study, we estimated that 31% of measurable genes exhibit allelic differences in RNA decay rate. We discovered that genes with significant allele-specific differences in RNA decay rate have higher levels of polymorphism compared to

other genes, with all gene regions (i.e. 5' untranslated region, coding region, and 3' untranslated region) contributing to allelic differences in RNA decay rate. Additionally, our results highlighted an important role for variants that affect RNA secondary structure. Finally, we found widespread evidence for compensatory evolution, such that variants influencing transcription initiation and decay have opposite effects, suggesting steady-state gene expression levels are subject to stabilizing selection.

During the time that elapsed while I conducted my study of the *cis*-regulatory variation affecting RNA decay rates, the number of functional genomics technologies that could measure interactions between DNA, RNA, and protein, as well as properties of these biomolecules, like structure or dynamic behavior, expanded rapidly. To further explore how genetic variation among individuals shapes variation in organismal phenotypes, we thus comprehensively investigated allele-specific differences in six molecular phenotypes traversing the conversion of genotype into phenotype, including chromatin accessibility, transcription rate, RNA binding protein-RNA (RBP-RNA) interactions, RNA secondary structure, RNA stability, and translation efficiency in a diploid yeast produced by mating two genetically diverse wild-derived haploid strains of *S. cerevisiae*: the Sake strain Y12 and the West-African strain DBVPG6044. To facilitate accurate allele-specific read mapping in the many functional genomics assays we conducted, a supplementary aspect of our project was to generate, and, subsequently, annotate high quality *de novo* genome assemblies of the Y12 and DBVPG6044 strains. Our study demonstrated that *cis*-regulatory variation has extensive influence on all molecular phenotypes examined, and we uncovered widespread pleiotropy of *cis*-regulatory variation. Our results revealed that relationships between allelic differences in

the measured molecular phenotypes are highly complex, and highlighted the need for additional, integrated studies of the effects of *cis*-variation on gene and protein expression, and ultimately, phenotype.

Moving forward, continuing reductions in the costs of next-generation and long-read sequencing, advances in computational algorithms, and innovations in functional genomics technologies will make obtaining high quality genome assemblies and, subsequently, analyzing of the functional consequences of the variation present in those genomes increasingly feasible. These new data will empower genotype-phenotype inferences as well as testing of hypotheses about the evolution of molecular and organismal traits, and can, thus, be used to build predictive models of how genetic variation ultimately produces phenotypic variation, expanding the foundation I have laid through my thesis research.

One area of research critically needed for accurate prediction of phenotypic variation from genotypic variation involves understanding the processes underlying RNA expression differences. In our cradle-to-grave analysis of *cis*-regulatory variation in yeast, we found a poor correlation between allele-specific chromatin accessibility differences and allele-specific transcriptional differences, suggesting that predicting RNA transcription rate will require knowledge of how variation affects the binding of specific transcription factors. Recently, Hillenbrand *et al.* showed that the rate of mRNA synthesis could be determined over the course of the cell cycle for target genes in the *CLB2* gene cluster in yeast based on the input transcription factor network alone (2016). However, the effects of genetic variation within and between species on transcription factor binding in relation to RNA synthesis remain to be explored. Additionally, while we

and others have studied the interaction of genetic variation with transcription rates and RNA stability in a small number of individuals within or between species (Dori-Bachash *et al.* 2011; Pai *et al.* 2012; Andrie *et al.* 2014), our ability to precisely dissect how variation influences these processes, as well as to understand the evolutionary forces governing such variation, is vitally dependent on broader study of intra- and inter-specific variation in transcription and RNA decay rates. Relatedly, future studies should include both *cis*-regulatory variation, which I focused on in my work, and *trans*-regulatory variation. One classic method for estimating both types of effects is to compare measurements of the molecular phenotypes of interest between the parents and the hybrid diploid. Furthermore, we also need to more widely investigate gene-environment (GxE) interactions by examining transcription rate and RNA decay rate in the same individuals, but across many growth conditions. In addition to more wide-ranging exploration of transcription and RNA stability, our ability to create accurate prediction models for how genotype shapes phenotype will also require comprehensive knowledge of RBP-RNA interactions, RNA secondary structure, and RNA localization. We need to deeply examine how genetic variation can change these molecular phenotypes and what effects those changes exert on RNA synthesis and decay as well as translation efficiency.

While predicting RNA expression levels is undoubtedly a crucial part of modeling phenotype from genotype, previous research has shown that variation in mRNA and protein expression levels are often uncorrelated (Foss *et al.* 2007; Ghazalpour *et al.* 2011). Studies in humans and yeast, including our own cradle-to-grave study, have come to conflicting conclusions on the role of variation in RNA abundance, translation efficiency, and other post-translational processes in determining variation in protein

abundance (Albert *et al.* 2014; Battle *et al.* 2015; Cenik *et al.* 2015). Therefore, a critical area for future research aimed at determining how genotype produces phenotype will be to more broadly explore the relationship between RNA abundance and protein abundance. In particular, as with transcription and RNA decay rates, to better catalog variation influencing translational efficiency, we need to consider both *cis* and *trans* effects, more widely examine intra- and inter-specific translation rate differences, and investigate GxE interactions. Furthermore, we need to couple such exploration with proteomic-based studies of protein expression variation, so that we can accurately assess the contribution of translational differences to protein expression level differences. Along these lines, two studies in humans recently examined inter-individual variation in human populations in steady-state RNA expression levels, translation rate, and protein abundance (Battle *et al.* 2015; Cenik *et al.* 2015); however, these studies were limited, in part, by the accuracy of the quantitative mass-spectrometry methods they used. As evidenced by these studies, development of improved mass-spectrometry methods for quantifying variation within and between individuals in protein levels will empower our ability to truly understand the relationship between RNA expression, translation, and protein level differences, and will be an important component of future research directed at understanding how phenotypic variation is produced from genotypic variation. Additionally, improved mass-spectrometry methodology will enable higher quality studies of how genetic variation influences post-translational protein modifications and protein localization, as well as protein degradation rates.

To synthesize the vast amount of information produced by studies investigating the effects of heritable variation on RNA and protein expression variation, and,

ultimately, achieve the goal of being able to accurately predict organismal phenotype from genomic sequence, an obvious and essential area of future research will also be how to best create predictive models. For example, innovative computational algorithms will be necessary in order to sift through the findings of disparate studies across different populations and in different molecular phenotypes and precisely discern the effects genetic variation on phenotypic variation.

Overall, my thesis research represents the most comprehensive analysis conducted to date of how genetic variation affects gene expression, and, therefore, provides an important advancement toward the understanding of how genotype produces phenotype. Much remains to be learned though, and it will be exciting to see what future studies examining *cis*- and *trans*-regulatory variation, as well the interaction of genetic variation with the environment, reveal about how genetic variation influences variation in RNA expression, protein abundance, and, ultimately, phenotype.

REFERENCES

Albert, F. W., D. Muzzey, J. S. Weissman, and L. Kruglyak, 2014 Genetic influences on translation in yeast. PLoS Genet. 10: e1004692.

Andrie, J. M., J. Wakefield, and J. M. Akey, 2014 Heritable variation of mRNA decay rates in yeast. Genome Res. 24: 2000-2010.

Battle, A., Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford *et al.*, 2015 Impact of regulatory variation from RNA to protein. Science 347: 664-667.

Cenik, C., E. S. Cenik, G. W. Byeon, F. Grubert, S. I. Candille *et al.*, 2015 Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. Genome Res. 25: 1610-1621.

Dori-Bachash, M., E. Shema, and I. Tirosh, 2011 Coupled evolution of transcription and mRNA degradation. PLoS Biol. 9: e1001106.

Foss, E. J., D. Radulovic, S. A. Shaffer, D. M. Ruderfer, A. Bedalov et al., 2007 Genetic basis of proteome variation in yeast. Nat Genet. 39: 1369-1375.

Ghazalpour, A., B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian *et al.*, 2011 Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. 7: e1001393.

Goodwin, S., J. D. McPherson, and W. R. McCombie, 2016 Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17: 333-351.

Hillenbrand, P., K. C. Maier, P. Cramer, and U. Gerland, 2016 Inference of gene regulation functions from dynamic transcriptome data. Elife 5: e12188.

Pai, A. A., C. E. Cain, O. Mizrahi-Man, S. De Leon, N. Lewellen, *et al.*, 2012 The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. PLoS Genet. 8: e1003000.