Extensive and coordinated control of allele-specific expression by both transcription and translation in *Candida albicans*

Dale Muzzey,1 Gavin Sherlock,2,3 and Jonathan S. Weissman1,3

1Department of Cellular and Molecular Pharmacology, California Institute for Quantitative Biomedical Research, Center for RNA Systems Biology, and Howard Hughes Medical Institute, University of California, San Francisco, California 94117, USA; 2Department of Genetics, Stanford University, Stanford, California 94305, USA

Though sequence differences between alleles are often limited to a few polymorphisms, these differences can cause large and widespread allelic variation at the expression level. Such allele-specific expression (ASE) has been extensively explored at the level of transcription but not translation. Here we measured ASE in the diploid yeast *Candida albicans* at both the transcriptional and translational levels using RNA-seq and ribosome profiling, respectively. Since *C. albicans* is an obligate diploid, our analysis isolates ASE arising from *cis* elements in a natural, nonhybrid organism, where allelic effects reflect long-term evolutionary forces. Importantly, we find that ASE arising from translation is of a similar magnitude to transcriptional ASE, both in terms of the number of genes affected and the magnitude of the bias. We further observe coordination between ASE at the levels of transcription and translation for single genes. Specifically, reinforcing relationships—where transcription and translation favor the same allele—are more frequent than expected by chance, consistent with selective pressure tuning ASE at multiple regulatory steps. Finally, we parameterize alleles based on a range of properties and find that SNP location and predicted mRNA-structure stability are associated with translational ASE in *cis*. Since this analysis probes more than 4000 allelic pairs spanning a broad range of variations, our data provide a genome-wide view into the relative impact of *cis* elements that regulate translation.

[Supplemental material is available for this article.]
We reasoned that such errors would manifest as systematic disassembly, and (3) errors in the scripts we used to process the data. Pertinent to its homolog may nonetheless have higher FP signal in a particular SNP window if the sequence or structural features of the allele-specific reads. For each allele, we summed the mRNA and FP reads to our phased diploid assembly of the C. albicans genome, requiring a perfect match at SNP positions in order to identify allele-specific reads. For each allele, we summed the mRNA and FP reads from both replicates across all SNPs; the allelic ratio of mRNA reads reflects transcriptional ASE, and the allelic ratio of translational efficiency ("TE" where TE = FP/mRNA) represents translational ASE.

We validated our data by several analyses. First, if our data were reliable, we would expect to observe a consistent ASE bias across many SNPs within the same gene. Figure 1B shows CHO2, a representative gene with 17 discrete SNP windows, where a SNP window contains all nucleotides that are 30 or fewer nucleotides upstream of a SNP, since—with our average read length of 30 nt—all such positions will contain allele-specific expression information. The sum across SNP windows indicates favored translation of the B allele and comparable transcription of both alleles. Importantly, these features are also apparent within the large majority (15 out of 17) of the 17 SNP windows (Fig. 1B,C). For CHO2, the agreement among SNP windows in terms of which allele was favored at the FP level was 88% (15/17 = 0.88); such agreement is very unlikely to occur if SNP windows randomly favored either allele (P = 0.0023, binomial test). Across all genes, we find average signal coherence across multiple SNPs of 80%–89% for mRNAs and 76%–80% for FPs (ranges depend on sequence coverage, with higher coverage giving higher coherence; see Methods). Coherence among FPs may be slightly lower than for mRNAs due to ribosome-progression effects: For instance, an allele with low overall TE relativity to its homolog may nonetheless have higher FP signal in a particular SNP window if the sequence or structural features of the mRNA at that SNP window cause the ribosome to pause or slow down.

We next considered the impact of alignment error potentially arising from three sources: (1) allele-specific biases in library generation and sequencing, (2) inadequacies with the diploid genome assembly, and (3) errors in the scripts we used to process the data. We reasoned that such errors would manifest as systematic disparities between the levels of allele-specific reads and nonspecific reads. For instance, if our phased assembly only contained half of the real SNPs in the genome, then reads harboring the neglected SNPs would fail to align, leading to lower counts in our ostensibly nonspecific regions. We calculated the coverage by allele-specific reads and compared it with the level of coverage by the nonspecific reads that lack SNPs, and indeed they were closely matched for both CHO2 (Fig. 1D) (P > 0.05, KS test) and across all SNP-containing genes more generally (Fig. 1E) (regression slope = 1.007).

Assessing the significance of ASE
To determine which genes have significant translational ASE bias, we used a bootstrapping strategy. In assigning significance it is important to note that since our measurement of ASE involves a ratio of two sums, our calculations could be highly susceptible to noise. We wanted to distinguish among several scenarios, depicted schematically in Figure 2A for a mock gene with a single SNP. In particular, we aimed to differentiate those genes that have allelic bias reproducibly across allele-specific positions in the gene (scenario #1) from those whose calculation of TE is dominated by a minority of positions (scenario #2), or those that simply lack allelic bias (scenario #3). For each gene, we compiled the list of all nucleotide positions in SNP windows (i.e., all reads mapping to these positions contain allele-specific information) that span the open reading frame (see Methods). Since each SNP window is ~30 nt in length, this list generally contains ~30 positions per SNP window. For a list of length x, we next created 5000 new lists, each containing x positions that were chosen at random and with replacement from the original list. For each random list, the TE was calculated from the mRNA and FP reads at the randomly chosen positions and compiled into a “bootstrap distribution.” The bootstrap distribution shows the likelihood of a given ASE bias based on the empirically observed read counts. For scenarios #1 and #3, the bootstrap distributions are narrow, since nearly all positions have the same balance between FP_d, FP_b, mRNA_d, and mRNA_b. Thus, random lists comprised of these positions will tend to report similar TE values. Conversely, for scenario #2, the bootstrap distribution is wide, since the eighth position—which alone has strong allelic disparity—may be included once, many times, or not at all, in the random lists. We used the means and standard deviations of the bootstrap distributions to classify whether or not genes have a significant translational ASE bias. The bootstrap distributions for the majority of genes in the genome have both low means and low standard deviations, consistent with a high confidence of no allelic bias (Fig. 2B, high gene density in lower left). However, a subset of genes has bootstrap distributions with means that are many standard deviations away from zero, suggestive of allelic bias. To identify such genes systematically, for each gene we calculated a P-value from a Z-score reflecting the number of standard deviations separating the bootstrap mean from zero. From the resulting distribution of P-values (Supplemental Fig. S3), we selected a cutoff such that the false-discovery rate is 5%. These analyses adjust for multiple hypothesis testing and revealed that 139 genes—4.2% of those tested—show strong and high-confidence allelic bias in TE. We used FuncAssociate (Berriz et al. 2003) and the Candida Genome Database GO Term Finder (Inglis et al. 2012) to identify significantly enriched functional categories among genes with allelic TE bias (see Methods) and found an unexpected abundance of nuclear-encoded transcripts for mitochondrially targeted proteins.
Figure 1. Sensitive detecting ASE at translational level with ribosome profiling. (A) Schematic of the approach. For a given gene with two SNPs, transcripts from the B allele may be more abundant, whereas translation favors the A allele, as indicated by increased density of ribosomes, shown in green. These biases are revealed by RNA-seq and ribosome profiling, respectively. Allele-specific reads are summed across all SNPs in the gene, and translational efficiency (“TE”) is calculated from the mRNA and footprint (“FP”) levels. (B, C) Signal is consistent across many SNPs. There are 17 distinct SNP windows in orf19.169/CHO2, and the majority indicates a translational bias toward the B allele, but roughly equal transcript levels (B), with little error across SNPs (C); error bars, ± SEM. (D) The sum of allele-specific reads (red and blue bar) matches the level of nonspecific reads that do not include SNPs (gray bar) for orf19.169; error bars, ± SEM. (E) Across all genes, the fraction of SNP-containing reads corresponds strongly to the fraction of gene length comprised of SNP-containing windows.
Transcription and translation have comparable effects on ASE

Since our data set measures ASE at both the transcriptional and translational levels, we sought to determine the relative ASE contributions of these two mechanisms. For genes with bootstrap standard deviations below 0.4 at the TE level (Fig. 2B), we compared the respective allelic levels of mRNA, FP, and TE (Fig. 3A–C) and further compiled these values into histograms that depict allelic log-fold-difference ("ALFD") (Fig. 3D). The standard deviations of ASE differences between biological replicates were less than half of the respective standard deviations of the observed ALFD distributions (Fig. 3D; Supplemental Fig. S1), arguing against a common noise floor that similarly affects each distribution. Additionally, noise in ALFD from low-sequencing coverage is not a main driver of the distribution widths, since median read counts for genes at the extremities of the distributions are in excess of 50 (Supplemental Fig. S4).

The ALFD histograms for all three metrics were largely superimposable (\( P > 0.05 \) for all three pairwise comparisons among distributions, Mann-Whitney U-test), indicating that allelic variability at all three levels affects a comparable number of genes by a similar amount. However, such a perfect correlation would lead to a delta function for the TE ALFD distribution, since \( \text{TE}_{\text{ALFD}} = \text{FP}_{\text{ALFD}} - \text{mRNA}_{\text{ALFD}} \). If distributions for mRNA ALFD and FP ALFD overlapped but the values were uncorrelated, then the distribution width for \( \text{TE}_{\text{ALFD}} \) would be higher than the respective widths for mRNA ALFD and FP ALFD, since the variances for independent values would be additive. Indeed, the congruence among all three distributions reflects the positive but not perfect correlation (\( R = 0.50 \)) that we observe between mRNA ALFD and FP ALFD levels.

To further explore our results, we implemented a simple simulation that models transcription and translation of two alleles (see Methods). In brief, for a given gene, we chose its transcriptional and translational propensities randomly from lognormal distributions. To get allelic values for mRNA and FP, we scaled the respective propensities by normally distributed noise terms to get mRNA A, mRNA B, FP A, and FP B values. The parameters for transcription were drawn directly from the empirical distribution in Figure 3A, and the mean translational propensity was derived from Figure 3B. Two fit parameters—one for the width of the lognormal translational propensity and the other for translational noise—allowed the model to match the dispersion of points in both the FP and TE 2D histograms (Fig. 3F,G). In total, the simulation corresponds well with the observed data both at the single-gene level (Fig. 3E–G) (\( R^2 = 0.946 \); parameters listed in Supplemental Table S1) and in aggregate in the ALFD histograms (Fig. 3H), and we performed cross-validation to exclude the possibility of overfitting (Supplemental Fig. S5). As expected from the observed overlap in mRNA ALFD and FP ALFD distributions, the noise-strength parameters that drive allelic differences for transcription and translation in the

Figure 2. A total of 4.2\% of genes show translational allelic bias. (A) Schematic of the bootstrapping procedure. (Top) For a mock gene containing a single SNP, \(-30\) consecutive positions contain allele-specific information, and three scenarios for read distributions are shown: (#1) shows reproducible bias toward allele A in blue; (#2) shows how a single position could suggest a bias that is not supported by other positions, and (#3) shows a consistently reported lack of bias. (Middle) For each of 5000 iterations, 30 positions are selected randomly and with replacement, and a TE value is calculated from the mRNA and FP reads from those positions. (Bottom) The results are tabulated into a histogram, where the mean and standard deviation of the bootstrap distribution reflect the magnitude and confidence, respectively, of allele-specific bias in TE. (B) Scatterplot and accompanying histograms (top and right) showing the bootstrap means and standard deviations for the 3285 genes with at least five reads for mRNA A, mRNA B, FP A, and FP B (shading indicates the metric with the fewest read counts, as shown in the legend at bottom). Blue-rimmed circles indicate genes that pass the 5\% FDR threshold.
Allelic TE bias tends to reinforce transcriptional bias more than expected by chance

Our measurements of ASE at both the transcriptional and translational levels allowed us to explore whether the two processes interact in a systematic fashion to influence allelic bias. For instance, allelic disparity at the transcriptional level that favors one allele may be offset by higher TE for the opposite allele, a compensatory interaction that stabilizes FP levels (Fig. 4A). Alternatively, if transcription and TE favor the same allele in a reinforcing manner, FP levels are highly disparate (Fig. 4B). We sought to assess the prevalence of such interactions from a scatter plot of mRNA_ALFD versus TE_ALFD (Fig. 4C), where genes with reinforcing interactions have fold changes of the same sign, and those with compensatory interactions have opposite signs. There is a negative correlation of −0.33 between the twofold changes, which would seem to suggest a prevalence of compensatory interactions, as reported recently from analysis of comparable plots (Artieri and Fraser 2014; McManus et al. 2014). However, since TE_ALFD is a monotonically decreasing function of mRNA_ALFD, this negative correlation is to be expected and must be factored into subsequent analyses of enrichment.

To determine whether compensatory or reinforcing interactions are enriched, we performed a permutation analysis. First, from the empirical plot in Figure 4C, we demarcated regions shown in purple and green that represent reinforcing and compensatory interactions, respectively. Next, we generated 1000 randomized null data sets in which the empirical mRNA_ALFD distribution was used to create randomized FP_ALFD and TE_ALFD distributions that match key features of the empirical data (e.g., mean, standard deviation, and correlation with mRNA_ALFD) but lack bias toward compensatory or reinforcing relationships (see Methods). From each randomized data set, we counted the number of genes in the compensatory and reinforcing colored regions and compiled the results into histograms (Fig. 4D,E). This analysis revealed that compensatory interactions are not more abundant than expected by chance. However, reinforcing interactions are significantly enriched (P < 0.01), suggesting that allele-specific bias can be coordinated at multiple expression levels, with a trend toward maximizing expression differences.

Identification of cis elements that affect translational efficiency

We sought to identify cis features of mRNAs that affect how well they are translated. Specifically, from our set of 2132 allelic pairs that have high-confidence measurements of allele-specific TE (i.e., all genes in Fig. 3), we wanted to determine sequence properties that distinguish genes with strong allelic bias in TE from those lacking such bias.

We first investigated how well codon bias might explain differences in allelic translational efficiency. We considered three gene sets based on their TE_ALFD values (Fig. 5A), calculated the codon adaptation indices (“CAI”) (Sharp and Li 1987) for alleles in each set, and found no significant disparity among the resulting distributions (Fig. 5B), consistent with codon bias playing a minor role, if any, in determining allelic TE bias in C. albicans. We similarly observed no correspondence between codon-pair bias and TE_ALFD (Supplemental Fig. S6).

We next considered whether the computed stability of mRNAs near the start codon was predictive of TE_ALFD in C. albicans, since it has been found to play a role in bacteria (Kudla et al. 2009). We focused on the 60-nucleotide window centered on the start codon; for each allele with at least one SNP in this window, we used Mfold
Allelic pairs with the expected relationship between 5' TEALFD is twofold greater than the number with the unexpected. From the low-|TEALFD|averaged PDF, we found that the high-|TEALFD|averaged PDF has a significantly elevated sum-squared deviation compared to the low-|TEALFD|set (see Methods). We next considered whether these averaged PDFs differed significantly from each other and found that the high-|TEALFD|averaged PDFs from randomly subsampled gene subsets of the same gene (Gregg et al. 2010; DeVeale et al. 2012; Kelsey and Bartolomei 2012), our RNA-seq and ribosome profiling data sets have high consistency across phased SNPs within a gene. The reproducibility across biological replicates further supports the quality of our data, and we were careful to filter genes using our bootstrapping approach to ensure that our analyses focused on genes with high-confidence measures of ASE.

Figure 4. Biases in transcription and translation are often coordinated, with interactions favoring compensation over reinforcing. (A,B) Specific examples of compensatory (A) and reinforcing (B) interactions between transcription and translation. Genes with compensatory interactions have higher allelic difference at the mRNA level than at the FP level, whereas those with reinforcing interactions differ more at the FP level than at the mRNA level. (C) Scatterplot of mRNAALFD and TEALFD levels, where genes from A and B are indicated as darkened circles, and shaded regions indicate compensatory and reinforcing interactions, as indicated. The purple and green regions' curved portions reflect the two-standard-deviation spread of the data along the y = x and y = -x lines, respectively, and straight segments are based on a heuristic chosen to ensure that both the mRNAALFD and TEALFD values are nonzero (see Methods). (D,E) PDFs indicating the number of reinforcing (D) and compensatory (E) genes from permuted data; arrows indicate the number of genes from the observed data.

Discussion

Expression variation among alleles in diploid organisms is important for both clinical and research applications. Studies of allele-specific expression have appreciably advanced our understanding of transcriptional control, both in cis and trans. Here we have performed the first highly sensitive, genome-wide, allele-specific assays of both transcription and translation in a natural organism. Though previous sequencing-based assays of allele-specific expression have been cast in doubt due to inconsistencies in signal across SNPs of the same gene ( Gregg et al. 2010; DeVale et al. 2012; Kelsey and Bartolomei 2012), our RNA-seq and ribosome profiling data sets have high consistency across phased SNPs within a gene. The reproducibility across biological replicates further supports the quality of our data, and we were careful to filter genes using our bootstrapping approach to ensure that our analyses focused on genes with high-confidence measures of ASE.

There is consensus between our work and two recent investigations of translational ASE (Artieri and Fraser 2014; McManus et al. 2014) that the prevalence and magnitude of ASE bias are similar for both transcription and translation. In our data, this is indicated both by the overlap among the ALFD distributions of mRNA, FP, and TE by the allelic-bias parameters of our simulations. Among the many implications of this observation is that the steps are not optimized for specific size biases. In other words, for a gene to achieve a particular allelic expression bias, either its transcriptional or translational control can be tuned. Given their
comparable magnitude, we anticipated that there would be widespread compensation between transcription and translation, such that alleles tend to be comparably expressed. However, statistical analysis using our null model indicated that compensatory interactions were no more prevalent than expected by chance, and instead it is the reinforcing interactions that were overrepresented. Though we found no functional enrichments among the reinforcing genes, it will be interesting to attribute the transcriptional and translational effects to particular SNPs, to assess their evolutionary history, and to determine whether the reinforcing interaction confers a fitness advantage to the cell.

A priori, we expected allelic bias at the translational level to be indiscriminate with respect to function, and thus we were surprised to see enrichment for allelic bias among genes with mitochondrial roles. Importantly, our analyses only considered genes encoded on the eight nuclear chromosomes, not those in the mitochondrial genome. Since mitochondrial genomes are known to be fast evolving relative to nuclear genomes (Brown et al. 1979), we speculate that mitochondrially targeted proteins encoded on the nuclear chromosomes may need to evolve at a similarly rapid rate to keep pace functionally with their counterparts encoded in the mitochondrial genome. This possible source of evolutionary pressure may manifest as the observed ASE bias, and, interestingly, this ASE bias is restricted to translational control, since mRNA allelic variability was the same for mitochondrial genes as for other genes.

Because of the above, we do not observe any translation-based differences among the genes in our study. Instead, it is the reinforcing interactions that were overrepresented. Though we found no functional enrichments among the reinforcing genes, it will be interesting to attribute the transcriptional and translational effects to particular SNPs, to assess their evolutionary history, and to determine whether the reinforcing interaction confers a fitness advantage to the cell.
It is likely that each gene’s ASE bias is a combination of multiple determinants, many of which we may not have yet identified. Resolving the effect of mRNA secondary structure on allele-specific TE will be greatly facilitated by recent techniques that probe RNA structure genome-wide (Rouskin et al. 2014; Wan et al. 2014). However, for SNPs that do not affect mRNA structure, it is likely that the variants in cis sequences alter the interaction with trans factors in a way that influences ASE. Thus, additional work will be needed to search for disrupted RNA-binding protein motifs in alleles with high translational ASE bias and to verify a role for RNA-binding proteins via perturbation of the trans-factor composition of the cell. C. albicans is a well-suited model organism for these trans-factor perturbations because it can exist as a homzygous diploid for certain chromosomes (thereby only expressing a subset of trans-factor alleles), or as a haploid amenable to genetic manipulation (Hickman et al. 2013). Future studies of both cis- and trans-regulation of ASE also have a promising future in human cells, where the number of phased genomes is growing rapidly (Ma et al. 2010; Fan et al. 2011; Peters et al. 2012) and haploid cell lines are now available (Carette et al. 2009).

Methods

RNA-seq and ribosome profiling

Libraries for RNA-seq and ribosome profiling from the SC5314 strain of C. albicans were prepared as described in Brar et al. (2011), with a few exceptions. Cells were harvested via filtration in a 30°C room and then flash frozen in liquid nitrogen without the pre-treatment of a translational inhibitor. To guard against post-lysis translation in the ribosome-profiling sample, we supplemented the lysis buffer with 50 μg/mL GMPPNP and 10 μg/mL Blastidin S (InvivoGen), since cycloheximide does not affect the ribosomes of C. albicans (Yamaguchi and Iwata 1970). Oligos used for rRNA subtraction are listed in Supplemental Table S2. Deep sequencing was performed on the Illumina Genome Analyzer Ix and HiSeq 2000. Reads from the two biological replicates were pooled to enhance total signal.

Alignment

Reads were aligned using Bowtie v0.12.7 (Langmead et al. 2009), allowing no mismatches but up to 16 match locations per read. The Bowtie library was built from the diploid FASTA file of the C. albicans genome containing SNPs but not indels (Muzzey et al. 2013); indels are almost exclusively noncoding and were excluded to facilitate analysis. Custom programs in C and scripts in PHP were used to parse Bowtie output and assign reads to their respective alleles based on their SNP composition. To minimize cloning-bias artifacts, reads in which the SNP occurred at the first or last position were not assigned to a particular allele. A two-state, strand-specific hidden Markov model (“HMM”) was used to delineate gene boundaries based on our observed FP sequencing data. This unsupervised method of gene-boundary detection was used to enhance ASE signal measurements by maximizing the number of nucleotide positions that have FP density contiguous with a gene’s annotated coding region. Thus, the HMM performs two main functions. First, it corrects for misannotations in gene boundaries that could underestimate gene length and thereby compromise the accuracy of ASE measurements. Second, since FP reads range in length from 28 to 32 nt and were mapped to the genome based on their 5’ ends as in Ingolia et al. (2009), the HMM automatically accounts for the fact that there is a variable offset between the 5’ end of a read and the part of the read that was in the P site; rather than readjust 5’-end alignment positions by subtracting a fixed and potentially flawed offset value to approximate the P-site position, the HMM simply identifies the empirical gene boundary based on the density of the 5’ ends of reads. The HMM had two states—“Coding” and “Noncoding”—and focused only on strand-specific FP reads, since FP regions are embedded within mRNA regions by definition. For simplicity, it converted the FP signal to a binary string, where positions with zero read counts were called “0” and those with nonzero read counts were assigned “1.” Transition and emission probabilities were tuned with two considerations: (1) maximizing the number of open-reading frames spanned by a single, contiguous “Coding” region, and (2) minimizing the number of contiguous “Coding” regions that spanned multiple open-reading frames. The parameters noted in Table 1 were used in the final HMM, with performance illustrated in Supplemental Fig. S9:

A gene’s final boundaries were determined as follows:

- 5’ boundary: If the start-codon position from the Assembly 21 GFF file ([Inglis et al. 2012] and http://www.candidagenome.org/) was embedded in an HMM-identified “Coding” region, the 5’ end of the “ Coding” region was used; otherwise the start-codon position was used.
- 3’ boundary: If the stop codon was embedded in an HMM-identified “Coding” region, the 3’ end of the “Coding” region was used; otherwise the stop-codon position was used.

In general, the HMM-demarcated boundaries differed little from the annotated gene boundaries, with a median length increase of 1.7%. Read counts per gene are included in Supplemental Table 3.

Signal coherence across SNPs

The main goal of our assessment of signal agreement across SNPs is to ensure that SNPs reporting a large mRNA or FP bias in favor of one allele do not reside in the same gene as other SNPs strongly favoring the opposite allele. To this end, we restricted our analysis to genes where at least 75% of SNPs each report an allelic difference of 25% or more, giving 185 genes at the mRNA level and 320 genes at the FP level. For such genes, we counted the number of all SNPs that favor the A allele and divided by the total number of SNPs in the gene to get the A-allele coherence. We calculated the B-allele coherence as 1-(A-allele coherence) and called the gene’s SNP coherence as 1-(A-allele coherence) and called the gene’s SNP coherence across SNPs with custom PHP scripts using the algorithm described in the main text. For each gene, the results from the 5000 iterations were written to a text file and processed in batch with Python to get the mean and standard deviation of bootstrap distributions. These data were then analyzed in Excel. Bootstrapping means and standard deviations are included in Supplemental Table 3.

Functional enrichment analysis

We performed two functional-enrichment tests, one on the set of genes shown to have strong ASE bias in TE at a 5% FDR (N = 139),
and a larger set in which we included the top 1000 genes based on their Z-score (i.e., how many standard deviations the bootstrap mean is from zero). The gene universe was all genes for which \( \min(\text{mRNA}_{\text{ALFD}}) \), \( \text{mRNA}_{\text{ALFD}} \), \( \text{FP}_{\text{ALFD}} \), \( \text{FP}_{\text{ALFD}} \) > 5 reads (N = 3285 genes). The large set—but not the small set—had significant functional enrichments after correcting for multiple hypothesis testing in FuncAssociate (Berriz et al. 2003) and on the GO Term Finder on the Candida Genome Database (Inglis et al. 2012).

### Simulation and fitting

The stochastic simulation was implemented and fit using a custom Python script with the following equations:

\[
\begin{align*}
    n_{\text{trans}} &= e^{k_1 + k_2 + \text{randn}(\text{)}}, \\
    n_{\text{ntnl}} &= e^{k_3 + k_4 + \text{randn}(\text{)}}, \\
    \text{mRNA}_A &= \max(1, n_{\text{trans}} \times (1 + k_5 + \text{randn}())) , \\
    \text{mRNA}_B &= \max(1, n_{\text{trans}} \times (1 + k_6 + \text{randn}())) , \\
    \text{FP}_A &= \max(0, \text{mRNA}_A \times n_{\text{ntnl}} \times (1 + k_7 + \text{randn}()) ) , \\
    \text{FP}_B &= \max(0, \text{mRNA}_B \times n_{\text{ntnl}} \times (1 + k_8 + \text{randn}()) ) , \\
    \text{TE}_A &= \text{FP}_A / \text{mRNA}_A , \\
    \text{TE}_B &= \text{FP}_B / \text{mRNA}_B ,
\end{align*}
\]

where \( n_{\text{trans}} \) and \( n_{\text{ntnl}} \) are the log-normally distributed transcription and translation propensities, respectively; \( k_i \) and \( k_j \) are the mean and standard deviation of the normal distribution exponentiated to yield \( n_{\text{trans}} \), and \( k_4 \) and \( k_8 \) are the corresponding parameters for \( n_{\text{ntnl}} \). \( k_3 \) and \( k_6 \) represent the allelic noise variables for transcription and translation, respectively; finally, \( \text{randn}(\text{)} \) is a normally distributed random variable with mean 0 and variance of 1. Only \( k_4 \) and \( k_8 \) were fit parameters; the others were calculated directly from the data as follows: \( k_1 \) is the mean of the empirical ln(mRNA) data (Fig. 3A); \( k_2 \) is the standard deviation of the ln(mRNA) data (Fig. 3B); \( k_3 \) is the standard deviation of the distances of \( \text{ln}(\text{mRNA}) \) data from \( y = x \) (Fig. 3A), and \( k_4 \) is the mean of the ln(FP) data divided by \( k_1 \) (Fig. 3B). All parameter values are summarized in Supplemental Table S1. Each run of the simulation iteratively executed equations 1 through 8 3285 times, once for each gene in the observed data set. The mRNA, FP, and TE allelic log-fold differences were then plotted as a 2D PDF, shown as heatmaps in Figure 3, E–G. Error for a single run of the simulation was calculated as the sum of squared differences between the simulated and observed values for each bin in the heatmap grids shown in Figure 3, E–G. For each set of parameter values \( k_5 \) and \( k_6 \) (each parameter sampled over the range [0,5]), the model was run 10 times to yield an average error, and the parameters generating the smallest average error are reported in Supplemental Table S1, and correspondence between the observed and fit data is in Supplemental Figure S5.

### Significance testing of compensatory and reinforcing interactions

For each gene \( i \) with observed allelic log-fold difference (“ALFD”) value of mRNA\(_{\text{ALFD}}[i] \), a random FP\(_{\text{ALFD}}[i] \) (with \( j \neq i \)) was selected from the empirical FP\(_{\text{ALFD}} \) distribution and then scaled to yield FP\(_{\text{ALFD}}[i] \) such that the correlation between mRNA\(_{\text{ALFD}} \) and FP\(_{\text{ALFD}} \) matched the observed correlation between mRNA\(_{\text{ALFD}} \) and FP\(_{\text{ALFD}} \) (\( r = 0.50 \)) using the following equation: \( \text{FP}_{\text{ALFD}}[i] = r \times \text{mRNA}_{\text{ALFD}}[i] + (1 - r^2) \times \text{FP}_{\text{ALFD}}[i] \). Importantly, after this scaling, the mean and standard deviation of FP\(_{\text{ALFD}} \) and FP\(_{\text{ALFD}} \) are equivalent. Finally, we calculated TE\(_{\text{ALFD}}[i] = \text{mRNA}_{\text{ALFD}}[i] - \text{FP}_{\text{ALFD}}[i] \). The number of genes falling in shaded regions delineated in Figure 4C were counted from the scatterplot of mRNA\(_{\text{ALFD}} \) and TE\(_{\text{ALFD}} \). These regions were specified based on the spread of the observed data. Specifically, for compensatory shaded regions, all datapoints were projected onto the line \( y = -x \), the standard deviation of the projected data was calculated, and the green shaded region indicates distances of at least two standard deviations from the origin. For reinforcing interactions, corresponding steps were performed after the data were projected onto the line \( y = x \). Finally, we imposed the heuristic requirement that abs(log10(abs(mRNA\(_{\text{ALFD}}[i] \) / TE\(_{\text{ALFD}}[i] \))) < 0.5 such that we focus specifically on genes for which both mRNA\(_{\text{ALFD}} \) and TE\(_{\text{ALFD}} \) change.

### CAI and mRNA structure prediction

With a custom PHP script, CAI and codon-pair-bias scores were calculated for each gene using their frequencies in coding genes. mRNA structure predictions were generated using Quikfold (http://mfold.maa thuộc University of California at Albany (http://mfold.maa thuộc University of California at Albany Quikfold), with the output processed with custom scripts.

### SNP positional enrichment at 5’ and 3’ ends

Two hundred genes with TE\(_{\text{ALFD}} > 0.58 \) (a 50% difference between alleles) were defined as the “high|TE\(_{\text{ALFD}}|| \) group, and the 1000 genes with lowest|TE\(_{\text{ALFD}}|| \) were defined as the “low|TE\(_{\text{ALFD}}|| \) group. In each group, for each gene with at least three SNPs a five-bin PDF of SNP location was calculated, and all such PDFs were summed and normalized to yield the averaged PDFs in Figure 5E. Deviation of averaged PDFs from the uniform distribution was determined by tabulating the empirical number of genes with at least three SNPs from each group (i.e., 200 genes in the high TE\(_{\text{ALFD}}|| \) group and 1000 genes in the low TE\(_{\text{ALFD}}|| \) group) as well as their empirical number of SNPs per gene, and then generating random single-gene PDFs based on uniform distribution of SNPs among the bins. As with the empirical data, these random PDFs were summed, normalized, and then compared with a uniform distribution using the sum-squared deviation across the five bins. Ten thousand iterations of this procedure for each group yielded the distributions in Figure 5, F and G. Deviation of the two averaged PDFs from each other was determined by creating 1000 random subsampled sets from the low|TE\(_{\text{ALFD}}|| \) group (N = 1000), with each subsampled set having 200 genes such that the subsampled sets had the same number of genes as the high|TE\(_{\text{ALFD}}|| \) group (N = 200 genes). For averaged PDFs from all subsampled sets as well as the high|TE\(_{\text{ALFD}}|| \)
group, the sum-squared deviation was calculated relative to the averaged PDF of the entire low-|TEALFD| group (i.e., gray trace in Fig. 5E) and shown in Supplemental Fig. S8.

Data access
All raw and processed data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE52236.

Acknowledgments
We thank Suzanne Komili, Frank Albert, and Leonid Kruglyak for helpful advice, and Murat Tasan and Fritz Roth for FuncAssociate support. D.M. is an HHMI Fellow of the Damon Runyon Cancer Research Foundation. This work was also supported by the Howard Hughes Medical Institute (J.S.W.) and NIH grants R01-HG003468 (G.S.) and R01-DE015873 (G.S.).

References
Artieri CG, Fraser HB. 2014. Evolution at two levels of gene expression in yeast. Genome Res 24: 411–421.
Bai Y, Zhou K, Doudna JA. 2013. Hepatitis C virus 3'UTR regulates viral translation through direct interactions with the host translation machinery. Nucleic Acids Res 41: 7861–7874.
Barbosa C, Peixeiro I, Romo R. 2013. Gene expression regulation by upstream open reading frames and human disease. PLoS Genet 9: e1003529.
Berthier GR, King OD, Bryant B, Sander C, Roth FP. 2003. Characterizing gene sets with FuncAssociate. Bioinformatics 19: 2502–2504.
Bolze A, Mahlaoui N, Bley M, Turner B, Trede N, Ellis SR, Abhayankar A, Artieri CG, Fraser HB. 2014. Evolution at two levels of gene expression in Candida albicans and Candida glabrata. Nucleic Acids Res 40: D667–D674.
Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218–223.
Jacobson EM, Conception E, Oash T, Turner Y. 2005. A Graves' disease-associated Kozak sequence single-nucleotide polymorphism enhances the efficiency of CD40 gene translation: a case for translational pathophysiology. Endocrinology 146: 2684–2691.
Kellis M, Patterson N, Erdin D, Birren B, Lander ES. 2004. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 428: 241–254.
Kelsey G, Bartolomei MS. 2012. Imprinted genes … and the number is? PLoS Genet 8: e1002601.
Khan Z, Bloom JS, Amini S, Singh M, Perlman DH, Caudy AA, Kruglyak L. 2012. Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. Mol Syst Biol 8: 602.
Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science 324: 255–258.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to a genome. Genome Biol 10: R25.
Lee K-H, Kim S-H, Kim H-J, Kim W, Lee H-R, Jung Y, Choi J-H, Hong KY, Jang SK, Kim K-T. 2014. AUFI contributes to Cryptochrome mRNA degradation and rhythmic translation. Nucleic Acids Res 42: 3590–3606.
Lelefvre BJ, Vello E, Ge B, Montgomery SB, Dermitzakis ET, Pastinen T, Labuda D. 2012. Genotype-based test in mapping cis-regulatory variants from allele-specific expression data. PLoS ONE 7: e38667.
Ma L, Xiao Y, Huang H, Wang Q, Rao W, Feng Y, Zhang K, Song Q. 2010. Direct determination of molecular haplotypes by chromosome microdissection. Nat Methods 7: 299–301.
Mangus DA, Evans MC, Jacobson A. 2003. Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. Genome Biol 4: 223.
McMains J, May GE, Spealman P, Shityeyman A. 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. Genome Biol 24: 422–430.
Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777.
Muzzey D, Schwartz K, Weissman JS, Sherlock G. 2013. Assembly of a phased diploid Candida albicans genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. Genome Biol 14: R97.
Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet 11: S33–S38.
Peters BA, Kermami BG, Sparks AB, Afferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang Y, Haas J, et al. 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487: 190–195.
Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Nisbett J, Haltion J, Pritchard JK, et al. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772.
Press I, Hentze MW. 1999. From factors to mechanisms: translation and translational control in eukaryotes. Curr Opin Genet Dev 9: 515–521.
Rouskin S, Zubratić M, Washietl S, Kells M, Weissman JS. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature 513: 701–705.
Sharp PM, Li WH. 1987. The codon-anticodon recognition process. Curr Opin Genet Dev 9: 515–521.
Stumpf CR, Ruggero D. 2011. The cancerous translation apparatus. Curr Opin Genet Dev 21: 474–483.
Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, et al. 2014. Landscape and variation of RNA secondary structure across the human transcriptome. Nature 505: 706–709.
Warner JR. 1999. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci 24: 437–440.
Yamaguchi H, Iwata K. 1970. In vitro and in vivo protein synthesis in Candida albicans. 3. Protein synthesis and inhibition. Sabouraudia 8: 201–211.
Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31: 3406–3415.

Received October 25, 2013; accepted in revised form April 11, 2014.