bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genetic influences on translation in yeast

Frank W. Albert[1], Dale Muzzey[2], Jonathan Weissman[2], and Leonid Kruglyak[1,3,4]

[1] Department of Human Genetics, University of California, Los Angeles, Los Angeles, California, USA

[2] Department 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, San Francisco, California, USA

[3] Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, California, USA

[4] Howard Hughes Medical Institute, University of California, Los Angeles, Los Angeles, California, USA

Contact information:

Leonid Kruglyak
Gonda Research Center
695 Charles E. Young Drive South
Los Angeles, CA 90095
USA
Phone: 310-825-5486
Fax: 310-795-5446
Email: LKruglyak@mednet.ucla.edu

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Heritable differences in gene expression between individuals are an important source of phenotypic variation. The question of how closely the effects of genetic variation on protein levels mirror those on mRNA levels remains open. Here, we addressed this question by using ribosome profiling to examine how genetic differences between two strains of the yeast *S. cerevisiae* affect translation. Strain differences in translation were observed for hundreds of genes, more than half as many as showed genetic differences in mRNA levels. Similarly, allele specific measurements in the diploid hybrid between the two strains revealed roughly half as many *cis*-acting effects on translation as were observed for mRNA levels. In both the parents and the hybrid, strong effects on translation were rare, such that the direction of an mRNA difference was typically reflected in a concordant footprint difference. The relative importance of *cis* and *trans* acting variation on footprint levels was similar to that for mRNA levels. Across all expressed genes, there was a tendency for translation to more often reinforce than buffer mRNA differences, resulting in footprint differences with greater magnitudes than the mRNA differences. A reanalysis of two earlier studies which reported translational buffering between two yeast species showed that translational reinforcement is in fact more common between these species, consistent with our results. Finally, we catalogued instances of premature translation termination in the two yeast strains. Overall, genetic variation clearly influences translation, but primarily does so by subtly modulating differences in mRNA levels. Translation does not appear to create strong discrepancies between genetic influences on mRNA and protein levels.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Many genetic differences among individuals influence gene expression levels. Such regulatory variants are responsible for a large fraction of the variation in disease risk among humans and are also thought to be important for the evolution of phenotypes [1-3]. Regulatory variants can be mapped as expression quantitative trait loci (eQTL). Due to the relative ease and low cost of mRNA quantification, most eQTL studies have used levels of mRNA, rather than protein, as a measure of gene expression. The few initial studies that examined genetic influences on protein levels reported surprisingly different genetic architectures for protein and mRNA levels [4-6]. For a given gene, many eQTL did not correspond to a protein QTL (pQTL) and vice versa. Some analyses even suggested that eQTL and pQTL for certain groups of genes have significantly less overlap than expected by chance [5]. While more recent work [7-9] has found that eQTL and pQTL are more concordant than seen in the initial studies, numerous discrepancies remain. Together, these results have been taken to suggest that there must be substantial genetic variation acting on posttranscriptional processes.

Translation is an important determinant of cellular protein abundance ([10], but see [11]) and the rate of translation was shown to be a better predictor of protein levels than mRNA abundance [12]. Therefore, genetic variants that specifically influence translation are a potential explanation for the reported discrepancies between eQTL and pQTL.

Differences in gene expression between individuals can be caused by genetic variants that act in *cis* or by variants that act in *trans* [2]. Variants that act in *cis* influence the expression of alleles to which they are physically linked. In a diploid organism, *cis* acting variants can be detected as preferential expression of one allele compared to the other ("allele-specific expression", ASE) [13-18]. By contrast, *trans* acting variants influence the expression of both alleles of a gene to a similar extent.

Both *cis* and *trans* acting variants might have effects on translation. To affect translation in *cis*, a variant needs to reside within the mRNA of the given gene. By contrast, genetic variation in the various translation factors [19] might influence

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

translation in *trans*. Further, mutations in ribosomal proteins can lead to highly specific differences in translation of small groups of mRNAs during mouse development [20], suggesting that genetic differences in genes beyond classic translation factors could affect translation in *trans*.

In this paper, we explored the influence of genetic variation on translation. We measured genome-wide translational activity in two genetically different strains of the yeast *S. cerevisiae* – the laboratory strain BY and the wine strain RM – as well as their diploid hybrid. Translation was measured by massively parallel sequencing of "ribosome footprints", i.e. of mRNA fragments that are associated with translating ribosomes [12,21]. By comparing the footprint data to measures of mRNA abundance gathered in parallel, we determined translation-specific influences on gene expression. In what follows, we distinguish three quantities. "mRNA abundance" quantifies RNA fragments from the polyadenylated transcriptome, irrespective of whether these molecules are translated. We denote as "footprint abundance" the number of RNA fragments bound by ribosomes, which is a measure of the total protein production for the given gene [12]. Finally, we refer to the ratio of footprint abundance to mRNA abundance as "translational efficiency" (TE) [12]). TE measures the extent to which the mRNA molecules of a given gene are translated.

We found that the differences in footprint abundance between BY and RM were highly correlated with the differences in mRNA abundance, both when comparing the parents and for ASE in the hybrid. Against this largely concordant backdrop, there were a small number of genes with evidence for strong translation-specific genetic effects on their expression, and hundreds of genes with more modest effects. Further, we detected a tendency for translation to more often reinforce (rather than buffer) differences in mRNA levels, such that differences in protein synthesis often exceed differences in mRNA abundance. We found a similar pattern in a reanalysis of published data from studies that compared two yeast species [22,23] and had reported widespread buffering of mRNA differences by translational processes. Overall, genetic effects at the level of translation primarily result in subtle modulation of mRNA differences.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESULTS

Differences in translation between two yeast strains

We used ribosome profiling and mRNA sequencing to compare genome-wide patterns of translation in protein coding regions between the BY and the RM yeast strains. Alignment statistics are presented in Supplementary Table S1 and discussed in Supplementary Note 1. There was excellent agreement between our measures in BY and those obtained by re-aligning the reads from a published yeast ribosome profiling dataset [12] (Supplementary Figure S1). This agreement is in spite of different growth media, slightly different strain backgrounds, several minor differences between library protocols, and substantially deeper sequence coverage in the current dataset (Methods).

mRNA abundance across the 6,697 genes annotated as coding or potentially coding in the yeast genome database varied by 3 – 4 orders of magnitude for the central 95% of genes. Footprint abundance spanned 4 – 5 orders of magnitude and was highly correlated with mRNA abundance (Figure 1). TE varied by ~100 fold among the central 95% of genes, in line with previous observations in yeast [12]. We observed that across genes, footprint abundance increased more rapidly than mRNA abundance such that genes with high mRNA abundance tended to have higher TE, while genes with lower mRNA abundance tended to have lower TE (Figure 1). At the extreme end of this distribution were open reading frames (ORFs) categorized as "dubious" in the yeast genome database (Figure 1). Low or absent translation for dubious ORFs is consistent with the definition of ORFs in this category as "unlikely to encode an expressed protein" (www.yeastgenome.org).

mRNA and footprint levels in BY were highly correlated with those in RM (Figure 2A – B). Consequently, while 52% and 43% of genes had significant (G-test, Bonferroni corrected $p < 0.05$) mRNA and footprint differences between the strains, more than 90% of these differences had small magnitudes of less than 2-fold (Table 1 & Figure 2A – B). That the majority of genes shows significant expression differences between BY and RM is in line with published estimates for mRNA levels – for example, 69% of genes were differentially expressed in a microarray based experiment [24].

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To explore differences in translation between strains, we compared the mRNA differences to the footprint differences. The magnitudes of these differences were correlated (rho = 0.69, Figure 3A), and this correlation became stronger when restricting the analyses to genes with a significant mRNA or footprint difference (rho = 0.78, Figure 3B). As expected from this correlation, genes with a significant mRNA difference were highly likely to also have a significant footprint difference (Fisher's exact test (FET): odds ratio = 5, $p < 2.2e-16$) and the direction of differences agreed for 84% (3,157 out of 3,778) of genes with a significant mRNA and / or footprint difference. Thus, a gene that differs significantly in mRNA abundance between BY and RM typically also has a significant footprint difference in the same direction. For most genes, translation carries forward mRNA differences to differences in protein synthesis.

This observation leaves open the possibility that translation may exert subtle, quantitative effects. Indeed, 37% of genes showed significant (Bonferroni corrected $p < 0.05$) differences in TE; i.e., the ratio of footprint levels between strains differed from the respective mRNA ratio. As expected given the concordance between mRNA and footprint differences noted above, most of the TE differences were of small magnitude (Figure 2C & Table 1). Thus, while translation typically does not override mRNA differences between BY and RM, for many genes it subtly alters the degree to which mRNA differences are reflected at the level of protein synthesis.

Allele specific translation in the BY / RM hybrid

To investigate the extent to which translation is influenced by *cis*-acting variants, we gathered ribosomal footprint and mRNA data from the diploid hybrid between BY and RM. Reproducibility of the allele-specific measurements was excellent, as judged by comparing two biological replicates processed at the same time (Supplementary Figure S2). As expected, the number of genes with significant ASE was less than the number of genes with differences between the BY and RM parental strains (Figure 2 & Table 1): we detected significant (Bonferroni corrected $p < 0.05$) ASE for the mRNA levels of 6% of genes and for the footprint levels of 7% of genes. The mRNA estimate is lower than

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

previous estimates for these two yeast isolates (~14% [16] – ~20% [25]) because of the stringent Bonferroni correction we applied here.

Variants that act in *cis* to alter translation should result in footprint ASE that is not a direct reflection of mRNA ASE. By contrast, we found that the magnitudes of mRNA and footprint ASE were highly correlated (rho = 0.80) when considering genes with significant mRNA or footprint ASE (Figure 3E). Consequently, genes with significant mRNA ASE were very likely to also have significant footprint ASE (FET odds ratio = 49, $p < 2.2e-16$) and 90% (197 / 219) of genes with significant mRNA or footprint ASE agreed in the direction of ASE. Thus, allele specific footprint levels mostly reflect allele specific mRNA expression, suggesting that most genes do not carry *cis* acting variants with strong effects on translation.

To search for genes that may carry *cis* acting variants with more subtle effects on translation, we tested for allele-specific TE, i.e. for genes where the ratio of mRNA ASE differs from the ratio of footprint ASE. Significant (Bonferroni corrected $p < 0.05$) allele-specific TE was found for 3% (n = 85) of genes (Table 1). While most of these effects had small magnitude (Figure 2F), 13 genes had significant allele-specific TE greater than 2-fold (Table 2). For three of these genes, allele-specific TE is due to premature translation termination in one strain compared to the other (Table 2). The remaining ten genes appear to carry *cis* acting variants that substantially alter their translation rate, while several dozen additional genes carried variants with more modest effects.

Cis *and* trans *contributions to translation*

The overall contribution of *cis* vs. *trans* acting variants to gene expression variation can be determined by comparing differences in expression between two strains to allele-specific expression in their diploid hybrid [13,16]. For genes that are entirely regulated in *cis*, the parental difference should be completely recapitulated by allele-biased expression in the hybrid. For genes that are entirely regulated in *trans*, expression of the two alleles should be the same in the hybrid, irrespective of the parent difference. Consequently, if all genes in the genome were exclusively affected by *cis*-acting variants the slope of the

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

relationship between allelic differences and parental differences should equal one. If all genes were exclusively regulated in *trans*, the slope should equal zero.

In our data, the slopes of these relationships (calculated using major axis estimation [26]) were 0.37 for mRNA and 0.33 for footprint abundance (Figure 4A). Bootstrapped distributions of the mRNA and footprint slope estimates overlapped substantially (Figure 4C). Thus, the relative contributions of *cis*- and *trans*-acting variants on mRNA abundance are faithfully represented in footprint abundance.

Next, we asked if the effects of previously identified eQTL (i.e., individual loci identified through their effects on mRNA levels) are reflected in our data. We stratified genes according to whether they were influenced by a local (likely *cis*-acting) eQTL, a distant (likely *trans*-acting) eQTL, or by both types of eQTL, using eQTL reported in [24]. Genes with local eQTL had a significantly steeper slope (the 95% confidence intervals from 1,000 bootstraps did not overlap) than genes with distant eQTL (Figure 4B & C). Thus, the effects of known genetic variants are recapitulated in our mRNA data. The effects of known eQTL were also seen in footprint abundance (Figure 4B & C). Together, these analyses suggest that the relative importance of *cis* vs. *trans* acting genetic variation on footprint abundance is largely similar to that on mRNA abundance.

Evidence for translational reinforcement of mRNA differences

Translation can interact with differences in mRNA abundances in four ways. A given mRNA difference can be "reinforced", resulting in a larger difference in footprint abundance. Conversely, translation can "buffer" an mRNA difference such that the difference in footprint abundance is reduced or even absent. Translation may also invert an mRNA difference, such that the resulting footprint difference is in the opposite direction of the mRNA difference. Finally, when translation faithfully carries forward the mRNA difference, differences in footprint abundance are a direct reflection of the difference in mRNA abundance. To examine the relative frequencies of these scenarios, we performed two sets of analyses.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

First, we partitioned the genes with significant TE differences between BY and RM ("TE genes", 36% of genes) based on the direction and significance of the respective mRNA and footprint differences (Table 3 & Figure 3C). For 8% of the TE genes, neither the mRNA nor the footprint difference was significant, providing little information on the relative importance of buffering or reinforcing interactions. Genes with opposite direction of the differences were relatively uncommon: in the parent data, 6% of the TE genes showed a significant mRNA difference and a significant footprint difference with opposite sign. Fifty-four percent of the TE genes showed a pattern consistent with reinforcement in which a significant footprint difference was larger than the corresponding mRNA difference (including cases where the mRNA difference was not significant). Conversely, 31% of the TE genes showed a pattern consistent with buffering, such that a significant mRNA difference corresponded to a smaller or non-significant footprint difference. The difference between genes showing reinforcing vs. buffering interactions was significant (Chi squared test with one degree of freedom, χ^2 = 138, $p < 2e\text{-}16$), suggesting that translation more often leads to an increase, rather than a reduction, in the magnitude of footprint differences compared to mRNA differences between the BY and RM strain.

In the hybrid data, mRNA differences were associated with larger footprint differences at 29% of the 85 TE genes (Table 3 & Figure 3F) and associated with smaller footprint differences at 33% of TE genes. The difference between these two groups was not significant (χ^2 = 0.2, p = 0.7). Therefore, this analysis provides no evidence that allele-specific mRNA differences are preferentially reinforced or buffered by translational effects.

Next, we considered all expressed genes (irrespective of whether they had significant mRNA, footprint, or TE differences) and calculated the slope of the relationship between footprint differences on mRNA differences. Because we are interested in measuring the relationship between these quantities rather than predicting the one from the other, we calculated the slopes using major axis estimation rather than linear regression [26]. For both parental (slope = 1.16) and hybrid data (slope = 1.52),

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these slopes were greater than one (Figure 3A, D and Figure 5A, D), as expected if reinforcing interactions predominate.

To test if these observed slopes are significantly steeper or flatter than expected by chance, we constructed 1,000 randomized null data sets that have no preference for reinforcing or buffering translational effects (Methods, Supplementary Figure S3). In each of these randomized datasets, we calculated the slope between footprint differences and mRNA differences. The randomized slopes were on average close to one. In both the parental and the hybrid data, the slope observed in the real data was significantly steeper than those seen in the randomized data ($p < 0.001$, Figure 5A & D). Thus, across all genes, there was an excess of reinforcing interactions in which translation enhances an existing mRNA difference.

To test if this translational reinforcement might manifest itself at the protein level, we examined published estimates of strain differences between BY and RM for proteins [4] and mRNAs [24] of 589 genes. These differences were correlated with each other (Spearman rank correlation rho = 0.48, $p < 2.2e-16$). Among the 318 genes for which the protein and mRNA differences were in the same direction, the magnitudes of the protein differences were slightly but significantly larger than those of the mRNA differences (median protein difference = 1.31 fold, median mRNA difference = 1.25 fold; Wilcoxon Rank Test $p = 0.008$). Published protein and mRNA differences between BY and RM are therefore consistent with an overall tendency for translation to reinforce mRNA differences.

No evidence of translational buffering of mRNA differences between two yeast species

Recent work comparing the two yeast species *S. cerevisiae* (two isolates of which we study here) and *Saccharomyces paradoxus* reported widespread translational buffering of mRNA differences, such that a higher mRNA level in one species was typically accompanied by a difference in translation that counteracted this mRNA difference [22,23]. In both studies, this inference was primarily based on the observation that TE differences showed a highly significant negative correlation with mRNA differences.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

However, this inference is problematic because the TE difference is the ratio of the footprint difference and the mRNA difference. It has long been noted [27,28] that comparisons between ratios and their components induce "spurious" correlations. An illustration of this effect is provided in Supplementary Figure S4. A negative correlation between TE differences and mRNA differences is therefore not sufficient to infer translational buffering.

The published inferences of buffering were further based on the observation that the slopes of the regression of footprint differences on mRNA differences were less than one. Intuitively, this might be considered evidence for translational buffering because for any given mRNA difference, the associated footprint difference is predicted to be of smaller magnitude. However, this intuition is misleading due to regression to the mean ([29] p. 58). In the presence of measurement noise, and when two observations are on similar scales (as is the case for mRNA and footprint differences), regression slopes are less than one. Therefore, this observation is also not sufficient to infer buffering.

To test if translation indeed buffers mRNA differences between yeast species, we analyzed the data from [22] and [23]. We first calculated slopes using major axis estimation [26]. In all four cases (two species comparisons and two hybrid comparisons), the resulting slopes were steeper than one (Figure 5 B, C, E, F). We then constructed randomized datasets following the same procedure as for our own data. None of the four datasets showed an excess of buffering compared to the randomized distributions. Instead, the observed slopes were steeper than expected by chance in all four datasets ($p = 0.01$ for the interspecies comparison in [23], $p < 0.001$ otherwise, Figure 5). Thus, in our re-analyses, the published between-species data showed more reinforcing, rather than buffering, effects of translation, consistent with our within-species results.

mRNA and footprint levels are correlated with genetic influences on protein levels

Next, we asked if the mRNA and footprint differences correspond to differences in protein abundance that are due to individual genetic loci. We recently used a bulk-segregant approach to map *trans*–acting pQTL in BY and RM [7]. The high statistical

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

power of that approach resulted in the identification of multiple pQTL for many of the analyzed genes. Of the 160 genes in our earlier study, 117 could be analyzed in the present parent and hybrid datasets. For each of these 117 genes, we summed the allele frequency differences at the pQTL, a measure related to the genetic effects of the pQTL (Methods). These summed measures provide a rough expectation of protein level differences between BY and RM that are due to *trans*-acting variation.

As expected given that all the pQTL considered here act in *trans*, there was no correlation between the predicted protein differences and ASE in the hybrid data (Figure 6C & D). By contrast, we found that the summed pQTL effects correlated significantly with differences in mRNA abundance between BY and RM (Spearman rank correlation rho = 0.38, p = 1e-6; Figure 6A). Thus, the aggregate effects of loci that were detected through their effects on protein levels are reflected in mRNA differences between strains. The summed pQTL effects were also correlated with strain differences in footprint abundance (rho = 0.47, p = 9e-10; Figure 6B). This correlation with footprint differences was significantly higher than that with mRNA differences (only 29 of 1,000 bootstrapped datasets showed a larger mRNA than footprint correlation). These observations are consistent with the hypothesis that most pQTL reflect genetic influences on mRNA levels, and that on average, translation amplifies these mRNA differences to result in larger differences in protein levels.

Effects of nonsense and frameshift mutations on translation

The single base pair resolution of ribosome-profiling data permits detailed examination of footprint abundance along the length of a gene, and of how these patterns are affected by genetic variation. In particular, variants that create or disrupt a stop codon are of interest due to the potentially large phenotypic consequences. In addition, insertions and deletions (indels) that lead to a shift in the reading frame of a coding gene will usually result in premature translation termination or, more rarely, in translation proceeding beyond the original stop codon. All these types of variants should lead to detectable differences in the pattern of footprint coverage.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Among our set of high quality coding SNPs in 3,376 genes that had a status of "verified" according to the SGD database, we identified 18 sites where RM relative to BY has gained a premature stop codon and 10 sites where RM has lost the annotated stop codon (Supplementary Data S1). In addition, we catalogued 32 short indels predicted to lead to a frameshift. We visually examined the footprints patterns in BY and RM at these sites.

Of the 18 gained stops in RM, three were in genes that were not expressed and could not be analyzed. Of the 15 sites in expressed genes, nine led to clearly visible premature termination (Figure 7A for an example). Four putative premature stop codons were located in a part of the ORF that, while annotated as part of the coding sequence, is in fact upstream of the region we found to be translated in both BY and RM (Figure 7B). These four sites therefore do not affect the protein, but reflect errors in gene annotation. The remaining two sites were situated in the translated part of the coding sequence, but did not lead to a visible reduction in translation. Closer inspection of these two SNPs showed that both of them are part of multi-base substitutions that together lead to an amino acid substitution instead of a nonsense mutation. Thus, only 60% (9 / 15) of our list of predicted nonsense mutations in expressed genes had detectable effects on protein sequence. We note that six of these truncating mutations were close to the 3' end of the coding sequence, where they may be less likely to severely disrupt protein function [30].

Of the ten sites where a BY stop codon was absent in RM, three resided in genes with no or very low expression. Four did lead to visible ribosomal readthrough, and two of these sites are in fact known instances of difference in primary protein structure between BY and RM. For example, the gene *NIT1* and the gene annotated to lie immediately downstream (YIL165C) form a single ORF in other yeast species [31] and other *S. cerevisae* strains [32]. The two genes are annotated as two separate ORFs because the yeast genome annotation is primarily based on BY, which carries a premature stop codon inside the ORF (Figure 7C). The remaining three lost stop codons did not visibly result in translational readthrough. Closer examination of the sequence context reveals that these codons are immediately followed by a secondary stop codon that compensates for the lost stop codon [33].

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We made similar observations for the 32 putative frame shifting indels: ten were in genes with low expression, five were in untranslated regions erroneously annotated as coding, three were in repetitive regions and may be due to alignment errors, six were close to the end of the ORF, one was downstream of a premature stop and therefore of no consequence itself, and only seven led to visible early termination or extension of the frame-shifted protein.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCUSSION

We used ribosome profiling [12] to explore how genetic differences between the two yeast strains BY and RM influence mRNA abundance and translation. We found that most genes with significant differences in mRNA levels had footprint differences in the same direction. Thus, translation typically carries forward genetic influences on mRNA levels into differences in protein synthesis. While we did detect hundreds of genes that showed evidence for genetic effects on translation, most of these effects subtly modulate rather than override mRNA differences. BY and RM do not appear to carry genetic variants that in aggregate strongly affect translation.

We made similar observations in the hybrid between BY and RM. Significant ASE in mRNA was highly correlated with ASE in footprints. Therefore, with a few exceptions (e.g. those listed in Table 2), most genes do not carry *cis*-acting variants that have large, specific influences on translation.

By comparing the parental differences to ASE in the hybrid [13], we found that the relative contribution of *cis*- vs. *trans*-acting variants on footprint levels was similar to that on mRNA levels. Further, individual local and distant eQTL that had earlier been identified based on their effects on mRNA levels [24] influence the *cis* vs. *trans* contriution in both the mRNA and footprint data presented here. These eQTL therefore are carried forward to translation and would be expected to also affect protein levels.

Analyses of a mass spectrometry dataset have reported substantial discrepancies between genetic influences on mRNA and protein differences between BY and RM [4,5]. Our ribosome profiling data provides little evidence that genetic effects on translation might be responsible for these discrepancies. This observation is in line with recent pQTL studies in yeast that leveraged improvements in protein measurements and experimental design [7,8] and found that eQTL and pQTL are not as discordant as reported previously. To the extent that the remaining discrepancies between eQTL and pQTL are real (as opposed to, for example, due to experimental variation [34]), our results here suggest that they are more likely caused by genetic influences on protein degradation rather than on translation.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We made multiple observations supporting a tendency for mRNA differences to be reinforced, rather than buffered, at the footprint level, in agreement with recent findings for allele-specific translation in *Candida albicans* (Muzzey, Sherlock & Weissman, in revision). This observation implies that a typical eQTL with effects on mRNA levels should not only also influence protein levels (and thus also be a pQTL), but that its effects on protein levels should be larger than those on mRNA levels. While this prediction remains to be rigorously tested, it is tentatively supported by two observations. First, published protein strain differences between BY and RM (which are caused by *cis*- and / or *trans*-acting QTL) were on average larger than mRNA differences. Second, footprint differences correlated slightly better with the summed effects of multiple pQTL [7] than did mRNA differences. Thus, these pQTL may reflect genetic effects that cause differences in mRNA levels that are reinforced by translation to result in larger protein differences.

Two recent papers examined the evolution of mRNA and footprint levels between the yeast species *S. cerevisiae* and *S. paradoxus*. Both studies reported that mRNA differences are more often buffered than reinforced by translation. This is in apparent contrast to our observation that among strains of *S. cerevisiae*, translation tends to reinforce mRNA differences. However, we showed that the reported inferences of widespread translational buffering between species are likely explained by negative correlations that arise when comparing a ratio (the TE difference) to the ratio's denominator (the mRNA differences). In our reanalysis of the inter-species data, we found no evidence for widespread translational buffering. Instead, translation more often appeared to reinforce mRNA differences, similar to what we observe for genetic variation within *S. cerevisiae*. This tendency is at odds with observations that protein levels are more highly conserved than mRNA levels across highly divergent taxa [35,36] and more closely related species [37,38]. Higher conservation of protein levels over longer evolutionary time may primarily be caused downstream of translational regulation, for example at the level of protein stability and degradation. More work, including improved methods to quantify protein abundance and degradation, is needed to resolve these fundamental questions.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A further open question is what causes the observed tendency of translation to reinforce mRNA differences. Gene-specific coordinated tuning of mRNA levels and translation by genetic variation seems *a priori* unlikely between the BY and the RM strains because it would require, for each gene, at least two variants to act in concert: one altering mRNA levels and another altering translation.

A more parsimonious explanation may be related to the observation that translation can stabilize mRNA molecules by protecting them from degradation [39,40]. Therefore, a higher translation rate *per se* can result in higher mRNA levels at steady state. A sequence variant that increases TE of a given gene (in *cis* or in *trans*) could then not only result in higher footprint levels but also increase mRNA levels, even if the variant has no effect on transcription. Depending on the strength of coupling between TE and the protection from mRNA degradation, such a variant can produce a larger increase in footprints than in mRNA levels, consistent with our observations. If this hypothesis holds, translational "reinforcement" would then not begin with an mRNA difference that is modulated at the level of translation, but rather arise due to genetic influences on translation that feed back to mRNA levels. One attraction of this hypothesis is that it requires only a single sequence change to explain the observed coordination between mRNA and footprint differences. An interesting correlate of the hypothesis is that some eQTL might in fact be secondary consequences of genetic effects on translation that then feed back to mRNA levels. Careful study of the dynamics of translation (e.g. [41]) will be needed to further address this question.

Finally, our analyses of nonsense and frameshift polymorphisms showed that these variants indeed result in detectable differences in translation. However, the results serve as a reminder to exercise caution when interpreting the potential functional impact of variants identified in next generation sequencing datasets, especially for variants with putative large effects [30]. Sequence context (e.g. secondary stop codons downstream of a lost stop [33]) and multi-base substitutions can obscure the true consequences of a variant called from a high-throughput pipeline when considered in isolation. Further, even in an extremely well annotated genome such as that of *S. cerevisiae*, errors in gene annotation can generate the illusion of severe differences in protein sequence between

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

strains when in fact the corresponding variants reside outside of coding regions. Our list of variants between BY and RM with validated effects on translation (Supplementary Data S1) can be useful to assess the consequences of genetic differences between these yeast strains.

Molecular phenotypes such as mRNA and protein levels (as well as others [42,43]) provide crucial intermediates for connecting DNA sequence variation to organismal phenotypes. New measurement technologies will allow an increasingly fine-grained view of the mechanistic connections between the levels of molecular traits and illuminate how genetic variation shapes organisms.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MATERIALS AND METHODS

Yeast strains

We studied the same strains as in Bloom *et al.* [44]. The common laboratory BY strain we used had mating type *MAT**a***. The RM strain was originally isolated from a vineyard. Our RM strain had genotype *MAT*α *ho*Δ*::*hphMX4 *flo8*Δ::natMX4 *AMN1-BY*. Both strains were prototrophic, *i.e.* they did not carry any engineered deletions of metabolic genes. These deletions are commonly used as genetic markers that can have strong effects on gene expression [1]. The haploid parental strains were crossed to generate the diploid hybrid. BY and RM differ in cycloheximide resistance at a dose several orders of magnitude lower than those used in the ribosome profiling protocol [12,44]. To confirm that the parents and the hybrid were equally sensitive to the high cycloheximide dose used here to block translation, we attempted to grow them at 30°C in triplicates in liquid yeast nitrogen base (YNB) medium with a range of cycloheximide concentrations centered on the dose used in the ribosome profiling protocol. While growth was normal in negative controls without cycloheximide, there was no growth within 24 hours in any of the cycloheximide doses tested (data not shown).

Ribosome profiling and sequencing

Libraries for RNA-seq and ribosome profiling were prepared as described in [12], with the following exceptions: (1) cells were cultured in YNB, (2) the reverse-transcription step was primed by ligating miRNA Cloning Linker 1 (IDT) onto the RNA fragments, and (3) highly abundant rRNA species were hybridized to biotinylated oligos and subtracted using streptavidin-coated DynaBeads (Invitrogen) as in [45]. Deep sequencing was performed on the Illumina HiSeq 2000 platform. Raw reads are available in the NCBI Gene Expression Omnibus under accession GSE55400.

SNP set for allele specific quantification

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We employed a set of filters to ensure unbiased estimates of ASE. We used the program BWA [46] to align high coverage (> 50X) 94 bp paired-end whole genome Illumina sequencing data from the BY and the RM strains used in this study [44] to the reference yeast genome version sacCer3 downloaded from the UCSC genome browser (http://genome.ucsc.edu). We used a custom python script kindly provided by Martin Kircher to remove PCR duplicates. Samtools [47] was used to extract a preliminary set of SNPs with variant quality score > 30 and with an estimated alternative allele frequency of 1 ("AF1=1" flag in the vcf file). Next, we retained only biallelic SNPs where our RM strain carries an alternative allele and our BY strain carries the genome reference allele. There were 43,154 SNPs in this initial set.

We sought to restrict this set to those SNPs where short sequencing reads (such as those obtained in ribosome profiling) can be aligned to unique positions in both the BY and the RM reference genome. For each SNP, we extracted the 30 bp up- and downstream sequence from the BY genome reference (sacCer3), from both the plus and the minus strand. The SNP allele itself was set to the RM allele. The resulting 61 bp sequences were aligned to the RM reference genome downloaded from the Broad Institute (http://www.broadinstitute.org/annotation/genome/saccharomyces_cerevisiae.3/Info.html) using BWA [46]. We removed SNPs whose flanking sequences mapped to more than one position in the RM genome as well as SNPs where multiple SNPs mapped to the same position in the RM genome. The number of SNPs after these filters was 38,706.

Next, we sought to remove SNPs with alignment biases towards one or the other reference genome by examining the alignment behavior of publicly available DNA sequence data obtained from a BY / RM hybrid [48]. Any allelic bias seen in hybrid DNA sequences necessarily is of technical origin and indicates problematic SNPs. We trimmed the hybrid DNA reads to 30 bp single end, aligned them to both the BY and RM reference genomes and counted the number of reads that overlapped the BY or RM reference alleles at each SNP, exactly as described below for our mRNA and footprint reads. To identify SNPs with allelic bias beyond that expected by chance, we simulated an unbiased dataset as follows. For each SNP, we generated allele counts assuming a

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

binomial distribution with p = 0.5, and at a depth of coverage drawn from the observed data. We determined criteria for SNP exclusion based on visual comparison of the observed hybrid DNA dataset to the simulated unbiased data. We removed SNPs with very high (> 100 fold) and very low (< 30 fold) coverage, as well as any remaining SNPs with allelic bias towards either genome more extreme than 0.3. After these filtering steps, 36,089 SNPs remained.

We noted a population of SNPs with hybrid DNA allelic ratio centered at ~1/3, i.e. a 2:1 bias towards the RM genome. Further inspection revealed that these SNPs all resided in regions where DNA sequencing coverage in our RM parent was twice as high as that in our BY parent. Nearly all of these regions were situated at chromosome ends and likely reflect segmental duplications of these distal regions in the RM strain compared to the BY reference genome. These regions extended for several kb and contained annotated protein coding genes, in line with the recent observation that subtelomeric regions contain large structural variants that segregate among wild yeast [49]. We visually examined the coverage across our BY and RM parent DNA sequences and excluded any regions with evidence for segmental duplications in the RM but not the BY parent. This removed 821 SNPs, for a remaining set of 35,268.

Finally, because we are interested in quantifying expression of protein coding genes, we retained only the 23,412 SNPs in ORFs annotated in the SGD database (www.yeastgenome.org, accessed on 06 / 28 / 2013). SNPs in ORFs annotated as overlapping on the same strand were removed. However, because the mRNA and footprint data are strand-specific, we were able to retain 395 SNPs that overlap ORFs on different strands for a total of 23,807 quantifiable positions in 4,462 ORFs (Supplementary Data S2).

Read processing and alignments

Because the reference yeast genome is based on a strain with the BY background, sequence differences between the reference and RM make read alignments from an RM sample more difficult, especially with short reads such as the ~32 base pair (bp)

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ribosomal footprint fragments. To counter this problem, we implemented a computational pipeline that uses "personalized" genome references for the BY and the RM strain to allow unbiased read mapping.

Prior to mapping, we removed sequences corresponding to the Illumina adapter sequence (CTGTAGGCACCATCAAT) opposite the sequencing priming site and discarded all reads that did not contain these adapter sequences. We also removed the first base from each read as these often corresponded to adenosines introduced during ligation in the library preparation protocol.

The trimmed reads were mapped using BWA [46] as follows (see Supplementary Table 1 for alignment statistics). For the comparisons between the BY and RM strain, reads can be considered irrespective of whether they cover a SNP or not. Reads from the BY strain were mapped to the BY reference genome (version sacCer3). Reads from the RM strain were mapped to a modified version of the BY reference where the 43,154 SNPs between BY and RM as described the section above were set to the RM allele. The rationale for using this strategy was to maximize the number of RM reads that can be mapped to the BY reference without penalizing reads that contain a sequence difference between BY and RM, while still being able to directly use the BY gene annotations. We counted only uniquely mapping reads on the correct strand in genes.

For the ASE analyses, we are only interested in reads that span a SNP between the BY and RM strains. We noted that the short reads produced in ribosome profiling are heavily biased against mapping RM reads to the BY reference (not shown). We therefore mapped all reads to both the BY reference and the RM reference available from the Broad Institute. We considered only reads that mapped to one of these two reference sequences uniquely and without mismatch. This strategy guarantees that reads that span a sequence difference between BY and RM can be unambiguously assigned to the parental chromosome they originated from. At each of the 23,807 high quality coding SNPs (s. section above), we counted the number of reads that mapped to the correct strand of the BY or the RM genome. When a read overlapped multiple closely linked SNPs, it was randomly counted towards one of them. Because we excluded reads with mismatches, our strategy excludes all reads with sequencing errors. For comparison of ASE in the hybrid

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to differences between the parent strains, we re-mapped the BY and RM parent reads and quantified allele-specific expression as described in this section.

Quantification of mRNA and footprint abundance

For determining the genomic source of reads in the libraries (ORFs, UTRs, ncRNAs, etc.) as well as for the comparison between the parent strains reported in the main text, we used htseq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html) and annotations extracted from SGD (www.yeastgenome.org). We analyzed the 5,964 genes where none of the four parental samples (mRNA and footprints for BY and RM, respectively) had a count of zero.

For the analyses of allele specific expression, we added the allele counts for all SNPs in a gene. For the hybrid, we summed the counts from the two replicates. Unless otherwise noted, we excluded genes where any of the BY parent, RM parent or the summed hybrid samples had a count of zero in either mRNA or footprint data. This resulting set comprised quantifications for 3,693 ORFs. While all statistical analyses were performed directly on count data (s. below), the figures show gene abundance as log10-transformed fractions of total counts for the given sample. Translation efficiency (TE) for a gene was calculated as the difference between the log10-transformed mRNA fraction and the log10-transformed footprint fraction. All quantifications, both for whole ORF and SNPs, are available in Supplementary Data S2.

Statistical analyses

All statistical analyses were performed in the R programming language (www.r-project.org). Unless stated otherwise, we calculated slopes using major axis estimation [16,26] as implemented in [50]. Sequencing-based expression data takes the form of discrete counts rather than continuous values. We therefore used count-based G-tests to test for differential expression. Count-based tests have higher power when the absolute number of counts is high. Our read coverage was different between different samples and

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between different data types (specifically, the parental footprint libraries had 30% - 70% more reads than the parental mRNA libraries). We sought to minimize these differences in total read counts by downscaling all samples to the sample with the lowest total count. We calculated the counts y_{down} that would be expected for a given gene i with observed counts y_{obs} if the total read count was T_{down} instead of the actual counts T_{obs}:

$$y_{down} = (y_{obs} / T_{obs})\ T_{down}$$

The resulting values y_{down} were rounded to the nearest integer.

For each gene, we compared the downscaled BY and RM counts (for either mRNA or footprints) to the counts that would be expected under no gene expression difference. For a given gene i, G-tests were thus performed on 2x2 tables of the form:

	Observed	Expected
BY	$y_{i,BY}$	$(y_{i,BY} + y_{i,RM}) / 2$
RM	$y_{i,RM}$	$(y_{i,BY} + y_{i,RM}) / 2$

where $y_{i,strain}$ is the number of downscaled counts for gene i in *strain* (BY or RM). The test G-test p-values were corrected for multiple tests using Bonferroni correction. To estimate π_1, we used the R package qvalue [51] with the "bootstrap" option for estimating π_0.

To test for differential TE, we tested if the difference between (downscaled) footprint counts differed from that between mRNA counts. G-tests were performed on 2x2 tables of the form:

	Footprints	mRNA
BY	$y_{i,\ BY\ footprints}$	$y_{i,\ BY\ mRNA}$
RM	$y_{i,\ RM\ footprints}$	$y_{i,\ RM\ mRNA}$

All analyses, including downscaled counts, p-values and fold changes, are available in Supplementary Data S2.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Randomization test for buffering vs. reinforcing translational effects

Our goal was to generate datasets that retain all features of the real data but do not contain an excess of reinforcing or buffering effects of translation (Supplementary Figure S3). For each gene, we randomly switched the direction of the observed TE difference towards buffering or reinforcement. To do this, we summed the mRNA counts from BY and from RM (or from the two alleles in the hybrid) to obtain the "summed mRNA counts" for that gene. We equivalently obtained "summed footprint counts". We then divided up the summed mRNA counts either according to the observed mRNA fold change or according to the observed footprint fold change. The summed footprint counts were then divided up according to the remaining fold change (footprint in the first case, mRNA in the second). In the first case, the observation for that gene remains unchanged compared to the real data. In the second case, the mRNA and footprint fold changes (and therefore the TE difference) are inverted. Thus, in these randomized datasets most aspects of the data (expression levels as well as the magnitudes of mRNA, footprint and TE fold changes) are retained, and only the direction of the TE fold change is randomized. In each of 1,000 datasets generated this way, we calculate the slope of the regression of the footprint differences on the mRNA differences. Spearman's rank correlation between the mRNA and the footprint differences in the randomized data closely resembled those in the real data (not shown). P-values were calculated as the fraction of randomized datasets where the regression slope exceeded that seen in the real data.

Comparison of mRNA and footprint differences to pQTL effects

Recently, we showed that the expression of many proteins is influenced by multiple loci that segregate between the BY and the RM isolates [7]. The Albert *et al.* and the present datasets overlap for 117 proteins when considering only genes that can be analyzed in the hybrid data (i.e., that are expressed and that contain at least one SNP). To generate a rough expectation for the aggregate effect that the multiple pQTL have on a given protein, we added their effects. The pQTL in our earlier study were obtained by

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

comparing allele frequencies in populations of cells with high and low protein expression, so that direct estimates of QTL effects (i.e. the expected magnitude by which protein expression differs between the different alleles) are not available. Instead, we used the observed difference in allele frequency at the pQTL location as a measure of effect size. Note that the locus effects can cancel each other: two pQTL with the same absolute allele frequency difference, but with opposite sign will result in an expected aggregate effect of zero. The summed pQTL effects were compared to mRNA and footprint differences from the present study using nonparametric Spearman rank correlation.

To test whether the footprint differences or the mRNA differences correlated better with the pQTL effects than the mRNA differences, we constructed 1,000 bootstrap datasets. In each of these datasets, we randomly sampled from the 117 genes with replacement and calculated both the mRNA and footprint correlations. We calculated the p-value as the fraction of bootstrap datasets where the mRNA difference / pQTL effect correlation exceeded the footprint difference / pQTL effect correlation.

Analyses of published mRNA and protein levels in the BY and RM strains

Published microarray-based mRNA expression data was obtained from Smith & Kruglyak 2008 [24], and we calculated fold changes as the difference between the mean expression values (reported as log2-transformed values) from growth in glucose. For protein levels, we used data from [4] and calculated the difference between the parental means.

Published footprint data for S. cerevisiae *and* S. paradoxus

For Artieri & Fraser [22], we used the RPKM values provided in their Supplementary Table S1 and rounded them to the nearest integer value. For McManus *et al.* [23] we used the read counts provided in their Supplementary Table S5. For both studies, we added the counts for the two replicates for each sample. We analyzed the respective parent and

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hybrid data separately, downscaled the data as in our analyses and excluded genes where any of the respective four samples had a count of zero. This resulted in 4,306 genes in the parent analyses and 4,267 genes in the hybrid analyses of the Artieri & Fraser data, and 4,864 genes in both the parent and hybrid analyses of the McManus *et al.* data. Fold changes, correlations, slopes were calculated and bootstraps and null models constructed exactly as for our data.

Analyses of nonsense and frameshift mutations

We restricted these analyses to coding genes with a "verified" status according to the SGD database. For nonsense SNPs, we considered only the set of high-quality SNPs we used in our ASE analyses. For indels, we considered indel calls produced by samtools mpileup [47]. Because our sequencing coverage of the BY and RM strains is very deep and often exceeds the default limit of 250 fold coverage, we used mpileup with parameters –d 1000 and –L 1000. The resulting indels were further filtered using the bcftools vcfutils.pl varFilter script to only retain indels with coverage of at least 10-fold, variant quality of at least 30, an estimated allele frequency in the sample of 1 (as the sequenced strain was haploid). Finally, we only considered indels in "verified" genes. To identify nonsense SNPs and frameshift indels, we used the standalone perl version of the Ensembl Variant Effect Predictor tool [52] with Ensembl cache data for the yeast sacCer3 genome build available in Ensembl build 72. To examine the effects on translation of the resulting nonsense and frameshift variants, we generated custom tracks in the bedGraph format for display in the UCSC genome browser [53]. These tracks show the start coordinate of each read. The track plots in Figure 7 were generated using the Gviz R package [54].

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

We are grateful to Joshua Bloom, Sebastian Treusch, J.J. Emerson and members of the Kruglyak lab for comments on the analyses and the manuscript.

References

1. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic Dissection of Transcriptional Regulation in Budding Yeast. Science (New York, NY) 296: 752–755. doi:10.1126/science.1069516.

2. Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nature Reviews Genetics 7: 862–872. doi:10.1038/nrg1964.

3. Lappalainen T, Sammeth M, Friedländer MR, Hoen PAC', Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506–511. doi:doi:10.1038/nature12531.

4. Foss EJ, Foss EJ, Radulovic D, Radulovic D, Shaffer SA, et al. (2007) Genetic basis of proteome variation in yeast. Nature Genetics 39: 1369–1375. doi:10.1038/ng.2007.22.

5. Foss EJ, Radulovic D, Shaffer SA, Goodlett DR, Kruglyak L, et al. (2011) Genetic Variation Shapes Protein Networks Mainly through Non-transcriptional Mechanisms. PLoS Biology 9: e1001144. doi:10.1371/journal.pbio.1001144.s010.

6. Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, et al. (2011) Comparative Analysis of Proteome and Transcriptome Variation in Mouse. PLoS Genetics 7: e1001393. doi:10.1371/journal.pgen.1001393.s015.

7. Albert FW, Treusch S, Shockley AH, Bloom JS, Kruglyak L (2014) Genetics of single-cell protein abundance variation in large yeast populations. Nature 506: 494–497. doi:doi:10.1038/nature12904.

8. Skelly DA, Skelly DA, Merrihew GE, Merrihew GE, Riffle M, et al. (2013) Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Research 23: 1496–1504. doi:10.1101/gr.155762.113.

9. Wu L, Candille SI, Choi Y, Xie D, Jiang L, et al. (2013) Variation and genetic control of protein abundance in humans. Nature 499: 79–82. doi:10.1038/nature12223.

10. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. (2011) Global

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

quantification of mammalian gene expression control. Nature 473: 337–342. doi:doi:10.1038/nature10098.

11. Li JJ, Bickel PJ, Biggin MD (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ 2: e270. doi:10.7717/peerj.270.

12. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science (New York, NY) 324: 218–223. doi:10.1126/science.1168978.

13. Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. Nature 430: 85–88. doi:10.1038/nature02698.

14. Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between Drosophila species. Nature Genetics 40: 346–350. doi:10.1038/ng.77.

15. Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics 10: 221. doi:10.1186/1471-2164-10-221.

16. Emerson JJ, Hsieh L-C, Sung H-M, Wang T-Y, Huang C-J, et al. (2010) Natural selection on cis and trans regulation in yeasts. Genome Research 20: 826–836.

17. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772. doi:doi:10.1038/nature08872.

18. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777. doi:10.1038/nature08903.

19. Sonenberg N, Hinnebusch AG (2009) Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell 136: 731–745. doi:10.1016/j.cell.2009.01.042.

20. Kondrashov N, Pusic A, Stumpf CR, Shimizu K, Hsieh AC, et al. (2011) Ribosome-Mediated Specificity in Hox mRNA Translation and Vertebrate Tissue Patterning. Cell 145: 383–397. doi:10.1016/j.cell.2011.03.028.

21. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc 7: 1534–1550. doi:10.1038/nprot.2012.086.

22. Artieri CG, Fraser HB (2014) Evolution at two levels of gene expression in yeast.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome Research 24: 411–421. doi:10.1101/gr.165522.113.

23. McManus CJ, May GE, Spealman P, Shteyman A (2014) Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. Genome Research 24: 422–430. doi:10.1101/gr.164996.113.

24. Smith EN, Kruglyak L (2008) Gene–Environment Interaction in Yeast Gene Expression. PLoS Biology 6: e83. doi:10.1371/journal.pbio.0060083.st004.

25. Ronald J, Brem RB, Whittle J, Kruglyak L (2005) Local Regulatory Variation in Saccharomyces cerevisiae. PLoS Genetics 1: e25. doi:10.1371/journal.pgen.0010025.

26. Warton DI, Wright IJ, Falster DS, Westoby M (2006) Bivariate line-fitting methods for allometry. Biological Reviews 81: 259. doi:10.1017/S1464793106007007.

27. Pearson K (1897) Mathematical Contributions to the Theory of Evolution.--On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. Proceedings of the Royal Society of London 60: 489–498.

28. Jackson DA, Somers KM (1991) The spectre of "spurious" correlations. Oecologia 86: 147–151. doi:10.1007/BF00317404.

29. Gelman A, Hill J (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

30. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science (New York, NY) 335: 823–828. doi:10.1126/science.1215040.

31. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254. doi:10.1038/nature01644.

32. Godard P, Urrestarazu A, Vissers S, Kontos K, Bontempi G, et al. (2007) Effect of 21 Different Nitrogen Sources on Global Gene Expression in the Yeast Saccharomyces cerevisiae. Molecular and Cellular Biology 27: 3065–3086. doi:10.1128/MCB.01084-06.

33. Williams I, Richardson J, Starkey A, Stansfield I (2004) Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae. Nucleic Acids Research 32: 6605–6616. doi:10.1093/nar/gkh1004.

34. Wallace EWJ, Airoldi EM, Drummond DA (2013) Estimating Selection on Synonymous Codon Usage from Noisy Experimental Data. Mol Biol Evol 30: 1438–1453. doi:10.1093/molbev/mst051.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35. Laurent JM, Vogel C, Kwon T, Craig SA, Boutz DR, et al. (2010) Protein abundances are more conserved than mRNA abundances across diverse taxa. Proteomics 10: 4209–4212. doi:10.1002/pmic.201000327.

36. Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, et al. (2009) Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes. PLoS Biology 7: e1000048. doi:10.1371/journal.pbio.1000048.

37. Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, et al. (2013) Primate Transcript and Protein Expression Levels Evolve Under Compensatory Selection Pressures. Science (New York, NY) 342: 1100–1104. doi:10.1126/science.1242379.

38. Khan Z, Bloom JS, Amini S, Singh M, Perlman DH, et al. (2012) Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. Mol Syst Biol 8. doi:10.1038/msb.2012.34.

39. Parker R (2012) RNA degradation in Saccharomyces cerevisae. Genetics 191: 671–702. doi:10.1534/genetics.111.137265.

40. Lackner DH, Beilharz TH, Marguerat S, Mata J, Watt S, et al. (2007) A Network of Multiple Regulatory Layers Shapes Gene Expression in Fission Yeast. Molecular Cell 26: 145–155. doi:10.1016/j.molcel.2007.03.002.

41. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB (2013) Rate-limiting steps in yeast protein translation. Cell 153: 1589–1601. doi:10.1016/j.cell.2013.05.049.

42. Dermitzakis ET (2012) Cellular genomics for complex traits. Nature Reviews Genetics. doi:10.1038/nrg3115.

43. Gaffney DJ (2013) Global Properties and Functional Complexity of Human Gene Regulatory Variation. PLoS Genetics 9: e1003501. Available: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1003501.

44. Bloom JS, Ehrenreich IM, Loo WT, Lite T-LV, Kruglyak L (2013) Finding the sources of missing heritability in a yeast cross. Nature 494: 234–237. doi:10.1038/nature11867.

45. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, et al. (2012) High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. Science (New York, NY) 335: 552–557. doi:10.1126/science.1215110.

46. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. doi:10.1093/bioinformatics/btp324.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. doi:10.1093/bioinformatics/btp352.

48. 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 Research 21: 1728–1737. doi:10.1101/gr.119784.110.

49. Bergstrom A, Simpson JT, Salinas F, Barre B, Parts L, et al. (2014) A high-definition view of functional genetic variation from natural yeast genomes. Mol Biol Evol. doi:10.1093/molbev/msu037.

50. Warton DI, Duursma RA, Falster DS, Taskinen S (2011) smatr 3- an R package for estimation and inference about allometric lines. Methods in Ecology and Evolution 3: 257–259. doi:10.1111/j.2041-210X.2011.00153.x.

51. Dabney A, Storey JD (2004) Q-value estimation for false discovery rate control. Medicine 344: 539–548.

52. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070. doi:10.1093/bioinformatics/btq330.

53. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC Genome Browser Database. Nucleic Acids Research 31: 51–54. doi:10.1093/nar/gkg129.

54. Hahne F, Durinck S, Ivanek R, Mueller A, Lianoglou S, et al. (n.d.) Gviz: Plotting data and annotation information along genomic coordinates. R package version 1.7.4.

55. Warner JR (1999) The economics of ribosome biosynthesis in yeast. Trends in Biochemical Sciences 24: 437–440. doi:10.1016/S0968-0004(99)01460-7.

56. Venema J, Tollervey D (1999) Ribosome synthesis in Saccharomyces cerevisiae. Annual Review of Genetics 33: 261–311.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tables

Table 1 – Differential expression statistics

Comparison	Reads	Data	Analyzed genes	2-fold	G-test[1]	Intersect
Parent	All	mRNA	5,964	546	3,118	422 (7%)
Parent	All	Footprint	5,964	829	2,566	493 (8%)
Parent	All	TE	5,964	370	2,182	124 (2%)
Parent	SNP	mRNA	3,693	251	475	179 (5%)
Parent	SNP	Footprint	3,693	460	481	205 (6%)
Parent	SNP	TE	3,693	270	253	62 (2%)
Hybrid	SNP	mRNA	3,693	116	152	40 (1%)
Hybrid	SNP	Footprint	3,693	185	180	44 (1%)
Hybrid	SNP	TE	3,693	194	85	13 (0.4%)

[1]Bonferroni corrected $p < 0.05$

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2 – Strong *cis* effects on translation

	TE		mRNA		FP			
Gene	fold change	p	fold change	p	fold change	p	Type	Notes
YBR012C	-3.20	5e-7	-1.10	4e-8	-4.30	2e-9	Reinforce	
YBR107C (*IML3*)	1.23	2e-6	-1.29	1e-13	-0.07	0.8	Remove	
YBR114W (*RAD16*)	1.20	3e-6	-0.81	1e-4	0.39	0.2	NS	
YDL231C (*BRE4*)	-3.07	5e-8	-0.16	0.5	-3.23	3e-6	Introduce	
YDR133C	-2.45	< 2e-16	-1.20	< 2e-16	-3.65	<2e-16	Reinforce	(1)
YEL066W (*HPA3*)	1.59	1e-10	-0.54	0.02	1.05	5e-5	NS	
YIL165C	-3.68	< 2e-16	-0.22	0.04	-3.90	<2e-16	Introduce	(2)
YJR015W	4.92	1e-15	1.17	3e-6	6.09	<2e-16	Reinforce	(3)
YJR072C (*NPA3*)	-1.97	< 2e-16	1.39	< 2e-16	-0.58	0.001	Remove	
YLL007C	1.55	3e-6	-1.31	2e-4	0.25	0.4	NS	
YML048W (*GSF3*)	-1.03	1e-7	1.02	6e-8	-0.01	0.9	Remove	
YMR091C (*NPL6*)	1.08	5e-7	-1.23	1e-11	-0.15	0.5	Remove	
YOR304W (*ISW2*)	1.05	7e-8	-0.84	3e-8	0.22	0.3	Remove	

For consistency with the figures, the fold changes are log2-transformed so that zero indicates no difference and one indicates a two-fold difference. Positive values indicate higher abundance in BY compared to RM. FP: footprints. NS: neither mRNA nor footprint difference was significant. (1) "Dubious" ORF, footprint data shows translated region only partially overlaps with annotation. The TE difference is due to a nonsense SNP in BY that results in early termination compared to RM. (2) "Dubious" ORF immediately downstream of *NIT1*; in RM, these two ORFs form a single, consistently translated ORF (Figure 7C). (3) Putative protein with frameshift in RM that leads to premature termination. Note that "dubious" ORFs were not included in our analyses of nonsense SNPs so that YDR133C and YJR015W were not included in those analyses.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3 – Effects of translation in genes with significant TE

Significant difference	Direction of difference	Magnitude of difference	Parent	Hybrid
mRNA and footprint	same	Footprint > mRNA	680 (31%)	11 (13%)
Footprint only	–	–	512 (23%)	14 (16%)
mRNA and footprint	same	mRNA > footprint	198 (9%)	5 (6%)
mRNA only	–	–	486 (22%)	23 (27%)
mRNA and footprint	opposite	–	122 (6%)	1 (1%)
neither	–	–	184 (8%)	31 (36%)
Sum	–	–	2,182	85

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Tables

Supplementary Table S1 – Sequencing and alignment statistics

Strain	Data type	Raw reads (million)	Parent Strain comparison: unique alignments		Allele-specific analyses: unique & no mismatch	
Reference genome			BY	Edited BY[1]	BY	RM
BY parent	Footprint	189	82	-	74	69
BY parent	mRNA	146	53	-	46	42
RM parent	Footprint	222	-	151	126	129
RM parent	mRNA	129	-	52	43	46
BY/RM diploid 1	Footprint	103	-	-	33	32
BY/RM diploid 1	mRNA	98	-	-	27	26
BY/RM diploid 2	Footprint	108	-	-	44	42
BY/RM diploid 2	mRNA	113	-	-	28	28

[1]A version of the BY reference genome where all known single nucleotide differences to RM were set to the RM allele

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table S2 – Genomic sources of mRNA and footprint reads in the BY parent

	mRNA		Ribosomal footprints	
	unique	repetitive	unique	repetitive
CDS	44.4M (84%)	5.8M (6.6%)	79.2M (97%)	12.2M (12%)
UTRs[1]	8.9M (17%)	66k (0.1%)	4.7M (5.7%)	145k (0.1%)
5'UTR	3.1M (35%)[1]	34k (52%)[1]	3.2M (68%)[1]	69k (48%)[1]
3'UTR	5.9M (66%)[1]	32k (48%)[1]	1.5M (32%)[1]	75k (52%)[1]
rRNA	246k (0.5%)	79M (90%)	950k (1.2%)	88M (85%)
tRNA	73k (0.1%)	466k (0.5%)	228k (0.3%)	1.2M (1.2%)
Other noncoding: snoRNA, snRNA, ncRNA	468k (0.9%)	727	217k (0.3%)	264
Total	53M	88M	82M	103M

Percentages can sum to more than 100 due to overlapping annotations

[1]percent of all UTRs

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

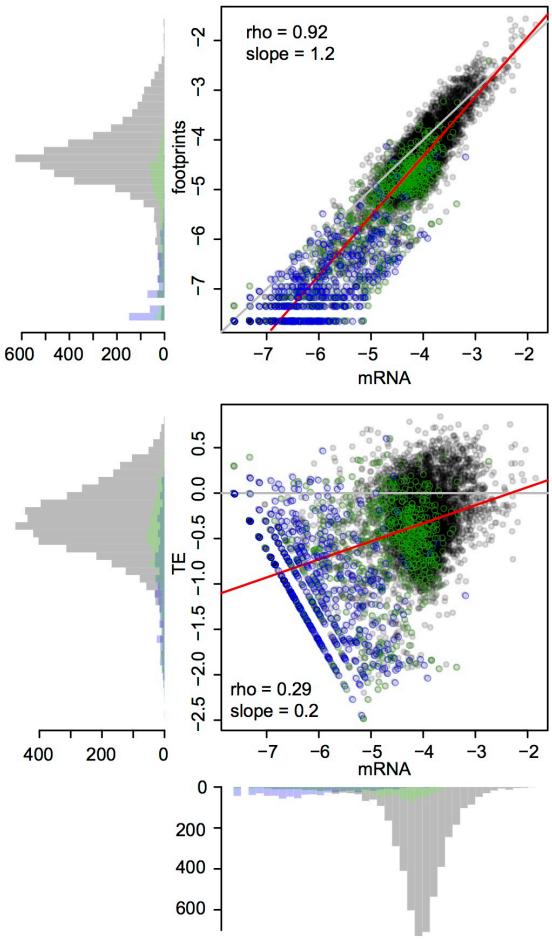


Figure 1 – Global mRNA and footprint abundance

Shown are log10-transformed normalized read counts in the BY strain. Top panel: mRNA vs. footprint abundance. The red line shows the regression of footprint on mRNA abundance. The grey line indicates identity. Bottom panel: TE as a function of mRNA abundance. The grey line denotes identity between footprint and mRNA levels (i.e. log10(TE) = 0). The red line shows the regression of TE on mRNA abundance. Throughout the figure, transparent grey points are "verified" ORFs, green points are "uncharacterized" ORFs and blue points are "dubious" ORFs.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

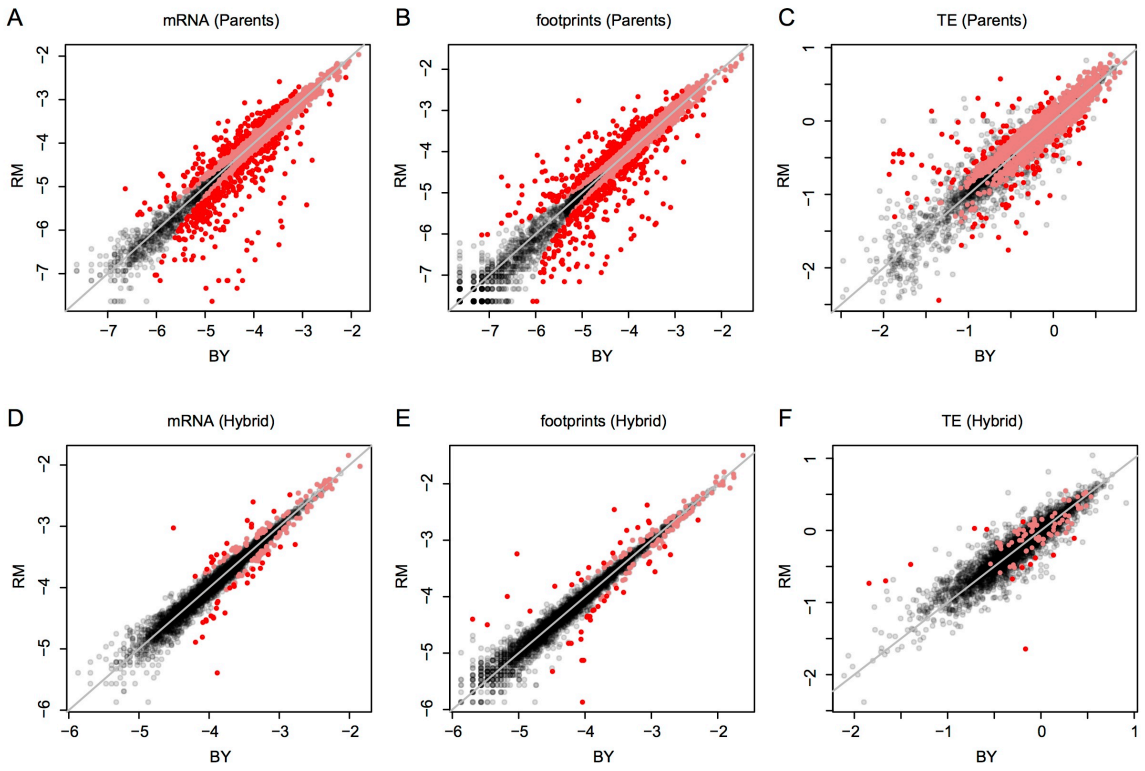


Figure 2 – Expression in BY vs. RM and ASE in the hybrid

Shown are log10-transformed normalized read counts. Grey lines mark identity. Light red points are genes with significant differences ($p <$ Bonferoni cutoff), darker red points are significant genes with a fold change ≥ 2. A-C: parental comparisons, D-F: hybrid ASE

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

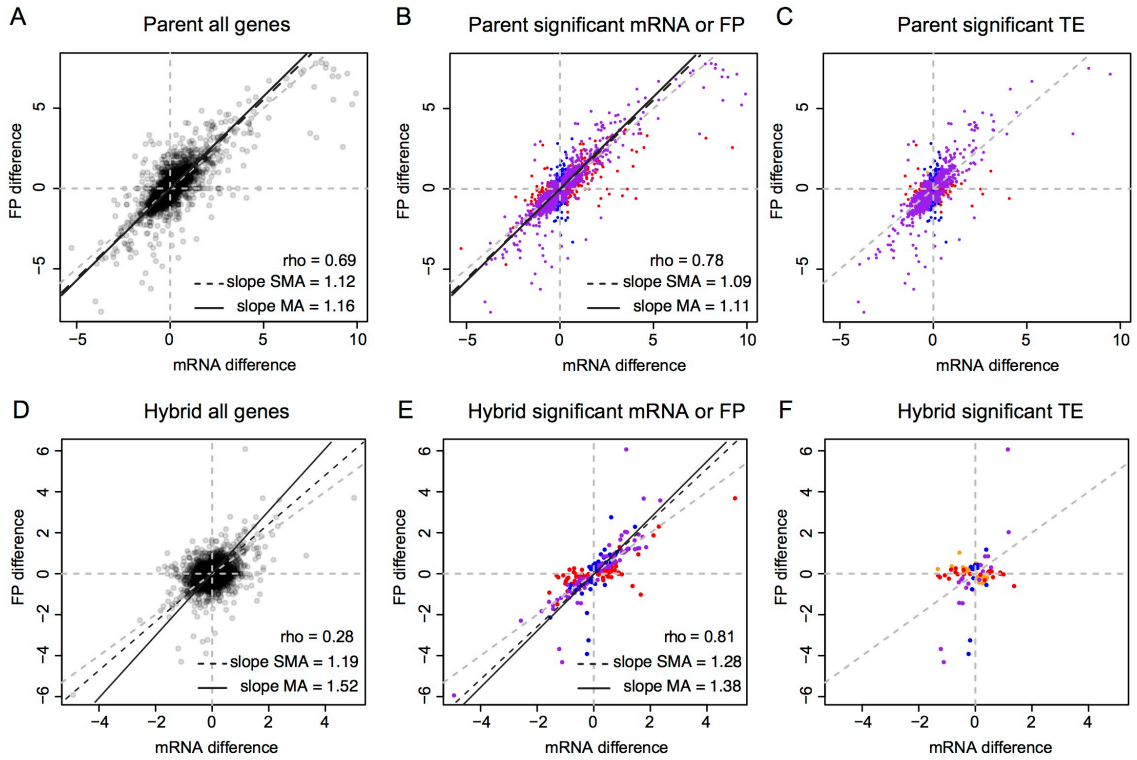


Figure 3 – mRNA vs. footprint differences

Shown are log2-transformed fold changes. A-C: parents, D-F: hybrid ASE. Grey dashed lines are the diagonal. Black lines show the slope of the relationship between footprint differences and mRNA differences for the genes in the respective plot. Slopes were calculated using major axis estimation (MA) and standardized major axis estimation (SMA). Left column: all genes. Middle column: genes with a significant ($p <$ Bonferoni cutoff) mRNA (red), footprint (blue) or both mRNA and footprint (purple) difference. Right column: genes with a significant TE difference. Red: genes with only a significant mRNA difference, blue: genes with only a significant footprint difference, purple: genes with both a significant mRNA and footprint difference, orange: genes with neither a significant mRNA nor a significant footprint difference.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

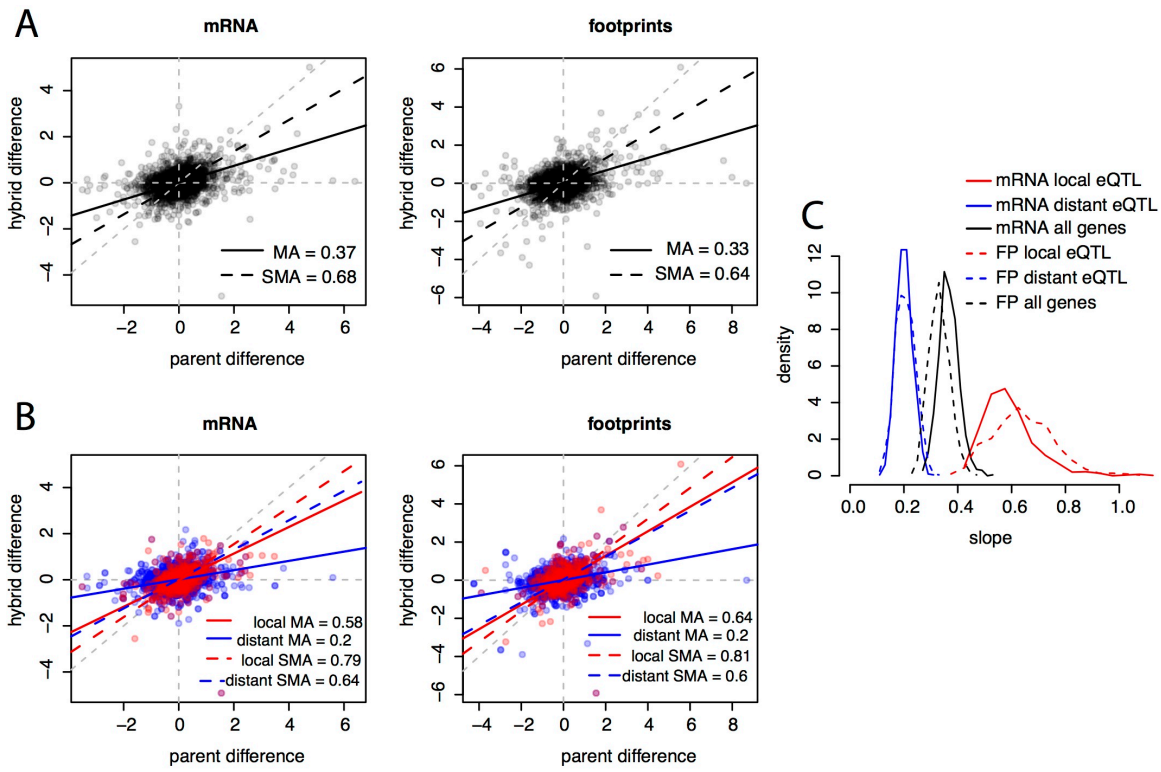


Figure 4 – *Cis* and *trans* effects

A. Parental differences (estimated based on SNP allele counts) on the x-axes, and hybrid differences on the y-axes, for all genes. Black lines show the slope of the relationship between hybrid and parental differences. The legends indicate the values of these slopes. MA: major axis estimate; SMA: standardized major axis estimate. B. as in A), but only for genes with eQTL in [24]. Red: genes with a local but no distant eQTL, blue: genes with a distant but no local eQTL, purple: genes with both a local and a distant eQTL. Colored lines show the respective regressions of hybrid on parental differences. C. bootstrapped distributions of MA slope estimates. Results from SMA were qualitatively similar.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

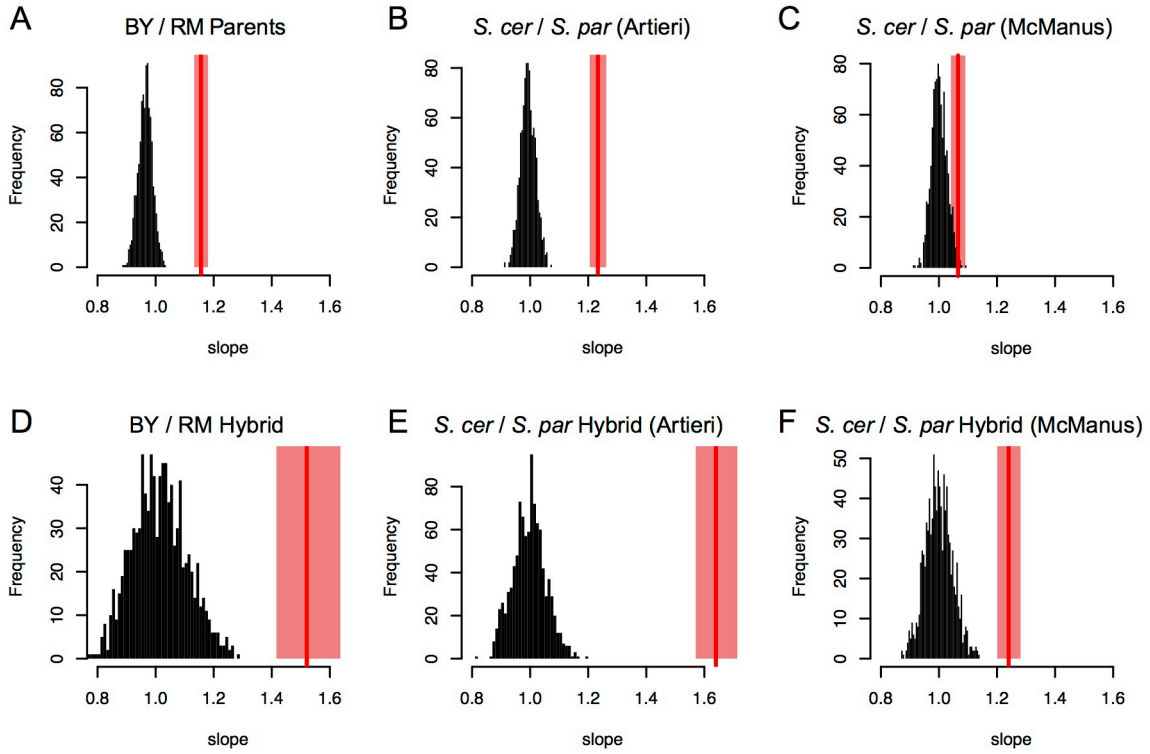


Figure 5 – Randomized and observed slopes of the relationship between footprint differences and mRNA differences

In each plot, the black histogram shows the distribution of slopes obtained by major axis estimation in 1,000 randomized datasets. Standardized major axis estimation yielded similar results. Red lines indicate the slope observed in the real data, and the light red box indicates the corresponding confidence intervals. *S. cer*: *Saccharomyces cerevisiae*. *S. par*: *Saccharomyces paradoxus*. The interspecies data were analyzed from published datasets. Artieri: [22], McManus: [23].

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

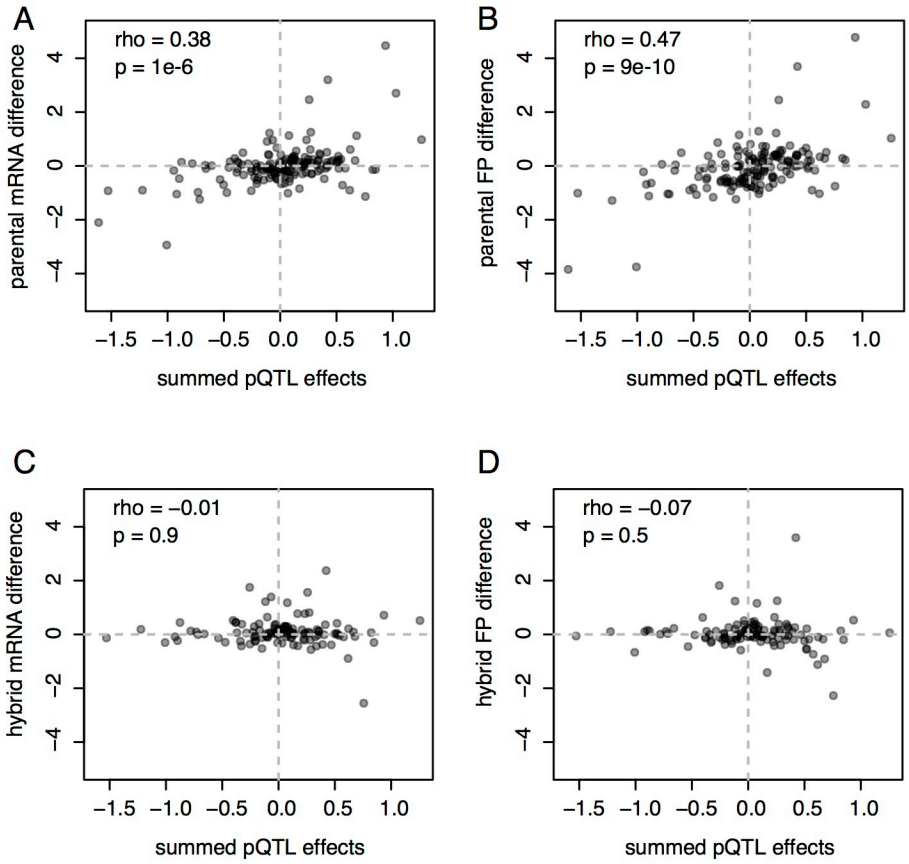


Figure 6 – Comparison of mRNA and footprint differences to pQTL effects

For each gene, the pQTL effects are shown as the sum of allele frequency differences at all pQTL identified by a bulk segregant approach [7].

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

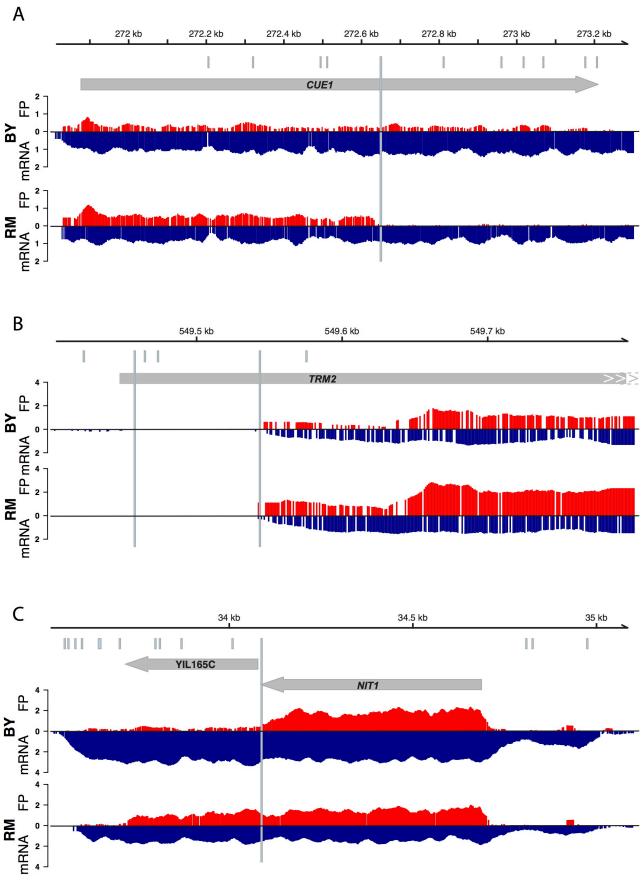
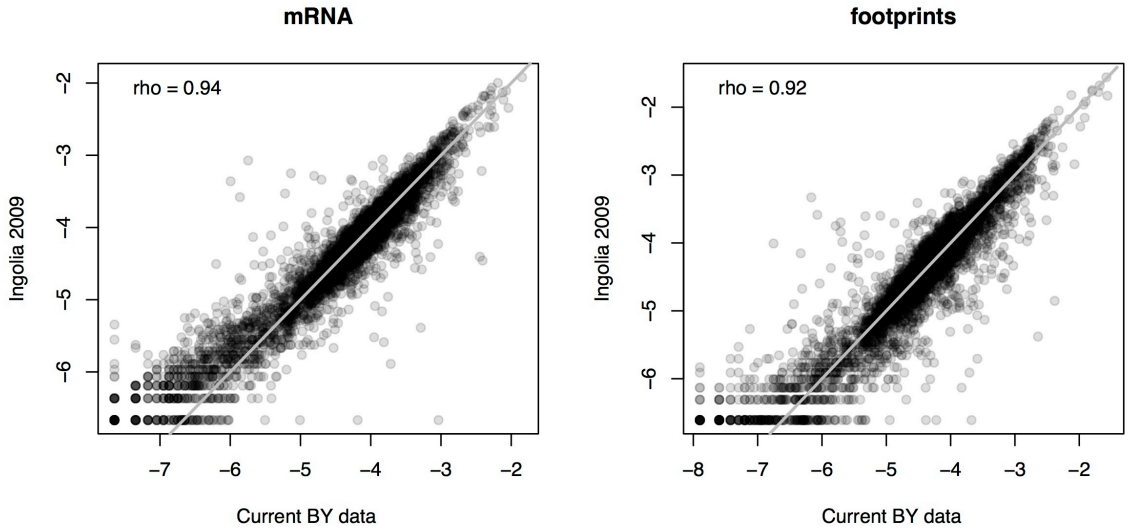


Figure 7 – Examples of patterns of translation at putative premature stop codons

Grey arrows indicate the position and strand of ORFs. Footprints (red) and mRNA (blue, inverted scale) for BY and RM are plotted beneath. The positions of putative premature stop codons in BY or RM are shown as light blue, longer horizontal bars, while all sequence differences between BY and RM are shown as light blue tickmarks above the ORF. The mRNA and footprint densities are shown as log transformed numbers of read starts in 30 bp wide smoothed windows. They are only shown for the strand of the displayed ORFs. A. An example of a premature translation termination in *CUE1* in RM compared to BY. B. Two putative nonsense SNPs in *TRM2* are in fact upstream of the translated and transcribed ORF. C. The gene *NIT1* in BY is the result of a premature termination of a full length ORF that in RM includes the downstream ORF YIL165C.

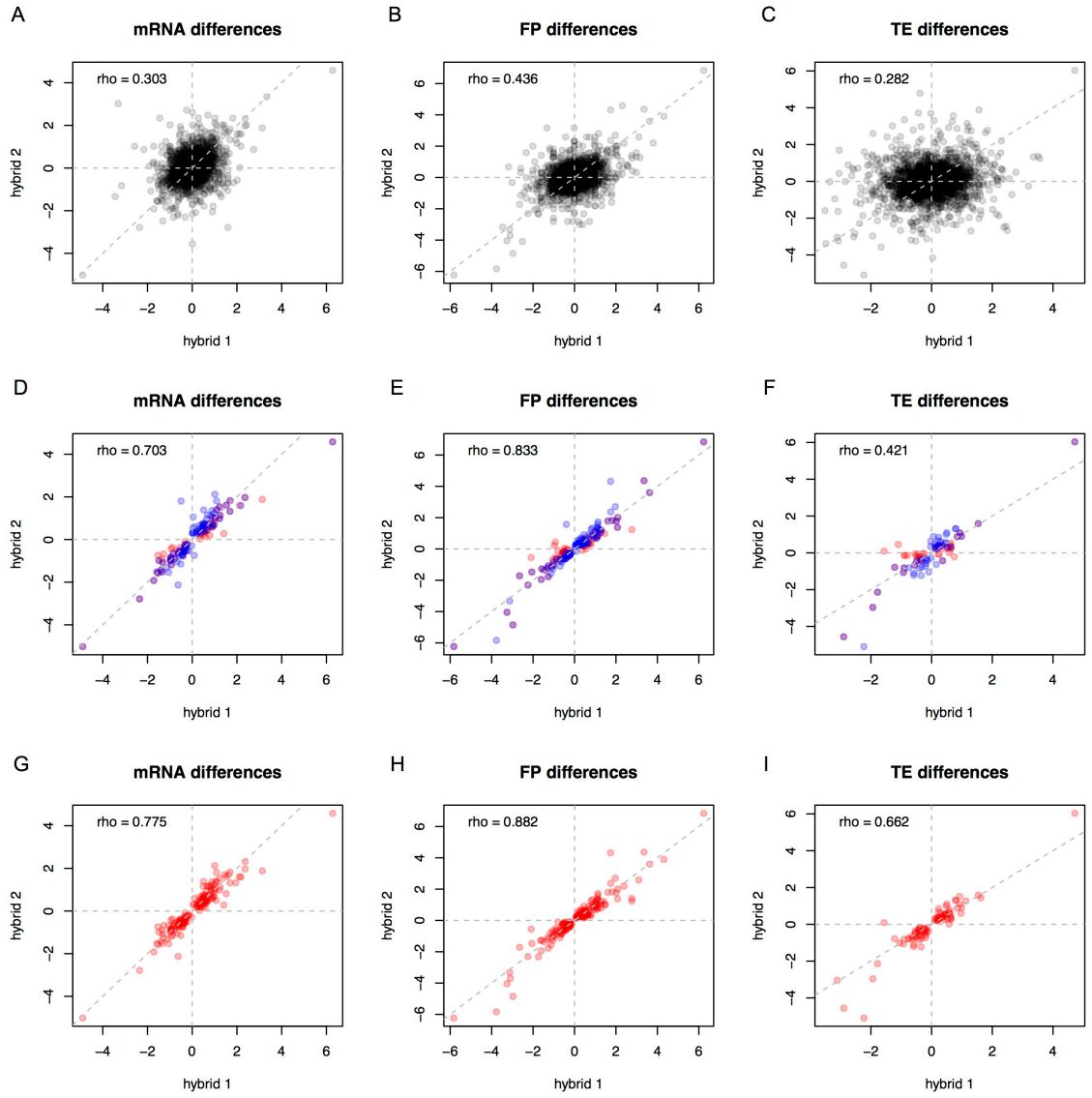
bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure S1 – Comparison to Ingolia 2009 data

Shown are log10 transformed normalized read counts. The grey line marks identity.

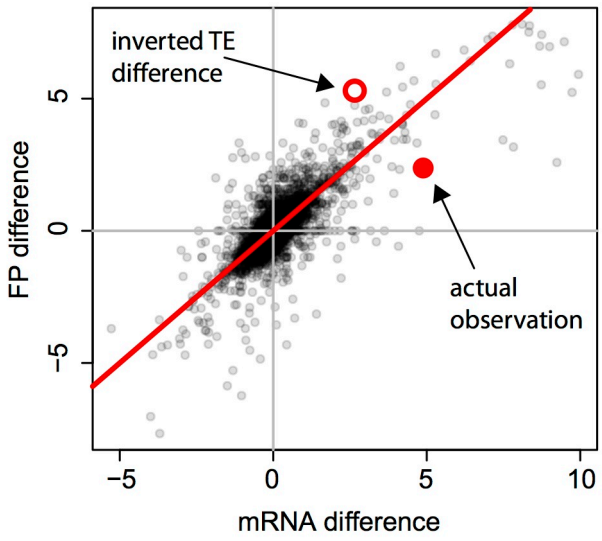
bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure S2 – Reproducibility of hybrid measurements

Shown are log2-transformed fold changes. Grey diagonals mark identity. Top row: all genes. Middle row: significant genes in one and / or the other replicate. Bottom row: significant genes in the combined hybrid data. Spearman correlation coefficients between replicates are given in each panel.

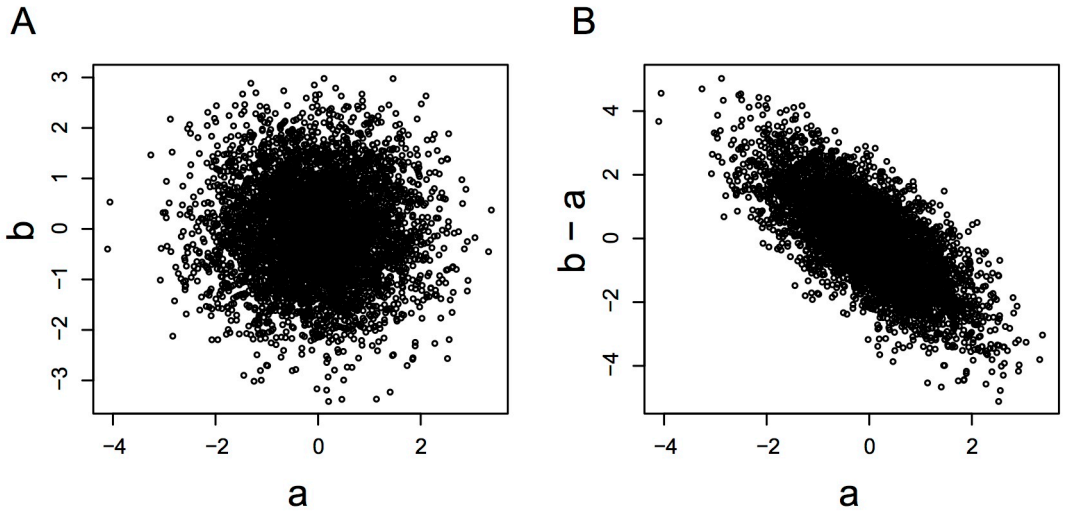
bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure S3 – Illustration of the TE randomization strategy

Shown are the parental mRNA and footprint fold changes, with one gene highlighted in red. The observed data for the gene is shown as a filled circle. The position with inverted TE, where the mRNA and footprint fold changes have been swapped, is shown as an open circle. In the randomization procedure, each gene has a 50% chance of retaining its observed data and of having its TE effect inverted. On average, there is thus no preference for reinforcing or buffering effects of translation in datasets constructed this way.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure S4 – Spurious correlations induced by correlations between a log ratio and its denominator

A. A scatterplot of two random samples *a* and *b* of size 5,000 from a standard normal distribution with mean = 0 and standard deviation = 1. Note that *a* and *b* are entirely uncorrelated. B. The correlation between the quantity *b* - *a* and *a* is negative and highly significant because of regression to the mean. For example, when *a* happens to be large by chance, the corresponding value of *b* will usually be closer to the mean than *a* because it is unlikely that a large value is sampled two times by chance. Therefore, the quantity *b* - *a* is systematically more likely to be less than zero for $a > 0$. If *a* and *b* are interpreted as the logarithms of mRNA and footprint differences, *b* - *a* is equivalent to the corresponding TE differences. A negative correlation between TE differences and mRNA differences is thus not by itself sufficient to infer translational buffering.

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Note 1 – Genomic sources of mRNA and footprint fragments

We generated 189 million (M) and 222 M ribosomal footprint reads from BY and RM, respectively (see Supplementary Table S1 for read and alignment statistics for all samples). In parallel, we gathered mRNA data from the same yeast cultures. Of the reads, 82 M (43%) for BY and 151 M (68%) for RM mapped to unique positions in the genome. The difference in the percentage of uniquely mapping reads is likely due to differences in the efficacy of ribosomal RNA (rRNA) depletion during library construction. rRNA is transcribed from a highly repetitive region of the genome (s. below) and greatly outnumbers mRNA in exponentially growing yeast cells [55] so that even minor differences in the efficacy of rRNA depletion can lead to large differences in rRNA retention. As a consequence, more uniquely mapping mRNA sequencing reads are available in samples with more effective rRNA depletion.

We determined which genomic features were the source of the footprint and mRNA reads. These analyses were conducted on the data from the BY strain, in order to allow us to directly use the reference genome annotation (Supplementary Table S2). Of uniquely mapping mRNA reads, the vast majority (84%) corresponded to either protein coding sequences (CDS) or untranslated regions (UTRs; 17%, the sum can be more than 100% because of overlapping annotations). For footprints, 97% of uniquely mapping reads aligned to CDS, while 6% mapped to UTRs. The higher fraction mapping to CDS in footprints fits the expectation that translating ribosomes should be preferentially found on coding regions, rather than on UTRs. Notably, of mRNA reads that uniquely mapped to UTRs, 35% mapped to 5' UTRs, while this fraction was nearly twice as high (68%) in the footprints. A higher density of ribosomes in 5'UTRs than expected based on mRNA abundance may be due to ribosomes that translate upstream open reading frames (uORFs) on 5'UTRs but not 3'UTRs [12,45]. Of the reads that mapped to multiple locations in the genome, many fewer mapped to CDS and UTRs (Supplementary Table S2). These "repeat reads" were heavily dominated by the ribosomal rRNA genes: 90% of mRNA and 85% of footprint reads mapped there. The rRNA genes are each represented by a small number of gene annotations in the yeast reference genome sequence (two

bioRxiv preprint doi: https://doi.org/10.1101/003319; this version posted March 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

annotations for 35S pre-rRNA and six annotations for different variants of 5S rRNA) which represent 100-200 tandem repeats of the rDNA locus [56]. The large number of rRNA reads therefore reflects rRNA transcription across all rDNA repeats and is in line with the high amounts of rRNAs in exponentially growing yeast [55]. In this work, we considered only reads with unique alignments.