Positive Selection for Unpreferred Codon Usage in Eukaryotic Genomes

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

| Citation       | Neafsey, Daniel E., and James E. Galagan. 2007. Positive selection for unpreferred codon usage in eukaryotic genomes. BMC Evolutionary Biology 7:119. |
|----------------|-------------------------------------------------------------------------------------------------------------------------------------|
| Published Version | doi:10.1186/1471-2148-7-119                                                                                                           |
| Citable link    | http://nrs.harvard.edu/urn-3:HUL.InstRepos:4595137                                                                                   |
| Terms of Use    | This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA |
Positive selection for unpreferred codon usage in eukaryotic genomes
Daniel E Neafsey* and James E Galagan

Abstract

**Background:** Natural selection has traditionally been understood as a force responsible for pushing genes to states of higher translational efficiency, whereas lower translational efficiency has been explained by neutral mutation and genetic drift. We looked for evidence of directional selection resulting in increased unpreferred codon usage (and presumably reduced translational efficiency) in three divergent clusters of eukaryotic genomes using a simple optimal-codon-based metric ($K_p/K_u$).

**Results:** Here we show that for some genes natural selection is indeed responsible for causing accelerated unpreferred codon substitution, and document the scope of this selection. In Cryptococcus and to a lesser extent Drosophila, we find many genes showing a statistically significant signal of selection for unpreferred codon usage in one or more lineages. We did not find evidence for this type of selection in Saccharomyces. The signal of positive selection observed from unpreferred synonymous codon substitutions is coincident in Cryptococcus and Drosophila with the distribution of upstream open reading frames (uORFs), another genic feature known to reduce translational efficiency. Functional enrichment analysis of genes exhibiting low $K_p/K_u$ ratios reveals that genes in regulatory roles are particularly subject to this type of selection.

**Conclusion:** Through genome-wide scans, we find recent selection for unpreferred codon usage at approximately 1% of genetic loci in a Cryptococcus and several genes in Drosophila. Unpreferred codons can impede translation efficiency, and we find that genes with translation-impeding uORFs are enriched for this selection signal. We find that regulatory genes are particularly likely to be subject to selection for unpreferred codon usage. Given that expression noise can propagate through regulatory cascades, and that low translational efficiency can reduce expression noise, this finding supports the hypothesis that translational efficiency may be suppressed in some cases to reduce stochastic noise in gene expression.

**Background**

It is generally accepted that natural selection operates to increase translational efficiency in the genomes of unicellular as well as some multicellular organisms [1-4]. The observation that codons translated by common tRNAs are used more frequently than synonymous codons translated by rare tRNAs, and that this usage bias strengthens with gene expression level, is interpreted as a signal of selection to increase translation rate [5] and/or accuracy[6]. The observation that codons translated by rare...
tRNAs can significantly decrease translation rate [7,8] has likewise motivated many reports claiming evidence for the downward modulation of expression level through the use of inefficiently translated codons [9-13].

Though this analogous argument for down-regulation of expression through codon usage, known as the 'expression-regulation theory,' has a symmetric appeal, it has been strongly and repeatedly challenged [9,14,15]. Among the principal objections to the theory are simple pragmatism; because the process of transcription consumes cellular resources and energy, it would therefore be natural to assume that most constitutive (non-regulatory) constraints on gene expression levels would be imposed before, rather than after transcription. Indeed, it has even been suggested that such constraints would be more easy to evolve at the transcriptional rather than the translational level [14]. Further, it was assumed that synonymous substitutions creating inefficiently translated codons would have no perceptible phenotypic effect except in very highly expressed genes, making such mutations effectively invisible to natural selection [9,14]. Many early studies in favor of the expression-regulation theory also failed to document a significant enrichment of translationally inefficient codons in genes thought to be subject to translational repression [14], leading to the 'selection-mutation-drift theory' that weak codon bias results from an absence of selection for translational efficiency, rather than from selection in the opposite direction.

Recently, however, positive selection for inefficiently translated codons has been reported for several exons of a gene in the fruitfly Drosophila melanogaster [16,17], and a deficit of translationally efficient codons has been detected in some human genes [18], suggesting that some cases of inefficient translation may indeed be an evolutionarily deliberate strategy. A larger than expected in vivo phenotypic effect of translationally inefficient codon substitutions has also been observed at the Adh locus in D. melanogaster [8], suggesting that such mutations might be more evolutionarily labile than previously believed.

Motivated by these findings, we conducted a genome-wide scan for selection for inefficient translation in two widely divergent fungal lineages and Drosophila, using clusters of three to four closely related species from each lineage to observe and root recent synonymous substitutions. We use the rate of unpreferred codon substitution as a measure of selection for translational inefficiency. While this signal may also be interpretable as a measure of selection for translational inaccuracy, we know of no hypotheses predicting such selection. Further, we find that in Cryptococcus and Drosophila, the signal of positive selection observed from synonymous codon substitutions in certain genes is concordant with the distribution and conservation of upstream open reading frames (uORFs), another generic feature known to reduce translational efficiency. We find that natural selection does in fact operate at many loci across the genome in Cryptococcus and several loci in Drosophila to reduce preferred codon usage and presumably translational efficiency, but we fail to find such a signal in Saccharomyces. Based on functional enrichment analysis of genes showing accelerated rates of mutation towards translationally inefficient codons, we suggest this selection may be acting to minimize stochastic noise in gene expression.

Results

Synonymous Codon analysis

To perform genome-wide scans for selection for inefficient translation, we employed a straightforward metric that detects recent mutation-selection disequilibrium for codon usage, using partitioned counts of synonymous sites and substitutions as described by Bauer Dumont et al. [16]. To calculate this metric, synonymous codons were assigned "preferred," "unpreferred," or "equal" status according to whether their usage differed significantly between gene sets exhibiting high or low overall codon bias in each genome (Methods; Additional Files 1, 2, 3). We then classified synonymous substitutions in aligned orthologous genes within each genus as preferred or unpreferred according to the status of the ancestral and derived codons [19], using a maximum likelihood approach to infer ancestral states (Methods). For example, an unpreferred codon that changes into a preferred synonymous codon would be classified as a preferred synonymous substitution, and the opposite directionality of change would be classified as an unpreferred synonymous substitution. Counts of synonymous substitutions were corrected for multiple substitutions [20].

To create normalized rates of synonymous substitution, we divided each substitution count by the number of ancestral synonymous ‘sites’, or opportunities for mutation available for each class of substitution. Note that on average, preferred codons will tend to exhibit more unpreferred sites, or opportunities for unpreferred changes, than unpreferred codons, and likewise unpreferred codons will exhibit more preferred sites than preferred codons. Normalizing the counts of preferred and unpreferred substitutions by the number of ancestral preferred and unpreferred sites, respectively, therefore allows one to make fair comparisons of the relative rate of preferred to unpreferred mutations among genes exhibiting differing degrees of ancestral codon usage bias. We define Kp as (# of preferred synonymous substitutions)/(# of preferred ancestral sites) and Ku as (# of unpreferred synonymous substitutions)/(# of unpreferred ancestral sites). Taking the quotient of Kp and Ku (Kp/Ku) then yields a metric for
We calculated the \( K_p/K_u \) metric for 5,450 Cryptococcus genes, 5,921 Drosophila genes, and 5,158 Saccharomyces genes (Figure 1). Gene set sizes were determined by the number of all-way reciprocal-best-BLAST hits that were obtained within each clade (Methods). Cryptococcus exhibited the strongest signal of accelerated unpreferred substitution. We found 125 Cryptococcus genes exhibiting \( K_p/K_u \) ratios less than 1 at a \( p \) value \( \leq 0.01 \) (1-tailed Fisher’s exact test; Additional File 4). We also found 69 Drosophila genes and 36 Saccharomyces genes exhibiting ratios with such low \( p \) values (Additional Files 5 &6). Q-value analysis [21] to account for multiple testing suggests a false discovery rate of 32% among the set of 125 Cryptococcus genes, 91% among the 96 Drosophila genes, and 100% among the 33 Saccharomyces genes, yielding approximately 85 genes in Cryptococcus and 6 genes in Drosophila that reflect strong selection for translational inefficiency mediated via accelerated unpreferred synonymous substitution in one or more lineages.

Analysis of substitution patterns in the 5’ leader and 3’ trailer sequences flanking these genes indicates the observed selection signal does not derive from selection on local nucleotide composition or biased mutation rates. In Cryptococcus and Drosophila, unpreferred codons uniformly exhibit an A or U and preferred codons uniformly exhibit a G or C in the third position within the tyrosine, histidine, glutamine, asparagine, lysine, aspartic acid, and glutamic acid two-fold degenerate synonymous codon families (Additional Files 1 &2). A lesser GC bias also exists among preferred codons in Saccharomyces (Additional File 3). This creates the possibility that genes with a low Kp/Ku ratio reflect localized selection for lower GC content or regional mutation bias rather than selection for translational inefficiency in these genera. To test for this, we compared patterns of nucleotide substitution in the 5’ leader and 3’ trailer sequences of two sets of genes from each taxonomic cluster: a set exhibiting the lowest observed Kp/Ku ratios (5th percentile and below), and a set exhibiting the highest Kp/Ku ratios (95th percentile and above). Ancestral and derived states were inferred for substitutions in these flanking regions in the same manner as for substitutions at synonymous coding sites (Methods).

In both fungal genera and in fruitflies, the ratios of A/T-to-G/C and G/C-to-A/T substitution rates did not significantly differ between the high and low \( K_p/K_u \) gene sets (Table 1; \( \chi^2 \) test; Cryptococcus \( p = 0.68 \); Saccharomyces \( p = 0.59 \); Drosophila \( p = 0.50 \)). Thus we infer that genes exhibiting \( K_p/K_u \) ratios significantly less than or greater than 1 likely reflect selection on the translational properties of codons rather than local selection for nucleotide composition or a locally biased mutation profile.

While Saccharomyces yielded few genes with \( K_p/K_u \) ratios significantly less than 1, there remains evidence that the statistic is an estimator of selection on translational efficiency in this organism. We find a highly significant association between \( K_p/K_u \) ratio and empirical measurements of translational efficiency based on ribosome density on transcripts [22] (Additional File 7; Spearman’s rho = 0.28; \( p < 0.00001 \)).

We observed that the recent selective forces on genes reflected by \( K_p/K_u \) are in most cases concordant with historical selection pressures on genes, as measured by codon bias. Genes exhibiting an excess of preferred synonymous substitutions (high \( K_p/K_u \)) tend to exhibit stronger codon bias in Cryptococcus, Saccharomyces, and Drosophila (Figure 2), suggesting that selection is continuing to strengthen or reinforce codon bias in those genes. However, genes exhibiting the most extreme unpreferred synonymous substitution rates (\( K_p/K_u < 0.25 \)) demonstrate stronger average codon bias than genes exhibiting more moderate unpreferred substitution rates (\( K_p/K_u 0.25–0.75 \)), particularly in Cryptococcus. A similar pattern was recently observed in the human genome, where genes exhibiting the lowest incidence of optimal codons (where optimality was determined by tRNA gene counts) exhibit stronger codon bias than genes exhibiting intermediate levels of optimal codon usage [18]. In Cryptococcus, however, we found that genes exhibiting \( K_p/K_u \) ratios less than 0.25 actually exhibit a higher ratio of preferred to unpreferred codon incidence (1.47 vs. 1.40; \( \chi^2 = 9.32, p = 0.002 \)) than genes exhibiting more moderate \( K_p/K_u \) ratios (0.25–0.75), as well as a higher average count of genomic tRNAs/codon (4.66 vs. 4.36; 2-tailed \( t \) test, \( p = 8.4E^{-11} \)). These results suggest that the concave shape of the curves in Figure 2 is most likely due to asymmetric variance in \( K_p/K_u \) for genes exhibiting strong codon bias.
at genes fitting these requirements in Cryptococcus strain JEC21, 642 uORFs in D. melanogaster, and 403 uORFs in S. cerevisiae. Of these total counts, 107, 417, and 102 uORFs were respectively conserved in all species of Cryptococcus, Drosophila, and Saccharomyces (Additional File 8). Using a Mann-Whitney U test, we found a statistically significant association between K_p/K_u and uORF presence in Cryptococcus and Drosophila, both for ‘all’ uORFs (present in reference species for genus) and ‘conserved’ uORFs (present in all species in genus; Table 2). This suggests that for some genes, selection may be operating in parallel on synonymous codon usage and uORFs to reduce translational efficiency, and that suppression of translation efficiency may be a more important mechanism of eukaryotic gene regulation than currently appreciated.

**Functional enrichment analysis**

In all three genera, we identified the Gene Ontology Biological Processes that were significantly enriched for low K_p/K_u ratios using a Mann-Whitney U test (Table 3). All categories that were significantly enriched for low K_p/K_u ratios, and presumably inefficient translation, were either explicitly or potentially associated with regulatory or signal transduction roles.

**Discussion**

We report strong evidence of natural selection for unpreferred codon usage across dozens of genes in Cryptococcus and Drosophila. We find an association between the simple K_p/K_u selection metric and translational efficiency in Saccharomyces, as well as a significant association between K_p/K_u and uORFs in Cryptococcus and Drosophila, suggesting selection is acting on translational efficiency as opposed to accuracy. Further, we report an enrichment of this selection signal in genes regulating transcription or cellular processes in all three genera. Selection for unpreferred codon usage, and potentially reduced translational efficiency, is likely to be more common than previously thought among Eukaryotes, and is a factor that must be considered in extrapolating ultimate gene expression levels from the vast amounts of transcription data now available.

Though we find that mutational bias or selection for nucleotide composition are not likely to cause the selection signal we see in Cryptococcus and Drosophila, it is possible that some non-translational factor is driving the selection pattern we observe. Recent work has shown that synonymous codon usage may be subject to constraint imposed by mRNA secondary structure[26,27], exonic splicing enhancers[28,29], and even microRNA binding[30]. Given that unpreferred codons in the lineages we examined uniformly end in an A or U nucleotide, if G/C-to-A/U mutations are more likely to induce an advantageous change in mRNA secondary structure or some other
molecular phenotype than mutations in the opposite direction, such a phenotype may be the true subject of selection. These additional functional roles imposed on silent coding sites might be expected to generally reduce synonymous substitution rates, in contrast to the accelerated substitution rates that we see in selected genes, but could contribute to occasional bouts of disequilibrium in substitution patterns.

Another caveat derives from the fact that the method we use to assign preferred and unpreferred codon status assumes consistent codon preference across growth/developmental phases, tissue types, etc. Codon preferences are known to be variable within an organism (e.g. [31]) in accordance with spatially or temporally fluctuating tRNA levels. So, the possibility exists that some of the genes exhibiting accelerated unpreferred substitution in fact have inverted usage preferences and are undergoing accelerated preferred substitutions. The accelerated signals of unpreferred substitution we report do not occur within single synonymous codon families, however, but across many amino acids. For a preference inversion to generate a significant acceleration in the opposing direction (as opposed to a nonsignificant result) would thus require a consistent preference inversion across many synonymous codon families, which we consider mechanistically unlikely.

The present analysis may be considered a conservative assessment of the extent of selection for unpreferred codon usage for several additional reasons. The selection signal we report represents an average level of mutation-selection disequilibrium in codon usage across multiple species within each taxonomic group. This approach increases statistical power by boosting the observed number of synonymous substitutions per gene, but may obscure species-specific selection. If only one species in the group has undergone selection for unpreferred codon usage/inefficient translation, the signal of that selection will be diluted by the sister lineages that are in selection-equilibrium for synonymous codon usage. Measuring $K_p/K_u$ ratios for each species, as opposed to across several species at a time, yields distinct but overlapping sets of genes that show significant evidence of selection for inefficient translation (results not shown). This indicates that some of the genes we identify as having $K_p/K_u$ ratios significantly less than 1 may be undergoing selection in multiple species, but also that we are overlooking some genes subject to species-specific selection. Power may also be slightly compromised in our combined-species analysis by minor divergences among species in mutation profiles or preferred/unpreferred codon assignments. Nielsen et al. [17] recently published a likelihood-based estimator of selection on codon usage that may be more suited to detailed studies of selection at loci of interest.

Despite these considerations, we found that on the order of 1% of genes in the Cryptococcus genome and several genes in the Drosophila genome exhibit a statistically significant signal of selection for unpreferred codon usage. Given that these genes are enriched for uORFs, we interpret this as selection for translational inefficiency rather than inaccuracy. There may ultimately be many reasons to select for reduced translational efficiency. Several hypotheses have recently been advanced to explain reduced translational efficiency, including facilitation the maintenance of open chromatin structure via low-level transcription [32], facilitation of domain folding during translation via reduced rates of elongation [33], as well as minimization of stochastic gene expression noise [34-36]. The expression noise hypothesis derives from recent analytical and empirical findings that low translational efficiency results in less noisy gene expression [34,36-38]. Fraser et al. [35] recently found evidence that gene expression noise may be a trait subject to natural selection, as essential genes and genes that encode subunits of protein complexes in yeast (two proposed classes of genes particularly requiring precise expression) are expressed with less predicted noise than most other genes.

| Mutation Type | A/T->G/C | GC->A/T | Chi. Sq. | P value |
|---------------|---------|--------|---------|--------|
| Cryptococcus  |         |        |         |        |
| low $K_p/K_u$ | 319     | 333    | 0.2     | 0.66   |
| high $K_p/K_u$| 360     | 394    |         |        |
| Saccharomyces |         |        |         |        |
| low $K_p/K_u$ | 165     | 128    | 0.29    | 0.59   |
| high $K_p/K_u$| 310     | 260    |         |        |
| Drosophila    |         |        |         |        |
| low $K_p/K_u$ | 219     | 269    | 0.46    | 0.5    |
| high $K_p/K_u$| 132     | 179    |         |        |
The noise minimization hypothesis generates a testable prediction in the context of the current results, as analytical and empirical models show that expression noise can be propagated through gene regulatory cascades [34,38,39]. Genes in regulatory cascades that are noisily expressed might therefore engender more severe fitness consequences than noisily expressed genes in non-regulatory roles, as regulator noise may be amplified at downstream targets. We predicted, therefore, that transcriptional regulators and other classes of regulatory genes should be insulated from noise in order to prevent the propagation and amplification of expression noise through a genetic cascade.

When we looked for biological processes significantly enriched for genes exhibiting low \( K_p/K_u \) ratios in \textit{Saccharomyces}, \textit{Cryptococcus}, and \textit{Drosophila}, we found transcriptional regulators, protein kinases, DNA/nucleotide binding genes, and proteins involved in signal transduction enriched across the three taxonomic groups (Table 3). The presence of transcriptional regulators in this list directly confirms our hypothesis. Many genes annotated as nucleotide binding are potentially regulatory as well. Protein kinases are involved in the regulation of many cellular processes, and may therefore likely command expression levels as precise as those genes regulating transcription or transducing molecular signals.

**Conclusion**

Using patterns of synonymous substitution, we detect evidence of recent selection for unpreferred codon usage at dozens of genetic loci in both a fungal and insect lineage. The accelerated unpreferred synonymous substitution rates we see may result from selection for translational inefficiency or inaccuracy, or may also represent selection on mRNA secondary structure or some other molecular phenotype. This signal of positive selection is concordant with purifying selection observed at uORFs. As uORFs are another genic feature known to reduce translational efficiency, this lends credence to the hypothesis that translational inefficiency is the driver of selection, perhaps to limit expression noise. The similar functional enrichment profile observed across the fruitfly lineage and two deeply divergent fungal lineages suggests that selection to moderate expression noise of genes involved in signaling, activation, or regulation of other genes may be a general phenomenon in eukaryotic genomes, and that this noise moderation is attainable through unpreferred codon usage, uORFs, and possibly other mechanisms as yet undiscovered.

**Methods**

**Sequence resources and 5’ Leader/3’ Trailer Mapping**

We obtained the genome assemblies of four species belonging to the \textit{Cryptococcus neoformans} species complex from the websites of the sequencing centers that produced them (strain JEC21: TIGR; strain WM276: Michael Smith Genome Center; strains H99 and R265: Broad Institute). We used gene calls from TIGR for strain JEC21. We used gene calls produced by Jason Stajich [40] for strains R265, H99, and WM276. We defined the most distal extent 5’ leader sequences and 3’ trailer sequences in the alignments using a library of 23,000 full-length cDNAs from strain JEC21 produced by TIGR [41]. We retained for analysis only those 5’ and 3’ leader/trailer sequences that showed no evidence of introns and exhibited conserved genic start/stop codons.
For *Drosophila*, we obtained the *D. melanogaster* release 4.3 assembly and annotation from FlyBase [42]. We obtained the most recent *D. simulans*, *D. sechellia*, and *D. yakuba* assemblies and annotations from the UCSC Genome Bioinformatics webpage [43]. We defined 5' leader and 3' trailer regions according to the *D. melanogaster* release 4.3 annotations.

For *Saccharomyces*, we obtained the most recent *S. cerevisiae* strain S288C assembly and gene calls from SGD [44]. We obtained genome assemblies for *S. paradoxus* and *S. mikatae* from the Broad Institute website [45]. We defined the most distal extent of *Saccharomyces* 5' leader and 3' trailer regions using 5' SAGE data [46] as well as expression tiling array data [47]. In cases where these two data sources cited different leader lengths for the same gene, we favored the longer estimate.

**Synonymous Codon analysis**

Codon analyses were performed on clusters of aligned, orthologous genes. Orthology was determined within each clade using a reciprocal-best-BLAST hit criterion implemented with a custom Perl script. Orthologs were aligned using ClustalW [48].

We evaluated codon bias for each gene in all genomes with the ENC' statistic[49]. We conferred preferred, unpreferred, and equal status on each gene after the method of Sharp and Lloyd [50], using genes that scored below the 10th percentile and above the 90th percentile of codon bias as ‘highly’ and ‘lowly’ biased gene sets for evaluation of relative codon usage. A heterogeneity chi square test ($\chi^2$ highly biased + $\chi^2$ lowly biased – $\chi^2$ pooled) was used to identify divergent codon usage patterns between the two gene sets. Codons exhibiting heterogeneity $\chi^2$ values greater than 7.88 ($p < 0.005$) among the highly and lowly biased gene sets were assigned preferred or unpreferred status; less significant usage differences were interpreted as equal status. Codon usage preferences as determined by this method are nearly identical within the *Cryptococcus*, *Drosophila*, and *Saccharomyces* genera (Additional Files 1, 2, 3). We computed tallies for each subclass of synonymous site in each gene as described in Bauer DuMont et al. [16], using an empirical substitution rate matrix derived from substitutions observed among orthologous 5’ leader and 3’ trailer sequences (Additional File 9). The *Drosophila* empirical rate matrix we derived using this method is very similar to that reported by Petrov and Hartl [51]. Ancestral codon states for synonymous differences observed were inferred using a maximum likelihood approach implemented in the codeml program of PAML 3.14[52]. The codon with the highest posterior probability under the marginal reconstruction approach was assumed ancestral. To avoid ambiguity, only synonymous codons differing by a single base change between their ancestral and derived states were utilized, and orthologous codons exhibiting signs of nonsynonymous change in any lineage were discarded.

We identified $K_p/K_u$ ratios significantly less than one using a 1-tailed Fisher’s exact test (FET). We performed false discovery rate analysis [21] on the FET $p$ value distribution.

**Table 2: Association between $K_p/K_u$ $p$ values and uORFs.**

|              | P value for $K_p/K_u$ association w/uORFs$^a$ |
|--------------|---------------------------------------------|
|              | all uORFs | conserved uORFs |
| Cryptococcus | 0.047     | 0.0069          |
| Saccharomyces| 0.19      | 0.68            |
| Drosophila  | 0.017     | 0.05            |

$^a$ Analyses performed using Mann-Whitney U test on FET 1-tailed $p$ values

**Table 3: Gene Ontology Biological Processes significantly enriched for genes exhibiting low $K_p/K_u$ ratios.**

| GO Biological Function Category | Bonferroni-corrected $P$ value |
|--------------------------------|--------------------------------|
| Cryptococcus                   |                                |
| GO:0030528 Transcription regulator activity | 0.00052                        |
| GO:0084672 Protein kinase activity          | 0.05                            |
| Saccharomyces                  |                                |
| GO:0003677 DNA-binding           | 0.00079                         |
| GO:0030528 Transcription regulator activity | 0.0088                         |
| GO:0084672 Protein kinase activity          | 0.024                           |
| Drosophila                     |                                |
| GO:0000166 Nucleotide binding      | 0.003                           |
| GO:0004871 Signal Transducer Activity   | 0.02                            |
using the QVALUE software package[53]. This software uses a density histogram of p values to calculate the incidence of false positive results for p values less than or equal to a given value.

We analyzed the nature of the codon bias exhibited by genes showing extremely low Kp/Ku ratios in Cryptococcus using counts of different tRNAs in the C. neoformans JEC21 genome and the ratio of preferred to unpreferred codons in JEC21 genes. The number of tRNA genes per codon was calculated by counting the tRNAs identified in the official TIGR annotation of JEC21 and applying standard eukaryotic wobble pairing rules. The genes of JEC21 for which Kp/Ku could be calculated were then analyzed to determine the arithmetic average number of tRNAs/codon.

uORF analysis
We conducted all analyses on uORFs with custom Perl scripts. For the purposes of this analysis we defined a uORF as an AUG triplet followed by at least one intervening codon and a stop codon (UAG, UAA, or UGA). uORFs were permitted to overlap with each other. We required uORFs to be either contained entirely within the 5' leader sequence or to overlap with the downstream coding ORF by at most a single base. We considered a uORF to be conserved if, in the multiple alignment of orthologous leader sequences, all strains exhibited a start codon and a stop codon in the same position, and those start and stop codons were in the same frame relative to each other.

Functional enrichment analysis
We performed functional enrichment analysis for the Kp/Ku results using a Mann-Whitney U test. Gene Ontology annotations for Drosophila melanogaster, Saccharomyces cerevisiae, and Cryptococcus neoformans strain JEC21 were respectively obtained from FlyBase, SGD, and TIGR. Gene Ontology Biological Process categories annotated to at least 10 genes in the reference genome of each clade were used for enrichment analysis. Categories that were functionally overlapping or nested in each annotation were condensed using the GOSLIM algorithm[54] to minimize the multipletesting penalty. Speciesspecific slims were used for Saccharomyces and Drosophila; a generic slim was used for Cryptococcus. Bonferroni correction was applied to enrichment p values to compensate for the testing of multiple categories.

Authors’ contributions
DN conceived of the study, assembled the data set, carried out the analyses, and drafted the manuscript. JG assisted with guiding the analyses, interpretation of the results, and drafting of the manuscript. Both authors read and approved the final manuscript.

Appendix 1: Rationale behind the Kp/Ku Statistic
That Kp/Ku statistic may be used to compare the relative rates of preferred and unpreferred synonymous substitution. Because an ancestral sequence may contain different numbers of preferred and preferred ‘sites’ (opportunities for each type of mutation), it is necessary to normalize the count of each class of substitution by the number of ancestral sites in each class. Kp is defined as (no. of preferred subs/no. preferred ancestral sites), and Ku is likewise defined as (no. of unpreferred subs/no. of unpreferred ancestral sites). Ancestral codons that are already in a preferred state tend to exhibit a higher ratio of unpreferred to preferred sites, just as unpreferred ancestral codons exhibit a higher ratio of preferred to unpreferred sites.

As an example, consider a gene encoding only one type of amino acid, tyrosine, which is encoded by only two possible synonymous codons, TAT and TAC. Further, let us assume that TAC is a preferred codon, TAT is an unpreferred codon, and that a Jukes-Cantor model of substitution applies.

We can then calculate that each TAC codon exhibits 1/3 of an unpreferred synonymous site [0 (pos. 1) + 0 (pos. 2) + 1/3 (pos. 3)], and 0 preferred synonymous sites. Likewise, TAT codons exhibit 1/3 of a preferred site, and 0 unpreferred sites. See methods section for further detail on estimating counts of sites.

If the hypothetical gene ancestrally contains 30 TAC codons and 30 TAT codons, then it exhibits (30 * 1/3 + 30 * 0) = 10 unpreferred sites and (30 * 0 + 30 * 1/3) = 10 preferred sites. That is, the ancestral sequence offers equal opportunities for both preferred and unpreferred synonymous substitutions to occur in descendant lineages.

Under conditions without selection for codon usage, and assuming no nonsynonymous mutations are tolerated, one would therefore expect on average to observe roughly equal numbers of preferred and unpreferred substitutions in descendant lineages. If 5 TAC codons turn into TAT codons, that would constitute 5 unpreferred substitutions, and Ku would be 5/10 = 0.5. If 5 TAT codons also turn into TAC codons, Kp is similarly 5/10 = 0.5, so the Kp/Ku ratio would be equal to 1 on average.

If, however, there is selection for increased unpreferred codon usage, then one might be more likely to see mutations that change TAC codons into TAT rather than vice versa. If 8 TAC-to-TAT changes and only 2 TAT-to-TAC changes occur, Kp would be 8/10 = 0.8 and Ku would be 2/10 = 0.2, yielding a Kp/Ku ratio less than 1 (0.25). Selection for greater preferred codon usage would similarly yield a Kp/Ku ratio that is greater than 1.
Fisher’s exact test, or similar statistical tests, may be used to identify significant deviations from the equilibrium expectation that \(K_p/K_u = 1\). Note that \(K_p/K_u\) is independent of ancestral codon usage bias, so that genes exhibiting either ancestrally high or low codon usage bias may both be expected to yield \(K_p/K_u\) ratios close to 1 if there is no change in selection for synonymous codon usage in any descendant lineages.

**Additional material**

**Additional file 1**
Synonymous codon class assignments in Cryptococcus spp. (a table of preferred, unpreferred, and equal codon assignments)
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S1.pdf](http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S1.pdf)

**Additional file 2**
Synonymous codon class assignments in Drosophila spp. (a table of preferred, unpreferred, and equal codon assignments)
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S2.pdf](http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S2.pdf)

**Additional file 3**
Synonymous codon class assignments in Saccharomyces spp. (a table of preferred, unpreferred, and equal codon assignments)
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S3.pdf](http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S3.pdf)

**Additional file 4**
Cryptococcus genes exhibiting \(K_p/K_u < 1\) at \(p < 0.01\) (a list of genes potentially exhibiting significantly accelerated unpreferred substitution)
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S4.pdf](http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S4.pdf)

**Additional file 5**
Drosophila genes exhibiting \(K_p/K_u < 1\) at \(p < 0.01\) (a list of genes potentially exhibiting significantly accelerated unpreferred substitution)
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S5.pdf](http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S5.pdf)

**Additional file 6**
Saccharomyces genes exhibiting \(K_p/K_u < 1\) at \(p < 0.01\) (a list of genes potentially exhibiting significantly accelerated unpreferred substitution)
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S6.pdf](http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S6.pdf)

**Additional file 7**
Translational efficiency vs. \(K_p/K_u\) in Saccharomyces cerevisiae (a graph comparing empirical measurements of translational efficiency with \(K_p/K_u\))
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S7.pdf](http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S7.pdf)

**Additional file 8**
Lists of genes with uORFs in Cryptococcus, Drosophila, and Saccharomyces (lists of uORFs occurring in reference species and uORFs conserved in each genus)
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S8.pdf](http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S8.pdf)

**Additional file 9**
Mutation profiles derived from 5’ leader and 3’ flanking regions (a table of rooted mutations observed in genomic flanking regions of all 3 genera)
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S9.pdf](http://www.biomedcentral.com/content/supplementary/1471-2148-7-119-S9.pdf)

**Acknowledgements**
We thank Bruce Birren, Matt Sachs, Raya Jiang, and two anonymous reviewers for feedback and comments on this work. This work was supported in part by funds from the National Science Foundation (JG) and the National Institute of Allergy and Infectious Diseases.

**References**
1. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R: Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 1981, 9(1):43-74.
2. Akashi H: Gene expression and molecular evolution. Curr Opin Genet Dev 2001, 11(6):660-666.
3. Duret L: Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 2002, 12(6):640-649.
4. Chamary JV, Parmley JL, Hurst LD: Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 2006, 7(2):98-108.
5. Ikemura T: Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 1985, 2(1):13-34.
6. Akashi H: Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 1994, 136(3):927-935.
7. Robinson M, Lilley R, Little S, Emmage JS, Yarranton G, Stephens P, Millican A, Eston M, Humphreys G: Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res 1984, 12(17):6663-6671.
8. Carlini DB, Stephan W: In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. Genetics 2003, 163(1):239-243.
9. Andersson SG, Kurland CG: Codon preferences in free-living microorganisms. Microb Evol 1990, 54(2):198-210.
10. Grosjean H, Fiers W: Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 1982, 18(3):199-209.
11. Konigsberg W, Godson GN: Evidence for use of rare codons in the dnaG gene and other regulatory genes of Escherichia coli. Proc Natl Acad Sci U S A 1983, 80(3):687-691.
12. Burns DM, Beacham IR: Rare codons in E. coli and S. typhimurium signal sequences. FEBS Lett 1985, 189(2):318-324.
13. Hinds PW, Blake RD: Delineation of coding areas in DNA sequences through assignment of codon probabilities. J Mol Biol Struct Dyn 1985, 3(3):543-549.
14. Sharp PM, Li WH: Codon usage in regulatory genes in Escherichia coli does not reflect selection for ‘rare’ codons. Nucleic Acids Res 1986, 14(19):7737-7749.
15. Bulmer M: The selection-mutation-drift theory of synonymous codon usage. Genetics 1991, 129(3):897-907.
16. Bauer DuMont V, Fay JC, Calabrese PP, Aquadro CF: DNA variability and divergence at the notch locus in Drosophila melanogaster.
nogaster and D. simulans: a case of accelerated synonymous site divergence. Genetcs 2004, 167(1):171-185.
17. Nielsen R, Hruska C, Dumont VL, Hubisz MJ, Aquadro CF: Research Article: Maximum Likelihood Estimation of Ancestral Codon Usage Bias Parameters in Drosophila. Mol Biol Evol 2006.
18. Lavner Y, Kodlar D: Codon bias as a factor in regulating expression via translation rate in the human genome. Gene 2005, 345(1):127-138.
19. Akashi H: Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 1995, 139(2):1067-1076.
20. Jukes TH, Cantor, C. R.: Evolution of protein molecules. In Mammalian Protein Metabolism Edited by: Munro HN. New York, Academic Press; 1969:21-123.
21. Storey JD, Tibshirani R: Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003, 100(16):9440-9445.
22. Parmley JL, Hurst LD: Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 2005, 6(9):R75.
23. Shabalina SA, Ogurtsov AY, Spiridonov NA: A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res 2006, 34(9):2428-2437.
24. Parmley JL, Chamary JV, Hurst LD: Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol 2006, 23(2):301-309.
25. Blakesly PW, KAE, Cantor CR, Collins JS: Noise in eukaryotic gene expression. Nature 2003, 422(6932):633-637.
26. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB: Noise minimization in eukaryotic gene expression. PLoS Biol 2004, 2(6):e137.
27. Mardis ER, Cantor CR: RNA-mediated regulation on coding sequence evolution in human and mouse. J Mol Biol 2006, 363(2):174-182.
28. Dong H, Nilsson L, Kurland CG: Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J Mol Biol 1996, 260(5):649-663.
29. Law GL, Bicket KS, Mackay VL, Morris DR: The undertranslated transcriptome reveals widespread translational silencing by alternative 5' transcript leaders. Genome Biol 2005, 6(13):r111.
30. Thanaraj TA, Argos P: Ribosome-mediated translational pause and protein domain organization. Protein Sci 1996, 5(8):1594-1612.
31. McAdams HH, Arkin A: Stochastic mechanisms in gene expression. Proc Natl Acad Sci U S A 1997, 94(3):814-819.
32. Jones T, Davis RW, Steinmetz LM: The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. J Biol Chem 2001, 276(11):8165-8172.
33. Kierzek AM, Zaim J, Zielenkiewicz P: The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. J Biol Chem 2001, 276(11):8165-8172.
34. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A: Regulation of noise in the expression of a single gene. Nat Genet 2002, 31(1):69-73.
35. Thattai M, van Oudenaarden A: Intrinsic noise in gene regulatory networks. Proc Natl Acad Sci U S A 2001, 98(15):8614-8619.
36. Resources for fungal comparative genomics [http://fun-gal.genome.duke.edu/]
37. TIGR database: Cryptococcus neoformans Genome Project [http://www.tigr.org/dtlb/e2k1/cna1/l]
38. Flybase [http://www.flybase.org]