Decomposing Mutation and Selection to Identify Mismatched Codon Usage

Cedric Landerer *

Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville TN 37996

National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996

Brian C. O'Meara

Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville TN 37996

National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996

Russell Zaretzki

Department of Business Analytics and Statistics, University of Tennessee, Knoxville TN 37996

National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996

Michael A. Gilchrist

Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville TN 37996

National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996

December 13, 2018

*Corresponding author: cedric.landerer@gmail.com

Abstract

For decades, codon usage has been used as a measure of adaptation for translational efficiency of a gene's coding sequence. These patterns of codon usage reflect both the selective and mutational environment in which the coding sequences evolved. Over this same period, gene transfer between lineages has become widely recognized as an important biological pheonmena. Nevertheless, most studies of codon usage implicitly assume that all genes within a genome evolved under the same selective and mutational environment, an assumption violated when introgression occurs. In order to better understand the effects of introgression on codon usage patterns and vice versa, we examine the patterns of codon usage in the yeast which has experienced a large introgression, *Lachancea kluyveri*. We quantify the effects of mutation bias and selection for translation efficiency on the codon usage pattern of the endogenous and introgressed exogenous genes using a Bayesian mixture model, ROC SEMPPR, which is built on mechanistic assumptions of protein synthesis and grounded in population genetics. We find substantial differences in codon usage between the endogenous and exogenous genes, and show that these differences can be largely attributed to a shift in mutation bias from A/T ending codons in the endogenous genes to C/G ending codons in the exogenous genes. Recognizing the two different signatures of mutation and selection bias improves our ability to predict protein synthesis rate by 17% and allowed us to accurately assess codon preferences. In addition, using our estimates of mutation and selection bias, we to identify *Eremothecium gossypii* as the most likely source lineage, estimate the introgression occurred $\sim 6 \times 10^8$ generation ago, and estimate its historic and current genetic load. Together, our work illustrates the advantage of mechanistic, population genetic models like ROC SEMPPR and the quantitative estimates they provide when analyzing sequence data.

Introduction

Synonymous codon usage patterns varies within a genome and between taxa, reflecting differences in mutation bias, selection, and genetic drift. The signature of mutation bias is largely determined by the organism's internal or cellular environment, such as their DNA repair genes or UV exposure. While this mutation bias is an omnipresent evolutionary force, its impact can be obscured or even amplified by selection. The signature of selection on codon usage is also largely determined by an organism's cellular environment, such as its tRNA species, their copy number, and post-transcriptional modifications. The strength of selection on the codon usage of an individual gene is largely determined by its expression level which, in turn, is also largely determined by the organism's external environment. In general, the strength of selection on codon usage increases with its expression level (Gouy and Gautier, 1982; Ikemura, 1985; Bulmer, 1990), specifically its protein synthesis rate (Gilchrist, 2007). Thus as gene expression increases, codon usage shifts from a process dominated by mutation to a process dominated by selection. The overall efficacy of selection on codon usage is a function of the organism's effective population size N_e which, in turn, is largely determined by its external environment. By explicitly modeling the combined forces of mutation, selection, and drift, ROC SEMPPR allows us disentangle the evolutionary forces responsible for the patterns of codon usage bias (CUB) encoded in an species' genome (Gilchrist, 2007; Shah and Gilchrist, 2011; Wallace *et al.*, 2013; Gilchrist *et al.*, 2015), should provide biologically meaningful information about the lineage's historical cellular and external environment.

Most studies implicitly assume that the CUB of a genome is shaped by a single cellular environment. As genes are horizontally transferred, introgress, or combined to form novel hybrid species, one would expect to see the influence of multiple cellular environments on a genomes codon usage pattern (Mdigue *et al.*, 1991; Lawrence and Ochman, 1997). Given that transferred genes are likely to be less adapted than endogenous genes to their new cellular environment, we expect a greater genetic load of transferred genes if donor and recipient environment differ greatly in their selection bias, making such transfers less likely. More practically, if differences in codon usage of transferred genes are unaccounted for, they may distort parameter estimates. Such distortion could lead to the wrong codon preference for an amino acid, underestimate the variation in protein synthesis rate, or bias mutation estimates when analyzing a genome.

To illustrate these ideas, we analyze the CUB of the genome of *Lachancea kluyveri*, which is sister to all other Lachancea. The Lachancea clade diverged from the Saccharomyces clade, prior to its whole genome duplication $\sim$ 100 Mya ago (Marcet-Houben and Gabaldn, 2015; Beimforde *et al.*, 2014). Since that time, *L. kluyveri* has experienced a large introgression of exogenous genes found in all populations (Friedrich *et al.*, 2015). The introgression replaced the left arm of the C chromosome and displays a 13%

higher GC content than the endogenous *L. kluyveri* genome (Payen *et al.*, 2009; Friedrich *et al.*, 2015). These characteristics make *L. kluyveri* an ideal model to study the effects of an introgressed cellular environment and the resulting mismatch in codon usage.

Using ROC SEMPPR, a Bayesian population genetics model based on a mechanistic description of ribosome movement along an mRNA, allows us to quantify the cellular environment in which genes have evolved by separately estimating the effects of mutation bias and selection bias on codon usage. ROC SEMPPR's resulting predictions of protein synthesis rates have been shown to be on par with laboratory measurements (Shah and Gilchrist, 2011; Gilchrist *et al.*, 2015). In contrast to often used heurisitc approaches to study codon usage (Sharp and Li, 1987; dos Reis *et al.*, 2004), ROC SEMPPR explicitly incorpoates and distinguishes between mutation and selection effects on codon usage. We use ROC SEMPPR to independently describe two cellular environments reflected in the *L. kluyveri* genome; the signature of the current environment in the endogenous genes and the decaying signature of the exogenous environment in the introgressed genes. Our results indicate that the difference in GC content between endogenous and exogenous genes is mostly due to the differences in mutation bias of their ancestral environments. Accounting for these different signatures of mutation bias and selection bias of the endogenous and exogenous sets of genes substantially improves our ability to predict present day protein synthesis rates. These endogenous and exogenous gene set specific estimates of mutation bias and selection bias, in turn allow us to address more refined questions of biological importance. For example, it allows us to identify *E. gossypii* as the most likely source of the introgressed genes out of the 38 yeast lineages with sequenced genomes, estimate the age of the introgression to be on the order of 0.2-1 Mya, estimate the genetic load of these genes, both at the time of introgression and now, as well as make predictions about how the CUB of the introgressed genes will evolve in the future.

Results

The Signatures of two Cellular Environments within *L. kluyveri*'s Genome

We used our software package AnaCoDa (Landerer *et al.*, 2018) to compare model fits of ROC SEMPPR to the entire *L. kluyveri* genome and its genome partitioned into two sets of 4,864 endogenous and 497 exogenous genes. AIC values strongly support the hypothesis that the *L. kluyveri* genome consists of genes with two different and distinct patterns of codon usage bias (ΔAIC $= 75,462$; Table 1). We find additional support for this hypothesis when we compare our predictions of gene expression to empirically observed values. Specifically, the explanatory power between our predictions and observed values improved by $\sim 42\%$, from $R^2 = 0.33$ to 0.46 (Figure 1).

Table 1: Model selection of the two competing hypothesis. Reported are the log-likelihood, $\log(\mathcal{L})$, the number of parameters estimated n, AIC, and ΔAIC values.

Hypothesis	$\log(\mathcal{L})$	n	AIC	ΔAIC
Separated	-2,612,397	5,402	5,235,598	0
Combined	-2,650,047	5,483	5,311,060	75,462

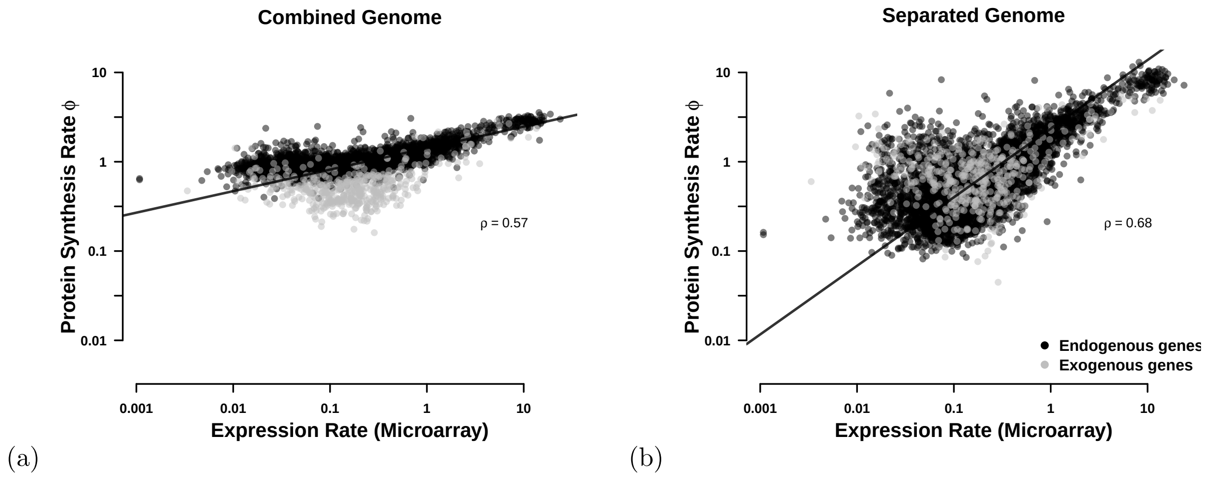


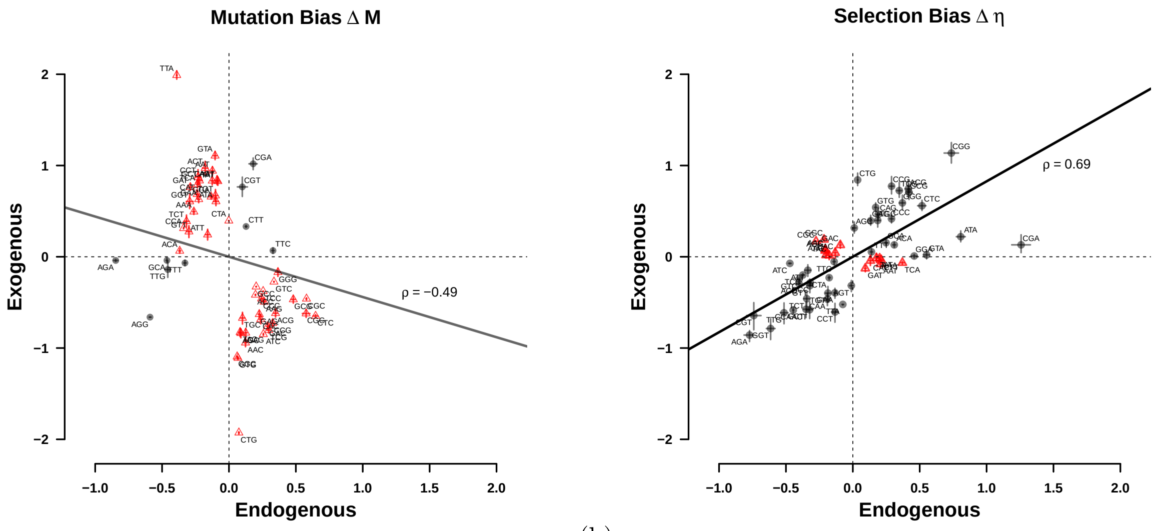
Figure 1: Comparison of predicted protein synthesis rate ϕ to microarray data from Tsankov *et al.* (2010) for (a) the combined genome and (b) the separated endogenous and exogenous genes. Endogenous genes are displayed in black and exogenous genes in gray. Black line indicates type II regression line (Sokal and Rohlf, 1981).

Comparing Differences in the Endogenous and Exogenous Codon Usage

To better understand the differences in the endogenous and exogenous cellular environments, we compared our parameter estimates of mutation bias ΔM and selection $\Delta\eta$ for the two sets of genes. Our estimates of ΔM for the endogenous and exogenous genes were negatively correlated ($\rho = -0.49$), indicating weak concordance of $\sim 5\%$ between the two mutation environments (Figure 2). For example, the endogenous genes show a mutational preference for A and T ending codons in $\sim 95\%$ of the codon families. In contrast, the exogenous genes display an equally consistent mutational preference towards C and G ending codons (Table S1). As a result, only the two codon amino acid Phenylalanine (Phe, F) shares the same rank order across the endogenous and exogenous ΔM estimates.

In contrast, our estimates of $\Delta\eta$ for the endogenous and exogenous genes were positively correlated ($\rho = 0.69$) and showing concordance of $\sim 53\%$ between the two selection environments (Figure 2). ROC SEMPPR constraints $E[\phi] = 1$, allowing us to interpret $\Delta\eta$ as selection on codon usage of the average gene with $\phi = 1$ and gives us the ability to compare the efficacy of selection sN_e across genomes. We find that the strength of selection within each codon family differs between sets of genes. Overall, the endogenous genes only show a selection preference for C and G ending codons in $\sim 58\%$ of the codon families. In contrast, the exogenous genes display a strong preference for A and T ending codons in $\sim 89\%$ of the codon families.

The difference in codon usage between endogenous and exogenous genes is striking. As a result, our estimates of the optimal codon differ in nine cases between endogenous and exogenous genes (Table S2).



(a)

(b)

Figure 2: Comparison of (a) mutation bias ΔM and (b) selection bias $\Delta\eta$ parameters for endogenous and exogenous genes. Estimates are relative to the mean for each codon family. Black dots indicate ΔM or $\Delta\eta$ parameters with the same sign for the endogenous and exogenous genes, red dots indicate parameters with different signs. Black line shows the type II regression line (Sokal and Rohlf, 1981). Dashed lines mark quadrants.

Fits to the complete *L. kluyveri* genome reveal that the relatively small exogenous gene set ($\sim$ 10% of genes) has a disproportional effect on the model fit. We find that the complete *L. kluyveri* genome is estimated to share the mutational preference with the exogenous genes in $\sim$ 78% of the 19 codon families that are discordant between the endogenous and exogenous genes. In two cases, Isoleucine (Ile, I) and Arginine (Arg, R), the strong discordance in mutation preference results in an estimated codon preference in the complete *L. kluyveri* genome that differs from both the endogenous, and the exogenous genes.

The effect of the small exogenous gene set on the fit to the complete *L. kluyveri* genome is smaller in our estimates of selection bias $\Delta\eta$ than ΔM, but still large. We find that the complete *L. kluyveri* genome is estimated to share the selection preference with the exogenous genes in $\sim$ 60% of codon families that show discordance between endogenous and exogenous genes. These results clearly show that it is important to recognize the difference in endogenous and exogenous genes and treat these genes as separate sets to avoid the inference of incorrect synonymous codon preferences and better predict protein synthesis.

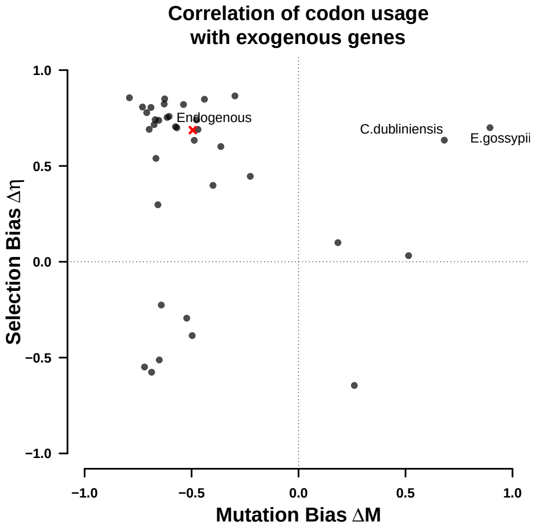


Figure 3: Correlation coefficients of ΔM and $\Delta\eta$ of the exogenous genes with 38 examined yeast lineages. Dots indicate the correlation of ΔM and $\Delta\eta$ of the lineages with the endogenous and exogenous parameter estimates. All regressions were performed using a type II regression (Sokal and Rohlf, 1981).

Determining Source of Exogenous Genes

We combined our estimates of mutation bias ΔM and selection bias $\Delta\eta$ with synteny information and searched for potential source lineages of the introgressed exogenous region. We examined 38 yeast lineages (Table S3) of which two (*Eremothecium gossypii* and *Candida dubliniensis*) showed a strong positive correlation in codon usage (Figure 3). The endogenous *L. kluyveri* genome exhibits codon usage very similar to most yeast lineages examined, indicating little variation in codon usage among the examined yeasts (Figure S1). Four lineages show a positive correlation for ΔM and $\Delta\eta$ with the exogenous genes and have a weak to moderate positive correlation in selection bias with the endogenous genes; but, like the exogenous genes, tend to have a negative correlation in ΔM with the endogenous genes.

Comparing synteny between the exogenous genes, which are restricted to the left arm of chromosome C, and *E. gossypii* and *C. dubliniensis* as well as closely related yeast species we find that *E. gossypii* displays the highest synteny (Figures S3 & S4). *C. dubliniensis*, even though it displays similar codon usage does not show synteny with the exogenous region. Furthermore, the synteny relationship between the exogenous region and other yeasts appears to be limited to the Saccharomycetacease clade (Figure S4). Given these results, we conclude that of the 38 examined yeast lineages the *E. gossypii* lineage is the most likely source of the introgressed exogenous genes.

Estimating Introgression Age

We modeled the change in codon frequency as a model of exponential decay, we estimated the age of the introgression assuming that *E. gossypii* still represents the mutation bias of its ancestral source lineage at the time of the introgression and a constant mutation rate. We infer the age of the introgression to be on the order of $6.2 \pm 1.2 \times 10^8$ generations. Assuming *L. kluyveri* experiences between one and eight generations per day, we estimate the introgression to have occurred between $212,000$ to $1,700,000$ years ago. Our estimate places the time of the introgression earlier than previously assumed (Friedrich *et al.*, 2015).

Using the same approach, we also estimated the persistence of the signal of the exogenous cellular environment. We assume that differences in mutation bias will decay more slowly than differences in selection bias to be able to utilize our bias free estimates of ΔM. We predict that the ΔM signal of the source cellular environment will have decayed to be within one percent of the *L. kluyveri* environment in $\sim 5.4 \pm 0.2 \times 10^9$ generations, or between $1,800,000$ and $15,000,000$ years. Together, these results indicate that the mutation signature of the exogenous genes will persist for a very long time.

Genetic Load due to Mismatching Codon Usage of the Exogenous Genes

We define genetic load as the difference between the fitness of an expected, replaced endogenous gene and the exogenous gene, $s \propto \phi\Delta\eta$ due to the mismatch in codon usage parameters (See Methods for details). Estimates of selection bias for the exogenous genes show that, while well correlated with the endogenous genes, only nine amino acids share the same optimal codon. Exogenous genes are, therefore, expected to represent a significant reduction in fitness, or genetic load for *L. kluyveri* due to this mismatch in codon usage. As the introgression occurred before the diversification of *L. kluyveri* and has fixed throughout all populations (Friedrich *et al.*, 2015), we can not observe the original endogenous sequences that have been replaced by the introgression. Using our estimates of ΔM and $\Delta\eta$ from the endogenous genes and assuming hat the current exogenous amino acid composition of genes is representative of the replaced endogenous genes, we estimate the genetic load of the exogenous genes at the time of introgression (Figure 4a) and currently (Figure 4b). We find that the genetic load due to mismatched codon usage was -0.0008 at the time of the introgression and still represents a genetic load of -0.0003 today.

In order to account for differences in the efficacy of selection on codon usage between the donor lineage and *L. kluyveri* using a linear scaling factor κ (See Methods for details). We predict that a small number of low expression genes ($\phi < 1$) were weakly exapted at the time of the introgression (Figure 4a). High expression genes ($\phi > 1$) are predicted to have carried the largest genetic load in the novel cellular environment. These highly expressed genes are inferred to have the greatest degree of adaptation since

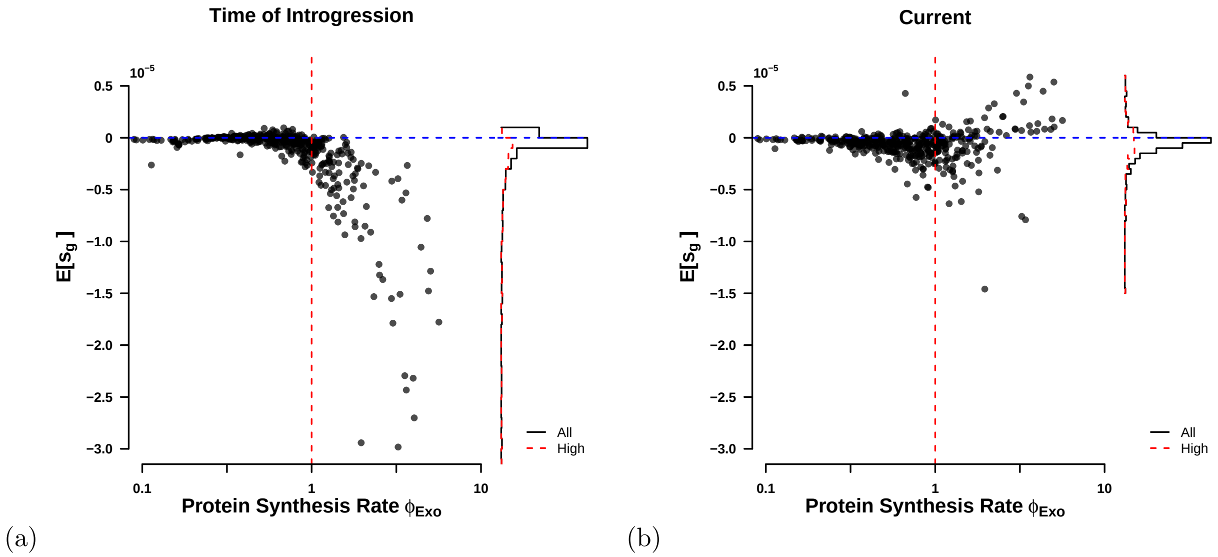


Figure 4: Genetic load $s = \Delta\eta\phi$ (a) at the time of introgression ($\kappa = 5$), and (b) currently ($\kappa = 1$).

the time of the introgression to the *L. kluyveri* cellular environment (Figures 4a & S6).

Discussion

In order to study the evolutionary effects of an introgression, we used ROC SEMPPR, a mechanistic model of ribosome movement along an mRNA. Our parameter estimates indicate that the *L. kluyveri* genome contains distinct signatures of mutation and selection bias from both an endogenous and exogenous cellular environment. By fitting ROC SEMPPR separately to *L. kluyveri*'s endogenous and exogenous sets of genes we generate a quantitative description of their signatures of mutation bias and natural selection for efficient protein translation. Our results indicate that the difference in GC content between endogenous and exogenous genes is mostly due to differences in mutation bias, but we also show that the strength and rank order of selection within a codon family differ between endogenous and exogenous cellular environments. Even though the exogenous genes make up only $\sim 10\%$ of the *L. kluyveri* genome, when we fail to recognize these differences our estimates of ΔM and $\Delta\eta$ deviate substantial from their actual values (Figure S2). While this sensitivity of our parameters to a second cellular environment may be surprising, it highlights the importance of recognizing different cellular environments reflected by a genome. Furthermore, our results indicate that we can attribute the increased GC content in the exogenous genes mostly to differences in mutation bias favoring G/C ending codons rather than selection.

The separation of the endogenous and exogenous genes improves our estimates of protein synthesis rate ϕ by 42% relative to the full genome estimate ($R^2 = 0.32$ vs. 0.46, respectively). Furthermore, failing to separately analyze the endogenous and exogenous genes results in an unrealistically small amount of intergenic variation in ϕ (compare Figure 1a & b). This behavior is due, in part, to constraining $E[\phi] = 1$ which allows us to compare the efficacy of selection sN_e across genomes. Extremely small variances in the ϕ values estimated by ROC SEMPPR could indicate that a genome contains the signature of multiple cellular environments.

The mutation and selection bias parameters ΔM and $\Delta\eta$ of the introgressed exogenous genes contain information, albeit decaying, about its previous cellular environment. We, therefore, utilize ΔM and $\Delta\eta$ to identify potential source lineages. The *E. gossypii* and *C. dubliniensis* lineages stand out from the other 36 yeast lineages in that the correlation coefficients between their ΔM and $\Delta\eta$ parameters and those of the exogenous genes are > 0.5 (Figure 2). In terms of gene order, we found that synteny with the exogenous genes is limited to the Saccharomycetaceae clade, which *C. dubliniensis* is outside of. Overall, the synteny coverage extends along the whole exogenous regions with the exception of the 3' and 5' ends of the exogenous region (Figure S4b). Further, of the 38 species examined, *E. gossypii* is the only genome with a GC content $> 50\%$, making it most similar to the exogenous genes. Thus, only the *E. gossypii* genome displays strong correlations in ΔM and $\Delta\eta$, synteny, and similar GC content with the exogenous genes.

With *E. gossypii* identified as potential source lineage of the introgressed region, we inferred the time since the introgression occurred using our estimates of mutation bias ΔM. Our ΔM estimates are well suited for this task as they are free of the influence of selection and unbiased by N_e and other scaling terms, which is in contrast to our estimates of $\Delta\eta$ (Gilchrist *et al.*, 2015). Our estimated age of the introgression of $6.2 \pm 1.2 \times 10^8$ generations is ~ 10 times longer time than a previous minimum estimate by Friedrich *et al.* (2015) of 5.6×10^7 generations. Our estimate assumes that the current *E. gossypii* and *L. kluyveri* cellular environment reflect their ancestral states at the time of the introgression. If the ancestral mutation environments were more similar (dissimilar) at the time of the introgression than now our result is an overestimate (underestimate).

In order to estimate the introgression's genetic load due to codon mismatch, we had to make three key assumptions: 1) at the time of introgression the amino acid sequences of the endogenous genes and exogenous genes where highly similar, 2) the current *L. kluyveri* cellular environment is reflective of the cellular environment at the time of the introgression, and 3) the *E. gossypii* cellular environment reflects its ancestral environment at the time of the introgression. In general due to their very nature, low expression genes contribute little to the genetic load. Indeed, $\sim 30\%$ of low expression exogenous genes ($\phi < 1$) appeared to be exapted at the time of the introgression. These exapted genes are likely due to the mutation bias in the endogenous genes matching the selection bias in the exogenous genes for G/C ending codons. In contrast, highly expressed genes are predicted to have imposed a large genetic load. Many of these genes appear to still represent a significant genetic load. Overall, our estimates of codon mismatch genetic load, therefore, suggest strong selection against the introgression.

It is hard to contextualize the probability of this introgression being fixed as we are not aware of any estimates of the frequency at which such large scale introgressions of genes occur. A related example of a large scale merger of genomic material can be found in *S. pastorianus*, which is currently believed to be a hybrid of *S. cerevisiae* and *S. eubayanus* lineages, (Baker *et al.*, 2015). Unlike with *L. kluyveri* and *E. gossypii*, the progenitor lineages of *S. pastorianus* have similar codon usage parameters. The correlation between ΔM and $\Delta\eta$ for these two lineages are $\rho = 0.83$ and 0.98 (data not shown). These similarities in ΔM and $\Delta\eta$ parameters suggest that the genetic load for *S. pastorianus* due to codon usage mismatch is small relative to the exongenous genes considered here. The large genetic load of the exogenous genes due to codon mismatch at the time of the introgression would seem to indicate that the fixation of the introgression was either a fluke event or the codon mismatch genetic load was countered by one or more highly advantageous loci within the introgression.

Under the first scenario, our best estimate of the selection coefficient against the introgression based on expected codon mismatch at that time is $s = -0.0008$ and an effective population size N_e on the order of 10^8 (Wagner, 2005) yields an approximate fixation probability of $(1-\exp[-s])/(1-\exp[2-sN_e]) \approx 10^{-6950}$ (Sella and Hirsh, 2005). Even though *L. kluyveri* diverged from the rest of the Lachancea clade around 85 Mya (Kensche *et al.*, 2008; Marcet-Houben and Gabaldn, 2015), if we assume 1 to 8 generations/day, which implies 10^{10} to 10^{11} generations since the time of divergence, one round of meiosis for every 1000 rounds of mitosis based on *S. paradoxus* (Tsai *et al.*, 2008), and $N_e \approx 10^8$ there were only 10^{15} to 10^{16} opportunities for such an introgression to have occurred and fixed. Clearly, unless there was a severe bottleneck with $N_e < 1/|s| \approx 1,250$ around the time of introgression, which conceivably could have been triggered by a speciation event, this scenario seems very unlikely.

In the second scenario, where we assume the introgression contained advantageous loci, one may wonder why recombination events did not limit the introgression to only the adaptive loci. Payen *et al.* (2009) found that the exogenous region has a lower rate of recombination, presumably due to the dissimilarity in GC content and/or a lower than average sequence homology between the exogenous region and the one it replaced. Compatible with this explanation is the possibility of several highly advantageous loci distributed across the region which then drove a rapid selective sweep and/or the population through a bottleneck speciation process A careful analysis of intra-specific genetic variation within the endogenous and exongenous regions could provide help us distinguish between these various scenarios.

Overall, our results show the usefulness of the separation of mutation bias and selection bias and the importance of recognizing the presence of multiple cellular environments in the study of codon usage. We also illustrate how a mechanistic model like ROC SEMPPR and the quantitative estimates it provides can be used for more sophisticated hypothesis testing in the future. In contrast to other approaches used to study codon usage like CAI (Sharp and Li, 1987) or tAI (dos Reis *et al.*, 2004), ROC SEMPPR incorporates the effects of mutation bias and amino acid composition explicitly Cope *et al.* (2018). We highlight potential issues when estimating codon preferences, as estimates can be biased by the signature of a second, historical cellular environment. In addition, we show how quantitative estimates of mutation bias and selection relative to drift can be obtained from codon data and used to infer the fitness cost of an introgression as well as its history and potential future.

Materials and Methods

Separating Endogenous and Exogenous Genes

A GC-rich region was identified by Payen *et al.* (2009) in the *L. kluyveri* genome extending from position 1 to 989,693 of chromosome C. This region was later identified as an introgression by Friedrich *et al.* (2015). We obtained the *L. kluyveri* genome from SGD Project `http://www.yeastgenome.org/download-data/` (on 09-27-2014) and the annotation for *L. kluyveri* NRRL Y-12651 (assembly ASM14922v1) from NCBI (on 12-09-2014). We assigned 457 genes located on chromosome C with a location within the $\sim 1Mb$ window to the exogenous gene set. All other 4864 genes of the *L. kluyveri* genome were assigned to the exogenous genes. All genes could be uniquely assigned to one or the other gene set.

Model Fitting with ROC SEMPPR

ROC SEMPPR was fitted to each genome using AnaCoDa (0.1.1) (Landerer *et al.*, 2018) and R (3.4.1) (R Core Team, 2013). ROC SEMPPR was run from multiple starting values for at least 250,000 iterations, only every 50th step was collected as a sample to reduce autocorrelation. After manual inspection to verify that the MCMC had converged, parameter posterior means were estimated from the last 500 samples.

Comparing Codon Specific Parameter Estimates

Choice of reference codon does reorganize codon families coding for an amino acid relative to each other, therefore all parameter estimates are relative to the mean for each codon family.

$$\Delta M^c_{i,a} = \Delta M_{i,a} - \overline{\Delta M_a} \tag{1}$$

$$\Delta\eta^{c}_{i,a} = \Delta\eta_{i,a} - \overline{\Delta\eta_a} \tag{2}$$

Comparison of codon specific parameters (ΔM and $\Delta\eta = 2N_e q(\eta_i - \eta_j)$) was performed using the function lmodel2 in the R package lmodel2 (1.7.3) (Legendre, 2018) and R version 3.4.1 (R Core Team, 2013). The parameter $\Delta\eta$ can be interpreted as the difference in fitness between codon i and j for the average gene with $\phi = 1$ scaled by the effective population size N_e, and the selective cost of an ATP q (Gilchrist, 2007; Gilchrist *et al.*, 2015). Type II regression was performed with re-centered parameter estimates, accounting for noise in dependent and independent variable (Sokal and Rohlf, 1981).

Synteny Comparison

We obtained complete genome sequences from NCBI (on: 02-05-2017). Genomes were aligned and checked for synteny using SyMAP (4.2) with default settings (Soderlund *et al.*, 2006, 2011). We assess synteny as percentage coverage of the exogenous gene region (Figure S4b).

Estimating Age of Introgression

We modeled the change in codon frequency over time using an exponential model for all two codon amino acids, and describing the change in codon c_1 as

$$\frac{dc_1}{dt} = -\mu_{1,2}c_1 - \mu_{2,1}(1 - c_1) \tag{3}$$

where $\mu_{i,j}$ is the rate at which codon i mutates to codon j and c_1 is the frequency of the reference codon. Our estimates of ΔM_{endo} can be used to calculate the steady state of equation 3.

$$\frac{\mu_{2,1}}{\mu_{1,2} + \mu_{2,1}} = \frac{1}{1 + \exp[\Delta M_{\text{endo}}]} \tag{4}$$

Solving for $\mu_{1,2}$ gives us $\mu_{1,2} = \Delta M_{\text{endo}} \exp[\mu_{2,1}]$ which allows us to rewrite and solve equation 3 as

$$c_1(t) = \frac{\exp[-t(1 + \Delta M_{\text{endo}})\mu_{2,1}]\exp[t(1 + \Delta M_{\text{endo}})\mu_{2,1}] + (1 + \Delta M_{\text{endo}})K}{1 + \Delta M_{\text{endo}}} \tag{5}$$

where K is

$$K = c_1(0) - \frac{1}{1 + \Delta M_{\text{endo}}} \tag{6}$$

Equation 5 was solved with a mutation rate $m_{2,1}$ of 3.8×10^{-10} per nucleotide per generation (Lang and Murray, 2008). Initial codon frequencies $c_1(0)$ for each codon family where taken from our mutation parameter estimates for *E. gossypii* ΔM_{gos}. Current codon frequencies for each codon family where taken from our estimates of ΔM from the exogenous genes. Mathematica (11.3) (Wolfram Research Inc., 2017) was used to calculate the time t_{intro} it takes for the initial codon frequencies $c_1(0)$ for each codon family to equal the current exogenous codon frequencies. The same equation was used to determine the time t_{decay} at which the signal of the exogenous cellular environment has decayed to within 1% of the endogenous environment.

Estimating Genetic Load

To estimate the genetic load due to mismatched codon usage, we made three key assumptions. First, we assumed that the current exogenous amino acid sequence of a gene is representative of its ancestral state and the replaced endogenous gene it replaced. Second, we assume that the currently observed cellular environment of *E. gossypii* reflects the cellular environment that the exogenous genes experienced before transfer to *L. kluyveri*. Lastly, we assume that the difference in the efficacy of selection between the cellular environments due to differences in either effective population size N_e or the selective cost of an ATP q of the source lineage and *L. kluyveri* can be expressed as a scaling constant and that protein synthesis rate ϕ has not changed between the replaced endogenous and the introgressed exogenous genes. Using estimates for $N_e = 1.36 \times 10^7$ (Wagner, 2005) for *Saccharomyces paradoxus* we scale our estimates of $\Delta\eta$ and define $\Delta\eta' = \frac{\Delta\eta}{N_e}$.

We scale the difference in the efficacy of selection on codon usage between the donor lineage and *L. kluyveri* using a linear scaling factor κ. As $\Delta\eta$ is defined as $\Delta\eta = 2N_e q(\eta_i - \eta_j)$, we can not distinguish if κ is a scaling on protein synthesis rate ϕ, effective population size N_e, or the selective cost of an ATP q (Gilchrist, 2007; Gilchrist *et al.*, 2015). We calculated the genetic load each gene represents due to its mismatched codon usage assuming additive fitness effects as

$$s_g = \sum_{i=1}^{n_g} -\kappa \phi_g \Delta\eta'_i \tag{7}$$

where s_g is the overall strength of selection for translational efficiency on gene, g in the exogenous gene set, κ is a constant, scaling the efficacy of selection between the endogenous and exogenous cellular environments, n_g is length of the protein, ϕ_g is the estimated protein synthesis rate of the gene in the endogenous environment, and $\Delta\eta'_i$, is the $\Delta\eta$' for the codon at position i. As stated previously, our $\Delta\eta$ are relative to the mean of the codon family. We find that the genetic load of the introgressed genes is minimized at $\kappa \sim 5$ (Figure S5b). Thus, we expect a five fold difference in the efficacy of selection between *L. kluyveri* and *E. gossypii*, either due to differences in either protein synthesis rate ϕ, effective population size N_e, or the selective cost of an ATP q. Therefore, we set $\kappa = 1$ if we calculate the s_g for the endogenous and the current exogenous genes, and $\kappa = 5$ for s_g for the genetic load at the time of introgression.

Since we are unable to observe codon counts for the replaced endogenous genes and for the exogenous genes at the time of introgression, we calculate expected codon counts

$$E[n_{g,i}] = \frac{\exp[-\Delta M_i - \Delta\eta_i\phi_g]}{\sum_j^C \exp[-\Delta M_j - \Delta\eta_j\phi_g]} \times m_{a_i} \tag{8}$$

m_{a_i} is the number of occurrences of amino acid a that codon i codes for. We report the genetic load due to mismatched codon usage of the introgression as $E[s_g] = s_{\text{intro},g} - s_{\text{endo},g}$ where $s_{\text{intro},g}$ is the genetic load of an introgressed gene g either at the time of the introgression or presently.

Acknowledgments

This work was supported in part by NSF Awards MCB-1120370 (MAG and RZ) and DEB-1355033 (BCO, MAG, and RZ) with additional support from The University of Tennessee Knoxville. CL received support as a Graduate Student Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award DBI-1300426, with additional support from UTK. The authors would like to thank Alexander Cope for helpful criticisms and suggestions for this work.

References

Baker, E., Wang, B., Bellora, N., *et al.* 2015. The genome sequence of saccharomyces eubayanus and the domestication of lager-brewing yeasts. *Molecular Biology and Evolution*, 32(11): 2818–2831.

Beimforde, C., Feldberg, K., Nylinder, S., *et al.* 2014. Estimating the phanerozoic history of the ascomycota lineages: combining fossil and molecular data. *Mol. Phylogenet. Evol.*, 78: 386–398.

Bulmer, M. 1990. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129: 897–907.

Cope, A., Hettich, R., and Gilchrist, M. 2018. Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1860(12): 2479–2485.

dos Reis, M., Savva, R., and Wernisch, L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research*, 32(17): 5036–5044.

Friedrich, A., Reiser, C., Fischer, G., and Schacherer, J. 2015. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Molecular Biology and Evolution*, 32(1): 184 – 192.

Gilchrist, M. 2007. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Molecular Biology and Evolution*, 24(11): 2362–2372.

Gilchrist, M., Chen, W., Shah, P., Landerer, C., and Zaretzki, R. 2015. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. *Genome Biology and Evolution*, 7: 1559–1579.

Gouy, M. and Gautier, C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, 10: 7055–7074.

Ikemura, T. 1985. Codon usage and trna content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, 2: 13–34.

Kensche, P., Oti, M., Dutilh, B., and Huynen, M. 2008. Conservation of divergent transcription in fungi. *Trends Genet.*, 5(24): 207–211.

Landerer, C., Cope, A., Zaretzki, R., and Gilchrist, M. A. 2018. Anacoda: analyzing codon data with bayesian mixture models. *Bioinformatics*, 34(14): 2496–2498.

Lang, G. I. and Murray, A. W. 2008. Estimating the per-base-pair mutation rate in the yeast saccharomyces cerevisiae. *Genetics*, 178(1): 67 – 82.

Lawrence, J. and Ochman, H. 1997. Amelioration of bacterial genomes: Rates of change and exchange. *Journal of Molecular Miology*, 44: 383–397.

Legendre, P. 2018. *lmodel2: Model II Regression*. R package version 1.7-3.

Marcet-Houben, M. and Gabaldn, T. 2015. Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology*, 13(8): e1002220.

Mdigue, C., Rouxel, T., Vigier, P., Hnaut, A., and Danchin, A. 1991. Evidence for horizontal gene transfer in escherichia coli speciation. *Journal of Molecular Miology*, 222(4): 851–856.

Payen, C., Fischer, G., Marck, C., *et al.* 2009. Unusual composition of a yeast chromosome arm is associated with its delayed replication. *Genome Research*, 19(10): 1710–1721.

R Core Team 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sella, G. and Hirsh, A. 2005. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102: 9541–9546.

Shah, P. and Gilchrist, M. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proccedings of the National Academy of Sciences U.S.A*, 108(25): 10231–10236.

Sharp, P. and Li, W. 1987. The codon adaptation index - a meassure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15: 1281–1295.

Soderlund, C., Nelson, W., Shoemaker, A., and Paterson, A. 2006. Symap A system for discovering and viewing syntenic regions of fpc maps. *Genome Research*, 16: 1159 – 1168.

Soderlund, C., Bomhoff, M., and Nelson, W. 2011. Symap v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Research*, 39(10): e68.

Sokal, R. and Rohlf, F. 1981. *Biometry - The principles and practice of statistics in biological*, pages 547–555. W. H. Freeman.

Tsai, I., Bensasson, D., Burt, A., and Koufopanou, V. 2008. Population genomics of the wild yeast saccharomyces paradoxus: quantifying the life cycle. *Proc Natl Acad Sci U.S.A.*, 105: 4957–4962.

Tsankov, A., Thompson, D., Socha, A., Regev, A., and Rando, O. 2010. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol*, 8(7): e1000414.

Wagner, A. 2005. Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution*, 22: 1365–1374.

Wallace, E., Airoldi, E., and Drummond, D. 2013. Estimating selection on synonymous codon usage from noisy experimental data. *Molecular Biology and Evolution*, 30: 1438–1453.

Wolfram Research Inc. 2017. *Mathematica 11*.

Supplementary Material

Supporting Materials for *Decomposing Mutation and Selection to Identify Mismatched Codon Usage* by Landerer *et al.*.

Table S1: Synonymous codon preference in the various data sets based on our estimates of ΔM

Amino Acid	*E. gossypii*	Endogenous	Exogenous	*L. kluyveri*
Ala A	GCG	GCA	GCG	GCG
Cys C	TGC	TGT	TGC	TGC
Asp D	GAC	GAT	GAC	GAC
Glu E	GAG	GAA	GAG	GAG
Phe F	TTC	TTT	TTT	TTT
Gly G	GGC	GGT	GGC	GGC
His H	CAC	CAT	CAC	CAC
Ile I	ATC	ATT	ATC	ATA
Lys K	AAG	AAA	AAG	AAA
Leu L	CTG	TTG	CTG	CTG
Asn N	AAC	AAT	AAC	AAT
Pro P	CCG	CCA	CCG	CCG
Gln Q	CAG	CAA	CAG	CAG
Arg R	CGC	AGA	AGG	CGG
Ser_4 S	TCG	TCT	TCG	TCG
Thr T	ACG	ACA	ACG	ACG
Val V	GTG	GTT	GTG	GTG
Tyr Y	TAC	TAT	TAC	TAC
Ser_2 Z	AGC	AGT	AGC	AGC

Table S2: Synonymous codon preference in the various data sets based on our estimates of $\Delta\eta$

Amino Acid	*E. gossypii*	Endogenous	Exogenous	*L. kluyveri*
Ala A	GCT	GCT	GCT	GCT
Cys C	TGT	TGT	TGT	TGT
Asp D	GAT	GAC	GAT	GAT
Glu E	GAA	GAA	GAA	GAA
Phe F	TTT	TTC	TTC	TTC
Gly G	GGA	GGT	GGT	GGT
His H	CAT	CAC	CAT	CAT
Ile I	ATA	ATC	ATT	ATT
Lys K	AAA	AAG	AAA	AAG
Leu L	TTA	TTG	TTG	TTG
Asn N	AAT	AAC	AAT	AAC
Pro P	CCA	CCA	CCT	CCA
Gln Q	CAA	CAA	CAA	CAA
Arg R	AGA	AGA	AGA	AGA
Ser$_4$ S	TCA	TCC	TCT	TCT
Thr T	ACT	ACC	ACT	ACT
Val V	GTT	GTC	GTT	GTT
Tyr Y	TAT	TAC	TAT	TAC
Ser$_2$ Z	AGT	AGT	AGT	AGT

Table S3: Overview of yeast lineages used in this study.

Taxon	Abbreviation	NCBI taxonomic ID	Codon Table	% GC
Candida albicans	Calb	5476	12	34
Saccharomyces bayanus	Sbay	4931	1	40
Trichophyton benhamiae	Tben	63400	1	49
Tetrapisispora blattae	Tbla	1071379	1	32
Saccharomyces castellii	Scas	27288	1	37
Saccharomyces cerevisiae	Scer	4932	1	38
Eremothecium cymbalariae	Ecym	45285	1	40
Torulaspora delbrueckii	Tdel	4950	1	42
Candida dubliniensis	Cdub	42374	12	33
Lodderomyces elongisporus	Lelo	36914	1	37
Saccharomyces eubayanus	Seub	1080349	1	40
Debaryomyces fabryi	Dfab	58627	1	36
Candida glabrata	Cgla	5478	1	39
Eremothecium gossypii	Egos	33169	1	52
Meyerozyma guilliermondii	Mgui	4929	12	44
Debaryomyces hansenii	Dhan	4959	12	36
Lachancea kluyveri	Lku	4934	1	40/53
Saccharomyces kudriavzevii	Skud	114524	1	41
Kluyveromyces lactis	Klac	28985	1	39
Lachancea lanzarotensis	Llan	1245769	1	44
Yarrowia lipolytica	Ylip	4952	1	49
Clavispora lusitaniae	Clus	36911	12	45
Kluyveromyces marxianus	Kmar	4911	1	40
Saccharomyces mikatae	Smik	114525	1	38
Sphaerulina musiva	Smus	85929	1	51
Kazachstania naganishii	Knag	588726	1	46
Saccharomyces paradoxus	Spar	27291	1	38
Candida parapsilosis	Cpar	5480	12	38
Spathaspora passalidarum	Spas	340170	12	38
Tetrapisispora phaffii	Tpha	113608	1	34
Vanderwaltozyma polyspora	Vpol	36033	1	33
Lachancea quebecensis	Lque	1654605	1	47
Zygosaccharomyces rouxii	Zrou	4956	1	40
Scheffersomyces stipitis	Ssti	4924	12	41
Lachancea thermotolerans	Lthe	381046	1	47
Candida tropicalis	Ctro	5482	12	33
Lachancea waltii	Lwal	4914	1	44
Cladophialophora yegresii	Cyeg	470704	1	54

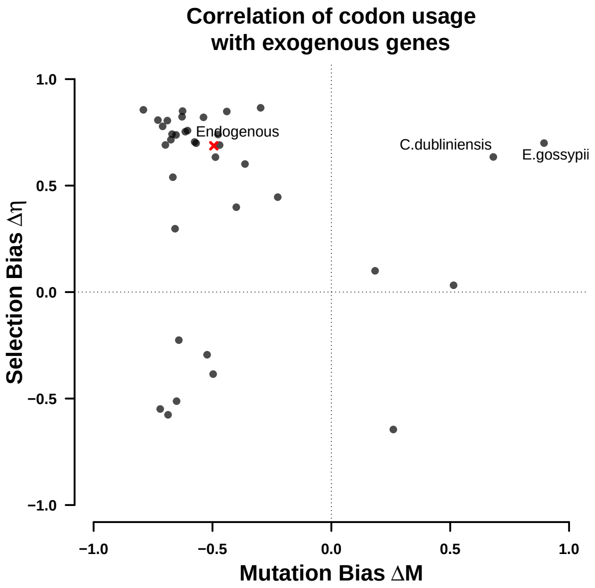


Figure S1: Correlation coefficient of ΔM and $\Delta\eta$ of the endogenous genes with 38 examined yeast lineages. Dots indicate the correlation of ΔM and $\Delta\eta$ of the lineages with the endogenous and exogenous parameter estimates. All regressions were performed using a type II regression line (Sokal and Rohlf, 1981).

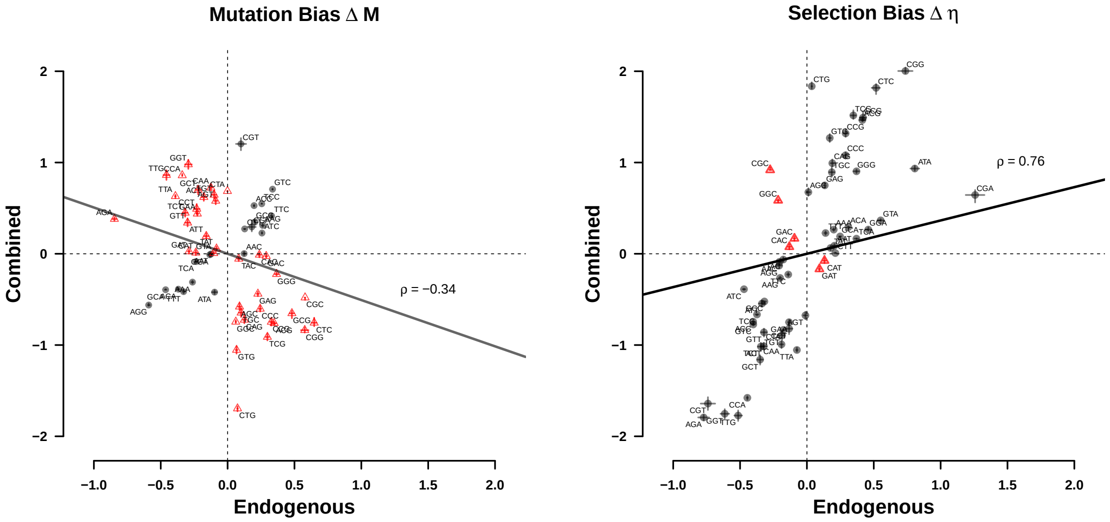


Figure S2: Comparison of (a) mutation bias ΔM and (b) selection bias $\Delta\eta$ parameters for endogenous genes and combined gene sets. Estimates are relative to the mean for each codon family. Black dots indicate ΔM or $\Delta\eta$ parameters with the same sign for the endogenous and exogenous genes, red dots indicate parameters with different signs. Black line shows the type II regression line (Sokal and Rohlf, 1981). Dashed lines mark quadrants.

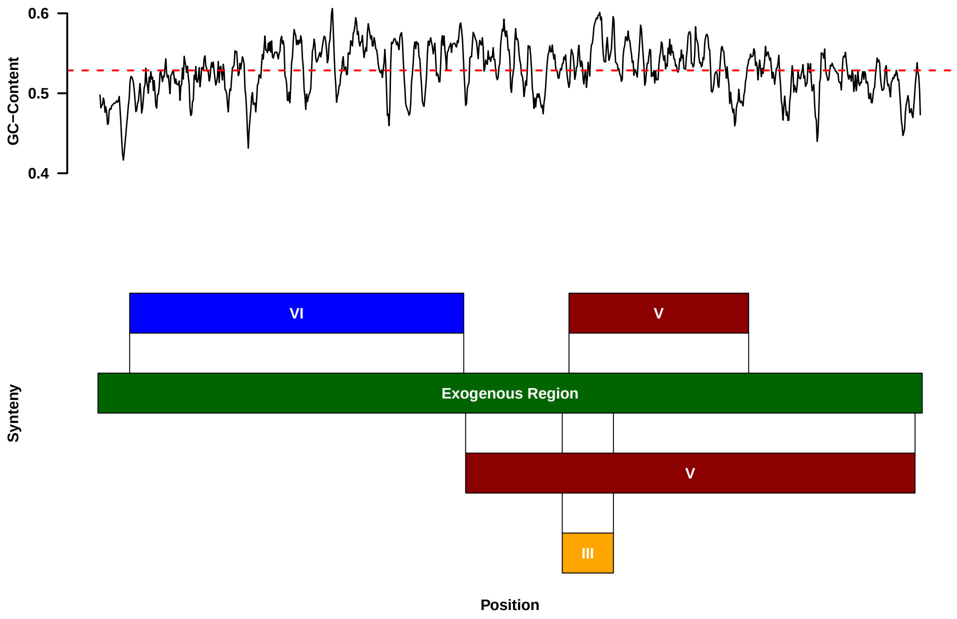


Figure S3: Synteny relationship of *E. gossypii* and the exogenous genes. Indicated is the GC content along the introgression.

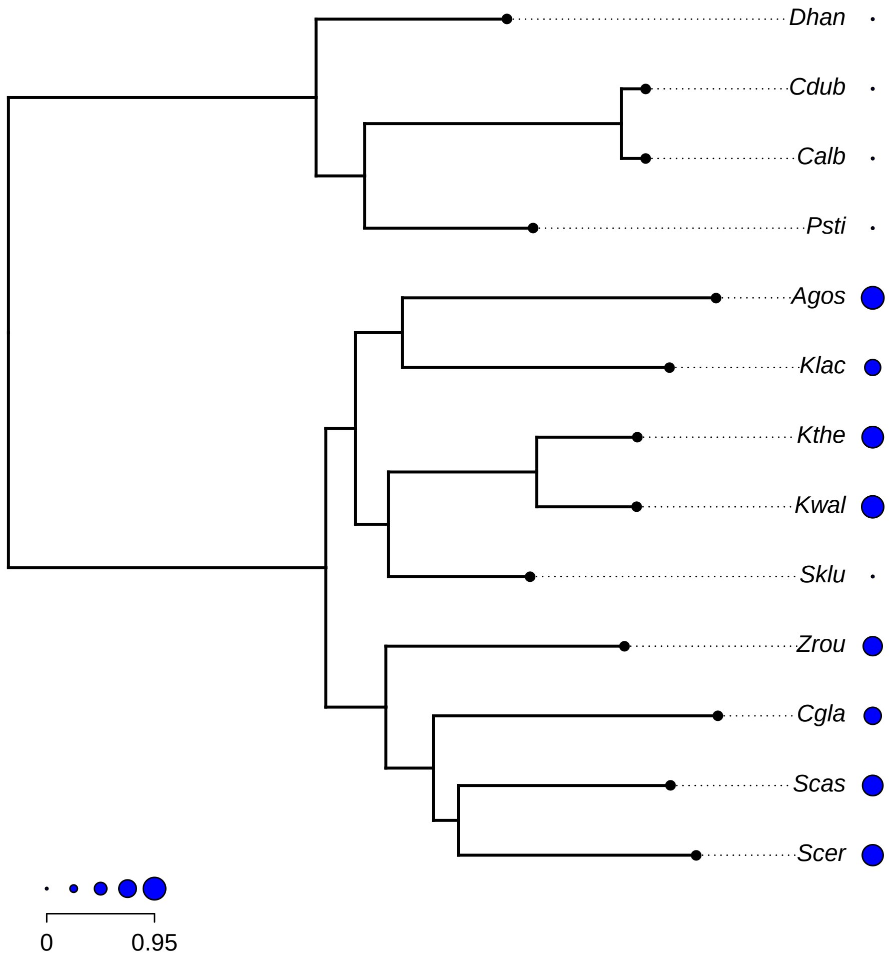


Figure S4: Amount of synteny for each species in units of standard deviations for selected species.

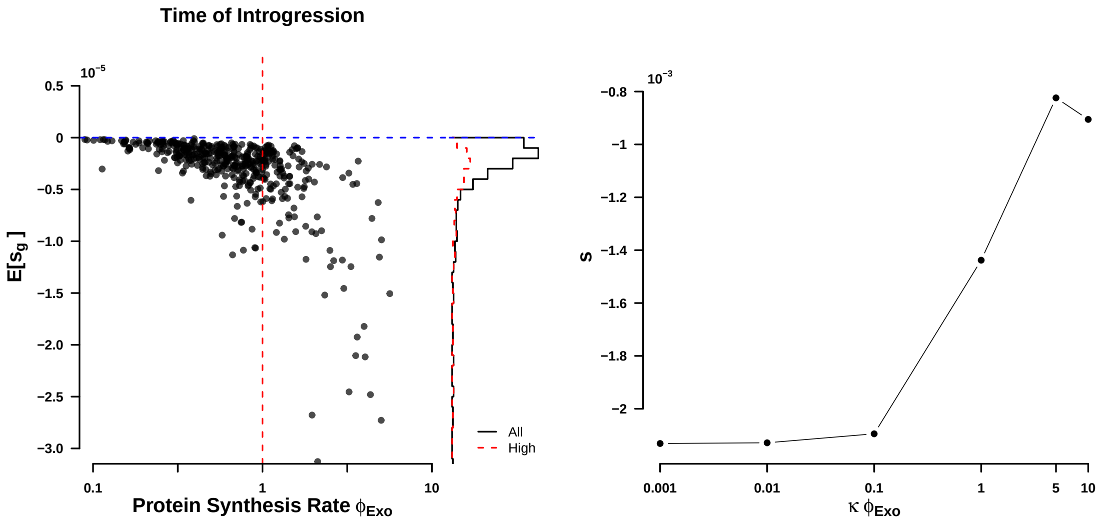


Figure S5: Genetic load (left) without scaling of ϕ per gene, and change of total genetic load with scaling κ between *E. gossypii* and *L. kluyveri* (right)

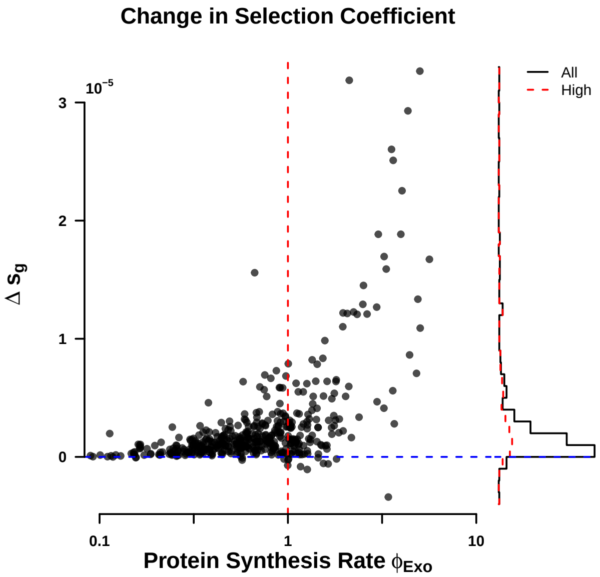


Figure S6: Total amount of adaptation estimated to have occurred between time of introgression and currently observed per gene.

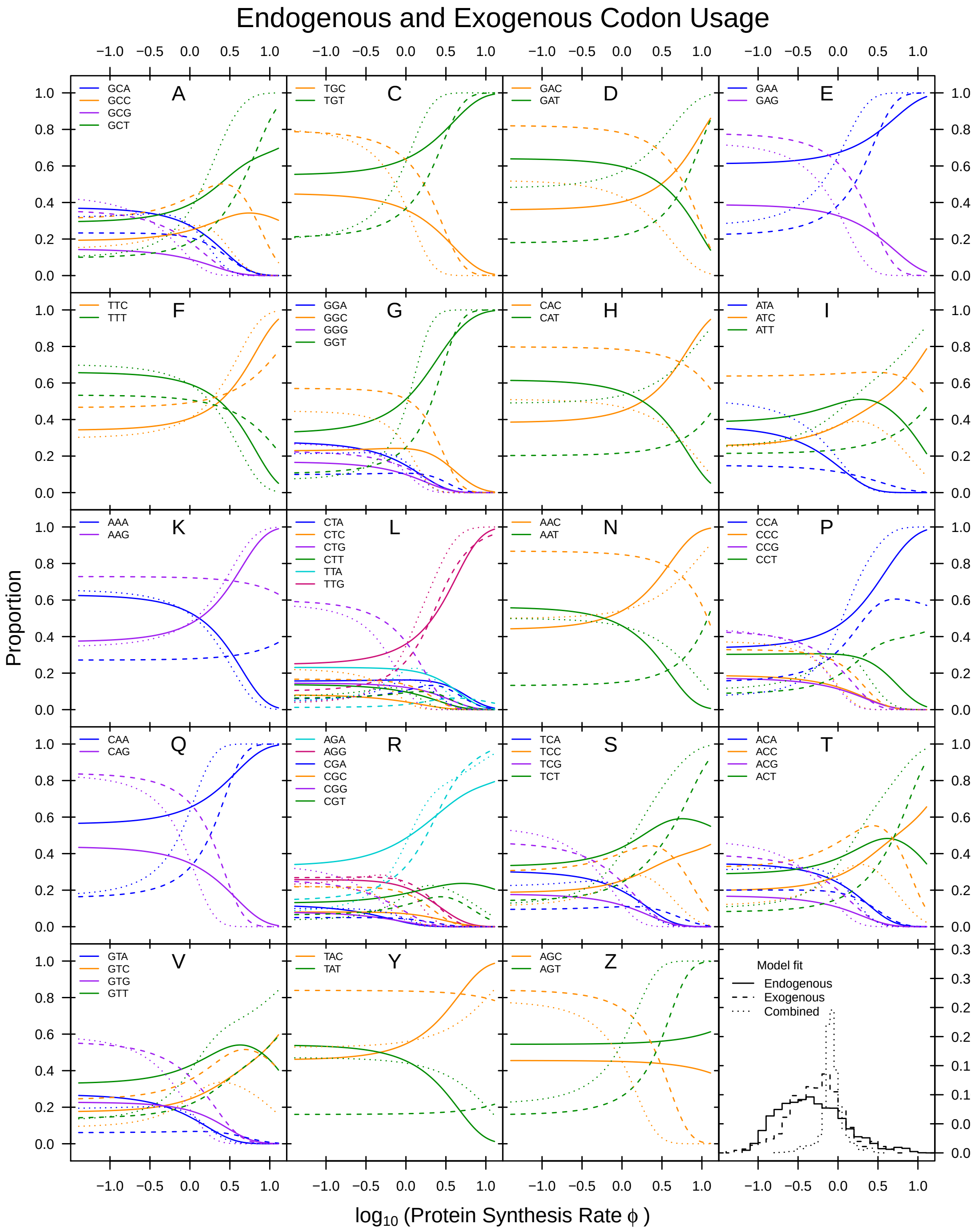


Figure S7: Codon usage patterns for 19 amino acids. Amino acids are indicated as one letter code. The amino acids Serine was split into two groups (S and Z) as Serine is coded for by two groups of codons that are separated by more than one mutation. Solid line indicates the endogenous codon usage, dashed line indicates the exogenous codon usage, dotted line indicates the combined codon usage.