Decomposing mutation and selection to identify mismatched codon usage

CEDRIC LANDERER[1,2,*], RUSSELL ZARETZKI[3], AND MICHAEL A. GILCHRIST[1,2]

[1]Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610

[2]National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996-3410

[3]Department of Business Analytics & Statistics, Knoxville, TN 37996-0532

[*]Corresponding author. E-mail: cedric.landerer@gmail.com

Version dated: October 12, 2018

Abstract

Here we examine variation in codon usage patterns of endogenous and exogenous genes in the yeast *Lachancea kluyveri*. Previous studies indicate that the left arm of chromosome C, or $\sim$ 10% of the *L. kluyveri* genome, is the result of an large introgression of exogenous genes. Thus, the *L. kluyveri* genome provides an opportunity to study the adaptation of these exogenous to a novel cellular environment and estimate how the genetic load of these genes changes over time. In order to quantiatively describe *L. kluyveri*'s codon usage environment, we fitted a bayesian, mechanistic model of codon usage bias evolution, ROC-SEMPPR, to *L. kluyveri*'s endogenous gene in order to estimate the strength of mutation bias and selection on codon usage. We then compared these parameter estimates to those we obtained by fitting ROC-SEMPPR to *L. kluyveri*'s exogenous genes, which provides a biased estimate of the ancestral environment of the exogenous genes. Our results indicate the differences in codon usage between *L. kluyveri*'s endogenous and exogenous genes are largely due to differences in mutation bias, rather than selection. Estimating mutation and selection parameters separately for the endogenous and exogenous genes improved our ability to predict empirical estimates of protein synthesis by 17% and avoided errors in identifying *L. kluyveri*'s selectively favored or 'optimal' codons. By comparing our mutation and selection parameters to those estimated for other yeast species, we identified *Eremothecium gossypii* as the most likely source of *L. kluyveri*'s exogenous genes. Using these parameters and available estimates of mutation rates in yeast, we estimated the age of the introgression to be on the order of 6×10^8 generation. Finally, we estimated the genetic load of the exogeneous genes both at the time of introgression and currently. In summary, our work shows the advantage of using mechanistic models that separate the effects of selection and mutation on codon usage.

Introduction

Synonymous codon usage patterns often 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. 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. In contrast, 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 turn, the 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. Thus, disentangling the evolutionary forces responsible for the patterns codon usage bias (CUB) encoded in an species genome, should provide biologically meaningful information about the lineage's historical cellular and external environment.

In order to disentangle the forces of mutation, selection, and drift behind CUB we utilize a quantiative, population genetics based approach after Bulmer [1991]. More specifically, we utilize the Ribosome Overhead Cost (ROC) version of Shah and Gilchrist [2011] of the more general Stochastic Evolutionary Model of Protein Production Rates (SEMPPR) introduced in Gilchrist [2007] using the R software package AnaCoDaLanderer et al. [2018]. The population genetics mutation-selection-drift framework of ROC SEMPPR allows us to quantiatively describe the environment in which genes evolve with respect to mutation bias and selection bias, which are the codon specific selection terms implicitly scaled by N_e and explicitly scaled by the average expression level of a gene [See Gilchrist et al., 2015a, for more details], using only coding sequenced data. Here we expand upon our previous work with ROC to accommodate the additional complications of gene introgression.

Most studies implicitly assume that synonymous codon usage of a genome is reflects the single mutational and selective cellular environment of the organism. However, any

introgressed genes, whether the result of hybridization or horizontal gene transfer, should carry the signature of the exogenous cellular environment whence they came and, in turn, impose a genetic load on the recipient lineage. The magnitude of the exogenous genes' genetic load on the recipient or endogenous lineage should increase as the mutation and selective environments differ between the donor and recipient lineages as well as with the expression level of the genes in the recipient cells. Thus codon usage patterns likely play a critical role in the rates of introgression between lineages and, as a result, can serve as an important source of information about such events.

To illustrate these ideas, here we analyze the synonymous codon usage of the genome of *Lachancea kluyveri*, the earliest diverging lineage of the Lachancea clade. The Lachancea clade diverged from the Saccharomyces clade about 100 Mya ago, predating Saccharomyces most recent genome duplication. Since its divergence from the other Lachancea, *L. kluyveri* has experienced a large introgression of exogenous genes, replacing the $\sim$ 500 on the left arm of *L. kluyveri*'s C chromosome. This introgression of exogenous genes was previously identified by its $\sim$ 13% elevation in GC content content relative to *L. kluyveri*'s remaining $\sim 5,000$ endogenous genes [Payen et al., 2009, Friedrich et al., 2015]. Taking into account the different signatures of mutation bias and selection bias of these endogenous and exogenous sets of genes substantially improves our ability to predict present day protein synthesis rate ϕ. It also 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, hypothesize about the genetic load of these genes, both at the time of introgression and now, as well as make predictions about the CUB of the introgressed genes will evolve in the future.

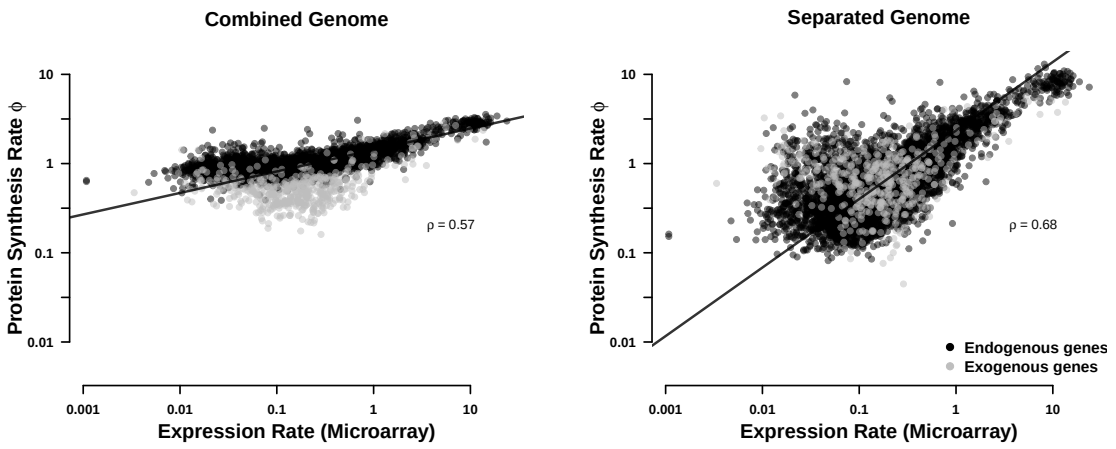

Figure 1: Comparison of predicted protein synthesis rate ϕ to Microarray data from Tsankov et al. [2010] for (a) the combined genome and (b) the separated endogenous and exogenous genes. Endogenous genes are displayed in black and exogenous genes in red. Black line indicates type II regression line.

Results

L. kluyveri's Genome Contains Signatures from Two Cellular Environments

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 separated into two sets of 4,864 endogenous and 497 exogenous genes. AIC values ($\Delta\text{AIC} = 75,462$; Table 1) strongly support the hypothesis that the *L. kluyveri* genome consists of genes with two different and distinct patterns of codon usage bias. We found additional support for this hypothesis when we compared our predictions of gene expression to empirically observed values. Specfically, the correlation between our predictions and observed values improved by almost 20%, from $\rho = 0.57$ to 0.68 (Figure 1).

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) has the same rank order for the endogenous and exogenous ΔM values.

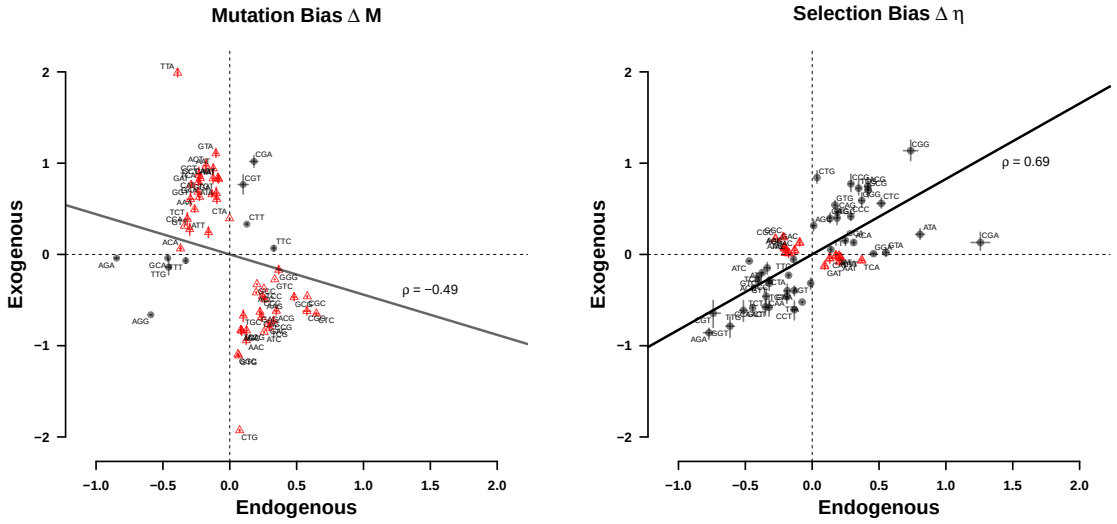

Figure 2: Comparison of (a) mutation bias ΔM and (b) selection bias $\Delta\eta$ parameters for endogenous and exogenous genes. Estimates are relative to the mean for each codon family. Black dots indicate ΔM or $\Delta\eta$ parameters with the same sign for the endogenous and exogenous genes, red dots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

Our estimates of $\Delta\eta$ for the endogenous and exogenous genes were positively correlated ($\rho = 0.69$), indicating increased concordance of $\sim 53\%$ between the two selection environments (Figure 2). Nevertheless, 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 preference between endogenous and exogenous genes is striking. Fits to the complete *L. kluyveri* genome reveal that the relatively small exogenous gene set 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 codon families with discordance between endogenous and exogenous genes. In two cases, Isoleucine (Ile, I) and Arginine (Arg, R), the strong discrodance in mutation preference results in a estiamted codon preference in the complete *L. kluyveri* genome that is not reflected by either endogenous nor exogenous genes.

The impact of the small exogenous gene set on the fit to the complete *L. kluyveri* genome is less prevalent in our estimates of selection bias $\Delta\eta$ but still strong. 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 with discordance between endogenous and exogenous genes. Therefore, it is important to recognize and treat endogenous and exogenous genes as separate sets to avoid the inference of incorrect synonymous codon preferences.

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 region. We examined 38 yeast lineages of which two (*Eremothecium gossypii* and *Candida dubliniensis*) showed a strong positive correlation in codon usage (Figure 3).

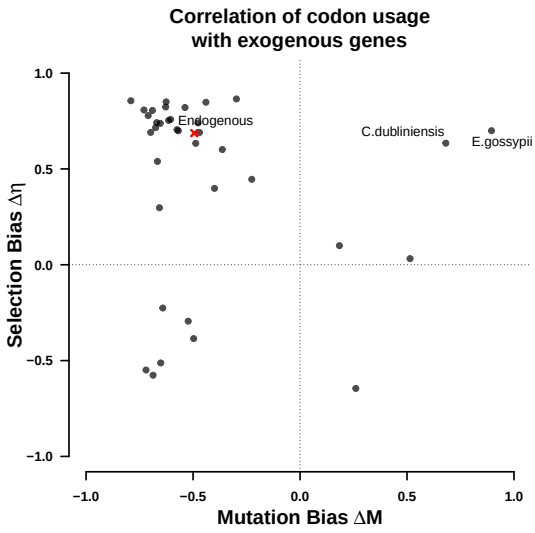

Figure 3: Correlation of ΔM and $\Delta\eta$ of the exogenous genes with 38 examined yeast lineages. Dots indicate the correlation of ΔM and $\Delta\eta$ of the lineages with the endogenous and exogenous parameter estimates. All regressions were performed using a type II regression.

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 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 coverage (Figure S2, S3). *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 group(Figure S3). Given these results, we conclude that the *E. gossypii* lineage is the most likely source of the introgressed exogenous genes.

Estimating Introgression Age

We estimated the introgression assuming that *E. gossypii* is still representative of the mutation bias of its ancestral source lineage at the time of the introgression. We infer the age of the introgression to be on the order of $6.2 \pm 1.2 \times 10^8$ generations. *L. kluyveri* experiences between one and eight generations per day, we therefore expect the introgression to have occurred between $205,000$ to $1,600,000$ years ago. This estimates the introgression to be older than previous assumed Friedrich et al. [2015]. However, our estimates are likely overestimates as they assume a purely neutral decay.

We also estimated the persistence of the signal of the foreign cellular environment. Assuming that differences in mutation bias will decay more slowly than differences in selection bias, we predict that the ΔM signal of the source cellular environment will have decayed to be within one percent of the *L. kluyveri* environment within about $5.4 \pm 0.2 \times 10^9$ generations.

Genetic Load of the Exogenous Genes

Estimates of selection bias for the exogenous genes show that, while well correlated with the endogenous genes, only nine amino acids share the optimal codon. We therefore expect that the introgressed genes represent a significant reduction in fitness, or genetic load for *L. kluyveri*, and even more so at the time of introgression. As the introgression occurred before the diversification of *L. kluyveri* and has fixed since then throughout the various populations, we are left without the original chromosome arm [Friedrich et al., 2015]. Using our estimates of ΔM and $\Delta\eta$ from the endogenous genes, we can estimate the genetic load of the exogenous genes relative to an expected gene set. We define genetic load as the difference between the fitness of an expected, replaced endogenous gene and the inferred introgressed gene relative to drift $s \propto \phi\Delta\eta$ (See Methods for details).

We estimate the genetic load of the exogenous genes at the time of introgression (Figure 4a) and currently (Figure 4b). 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 the value of an ATP

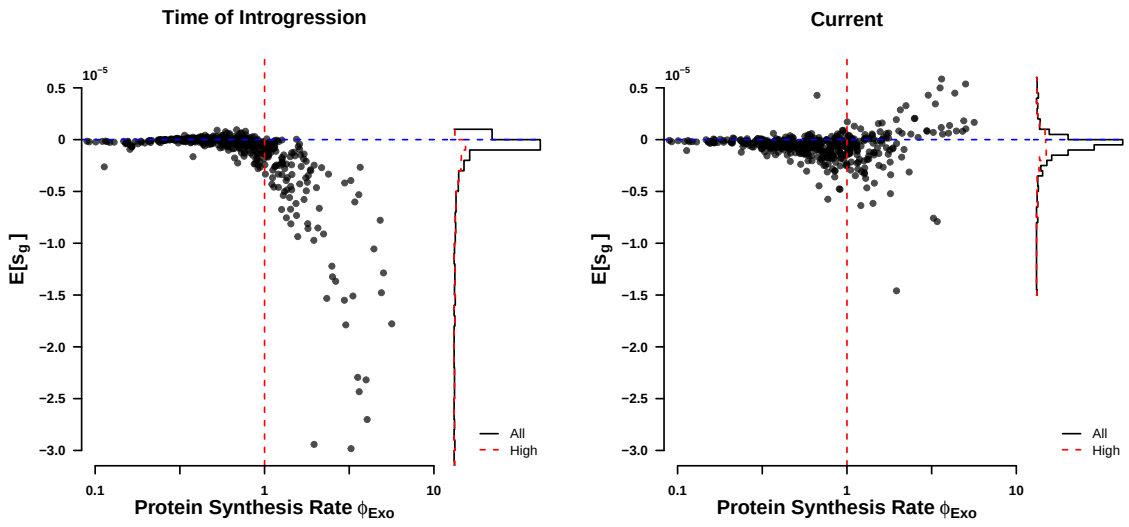

Figure 4: Fitness burden ΔsN_e (a) at the time of introgression ($\kappa = 5$), and (b) currently ($\kappa = 1$).

q[Gilchrist et al., 2015b].

At the time of the introgression, we predict that only a few genes were weakly exapted (Figure 4a) with all high expression genes ($\phi > 1$) being maladapted to the novel cellular environment. However, these highly expressed genes show the greatest rate of adaptation to the *L. kluyveri* cellular environment (Figures 4a, S5).

Discussion

Using ROC SEMPPR we show that the *L. kluyveri* genome contains two distinct signatures of cellular environments, its own endogenous and a foreign exogenous one obtained by an introgression event ($\Delta AIC = 78,000$). Following Payen et al. [2009], who defined the boundary of the anomalous chromosome region based on its elevated GC content, we partitioned the *L. kluyveri* genome into an endogenous and an exogenous gene set using gene location. We estimated the codon usage of the entire *L. kluyveri* genome and the separated endogenous and exogenous gene sets (Figure S6). Both, Mutation bias and selection bias differ between endogenous and exogenous genes. The endogenous genes show a strong mutation

bias towards A/T ending codons, while the exogenous genes show mutation is bias towards G/C ending codons. We observed the reversed to be true in selection bias, leading to a strong mismatch in codon usage between the gene sets, supporting our notion of two distinct signatures of codon usage.

Only half of the codon families share the same optimal codon in the endogenous and exogenous gene sets. However, we find that the strength of selection within a codon family differs between gene sets, causing a change in rank order. Nevertheless, we find a high correlation for our estimates of selection bias $\Delta\eta$ between the two gene sets. Our estimates of the optimal codon differ in nine cases between endogenous and exogenous genes. Interestingly, when the difference in codon usage is ignored, we find that in seven out of these nine cases the exogenous codon preference is inferred as optimal(Table S2). We find even greater discordance in our estimates of ΔM between endogenous and exogenous gene sets(Table S1). Without recognizing this difference in codon preference our estimates would not have been reflective of the actual codon usage of the *L. kluyveri* genome but of a relatively small introgressed gene set. This shows that a small number of exogenous genes ($\sim 9\%$ of genes) can have a disproportional impact on our estimates of ΔM and $\Delta\eta$ when fitting ROC SEMPPR to the entire *L. kluyveri* genome. While this is surprising, it highlights the importance to recognize differences in codon usage within a genome. Our results also indicate that we can attribute the higher GC content in the exogenous genes mostly to differences in mutation bias favoring G/C ending codons rather than a novel selective force.

Separating the endogenous and exogenous genes improves our estimates of protein synthesis rate ϕ by 17% relative to the full genome estimate ($\rho = 0.59$ vs. $\rho = 0.69$, respectively). Furthermore, we find that the variation in our estimates of ϕ is more consistent with the current understanding of gene expression (compare Figure 1a and b). Small variation in ϕ estimates may serve as an indicator for the presents of the signature of multiple cellular environments in future work. In the case of the *L. kluyveri* genome, finding a severe mismatch in ΔM causes ϕ values for low expression genes ($\phi < 1$) to increase towards the inflection

point where the dominance of mutation gives way to selection. In the case of the two codon amino acids, the inflection point represents the point at which mutation and selection are contributing equally to the probability of a codons occurrence. We find this inflection point around $\phi = 1$ for most amino acids (Figure S6). However, ROC SEMPPR assumes that estimates of ϕ follow a log-normal distribution with an expected value $E[\phi] = 1$. This assumption allows us to interpret $\Delta\eta$ as the strength of selection relative to drift (sN_e) for a codon in a gene with the average protein synthesis rate $\phi = 1$. However, tying the mean and standard deviation of the prior distribution together. Therefore, an increase in ϕ for low expression genes has to be meet with a decrease of ϕ for high expression genes, reducing the overall variance in ϕ (see Gilchrist et al. [2015b] for details).

Having shown that the introgressed exogenous genes reflect a foreign cellular environment, we used the quantitative estimates of mutation bias ΔM and selection bias $\Delta\eta$ from ROC SEMPPR to identify potential source lineages. The comparison of the endogenous and exogenous ΔM and $\Delta\eta$ estimates to 38 other yeast lineages revealed that most yeasts examined share similarity in mutation bias (Figure 2). Similar, we find strong similarities in selection bias between examined yeasts, potentially indicating stabilizing selection on codon usage. However, the exogenous genes do not share this commonality (Figure 2a), as their mutation bias strongly deviates from the endogenous genes and most other yeast species examined. This large difference in mutation bias between endogenous and exogenous genes allowed us to limit our candidate list to only two likely lineages, *C. dubliniensis* and *E. gossypii*. Interestingly, we did not find *Lachancea thermotolerance*, a thermophilic lineage closely related to *L. kluyveri*, as a potential candidate. While *L. thermotolerance* does have a strong synteny relationship with *L. kluyveri*, it does not show similarity in codon usage with the exogenous genes and does not share their high GC content.

Inference of synteny relationships between the exogenous region and *C. dubliniensis* and *E. gossypii* as well as closely related species showed that synteny relationship is limited to the Saccharomycetaceae clade (Figure S3b). *E. gossypii* showed the highest syntenty coverage

and is the only species with similar codon usage. Furthermore, *E. gossypii* is the only species examined with a GC content $> 50\%$ like it is observed in the exogenous region. The synteny coverage extends along the whole exogenous regions with the exception to the very 3' and 5' end of the region. The lack of synteny at the ends of the region also coincides with a drop in GC content, potentially indicating remains of the original replaced region or increased adaptation. The ancestral introgressed region may have also broken up in *E. gossypii* as we find non overlapping synteny with chromosomes *VI* and *V* as well as have indication that the C chromosome of *L. kluyveri* very robust to recombination events [Payen et al., 2009, Vakirlis et al., 2016].

With *E. gossypii* identified as potential source lineage of the introgressed region, we inferred the time past since the introgression occurred using our estimates of mutation bias ΔM. The 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., 2015b]. We estimated the time since introgression to be on the order of 6×10^8 generations, which is ~ 10 times longer time than a previous estimate by Friedrich et al. [2015] of a minimum of 5.6×10^7 generations . However, our estimate implicitly assumes all mutations are neutral, it is therefore a conservative estimate, potentially overestimating the time since introgression. Our estimate also depend on the assumption that the *E. gossypii* cellular environment reflects the ancestral environment at the time of the introgression. If the the ancestral mutation environment was more similar to the *L. kluyveri* environment at the time of the introgression than the *E. gossypii* environment is today, we would overestimate this time. On the other hand, we would underestimate the time since introgression if the two cellular environments were more dissimilar. We could have attempted to reconstruct the ancestral state of *E. gossypii*, however, as methods for ancestral state reconstruction are phenomenological, assumptions would be unclear.

The estimates of mutation bias ΔM also allow us the infer the time until the signature of the exogenous cellular environment will have decayed to be indistinguishable at about one

percent difference. Our estimate of decay is an order of magnitude greater than our estimate of the time since introgression (5×10^9 and 6×10^8 generations). Estimates of decay based on ΔM are more conservative as we expect differences in $\Delta\eta$ to decay before due to selection favoring the decay.

As we have determined that the introgression event has a long persisting exogenous signature, it is important to understand the fitness consequences of such an event. We estimated the genetic load that the exogenous genes represent assuming that the replaced endogenous genes and the new exogenous genes had the same amino acid composition. This assumption, along with the assumption that the current *L. kluyveri* cellular environment is reflective of the cellular environment at the time of the introgression is necessary to estimate the expected endogenous sequence that was replaced. Our results show that individual low expression genes contribute little to the genetic load, and show less adaptation to the novel cellular environment (Figure 4, S5). A small number of low expression genes even appear exapted, likely due to the mutation bias in the endogenous genes matching the selection bias in the exogenous genes for G/C ending codons. Highly expressed genes on the other hand have greatly adapted to the *L. kluyveri* cellular environment. This, however, does not mean that these genes show a higher rate of evolution, but that small changes in their sequence have large impacts on the fitness burden these sequences represent. To this day, the exogenous genes represent a significant fitness burden on *L. kluyveri*. However, our estimates are conservative as we do not account for potential changes in the codon usage of *E. gossypii*. While divergent evolution in codon usage between *E. gossypii* and *L. kluyveri* would cause us to overestimate the genetic load, convergent evolution, on the other hand, would cause us to underestimate the genetic load. However, as the introgression appears to have reached fixation [Friedrich et al., 2015], the genetic load relative to the replaced chromosome arm is only of theoretical interest.

The large genetic load the exogenous genes represented at the time of the introgression indicates that the fixation of the introgression was a very unlikely event in a population

with a large N_e as it is typical for yeasts. 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 with very different signatures of codon usage occur. One example is *Saccharomyces bayanus*, a hybrid of *Saccharomyces uvarum*,*Saccharomyces cerevisiae*, and *Saccharomyces eubayanus*. However, unlike with *L. kluyveri* and *E. gossypii* it appears that the donor lineages show similar codon usage. *Saccharomyces cerevisiae* and *Saccharomyces eubayanus* show a very strong correlation between selection bias $\Delta\eta$ of $\rho = 0.98$ and a strong correlation between mutation bias ΔM of $\rho = 0.83$ We were unable to identify codon usage for *Saccharomyces uvarum*. However, *L. kluyveri* diverged about 85 Mya ago from the rest of the Lachancea clade. This represents between 10^{10} to 10^{11} generations. Assuming for yeasts typical effective population size on the order of 10^8, we are left with 10^{18} to 10^{19} opportunities for such an event to occur. In addition, the strong mutation bias towards G/C ending codons in the exogenous genes may have contributed to the fixation of this introgression (include figure of ΔM v $\Delta\eta$). It is, on the other hand, also possible that despite their mismatch in codon usage, the exogenous genes have represented a fitness increase due to external environmental factors resulting in the fixation of the introgression.

In conclusion, 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, 1987] or tAI [dos Reis et al., 2004], ROC SEMPPR is sensitive to differences in mutation bias. We highlight potential pitfalls 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.

Text from Intro That Might Be Useful in Discussion

In general, the strength of selection on codon usage increases with gene expression [Ikemura, 1985, Gouy and Gautier, 1982]. Conversely, the impact of mutation bias on codon usage declines with gene expression. Thus, we can easily imagine that with increasing gene expression, codon usage shifts from a process dominated mutation to a process dominated by selection. Together, the mutation process favoring specific synonymous codons - or mutation bias - and the selection for translation efficiency scaled by gene expression and effective population size - or selection bias - shape codon usage in a genome.

In order to study the effects of introgression and the resulting mismatches in codon usage in the *L. kluyveri* genome, we use ROC SEMPPR, a mechanistic model of codon usage bias evolution grounded in population genetics. ROC SEMPPR, which uses a bayesian MCMC method for model fitting, allows us to quantify the contributions of mutation bias and selection on to the codon usage patterns of a set of genes. ROC SEMPPR also allows us to predict a gene's average predicting protein production rate based on its individual codon usage pattern with a precision comprable to that of more direct empirical methods [Gilchrist et al., 2015b]. By fitting ROC SEMPPR separately to *L. kluyveri*'s endogenous and exogenous sets of genes we generate a quantiative 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 mostly to differences in mutation bias. In addition, we show that separately fitting ROC SEMPPR to endogenous and exogenous gene sets substantially improves our ability to predict empirical estimates of protein synthesis rates over fitting ROC to a combined dataset of endogenous and exogenous genes.

In order to identify a potential source lineage for *L. kluyveri*'s exogenous gene set we fit ROC SEMPPR to the genomes of 38 yeastspecies. We then compared ROCs parameter estimates of mutation bias and selection of *L. kluyveri*'s exogenous genes to these species and found a strong correlation in only two species, *E. gossypii* and *C. dubliniensis*. We

also compared the synteny of *L. kluyveri*'s exogenous genes to these lineages. We found strong synteny in a number of cases, most notably in *E. gossypii* but not *C. dubliniensis*. As a result, we conclude that of the yeast species we examined, the *E. gossypii* lineage is most closely related to the the donor of *L. kluyveri*'s exogenous genes. Assuming that *E. gossypii*'s mutation bias is similar to the source of the exogenous genes, we estimated the introgression occurred approximately 6×10^8 generations ago using a model of exponential decay to describe the shift in mutation bias of the exogenous genes. Finally, we estimate the selective cost or genetic load of the exogenous genes due to codon usage mismatch using our estimates of the selection parameters from *L. kluyveri*'s endogenous genes and the our estimates of the protein synthesis rate of the exongenous genes.

Need to discuss introgression

- Load calculations
  - Some genes pre-adapted to new environment
  - Most genes not
  - Load estimate indicates strong selection against introgression sequences alone
- Explaining introgression
  - Assuming introduction is continuous as it appears, indicates little recombination during spread
  - Data suggests introgression spread quickly
  - Potential explanations
    * Identified wrong source, though even current load is quite large.
    * Major flaws in our calculation of fitness costs.
    * Failure for positive selection on at amino acid or regulatory sequence at one or more loci, countering the selection on CU

    * Introgression triggered speciation event, thus N_e was very small (< 100) so even if strongly selected against it still had a reasonable probability of fixing.
    * Unlikely event, but introgressions happen frequently. Note here mutation is actually an introgression event, not a nt change. Although pop gen predicts fixation probability is very low. However, pop gen also tells us that if such an unlikely fixation occurs, it is very likely to happen quickly. Thus, continuous nature of introgression also consistent with a rare, maladaptive fixation event.
    * Other adaptive effects of introgression seems most plausible, but since we don't know have reasonable estimates about frequently hybridiations occur nor accurate estimate of how frequent such introgressions fix, the maladaptive explanation is hard to evaluate.
    * Combination of most maladaptive, some adaptive alleles, and speciation could also be a feasible hypothesis.

- Terminiology
    - Codon families?

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/` (last accessed: 09-27-2014) and the annotation for *L. kluyveri* NRRL Y-12651 (assembly ASM14922v1) from NCBI (last accessed: 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). ROC SEMPPR was run from multiple starting values for at least 250,000 iterations, every 50th sample was collected 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

Because our ΔM and $\Delta\eta$ are meaningful only for comparisons between synonymous codons, ROC SEMPPR returns mutation bias ΔM and selection bias $\Delta\eta$ parameter values relative to a reference codon. While ROC SEMPPR's choice of the reference codon is largely arbitrary, changes in the reference codon affect [NEED TO COMPLETE] To circumvent this issue, we express our estimates relative to the mean for each codon family.

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.

$$\Delta M^{c}_{i,a} = \Delta M_{i,a} - \Delta\bar{M}_{a} \tag{1}$$

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

Comparison of codon specific parameters (ΔM and $\Delta\eta$) was performed using the function lmodel2 in the R package lmodel2 (1.7.3) and R version 3.4.1. Type II regression was performed with re-centered parameter estimates, accounting for noise in dependent and independent variable.

Synteny

We obtained complete genome sequences from NCBI (last accessed: 02-05-2017). Genomes were aligned and checked for synteny using SyMAP (4.2) with default settings [Soderlund et al., 2006, 2011]. We assessed Synteny as percentage non-overlapping coverage of the exogenous gene region (Figure S3b).

Determining introgression timeline

We modeled the change in codon frequency over time using an exponential model for all two codon amino acids, and described 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 directly related to the steady state of equation 3.

$$\frac{\mu_{2,1}}{\mu_{1,2} + \mu_{2,1}} = \frac{1}{1 + \exp(\Delta M_{endo})} \tag{4}$$

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

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

where K is

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

Equation 5 was solved over time 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 estimates of ΔM_{gos} from *E. gossypii*. Current codon frequencies

for each codon family where taken from our estimates of ΔM from the exogenous genes. Mathematica (9.0.1.0) [Inc.] was used to calculate the time t_{exo} it takes for the initial codon frequencies $c_1(0)$ for each codon family to change to the current exogenous codon frequencies. The same equation was used to determine the time t_{endo} at which the signal of the exogenous cellular environment has decayed to within 1% of the endogenous environment.

Estimating Genetic Load

To estimate the fitness burden, we made three key assumptions. First, we assumed that the current exogenous amino acid composition of genes is representative of the replaced endogenous genes. 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 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 from XXX et al. we scale our estimates of $\Delta\eta$ and define $\Delta\eta' = \frac{\Delta\eta}{N_e}$. We calculated the fitness burden each gene represents assuming additive fitness effects as

$$s_g = \sum_{i}^{C} -\kappa\phi_g\Delta\eta'_i n_{g,i} \tag{7}$$

where s_g is the selection against translation inefficiency. ϕ_g is the estimated protein synthesis rate for gene g in the exogenous gene set. $n_{g,i}$ is the codon count of each codon i in the codon set C for each gene g. κ is a constant, scaling the efficacy of selection between cellular environments. Like stated previously, our $\Delta\eta$ are relative to the mean of the codon family. We find that the fitness burden of the introgressed genes is minimized at $\kappa \sim 5$ (Figure S4b). Thus, 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 fitness burden 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

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

435 m_{a_i} is the number of occurrences of amino acid a that codon i codes for.

436 We report the genetic load of the introgression as $\Delta s = s_{Intro} - s_{Endo}$ where s_{Intro} is
437 either the fitness burden at the time of the introgression or presently.

438 Acknowledgments

439 This work was supported in part by NSF Awards MCB-1120370 (MAG and RZ) and DEB-
440 1355033 (BCO, MAG, and RZ) with additional support from The University of Tennessee
441 Knoxville. CL received support as a Graduate Student Fellow at the National Institute
442 for Mathematical and Biological Synthesis, an Institute sponsored by the National Science
443 Foundation through NSF Award DBI-1300426, with additional support from UTK. The
444 authors would like to thank Brian C. O'Meara and Alexander Cope for their helpful criticisms
445 and suggestions for this work.

446 References

447 M. Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129:
448 897–907, 1991.

449 Premal Shah and Michael A. Gilchrist. Explaining complex codon usage patterns with se-
450 lection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the*
451 *National Academy of Sciences of the United States of America*, 108(25):10231–10236,
452 2011. doi: 10.1073/pnas.1016719108. URL `http://www.pnas.org/content/108/25/`
453 `10231.abstract`.

M. A. Gilchrist. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Molecular Biology and Evolution*, 24: 2362–2373, 2007. doi: doi:10.1093/molbev/msm169. URL `http://mbe.oxfordjournals.org/cgi/reprint/msm169v2.pdf`.

Cedric Landerer, Alexander Cope, Russell Zaretzki, and Michael A Gilchrist. Anacoda: analyzing codon data with bayesian mixture models. *Bioinformatics*, 34(14):2496–2498, 2018.

Michael A. Gilchrist, Wei-Chen Chen, Premal Shah, Cedric L. Landerer, and Russell Zaretzki. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. *Genome Biology and Evolution*, 7(6):1559–1579, 2015a. doi: 10.1093/gbe/evv087. URL `http://gbe.oxfordjournals.org/content/7/6/1559.abstract`.

Clia Payen, Gilles Fischer, Christian Marck, Caroline Proux, David James Sherman, Jean-Yves Coppe, Mark Johnston, Bernard Dujon, and Ccile Neuvglise. Unusual composition of a yeast chromosome arm is associated with its delayed replication. *Genome Research*, 19(10):1710–1721, 2009.

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

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

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

Nikolaos Vakirlis, Véronique Sarilar, Guénola Drillon, Aubin Fleiss, Nicolas Agier, Jean-Philippe Meyniel, Lou Blanpain, Alessandra Carbone, Hugo Devillers, Kenny Dubois, Alexandre Gillet-Markowska, Stéphane Graziani, Nguyen Huu-Vang, Marion Poirel, Cyrielle Reisser, Jonathan Schott, Joseph Schacherer, Ingrid Lafontaine, Bertrand Llorente, Cécile Neuvéglise, and Gilles Fischer. Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome research*, 26(7):918–32, 2016.

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

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

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

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

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

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

Gregory I. Lang and Andrew W. Murray. Estimating the per-base-pair mutation rate in the yeast saccharomyces cerevisiae. *Genetics*, 178(1):67 – 82, 2008. ISSN 0016-6731.

Wolfram Research, Inc. Mathematica, Version 9.0. Champaign, IL, 2012.

Hypothesis	L	n	AIC	ΔAIC
Endogenous & Exogenous			5,235,598	0
Combined			5,311,060	75,462

Table 1: L, number of model parameters n, AIC, and ΔAIC.

Figures

Table

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 S1: Synonymous codons with the greatest mutational preference (i.e. largest ΔM value). Bold face codons indicate synonyms whose ...

Supplementary Material

Supporting Materials for *Fitness consequences of mismatched codon usage* by Landerer *et al.*.

Tables

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

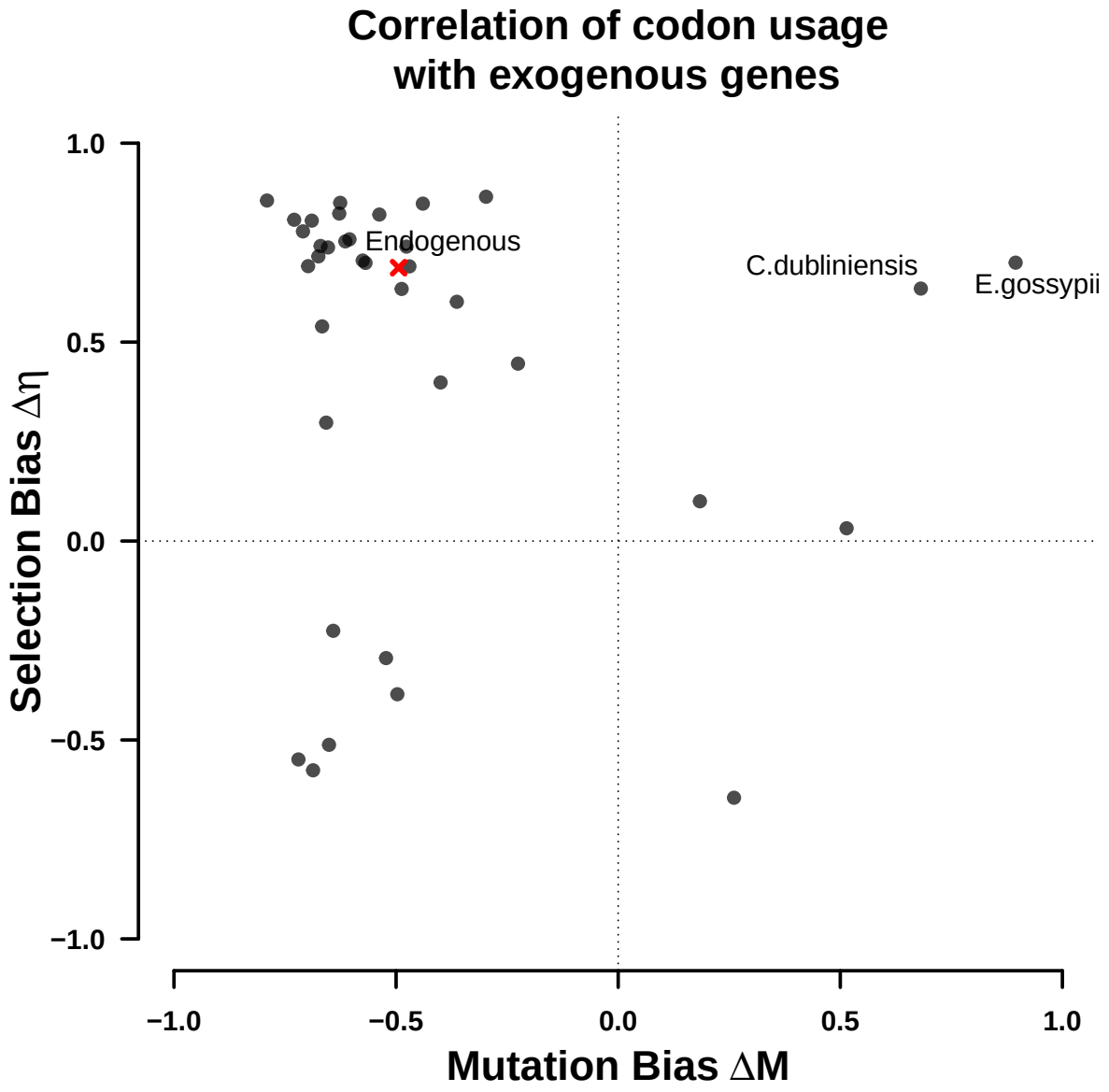

Figure S1: Correlation of ΔM and $\Delta\eta$ of the endogenous genes with 38 examined yeast lineages. Dots indicate the correlation of ΔM and $\Delta\eta$ of the lineages with the endogenous and exogenous parameter estimates. All regressions were performed using a type II regression.

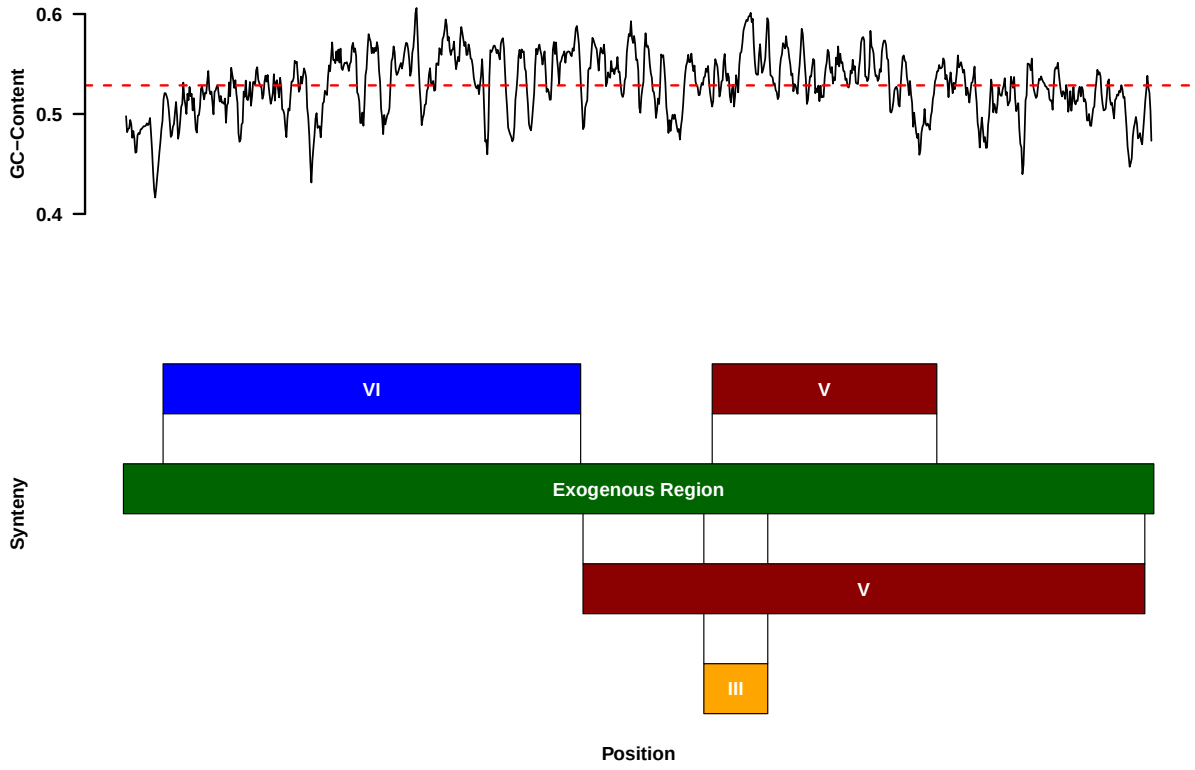

Figure S2: Synteny relationship of *E. gossypii* and the exogenous genes

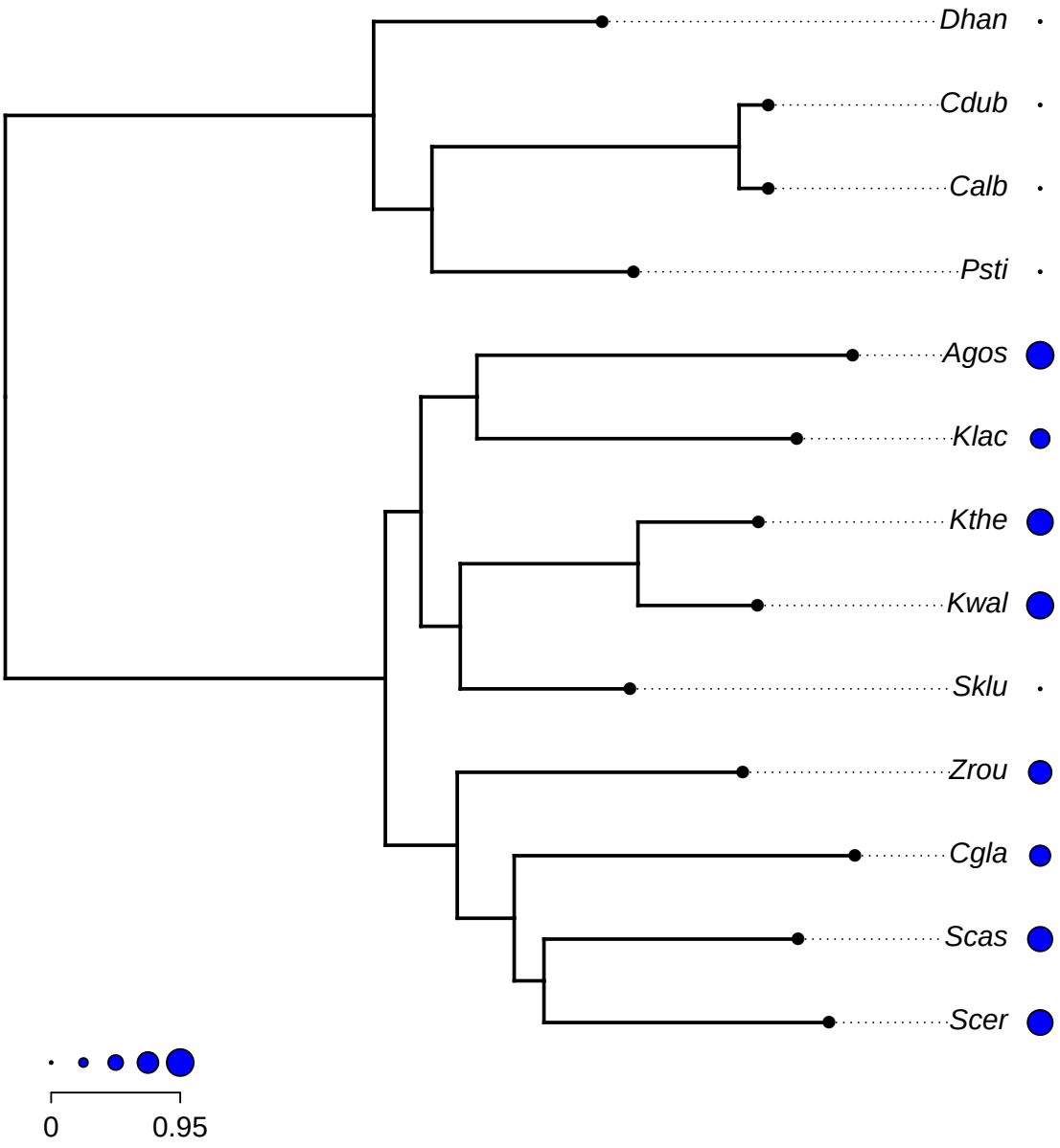

Figure S3: Amount of synteny for each species (Units of std dev) checked for synteny.

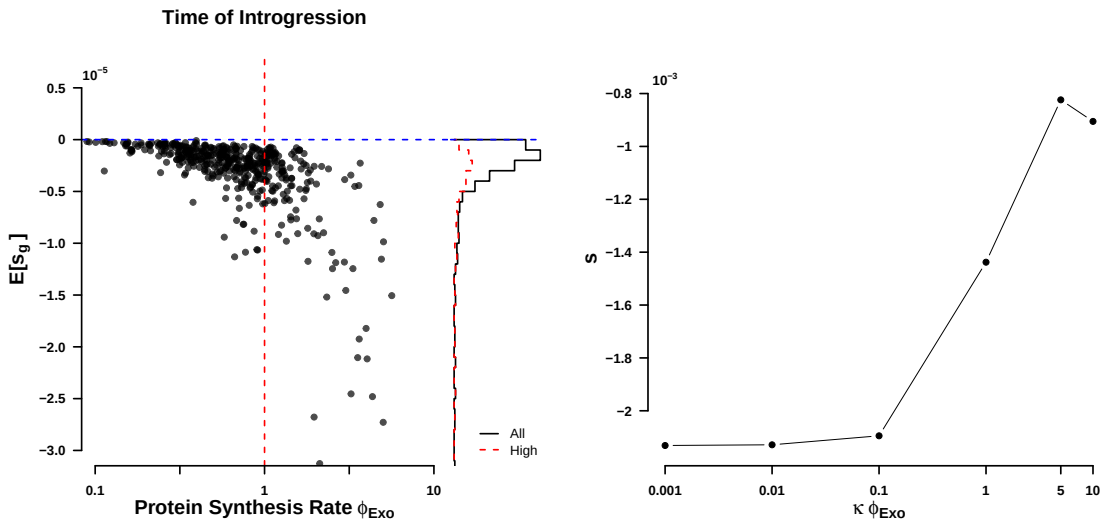

Figure S4: Suppl Fig: Fitness burden (left) without scaling of ϕ, and change of total fitness burden with scaling κ

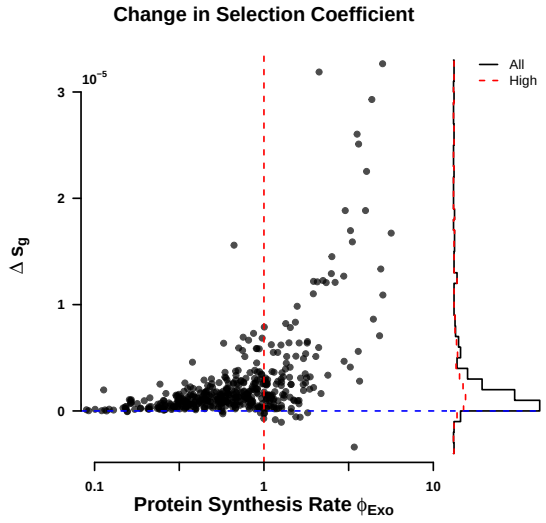

Figure S5: Total amount of adaptation between time of introgression and now

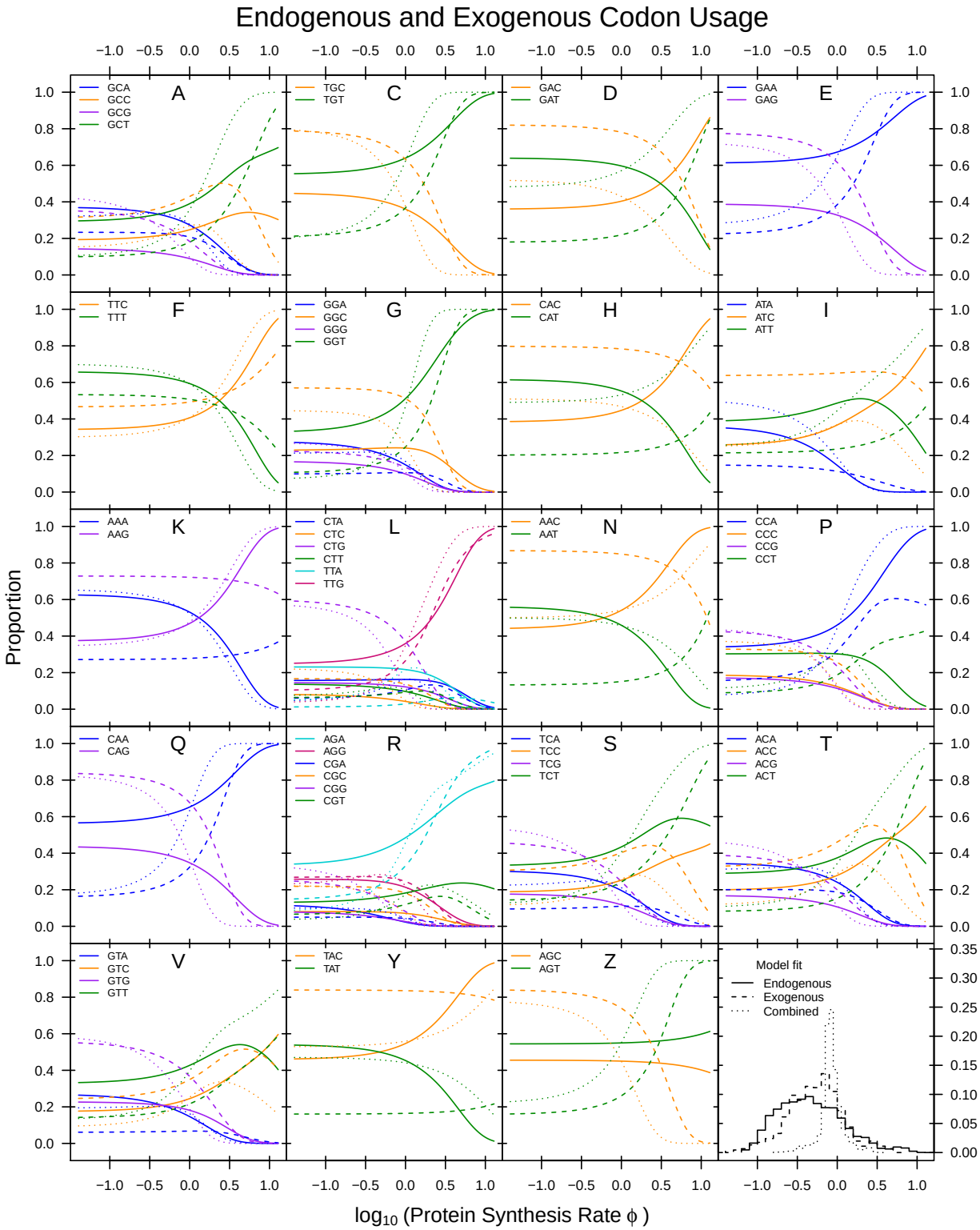

Figure S6: Suppl Fig