

2 **Decomposing mutation and selection to identify**  
3 **mismatched codon usage**

4 CEDRIC LANDERER<sup>1,2,\*</sup>, RUSSELL ZARETZKI<sup>3</sup>, AND MICHAEL  
5 A. GILCHRIST<sup>1,2</sup>

6 <sup>1</sup>Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-  
7 1610

8 <sup>2</sup>National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996-3410

9 <sup>3</sup>Department of Business Analytics & Statistics, Knoxville, TN 37996-0532

10 \*Corresponding author. E-mail: cedric.landerer@gmail.com

Version dated: August 6, 2018

## Abstract

Codon usage has been used as a measure for adaptation of genes to their cellular environment for decades. The introgression of genes from one cellular environment to another may cause well adapted genes to suddenly be less adapted due to them having evolved in a different environment. As a result, we expect that transferred genes result in a large fitness burden for the new host organism. Here we examine the yeast *Lachancea kluyveri* which has experienced a large introgression, replacing the left arm of chromosome C ( $\sim 10\%$  of its genome). The *L. kluyveri* genome provides an opportunity to study the evolution of introgressed genes to a novel cellular environment and estimate the fitness cost such a transfer imposes. We quantified the effects of mutation bias and selection against translation inefficiency on the codon usage pattern of the endogenous and introgressed exogenous genes using a Bayesian mixture framework. We found 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 improved our ability to predict protein synthesis rate by 17% and allowed us to accurately assess codon preferences. In addition we utilize the estimates of mutation bias and selection against translation inefficiency to determine *Eremothecium gossypii* as potential source lineage, estimate the time since introgression to be on the order of  $6 \times 10^8$  and assess the fitness burden across introgressed loci, showing the advantage of mechanistic models when analyzing codon data.

## Introduction

Mutation, selection and genetic drift can be used to quantify the environment a genome has evolved in. Mutation bias is purely determined by the cellular environment, while the strength and efficacy of selection relative to drift is determined by the cellular environment, e.g. tRNA abundance, and the natural environment e.g. gene expression. A lineages effective population size determines the efficacy of selection relative to drift. Synonymous codon usage, the non-uniform usage of codons encoding the same amino acid, is a reflection of both, the cellular and the natural environment. Decomposing codon usage, therefore, provides us the necessary information to describe the environment a genome has evolved in in terms of its codon usage.

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. This mutation-selection-drift balance model allows us to explicitly describe the environment in which genes evolve with respect to mutation and selection bias. Here we show that estimating the influence of mutation bias and selection bias on a gene's codon usage allows us to not only predict protein synthesis rate  $\phi$ , but also, to infer its history and make predictions about its future with respect to these forces.

Most studies implicitly assume that synonymous codon usage of a genome is shaped by a single cellular environment. However, it is easy to think about the influence of multiple cellular environments within a cell, as genes are horizontally transferred, introgress, or as species hybridize. Genes introduced via horizontal gene transfer, introgression, or hybridization may carry the signature of a different, foreign cellular environment. These transferred genes may

60 be less adapted to their new cellular environment, potentially imposing large fitness burdens  
61 to the organism. We expect a greater fitness burden of transferred genes if donor and recipi-  
62 ent environment differ greatly in their selection bias, making such transfers less likely. More  
63 practically, if transferred genes are unaccounted for, they may distort parameters by biasing  
64 estimates. This can lead to the conclusion of the wrong codon preference for an amino acid  
65 when analyzing a genome that has experienced such transfer events.

66 In this study, we analyze the synonymous codon usage of the genome of *Lachancea*  
67 *kluyveri*, the earliest diverging lineage of the Lachancea clade. The Lachancea clade di-  
68 verged from the Saccharomyces clade prior to the whole genome duplication, about 100 Mya  
69 ago. Since its divergence from the other Lachancea, *L. kluyveri* has experienced a large in-  
70 trogression of exogenous genes. The introgression replaced the left arm of the C chromosome  
71 and displays a 13% higher GC content than the endogenous *L. kluyveri* genome [Payen et al.,  
72 2009, Friedrich et al., 2015]. These characteristics make *L. kluyveri* an ideal model to study  
73 the effects of an introgressed cellular environment and the resulting mismatch codon usage.

74 Using ROC SEMPPR, a population genetics Bayesian model, allows us to quantify the  
75 cellular environment in which genes have evolved by separating and estimating effects of  
76 mutation bias and selection bias, and predicting protein production rate [Gilchrist et al.,  
77 2015]. We use ROC SEMPPR to describe two cellular environments reflected in the *L.*  
78 *kluyveri* genome, a native endogenous and an introgressed exogenous environment. Our  
79 results indicate that the difference in GC content between endogenous and exogenous genes  
80 mostly to differences in mutation bias. Recognizing the differences in codon usage between  
81 the endogenous and exogenous gene sets also improves our ability to predict protein synthesis  
82 rate from the sequence data alone.

83 With our improved model fits, we obtained more reliable estimates of mutation bias,  
84 selection bias and protein synthesis rate, allowing us to address more refined questions of  
85 biological importance. First we determine a potential source lineage of the exogenous genes  
86 using a combination of information in codon usage and gene synteny. We compared estimates

of mutation bias ( $\Delta M$ ) and selection bias ( $\Delta\eta$ ) for the exogenous genes to 38 yeast lineages and further investigated candidate lineages using synteny. Second, we estimate the time since introgression and the persistence of the signal of the exogenous cellular environment from our estimates of  $\Delta M$  using an exponential model of decay. Third, we estimate the selective cost of the mismatched codon usage for the introgression, using our estimates of  $\Delta\eta$  and protein synthesis rate  $\phi$ . Thus, in addition to being able to estimate codon preference and gene expression to describe codon usage patterns, we also gain insights into the evolution of genes that have been transferred between lineages.

## Results

Model selection and validation confirmed that the *L. kluyveri* genome contains signatures of at least two cellular environments. We compared model fits of ROC SEMPFR to the homogeneous *L. kluyveri* genome and the separated sets of endogenous and exogenous genes of 4864 and 497 genes respectively, using AnaCoDa [Landerer et al., 2018]. We compared estimates of the cellular environment to describe differences in endogenous and exogenous codon usage. Furthermore, we utilize the differences in model fit and parameters estimated from the endogenous and exogenous genes to explore the evolution of the exogenous gene set.

AIC indicates that parameter estimates for mutation bias ( $\Delta M$ ) and selection bias ( $\Delta\eta$ ) differ greatly between exogenous and endogenous gene sets. As a result, the partitioning of the *L. kluyveri* genome into an endogenous and exogenous gene set is clearly favored by model selection. The inclusion of 81 additional parameters (40 for  $\Delta M$ , 40 for  $\Delta\eta$ , and one for  $s_\phi$ ) necessary to describe both gene sets separately improves our model fit by  $\sim 78,000$  AIC units (5,313,110 for the combined gene set vs 5,235,598 for the separated gene sets).

In addition to model selection, we utilized independent information on gene expression to evaluate model fit. Recognizing differences in  $\Delta M$  and  $\Delta\eta$  for the endogenous and exogenous

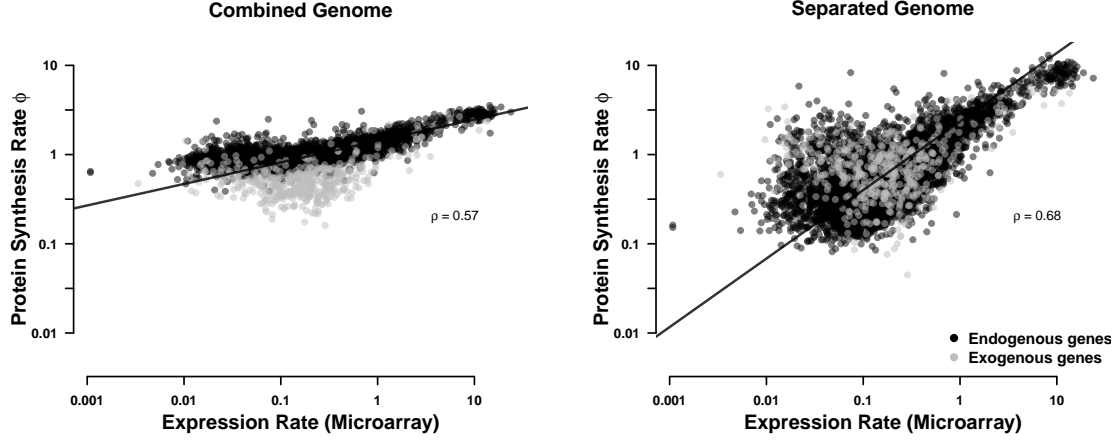


Figure 1: Comparison of predicted protein synthesis rate  $\phi$  to Microarray data from XX for (a) the combined genome and (b) the separated endogenous and exogenous genes. Endogenous genes are displayed in black and exogenous genes in red. Black line indicates type II regression line.

gene sets substantially improves our ability to predict protein synthesis rate  $\phi$  ( $\rho = 0.69$  vs.  $\rho = 0.59$  for the full genome; Figure 1).

## Differences in the Endogenous and Exogenous Codon Usage

As our estimates of parameters for a codon family coding for an amino acid are relative to a reference codon, changes in the reference codon will change the order between sets. To better compare our estimates of mutation bias ( $\Delta M$ ) and selection bias ( $\Delta \eta$ ) obtained from fitting ROC SEMPFR between the endogenous and exogenous gene sets, we express our estimates relative to the mean for each codon family. As we find larger differences between  $\Delta M$  than  $\Delta \eta$  (Figure 2). Estimates of  $\Delta M$  in the endogenous genes negatively correlate with the  $\Delta M$  estimates for the exogenous genes ( $\rho = -0.49$ ) indicating strong discordance in the mutation environment between *L. kluuyveri* and the donor lineage of the exogenous genes. For example,  $\sim 95\%$  of codon families show mutation preference for A/T ending codons, in contrast, the exogenous genes display an equally strong mutation bias towards C/G ending codons. Only the two codon amino acid Phenylalanine (Phe, F) shows complete



optimal codon between  $\Delta\eta_{complete}$ ,  $\Delta\eta_{exo}$ , and  $\Delta\eta_{endo}$  are in discordance for seven codon families. Thus, recognizing and treating endogenous and exogenous genes as separate sets avoids the inference of incorrect synonymous codon preferences.

## Determining Source of Exogenous Genes

We combined our estimates of mutation bias ( $\Delta M$ ) 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 3a). The endogenous *L. kluyveri* genome

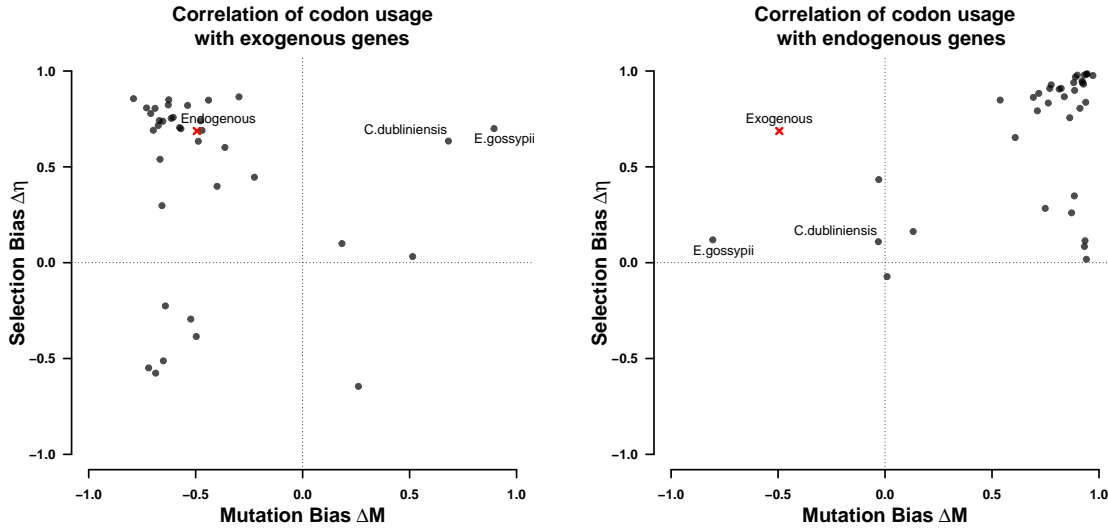


Figure 3: Correlation of  $\Delta M$  and  $\Delta\eta$  of the (a) exogenous and (b) endogenous genes with 38 examined yeast lineages. Dots indicate the correlation of  $\Delta M$  and  $\Delta\eta$  of the lineages with the endogenous and exogenous parameter estimates. All regressions were performed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

exhibits codon usage very similar to most yeast lineages examined, indicating little variation in codon usage among the examined yeasts (Figure 3b). Four lineages show a positive correlation for  $\Delta M$  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  $\Delta M$  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 S1a,b). *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 Saccharomycetaceae group (Figure S1b). 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 age using an exponential model of decay for mutation bias, by assuming that *E. gossypii* is still representative of the mutation bias of its ancestral source lineage at the time of the introgression. We utilize the  $\Delta M$  estimates for all two codon amino acids and infer the age of the introgression to be on the order of  $6.2 \pm 1.2 \times 10^8$  generations. We assume a mutation rate of  $3.8 \times 10^{-10}$  per nucleotide per generation, a value in line with other estimates [Zhu et al., 2014, Lang and Murray, 2008]. *L. kluyveri* experiences between one and eight generations per day, we therefore expect the introgression to have occurred about 205,000 to 1,600,000 years ago which is longer than previous estimates of Friedrich et al. [2015]. However, our estimates are likely overestimates as they assume a purely neutral decay.

Furthermore, we 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  $\Delta M$  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.

## Fitness Burden 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]. However, using our estimates of  $\Delta M$  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  $sN_e \propto \phi\Delta\eta$  (See Methods for details).

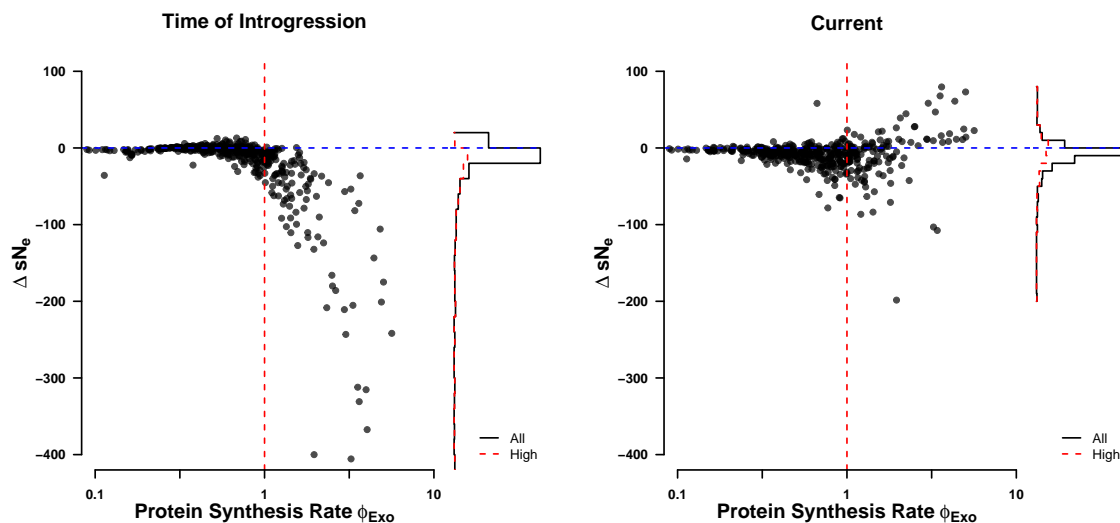


Figure 4: Fitness burden  $\Delta sN_e$  (a) at the time of introgression ( $\kappa = 5$ ), and (b) currently ( $\kappa = 1$ ).

We estimate the genetic load of the exogenous genes at the time of introgression (Figure 4a) and currently (Figure 4b). These estimates are dependent on three key assumptions. First, we assume again that the current cellular environment of *E. gossypii* is reflective of the ancestral environment. Second, we assume that the current amino acid composition of the

exogenous genes is the same as in the replaced endogenous genes. Third, we assume that the difference in the efficacy of selection between *E. gossypii* and *L. kluyveri* can be described with a simple scaling term we call  $\kappa$  (Figure S2b). As  $\Delta\eta$  is defined as  $\Delta\eta = 2N_e q(\eta_i - \eta_j)$ , we can not distinguish if  $\kappa$  is a scaling on protein synthesis rate  $\phi$ , effective population size  $N_e$  the value of an ATP  $q$  [Gilchrist et al., 2015].

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, S3).

## 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 S4). 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 corre-

lation 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  $\Delta M$  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  $\Delta M$  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  $\phi$  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  $\phi$  is more consistent with the current understanding of gene expression (compare Figure 1a and b). Small variation in  $\phi$  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  $\Delta M$  causes  $\phi$  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 S4). However, ROC SEMPPR assumes that estimates of  $\phi$  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  $\phi$  for low expression genes has to be met with a decrease of  $\phi$  for high expression genes, reducing the overall variance in  $\phi$  (see Gilchrist et al. [2015] for details).

Having shown that the introgressed exogenous genes reflect a foreign cellular environment, we used the quantitative estimates of mutation bias  $\Delta M$  and selection bias  $\Delta\eta$  from ROC SEMPPR to identify potential source lineages. The comparison of the endogenous and exogenous  $\Delta M$  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 S1b). *E. gossypii* showed the highest synteny 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  $\Delta M$ . The  $\Delta M$  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]. We estimated the time since introgression to be on the order of  $6 \times 10^8$  generations, which is  $\sim 10$  times longer time than a previous estimate by Friedrich et al. [2015] of a minimum of  $5.6 \times 10^7$  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  $\Delta M$  also allow us to 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 ( $5 \times 10^9$  and  $6 \times 10^8$  generations). Estimates of decay based on  $\Delta M$  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, S3). 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  $\Delta M$  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  $\Delta M$  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.

## 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



349 <http://www.yeastgenome.org/download-data/> (last accessed: 09-27-2014) and the anno-  
 350 tation for *L. kluyveri* NRRL Y-12651 (assembly ASM14922v1) from NCBI (last accessed:  
 351 12-09-2014). We assigned 457 genes located on chromosome C with a location within the  
 352  $\sim 1Mb$  window to the exogenous gene set. All other 4864 genes of the *L. kluyveri* genome  
 353 were assigned to the exogenous genes. All genes could be uniquely assigned to one or the  
 354 other gene set.

## 355 **Model Fitting with ROC SEMPFR**

356 ROC SEMPFR was fitted to each genome using AnaCoDa (0.1.1) [Landerer et al., 2018] and  
 357 R (3.4.1). ROC SEMPFR was run from multiple starting values for at least 250,000 itera-  
 358 tions, every 50th sample was collected to reduce autocorrelation. After manual inspection  
 359 to verify that the MCMC had converged, parameter posterior means were estimated from  
 360 the last 500 samples.

## 361 **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.

$$\Delta M_{i,a}^c = \Delta M_{i,a} - \Delta \bar{M}_a \quad (1)$$

$$\Delta \eta_{i,a}^c = \Delta \eta_{i,a} - \Delta \bar{\eta}_a \quad (2)$$

362 Comparison of codon specific parameters ( $\Delta M$  and  $\Delta \eta$ ) was performed using the function  
 363 `lmodel2` in the R package `lmodel2` (1.7.3) and R version 3.4.1. Type II regression was  
 364 performed with re-centered parameter estimates, accounting for noise in dependent and  
 365 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 S1b).

## Determining introgression timeline

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

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

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  $\Delta M_{endo}$  can be directly related to the steady state of equation 3.

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

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

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

where  $K$  is

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

Equation 5 was solved over time with a mutation rate  $m_{2,1}$  of  $3.8 \times 10^{-10}$  per nucleotide per generation [Lang and Murray, 2008]. Initial codon frequencies  $c_1(0)$  for each codon family were taken from our estimates of  $\Delta M_{gos}$  from *E. gossypii*. Current codon frequencies

for each codon family where taken from our estimates of  $\Delta M$  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 fitness burden

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  $\phi$  has not changed between the replaced endogenous and the introgressed exogenous genes.

We calculated the fitness burden each gene represents assuming additive fitness effects as

$$(sN_e)_g = \sum_i^C -\kappa\phi_g\Delta\eta_i n_{g,i} \quad (7)$$

where  $(sN_e)_g$  is the selection against translation inefficiency relative to drift.  $\phi_g$  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  $g$ .  $\kappa$  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 S2b). Thus, we set  $\kappa = 1$  if we calculate the  $(sN_e)_g$  for the endogenous and the current exogenous genes, and  $\kappa = 5$  for  $(sN_e)_g$  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

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

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

We report the fitness burden of the introgression as  $\Delta sN_e = (sN_e)_{Intro} - (sN_e)_{Endo}$  where  $(sN_e)_{Intro}$  is either the fitness burden 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 Brian C. O’Meara and Alexander Cope for their helpful criticisms and suggestions for this work.

## References

- 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.
- 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.

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, 2015.

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

Yuan O Zhu, Mark L Siegal, David W Hall, and Dmitri A Petrov. Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences*, 111(22):E2310–E2318, 2014.

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.

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 Poiriel, 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.

- 432 M dos Reis, R Savva, and L Wernisch. Solving the riddle of codon usage preferences: a test  
433 for translational selection. *Nucleic Acids Research*, 32(17):5036–5044, 2004.
- 434 C Soderlund, W Nelson, A Shoemaker, and A Paterson. Symap A system for discovering  
435 and viewing syntenic regions of fpc maps. *Genome Research*, 16:1159 – 1168, 2006.
- 436 C Soderlund, M Bomhoff, and W Nelson. Symap v3.4: a turnkey syteny system with  
437 application to plant genomes. *Nucleic Acids Research*, 39(10):e68, 2011.
- 438 Wolfram Research, Inc. Mathematica, Version 9.0. Champaign, IL, 2012.

Amino Acid	<i>E. gossypii</i>	Endogenous	Exogenous	<i>L. kluyveri</i>
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 <sub>4</sub> S	TCG	TCT	TCG	TCG
Thr T	ACG	ACA	ACG	ACG
Val V	GTG	GTT	GTG	GTG
Tyr Y	TAC	TAT	TAC	TAC
Ser <sub>2</sub> Z	AGC	AGT	AGC	AGC

Table S1: Synonymous codon preference in the various data sets based on our estimates of  $\Delta M$

## Supplementary Material

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

Amino Acid	<i>E. gossypii</i>	Endogenous	Exogenous	<i>L. kluyveri</i>
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 <sub>4</sub> S	TCA	TCC	TCT	TCT
Thr T	ACT	ACC	ACT	ACT
Val V	GTT	GTC	GTT	GTT
Tyr Y	TAT	TAC	TAT	TAC
Ser <sub>2</sub> Z	AGT	AGT	AGT	AGT

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



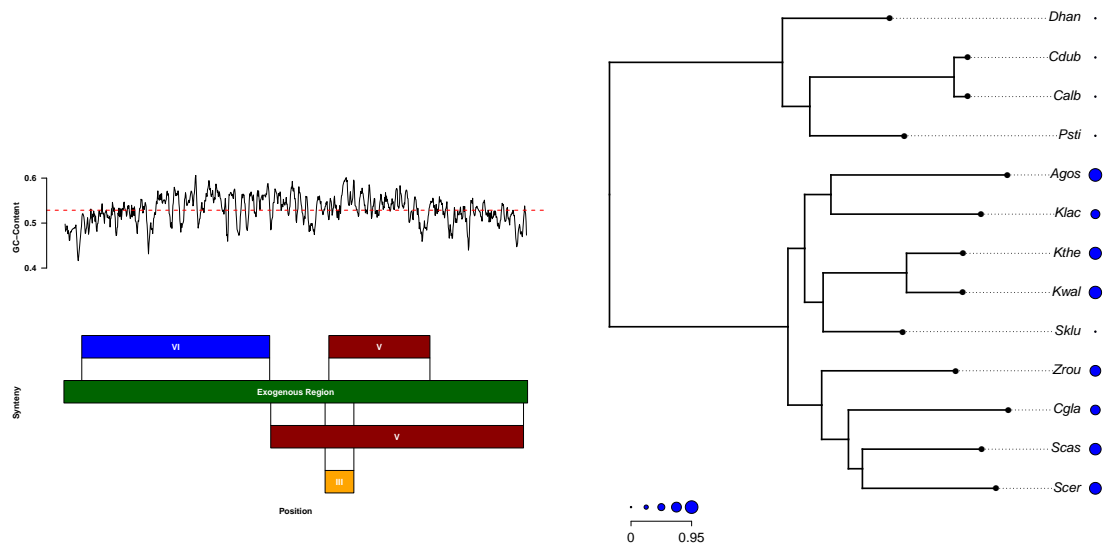


Figure S1: Suppl Fig: Synteny relationship of *E. gossypii* and the exogenous genes (left), Amount of synteny for each species (Units of std dev) checked for synteny.

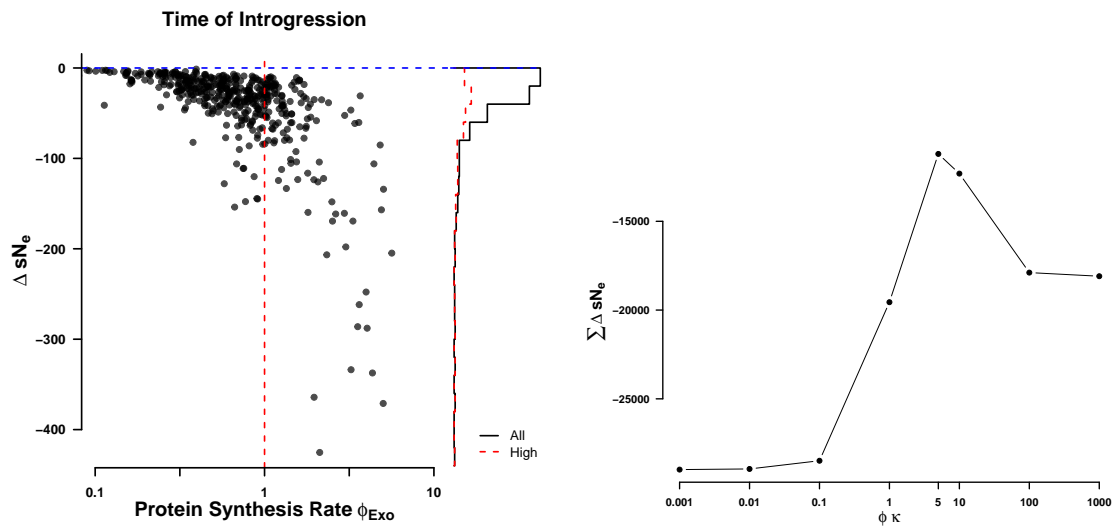
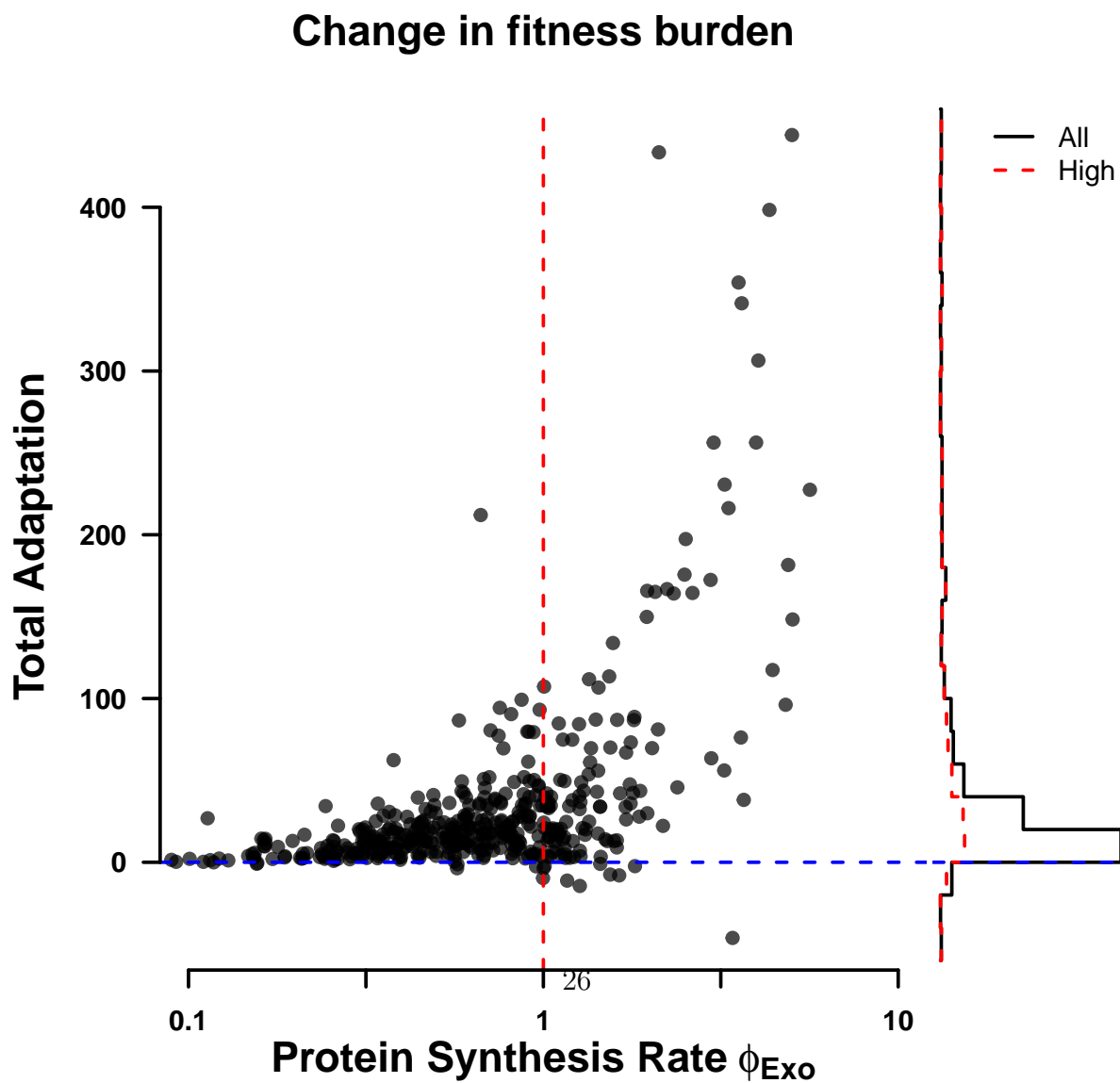


Figure S2: Suppl Fig: Fitness burden (left) without scaling of  $\phi$ , and change of total fitness burden with scaling  $\kappa$



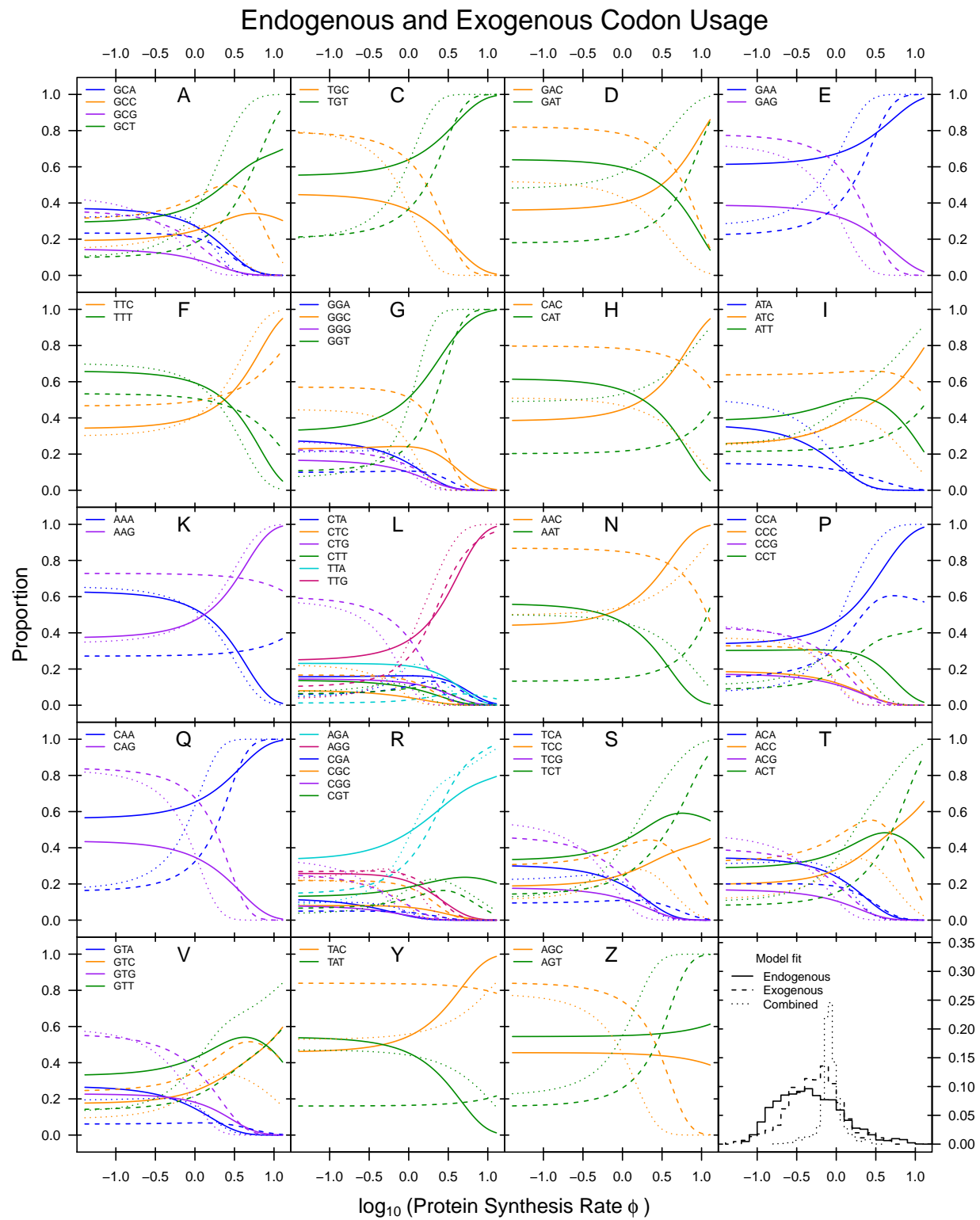


Figure S4: Suppl Fig