RH: LANDERER ET AL.— Intragenomic variation in codon usage

# Fitness consequences of 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: July 21, 2018

## Abstract

Codon usage has been used as a measure for adaptation of genes to their genomic environment for decades. The introgression of genes from one genomic environment to another may cause well adapted genes to suddenly be less adapted due to their signature of a foreign genomic environment. The reflection of a foreign genomic environment in transferred genes can 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 adaptation of introgressed genes to a novel genomic environment and estimate the fitness cost such a transfer imposes. The codon usage of the endogenous *L. kluyveri* genome and the exogenous genes were analyzed, using ROC SEMPPR which allows for the effects of mutation bias and selection bias on codon usage to be separated. 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 bias gained using ROC SEMPPR to determine a potential source lineage, estimate the time since introgression and assess the fitness burden the introgressed genes represent, showing the advantage of mechanistic models when analyzing codon data.

# Introduction

Synonymous codon usage is a reflection of the cellular environment. Mutation, selection and genetic drift can be used to quantify the cellular environemnt a genome has evolved in. Mutation bias is purely determined by the cellular environment, while the strength and efifcacy of selection relative to drift is determined by the cellular environment e.g. tRNA abundance, and the natural environment e.g. gene espression and effective population size.

Mutation, selection, and genetic drift, are three fundamental forces driving evolution, shaping the genomic environment every gene of an organism evolves in. The same forces shape the synonymous codon usage of genes. Codon usage, therefore, is a reflection of the genomic environment, giving us the opportunity to describe an organisms genomic environment in terms of its codon usage.

Link CUB and genomic environment better; first two sentences dont flow

In general, the strength of selection on codon usage increases with gene expression. Conversely, the impact of mutation bias on codon usage declines with gene expression. Thus, we can easily imagine codon usage to shift from a mutation dominated process to a selection driven process with increasing gene expression within a genome. 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 framework allows us to explicitly describe the cellular environment in which genes evolve with respect to these terms. Estimating the influence of mutation bias and selection bias on a gene also improves our understanding of its evolution; giving us the ability to describe 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 the product of a single genomic environment. While it can be argued that a cell only produces a single cellular environment - an assumption potentially violated by strand specific mutation bias and other factors [Arakawa and Tomita, 2012] - it is easy to think about the exhibition of multiple cellular environments in a cell only producing one. Genes introduced via horizontal gene transfer, introgression, or hybridization may carry the signature of a different, novel cellular environment. These transferred genes may be less adapted to the new cellular environment, with potentially large fitness consequences. We expect the fitness burden of transferred genes to be greater if donor and recipient environment differ greatly in their selection bias, making such transfers less likely. Furthermore, if unaccounted for, transferred genes may distort parameter estimates of mutation bias and selection bias describing codon usage - potentially causing us to conclude the wrong codon preference for an amino acid when analyzing a genome that has experienced such transfer events.

In this study, 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 prior to the whole genome duplication, about 100 Mya ago. Since its divergence from the other Lachancea, *L. kluyveri* has experienced a large introgression of exogenous genes. The introgression replaced the left arm of the C chromosome and displays a 13% higher GC content than the remaining endogenous *L. kluyveri* genome [Payen et al., 2009, Friedrich et al., 2015]. This makes *L. kluyveri* an ideal model to study the effects of multiple cellular environments and mismatching codon usage.

Using ROC SEMPPR Gilchrist et al. [2015] allows us to describe the cellular environment genes have evolved in by separating and estimating effects of mutation bias and selection bias, and predicting protein production rate. We use ROC SEMPPR to describe two cellular environment reflected in the *L. kluyveri* genome, a native endogenous and an introgressed exogenous environment. The separation of mutation bias and selection bias allows us to attribute the difference in GC content between endogenous and exogenous genes mostly to differences in mutation bias. Recognizing the differences in codon usage between the two gene sets also improves our ability to predict protein synthesis rate from the sequence data alone.

In addition to improvements to model fitting, we utilize the quantitative estimates of mutation bias, selection bias, and protein synthesis rate from ROC SEMPPR. First we determine a potential source lineage of the exogenous genes, comparing estimates of mutation bias ($\Delta M$) and selection bias ($\Delta\eta$) for the exogenous genes to 38 yeast lineages. 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. Third, we estimate the selective cost of the mismatched codon usage for the introgression, using our estimates of $\Delta\eta$ and protein synthesis rate $\phi$.

# Results

We compared model fits of ROC SEMPPR to the homogenous *L. kluyveri* genome and the separated set of endogenous and exogenous genes of X and 497 genes using AnaCoDa [Landerer et al., 2018]. We compared estimates of the cellular environment to describe differences in endogenous and exogenous codon usage. Furthermore, we contextualize differences in model fit and parameters estimated from the endogenous and exogenous genes.

## Separating Endogenous and Exogenous Genes Improves Model Fit and Prediction of Gene Expression

We find that the parameter estimates for mutation bias ($\Delta M$) and selection bias ($\Delta\eta$) differ significantly 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 ($40\ \Delta M + 40\ \Delta\eta + s_\phi$) necessary to describe both gene sets separately improves our model fit by $\sim 90,000$ AIC units (XXX for the combined gene set vs XXX 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 gene sets substantially improves our ability to predict protein synthesis rate $\phi$ ($\rho = 0.69$ vs. $\rho = 0.59$ for the full genome; Figure 1a,b).

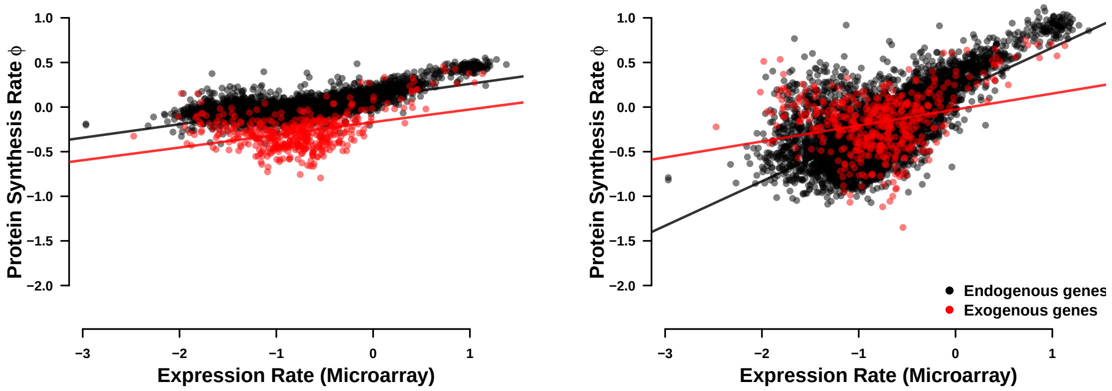


Figure 1: Put $\rho$ in plots, do I want two regression lines in plot? Only reporting overall regression so far. check if it is typeII regression

## Differences in the Endogenous and Exogenous Codon Usage

Model selection and validation confirmed that the *L. kluyveri* genome contains signatures of at least two cellular environments. We compared the quantitative estimates of mutation bias ($\Delta M$) and selection bias ($\Delta\eta$) obtained from fitting ROC SEMPPR to the endogenous and exogenous gene sets. We find larger differences between $\Delta M$ than $\Delta\eta$ (Figure 2). Estimates of $\Delta M$ in the endogenous genes negatively correlate with the exogenous genes ($\rho = -0.49$) indicating strong differences in the mutation environment between *L. kluyveri* and the donor lineage of the exogenous genes. For example, $\sim 95\%$ of codon families show mutation preference for A/T ending codons while, in contrast, the exogenous genes display an equally strong mutation bias towards C/G ending codons ($\sim 95\%$). Only the two codon amino acid Phenylalanine (Phe, F) shows complete concordance between endogenous and exogenous genes in mutation bias.

Estimates of $\Delta\eta$ show higher agreement between endogenous and exogenous genes ($\rho = 0.69$) than our estimates of $\Delta M$. For nine amino acids selection favors the same codon in endogenous and exogenous genes. Unlike the mutation bias, we find selection to be heavily biased towards A/T ending codons ($\sim 89\%$) in the exogenous genes. However, the selection environment in the endogenous genes is G/C biased ($\sim 58\%$). Thus, recognizing and

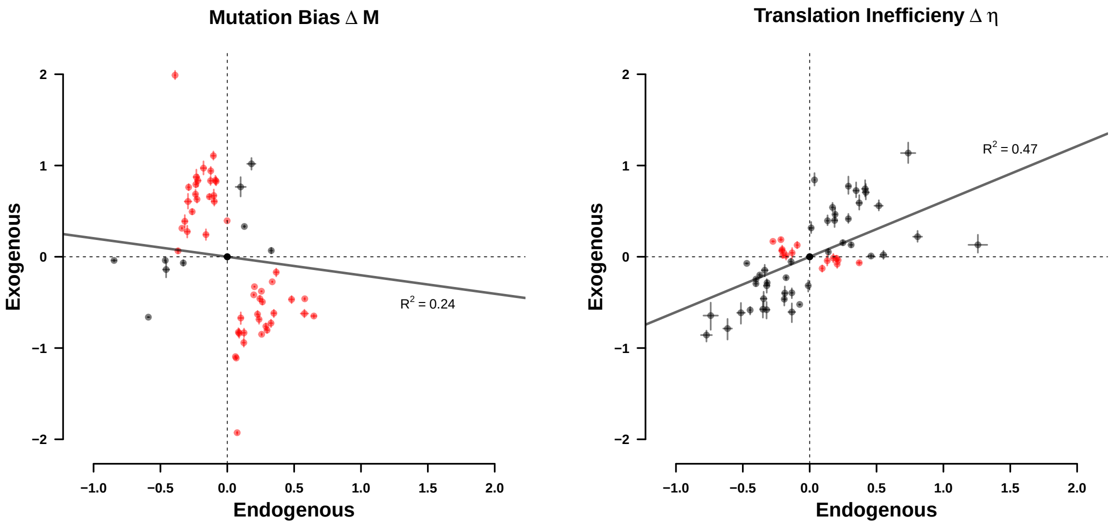


Figure 2: Parameters are relative to mean ($R^2$ doesn't make sense, change to $\rho$, check $\Delta M$ regression line, looks off), check if it is typeII regression; rename to selection bias

treating endogenous and exogenous genes as separate sets avoids the inference of incorrect synonymous codon preferences (Table S2).

## 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. Of the 38 examinded yeast lineages only two ((*Eremothecium gossypii* and *Candida dubliniensis*) showed a strong positive correlation in codon usage (Figure 3b). 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 3a). The four lineages showing a positive $\Delta M$ and $\Delta \eta$ correlation with the exogenous genes 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.

We compared synteny between the exogenous left arm of chromosome C and *E. gossypii* and *C. dubliniensis* as well as closely related yeast species using SyMAP [Soderlund et al.,

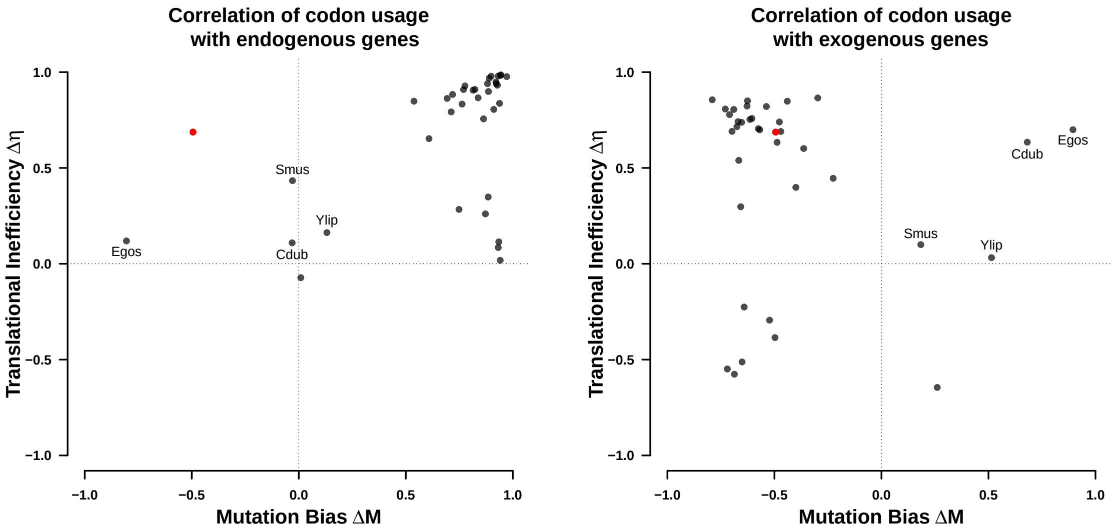


Figure 3: check if it is typeII regression

2006, 2011]. We find that *E. gossypii* has the highest synteny coverage of all examined lineages, covering nearly the whole exogenous region (Figure S1a) In contrast, *C. dubliniensis* does not have a synteny relationship with the exogenous region. Furthermore, the synteny relationship between the exogenous region and other yeasts appears to be limited to the Saccharomycetacease 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

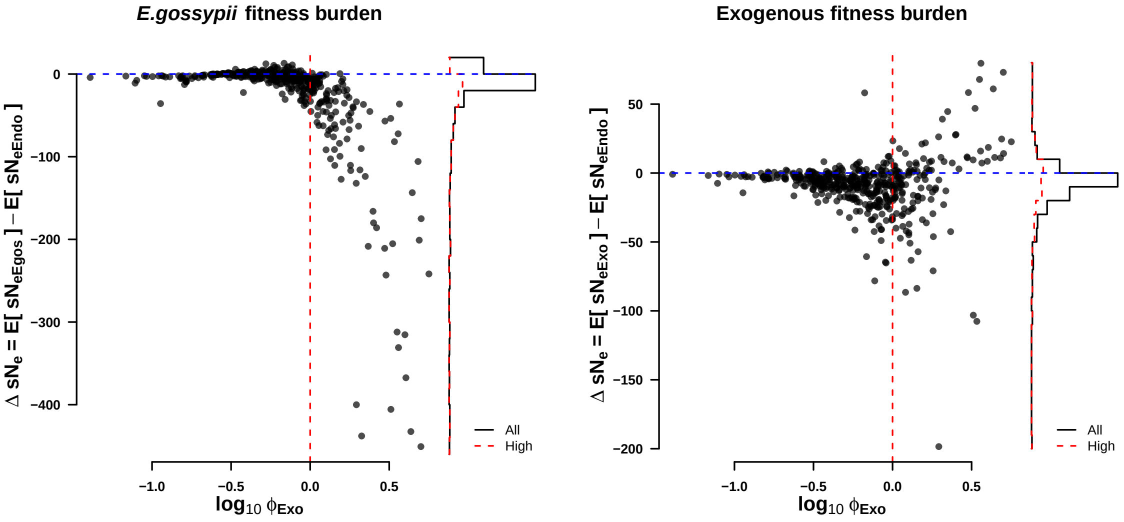
We estimated the introgression age using an exponential model of decay of mutation bias assuming that *E. gossypii* with its current mutation bias is still representative of the mutation bias in the source lineage. We utilize the $\Delta M$ estimates for all two codon amino acids and infered the age of the introgression to be on the order of $6 \times 10^8 \pm X \times 10^8$ 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 205,000-1,600,000 years ago, longer than previous estimates [Friedrich et al., 2015].

However, our estimates are likely overestimates as they assume a purely neutral decay.

Furthermore, we estimated the persistence of the foreign genomic environment. Assuming that differences in mutation bias will decay more slowly than differences in selection bias, we predict that the foreign genomic environment will have decayed to one percent of the *L. kluyveri* environment within about $5 \times 10^9$ 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 have the same codon preference. We therefore expect that the introgressed genes represent a significant reduction in fitness 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 fitness burden of the exogenous genes relative to an expected gene set.


***E.gossypii* fitness burden**

**Exogenous fitness burden**

Figure 4: Fitness burden at time of introgression (left) using scaled $\phi$, and currently (right). Simplify y-axis label, put in description

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 genomic 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 on protein synthesis rate $\phi$ (Figure S2b).

At the time of the introgression, only a few genes were weakly exapted (Figure 4a) with all high expression genes ($\phi > 1$) being mal-adapted to the novel cellular environment. However, these highly expressed genes show the greatest level of adaptationto the *L. kluyveri* cellular genomic environment (Figures 4a, S3).

# Discussion

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. 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, however, mutation is biases towards G/C ending codons in the exogenous genes. This tendency is reversed 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 nine codon families share the same optimal codon in the endogenous and egogenous gene sets. Nevertheless, we find that in both gene sets mostly the same codons are selected for, indicated by the high correlation of $\Delta\eta$ estimates between the two gene sets. However, the strength of selection within a codon family differes between gene sets, causing a change in rank order. Exceptions are e.g. XXX where XXX is favored by selection in the endogenous genes while XXX is favored in the exogenous genes. Out of the nine synonymous codon families with differing codon preferences, the entire *L. kluyveri* genome appears to share the exogenous codon preference seven times (Table S2). We find even greater discordance in our estimates of $\Delta M$ (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 further highlights the importance to recognize differences in codon usage within a genome. These 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.

Separating the endogenous and exogenous genes improves our estimates of protein synthesis rate $\phi$ by 17% relative to the full genome estimate. Furthermore, we find that the variation in our estimates of $\phi$ is more consitent 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 multiple genomic 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 aicds, the inflection point represents the point at which mutation and selection are contributing equally to the probability of a codons occurence. 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 $E[\phi] = 1$ allowing us to interpret $\Delta\eta$ as the average strength of selection relative to drift ($\overline{sNe}$) for the average gene, but also tying the mean and standard deviation of the prior distribution together. Therefore, an increase in $\phi$ for low expression genes has to be meet 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 genomic environment, we used the quantitative estimates of $\Delta M$ and $\Delta\eta$ 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 2ab). 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 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 coverage 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 a much longer time than a previous estimate by Friedrich et al. [2015] of a minimum of $55.5 \times 10^6$ generations . However, it must be highlighted that our estimate implicitly assumes neutrality and is therefore a conservative estimate, potentially overestimating the time since introgression. Our estimate also depend on the assumption that the *E. gossypii* genomic 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 genomic environments were more dissimilar.

The estimates of mutation bias $\Delta M$ also allow us the infer the time until the signature of the foreign genomic environment will have decayed. 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 foreign signature, it is important to understand the fitness consequences of such an event. In particular as it is an open question how codon usage changes. It is however, assumed that a selection has to favor shift in codon usage over a long period of time [Hershberg and Petrov, 2008], a situation clearly present in the *L. kluyveri* genome. We estimated the reduction in fitness that the exogenous genes represent assuming that the replaced endogenous genes and the new exogenous genes had a common 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 fitness cost, and show less adaptation to the novel cellular environment (Figure 4a,b, 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, as both are G/C biased. 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. In fact we have no evidence that the exogenous genes evolve faster than their endogenous counterparts. This is consistent with the wide body of work showing overall rates of change for high expression genes tend to be slower than in low expression genes. To this day, the exogenous genes represent a significant fitness burden on *L. kluyveri*. However, as the introgression appears to have reached fixation [Friedrich et al., 2015], the fitness burden relative to the replaced chromosome arm is only of theoretical interest.

Trying to think about it having reached fixation, there is no alternative cleft in the population to compete with.

The high fitness burden the exogenous genes represented at the time of the introgression indicates that the introgression was a very unlikely event to have reached fixation in a population with a large $N_e$ as it is typical for yeasts. It is hard to contextualize the probability of this introgression going to fixation 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. 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 a 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 the exogenous genes have represented a fitness increase due to external envrionemntal factors decpide their mismatch in codon usage; 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 presents of multiple genomic 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 not agnostic 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 genomic 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/` (last accessed: XX-XX-XXXX) and the annoation for *L. kluyveri* NRRL Y-12651 (assenbly ASM14922v1) from NCBI (last accessed: XX-XX-XXXX). We assigned XXX genes located on chromosome C with a location within the $\sim 1Mb$ window to the exogenous gene set. All other XXX genes of the *L. kluyveri* genome were assigned to the exogenous genes. All genes could be uniquly assigned to one or the other gene set.

## Fitting ROC SEMPPR

We used AnaCoDa [Landerer et al., 2018].

## Comparing codon specific parameter estimates

## Synteny

## Determining introgression timeline

ODE system solved with Mathematica

## Estimating fitness burden

# 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

Kazuharu Arakawa and Masaru Tomita. Measures of Compositional Strand Bias Related to Replication Machinery and its Applications. *Current Genomics*, 13(1):4, 2012. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3269016/.

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. doi: 10.1101/gr.090605.108. URL `http://genome.cshlp.org/content/19/10/1710.abstract`.

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*, page bty138, 2018. doi: 10.1093/bioinformatics/bty138. URL `http://dx.doi.org/10.1093/bioinformatics/bty138`.

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.

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. doi: 10.1534/genetics.107.071506. URL `http://www.genetics.org/content/178/1/67`.

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.

Ruth Hershberg and D A Petrov. Selection on Codon Bias. *Annual Review of Genetics*, 42 (1):287–299, 2008.

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.

| 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 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 | *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 $\Delta\eta$

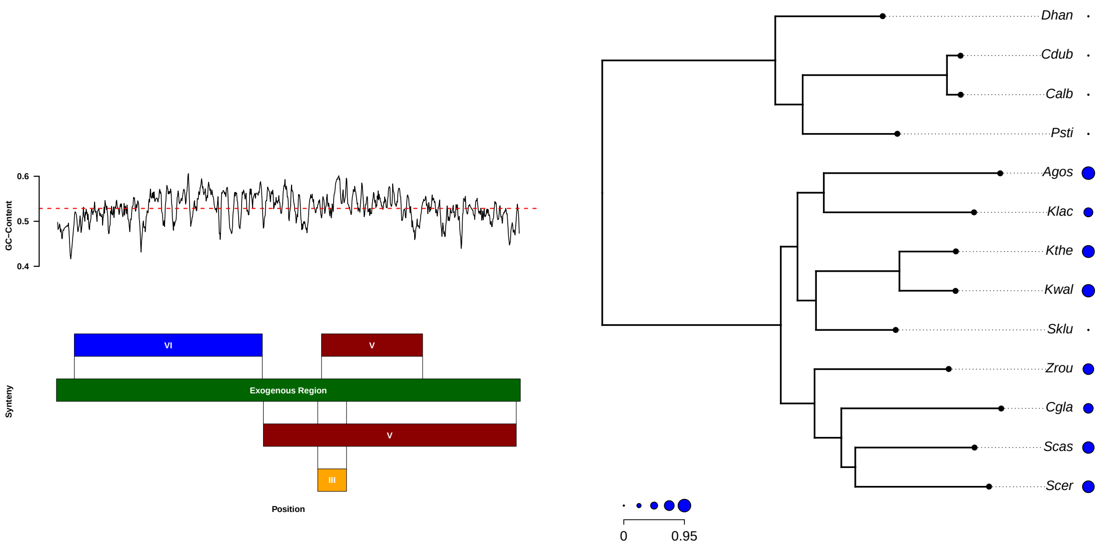


Figure S1: Suppl Fig: Synteny relationship of *E. gossypii* and the exogenous genes (left), Amount of synteny for each species (Units of std dev) checked for synteny.

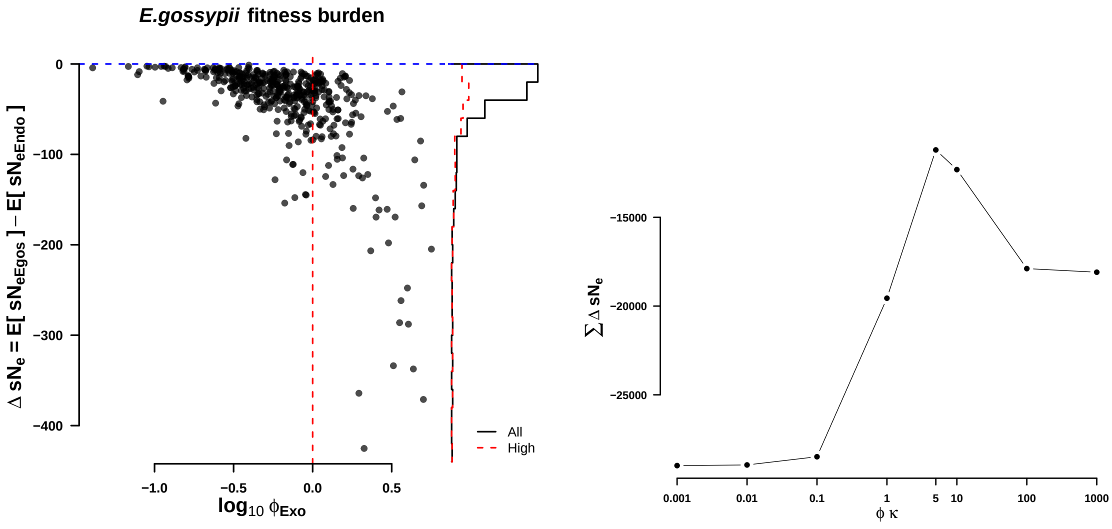
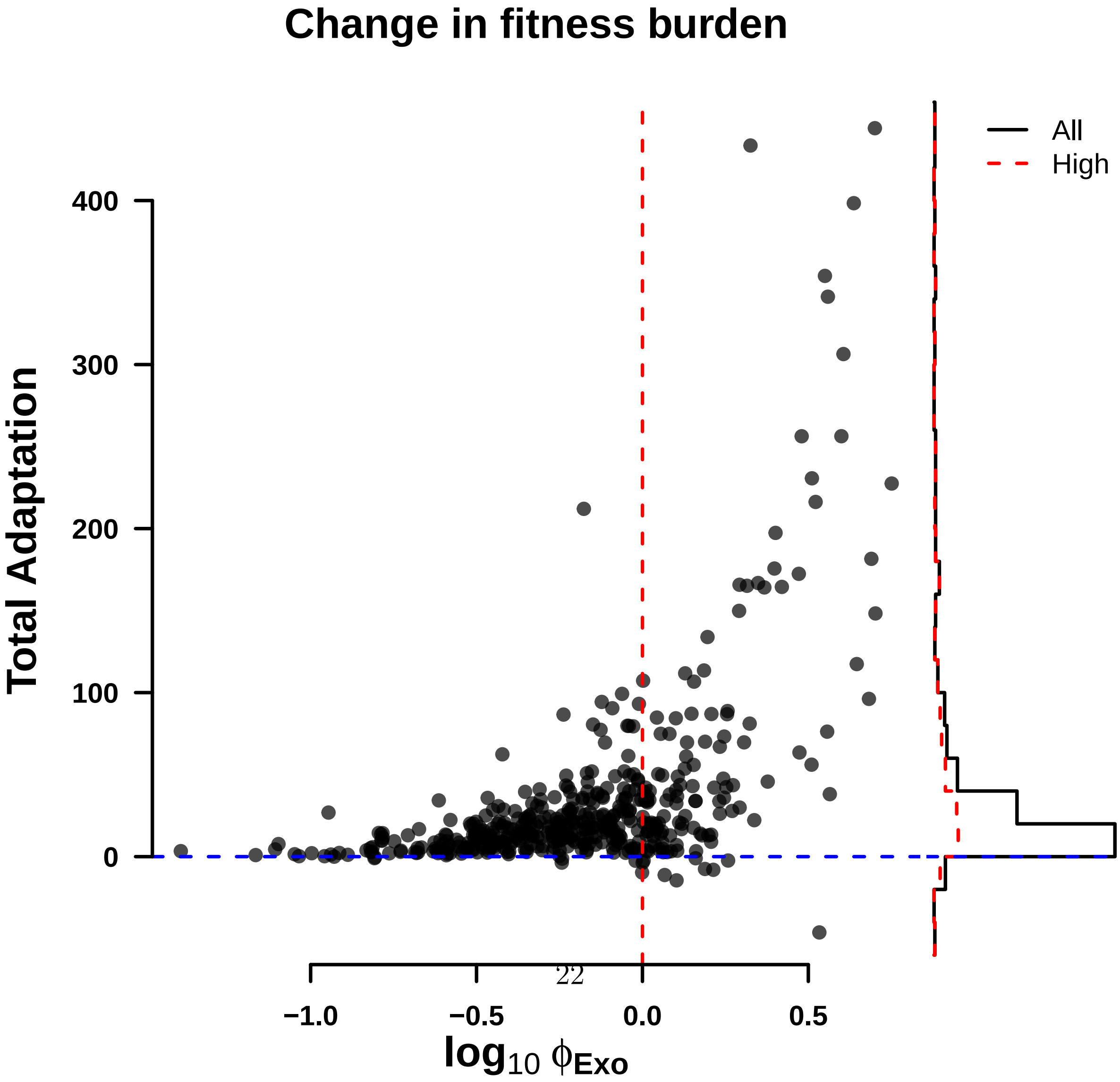


Figure S2: Suppl Fig: Fitness burden (left) without scaling of $\phi$, and change of total fitness burden with scaling $\kappa$



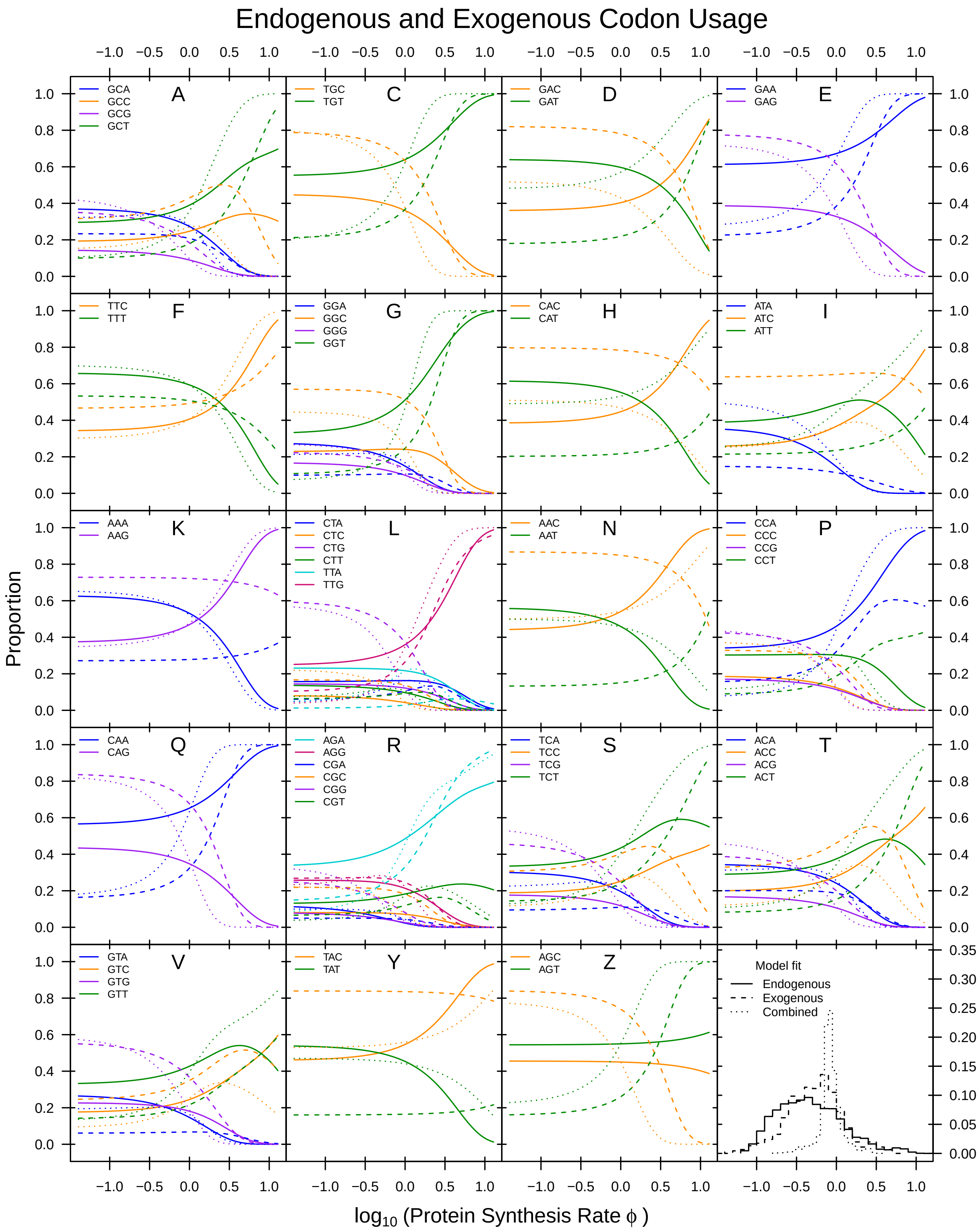


Figure S4: Suppl Fig