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

Differences in Codon Usage Bias between genomic regions in the yeast *Lachancea kluyveri*.

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: June 12, 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 be suddenly 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 gaines using ROC SEMPPR to determine a potential source lineage, estimate the time since introgression and asses the fitness burden the introgressed genes represent showing the advantage of mechanistic models have when analyzing codon data.

Introduction

- A genes codon usage reflects the genomic environment it has evolved in.
  - Mutation, selection, and drift are fundamental forces shaping the genomic environment.
  - The strength and efficacy of selection on codon usage is generally positively correlated with gene expression
  - On the other hand, the efficacy of mutation on codon usage is generally negatively correlated with gene expression.
  - Together, mutation driven bias - or mutation bias - and selection driven bias scaled by gene expression - or selection bias - shape codon usage in a genome; allowing us to describe the genomic environment in which genes evolve with respect to these terms.
  - Estimating the influence of mutation bias and selection bias on a gene improves our understanding of its evolution; giving us the ability to describe its history and making predictions about its future.
- Most studies implicitly assume that codon usage of a genome is the product of a single genomic environment.
  - This assumption however, can be violated by horizontal gene transfer, introgression, or hybridization.
  - The transfer of genes to a different genomic environment may cause them to be less adapted to the novel environment, with potentially large fitness consequences if the two genomic environments differ greatly in their selection bias, making such transfers less likely.
  - Furthermore, if unaccounted for, introgressed genes may distort parameter estimates describing codon usage and cause us to conclude wrong codon

- In this study we analyze the synonymous codon usage in the genome of *L. kluyveri*, the earliest diverging lineage of the Lachancea clade.
  - The Lachancea clade diverged from the Saccharomyces lineage prior to the whole genome duplication about 100 Mya ago.
  - Since diverging from the other Lachancea, *L. kluyveri* has experienced a large introgression of exogenous genes replacing the whole left arm of chromosome C.
  - This chromosome arm has a GC content $\sim 13\%$ higher than the endogenous *L. kluyveri* genome.
- Using ROC SEMPPR allows us to describe the genomic environment genes have evolved in by separating effects of mutation bias and selection bias, and predict protein synthesis rate.
  - We use ROC SEMPPR to describe two genomic environments reflected in the *L. kluyveri* genome, an endogenous and an exogenous environment.
    - We attribute most of the in GC bias to differences in mutation bias.
  - Recognizing differences in codon usage between endogenous and exogenous genes improves our ability to predict protein synthesis rate.
- In addition to improvements to model fitting, we show the utility of the quantitative estimates from ROC SEMPPR of mutation bias and selection bias, and protein synthesis rate by:
  - Determining a potential source lineage of the introgressed exogenous genes.
    - Comparing the estimates of ΔM and $\Delta\eta$ for the exogenous gene region to 38 yeasts and identified ancestors of *E. gossypii* and *C. dubliniensis* as most likely sources of the introgression.
  - Estimate the time since introgression and the persistence of the signal of the exogenous environment.

  - Estimate the selective cost of mismatched codon usage bias at gene and chromosome scale of the introgression.

Results

- We compared model fits of ROC SEMPPR to the full *L. kluyveri* genome, and the separated endogenous and exogenous genes.
  - We find that the partitioning of the *L. kluyveri* genome into endogenous and exogenous genes and estimating separate mutation and selection bias parameters is clearly favored ($\Delta AIC \sim 90,000$).
  - Treating endogenous and exogenous genes as separate sets:
    * improves our relative ability to predict protein synthesis rate by 17% ($\rho = 0.59$ vs. $\rho = 0.69$), (Figure 2).
    * avoids inference of incorrect codon preferences (Figure 1). .
- Larger differences in mutation bias than selection bias between the historical genomic environment of the endogenous and exogenous genes (Figure 1).
  - Mutation is biased towards A/T ending codons (11/19) in the endogenous genes and strongly biased towards C/G ending codons (17/19) in the exogenous genes.
  - Only two amino acids (A,F) showing complete concordance in mutation bias.
  - In contrast, we find the same codon preference in endogenous and exogenous genes for nine amino acids.
  - We also observe a stronger selection bias towards C/G ending codons reflected in the exogenous genes.
- We inferred potential source lineages by comparing codon usage bias and synteny of yeast lineages to the exogenous region

  - 33 out of 38 lineages showed similar selection bias
  - four lineages showed similarity in mutation bias (*E. gossypii*, *C. dubliniensis*, and *Sphaerulina musiva*, *Yarrowia lipolytica*).
  - Only *E. gossypii* and *C. dubliniensis* had a strong positive relationship in both mutation bias and selection bias.
  - synteny between *E. gossypii* and *C. dubliniensis* and closely related lineages and the exogenous genes left only *E. gossypii* as candidate.
    - Synteny with the exogenous region is limited to the Saccharomycetacease group.

- We predict the age of the introgression using an exponential model of decay.
  - We estimate the age to be about 6×10^8 generations.
  - The signature of the sources genomic environment will decay to one percent of the *L. kluyveri* genomic environment in about 5×10^9 more generations.

- We estimate the fitness burden of the introgressed region on *L. kluyveri* and compare it to the fitness burden at the time of introgression.
  - assuming constant *E. gossypii* genomic environment and amino acid usage.
  - We find that the exogenous genes:
    - were a large fitness burden at the time of introgression (Figure 4).
    - still represent a large reduction in fitness relative to the replaced expected endogenous genes (Figure 4).

Discussion

- Following Payen et al. (2009), we partitioned *L. kluyveri* into endogenous and exogenous genes using gene location.

- Estimating codon usage parameters using ROC SEMPPR, we find:
  - two gene sets show difference in mutation bias and selection bias
    - Endogenous genes tend to be generally biased towards A/T ending codons while exogenous genes are biased towards C/G ending codons.
    - We observe higher correlation between $\Delta\eta$ than ΔM, nevertheless we find the optimal codon differs between endogenous and exogenous genes in nine out of 19 synonymous codon families.
    - Without recognizing the difference in codon preference we would have inferred the preferred codon in the *L. kluyveri* genome wrong for seven amino acids.
  - We also improve our relative ability to predict protein synthesis rate when separating endogenous and exogenous genes by 17%.
- To identify a potential source lineage for the exogenous genes, we:
  - compared of ΔM and $\Delta\eta$ estimates from the exogenous genes to 38 other yeast lineages.
    - revealing 33 and five yeast lineages showing a positive relationship in selection bias and mutation bias, respectively.
  - compared synteny of the exogenous region:
    - finding consistency only within the Saccharomycetaceae clade.
    - Most of the eight species with synteny showed similarity in selection bias but not in mutation bias.
    - Only *E. gossypii* showed both synteny with the exogenous region and a similar mutation and selection bias.
- We estimated the age of the introgression to be on the order of 6.22×10^8 generation using our estimates of ΔM from the exogenous genes and *E. gossypii*.

  - Assuming a constant genomic environment in the *L. kluyveri* and *E. gossypii* genome.
  - The slower decay of mutation bias relative to selection bias also allowed us to estimate the time until the introgression will have decayed to be about 5.66×10^9 generations.
    - Differences in selection bias are expected to have decayed earlier.
    - This is consistent with the observation that HGT is more common between lineages with simlar codon preference, as most methods (CAI, tAI) focus only on selection. As a result, they are insensitive to differences in mutation bias.

- Assuming that the current amino acid sequence is representative of the ancestral, we estimate the fitness burden at the time of introgression of the exogenous genes on *L. kluyveri* compared to the endogenous genes with the expected ancestral codon usage.
  - As expected, low expression genes have a relatively small impact on fitness, and adapt more slowly to the endogenous environment.
  - Estimating the impact the exogenous genes had on the fitness of *L. kluyveri* at the time of the introgression revealed that this event was very unlikely to reach fixation.
  - 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 mutation and selection bias occur.
    - However, e.g. *L. kluyveri* diverged about 85 Mya, with one to eight generations per day, or 10^{10} to 10^{11} generations ago from the rest of the Lachancea clade. Assuming an 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.

  - We also show that the exogenous genes still represent a large decrease a fitness relative to the hypothesized ancestral endogenous genes.

- In conclusion, we show the usefulness of the separation of mutation bias and selection bias in this study of codon usage and we illustrate how ROC SEMPPR can be used for more sophisticated hypothesis testing in the future.

  - We highlight potential pitfalls when estimating codon preference, as estimates can be biased by the signature of a second, historical genomic environment.
  - In addition, we show how estimates of selection relative to drift can be obtained from codon data and used to infer the fitness cost of introgressed genes.

Figures

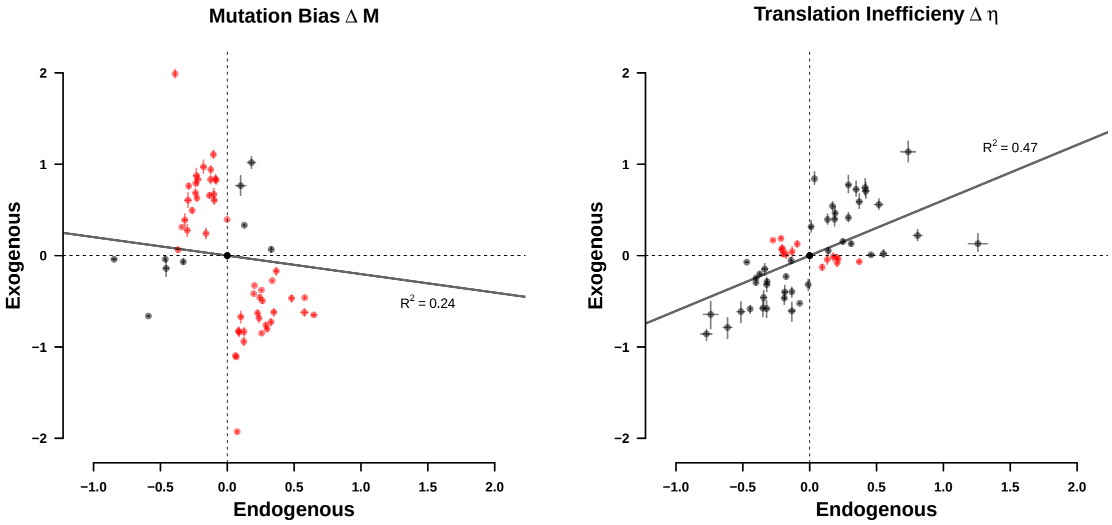

Figure 1: Parameters are relative to mean

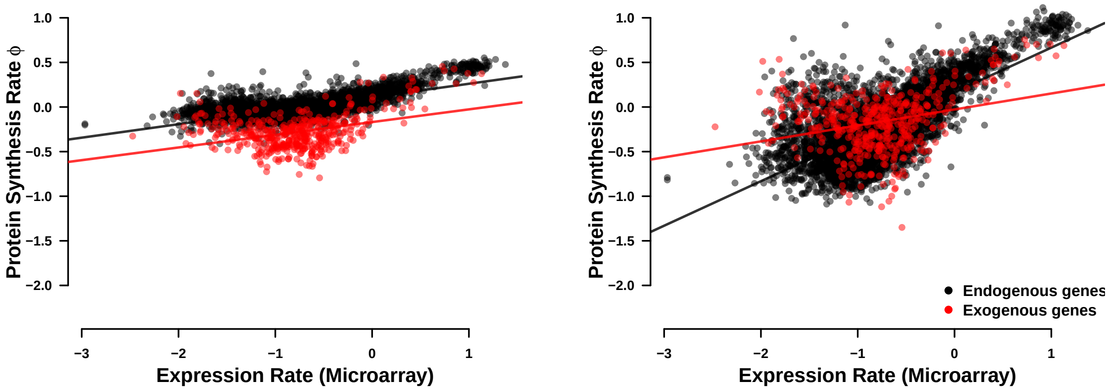

Figure 2

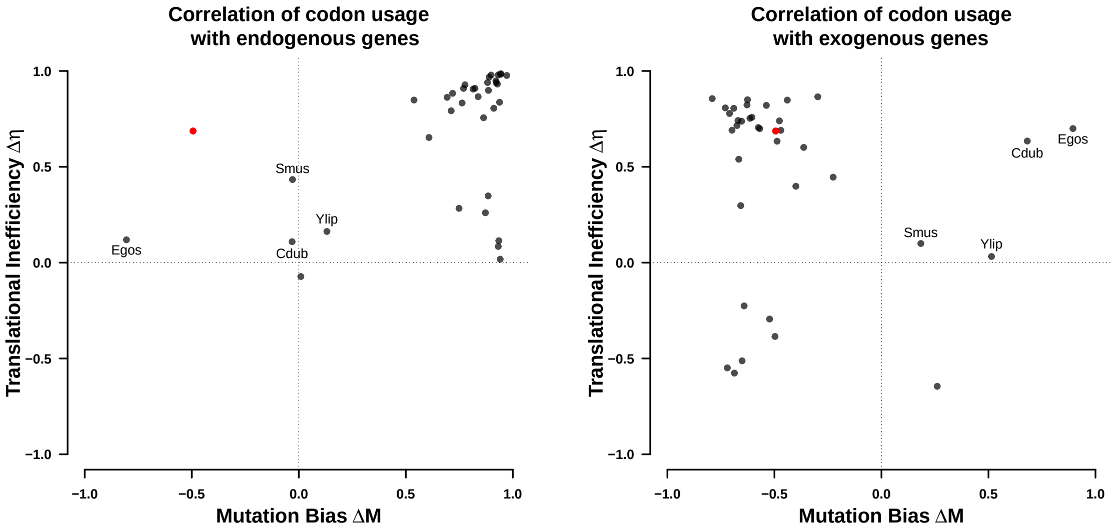

Figure 3

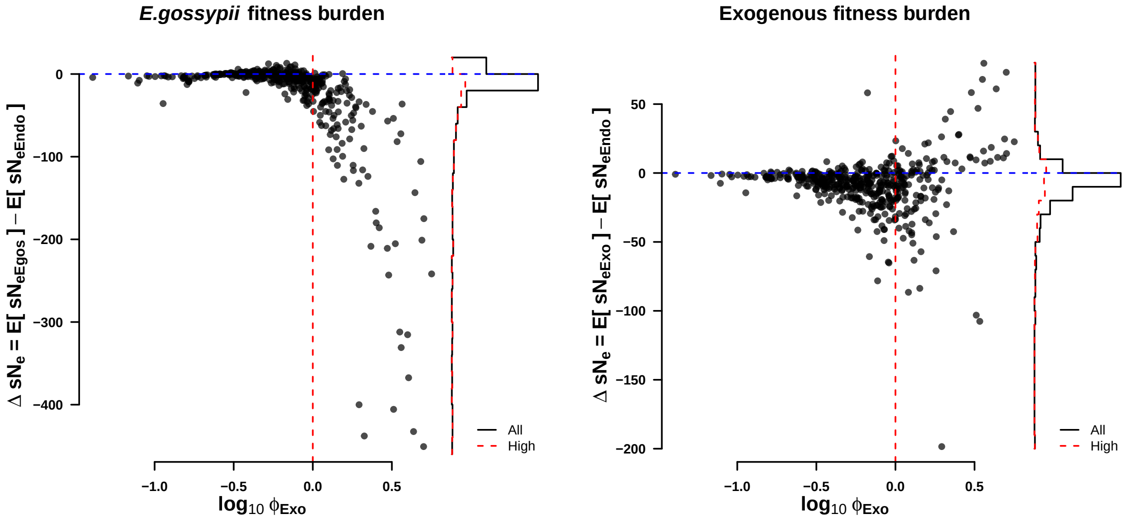

Figure 4: Fitness burden at time of introgression (left) using scaled ϕ, and currently (right).

Suppl Figures

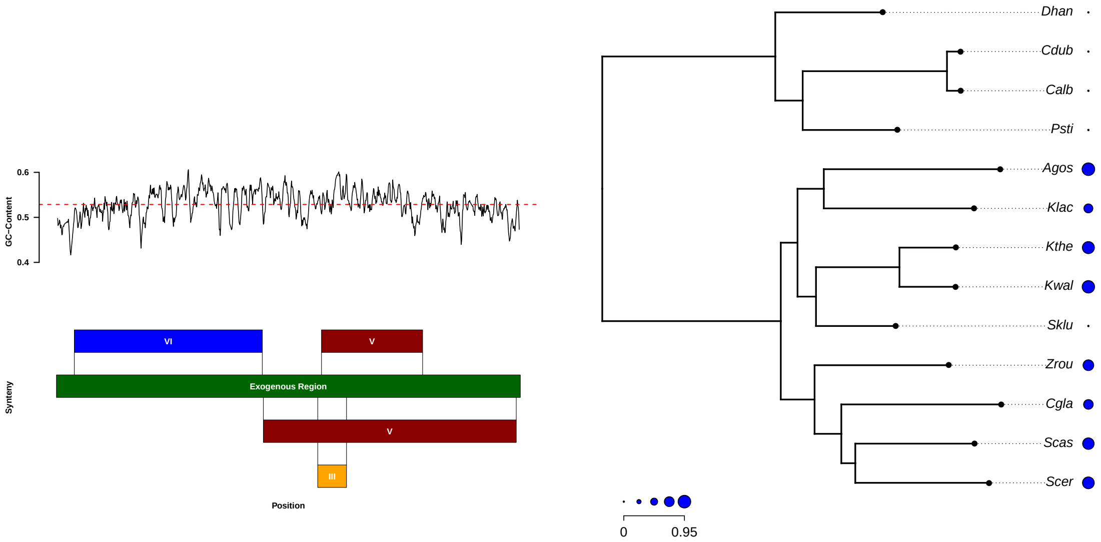

Figure 5: Suppl Fig: Synteny relationship of *E. gossypii* and the exogenous genes (left), Amount of synteny for each species (Units of std dev) checked for synteny.

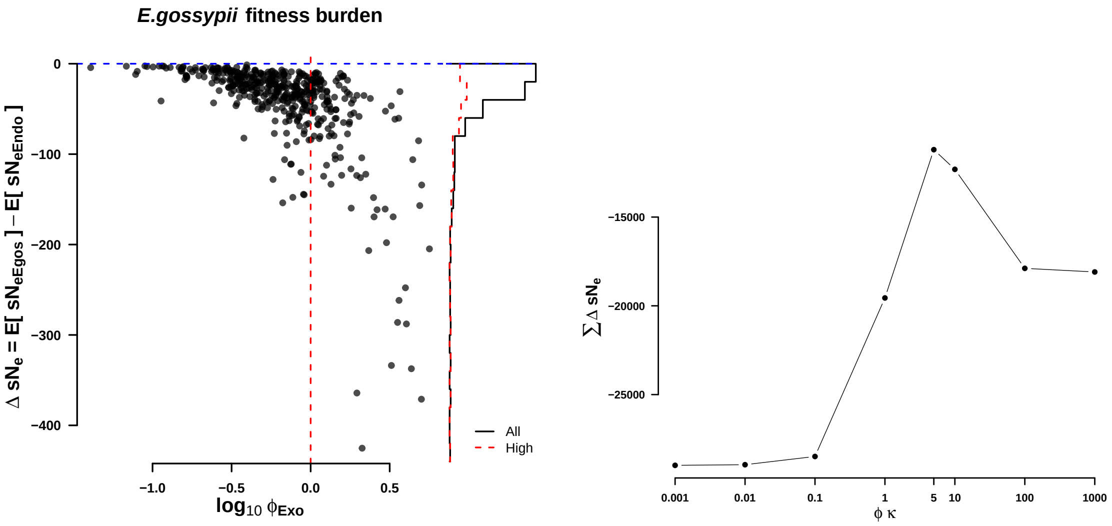

Figure 6: Suppl Fig: Fitness burden (left) without scaling of ϕ, and change of total fitness burden with scaling κ

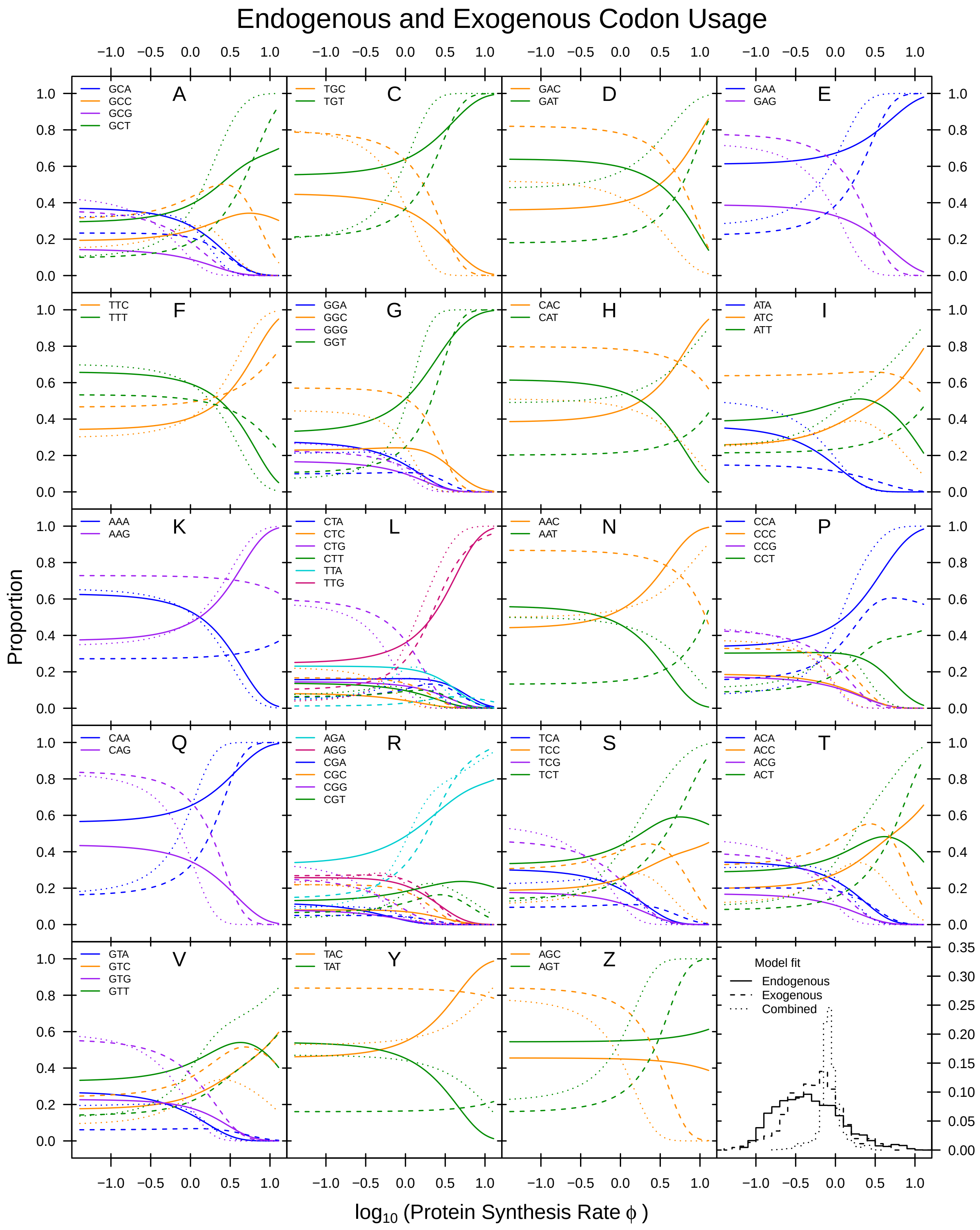

Figure 7: Suppl Fig