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: May 2, 2018

Abstract

Codon usage bias (CUB) and the contributions of mutation and selection to the evolution of CUB have been of interest for decades. Here we study the CUB of *Lachancea kluyveri* which has experienced a large introgression of the left arm of chromosome C of about 10% of it's genome. The *L. kluyveri* genome provides insights about the adaptation of introgressed regions to a novel genomic environment.

CUB of the endogenous *L. kluyveri* genome and the exogenous region were analyzed, while separating the effects of mutation bias and selection for translation efficiency on CUB. We found distinct codon preferences between the endogenous and exogenous regions of the *L. kluyveri* genome and show that these differences can be largely attributed to a shift in mutation bias from A/T to C/G ending codons.

The source of the exogenous genes has not yet been identified. We test if the shift in mutation bias is indicative of a potential source lineage. The estimation of codon preferences by mutation and selection across a variety of yeasts allowed us to identify two candidates, *Candida dubliniensis* and *Eremothecium gossypii*, as potential source lineages. Orthogonal information on synteny was used to validate the candidates we obtained using CUB.

Outline

Introduction

- CUB results from mutation, selection, and drift.
- Most studies assume that all genes have evolved in the same genomic environment.
  - This assumptions can be violated for multiple reasons, like introgression/horizontal gene transfer (HGT), population bottlenecks, etc.
- Genes with different signatures of genomic environments have previously primarily been studied in bacteria where transfer of a small number of genes via HGT is common.

- Hybridization/Introgression between species with different genomic environments can result in the misclassification of codon preference.

- In this study, we look at *L. kluyveri* which experienced a large introgression about $55.6e6$ generations ago, clearly marked by elevated GC content (13%) (three key results).

  - We found that codon preference differs between the introgressed exogenous region and the endogenous region.

    * We observe greater difference between the regions in mutation bias than in selection for translation inefficiency.

    * Taking this difference into account, we can increase our ability to extract biological information (like predicting gene expression).

  - We compared CUB parameters (ΔM and $\Delta\eta$) inferred from the exogenous genes to 45 other yeast species and identified the *E. gossypii* and *C. dubliniensis* lineages as likely sources of the exogenous genes.

    * A analuyses of synteny revealed that *C. dubliniensis* does not show any synteny, leaving *E. gossypii* as potential source.

  - We estimated the introgression occurred about $5e8$ generations ago.

Results

- We compared model fits of ROC SEMPPR to different partitionings of the *L. kluyveri* genome.

  - AIC clearly favored varying codon preferences between the endogenous and exogenous region of the *L. kluyveri* genome.

  - Prediction of protein synthesis ϕ was improved by varying codon preference (ρ: 0.59 vs 0.69) (Figure 1).

- Posterior estimates of codon specific parameters (ΔM and $\Delta\eta$) between regions show a negative correlation for ΔM and a positive correlation for $\Delta\eta$ (Figure 2).
  - Mutation preference in the two regions is distinct with the exception of the amino acids A,F which favor the same codon.
  - Selection preference overlaps in nine cases between the endogenous and exogenous region.
- ROC SEMPPR was used to infer ΔM and $\Delta\eta$ for several yeasts species.
  - We found three species with agreement in mutation bias (ΔM) and 33 species with agreement in selection bias ($\Delta\eta$) (Figure 5).
  - All three species identified via mutation bias, *E. gossypii* and *C. dubliniensis* and *Sphaerulina musiva*, showed agreement in selection bias (Figure 5).
- We validated our candidate list with orthogonal information on synteny.
  - We found eight species with synteny but only *E. gossypii* was supported by CUB (Figure 6).
- We estimated the time since the introgression occurred to be $3.32e8$ generations.
  - Our estimates overlap with the estimates of [1] ($19k$-$150k$ years), overlap is $114k$-$150k$ years.
  - Decay of the signature of the source environment to one percent of the *L. kluyveri* environment will occur in about $5.37e9$ generations.

Discussion

- Partitioning *L. kluyveri* based on the previously identified introgression allowed us to identify two distinct signatures of genomic environments, shaping CUB.

- We find that the endogenous region shows mutation bias towards T and A ending codons, the exogenous region is mutationally biased towards C and G ending codons
- We observe higher correlation between $\Delta\eta$ nevertheless we find the optimal codon differs between in endogenous and exogenous regions for most amino acids.
- Ignoring the difference in genomic environment between endogenous and exogenous region can lead to miss-classification of the optimal codon (D, H, I, S, V).
- Recognizing that ΔM and $\Delta\eta$ vary between endogenous and exogenous genes improves our ability to predict protein synthesis rate ϕ.

- We propose *E. gossypii* lineage as the source of the introgression.
  - We identified 33 yeast lineages with genomic environments resulting in similar $\Delta\eta$ of which three lineages show similar mutation bias as well.
  - Mutation bias is more informative: it would decay slower and most yeast species analyzed have similar selective environments, yet endogenous and exogenous genes differ in their optimal codon.
  - Synteny of the exogenous region was consistent with eight species, all within the Saccharomycetaceae group, none in the sister clade Debaryomycetaceae.
  - Only *E. gossypii* showed synteny with the exogenous region and a similar CUB.
- We estimated a time since introgression, assuming the *E. gossypii* lineage as source ($3.32e8$ generations) and the time until the signature of the source environment will have decayed to one percent to be $5.37e9$ generations.
  - We assume that *E. gossypii* exhibits the same genomic environment as it did during the transfer of the exogenous region to *L. kluyveri*.
  - Finding two amino acids with a negative estimated introgression time indicate that this assumption is violated.

- In conclusion, this study shows:
  - More than one set of optimal codons can be present in a genome, due to introgression, or other, internal factors; and recognizing it can prevent misclassification of optimal codons.
  - It is well accepted that CUB is driven by mutation, selection, and drift.
    * Here we illustrate again that it is important to include mutation, and that the mutation in CUB (disregarded by other approaches like CAI) provides valuable information.
  - While we used CUB to determine a potential origin of the exogenous region, this is just an example on how CUB and ROC SEMPPR can be used for more sophisticated hypothesis testing in the future.

References

[1] 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.

Figures and Tables

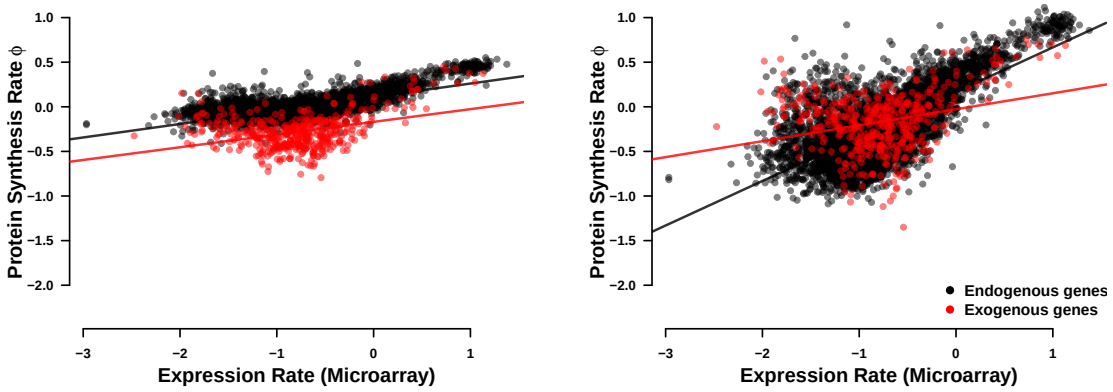

Figure 1: Person correlation of predicted protein synthesis rate ϕ with observed expression rate

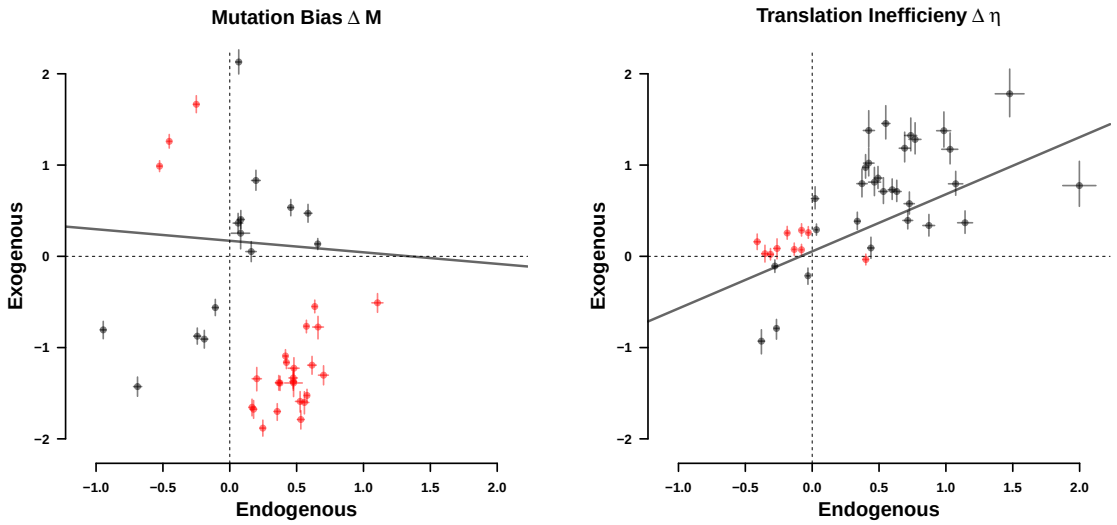

Figure 2: Person correlation for CUB parameters estimated from endogenous and exogenous genes (red = opposite sign, black = same sign)

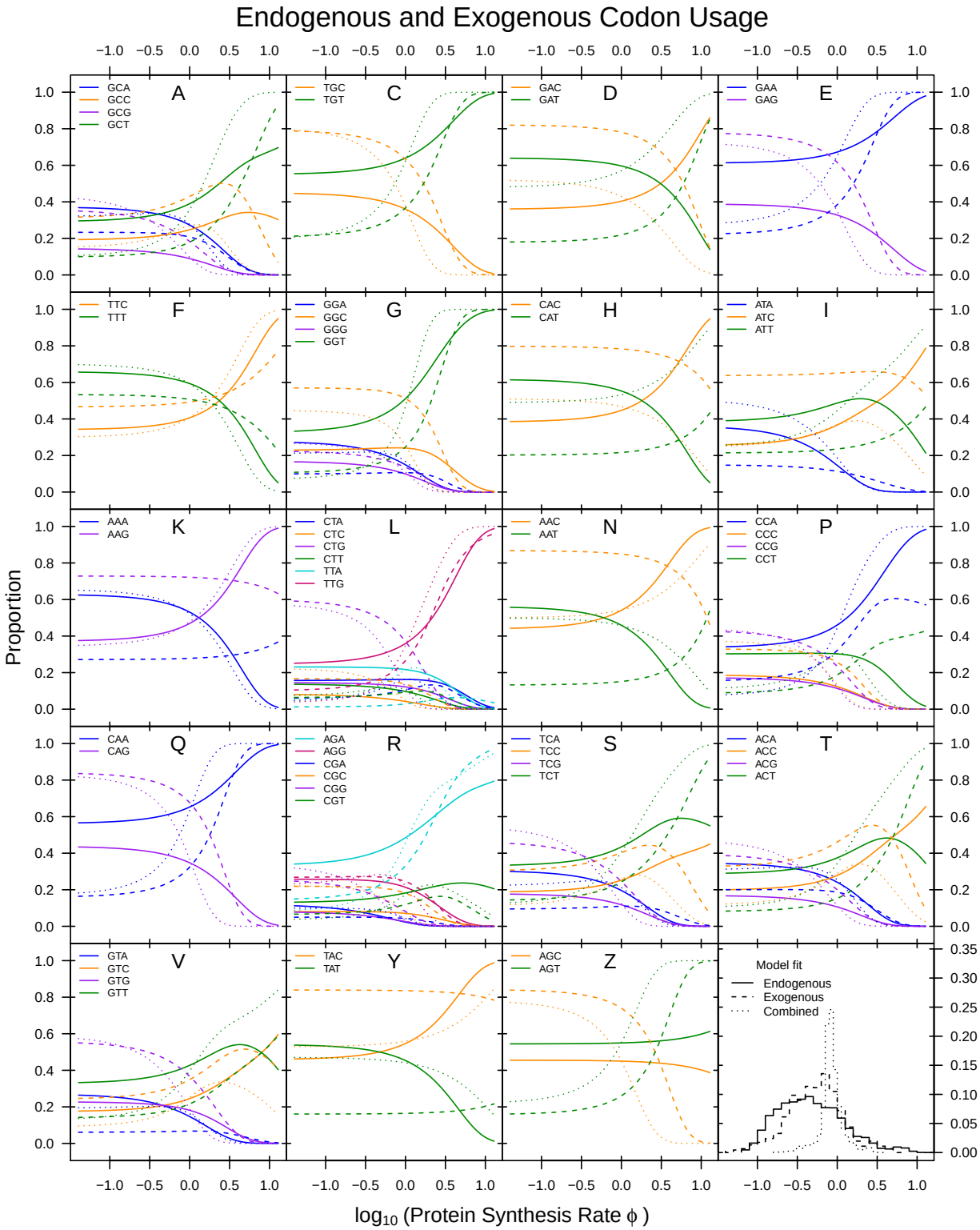

Figure 3: Codon Usage. Modify figure to indicate whether same AA is optimal in endogenous/exogenous region?

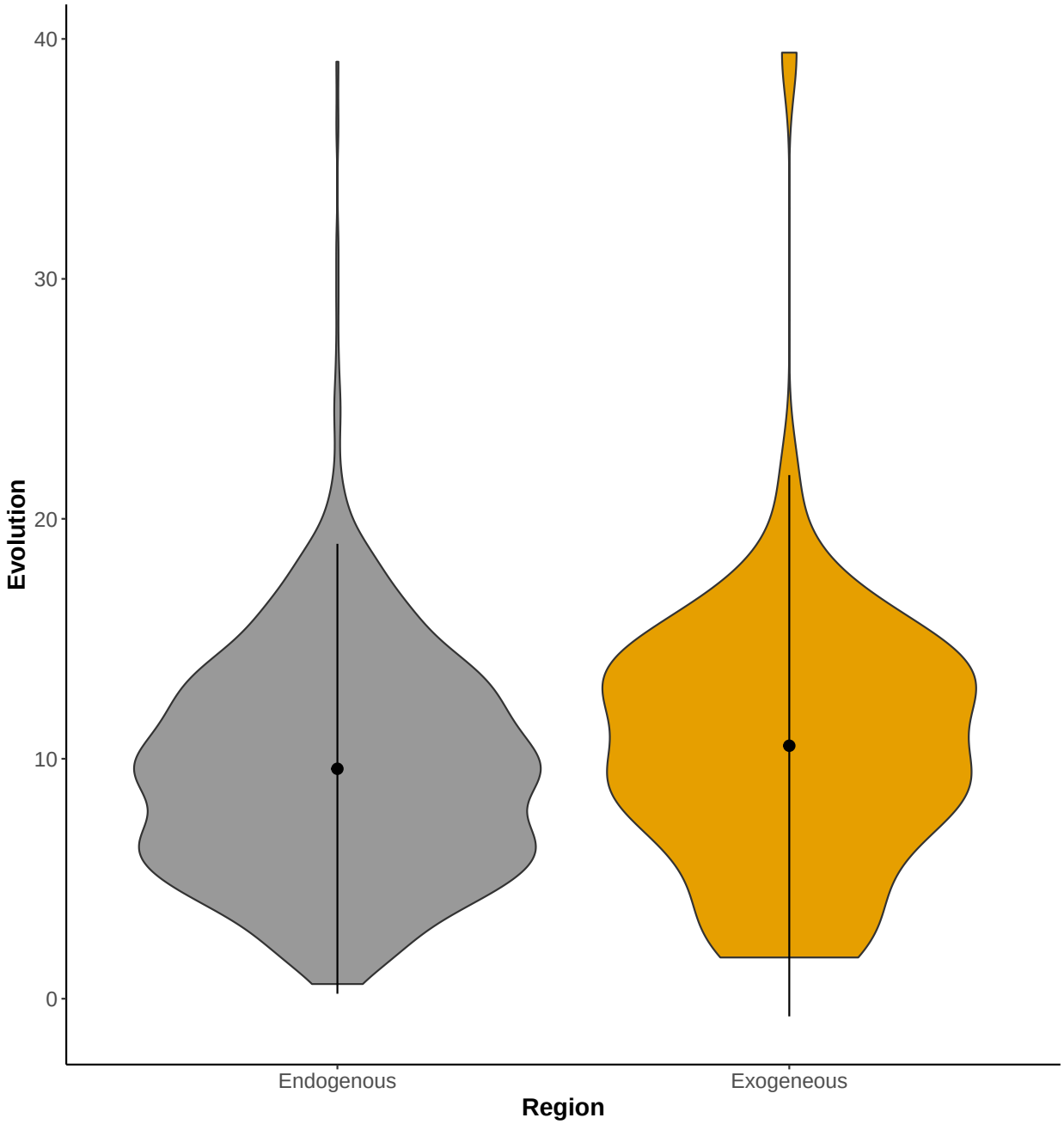

Figure 4: Overall time passed along gene tree

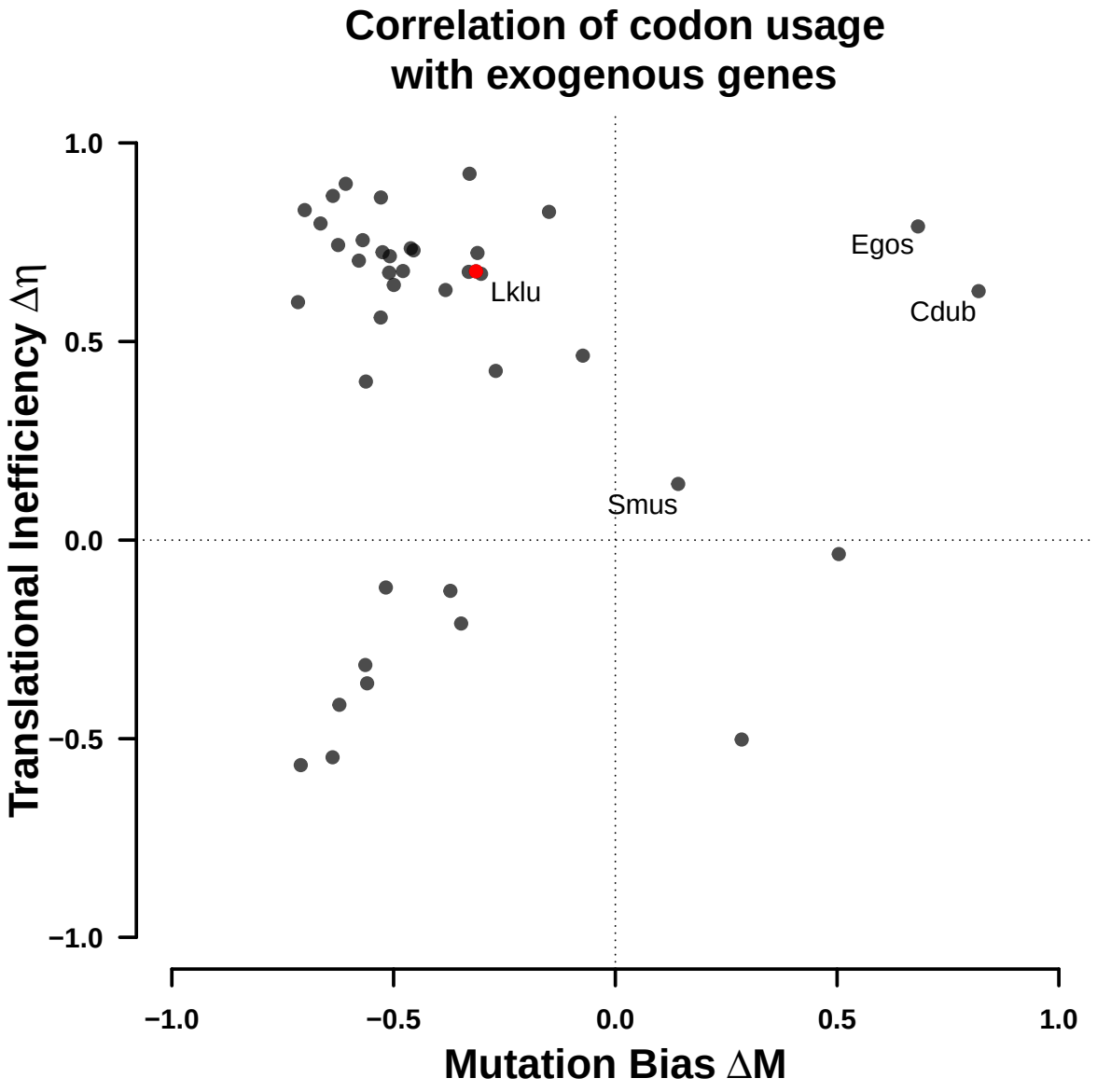

Figure 5: Codon Usage

Codon	Amino Acid	ΔM_{Egos}	ΔM_{Endo}	ΔM_{Exo}	T_{Intro}	T_{decay}
TGC	Cys (C)	-3.28	0.20	-1.34	$4.81e8$	$5.61e9$
GAC	Asp (D)	-2.57	0.58	-1.26	$2.88e8$	$4.79e9$
GAA	Glu (E)	2.47	0.45	1.26	$6.30e8$	$4.45e9$
TTC	Phe (F)	-1.46	0.66	0.14	$1.19e8$	$4.42e9$
CAC	His (H)	-2.31	0.48	-1.37	$2.41e8$	$4.96e9$
AAA	Lys (K)	0.96	-0.53	0.99	$-2.78e7$	$6.67e9$
AAC	Asn (N)	-1.28	0.25	-1.88	$-2.54e8$	$5.03e9$
CAA	Gln (Q)	2.98	-0.25	1.67	$3.57e8$	$6.68e9$
TAC	Tyr (Y)	-1.92	0.17	-1.65	$1.01e8$	$5.43e9$
AGC	Ser_2 (Z)	-3.11	0.18	-1.68	$3.10e8$	$5.63e9$
				Mean:	$3.32e8$ ($4.5e8$)	$5.37e9$ ($5.25e9$)
				Std Error:	$1.24e8$ ($1.07e8$)	$8.10e8$ ($2.38e8$)

Table 1: Mutation rate is $3.8e-10$ (Lang 2008), ignoring negative values in parenthesis. Decayed to 1%.

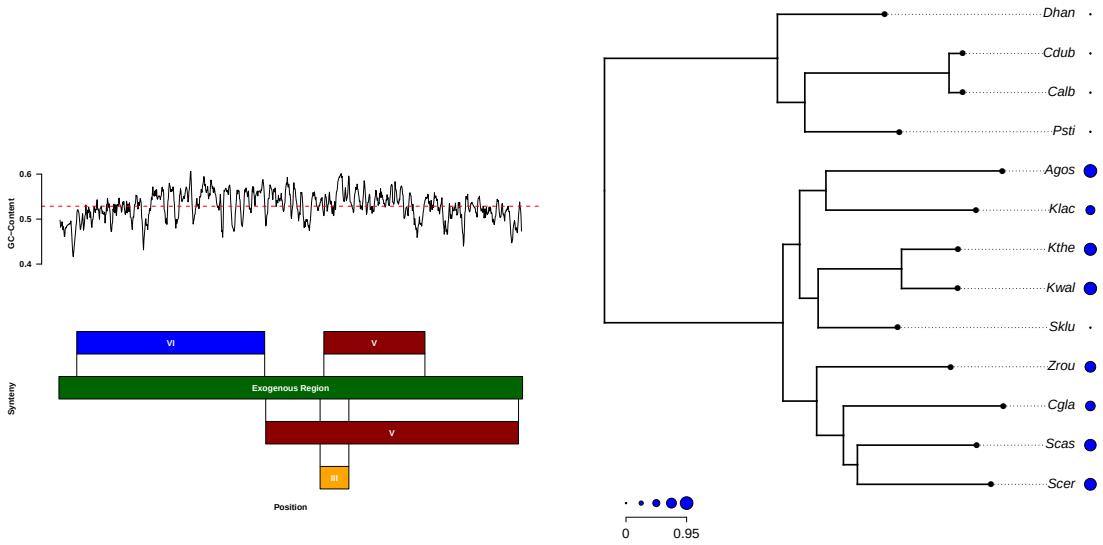

Figure 6: Synteny stuff