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: April 25, 2018

Abstract

Large efforts have been made to develop and explore models to understand intra-genomic variation in codon usage bias (CUB) and the contributions of mutation and selection to its evolution. Comparative studies have been undertaken to further our understanding of variation in codon usage between species. However, limited efforts have been made to understand how CUB is affected by, and in return effects hybridization or introgression events between species with potentially large differences in CUB. In this study, we explore the CUB of *Lachancea kluyveri* which has experienced a large introgression covering the whole left arm of chromosome C, affecting about 10% of all genes. The *L. kluyveri* genome provides insights about the adaptation of introgressed regions to a novel genomic environment, with potentially large differences in selection for translation efficiency due to factors like tRNA availability, effective population size, or differences in mutation environment.

We analyzed the CUB of the endogenous *L. kluyveri* genome and compared it to the CUB of the exogenous, introgressed, region while separating the effects of mutation bias and selection for translation efficiency on CUB. Our results show distinct CUB between the endogenous and exogenous regions of the *L. kluyveri* genome and we show that this differences can be mostly attributed to differences in mutation bias.

The introgression into the *L. kluyveri* genome is of additional interest as the source has not yet been identified. We explored if the understanding about CUB evolution gained in this study can be used to identify possible candidates for the origin of the introgression experienced by *L. kluyveri*. The estimation of CUB and its separation into contributions of mutation and selection across a variety of yeasts allowed us to identify two candidates, *Candida dubliniensis* and *Eremothecium gossypii*, for the origin of the exogenous genes. We used orthogonal information on synteny 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 environment for mutation, selection and drift with differing impact.
  - This assumptions can be violated for multiple reasons, like introgression/horizontal gene transfer (HGT), population bottlenecks, etc.
- Genes with different signatures of CUB environments have previously only been studied in bacteria where HGT is common.
  - HGT only transfers small number of genes, probably with little to no impact on overall CUB.
  - Hybridization/Introgression between species with different CUB environments should have a larger impact on CUB due to the amount of material transferred, possibly affecting the outcome of a study if ignored.
- In this study, we look at *L. kluyveri* which experienced an introgression, clearly marked by elevated GC content (13%) (three key results).
  - CUB differs between the introgressed exogenous region and the endogenous region.
    - Taking this difference into account, we can increase our ability to extract biological information (like predicting gene expression).
    - We observe greater difference between the regions in mutation bias than in selection for translation inefficiency.
  - We compares CUB parameters (ΔM and $\Delta\eta$) inferred from the exogenous genes to 45 other yeast species and identified *E. gossypii* and *C. dubliniensis* as potential origin of the exogenous genes.

    - ∗ Validation of our inference with synteny revealed that *C. dubliniensis* does not show any synteny, leaving *E. gossypii* as potential origin.
  - Assuming *E. gossypii* as origin for the exogenous region, we estimated a time since introgression from our estimates of mutation bias ($5e8$ generations).

Results

- We compared model fits of CUB for *L. kluyveri*.
  - Model selection by AIC favored varying CUB between the endogenous and exogenous region of the *L. kluyveri* genome.
  - Comparison of predicted protein synthesis ϕ of both fits with empirical estimates showed that varying CUB improved our ability to predict ϕ (ρ: 0.59 vs 0.69) (Figure 1).
- Comparison of posterior estimate of codon specific parameters (ΔM and $\Delta\eta$) between regions shows a negative correlation for ΔM and a positive correlation for $\Delta\eta$ (Figure 2).
  - Only two amino acids (A,F) favor the same codon by mutation in the two regions.
  - Nine amino acids share a preferred codon between endogenous and exogenous region
- CUB for several yeasts species was explored to determine if another yeast shows similar CUB and could have given rise to the exogenous region.
  - Comparison of CUB parameters yielded three species with agreement in mutation bias (ΔM) and 33 species with agreement in selection bias ($\Delta\eta$) (Figure 5).
  - Only three species, *E. gossypii* and *C. dubliniensis* and *Sphaerulina musiva* showed agreement in both, ΔM and $\Delta\eta$ (Figure 5).

- Synteny was used as an independent approach in an attempt to validate our candidate list.
  - The check revealed eight species but only *E. gossypii* was also supported by CUB (Figure 6).
- Assuming the exogenous region originated from *E. gossypii* ancestor, we estimated the time since introgression.
  - Based on the difference in mutation bias ΔM between *E. gossypii* and the endogenous region we estimated a time since introgression of $3.32e8$ generations.
  - Our estimates overlap with the estimates of [1] ($19k$-$150k$ years), overlap is $114k$-$150k$ years.
  - We also estimated that the exogenous regions CUB will have decayed to one percent of the endogenous CUB in about $5.37e9$ generations.

Discussion

- Partitioning *L. kluyveri* based on the previously identified introgression allowed us to identify two distinct signatures of CUB environments.
  - 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
  - While we find higher correlation between $\Delta\eta$ in both regions, most amino acids do not share their optimal codon
  - Ignoring the difference in CUB environment between endogenous and exogenous region can lead to miss-classification of the preferred codon (D, H, I, S, V).
  - Allowing CUB to vary between endogenous and exogenous genes also allows us to improve our ability to predict protein synthesis rate ϕ.

- We propose *E. gossypii* as the source of the introgression.
  - The estimation of CUB parameters for several closely related yeast species revealed 32 species that show agreement with the selective CUB component but only three with a similar mutation component.
  - Mutation bias is more informative: it would decay slower and most yeast species analyzed have similar selective environments.
  - This shows that the information about the mutation component in CUB (disregarded by other approaches like CAI) provides valuable information about the evolution of CUB and should not be ignored.
  - The check for synteny revealed eight species, all within the Saccharomycetaceae group, non 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 *E. gossypii* as origin ($3.32e8$ generations) and the time until CUB between endogenous and exogenous would be indistinguishable ($5.37e9$ generations)
  - Our approach assumes that gossypii has not evolved since 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.
  - If the exogenous region truly originated from gossypii, we can assume that the time since introgression is actually more recent than our estimate, bringing it closer to the estimate of [1].
- In conclusion, this study shows three things:
  - More than one CUB can be present in a genome, due to introgression, or other, internal factors; and ignoring it can lead to misinterpretation of results.

- It is well established that CUB is driven by Mutation, Selection, and Drift; Here we illustrate again that it is important to utilize all three factors to gain a complete picture.
- While we used CUB to determine a potential origin of the exogenous region, this is just an example using the better understanding of CUB evolution we gained in this study.

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

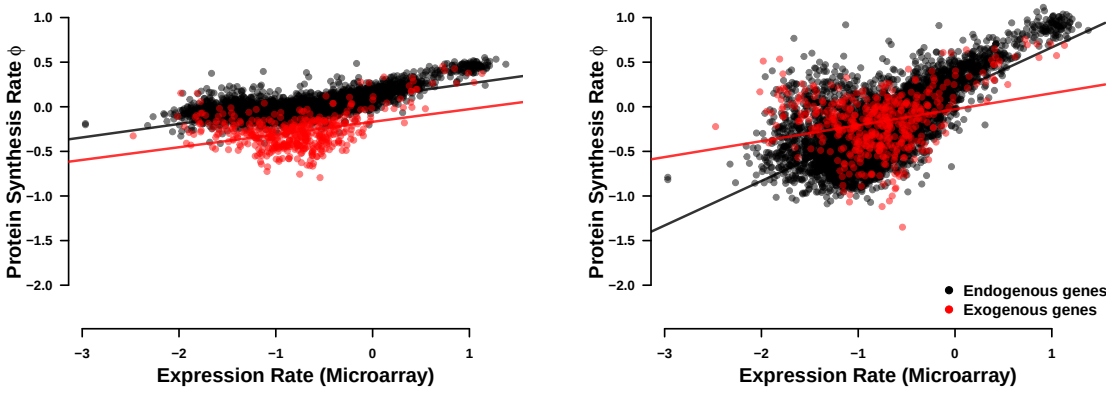

Figure 1: Person correlation of predicted protein synthesis rate ϕ with observed expression rate

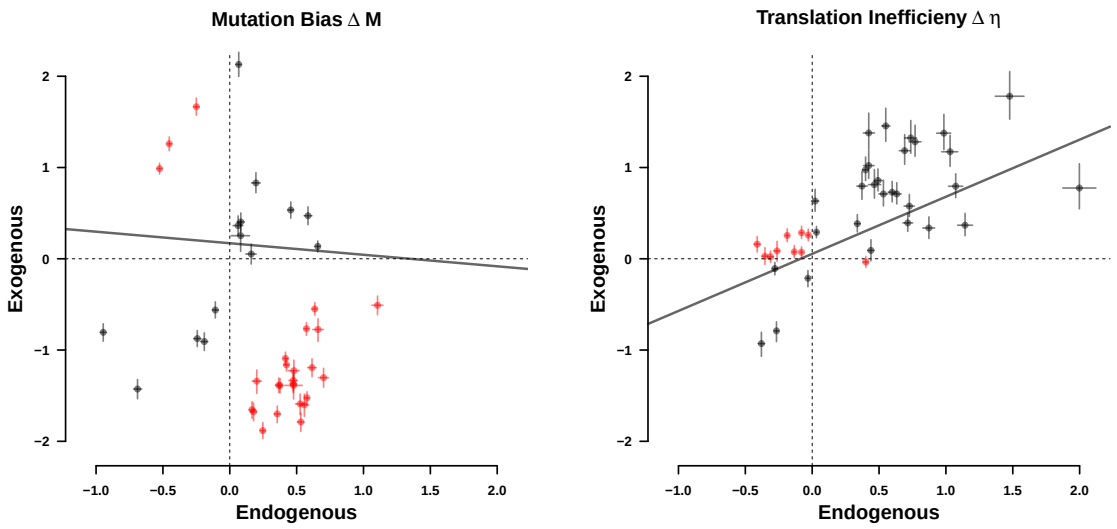

Figure 2: Person correlation for CUB parameters estimated from endogenous and exogenous genes (red = opposite sign, black = same sign)

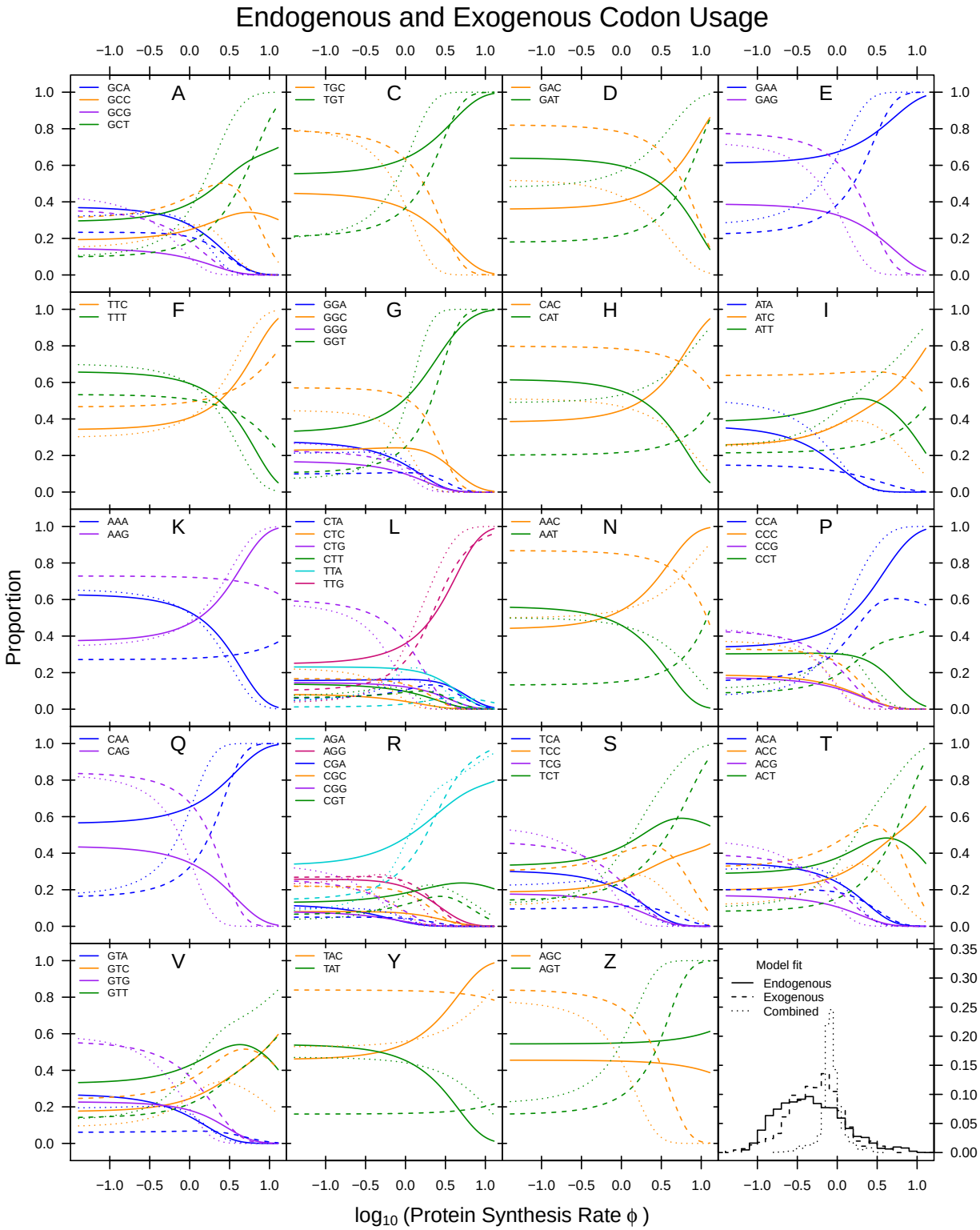

Figure 3: Codon Usage. Modify figure to indicate whether same AA is optimal in endogenous/exogenous region?

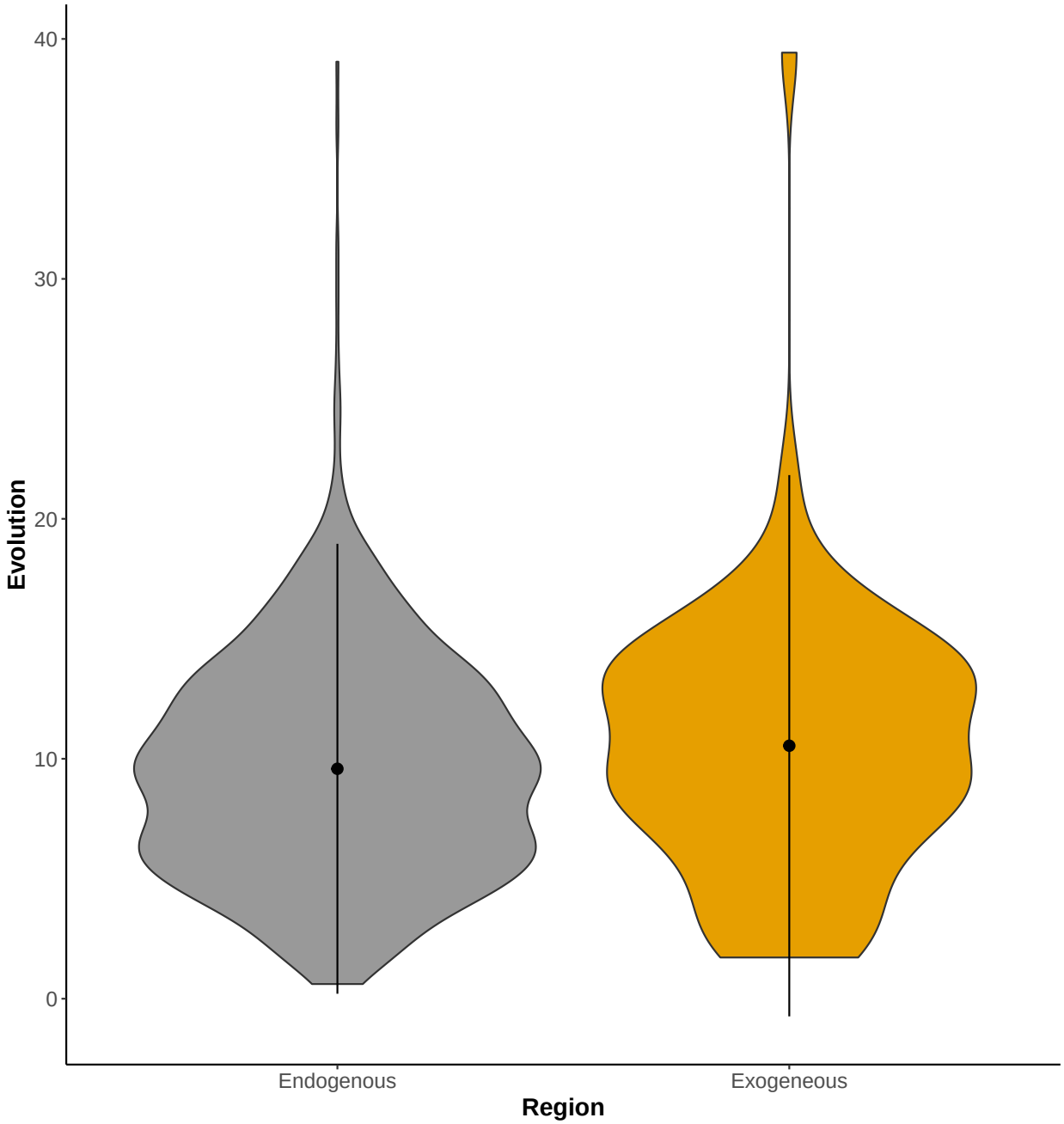

Figure 4: Overall time passed along gene tree

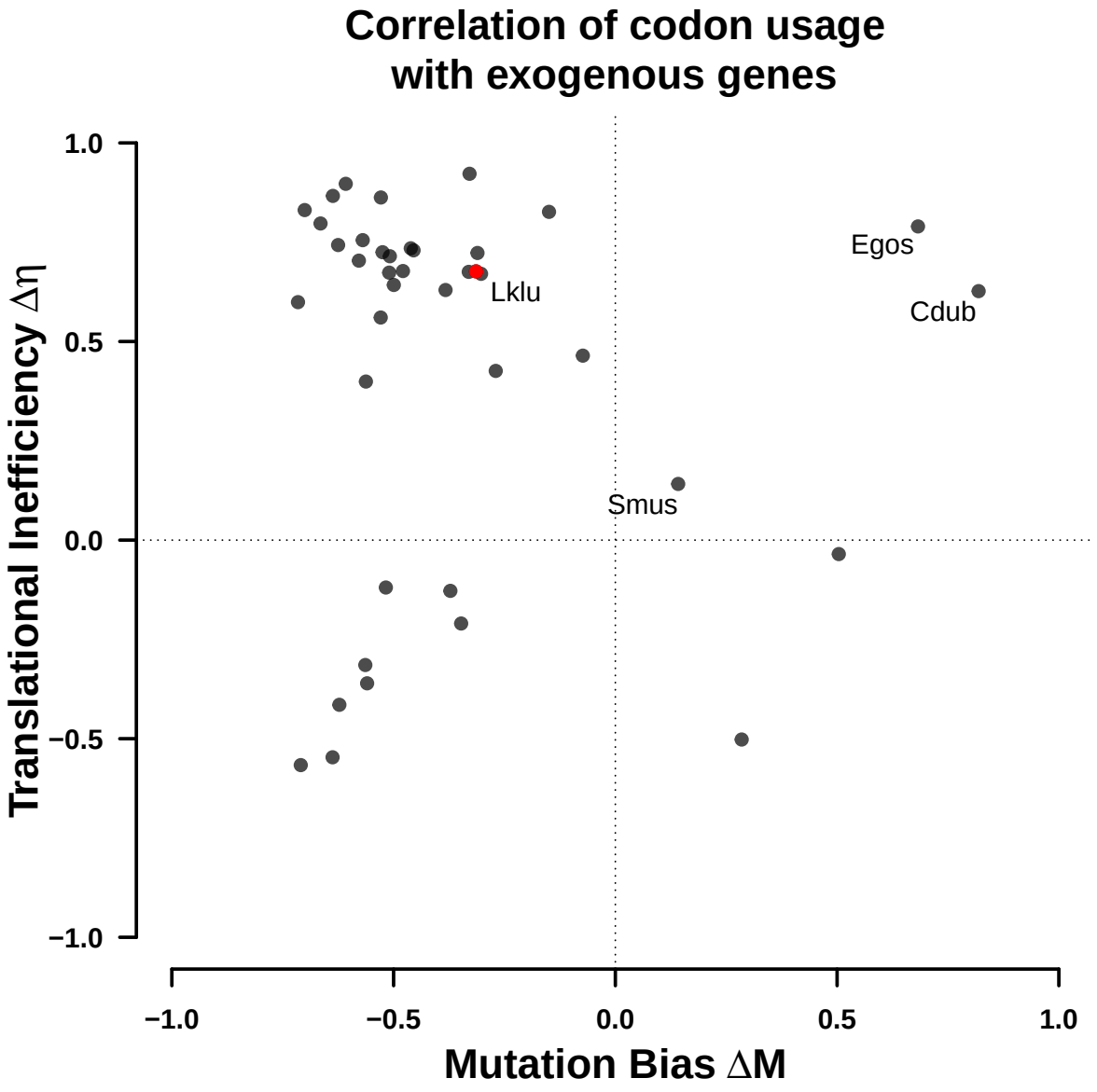

Figure 5: Codon Usage

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%.

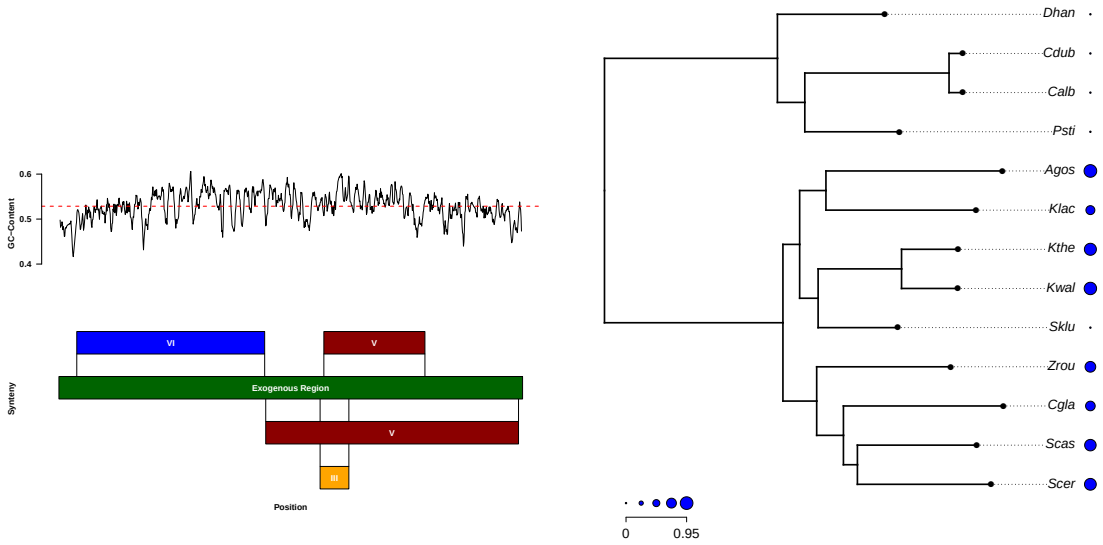

Figure 6: Synteny stuff