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 17, 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, 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 the 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. 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. Given our ability to clearly distinguish CUB between the exogenous and the endogenous region we explored if CUB can identify possible candidates for the origin of the introgression. The estimation of CUB and its separation into contributions of mutation and selection across a variety of yeasts allowed us to identify two candidates for the origin of the exogenous genes. We used orthogonal information about synteny to validate candidates obtained by matching CUB.

Outline

Introduction

- CUB changes due to differences in mutation, selection, and drift.
- most studies assume only one environment for mutation, selection and drift and therefore only one codon usage.
  - This assumptions can be violated for multiple reasons, like introgression/horizontal gene transfer (HGT), population bottlenecks, etc.
- Variation in CUB has previously only been studied in bacteria where HGT is common.
  - HGT only transfers small amount of genes, probably with little to no impact on overall CUB.
  - However, exogenous material can accumulate if HGT is frequent [2].
  - Previous studies have shown that genes with similar CUB are more likely to be transferred, potentially mitigating effects of accumulation [4].
  - Hybridization/Introgression 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* (three key results).
  - *L. kluyveri* has experienced a recent ($55.5e10$ generations) large scale introgression [1], clearly marked by elevated GC-content [3].
  - We expect that CUB differs between the introgressed exogenous region and the endogenous region due to the great (13%) difference in GC-content between the two regions.
    - We find differences in CUB between the two regions.

    * Taking this difference into account, we can increase our ability to extract biological information (predicting gene expression).
    * Thanks to our ability to distinguish between effects of mutation and selection on CUB, we are able to attribute most of the difference in CUB to mutation bias.
    * Figure 2 shows the CUB if we ignore the introgression (dotted), and for the endogenous (solid) and exogenous (dashed) respectively.
- At this point, the source of the introgression has not been identified.
    * Since we can clearly distinguish between the endogenous and exogenous CUB, can we use this information to find possible donor organisms?
    * We analyzed CUB for several yeasts and found several species with similar selection for translation efficiency, and a few with similar mutation bias, but only two with high agreement in both (gossypii and dubliensis, Figure 4).
    * We validated our findings with orthogonal information from synteny where analyzed a subset of our initial yeast set.
    * We found several closely related species with syntenious regions, but only one species that also showed agreement in CUB allowing us to exclude dubliensis since it does not show any synteny with *L. kluyveri* (Figure 5 right).
- Assuming gossypii as origin for the exogenous region, we estimated a time since introgression from our estimates of mutation bias.
    * Based on the two codon amino acids we estimated a time since introgression on the order of $10e8$
    * Assuming one to eight generations per day, we are finding an introgression age between 110k and 890k years, which overlaps with a previous estimate [1].

Results

- We compared model fits of CUB for *L. kluyveri* with a fit where we allowed CUB to vary between the endogenous and exogenous region.
  - 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 (0.59 vs 0.69) (Figure 1).
  - We also observed a decrease of the variation in estimated ϕ when assuming only one CUB environment.
- Comparison of posterior estimate between regions (ADD FIGURE).
  - The comparison estimates of mutation bias (ΔM) showed that X out of 40 parameters showed a difference in sign, meaning that different codons are favored by mutation in the two regions.
  - We find that only TTT is favored by mutation in both regions (CHECK 4 and 6 codon AA)
  - The comparison estimates of selection for translation inefficiency ($\Delta\eta$) showed that X out of 40 parameters showed a difference in sign, meaning that more of the same codons are favored by selection in both regions than in the mutation case.
- The exogenous region is assumed to be a recent introgression of unknown origin [1].
  - To determine a potential origin, we estimated the number of neutral substitutions that we expect to determine how different we can the exogenous region to be from its origin.

    * [1] argued that the introgression occurred about $55.5e6$ generations ago, and showed that it can be found in all studied populations.
    * Based on the length of the exogenous region ($1e6$), the mutation rate per nucleotide ($4e-10$) and the number of generations estimated ($55.5e6$) we expect about $22k$ neutral substitutions or about 2.2% of the introgressed region.
  - Estimates of gene trees with a fixed topology allowed us to determine that we do not observe accelerated evolution in the exogenous region when compared to the endogenous region (Figure 3).
  - these observations combined lead us to the expectation that the exogenous region should still reflect most of its original CUB environment.

- We explored CUB for several yeasts species to determine if another yeast shows similar CUB.
  - Comparison of CUB parameters yielded three species with agreement ($\rho > 0.5$) in mutation bias (ΔM) and 29 species with agreement in selection bias ($\Delta\eta$) (Figure 4).
  - Only two species, gossypii and dubliensis showed agreement in both, ΔM and $\Delta\eta$ (Figure 4).
  - musiva showed a positive correlation in in both ΔM and $\Delta\eta$ but did not satisfy our arbitrary cutoff.

- We used synteny as an independent approach as a means to validate our candidate list.
  - The check if a subset of our yeast (including our two candidates) shows synteny with the exogenous region revealed eight species (Figure 5).
    * dubliensis, a candidate based on CUB, did not show a synteny relationship with the exogenous region.

    - gossypii, the other candidate, was found to have a synteny coverage of 95% (Figure 5).
    - the other six yeasts with synteny showed agreement with only agreement in $\Delta\eta$ but not in ΔM (CHECK mutation/selection CORRELATION).

- Under the assumption that the exogenous region originated from gossypii, we estimated the time since introgression.
    - For simplicity, only the two codon amino acids were used.
    - We again assumed a mutation rate of $4e-10$.
    - Based on the difference in mutation bias ΔM between gosypii and the endogenous region we estimated a decay curve.
    - knowing the current ΔM parameters allowed us to place the xogenous region on that curve, providing us with an estimate of the time since introgression of about $4e8$ generations.
    - Assuming one to eight generations per day for *L. kluyveri* we estimate a time since introgression of about $110k$-$890k$
    - combining our estimates with the estimates of [1] ($19k$-$150k$) we date the age of the introgression to be between $110k$-$150k$.
    - Our time since introgression depends on gossypii being the origin and has not changed it's CUB since the introgression occurred.

Discussion

- based purely on selection, we would have not been able to identify the difference in CUB between the endogenous and exogenous region.
- With approaches like CAI or tAI, which are purely focused on selection we would not have been able to narrow the species down that much.

- However, in this particular case GC-content alone would have gotten us to the same result.
- GC-content would be faster, but can be expected to lead to more candidates since it is a more coarse grain approach.

Materials and Methods

- *L. kluyveri* genome preparation
  - We obtained the CDS for *L. kluyveri* from weird french site
  - We split the CDS into two partitions, an exogenous region, describing the introgression and an endogenous partition based on [3].
- *L. kluyveri* model fitting

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.

[2] JG Lawrence and H Ochman. Amelioration of bacterial genomes: Rates of change and exchange. *Journal of Molecular Miology*, 44:383–397, 1997.

[3] 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.

[4] T Tuller, Y Girshovich, Y Sella, A Kreimer, S Freilich, M Kupiec, U Gophna, and

E Ruppin. Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Research*, 39(11):4743–4755, 2011.

Figures and Tables

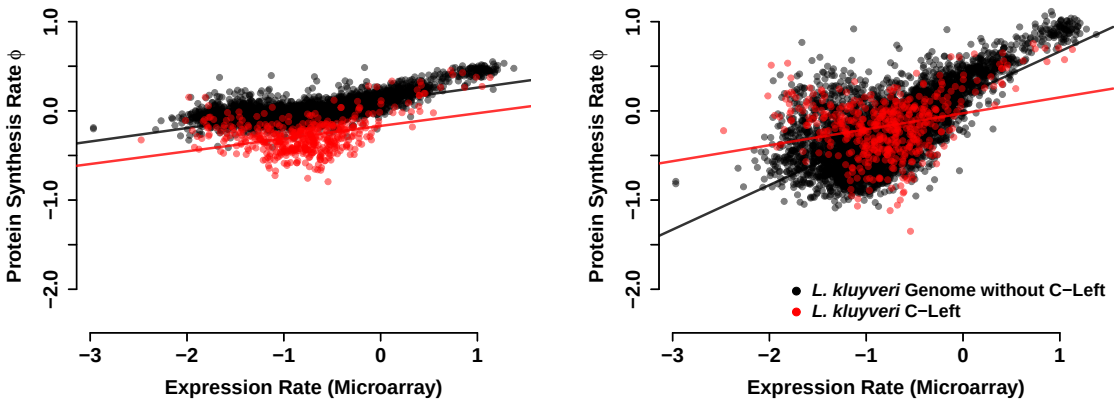

Figure 1: Person correlation of predicted protein synthesis rate ϕ with observed expression rate.)

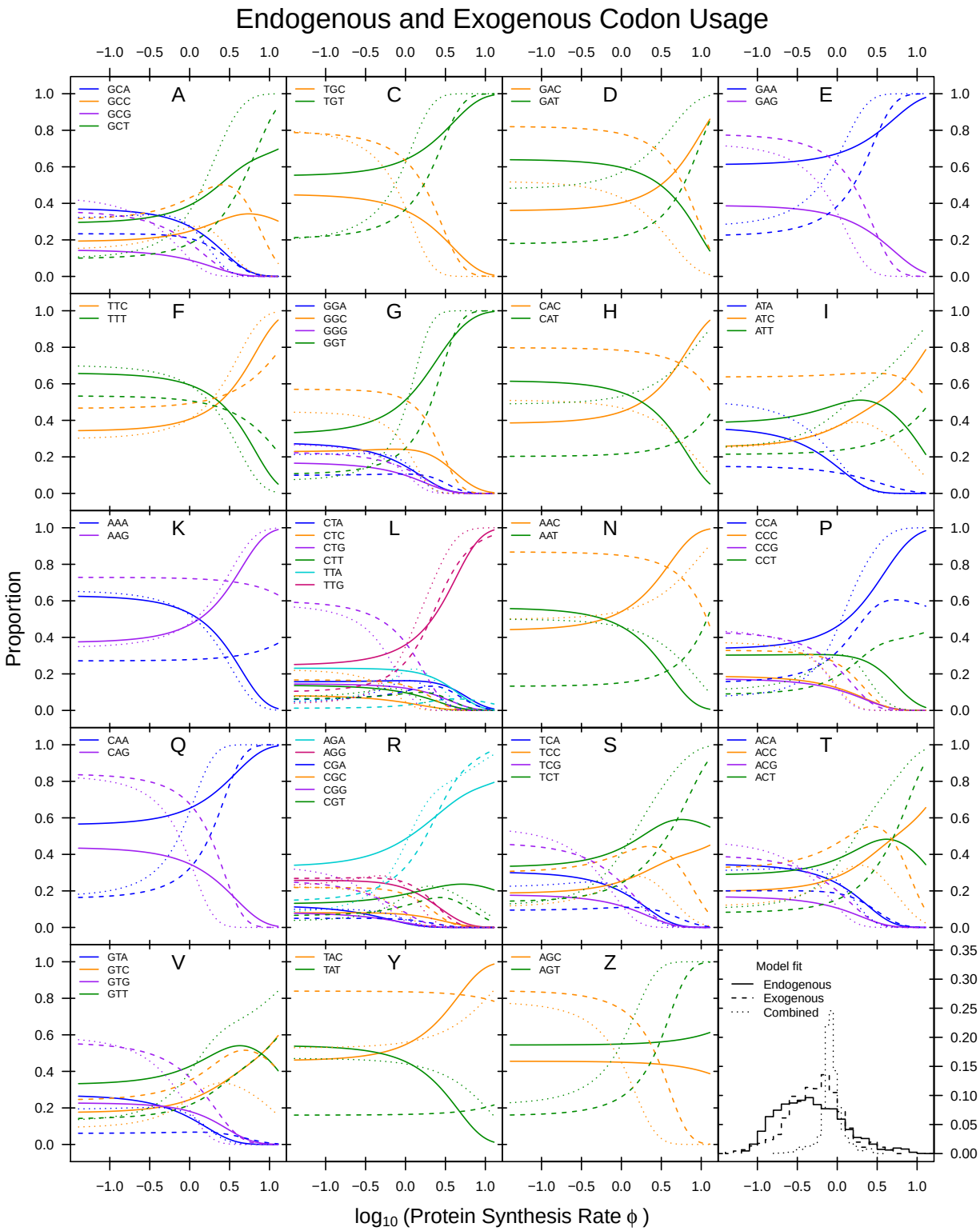

Figure 2: Codon Usage

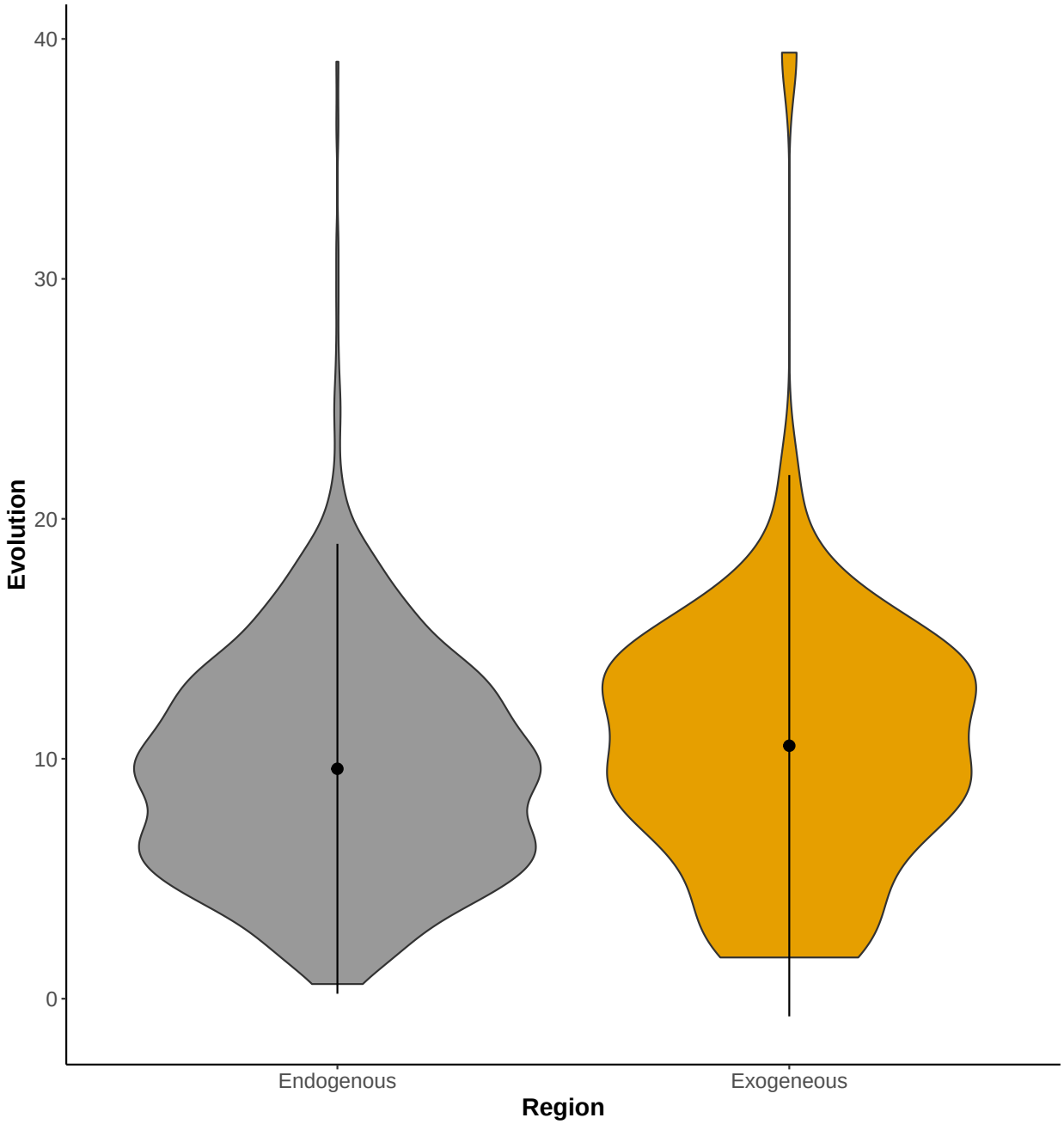

Figure 3: Overall time passed along gene tree

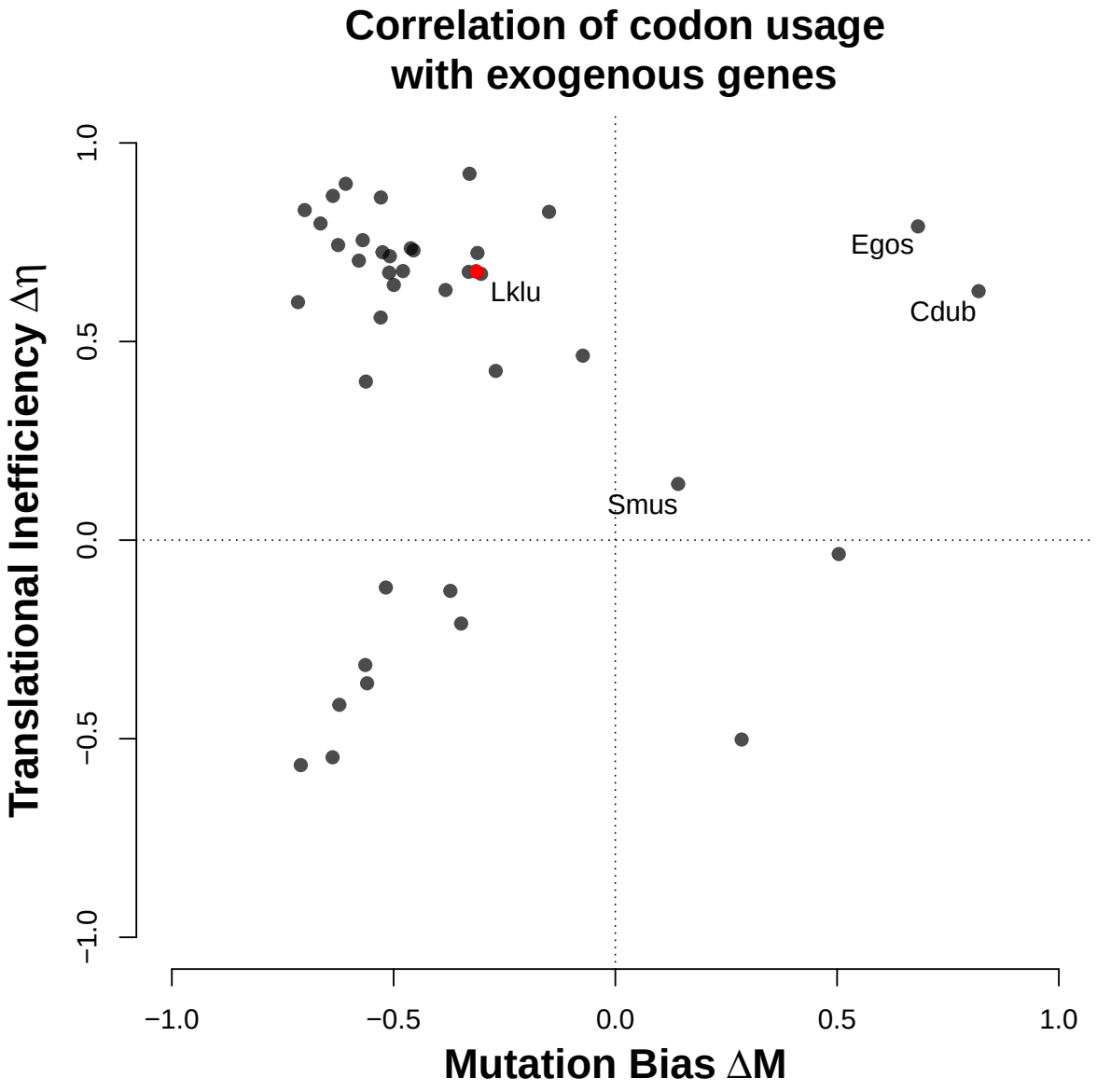

Figure 4: Codon Usage

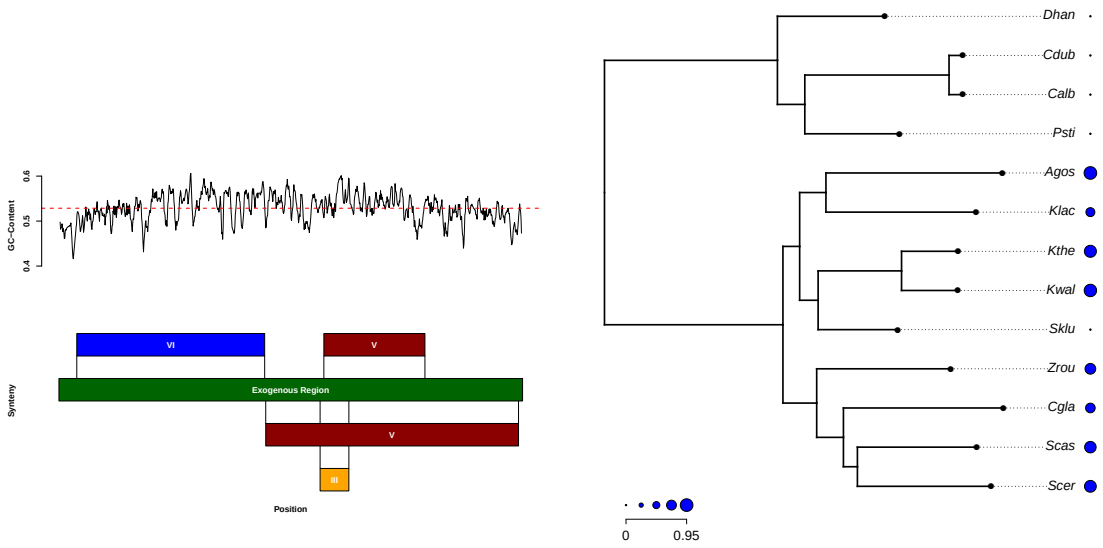

Figure 5: Synteny stuff