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 7, 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 efficacy of selection on codon usage is often assumed to be proportional to gene expression, where highly expressed genes show a greater level of adaptation than low expression genes.
    - Codon usage in low expression genes on the other hand, is dominated by mutation driven bias.
    - 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 it's history and making inferences of it's future.
- Most studies implicitly assume that codon usage in 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.

  - In contrast, similarities in the genomic environment could, under certain circumstances even promote the transfer of genomic material.
  - 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.
  - It appears that *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.
  - We utilize the ability to distinguish between effects of mutation bias and selection bias to describe two genomic environments reflected in the *L. kluyveri* genome, an endogenous and an exogenous environment.
    - We find that the 13% difference in GC content can be attributed to a shift in codon usage from A/T ending codon in the endogenous genes to C/G ending codons in the exogenous genes due to differences in mutation bias.
  - Recognizing the difference in codon usage between endogenous and exogenous genes allows us to infer the codon preference ($\Delta\eta$) for *L. kluyveri* without the exogenous genes distorting our estimates.
    - We also observe a relative improvement of 17% in our ability to predict protein synthesis rate.

- In addition to improvements to model fitting, we show the utility of the separation of mutation bias and selection bias 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.
        - Using orthogonal information on synteny with the exogenous genes, left *E. gossypii* as only potential source.
    - Estimating the time since introgression and the persistence of the signal using estimates of mutation bias.
    - Estimating the cost of the introgression using estimates of selection bias.

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 is clearly favored ($\Delta AIC \sim 90,000$).
    - We find that the disagreement in selection bias causes us to predict wrong codon preferences for seven amino acids if endogenous and exogenous genes are not treated separately.
- The comparison of parameter estimates reveals large differences in mutation bias and smaller differences in selection bias between endogenous and exogenous genes.
    - We find little agreement in mutation bias between endogenous and exogenous genes with only two amino acids (A,F) showing complete concordance.

  - 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.
  - However, we find the same codon preference in endogenous and exogenous genes for nine amino acids.
  - We also observe a stronger bias in selection towards C/G ending codons in the exogenous genes.
  - Furthermore, recognizing the differences in mutation driven bias and selection driven bias between endogenous and exogenous genes improves our relative ability to predict protein synthesis rate by 17% ($\rho = 0.59$ vs. $\rho = 0.69$).

- Inference of mutation bias ΔM and selection bias $\Delta\eta$ of 38 other yeasts revealed 33 species with similarities in selection bias and five species with similarities in mutation bias.
  - Only four species that showed similar mutation bias (*E. gossypii*, *C. dubliniensis*, and *Sphaerulina musiva*, *Yarrowia lipolytica*) showed a positive selection bias relationship.
  - Only *E. gossypii* and *C. dubliniensis* had a strong positive relationship in both mutation bias and selection bias.
  - We validated our candidate sources using orthogonal information on synteny.
  - We find that synteny with the exogenous genes is limited to the Saccharomycetacease group, eliminating *C. dubliniensis* as potential source, leaving us with *E. gossypii*.
  - Using our estimates of mutation bias we estimated the age of the introgression to be about 6.22×10^8 generations.
  - We predict that decay of the signature of the sources genomic environment to one percent of the *L. kluyveri* genomic environment will take about 5.66×10^9

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 no change in the *E. gossypii* genomic environment and constant amino acid usage.
  - We find that the exogenous genes were a large fitness burden at the time of introgression and still represent a large reduction in fitness relative to the replaced endogenous genes.

Discussion

- We partitioned *L. kluyveri* into endogenous and exogenous genes using information about a previously identified introgression event.
- After we inferred parameters describing codon usage using ROC SEMPPR, we find that the 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%.
- The comparison of ΔM and $\Delta\eta$ estimates from the exogenous genes to 38 other yeast lineages provided us with 33 yeast lineages showing a positive relationship in selection

bias and five lineages showing a positive relationship in mutation bias.

  - We expect differences in mutation bias to decay more slowly than differences in selection bias due to these differences being mostly neutral.
    - Therefore yeasts with similar estimates of $\Delta\eta$ are all likely sources.
    - However, the longer persistence of mutation bias should allow us to identify a source more reliably without the worry about signatures of selection to dissipate.
  - Synteny of the exogenous region was consistent with eight Saccharomycetaceae lineages.
    - Most of these eight species 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*.
  - 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) are insensitive to differences in mutation bias, focusing only on selection.
  - We acknowledge that our assumption about a constant genomic environment in the *E. gossypii* genome is unlikely.

- Assuming that the amino acid sequence has not changed over time, we infer the fitness burden at the time of introgression of the exogenous genes on *L. kluyveri* compared to the original endogenous genes from our estimates of selection bias.
  - 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.
  - However, 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, which may be indicative that we only observe these events very rarely.
  - 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 a mechanistic approach like 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.