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 14, 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 its genome. The *L. kluyveri* genome provides an opportunity to study the adaptation of an introgressed region to a novel genomic environment.

The CUB of the endogenous *L. kluyveri* genome and the exogenous region were analyzed, and the effects of mutation bias and selection for translation efficiency on CUB were separated. We found significant differences in 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.

In order to identify potential sources of the exogenous region we compared codon preferences across several yeast lineages. Our comparison identified two candidates, *Candida dubliniensis* and *Eremothecium gossypii*, as potential source lineages. We excluded *C. dubliniensis* using orthogonal information on synteny.

Introduction

Codon usage bias (CUB) - the non-uniform usage of synonymous codons - results from mutation, selection, and drift; creating a genomic environment in which all genes evolve. The efficacy of mutation and selection differs between genes. Genes with a low efficacy to selection will show a synonymous codon preference dominated by mutation, while selection will dominate synonymous codon preference in genes where selection efficacy is high. This variation in strength allows us to separate effects of mutation and selection on individual genes [1, 2].

It is often implicitly assumed that all genes in a genome have evolved within the same genomic environment [3, 4, 5]. This assumption, however, is easily violated by population bottlenecks, selective sweeps, horizontal gene transfer and introgression, or hybridization. The impact these events have on CUB is mostly unstudied. Selection on codon usage is often associated with factors contributing to the efficient translation of mRNA such as tRNA availability [4, 6, 7] and ribosome pausing times. Genes which evolved in a genomic environment were these factors differ can therefore be assumed to have a negative impact on the overall fitness of the organism [8]. To our knowledge, codon usage of genes that have evolved in different genomic environments has only been studied in bacteria where it is common for genes to be horizontally transferred between lineages. These transferred genes are not found to impact estimates of codon usage, likely because genes with CUB similar to their own are more likely to be taken up by organisms [9]. However, transfer of large genomic material between organisms that have evolved in differing genomic environments can lead to the misclassification of codon preferences.

In this study, we analyze the synonymous codon preferences in the genome of *L. kluyveri*, the earliest diverging lineage of the known *Lachancea* clade and diverged from the Saccharomyces lineage prior to the whole-genome duplication about 100 Mya ago [10]. *L. kluyveri* historically experienced a large introgression of about 1Mb of the left arm of chromosome C, clearly marked by elevated GC content [11]. The introgressed region contains about 10%

of all protein coding genes. Given the large number of introgressed genes, we would expect large fitness consequences if the genomic environment of the exogenous genes differs from *L. kluyveri*. We estimate parameters for mutation bias (ΔM) and selection against translation inefficiency ($\Delta\eta$) from the protein coding sequences using codon counts. We find that synonymous codon usage differs between the introgressed exogenous and the endogenous genes. We observe a greater difference in mutation bias than in selection against translation inefficiency between the exogenous and the endogenous genes. The exogenous genes exhibit a strong bias towards C/G ending codons consistent with the elevated GC content in that region. Taking into account the difference in codon usage improves our ability to predict protein synthesis rate and avoids misclassification of synonymous codon preferences.

A comparison of mutation bias and selection against translation inefficiency of the exogenous genes to 39 other yeast species within the Saccharomycetaceae and Debarymomycetaceae clades identified the *E. gossypii* and the *C. dubliniensis* lineages in the Saccharomycetaceae clade and the Debarymomycetaceae clade as most likely sources of the exogenous genes among the yeast species examined. Evaluation of our results with orthorgonal synteny information revealed that *C. dubliniensis* does not show any synteny with the exogenous region. Therefore, we propose *E. gossypii* as potential source lineage. With *E. gossypii* as potential origin, we were able to estimate the age of the introgression based on differences in mutation bias and find our estimates to be in agreement with previous work [12].

Materials and Methods

Results

Discussion

Partitioning of the *L. kluyveri* genome based on a previously identified introgression revealed two distinct signatures of genomic environments, reflected in the coding sequences by differences in synonymous codon usage. Our results complement and contrast therefore previous work on codon usage where it is common to assume that a genome displays only a single genomic environment.

The separation of effects of mutation and selection on the two signatures of genomic environments found in *L. kluyveri* revealed great differences in mutation bias between endogenous and exogenous genes. We find endogenous genes to exhibit mutation bias towards A and T ending codons, the introgressed exogenous genes in contrast, exhibit mutation bias towards C and G ending codons. Aspartate (Asp, D) is displaying a strong mutation bias ($\sim 78\%$) towards GAC in the exogenous genes in the absence of selection while in the endogenous genes we observe a strong mutation bias towards GAT ($\sim 65\%$). Similarly large changes in mutation bias are seen for the amino acids histidine (His, H) and Lysine (Lys, K). This shift in mutation bias towards C and G ending codons in the exogenous region is in line with the, by 13% increased, GC content in that region. Increased GC content content is associated with increased DNA stability [13, 14, 15, 16] and often found in thermophiles due to the increased stability of the base stacking of Cytosine and Guanine in comparison of the stacking of Adenine and Thymine [17]. While the three hydrogen bonds of a Cystosine/Guanine pair provides additional stability, this effect is independent of temperature [17]. Therefore, the high GC content found in the exogenous genes could hint towards a thermophilic source lineage.

We find a higher correlation in our estimates of parameters describing selection against translation inefficiency ($\Delta\eta$) between endogenous and exogenous genes than in correlation between our estimates of mutation bias (ΔM). The higher correlation and agreement in the optimal codon between the regions could be a result of faster decay of the selection environment relative to the mutation environment. Alternatively, we can not rule out that the donor lineage and *L. kluyveri* share a similar selective environment, resulting in a similar set of optimal codons. Nevertheless, we find that the optimal codon differs for 10 amino acids. We find preference for C and G ending codons for 17 amino acids in the exogenous genes; Phenylalanine (Phe, F) and Isoleucine (Ile, I) being the exception. Again adding to the elevated GC content in the exogenous region. Endogenous genes in contrast, show preference for A and T ending codons in 11 cases. Without the partitioning of the *L. kluyveri* genome, we would have inferred the optimal codon for seven amino acids wrong, i.e. in the case of Arginine (Arg, R) we would classify CGG as the optimal codon for *L. kluyveri* instead of CGA. In all cases were we would have misidentified the optimal codon, we find that the codon inferred represents the optimal codon for the exogenous genes.

A key feature of employing ROC SEMPPR is the prediction of the evolutionary average protein synthesis rate (ϕ) [2]. Recognizing that the parameter estimates for mutation bias (ΔM) and selection against translation inefficiency ($\Delta\eta$) differ between exogenous and endogenous genes allowed for improved prediction of ϕ. Using expression data as proxy for protein synthesis rate we find a Pearson correlation of $\rho = 0.59$ when ΔM and $\Delta\eta$ are shared between endogenous and exogenous genes, and $\rho = 0.69$ when parameters are allowed to be estimated independents. An improvement of 12 % in explained variation. Interestingly, the distribution of ϕ estimates is very narrow when parameters are shared between the endogenous and exogenous genes.

check type

Comparing estimates of mutation bias and selection against translation efficiency of 39 yeast lineages to the exogenous parameter estimates yielded several lineages with positive correlations. Most of the studied yeasts show a positive relationship in estimates of $\Delta\eta$ suggesting that variation in synonymous codon preference is small in the studied set of yeasts. The comparison of estimates of ΔM on the other hand, only shows four yeasts with positive relationship of our mutation bias estimates. Only two lineages show a high correlation in both ΔM and $\Delta \eta$. The Saccharomycetaceae *E. gossypii* showed the highest agreement in ΔM and $\Delta \eta$ with $\rho = 0.75$ and $\rho = 0.85$ respectively. We find a similar but weaker correlation with the Debarymomycetaceae *C. dubliniensis* with a Pearson correlation of $\rho = 0.68$ for ΔM and $\rho = 0.63$ for $\Delta \eta$. Combined we find that our estimates on mutation bias provide more information than our estimates on selection against translation inefficiency. In contrast to this expectation, it appears that most of the studied yeasts do experience a similar selective genomic environment. However, only a small sample of yeasts was analyzed.

Estimation of parameters describing codon usage allowed us to identify two likely candidates as source of the exogenous genes, *E. gossypii* and *C. dubliniensis*. Using orthogonal information on gene synteny of eight yeast species closely related to the two identified species and *L. kluyveri* we find that *C. dubliniensis* does not show any synteny relationship with the exogenous region. We find several other yeast species within the Saccharomycetaceae clade to show a synteny relationship with the exogenous region however, none of them display a similar genomic environment. The synteny relationship with the exogenous region is limited to species within the Saccharomycetaceae clade and does not exptend into the sister clade Debarymomycetaceae where we identified *C. dubliniensis* based on our estimates of ΔM and $\Delta \eta$. Taken together we propose *E. gossypii* as potential source of the introgressed exogenous region covering the left arm of chromosome C of the *L. kluyveri* genome.

feels short, think about what is missing

Our ability to identify regions that have evolved in different genomic environments depends on the time since the transfer of that region [18]. We estimated the time since the introgression of the exogenous region to be $3.32e8$ generations, using our estimates of mutation bias for all two codon amino acids. Mutation bias is well suited since, as previously mentioned, it provides more information about the exogenous genes than our estimates of selection against translation inefficiency. Furthermore, our estimates of mutation bias are free from influences of varying selection pressure or effective population size. However, other factors that are inconsistent with the model formulation of ROC SEMPPR could have been absorbed by the ΔM terms. Assuming between one and eight generations per day, we estimate the introgression to have occurred between $114,000$ and $910,000$ years ago, a time that is consistent with previous work [12]. Despite assuming the same mutation rate for each amino acid we observe large variation in our estimate of the time since the introgression event. This large variation can be caused by many factors, such as the uncertainty in our estimates of ΔM, noise in the data, or amino acid usage as we would expect rarely used amino acids to have a slower decaying signature of the genomic environment of the source lineage, and in turn overestimate the time since introgression. However, instead of finding amino acids with very large times, we find two amino acids, Lysine (Lys, K) and Asparagine (Asn, N), with a negative estimate of the time since introgression. We assume that *E. gossypii* did not evolved since the time of the introgression and still exhibits the same genomic environment. This assumption is likely violated as indicated by the two amino acids predicting a negative time since introgression. However, we can not rule out other factors such as shifts in amino acid composition [11] or non-linear dynamics of shift in codon usage . Further analyzing the time line of the introgression, we predict that the signature of the source lineage's genomic environment we identified in the codon usage of the exogenous genes will decay to one percent of the *L. kluyveri* genomic environment within $5.37e9$ generations. This time contextualizes our estimate of the time since the introgression occurred, showing how relatively recent this introgression occurred.

rethink word choice as both are two codon AA

In conclusion, this study shows that signatures of more than one genomic environment can be present in a genome. In the case of *L. kluyveri* this is due to an introgression event but other internal factors could lead to similar differences like strand bias. It was previously proposed that the difference in GC content found in the *L. kluyveri* genome was due to replication timing [11, 19]. Never mind the reason for the presents of multiple genomic environments; as this study shows, recognizing the presents of multiple genomic environments will help to prevent misintepretation of results and the misidentification of codon usage. Furthermore, this study highlights that it is important to take into account all factors driving codon usage; mutation, selection, and drift. Without information on mutation bias, we would have been unable to identify *E. gossypii* as a potential source lineage of the introgression. However, the information on mutation bias is often disregarded by other approaches used to analyze codon usage [20]. Lastly, while we used patterns of codon usage to determine a potential source lineage for the exogenous genes, our work highlights how ROC SEMPPR can be used for more sophisticated hypothesis testing in the future.

References

[1] EW Wallace, EM Airoldi, and DA Drummond. Estimating selection on synonymous codon usage from noisy experimental data. *Molecular Biology and Evolution*, 30:1438–1453, 2013.

[2] MA Gilchrist, WC Chen, P Shah, CL Landerer, and R Zaretzki. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. *Genome Biology and Evolution*, 7:1559–1579, 2015.

[3] R Grantham, C Gautier, M Gouy, M Jacobzone, and R Mercier. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research*, 9:43–74, 1981.

[4] T Ikemura. Codon usage and trna content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, 2:13–34, 1985.

[5] PM Sharp, E Cowe, DG Higgins, DC Shields, KH Wolfe, and F Wright. Codon usage patterns in escherichia coli, bacillus subtilis, saccharomyces cerevisiae, schizosaccharomyces pombe, drosophila melanogaster and homo sapiens; a review of the considerable within species diversity. *Nucleic Acids Research*, 16:8207–8211, 1988.

[6] H Dong, L Nilsson, and CG Kurland. Co-variation of trna abundance and codon usage in escherichia coli at different growth rates. *Journal of Molecular Miology*, 260:649–663, 1996.

[7] M dos Reis, R Savva, and L Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research*, 32(17):5036–5044, 2004.

[8] J Peter and J Schacherer. Population genomics of yeasta:towards a comprehensiveview across a broad evolutionary scale. *Yeast*, 33:73–81, 2016.

[9] 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.

[10] KH Wolfe and DC Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–7013, 1997.

[11] 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.

[12] 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.

[13] J Marmur and P Doty. Determination of the base composition of dioxyribonucleic acid from its thermal melting temperature. *J. Mol. Biol.*, 5:109–118, 1962.

[14] RM Wartell and AS Benight. Thermal denaturation of dna molecules. *Phys. Rep.*, 126:67–107, 1985.

[15] MC Williams and I Rouzina. Force spectroscopy of single dna and rna molecules. *Curr. Opin. Struct. Biol.*, 12:330–336, 2002.

[16] J SantaLucia and D Hicks. The thermodynamics of dna structural motifs. *Biomol. Struct.*, 33:415–440, 2004.

[17] P Yakovchuk, E Protozanova, and MD Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic Acids Research*, 34(2):564–574, 2006.

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

[19] Nicolas Agier, Orso Maria Romano, Fabrice Touzain, Marco Cosentino Lagomarsino, and Gilles Fischer. The spatiotemporal program of replication in the genome of lachancea kluyveri. *Genome Biology and Evolution*, 5(2):370–388, 2013.

[20] PM Sharp. The codon adaptatoin index - a meassure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15:1281–1295, 1987.