

RESEARCH

1
2
3
4
5
6
7
8

Unlocking a signal of introgression from codons in *Lachancea kluyveri* using a mutation-selection model

9 Cedric Landerer^{1,2,3*}, Brian C O'Meara^{1,2}, Russell Zaretzki^{2,4} and Michael A Gilchrist^{1,2}

10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	

Correspondence:
edric.landerer@gmail.com
Max-Planck Institute of
Molecular Cell Biology and
Genetics, Pfotenhauerstr. 108,
1307, Dresden, Germany
Full list of author information is
available at the end of the article
Correspondence

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33

Abstract

Background: For decades, codon usage has been used as a measure of adaptation for translational efficiency and translation accuracy of a gene's coding sequence. These patterns of codon usage reflect both the selective and mutational environment in which the coding sequences evolved. Over this same period, gene transfer between lineages has become widely recognized as an important biological phenomenon. Nevertheless, most studies of codon usage implicitly assume that all genes within a genome evolved under the same selective and mutational environment, an assumption violated when introgression occurs. In order to better understand the effects of introgression on codon usage patterns and vice versa, we examine the patterns of codon usage in *Lachancea kluyveri*, a yeast which has experienced a large introgression. We quantify the effects of mutation bias and selection for translation efficiency on the codon usage pattern of the endogenous and introgressed exogenous genes using a Bayesian mixture model, ROC SEMPPR, which is built on mechanistic assumptions about protein synthesis and grounded in population genetics.

Results: We find substantial differences in codon usage between the endogenous and exogenous genes, and show that these differences can be largely attributed to differences in mutation bias favoring A/T ending codons in the endogenous genes while favoring C/G ending codons in the exogenous genes. Recognizing the two different signatures of mutation bias and selection improves our ability to predict protein synthesis rate by 42% and allowed us to accurately assess the decaying signal of endogenous codon mutation and preferences. In addition, using our estimates of mutation bias and selection, we identify *Eremothecium gossypii* as the closest relative to the exogenous genes, providing an alternative hypothesis about the origin of the exogenous genes, estimate that the introgression occurred $\sim 6 \times 10^8$ generation ago, and estimate its historic and current selection against mismatched codon usage.

Conclusions: Together, our work illustrates the advantage of mechanistic, population genetic models like ROC SEMPPR and the quantitative estimates they provide when analyzing sequence data.

Keywords: codon usage; population genetics; introgression; mutation; selection

1 **Background**

2 Synonymous codon usage patterns varies within a genome and between taxa, re-
3 flecting differences in mutation bias, selection, and genetic drift. The signature of
4 mutation bias is largely determined by the organism's internal or cellular environ-
5 ment, such as their DNA repair genes or UV exposure. While this mutation bias
6 is an omnipresent evolutionary force, its impact can be obscured or amplified by
7 selection. The signature of selection on codon usage is largely determined by an or-
8 ganism's cellular environment alone, such as, but not limited to, its tRNA species,
9 their copy number, and their post-transcriptional modifications. In general, the
10 strength of selection on codon usage is assumed to increase with its expression level
11 [1–3], specifically its protein synthesis rate [4]. Thus as protein synthesis increases,
12 codon usage shifts from a process dominated by mutation to a process dominated
13 by selection. The overall efficacy of mutation and selection on codon usage is a
14 function of the organism's effective population size N_e . ROC SEMPPR allows us
15 to disentangle the evolutionary forces responsible for the patterns of codon usage
16 bias [5–7] (CUB) encoded in an species' genome, by explicitly modeling the com-
17 bined evolutionary forces of mutation, selection, and drift [4, 8–10]. In turn, these
18 evolutionary parameters should provide biologically meaningful information about
19 the lineage's historical cellular and external environment.

20 Most studies implicitly assume that the CUB of a genome is shaped by a single
21 cellular and external environment. However, this assumption is clearly violated to
22 increasing degrees via horizontally gene transfer, large scale introgressions, and hy-
23 brid specie formation. In these scenarios, one would expect to see the signature of
24 multiple cellular environments in a genome's CUB [11, 12]. Indeed, differences in
25 CUB between linages have been proposed to have a major effect on their rates of
26 gene transfer with rates declining with differences in their CUB. On a more practical
27 level, if differences in codon usage of transferred genes are not taken into account
28 for, they may distort the interpretation of codon usage patterns. Such distortion
29 could lead to the wrong inference of codon preference for an amino acid [8, 10], un-
30 derestimate the variation in protein synthesis rate, or distort estimates of mutation
31 bias when analyzing a genome.

32 To illustrate these ideas, we analyze the CUB of the genome of the yeast *Lachancea*
33 *kluyveri* using ROC SEMPPR, a population genetics based model of synonymous

1 codon usage evolution that accounts for and, in turn, can estimate the contribution
2 of mutation bias ΔM , selection bias. The mathematics of ROC SEMPPR are de-
3 rived on a mechanistic description of ribosome movement along an mRNA, although
4 the approximation of other biological mechanisms could also be consistent with the
5 model. Broadly speaking, ROC SEMPPR allows us to quantify the cellular environ-
6 ment in which genes have evolved by separately estimating the effects of mutation
7 bias and selection bias on codon usageDE between synonymous codons and pro-
8 tein synthesis rate ϕ to the patterns of codon usage observed within a set of genes.
9 Briefly, the set of ΔM for an amino acid quantifies the relative differences in muta-
10 tional stability or bias between the synonymous codons of the amino acid S . In the
11 absence of selection bias (or equivalently when gene expression $\phi = 0$), the equilib-
12 rium frequency of synonymous codon i is simply $\exp[-\Delta M_i] / \left(\sum_{j \in S} \exp[-\Delta M_j] \right)$.
13 Because the time units of protein production rate have no intrinsic time scale, we
14 define the average protein production rate for a set of genes to be one, i.e. $= 1$
15 by definition [10]. In order to facilitate comparisions between gene sets, we express
16 both, ΔM and $\Delta \eta$, as deviation from the mean of each synonymous codon family
17 (see Materials and Methods for details). Nevertheless, the difference $\Delta \eta$ describes
18 the difference in fitness between two synonymous codons relative to drift for a gene
19 whose protein production rate ϕ is equal to the the average rate of protein produc-
20 tion $\bar{\phi}$ across the set of genes. In other words, for a gene whose protein is expressed
21 at the average rate, for any two given synonymous codons i and j , $\Delta \eta_i - \Delta \eta_j = N_e s$.

22
23 The Lachancea clade diverged from the Saccharomyces clade, prior to its whole
24 genome duplication ~ 100 Mya ago [13, 14]. Since that time, *L. kluyveri*, which is
25 sister species to all other *Lachancea spp.*, has experienced a large introgression of
26 exogenous genes (1 Mb, 457 genes) which is found in all of its populations [15, 16],
27 but in no other known Lachancea species [17]. The introgression replaced the left
28 arm of the C chromosome and displays a 13% higher GC content than the en-
29 doogenous *L. kluyveri* genome [15, 16]. Previous studies suggest that the source of
30 the introgression is probably a currently unknown or potentially extinct Lachancea
31 lineage based on gene concatenation or synteny relationships [15–18]. These char-
32 acteristics make *L. kluyveri* an ideal model to study the effects of an introgressed
33 cellular environment and the resulting mismatch in codon usage.

1 While previous studies ROC SEMPPR's resulting predictions of protein synthesis
2 rates have been shown to be on par with laboratory measurements [8, 10]. In contrast
3 to often used heuristic approaches to study codon usage [5, 6, 19], ROC SEMPPR
4 explicitly incorporates and distinguishes between mutation and selection effects on
5 codon usage and properly weights its estimates by amino acid usage [20]. We use
6 ROC SEMPPR to separately describe the two cellular environments reflected in
7 the *L. kluyveri* genome; the signature of the endogenous environment reflected in
8 the larger set of non-introgressed genes and the decaying signature of the ancestral,
9 exogenous environment in the smaller set of introgressed genes. Our results indicate
10 that the current difference in GC content between endogenous and exogenous genes
11 is mostly due to the differences in mutation bias ΔM of their respective cellular
12 environments. Taking the different signatures of ΔM and selection bias $\Delta \eta$ of the
13 endogenous and exogenous sets of genes substantially improves our ability to predict
14 present day protein synthesis rates ϕ . These endogenous and exogenous gene set
15 specific estimates of ΔM and $\Delta \eta$, in turn, allow us to address more refined biological
16 questions. For example, we find support for an alternative origin of the exogenous
17 genes and identify *E. gossypii* as the nearest sampled relative of the source of the
18 introgressed genes out of the 332 budding yeast lineages with sequenced genomes
19 [21]. While this inference is in contrast to previous work [15–18], we find additional
20 phylogenetic support for via gene tree reconstruction and gene synteny. We also
21 estimate the age of the introgression to be on the order of 0.2 - 1.7 Mya, estimate
22 the selection against these genes, both at the time of introgression and now, and
23 predict a detectable signature of CUB to persist in the introgressed genes for another
24 0.3 - 2.8 Mya, highlighting the sensitivity of our approach.
25

26 Results

27 The Signatures of two Cellular Environments within *L. kluyveri*'s Genome

28 We used our software package AnaCoDa [22] to compare model fits of ROC
29 SEMPPR to the entire *L. kluyveri* genome and its genome partitioned into two
30 sets of 4,864 endogenous and 497 exogenous genes. These two set where initially
31 identified based on their striking difference in GC content [15], with very little over-
32 lap in GC content between the two sets (Figure S1a). ROC SEMPPR is a statistical
33 model that relates the effects of mutation bias ΔM , selection bias $\Delta \eta$ between syn-

1	Table 1 Model selection of the two competing hypothesis. Combined: mutation bias and selection	1
2	bias for synonymous codons is shared between endogenous and exogenous genes. Separated:	2
3	mutation bias and selection bias for synonymous codons is allowed to vary between endogenous	3
4	and exogenous genes. Reported are the log-likelihood, $\log(\mathcal{L})$, the number of parameters	4
5	estimated n, the log-marginal likelihood $\log(\mathcal{L}_M)$, Bayes Factor K, and the p-value of the	5
6	likelihood ratio test.	6
7		7

Hypothesis	$\log(\mathcal{L})$	n	$\log(\mathcal{L}_M)$	$\log(K)$	p
Combined	-2,650,047	5,483	-2,657,582	—	—
Separated	-2,612,397	5,402	-2,615,288	42,294	0

8 synonymous codons and protein synthesis rate ϕ , to explain the observed codon usage
 9 patterns. Thus, the probability of observing a synonymous codon is proportional
 10 to $p \propto \exp(-\Delta M - \Delta \eta \phi)$ [10]. Briefly, ΔM describes the mutation bias between
 11 two synonymous codons at stationarity under a time reversible mutation model.
 12 Because ROC SEMPPR only considers the stationary probabilities, only variation
 13 in mutation bias, not absolute mutation rates can be detected. $\Delta \eta$ describes the
 14 fitness difference between two synonymous codons relative to drift [10]. Since $\Delta \eta$ is
 15 scaled by protein synthesis rate ϕ , this term is dominant in highly expressed genes
 16 and tends towards 0 in low expression genes, allowing us to separate the effect of
 17 mutation bias and selection bias on codon usage. We express both, ΔM and $\Delta \eta$,
 18 as deviation from the mean of each synonymous codon family which prevents that
 19 the choice of the reference codon affects our results (see Materials and Methods for
 20 details).

21 Bayes factor strongly support the hypothesis that the *L. kluyveri* genome consists
 22 of genes with two different and distinct patterns of codon usage bias rather than a
 23 single ($K = \exp(42,294)$; Table 1). We find additional support for this hypothesis
 24 when we compare our predictions of protein synthesis rate to empirically observed
 25 mRNA expression values as a proxy for protein synthesis. Specifically, we improve
 26 the variance explained by our predicted protein synthesis rates by $\sim 42\%$, from
 27 $R^2 = 0.33$ ($p \approx 0$) to 0.46 ($p \approx 0$) (Figure 1). While the implicit consideration of GC
 28 content in this analysis certainly plays a roll, it does not explain the improvement
 29 in R^2 (Figure S1b).

30
 31 Comparing Differences in the Endogenous and Exogenous Codon Usage
 32 Because ROC SEMPPR defines $\bar{\phi} = 1$, it makes the interpretation of $\Delta \eta$ as selection
 33 on codon usage of the average gene with $\phi = 1$ straightforward and gives us the

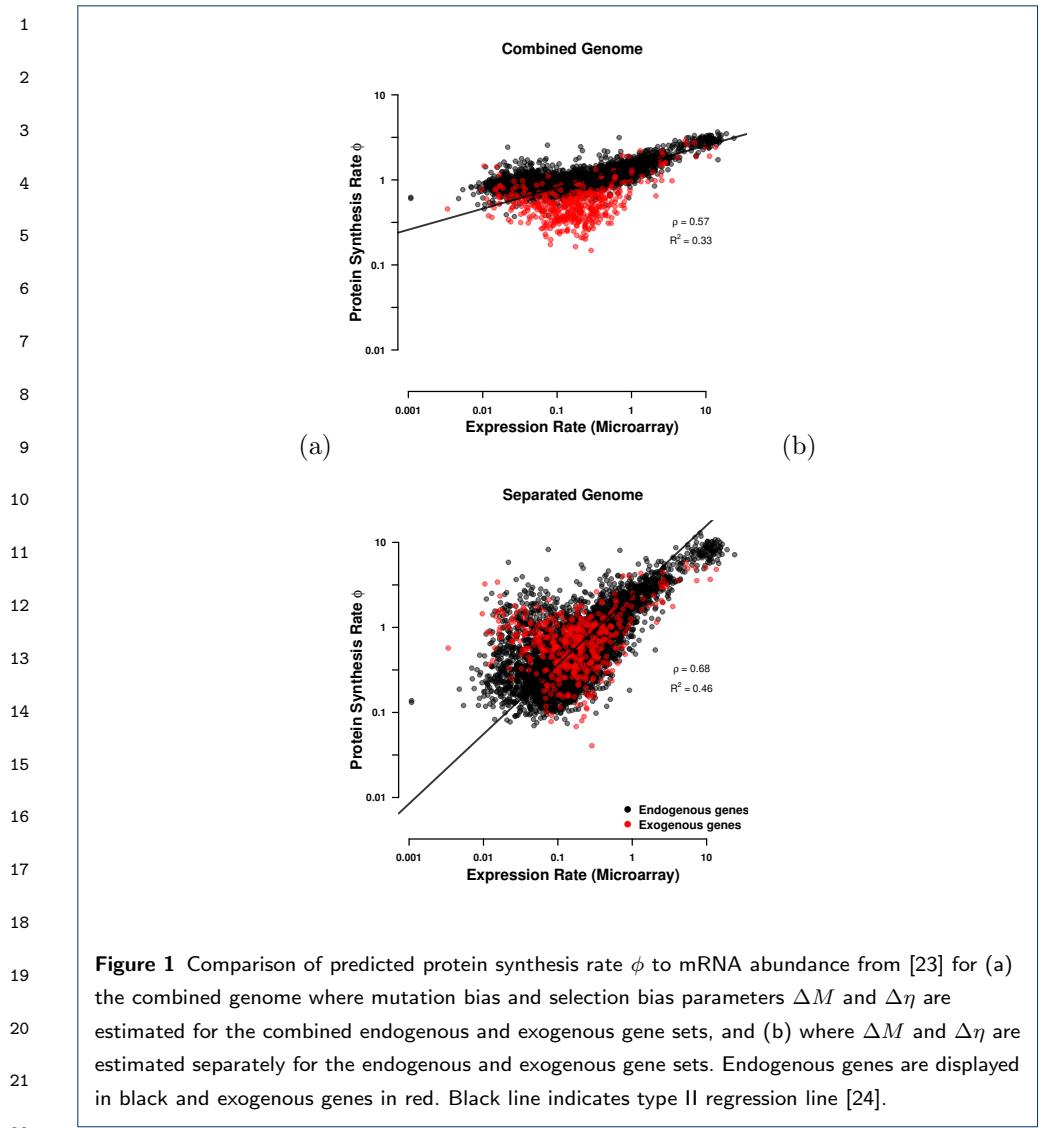
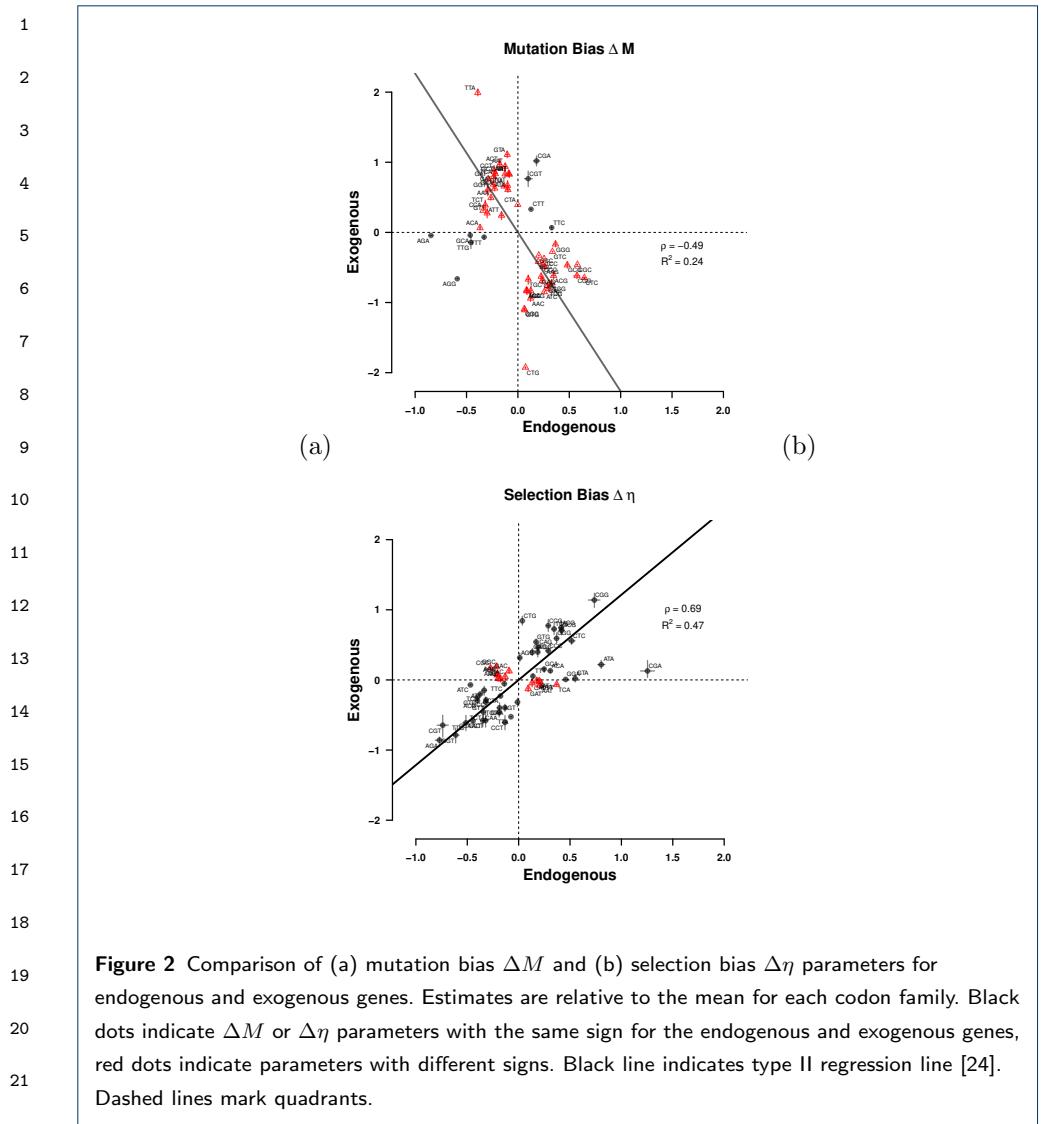


Figure 1 Comparison of predicted protein synthesis rate ϕ to mRNA abundance from [23] for (a) the combined genome where mutation bias and selection bias parameters ΔM and $\Delta \eta$ are estimated for the combined endogenous and exogenous gene sets, and (b) where ΔM and $\Delta \eta$ are estimated separately for the endogenous and exogenous gene sets. Endogenous genes are displayed in black and exogenous genes in red. Black line indicates type II regression line [24].

ability to compare the efficacy of selection sN_e across genomes. While it may be expected for the endogenous and exogenous genes to differ in their codon usage pattern due to the large difference in GC content it is not clear how much of this difference is due to differences in the mutation bias ΔM or selection bias $\Delta \eta$ between the gene sets. To better understand the differences in the endogenous and exogenous cellular environments, we compared our parameter estimates of ΔM and $\Delta \eta$ for the two sets of genes. Our estimates of ΔM for the endogenous and exogenous genes were negatively correlated ($\rho = -0.49$, $p = 3.56 \times 10^{-5}$), indicating weak similarity with only $\sim 5\%$ of the codons share the same sign between the two mutation environments (Figure 2a). Overall, mutation bias favors codons ending in



the purines A and T over codons ending in the pyrimidines G or C, respectively, as indicated by where the endogenous model fit curves intercept the left axis in Figure 3. One exception is Lys, L, where mutation appears to favor the codon TTG over TTA; however, this difference is very small and not statistically meaningful. The exogenous genes display behavior, favoring G over A and C over G and doing so more strongly, as indicated by shift in order and the greater distance between the exogenous model fit curves where they intercept the left axis in Figure 3). Please check and, if changes are correct, please revise $\Delta \eta$ results to mirror structure of ΔM section.

1 We find that the signature of selection bias $\Delta\eta$ also differs substantially between
2 the endogenous and exogenous gene sets. In terms of their magnitude relative to
3 mutation bias ΔM , $\Delta\eta$ for the endogenous gene set is substantially greater than
4 for the exogenous gene set as indicated by the fact that the protein production rate
5 ϕ where the effect of selection fails to override the effect of The difference in codon
6 usage between endogenous and exogenous genes is striking as some amino acids have
7 opposite codon preferences. For example, 'opposite codon preferences' is confusing,
8 is it that $\Delta\eta$ values have opposite signs between gene sets? As a result, our estimates
9 of the optimal codon differ in nine cases between endogenous and exogenous genes
10 (Figure 3, Table S2). For example, the usage of the Asparagine (Asn, N) codon AAC
11 is increased in highly expressed endogenous genes but the same codon is depleted in
12 highly expressed exogenous genes. For Aspartic acid (Asp, D), the combined genome
13 shows the same codon preference in highly expressed genes as the exogenous gene
14 set. Generally, fits to the complete *L. kluyveri* genome reveal that the relatively
15 small exogenous gene set ($\sim 10\%$ of genes) has a disproportionate effect on the
16 model fit (Figure S2, S3).

17 Of the nine cases in which the endogenous and exogenous genes show differences
18 in the selectively most favored codon five cases (Asp, D; His, H; Lys, K; Asn, N;
19 and Pro, P) the endogenous genes favor the codon with the most abundant tRNA.
20 For the remaining four cases (Ile, I; Ser, S; Thr, T; and Val, V), there are no
21 tRNA genes for the wobble free cognate codon encoded in the *L. kluyveri* genome.
22 However, the codon preference of these four amino acids in the exogenous genes
23 matches the most abundant tRNA encoded in the *L. kluyveri* genome. In contrast
24 to ΔM , our estimates of selection bias $\Delta\eta$ for the endogenous and exogenous genes
25 are positively correlated ($\rho = 0.69$, $p = 9.76 \times 10^{-10}$) and show the same sign in
26 $\sim 53\%$ of the cases (Figure 2).

27 This striking difference in codon usage was noted previously. For example, using
28 RSCU [5], GAA (coding for Glu, E) was identified as the optimal synonymous codon
29 in the whole genome and GAG as the optimal codon in the exogenous genes [15].
30 Our results, however, indicate that GAA is the optimal codon in both, endogenous
31 and exogenous genes, and that the high RSCU in the exogenous genes of GAG is
32 driven by mutation bias (Table S1 and S2). Similar effects are observed for other
33 amino acids.

1 The effect of the small exogenous gene set on the fit to the complete *L. kluyveri*
2 genome is smaller for our estimates of selection bias $\Delta\eta$ than ΔM , but still large.
3 We find that the complete *L. kluyveri* genome is estimated to share the selectively
4 preferred codon with the exogenous genes in ~ 60% of codon families that show dis-
5 similarity between endogenous and exogenous genes. We also find that the complete
6 *L. kluyveri* genome fit shares mutationally preferred codons with the exogenous
7 genes in ~ 78% of the 19 codon families showing a difference in mutational codon
8 preference between the endogenous and exogenous genes. In two cases, Isoleucine
9 (Ile, I) and Arginine (Arg, R), the strong dissimilarity in mutation preference results
10 in an estimated codon preference in the complete *L. kluyveri* genome that differs
11 from both the endogenous, and the exogenous genes. These results clearly show that
12 it is important to recognize the difference in endogenous and exogenous genes and
13 treat these genes as separate sets to avoid the inference of incorrect synonymous
14 codon preferences and better predict protein synthesis.

17 Can Codon Usage Help Determine the Source of the Exogenous Genes

18 Since the origin of the exogenous genes is currently unknown, we explored if the
19 information on codon usage extracted from the exogenous genes can be used to
20 identify a potential source lineage. We combined our estimates of mutation bias
21 ΔM and selection bias $\Delta\eta$ with synteny information and searched for potential
22 source lineages of the introgressed exogenous region. We used ΔM to identify can-
23 didate lineages as the endogenous and exogenous genes show greater dissimilarity
24 in mutation bias than in selection bias. We examined 332 budding yeasts [21] and,
25 identified the ten lineages with the highest correlation to the exogenous ΔM pa-
26 rameters as potential source lineages (Figure 4, Table 2). Two of the ten candidate
27 lineages utilize the alternative yeast nuclear code (NCBI codon table 12). In this
28 case, the codon CTG codes for Serine instead of Leucine. We therefore excluded the
29 Leucine codon family from our comparison of codon families; however, there was no
30 need to exclude Serine as CTG is not a one step neighbor of the remaining Serine
31 codons. A mutation between CTG and the remaining Serine codons would require
32 two mutations with one of them being non-synonymous, which would violate the
33 weak mutation assumption of ROC SEMPPR.

Table 2 Budding yeast lineages showing similarity in codon usage with the exogenous genes. $\rho_{\Delta M}$ and $\rho_{\Delta \eta}$ represent the Pearson correlation coefficient for exogenous ΔM and $\Delta \eta$ with the indicated species', respectively. GC content is the average GC content of the whole genome. Synteny is the percentage of the exogenous genes found in the listed lineage. Only one lineage (*E. gossypii*) shows a similar GC content > 50%.

Species	$\rho_{\Delta M}$	$\rho_{\Delta \eta}$	GC content	Synteny %	Distance [Mya]
<i>Eremothecium gossypii</i>	0.89	0.70	51.7	75	211.0847
<i>Danielozyma ontarioensis</i>	0.75	0.92	46.6	3	470.1043
<i>Metschnikowia shivogae</i>	0.86	0.87	49.8	0	470.1043
<i>Babjeviella inositovora</i>	0.83	0.78	48.1	0	470.1044
<i>Ogataea zsoltii</i>	0.75	0.85	47.7	0	470.1042
<i>Metschnikowia hawaiiensis</i>	0.80	0.86	44.4	0	470.1042
<i>Candida succiphila</i>	0.85	0.83	40.9	0	470.1042
<i>Middlehovenomyces tepae</i>	0.80	0.62	40.8	0	651.9618
<i>Candida albicans*</i>	0.84	0.75	33.7	0	470.1043
<i>Candida dubliniensis*</i>	0.78	0.75	33.1	0	470.1043

* Lineages use the alternative yeast nuclear code

The endogenous *L. kluyveri* genome exhibits codon usage very similar to most (77 %) yeast lineages examined, indicating that most of the examined yeasts share a similar codon usage (Figure S4). Only ~ 17% of all examined yeast show a positive correlation in both, ΔM and $\Delta \eta$ with the exogenous genes, whereas the vast majority of lineages (~ 83%) show a negative correlation for ΔM , only 21 % show a negative correlation for $\Delta \eta$.

Comparing synteny between the exogenous genes, which are restricted to the left arm of chromosome C, and the candidate yeast species we find that *E. gossypii* is the only species that displays high synteny (Table 2). Furthermore, the synteny relationship between the exogenous region and other yeasts appears to be limited to Saccharomycetaceae clade. Given these results, we conclude that, of the 332 examined yeast lineages the *E. gossypii* lineage is the most likely source of the introgressed exogenous genes. Previous studies which studied the exogenous genes and chromosome recombination in the Lachancea clade concluded that the exogenous region originated from within the Lachancea clade, from an unknown or potentially extinct lineage [15–17]. While it is not possible for us to dispute this hypothesis, our results provide a novel hypothesis about the origin of the exogenous genes.

To further test the plausibility of *E. gossypii* as potential source lineage, we identified 127 genes in our dataset [21] with homologous genes in *E. gossypii* and other Lachancea and used IQTree [25] to infer the phylogenetic relationship of the exogenous genes. Our results show that at least ~ 45% of exogenous genes (57/127) are

more closely related to *E. gossypii* than to other Lachancea S5. Interestingly, our results also indicate that codon usage does not necessarily correlate with phylogenetic distance (Table 2).

Estimating Introgression Age

If we assume that the exogenous genes originated from the *E. gossypii* lineage, we can estimate the age of the introgression based on our estimates of mutation bias ΔM . We modeled the change in codon frequency over time as exponential decay, and estimated the age of the introgression assuming that *E. gossypii* still represents the mutation bias of its ancestral source lineage at the time of the introgression and a constant mutation rate. We infer the age of the introgression to be on the order of $6.2 \pm 1.2 \times 10^8$ generations. Assuming *L. kluyveri* experiences between one and eight generations per day, we estimate the introgression to have occurred between 212,000 to 1,700,000 years ago. Our estimate places the time of the introgression earlier than the previous estimate of 19,000 - 150,000 years by [16].

Using our model of exponential decay model, we also estimated the persistence of the signal of the exogenous cellular environment. We predict that the ΔM signal of the source cellular environment will have decayed to be within one percent of the *L. kluyveri* environment in $\sim 5.4 \pm 0.2 \times 10^9$ generations, or between 1,800,000 and 15,000,000 years. Together, these results indicate that the mutation signature of the exogenous genes will persist for a very long time.

Estimating Selection against Codon Mismatch of the Exogenous Genes

We define the selection against inefficient codon usage as the difference between the fitness on the log scale of an expected, replaced endogenous gene and the exogenous gene, $s \propto \phi\Delta\eta$ due to the mismatch in codon usage parameters (See Methods for details). As the introgression occurred before the diversification of *L. kluyveri* and has fixed throughout all populations [16], we can not observe the original endogenous sequences that have been replaced by the introgression. Overall, we predict that a small number of low expression genes ($\phi < 1$) were weakly exapted at the time of the introgression (Figure 5a). Thus, they appear to provide a small fitness advantage due to the accordance of exogenous mutation bias with endogenous selection bias (compare Figure S2 and S3). High expression genes ($\phi > 1$) are predicted to have faced the largest selection against their mismatched codon usage in the novel cellular

1 environment. In order to account for differences in the efficacy of selection on codon
2 usage either due to the cost of pausing, differences in the effective population size,
3 or the decline in fitness with every ATP wasted between the donor lineage and *L.*
4 *kluyveri* we added a linear scaling factor κ to scale our estimates of $\Delta\eta$ between the
5 donor lineage and *L. kluyveri* and searched for the value that minimized the cost of
6 the introgression, thus giving us the best case scenario (See Methods for details).

7 Using our estimates of ΔM and $\Delta\eta$ from the endogenous genes and assuming the
8 current exogenous amino acid composition of genes is representative of the replaced
9 endogenous genes, we estimate the strength of selection against the exogenous genes
10 at the time of introgression (Figure 5a) and currently (Figure 5b). Estimates of
11 selection bias for the exogenous genes show that, while well correlated with the
12 endogenous genes, only nine amino acids share the same selectively preferred codon.
13 Exogenous genes are, therefore, expected to represent a significant reduction in
14 fitness for *L. kluyveri* due to mismatch in codon usage. Since $\Delta\eta$ is proportional
15 to the difference in fitness between the wild type and a mutant, we can use our
16 estimates of $\Delta\eta$ to approximate the selection against the exogenous genes Δs [10,
17 26]. We estimate that the selection against all exogenous genes due to mismatched
18 codon usage to have been $\Delta s \approx -0.0008$ at the time of the introgression and
19 ≈ -0.0003 today. This reduction in Δs is primarily due to adaptive changes to the
20 codon usage of the most highly expressed, introgressed genes (Figures 5a & S8).
21 Based on the selection against the codon mismatch at the time of the introgression
22 and assuming an effective population size N_e on the order of 10^7 [27], we estimate
23 a fixation probability of $(1 - \exp[-\Delta s]) / (1 - \exp[-2\Delta s N_e]) \approx 10^{-6952}$ [26] for the
24 exogenous genes. Clearly, the possibility of fixation under this simple scenario is
25 effectively zero. In order for the exogenous genes to have reached fixation one or
26 more exogenous loci must have provided a selective advantage not considered in
27 this study (See Discussion).

29 Discussion

30 In order to study the evolutionary effects of the large scale introgression of the left
31 arm of chromosome C, we used ROC SEMPPR, a mechanistic model of ribosome
32 movement along an mRNA. The usage of a mechanistic model rooted in popula-
33 tion genetics allows us generate more nuanced quantitative parameter estimates

1 and separate the effects of mutation and selection on the evolution of codon usage.
2 This allowed us to calculate the selection against the introgression, and provides *E.*
3 *gossypii* as a potential source lineage of the introgression which was previously not
4 considered. Our parameter estimates indicate that the *L. kluyveri* genome contains
5 distinct signatures of mutation and selection bias from both an endogenous and ex-
6 ogenous cellular environment. By fitting ROC SEMPPR separately to *L. kluyveri*'s
7 endogenous and exogenous sets of genes we generate a quantitative description of
8 their signatures of mutation bias and natural selection for efficient protein transla-
9 tion.

10 In contrast to other methods such as RSCU, CAI, or tAI, ROC SEMPPR does
11 not rely on external information such as gene expression or tRNA gene copy number
12 [5, 19]. Instead, ROC SEMPPR allows for the estimation of protein synthesis rate ϕ
13 and separates the effects of mutation and selection on codon usage. In addition, [20]
14 showed that approaches like CAI are sensitive to amino acid composition, another
15 property that distinguishes the endogenous and exogenous genes [15].

16 Previous work by [15] showed an increased bias towards GC rich codons in the
17 exogenous genes but our results provide more nuanced insights by separating the
18 effects of mutation bias and selection. We are able to show that the difference in GC
19 content between endogenous and exogenous genes is mostly due to differences in
20 mutation bias as 95% of exogenous codon families show a strong mutation bias to-
21 wards GC ending codons (Table S1). However, the exogenous genes show a selective
22 preference for AT ending codons for 90% of codon families (Table S2). Acknowl-
23 edging the increased mutation bias towards GC ending codons and the difference in
24 strength of selection between endogenous and exogenous genes by separating them
25 also improves our estimates of protein synthesis rate ϕ by 42% relative to the full
26 genome estimate ($R^2 = 0.46, p = 0$ vs. $0.32, p = 0$, respectively).

27 Previous studies showed that nucleotide composition can be strongly affected by
28 biased gene conversion, which, in turn would affect codon usage. Biased gene conver-
29 sion is thought to act similar to directional selection, typically favoring the fixation
30 of G/C alleles [28, 29]. Further, [30, Harrison & Charlesworth] suggested that bi-
31 ased gene conversion affects codon usage in *S. cerevisiae*. ROC SEMPPR, however,
32 does not explicitly account for biased gene conversion. If biased gene conversion is
33 independent of gene expression, as in the case of DNA repair, it will be absorbed

1 in our estimates of ΔM . If instead biased gene conversion forms hotspots, and
2 thus becomes gene specific, it will affect our estimates of protein synthesis ϕ . This
3 might be the case at recombination hotspots. Recombination, however, is very low
4 in the introgressed region (discussed below) [15, 18]. The low recombination rate
5 also indicates that the GC content had to be high before the introgression occurred.

6 The estimates of mutation and selection bias parameters, ΔM and $\Delta \eta$, are ob-
7 tained under an equilibrium assumption. Given that the introgression is still adapt-
8 ing to its new environment, this assumption is clearly violated. However, the adap-
9 tation of the exogenous genes progresses very slowly as a quasi-static process as
10 shown in this work as well as [16]. Therefore, the genome can be assumed to main-
11 tain an internal equilibrium at any given time. We see empirical evidence for this
12 behavior in our ability to predict gene expression and to correctly identify the low
13 expression genes (Figure 1b).

14 Despite the violation of the equilibrium assumption, the mutation and selection
15 bias parameters ΔM and $\Delta \eta$ of the introgressed exogenous genes contain informa-
16 tion, albeit decaying, about its previous cellular environment. We selected the top
17 ten lineages with the highest similarity in ΔM to see if our parameters estimates
18 would allow us to identify a potential source lineage. The synteny relationship of
19 these lineages with the exogenous genes was calculated as a point of comparison as
20 it provides orthogonal information to our parameter estimates. Synteny with the
21 exogenous genes is limited to the Saccharomycetaceae clade, excluding all of the
22 potential source lineages identified using codon usage but *E. gossypii* (Table 2). In-
23 terestingly, this also showed that similarity in codon usage does not correlate with
24 phylogenetic distance.

25 Previous work indicated that the donor lineage of the exogenous genes has to be
26 a, potentially unknown, Lachancea lineage [15–18]. These previous results, however,
27 are based on species rather than gene trees, ignoring the differential adaptation rate
28 to their novel cellular environment between genes or do not consider lineages outside
29 of the Lachancea clade. Considering the similarity in selection bias (Figure 2b) and
30 our calculation of selection on the exogenous genes (Figure 5b), both of which
31 are free of any assumption about the origin of the exogenous genes, a species tree
32 estimated from the exogenous genes will be biased towards the Lachancea clade.
33 Estimating individual gene trees rather than relying on a species tree provided

1 further evidence that the exogenous genes could originate from a lineage that does
2 not belong to the Lachancea clade. As we highlighted in this study, relatively small
3 sets of genes with a signal of a foreign cellular environment can significantly bias
4 the outcome of a study. The same holds true for phylogenetic inferences [31], and as
5 we showed the signal of the original endogenous cellular environment that shaped
6 CUB is at different stages of decay in high and low expression genes (Figure S8).
7 In summary, our work does not dispute an unknown Lachancea as possible origin,
8 but provides an alternative hypothesis based on the codon usage of the exogenous
9 genes, phylogenetic analysis, and synteny.

10 In terms of understanding the spread of the introgression, we calculated the ex-
11 pected selective cost of codon mismatch between the *L. kluyveri* and *E. gossypii*
12 lineages. Under our working hypothesis, the majority of the introgressed would have
13 imposed a selective cost due to codon mismatch. Nevertheless, $\sim 30\%$ of low expres-
14 sion exogenous genes ($\phi < 1$) appeared to be exapted at the time of the introgres-
15 sion. This exaptation is due to the mutation bias in the endogenous genes matching
16 the selection bias in the exogenous genes for GC ending codons. Our estimate of
17 the selective cost of codon mismatch on the order of -0.0008 . While this selective
18 cost may not seem very large, assuming *L. kluyveri* had a large N_e , the fixation
19 probability of the introgression is the astronomically small value of $\approx 10^{-6952} \approx 0$.
20 While this estimate heavily depends on the working hypothesis that the exogenous
21 genes originated from the *E. gossypii* lineage, we can also calculate the hypothetical
22 fixation probability if the current exogenous genes would introgress into *L. kluyveri*.
23 Our estimate of the current selective cost of the mismatch of codon usage is on the
24 order of -0.0003 . The fixation probability of the current exogenous genes would
25 still be astronomically small $\approx 10^{-2609} \approx 0$. These results are in accordance with
26 previous work, highlighting the necessity of codon usage compatibility between en-
27 dogenous and transferred exogenous genes [32, 33]. Thus, the basic scenario of an
28 introgression between two yeast species with large N_e and where the introgression
29 solely imposes a selective cost due to codon mismatch is clearly too simplistic.

30 One or more loci with a combined selective advantage on the order of 0.0008
31 or greater would have made the introgression change from disadvantageous to ef-
32 fectively neutral or advantageous. While this scenario seems plausible, it raises
33 the question as to why recombination events did not limit the introgression to

only the adaptive loci. A potential answer is the low recombination rate between the endogenous and exogenous regions [15, 18]. Estimates of the recombination rate as measured by crossovers (COs) for *L. kluyveri* are almost four times lower than for *S. cerevisiae* and about half that of *Schizosaccharomyces pombe* (≈ 1.6 COs/Mb/meiosis, ≈ 6 COs/Mb/meiosis, ≈ 3 COs/Mb/meiosis) with no observed crossovers in the introgressed region [18], and no observed transposable elements [15]. This is presumably due to the dissimilarity in GC content and/or a lower than average sequence homology between the exogenous region and the one it replaced. A population bottleneck reducing the N_e of the *L. kluyveri* lineage around the time of the introgression could also help explain the spread of the introgression. Compatible with these explanation is the possibility of several advantageous loci distributed across the exogenous region drove a rapid selective sweep and/or the population through a bottleneck speciation process.

Assuming *E. gossypii* as potential source lineage of the exogenous region, we illustrated how information on codon usage can be used to infer the time since the introgression occurred using our estimates of mutation bias ΔM . The ΔM estimates are well suited for this task as they are free of the influence of selection and unbiased by N_e and other scaling terms, which is in contrast to our estimates of $\Delta\eta$ [10]. Our estimated age of the introgression of $6.2 \pm 1.2 \times 10^8$ generations is ~ 10 times longer than a previous minimum estimate by [16] of 5.6×10^7 generations, which was based on the effective population recombination rate and the population mutation parameter [34]. Furthermore, these estimates assume that the current *E. gossypii* and *L. kluyveri* cellular environment reflect their ancestral states at the time of the introgression. Thus, if the ancestral mutation environments were more similar (dissimilar) at the time of the introgression then our result is an overestimate (underestimate).

Further, the presented work provides a template to explore the evolution of codon usage. This applies not only to species who experienced an introgression but is more generally applicable to any species.

Conclusion

Overall, our results show the usefulness of the separation of mutation bias and selection bias and the importance of recognizing the presence of multiple cellular

environments in the study of codon usage. We also illustrate how a mechanistic model like ROC SEMPPR and the quantitative estimates it provides can be used for more sophisticated hypothesis testing in the future. In contrast to other approaches used to study codon usage like CAI [5] or tAI [19], ROC SEMPPR incorporates the effects of mutation bias and amino acid composition explicitly [20]. We highlight potential issues when estimating codon preferences, as estimates can be biased by the signature of a second, historical cellular environment. In addition, we show how quantitative estimates of mutation bias and selection relative to drift can be obtained from codon data and used to infer the fitness cost of an introgression as well as its history and potential future.

11

11

Materials and Methods

Separating Endogenous and Exogenous Genes

A GC-rich region was identified by [15] in the *L. kluyveri* genome extending from position 1 to 989,693 of chromosome C. This region was later identified as an introgression by [16]. We obtained the *L. kluyveri* genome from SGD Project <http://www.yeastgenome.org/download-data/> (on 09-27-2014) and the annotation for *L. kluyveri* NRRL Y-12651 (assembly ASM14922v1) from NCBI (on 12-09-2014). We assigned 457 genes located on chromosome C with a location within the ~ 1 Mb window to the exogenous gene set. All other 4864 genes of the *L. kluyveri* genome were assigned to the exogenous genes.

22

22

Model Fitting with ROC SEMPPR

ROC SEMPPR was fitted to each genome using AnaCoDa (0.1.1) [22] and R (3.4.1) [35]. ROC SEMPPR was run from 10 different starting values for at least 250,000 iterations and thinned to every 50th iteration. After manual inspection to verify that the MCMC had converged, parameter posterior means, log posterior probability and log likelihood were estimated from the last 500 samples (last 10% of samples).

29

29

Model selection

The marginal likelihood of the combined and separated model fits was calculated using a generalized harmonic mean estimator [36]. A variance scaling of 1.1 was used to scale the important density of the estimator. Using the estimated marginal likelihoods, we calculated the Bayes factor to assess model performance. Increases

33

33

1 in the variance scaling increase the estimated Bayes factor, therefore we report a
2 conservative Bayes factor bases on a small variance scaling S9.
3

4 **Comparing Codon Specific Parameter Estimates and Selecting Candidate lineages** 4

5 As the choice of reference codon can reorganize codon families coding for an amino 5
6 acid relative to each other, all parameter estimates were interpreted relative to the 6
7 mean for each codon family. 7

8

$$9 \Delta M_i = \Delta M_{i,1} - \overline{\Delta M_i} \quad (1) \quad 9$$

10

$$11 \Delta \eta_i = \Delta \eta_{i,1} - \overline{\Delta \eta_i} \quad (2) \quad 11$$

12 Comparison of codon specific parameters (ΔM and $\Delta \eta = 2N_e q(\eta_i - \eta_j)$) was per- 12
13 formed using the function lmodel2 in the R package lmodel2 (1.7.3) [37] and R 13
14 version 3.4.1 [35]. The parameter $\Delta \eta$ can be interpreted as the difference in fitness 14
15 between codon i and j for the average gene with $\phi = 1$ scaled by the effective pop- 15
16 ulation size N_e , and the selective cost of an ATP q [4, 10]. Type II regression was 16
17 performed with re-centered parameter estimates, accounting for noise in dependent 17
18 and independent variable [24]. 18

19 Due to the greater dissimilarity of the ΔM estimates between the endogenous and 19
20 exogenous genes, and the slower decay rate of mutation bias, we decided to focus 20
21 on our estimates of mutation bias to identify potential source lineages. The top ten 21
22 lineages with the highest similarity in ΔM to the exogenous genes were selected as 22
23 potential candidates (Figure 2). 23

24

25 **Phylogenetic Analysis** 25

26 Using the dataset from [21], we first identified 129 alignments for exogenous genes 26
27 that further contained homologous genes for *E. gossypii*, and at least one other 27
28 Lachancea species. We excluded all species from the alignments that do not belong 28
29 to the Saccharomycetaceae clade. IQTree [25] was used to identify the best fit- 29
30 ting model for each gene and to estimate the individual gene trees. Each gene tree 30
31 was rooted using either *Saccharomyces cerevisiae*, *Saccharomyces uvarum*, *Saccha-* 31
romyces eubayanus as outgroup. We calculated the most recent common ancestor 32
33 (MRCA) of *L. kluyveri* and *E. gossypii* as well as the MRCA of *L. kluyveri* and the 33

1 remaining Lachancea. The distance between the MRCA and the root was used to
 2 asses which pairs (*L. kluyveri* and *E. gossypii*, or *L. kluyveri* and other Lachancea)
 3 have a more recent common ancestor.

5 Synteny Comparison

6 We obtained complete genome sequences for all 10 candidate lineages (Table 2)
 7 from NCBI (on: 02-05-2017). Genomes were aligned and checked for synteny using
 8 SyMAP (4.2) with default settings [38, 39]. We assess synteny as percentage coverage
 9 of the exogenous gene region.

10 Estimating Age of Introgression

11 We modeled the change in codon frequency over time using an exponential model
 12 for all two codon amino acids. While our approach is equivalent to [40], we want
 13 to explicitly state the relationship between the change in codon frequency c_1 as a
 14 function of mutation bias ΔM as

$$16 \quad \frac{dc_1}{dt} = -\mu_{1,2}c_1 - \mu_{2,1}(1 - c_1) \quad (3) \quad 17$$

18 where $\mu_{i,j}$ is the rate at which codon i mutates to codon j and c_1 is the fre-
 19 quency of the reference codon. Initial codon frequencies $c_1(0)$ for each codon
 20 family were taken from our mutation parameter estimates for *E. gossypii* where
 21 $c_1(0) = \exp[\Delta M_{\text{gos}}]/(1 + \exp[\Delta M_{\text{gos}}])$. Our estimates of ΔM_{endo} can be used to
 22 calculate the steady state of equation 3 were $\frac{dc_1}{dt} = 0$ to obtain the equality

$$24 \quad \frac{\mu_{2,1}}{\mu_{1,2} + \mu_{2,1}} = \frac{1}{1 + \exp[\Delta M_{\text{endo}}]} \quad (4) \quad 25$$

26 Solving for $\mu_{1,2}$ gives us $\mu_{1,2} = \Delta M_{\text{endo}} \exp[\mu_{2,1}]$ which allows us to rewrite and
 27 solve equation 3 as

$$28 \quad c_1(t) = \frac{1 + \exp[-X](K - 1)}{1 + \Delta M_{\text{endo}}} \quad (5) \quad 29$$

30 where $X = (1 + \Delta M_{\text{endo}})\mu_{2,1}t$ and $K = c_1(0)(1 + \Delta M_{\text{endo}})$.

31 Equation 5 was solved with a mutation rate $\mu_{2,1}$ of 3.8×10^{-10} per nucleotide per
 32 generation [41]. Current codon frequencies for each codon family where taken from
 33 our estimates of ΔM from the exogenous genes. Mathematica (11.3) [42] was used

1 to calculate the time t_{intro} it takes for the initial codon frequencies $c_1(0)$ for each
 2 codon family to equal the current exogenous codon frequencies. The same equation
 3 was used to determine the time t_{decay} at which the signal of the exogenous cellular
 4 environment has decayed to within 1% of the endogenous environment.
 5

6 Estimating Selection against Codon Mismatch

7 In order to estimate the selection against codon mismatch, we had to make three
 8 key assumptions. First, we assumed that the current exogenous amino acid sequence
 9 of a gene is representative of its ancestral state and the replaced endogenous gene
 10 it replaced. Second, we assume that the currently observed cellular environment of
 11 *E. gossypii* reflects the cellular environment that the exogenous genes experienced
 12 before transfer to *L. kluyveri*. Lastly, we assume that the difference in the efficacy
 13 of selection between the cellular environments due to differences in either effective
 14 population size N_e or the selective cost of an ATP q of the source lineage and *L.*
 15 *kluyveri* can be expressed as a scaling constant and that protein synthesis rate ϕ
 16 has not changed between the replaced endogenous and the introgressed exogenous
 17 genes. Using estimates for $N_e = 1.36 \times 10^7$ [27] for *Saccharomyces paradoxus* we
 18 scale our estimates of $\Delta\eta$ which explicitly contains the effective population size N_e
 19 [10] and define $\Delta\eta' = \frac{\Delta\eta}{N_e}$.

20 All of our genome parameter estimations are scaled by lineage specific effects such
 21 as N_e , the average, absolute gene expression level, and/or the proportionate fitness
 22 value of an ATP. In order to account for these genome specific differences in scaling,
 23 we scale the difference in the efficacy of selection on codon usage between the donor
 24 lineage and *L. kluyveri* using a linear scaling factor κ . As $\Delta\eta$ is defined as $\Delta\eta =$
 25 $2N_e q(\eta_i - \eta_j)$, we cannot distinguish if κ is a scaling on protein synthesis rate ϕ ,
 26 effective population size N_e , or the selective cost of an ATP q [4, 10]. We calculated
 27 the selection against each genes codon mismatch assuming additive fitness effects
 28 as

$$29 s_g = \sum_{i=1}^{L_g} -\kappa \phi_g \Delta\eta'_i \quad (6) \\ 30$$

31 where s_g is the overall strength of selection for translational efficiency on gene, g
 32 in the exogenous gene set, κ is a constant, scaling the efficacy of selection between
 33 the endogenous and exogenous cellular environments, L_g is length of the protein in

codons, ϕ_g is the estimated protein synthesis rate of the gene in the endogenous environment, and $\Delta\eta'_i$ is the $\Delta\eta'$ for the codon at position i . As stated previously, our $\Delta\eta$ are relative to the mean of the codon family. We find that the selection against the introgressed genes is minimized at $\kappa \sim 5$ (Figure S7b). Thus, we expect a five fold difference in the efficacy of selection between *L. kluyveri* and *E. gossypii*, due to differences in either protein synthesis rate ϕ , effective population size N_e , and/or the selective cost of an ATP q . Therefore, we set $\kappa = 1$ if we calculate the s_g for the endogenous and the current exogenous genes, and $\kappa = 5$ for s_g for selection calculations at the time of introgression.

However, since we are unable to observe codon sequences of the replaced endogenous genes and for the exogenous genes at the time of introgression, instead of summing over the sequence, we calculate the expected codon count $E[n_{g,i}]$ for codon i in gene g simply as the probability of observing codon i multiplied by the number of times the corresponding amino acids is observed in gene g , yielding:

$$\begin{aligned} E[n_{g,i}] &= P(c_i | \Delta M, \Delta\eta, \phi) \times m_{a_i} \\ &= \frac{\exp[-\Delta M_i - \Delta\eta_i \phi_g]}{\sum_j^C \exp[-\Delta M_j - \Delta\eta_j \phi_g]} \times m_{a_i} \end{aligned}$$

where m_{a_i} is the number of occurrences of amino acid a that codon i codes for. Thus replacing the summation over the sequence length L_g in equ. (6) by a summation over the codon set C and calculating s_g as

$$s_g = \sum_{i=1}^C -\kappa \phi_g \Delta\eta'_i E[n_{g,i}] \quad (7)$$

We report the selection due to mismatched codon usage of the introgression as $\Delta s_g = s_{\text{intro},g} - s_{\text{endo},g}$ where $s_{\text{intro},g}$ is the selection against an introgressed gene g either at the time of the introgression or presently.

Acknowledgments

The authors would like to thank Alexander Cope for helpful criticisms and suggestions for this work.

Availability of data and materials

Parameter estimates generated during this study are available from the corresponding author. All remaining data generated during this study are included in this published article as figures, tables.

Authors' contributions

CL and MAG initiated the study. CL collected and analyzed the data and wrote the manuscript. MAG and BCO edited the manuscript. CL, MAG, BCO, and RZ contributed to the data analysis and acquiring of funding. All Authors approved the final manuscript.

1	Funding	1
2	This work was supported in part by NSF Awards MCB-1120370 (MAG and RZ), MCB-1546402 (A. Von Arnim and	2
3	MAG), and DEB-1355033 (BCO, MAG, and RZ) with additional support from Department of Ecology &	3
4	Evolutionary Biology (EEB) at the University of Tennessee Knoxville (UTK) and the National Institute for	4
5	Mathematical and Biological Synthesis (NIMBioS), an Institute sponsored by the National Science Foundation	5
6	through NSF Award DBI-1300426. CL received support as a Graduate Student Fellow from NIMBioS with	
7	additional support from Departments of Mathematics and EEB at UTK.	
8	Ethics approval and consent to participate	6
9	Not applicable	7
10	Consent for publication	8
11	Not applicable	9
12	Competing interests	10
13	The authors declare that they have no competing interests.	11
14	Author details	12
15	¹ Department of Ecology & Evolutionary Biology, University of Tennessee, 37996, Knoxville, TN, USA. ² National	12
16	Institute for Mathematical and Biological Synthesis, 37996, Knoxville, TN, USA. ³ Max-Planck Institute of	13
17	Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307, Dresden, Germany. ⁴ Department of Business	
18	Analytics and Statistics, University of Tennessee, 37996, Knoxville, TN, USA.	14
19	References	15
20	1. Gouy, M., Gautier, C.: Codon usage in bacteria: correlation with gene expressivity. <i>Nucleic Acids Research</i> 10 , 7055–7074 (1982)	16
21	2. Ikemura, T.: Codon usage and tRNA content in unicellular and multicellular organisms. <i>Molecular Biology and Evolution</i> 2 , 13–34 (1985)	17
22	3. Bulmer, M.: The selection-mutation-drift theory of synonymous codon usage. <i>Genetics</i> 129 , 897–907 (1990)	18
23	4. Gilchrist, M.A.: Combining models of protein translation and population genetics to predict protein production	19
24	rates from codon usage patterns. <i>Molecular Biology and Evolution</i> 24 (11), 2362–2372 (2007)	
25	5. Sharp, P.M., Li, W.H.: The codon adaptation index - a measure of directional synonymous codon usage bias,	20
26	and its potential applications. <i>Nucleic Acids Research</i> 15 , 1281–1295 (1987)	
27	6. Wright, F.: The 'effective number of codons' used in a gene. <i>Genet</i> 87 , 23–29 (1990)	21
28	7. M, S.P., Stenico, M., Peden, J.F., Lloyd, A.T.: Codon usage: mutational bias, translational selection, or both? <i>Biochem Soc Trans.</i> 21 (4), 835–841 (1993)	22
29	8. Shah, P., Gilchrist, M.A.: Explaining complex codon usage patterns with selection for translational efficiency,	23
30	mutation bias, and genetic drift. <i>Proceedings of the National Academy of Sciences U.S.A</i> 108 (25), 10231–10236 (2011)	24
31	9. Wallace, E.W., Airoldi, E.M., Drummond, D.A.: Estimating selection on synonymous codon usage from noisy	25
32	experimental data. <i>Molecular Biology and Evolution</i> 30 , 1438–1453 (2013)	26
33	10. Gilchrist, M.A., Chen, W.C., Shah, P., Landerer, C.L., Zaretzki, R.: Estimating gene expression and	27
34	codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. <i>Genome Biology and Evolution</i> 7 , 1559–1579 (2015)	
35	11. Médigue, C., Rouxel, T., Vigier, P., Hénaut, A., Danchin, A.: Evidence for horizontal gene transfer in	28
36	<i>Escherichia coli</i> speciation. <i>Journal of Molecular Biology</i> 222 (4), 851–856 (1991)	
37	12. Lawrence, J.G., Ochman, H.: Amelioration of bacterial genomes: Rates of change and exchange. <i>Journal of</i>	29
38	<i>Molecular Biology</i> 44 , 383–397 (1997)	30
39	13. Marcet-Houben, M., Gabaldón, T.: Beyond the whole-genome duplication: Phylogenetic evidence for an ancient	31
40	interspecies hybridization in the baker's yeast lineage. <i>PLoS Biology</i> 13 (8), 1002220 (2015)	
41	14. Beimforde, C., Feldberg, K., Nylander, S., Rikkinen, J., Tuovila, H., Dörfelt, H., Gube, M., Jackson, D.J.,	32
42	Reitner, J., Seyfullah, L.J., Schmidt, A.R.: Estimating the phanerozoic history of the ascomycota lineages:	
43	combining fossil and molecular data. <i>Mol. Phylogenet. Evol.</i> 78 , 386–398 (2014)	33

- 1 15. Payen, C., Fischer, G., Marck, C., Proux, C., Sherman, D.J., Coppée, J.-Y., Johnston, M., Dujon, B.,
2 Neuvéglise, C.: Unusual composition of a yeast chromosome arm is associated with its delayed replication.
3 Genome Research **19**(10), 1710–1721 (2009)
4 16. Friedrich, A., Reiser, C., Fischer, G., Schacherer, J.: Population genomics reveals chromosome-scale
5 heterogeneous evolution in a protoploid yeast. Molecular Biology and Evolution **32**(1), 184–192 (2015)
6 17. Vakirlis, N., Sarilar, V., Drillon, G., Fleiss, A., Agier, N., Meyniel, J.-P., Blanpain, L., Carbone, A., Devillers, H.,
7 Dubois, K., Gillet-Markowska, A., Graziani, S., Huu-Vang, N., Poirel, M., Reisser, C., Schott, J., Schacherer,
8 J., Lafontaine, I., Llorente, B., Neuvéglise, C., Fischer, G.: Reconstruction of ancestral chromosome
9 architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. Genome
10 research **26**(7), 918–32 (2016)
11 18. Brion, C., Legrand, S., Peter, J., Caradec, C., Pfleiger, D., Hou, J., Friedrich, A., Llorente, B., Schacherer, J.:
12 Variation of the meiotic recombination landscape and properties over a broad evolutionary distance in yeasts.
13 PLoS Genetics **13**(8), 1006917 (2017)
14 19. dos Reis, M., Savva, R., Wernisch, L.: Solving the riddle of codon usage preferences: a test for translational
15 selection. Nucleic Acids Research **32**(17), 5036–5044 (2004)
16 20. Cope, A.L., Hettich, R.L., Gilchrist, M.A.: Quantifying codon usage in signal peptides: Gene expression and
17 amino acid usage explain apparent selection for inefficient codons. Biochimica et Biophysica Acta (BBA) -
18 Biomembranes **1860**(12), 2479–2485 (2018)
19 21. Shen, X.X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B., Wisecaver,
20 J.H., Wang, M., Doering, D.T., Boudouris, J.T., Schneider, R.M., Langdon, Q.K., Ohkuma, M., Endoh, R.,
21 Takashima, M., Manabe, R., Čadež, N., Libkind, D., Rosa, C., DeVirgilio, J., Hulfachor, A.B., Groenewald, M.,
22 Kurtzman, C., Hittinger, C.T., Rokas, A.: Tempo and mode of genome evolution in the budding yeast
23 subphylum. Cell **175**(6), 1533–154520 (2018)
24 22. Landerer, C., Cope, A., Zaretzki, R., Gilchrist, M.A.: AnaCoDa: analyzing codon data with bayesian mixture
25 models. Bioinformatics **34**(14), 2496–2498 (2018)
26 23. Tsankov, A.M., Thompson, D.A., Socha, A., Regev, A., Rando, O.J.: The role of nucleosome positioning in the
27 evolution of gene regulation. PLoS Biol **8**(7), 1000414 (2010)
28 24. Sokal, R.R., Rohlf, F.J.: Biometry - The principles and practice of statistics in biological, pp. 547–555. W. H.
29 Freeman, New York, NY (1981)
30 25. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q.: Iq-tree: A fast and effective stochastic algorithm
31 for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution **32**(1), 268–274 (2015)
32 26. Sella, G., Hirsh, A.E.: The application of statistical physics to evolutionary biology. Proceedings of the National
33 Academy of Sciences of the United States of America **102**, 9541–9546 (2005)
34 27. Wagner, A.: Energy constraints on the evolution of gene expression. Molecular Biology and Evolution **22**,
35 1365–1374 (2005)
36 28. Nagylaki, T.: Evolution of a finite population under gene conversion. Proc. Natl. Acad. Sci. U. S. A. **80**,
37 6278–6281 (1983)
38 29. Nagylaki, T.: Evolution of a large population under gene conversion. Proc. Natl. Acad. Sci. U. S. A. **80**,
39 5941–5945 (1983)
40 30. Harrison, R.J., Charlesworth, B.: Biased gene conversion affects patterns of codon usage and amino acid usage
41 in the *Saccharomyces* sensu stricto group of yeasts. Molecular Biology and Evolution **28**(1), 117–129 (2011)
42 31. Salichos, L., Rokas, A.: Inferring ancient divergences requires genes with strong phylogenetic signals. Nature
43 **497**, 327–331 (2013)
44 32. Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuela, P., Christen, J.A., Collado-Vides, J.: Successful lateral
45 transfer requires codon usage compatibility between foreign genes and recipient genomes. Molecular Biology
46 and Evolution **21**(10), 1884–1894 (2004)
47 33. Tuller, T., Girshovich, Y., Sella, Y., Kreimer, A., Freilich, S., Kupiec, M., Gophna, U., Ruppin, E.: Association
48 between translation efficiency and horizontal gene transfer within microbial communities. Nucleic Acids
49 Research **39**(11), 4743–4755 (2011). doi:10.1093/nar/gkr054
50 34. Ruderfer, D.M., Pratt, S.C., Seidl, H.S., Kruglyak, L.: Population genomic analysis of outcrossing and
51 recombination in yeast. Nature Genetics **38**(9), 1077–1081 (2006)
52 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53

- 1 35. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical
2 Computing, Vienna, Austria (2013). R Foundation for Statistical Computing. <http://www.R-project.org/>

3 36. Gronau, Q.F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D.S., Forster, J.J.,
4 Wagenmakers, E.J., Steingrover, H.: A tutorial on bridge sampling. *Journal of Mathematical Psychology* **81**,
5 80–97 (2017)

6 37. Legendre, P.: Lmodel2: Model II Regression. (2018). R package version 1.7-3.
7 <https://CRAN.R-project.org/package=lmodel2>

8 38. Soderlund, C., Nelson, W., Shoemaker, A., Paterson, A.: Symap A system for discovering and viewing syntenic
9 regions of fpc maps. *Genome Research* **16**, 1159–1168 (2006)

10 39. Soderlund, C., Bomhoff, M., Nelson, W.: Symap v3.4: a turnkey synteny system with application to plant
11 genomes. *Nucleic Acids Research* **39**(10), 68 (2011)

12 40. Marais, G., Charlesworth, B., Wright, S.I.: Recombination and base composition: the case of the highly
13 self-fertilizing plant *arabidopsis thaliana*. *Genome Biology* **5**, 45 (2004)

14 41. Lang, G.I., Murray, A.W.: Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*.
15 *Genetics* **178**(1), 67–82 (2008)

16 42. Wolfram Research Inc.: Mathematica 11. (2017). <http://www.wolfram.com>

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

1 **Supplementary Material**

2 Supporting Materials for *Unlocking a signal of introgression from codons in Lachancea kluveri using a*
mutation-selection model by Landerer et al..

3 **Table S1** Synonymous mutation codon preference based on our estimates of ΔM . Shown are the
 4 most likely codon in low expression genes for each amino acid in: *E. gossypii*, in the endogenous and
 5 exogenous genes of *L. kluveri*, and in the combined *L. kluveri* genome without accounting for the
 two cellular environments.

	Amino Acid	<i>E. gossypii</i>	Endogenous	Exogenous	Combined	
7	Ala A	GCG	GCA	GCG	GCG	6
8	Cys C	TGC	TGT	TGC	TGC	7
9	Asp D	GAC	GAT	GAC	GAC	8
10	Glu E	GAG	GAA	GAG	GAG	9
11	Phe F	TTC	TTT	TTT	TTT	10
12	Gly G	GGC	GGT	GGC	GGC	11
13	His H	CAC	CAT	CAC	CAC	12
14	Ile I	ATC	ATT	ATC	ATA	13
15	Lys K	AAG	AAA	AAG	AAA	14
16	Leu L	CTG	TTG	CTG	CTG	15
17	Asn N	AAC	AAT	AAC	AAT	16
18	Pro P	CCG	CCA	CCG	CCG	17
19	Gln Q	CAG	CAA	CAG	CAG	18
20	Arg R	CGC	AGA	AGG	CGG	19
21	Ser ₄ S	TCG	TCT	TCG	TCG	20
22	Thr T	ACG	ACA	ACG	ACG	21
23	Val V	GTG	GTT	GTG	GTG	22
24	Tyr Y	TAC	TAT	TAC	TAC	23
25	Ser ₂ Z	AGC	AGT	AGC	AGC	24
26						25
27						26
28						27
29						28
30						29
31						30
32						31
33						32

1		1				
2		2				
3		3				
4		4				
5		5				
6		6				
7		7				
8		8				
9		9				
10	Table S2 Synonymous selection codon preference based on our estimates of $\Delta\eta$. Shown are the most likely codon in high expression genes for each amino acid in: <i>E. gossypii</i> , in the endogenous and exogenous genes of <i>L. kluyveri</i> , and in the combined <i>L. kluyveri</i> genome without accounting for the two cellular environments.	10				
11		11				
12		12				
13	Amino Acid	<i>E. gossypii</i>	Endogenous	Exogenous	Combined	13
14	Ala A	GCT	GCT	GCT	GCT	
15	Cys C	TGT	TGT	TGT	TGT	
16	Asp D	GAT	GAC	GAT	GAT	
17	Glu E	GAA	GAA	GAA	GAA	
18	Phe F	TTT	TTC	TTC	TTC	
19	Gly G	GGA	GGT	GGT	GGT	
20	His H	CAT	CAC	CAT	CAT	
21	Ile I	ATA	ATC	ATT	ATT	
22	Lys K	AAA	AAG	AAA	AAG	
23	Leu L	TTA	TTG	TTG	TTG	
24	Asn N	AAT	AAC	AAT	AAC	
25	Pro P	CCA	CCA	CCT	CCA	
26	Gln Q	CAA	CAA	CAA	CAA	
27	Arg R	AGA	AGA	AGA	AGA	
28	Ser ₄ S	TCA	TCC	TCT	TCT	
29	Thr T	ACT	ACC	ACT	ACT	
30	Val V	GTT	GTC	GTT	GTT	
31	Tyr Y	TAT	TAC	TAT	TAC	
32	Ser ₂ Z	AGT	AGT	AGT	AGT	
33						33

