Published online 22 February 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 11* ***4743–4755***
doi:10.1093/nar/gkr054

Association between translation efficiency and horizontal gene transfer within microbial communities

Tamir Tuller[1,2,*], Yana Girshovich[3], Yael Sella[3], Avi Kreimer[3], Shiri Freilich[3,4], Martin Kupiec[5], Uri Gophna[5] and Eytan Ruppin[3,4]

[1]Faculty of Mathematics and Computer Science, [2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, [3]Blavatnik School of Computer Science, [4]School of Medicine, Tel Aviv University, Ramat Aviv 69978 and [5]Department of Molecular Microbiology and Biotechnology, Israel

Received November 4, 2010; Revised January 18, 2011; Accepted January 19, 2011

ABSTRACT

Horizontal gene transfer (HGT) is a major force in microbial evolution. Previous studies have suggested that a variety of factors, including restricted recombination and toxicity of foreign gene products, may act as barriers to the successful integration of horizontally transferred genes. This study identifies an additional central barrier to HGT—the lack of co-adaptation between the codon usage of the transferred gene and the tRNA pool of the recipient organism. Analyzing the genomic sequences of more than 190 microorganisms and the HGT events that have occurred between them, we show that the number of genes that were horizontally transferred between organisms is positively correlated with the similarity between their tRNA pools. Those genes that are better adapted to the tRNA pools of the target genomes tend to undergo more frequent HGT. At the community (or environment) level, organisms that share a common ecological niche tend to have similar tRNA pools. These results remain significant after controlling for diverse ecological and evolutionary parameters. Our analysis demonstrates that there are bi-directional associations between the similarity in the tRNA pools of organisms and the number of HGT events occurring between them. Similar tRNA pools between a donor and a host tend to increase the probability that a horizontally acquired gene will become fixed in its new genome. Our results also suggest that frequent HGT may be a homogenizing force that increases the similarity in the tRNA pools of organisms within the same community.

INTRODUCTION

Horizontal gene transfer (HGT), the passage of genetic material between organisms and lineages, is a major force in microbial evolution (1–3). It is estimated that >80% of gene families in prokaryotic genomes underwent HGT at some point of their evolutionary history (1). One fundamental question in this context is what determines the evolutionary success of HGT events and supports the fixation of the transferred genes. Previous studies in the field have identified several factors that influence the chance of successful HGT, including toxicity of the acquired gene product (2), the ability of the transferred segment to integrate into the host genome by recombination (4), the number of protein interactions of the transferred gene product (5) and the function of the gene(s), since genes that confer specific advantages are more likely to be retained. (6).

Recently, Kudla *et al*. (7) cloned differently encoded green fluorescent protein (GFP) variants into *Escherichia coli* and showed that over-expressing a single gene with incompatible codon usage can be severely deleterious to the organism. It is therefore perhaps not surprising that the codon usage in many viruses is adapted to the tRNA pools of their hosts (8). Although looking for genes with atypical codons is a common method for detecting HGT, Medrano-Soto and coworkers demonstrated that many horizontally-acquired genes were in fact compatible with the recipient genome's codon usage (9). Codon bias, however, is not necessarily driven by translation efficiency alone, and can sometimes be the product of constraints such as GC content, and/or amino acid usage.

*To whom correspondence should be addressed. Tel: +972 8 9344214; Fax: +972 8 934 4122; Email: tamirtul@post.tau.ac.il

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this work, we use a direct measure of translation efficiency and show (while controlling for other variables) that there is an association between translation efficiency and HGT. In the first three sections we show that this association exists at the levels of gene families, pairs of organisms and communities of organisms. Specifically, in the first section, we show that genes that have comparable tRNA adaptation values across organisms tend to be exchanged more often. In the second section, we show that organisms with similar tRNA pools tend to share genes more frequently, generating dynamic networks of increased gene exchange. In the third section, we show that pairs of organisms that share a community (and thus usually undergo more frequent HGT) have similar tRNA pools.

In the two last sections we study the potential forces of selection underlying the associations described above. Specifically, we suggest that the bi-directional relation between translation efficiency and HGTs is related to the relation between ribosome allocation and translation speed (7,10). In section four, we explain why the adaptation of the codon bias of the transferred gene to the tRNA pool of the host is a major determinant of the success of HGT. In section five we explain how frequent HGTs can homogenize the tRNA pools of organisms within a community.

MATERIALS AND METHODS

Data

Information about gene HGT and gene sharing. Data about gene sharing and inferred HGT between ancestral and extant organisms were obtained from Dagan *et al*. (1). Data about genes that underwent HGT (6) was kindly provided by Y. Nakamura. Additional data about gene sharing were obtained from (11).

Information about community structure. These data were downloaded from (12).

Coding sequences. All the coding sequences were downloaded from the NCBI site (http://www.ncbi.nlm.nih.gov/Ftp/).

tRNA copy numbers. Organismal tRNA copy numbers were downloaded from the Genomic tRNA Database (http://lowelab.ucsc.edu/GtRNAdb/) (13).

Gene expression in E. coli. These data were downloaded from (14).

The phylogenetic tree for reconstructing the ancestral tRNA pool

The phylogenetic tree was obtained from (15). This tree was reconstructed based on the rRNA operon (16S, 23S and 5S) from all 190 genomes using the maximum likelihood approach. The final phylogenetic tree includes 190 organisms. The list of organisms and their taxonomy appears in (1). We used Neyman's two state model (16), a version of Jukes Cantor (JC) model (17) for inferring the edge lengths of the tree (the probability of gain/loss of a gene family) by maximum likelihood, as implemented in PAML (18). The edge lengths correspond to the probabilities that a protein family will appear/disappear along the corresponding lineage.

Reconstruction of ancestral tRNA copy number

Let $p_{\alpha,\beta}$ denote the probability of gain/loss of a gene family along the edge tree edge (α,β). Let C_x denote the copy number of tRNA C in node (genome or ancestral genome)x of the evolutionary tree. We inferred the ancestral tRNA copy number using a generalized maximum parsimony method, whose penalty for a change in the number of genes corresponding to a tRNA C along the tree edge (α,β) is $-\log(p_{\alpha,\beta})^*|C_\alpha - C_\beta|$.

Computing tAI *and* tRs. tAI was computed following the work of dos Reis *et al*. (19), who defined this index. This measure gauges the availability of tRNAs for each codon.

The similarity in tRNA pools between two organisms, *tRs*, is defined as the non-parametric Spearman correlation between the two vectors (of length 61) of their codons' tAI values (denoted by *tRs*; see also Supplementary Note S6). The tAI of a gene is the geometric mean of the tAI of all its codons (additional technical details are provided in Supplementary Note S1).

Connecting evolutionary changes in tRNA copy numbers and HGT

Let (x,y) denote an edge in the evolutionary tree described above where x is the node that is closer to the root of the tree.

For each pair of edges, (x1,y1) and (x2,y2), in the evolutionary tree we computed the number of HGT along the corresponding evolutionary interval (x1 to y1 and x2 to y2; i.e. the HGT between x1 and x2 or between y1 and y2) and the change in the tRs along the edge. We correlated these values to show that along different paths of the evolutionary tree an increased number of HGT corresponds to increase in tRs. This also represents an additional control for phylogenetic proximity.

Estimating the number of HGTs a gene family has undergone. We ranked the gene families (COGs) according to their tendency to undergo HGT in the following way: first we mapped all the genes to their COGs; next we used the data of Nakamura *et al*. (6) to count the number of times (number of organisms) a gene corresponding to each COG was detected as a recently transferred gene.

Measures for the variability and robustness of the tAI of a gene family. We considered two measures for the robustness of the tAI of particular gene families:

The first measure, *VtAI*, is the standard deviation (SD) of the tAI of a COG across organisms. COGs with higher SD of tAI are those whose codon bias and/or the tRNA pool that recognizes their codons is more variable between organisms, suggesting different levels of translation efficiency of gene homologs in different organisms.

The second measure, tRNA robusteness, *RtAI*, was computed as follows: First, we computed for each codon the SD of the tAI score of that codon across all the analyzed organisms; let $t\mathrm{STD}_i$ denote the SD of the tAI of codon i. Next, we computed for each COG (C_i) in each organism O_j the mean SD of the tAI of its codons:

$$\mathrm{Etstdev}(C_i,O_j) = \frac{\sum_{k=1}^{l_{ci,oj}} t\mathrm{SD}_k}{l_{ci,oj}}$$

Where $t\mathrm{SD}_k$ is the mean SD of the tAI of the codon defined by the k'th triplet on gene g; and $l_{ci,oj}$ is the length of COG C_i in organism O_j.

The final *RtAI* score for COG C_i is the mean $Et\mathrm{SD}(C_i,O_j)$ over all the organisms in the database.

$$RtAI = \frac{\sum_{j=1}^{N} Et\mathrm{SD}(C_i,O_j)}{N}$$

Where N is the number of organisms in the database. A COG with higher *RtAI* has a codon bias that is less robust to changes in the tRNA pool (e.g. due to HGT). To control for the fact that some of the COGs appear only in very few organisms we considered only COGs that appear in at least 85 organisms (i.e. >50% of the organisms in the data set; higher cut-offs gave very similar results). In addition, the number of organisms possessing a COG was always included as one of the covariates in the partial correlations.

Comparison between the number of HGT events and the variability of the tAI among COGs

To compare the variability in *tAI* versus the number of HGT events that are related to each COG, we counted the HGT predictions that are related to each COG (6), and the mean *VtAI* and *RtAI* corresponding to each COG, and subsequently correlated these measures.

The RVtAI—a combination of *VtAI* and *RtAI*

RVtAI was computed in the following way: first, we ranked all the COGs by their *VtAI* and *RtAI*. Next, for each COG we averaged the ranks of these two parameters. Thus, this index is based 'both' on the variability of the tAI of a gene among the analyzed organisms (*VtAI*) and the robustness of the translation rate of the gene to changes in the tRNA pools (*RtAI*) and better reflects the robustness of the gene to changes in the tRNA pool than either *VtAI* or *RtAI* individually.

Controls for possible covariates

Controls when comparing the tRs of pairs of organism to the number of HGTs between them or to community similarity. The following variables were used as covariates (together) in the multivariate analysis of tRs versus the number of HGT events/sharing a niche (see more details below).

GC content. For each organism we computed the G+C (GC) content (the percentage of G+C in its genome). To control for the possibility that the correlation between tRs and the number of HGT events is related to GC content, we computed for each pair of organisms the absolute value of the difference in their GC content.

Amino acid usage. For each organism we computed the amino acid usage (the vector of frequencies of all the 20 amino acids in its coding sequences). To control for amino acid usage, we computed for each pair of organisms the correlation between their vectors of amino acid usage, ARs.

Phylogenetic distance. To control for phylogenetic distance, we computed for each pair of organisms their distance in the evolutionary tree based on the inferred edge lengths (see above).

Control for the correlation between codon bias and translation efficiency

We computed for each organism the correlation between the codon bias and tAI of codons; next, we computed for each pair of organisms the absolute value of the difference in their CB-tAI correlations.

Controls for growth rate

The generation times of 214 organisms were downloaded from (20). To control for the possibility that the correlation between tRs and the number of HGT events is related to growth rate, we computed for each pair of organisms the absolute value of the difference in their generation times.

Controls for genome size

We considered two variables: the difference between the number of genes in each pair of genomes and the sum of the number of genes in each pair of genomes.

Controls for the comparison of COGs. The following variables were used as covariates (together) in the multivariate analysis of *RVtAI* and the number of HGT of a COG (see more details below).

GC content control. We computed for each COG in each organism the GC content of the COG in the organism. Next, we computed the SD of the GC of the COG.

Amino acid control. We computed for each amino acid, in each organism, the expected tAI, EtAI, which is the weighted average of the tAI (taking into account the codon bias of the organism) of all the codons that code for this amino acid. The EtAI of a COG in a certain organism is the geometric mean of the EtAI of all the amino acids in the COG. Next, we computed the SD of the EtAI of the COG.

Control for expression levels. It is well know that the expression levels of genes is strongly correlated with their tAI (10,21,22). Thus, to control for the possibility that the correlation between *RVtAI* and the number of HGT

may be explained by the expression levels of the COG we first ranked the COGs according to their tAI in each organism (in each organism we computed the for each COG the percentage of COGs whose expression is higher than the expression level of the COG).

Next, we computed for each COG its mean tAI rank across all the organisms. For example, when we analyzed the 1594 COGs that appear in more than 50 organisms we found that the top 100 COGs with the highest mean tAI ranks include 16 ribosomal proteins (enrichment P-value $= 1.6 \times 10^{-7}$) and the lowest 100 COGs include no ribosomal proteins. It is known that the expression levels of ribosomal proteins are very high [see, for example, (23)] thus, this fact demonstrates that indeed our measure is a good approximation of expression levels.

Non-parametric multivariate analysis

Let X and Y denote two variables and $Z = [Z1, Z2, Z3,\ldots]$ denote a set of variables.

The non-parametric multivariate analysis that is reported in this paper includes partial Spearman correlations of the form $R(X,Y|Z)$. Roughly, if such a correlation is significant it means that there is a relation between X and Y that can not be explained by the variable Z.

In the first section of the paper we tried to find distinct correlation with 'genes shared among organisms' (X in our case) we considered each of the variables defined above as the explaining variables (Y, specifically, the *tRs*); where in each case the rest of the variables were used as covariates (Z). In the second and the third sections similar analyses were performed for the number of HGT of a COG (X, second section) and 'community co-membership' (X, third section).

Finally, note that in our case (as is usually the case when biological data are analyzed) the data points are not independent, and this may affect the results (increase or decrease the correlations).

Statistical tests and correlations. We used the following non-parametric tests: Spearman correlation, partial Spearman correlation and KS-test. The statistical analysis was performed in MATLAB. We reported both asymptotical P-values and empirical P-values for the Spearman correlations. The empirical P-values for a Spearman correlation, $R(X,Y|Z)$ were computed by randomly permuting X and $Yn = 1000$ times and counting the number of times the correlation of the permuted vectors was higher or equal to the original correlations (let m denote this number). The empirical P-value is $pe = m/n$.

A network of connections between pairs of organism with high *tRs*, modularity and significance

This network includes a node for each organism [a total of 648 organisms whose tRNA copy number was available (13)]; in this network, two organisms were connected by an edge if their *tRs* score was >97% of the *tRs* scores.

The modularity score of a network is a number between 0 and 1, where larger numbers correspond to stronger signal of modularity [see details in (24)]. The modularity score of the original network of tAI similarity was very high: 0.6518. We compared the modularity score of the original network to the score of 100 random networks that maintain the degree distribution of the original network; the random networks were generated similarly to the way such a network was generated in (25). In the case of the random networks the score was much lower (mean = 0.1235, SD = 0.0196). These results suggest that the modularity score of the network of tAI similarity is significantly higher than what is expected from random networks with similar properties (empirical $P < 0.01$).

We also compared the modularity score of the original network to the score of 20 random networks that were generated in a different way: each random network was generated as before but based on a permuted version of the tAI values of the codons of each organism. In this case also, the modularity score of the random networks was significantly lower (mean = 0.2451, SD = 0.0054; empirical $P < 0.05$).

Optimization of the correlation between the tAI and the expression levels

We computed the tRNA pool that optimizes the correlation between the tAI of genes and their mRNA levels [$R(tAI, mRNA)$] by performing a heuristic, hill climbing process. The procedure maintained the total sum of tRNA copy numbers and the tRNA copy numbers that are equal to zero [as was performed in (26)]. The starting point used was the original tRNA pool of the genome.

RESULTS

Gene families with more universally conserved tAI across organisms are exchanged more frequently

Based on the genomic tRNA copy number, a proxy for the expression levels of tRNAs [(27–30) and Supplementary Note S1], we computed the tRNA adaptation index (tAI) for 190 prokaryotes that appear in (1). The tAI of a codon [which is equal to its w-value (15); see 'Materials and Methods' section and Supplementary Note S1] is a measure of its translation efficiency (19,22) that combines the tRNA gene copy number with the efficiency of the codon–anticodon binding by the tRNA. This measure is more informative than the copy number of the tRNAs that recognize that codon ('Materials and Methods' section).

The tAI of a gene is the geometric mean of the tAI of its codons (19,22) which ranges between 0 (non-efficient translation) and 1 (highly efficient translation; see 'Materials and Methods' section and Supplementary Note S1). Unlike other measures such as codon bias or the codon adaptation index (CAI), the tAI is not only a good predictor of protein abundance and translation rate (10,19,21,22), but also measures the co-adaptation between codon usage and the tRNA pool directly (31) (Supplementary Table S1 provides the tRNA copy number and tAI scores of all the codons in all the analyzed organisms).

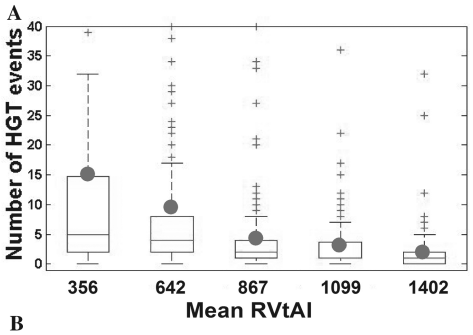
If a gene's tAI is a major determinant of its HGT frequency, then one would expect to see that genes with more

uniform levels of tAI across species tend to have higher levels of HGT. To test this hypothesis, we used the data of Nakamura *et al*. (6), who performed a large scale study of single genes that were horizontally transferred fairly recently [see also (32) and Supplementary Note S2]. We analyzed 4614 COGs (33) across 163 prokaryotic genomes, counting only the COGs that appear in more than half of the organisms (i.e. at least in 85 organisms). We considered two measures of tAI variance: the first one is simply the variance in the gene's tAI, i.e. to what degree does the level of translational efficiency of a gene vary across taxa (this measure was names *VtAI*, see 'Materials and Methods' section for details). Indeed, the correlation between the *VtAI* of a gene and the number of times that this gene has been horizontally transferred (controlling for the fact that some genes appear in fewer organisms; see 'Materials and Methods' section) is significant and negative ($r = -0.2958$; $P < 10^{-16}$; $pe < 0.001$; n = 917).

The second measure, *RtAI*, measures the robustness of the translation efficiency (tAI) of each COG to different tRNA pools (a higher RtAI means less robustness). This measure is based on the fact that some codons are more robust than others to changes in the tRNA pool across the analyzed genomes ('Materials and Methods' section). When we compared *RtAI* and the number of HGT events of each COG (again controlling for the fact that some genes appear in fewer organisms; see 'Materials and Methods' section) we, once more obtained a highly significant correlation ($r = -0.4125$ and $P < 10^{-16}$, $pe < 0.001$; $n = 917$). This result suggests that genes whose translation efficiency is less sensitive to changes in the tRNA pool (i.e. to the host where they are expressed) have been transferred more often.

To better understand how different features of a COG explain the number of times it has been horizontally transferred, we computed a combined and more robust measure of tAI robustness, RVtAI, (a combination of the two previous measures, where higher RVtAI means less robustness to changes in the tRNA pool, see 'Materials and Methods' section). As can be seen in Figure 1A, the correlation between the *RVtAI* of a COG and the number of times that COG was horizontally transferred is very high ($r = -0.4340$; $P < 10^{-16}$). Specifically, the 20% of the COGs with the lowest *RVtAI* scores have been horizontally transferred over 7-fold more frequently than the 20% of the COGs with the top *RVtAI* (Figure 1A) demonstrating again the strong relation between robustness to tRNA pools and HGT.

Thus, for example, the type I restriction-modification enzyme S subunit (COG0732V), known to be frequently mobile (34) and the thymydilate synthase protein ThyX (COG1351F), shown recently to be transferred multiple times in evolution (35), both have low *RVtAI*. In contrast, several ribosomal proteins (S2, S6, L7/12, L9, L21, L24P, L25, L36) considered to be rarely transferred (36), have high *RVtAI* values and poor robustness, although they tend to have high *tAI* values in most genomes (COGs with extreme *RVtAI* appear in Supplementary Table S2).



B

RVtAI	Varianve in GC content	Esimated mean expression	Variance in amino acid bias
-0.1125 ($p = 6.6*10^{-4}$; pe <0.001)	-0.11 ($p = 7.9*10^{-4}$; pe <0.001)	-0.2371 ($p = 3.9*10^{-13}$; pe < 0.001)	O02 (p = 0.5457; pe = O24)

C

No control	A	A + B	A + B + C
-0.434 ($p < 10^{-16}$; pe <0.001)	-0.16 ($p = 1.1*10^{-6}$; pe <0.001)	-0.14 ($p = 2.17*10^{-5}$; pe <0.001)	-0.1125 ($p = 6.64*10^{-4}$; pe <0.001)

Figure 1. (**A**) The number of HGT events (six) as a function of *RVtAI* across five bins of equal size: a Whisker plot with the means marked with red circles. (**B**) Correlations with the number of HGT events when controlling for all the other factors; *P* denotes asymptotic *P*-value, *pe* denotes empirical *P*-value ('Materials and Methods' section). (**C**) Correlation given increasing number of factors (A) estimated mean expression levels, (B) variance in GC content and (C) variance in amino acid bias.

As before, we considered other COG features that may potentially explain the correlation above: (i) An estimation of the mean expression level of the COG across the analyzed organisms [based on the tAI, which is often a better predictor of protein abundance than mRNA levels (21,22), which generally correlate well with tAI; see 'Materials and Methods' section]. This is pertinent as it is known that highly expressed genes are involved in less HGT events and are also under stronger selection for translation efficiency (7,10,37). (ii) The variation in the GC content of the COG across all organisms. (iii) The variation in the amino acid usage across these organisms (as the set of organisms used for computing the RVtAI is identical in all COGs there is no need here to control for the genome size or effective population size).

As depicted in Figure 1B, the correlation between the *RVtAI* of a COG and the number of times a COG was horizontally transferred remains significant even when controlling for all the three variables above 'together' ($r = -0.1125$; $P = 6.6 \times 10^{-4}$, $pe < 0.001$); most of the decrease in correlation is due to the estimation of expression levels (Figure 1C).

Organisms that are involved in more HGT have more similar tRNA pools

The translational efficiency pattern (TEP) of the codons in each organism (based on the tRNA pool) can be described by a vector of length 61 (the number of coding triplets), where each entry denotes the tRNA adaptation index (tAI) of the respective codon (which is more informative

then just the tRNA copy number, see 'Materials and Methods' section).

The similarity in the TEP of tRNA pools between any two organisms is then measured by the Spearman correlation between their corresponding TEP vectors, denoted as *tRs*. Accordingly, two organisms have a higher *tRs* if they have a more similar rank order of codons' translation efficiency.

We used the network of gene sharing from (1), in which a gene is defined as shared by two organisms if these organisms have relatively similar orthologs of that gene. Gene sharing between a pair of organisms can arise when the two are evolutionarily close and the gene was inherited vertically by both of them, or when a gene was transferred horizontally between these organisms or their ancestors. It is important to note that this network is protein-based, i.e. derived from 'amino acid alignments', and therefore is independent of codon information. Analyzing this network, we observed that *tRs* of organisms that share more genes tend to be more similar (Figure 2A and B; Spearman correlation above 0.32, $P < 10^{-30}$, empirical $P < 0.\ 001$). Similar findings were obtained when analyzing the data from Beiko *et al.* [(11); see Supplementary Note S3].

We aimed to distinguish between two possible explanations for this observation that are not necessarily mutually exclusive: (i) organisms that are involved in HGT are part of the same community and thus are under similar ecological constrains that select for the use of similar amino acids [e.g. see (38)] or a similar GC content, similar growth rates, etc. Due to this environmental similarity, the need to optimize the co-adaptation between their codon bias and the tRNA pool subsequently drives their tRNA pools to become similar. The association between HGT and tRNA similarity could therefore be merely a consequence of the shared niches and/or other variables. (ii) There are more successful HGT events between organisms with similar tRNA pools, by virtue of their codon compatibility, and this is not merely a by-product of the communities shared.

To distinguish between these two hypotheses we examined the association between *tRs* and gene sharing while controlling for six possible variables that may explain this association: (i) genome size (the sum of the genome sizes for each pair and the differences in genome size for each pair) which is correlated with population size of the corresponding organisms (39); (ii) phylogenetic distance (i.e. distance on the tree of life, see 'Materials and Methods' section); (iii) similarity in the amino acid usage; (iv) selection for translation efficiency (the correlation between codon bias and their translation efficiency); (v) similarity in the GC content; (vi) similarity in the optimal growth rate (20); and (vii) in addition, we controlled for a variable corresponding to the community associations of the organisms [a binary variable: 1/0—depending on whether or not these organisms have been shown to reside in the same community (12)].

Remarkably, the Spearman correlation between the number of shared genes and *tRs* for pairs of organisms remains significant also after controlling for all confounding factors defined above 'together' (as listed in the previous subsection). This result shows that both within and outside a community, HGT occurs more frequently between organisms with higher *tRs* ($r = 0.212$; $P = 1.18 \times 10^{-8}$; empirical *P*-value $pe < 0.001$ when controlling for all the variables above 'together'). Figure 2C and D details how the different variables above contribute to the relation between the number of HGT events [90% similarity cut-off for gene sharing (1)] and the *tRs* (similarity in the tRNA pools) of pairs of organisms. Similar correlations were obtained for lower similarity cut-offs definitions of gene sharing (1).

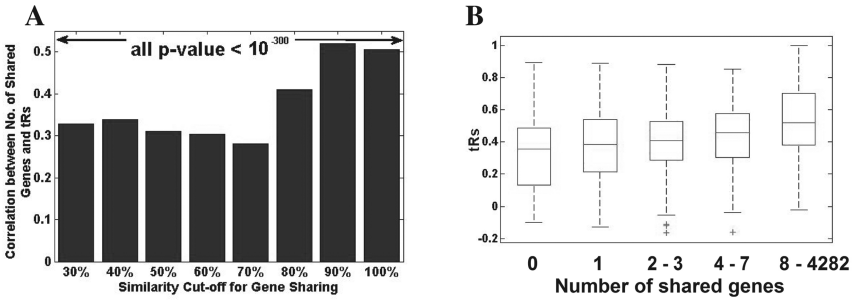
Based on extant species tRNA abundances, tRNA pools for ancestral organisms (i.e. ancestral nodes in the phylogenetic tree) can be reconstructed ('Materials and Methods' section). When we compared the ancestral) HGT frequencies between pairs of organism and their ancestral *tRs* along the analyzed phylogenetic tree ('Materials and Methods' section), the correlation obtained was 0.25 ($P < 10^{-16}$, $pe < 0.001$); see Figure 2E. Specifically, pairs of organisms whose *tRs* is high (mean $tRs = 0.82$) have been involved in 24 times more HGT events than organisms with very low *tRs* (mean $tRs = 0.76$; Figure 2E).

In addition, as an additional control for phylogenetic proximity, we computed the correlation between the HGT along pairs of edges and the change in the tRs along the pairs of edges (see details in the 'Materials and Methods' section). Indeed, we found a significant correlation between these variables showing again that more HGT corresponds to an increase in tRs ($r = 0.42$; $P < 10^{-16}$, $pe < 0.001$). The results reported above can therefore be extended to longer evolutionary timescales and to different periods of the evolution of the analyzed organisms.

Organisms that live in the same environment have similar tRNA pools

In the previous sections, we showed that there is an association between HGT and similarity in tRNA pool measured by the *tRs*. It is known that organisms that share the same environment are involved in more HGT (for example, the correlation between community co-membership and the number of shared genes when controlling for all the other variables defined in the pervious section is $r = 0.23$, $P = 2.4 \times 10^{-10}$, $pe < 0.001$). Thus, we aimed to compare the similarity in the tRNA pools, measured by the *tRs*, within and between communities, using the environmental partitioning established by Freilich *et al.* (12). Indeed, microorganisms that occupy similar niches have significantly higher *tRs* than those that live in different ones ($P = 2 \times 10^{-30}$; Figure 3A).

In the next stage, we aimed to distinguish between the two possible explanations for this observation that have already been discussed before (in the section about the relation between gene sharing and similar tRNA pool). To this end, we considered the variables defined in the previous section and performed a multivariate non-parametric analysis to understand how the different variables are associated with 'community co-membership'



C

tRs	*community co-membership*	*Sum of Genome sizes*	*Phylogenetic distance*	*Similarity in the GC content*	*Similarity in the AA bias*	*Similarity in the growth rate*	*Selection for translation efficiency*
0.212 ($p = 1.18*10^{-8}$; pe < 0.001)	0.21 ($p = 1.514*10^{-8}$; pe <0.001)	0.25 ($p = 4.88*10^{-12}$; pe <0.001)	-0.2826 ($p = 1.466*10^{-14}$; pe <0.001)	0.0891 (p = 0.0173; pe = 0.008)	0.063 (p = 0.093; pe = 0.039)	0.0434 (p = 0.247; pe = 0.124)	0.009 (p = 0.8095; pe = 0.3930)

D

No control	**A**	**A + B**	**A + B + C**	**A + B + C + D**	**A + B + C + D + E**	**A + B + C + D + E + F**	**A + B + C + D + E + F +G**	**A + B + C + D + E + F +G+H**
053 ($p < 10^{-16}$; pe < 0.001)	0.29 ($p = 3.89*10^{-15}$; pe < 0.001)	0.26 ($p = 7.45*10^{-13}$; pe < 0.001)	0.24 ($p = 8.9*10^{-11}$; pe < 0.001)	0.23 ($p = 2.46*10^{-10}$; pe <0.001)	0.224 ($p = 1.33*10^{-9}$; pe < 0.001)	0.21 ($p = 1.68*10^{-8}$; pe < 0.001)	0.21 ($p = 1.02*10^{-8}$; pe < 0.001)	0.21 ($p = 1.18*10^{-8}$; pe < 0.001)

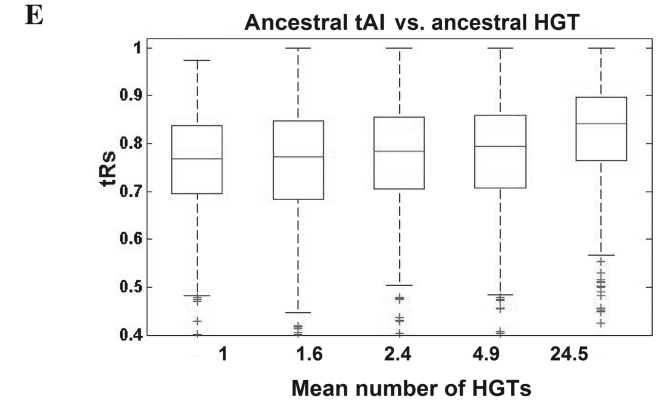
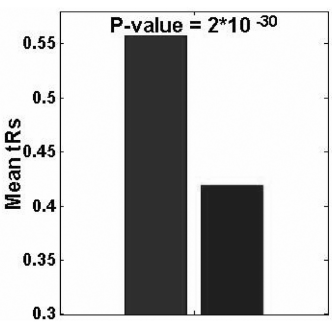


Figure 2. (**A** and **B**) The *tRs* values of organisms that share more genes are more similar. (A) Correlation between number of shared genes and *tRs* for different cutoffs of gene sharing (1). (B) Whisker plot (five bins of equal size) of *tRs* versus number of shared genes (cut-off of 70%; five bins equal size). (**C**) Spearman correlations of various variables with the number of shared genes when controlling for all the other factors; *P* denotes asymptotic *P*-value, *pe* denotes empirical *P*-value ('Materials and Methods' section). (**D**) Correlation given an increasing number of factors (A) phylogenetic distance, (B) sum of genome sizes, (C) community co-membership, (D) selection for translation efficiency, (**E**) similarity in the GC content, (F) similarity in the amino acid bias, (G) similarity in the growth rate and (H) difference in genome sizes; most of the decrease in correlation is due to genome sizes, phylogenetic distance and similarity in amino acid bias. (E) Whisker plot of the *tRs* for pairs of ancestral organisms versus the mean number of ancestral HGT between them (*x*-axis) for five bins of equal size.

and to study how they affect the relation between 'community co-membership' and *tRs* ('Materials and Methods' section). Figure 3B depicts the correlations between community co-membership and 'each of the variables' when controlling for the rest of the variables (i.e. the distinct correlation between community co-membership and each of the variables; 'Materials and Methods' section). As depicted in Figure 3B, the correlation between the community co-membership and tRs remains significant even when controlling for all the variables above combined ($r = 0.076$; $P = 0.001$, empirical $pe < 0.001$; 'Materials and Methods' section); similarly to the analysis reported in the previous subsection for HGT, most of the decrease in correlation is due to the contribution of genome sizes, phylogenetic distance and similarity in amino acid usage (Figure 3C). When we added to the

A



B

tRs	*Selection for translation efficiency*	*Sum Genome sizes*	*Phylogenetic distance*	*Similarity in the GC content*	*Similarity in the AA bias*	*Similarity in the growth rate*
0.076 (p= 0.001; pe <0.001)	-0.0518 (p = 0.032; pe = 0.011)	0.088 (p = 2.74*10^{-4}; pe < 0.001)	-0.085 (p = 3.86*10^{-4}; pe < 0.001)	0.023 (p= 0.34; pe = 0.162)	0.06 (p= 0.013 ; pe = 0.003)	0.005 (P= 0.84; pe = 0.428)

C

No control	A	A + B	A + B + C	A + B + C + D	A + B + C + D + E	A + B + C + D + E + F	A + B + C + D + E + F+G
0.4 8 (p < 10^{-16}; pe < 0.001)	0.1658 (p = 5.28* 10^{-12} ; pe < 0.001)	0.1 6 (p = 3. 59* 10^{-11} ; pe < 0.001)	0.0854 (p = 4.12* 10^{-4} ; pe < 0.001 (	0.07 6 (p= 0.001; pe < 0.001)	0.0769 (p= 0.001 ; pe < 0.001)	0.076 (p= 0.001 ; pe < 0.001)	0.076 (p= 0.001 ; pe < 0.001)

Figure 3. (**A**) The *tRs* of organisms that live in the same community [(12), left] are higher than the *tRs* of organisms from different communities (right). The *y*-axis is the mean *tRs* of the organisms in each group; we used the Kolmogorov–Smirnov (KS)-test for computing *P*-values. (**B**) Non-parametric correlations of several variables with niche-sharing, when controlling for all the other variables. *p* denotes asymptotic *P*-value, *pe* denotes empirical *P*-value ('Materials and Methods' section). (**C**) Correlation given increasing number of factors (A) sum genome of sizes, (B) phylogenetic distance, (C) similarity in the amino acid usage, (D) selection for translation efficiency, (E) similarity in the GC content, (F) similarity in the growth rates and (G) similarity in genome sizes.

list of covariate variables the number of shared genes (shown to be correlated with *tRs* in the previous sub-section) the correlation became not statistically significant($r = 0.018$; $P = 0.63$; $pe = 0.3030$). This is in contrast to the number of HGT events that was significantly correlated with *tRs* even when controlling for community co-membership (see the previous sub-section).

Finally, the fact that there is a significant relation between similarity in tRNA pools and a shared community, suggests that similarity in tRNA pools (e.g. the *tRs*) can be used for clustering organisms into their communities. Indeed, an initial analysis reported in Supplementary Note S4 supports this conjecture.

The effect of codon (dis)similarity of transferred genes on the recipient organism

Over-expressing a gene with maladapted codons is deleterious to the organism, because ribosomes would then spend too much time on its slow translation, leading to inefficient ribosome allocation and delayed growth (31). Accordingly, a significant positive correlation was previously shown between codon optimality of highly transcribed genes [measured by the CAI (31)] and fitness ($r = 0.54$; $P < 10^{-13}$) (7); see Figure 4A for the corresponding relation between tAI and fitness ($r = 0.52$; $P < 1.7 \times 10^{-11}$, $P < 0.001$). As a general rule, the fraction of the genes acquired by HGT that is highly transcribed is much lower than the fraction of highly expressed native genes [see, for example, (40,41)], mostly due to their foreign promoter regions. However, some horizontally acquired genes, especially of phage origin, can be highly transcribed [(14), see the next section]. Those genes can reduce host fitness if their codons are highly maladapted even if the protein they encode does not have a functional role in the recipient organism (42,43). In contrast, there are rare examples [Sorek *et al.* (2)] in which a 'specific' protein encoded by a foreign gene is somehow deleterious to the recipient organism. In these special cases, having highly compatible codons is likely to result in higher expression levels (more abundant deleterious proteins) leading to 'increased damage' to the host. Analyzing the data of Sorek *et al.*, who identified a relatively small number of genes (65 genes) that could not be cloned in *E. coli* due to the 'toxicity' of the corresponding proteins (2), we indeed found a significant 'positive' correlation between the number of genes from other organisms that cannot be cloned in *E. coli* and the similarity in the tRNA pools of donor organisms and *E. coli* ($r = 0.34$, $P = 0.0066$); see Figure 4B. Thus, as one would expect, the closer the tRNA pools are, the functional effects of genes whose proteins are toxic becomes larger, and hence the number of such 'forbidden', potentially toxic HGT events increases.

The results reported in the previous sections (the positive correlation between similarity the *tRs* and the number of HGTs) suggest that genes that produce highly deleterious proteins in the new host are 'relatively rare'. Thus, in general, it appears that beneficial genes and highly-transcribed neutral genes are frequent enough to create a selective pressure for similar tRNA pools in organisms from the same community and/or those that are involved in HGT, overcoming the inverse pressure exerted by the few genes whose proteins are 'deleterious' to the recipient organism.

HGTs may create a selective pressure for similar tRNA pools

In the previous sections we showed that organisms that share a community have more similar tRNA pools (i.e. more similar *tRs*), but how has this similarity evolved?

To try and answer that question we returned to the data set of Nakamura *et al.* (6), which contains 214 715 genes in the organisms analyzed (163 organisms in total) 12 897 of which underwent a relatively recent transfer [i.e. these genes have a foreign nucleotide composition and are not fully ameliorated (41)]. On average, in each of the organisms analyzed around 6% of the genes are 'recent' acquisitions. The method of Nakamura *et al.* (6) is based on codon bias; therefore, this is probably a 'lower bound' on the actual number of HT genes [see also (3)]. As we explain below, this 'non-negligible' fraction of genes that has been acquired can trigger changes in the copy number of some of the tRNA genes.

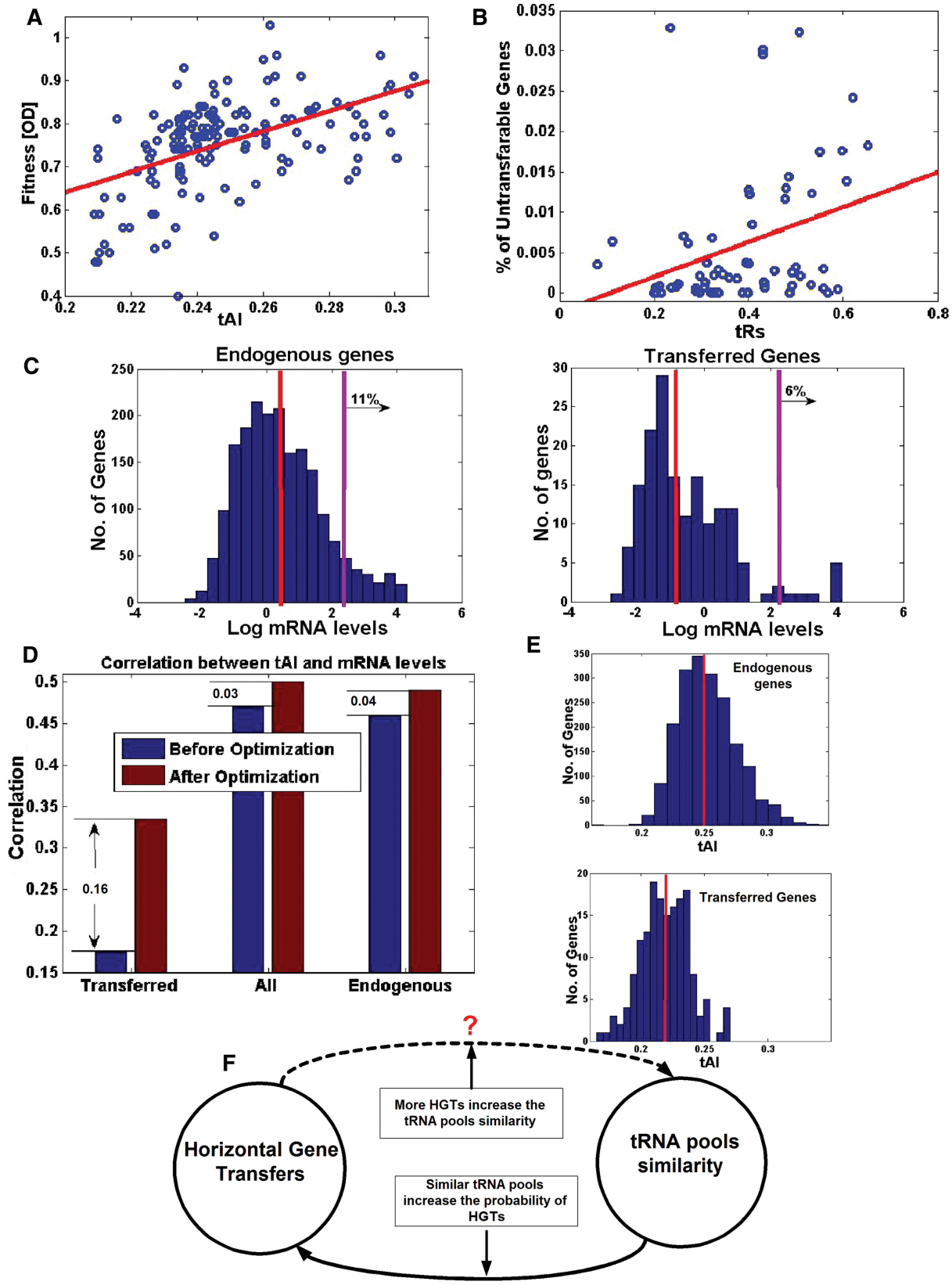


Figure 4. (**A**) Correlation between fitness (growth rate in OD) and the tAI, across a set of GFPs with different codon bias that have been expressed in *E. coli* [based on data from (7)]. (**B**) HGT events from different source organisms to *E. coli*: the percent of genes that are non-transferable to *E. coli* versus the corresponding *tRs* between the source organism and *E. coli* [based on data of Sorek *et al.* (2)]. (**C**) Gene expression of recently transferred genes and endogenous genes. The red lines denote the mean expression level in each group, the purple lines mark a threshold denoting high expression level [log(mRNA levels) >2]; 11% of endogenous genes (that did not undergo recent HGT) are highly expressed and 6% of the genes

(continued)

In all organisms that were analyzed before there is significant correlation between the expression levels of genes and translation efficiency [as measured, for example, by the tAI; see (10,19,21,22,26,37)]. One of the main reasons for this correlation is the fact that highly expressed genes potentially occupy more ribosomes. Thus, mutations that improve their translation efficiency are likely to have a higher effect on the organism's 'fitness'. As a result, these genes are under stronger selection for higher translation efficiency [see, for example, (7,10,37)]. In other words, improving the correlation between the expression level of genes and their translation efficiency (tAI) should improve the overall allocation of ribosomes and fitness of the organism. This correlation can be improved locally, by increasing the number of efficient codons in a highly expressed gene or globally, by changing the tRNA pool of the organism such that the correlation will increase. Changes in the tRNA pool may be due to duplication/deletion of a tRNA gene but also due to transfer of a tRNA(s) (possibly in the same HGT event) from organism(s) in the same environment/community.

To try and demonstrate how HGT can trigger selection for similar tRNA pools, we used *E. coli* [for which the mRNA levels for most genes, under known conditions, are available (14)] as a model organism. According to Nakamura *et al.* (6), 768 out of the 4376 *E. coli* genes underwent recent transfer. Figure 4C depicts the distribution of expression levels in *E. coli* genes, distinguishing between recently acquired genes and the rest of the genome. As can be seen, the average expression levels of recently acquired genes are lower than the mean expression level of the rest of the genes (Figure 4C). However, there is a non-negligible fraction of recently acquired genes that is highly expressed (6% of the recently acquired genes versus 11% of the rest of the genome). Hence, we expected that adjusting the tRNA pool to better fit the expression levels of these new genes, while maintaining the efficiency of the older and more established genes, should improve the fitness of the organism. A more efficient ribosome allocation should be reflected by a higher correlation between mRNA levels of genes and their tAI. In addition, it is important to note that even a 'small' change in this correlation may have a substantial effect on the fitness of the organism.

Quite surprisingly, in accordance with this hypothesis, the correlation between the expression levels and tAI is indeed higher when considering 'all' the genes ($r = 0.47$; $P < 10^{-16}$, $pe < 0.001$) than when considering 'only' the non-recently transferred genes ($r = 0.45$; $P < 10^{-16}$, $pe < 0.001$; Figure 4D; see Figure 4E for the distribution of *tAI* in the non-transferred genes and the transferred genes). This suggests that the tRNA pool of *E. coli* has underwent an adaptation to the new HT genes. Further support for this finding comes by observing that the correlation between the expression levels and tAI of all the genes (recently transferred and non-recently transferred) is significantly higher than when replacing the transferred genes with random groups of genes of similar size (whose codon biased underwent selection to fit their expression levels—controlling for the fact that the adaptation can be at the level of codon bias or that the HT genes have more compatible codon bias; see the previous subsection), or by randomly permuting the mRNA values of only the transferred genes (empirical P-value = 0.01 and 0.03, respectively). These results remain significant also when we sampled random groups with similar mean mRNA levels as the transferred genes (controlling for the fact that the recently acquired genes generally have lower mRNA levels; empirical P-value = 0.01).

In addition, we computed the tRNA pool that optimizes ('Materials and Methods' section) this correlation when considering all the genes and when considering only genes that 'did not' undergo recent HGT. These correlations were compared to the correlation obtained when using the actual tRNA pool. Again, the 'optimal' correlation was 33% 'closer' to the actual one obtained when considering 'all' the genes (Figure 4D).

These results may suggest that the tRNA pool of *E. coli* was shaped by the expression levels and codon bias of the transferred genes and not only its ancestral genes. More generally, these results hint that in practice there may be a sufficient level of HGT to trigger selection for changes in an organism's tRNA pool—i.e. such changes that make it more similar to the tRNA pools of its partners for gene exchange (see also Supplementary Note S5). The fact that we have observed this findings in recently transferred genes that have not had time to ameliorate, may suggest that such changes in the tRNA pool are relatively rapid. Specifically, these changes are faster than the time required for full amelioration, which was estimated to be around 300 million years (44).

DISCUSSION

This study shows that there is a bi-directional association between translation efficiency and HGT: genes tend to be transferred between organisms that have similar tRNA pools (tRs) and frequent HGT between organisms can in turn homogenize their tRNA pools. Likewise, genes whose tAI is similar among many organisms tend to be transferred more frequently. The fact that the relation between the similarity in tRNA pools and HGT was observed in all the analyzed data sets demonstrates the robustness of this relation.

Figure 4. Continued
that did undergo recent HGT are highly expressed. (**D**) The correlation between tAI and gene expression for the transferred genes, endogenous genes, and for all genes (blue) versus the correlation that is gained in each of these cases after optimizing the tRNA pool (brown); the optimal correlation when considering all the gene is closer to the actual one (a 'difference' of 0.03 between the optimal and actual correlations) than the correlation when considering only the non-transferred genes (a 'difference' of 0.04 between the optimal and actual correlations, 33% higher). (**E**) The distribution of tAI for endogenous genes (upper part: mean tAI is 0.25—the red line) and recently transferred genes (lower part: mean tAI is 0.22—the red line). (**F**) A schematic illustration of the possible bidirectional relation between HGT and similarity in the tRNA pools.

We show that this relation remains significant after controlling for many other possible variables (e.g. GC content, amino acid usage, phylogenetic distance and more). It is important to remember, however, that it is impossible to completely tease apart some of the variables mentioned above, as they are inherently inter-dependent. For example, similarity in tRNA pool (and thus codon composition) will be reflected in a similar GC content and amino acid usage; similarly, phylogenetic proximity can increase the number of HGT events due to similarity in the tRNA pools.

Thus, the correlations that were obtained after controlling for these variables (usually around 0.1–0.2) are only a 'lower bound' on the actual effect of translation efficiency on gene transfer. The fact that these correlations are significant, demonstrates that there is 'a distinct' effect of translation efficiency on HGT.

The correlation obtained without the controls (more than 0.4), on the other hand, represents an 'upper bound' for the effect of translation efficiency on HGT, suggesting overall that the actual association between HGT and compatibility of the tRNA pools is very substantial.

In summary, based on the results presented in this article, we suggest that when the tRNA pool of the donor organism is more similar to that of the recipient, this in turn increases the chance of successful HGT events (i.e. there is a higher probability that transferred genes will be fixed). Another possibility that is consistent with the data, is that when an organism receives genes from other organisms, most of them in its community/environment, its tRNA pool can also undergo selection to fit the new genes, especially those that are highly transcribed, thus improving the fitness of the organism. This scenario would be more likely when multiple genes are acquired from a single source together, as in the acquisition of a large plasmid. Thus, this alternative mechanism will also cause the tRNA pool of the organism to become more similar to other organisms in the community (Figure 4F). One can therefore speculate that a positive-feedback loop will exist, accounting for the increasing similarity in the tRNA pool of organisms in the same community/niche due to HGT, in turn promoting HGT between organisms in the same niche (Figure 4F).

Furthermore, this scenario is supported by reasonable population genetic considerations such as the effective size of bacteria, the fitness advantage of such changes in the tRNA pool, or the fitness disadvantage of receiving a gene with maladaptive codons (see details in Supplementary Note S5). These results encourage further studies in this direction, for example, by performing *in vitro* evolution experiments where bacteria receive a plasmid containing highly expressed genes and their tRNA pools are examined by genome re-sequencing every 1000 generations.

The results presented in this paper suggest that methods that detect (recent) HGT events based on difference in the codon bias of gene acceptor and the gene donor [see, for example, (6,45)] underestimate the number of horizontally transferred genes. Such methods search for genes whose codon bias is different from that of the host; however, as we report here, the donors of many of the horizontally transferred genes have tRNA pools (and similar codon bias) that are similar to the tRNA pools of the acceptors, making HGT detection difficult. This work therefore supports the conclusions of Medrano-Soto *et al.* (9), and stresses the importance of relying on phylogenetic tree reconstruction data as much as possible when detecting HGT, or when this is impossible, applying multiple HGT detection methods, as previously suggested (46). This is especially true when aiming to infer HGT between closely related organisms (with similar tRNA pool) that are either phylogenetically related or live in the same niche. Additionally, since a higher compatibility in tRNA pools facilitates gene transfer, this contributes to a higher level of gene exchange between related organisms. Thus, the apparent vertical phylogenetic signal one often associates with a tree of life can in fact be maintained by preferential bias of gene transfer among related taxa, as previously suggested by Gogarten *et al.* (47) for homologous recombination.

Finally, our results imply that considering that more HGTs are expected among organisms that have similar tRNA pools (and other cellular features), one can improve the current algorithms for phylogenetic network reconstruction (11,15,48,49) and for inferring ecological niches [see, for example, (12)], similarly to the way in which information about co-evolution (25,50) has been recently shown to improve ancestral gene reconstruction (51).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Tal Dagan and Yoji Nakamura for providing us with datasets of horizontal gene transfer. We also would like to thank Hila Gingold, Johann Peter Gogarten, Amos Tanay, Elchanan Mossel, and Yitzhak Pilpel for very helpful and useful discussions. We wish to assert here that UG and ER have equally contributed to the paper.

FUNDING

T.T. is a Koshland Scholar at Weizmann Institute of Science and is supported by a travel fellowship from EU grant PIRG04-GA-2008-239317. M.K. was supported by grants from the Israel Science Foundation and the German-Israel Binational Fund (GIF). U.G. is supported by a grant from the Israeli Ministry of Health. E.R. is supported by grants from the Israel Science Foundation. M.K., U.G. and E.R. are supported by a grant from the McDonnell Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Dagan,T., Artzy-Randrup,Y. and Martin,W. (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl Acad. Sci. USA*, **105**, 10039–10044.

2. Sorek,R., Zhu,Y., Creevey,C.J., Francino,M.P., Bork,P. and Rubin,E.M. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, **318**, 1449–1452.
3. Doolittle,W.F. and Bapteste,E. (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl Acad. Sci. USA*, **104**, 2043–2049.
4. Thomas,C.M. and Nielsen,K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.*, **3**, 711–721.
5. Wellner,A., Lurie,M.N. and Gophna,U. (2007) Complexity, connectivity, and duplicability as barriers to lateral gene transfer. *Genome Biol.*, **8**, R156.
6. Nakamura,Y., Itoh,T., Matsuda,H. and Gojobori,T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.*, **36**, 760–766.
7. Kudla,G., Murray,A.W., Tollervey,D. and Plotkin,J.B. (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science*, **324**, 255–258.
8. Bahir,I., Fromer,M., Prat,Y. and Linial,M. (2009) Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.*, **5**, 311.
9. Medrano-Soto,A., Moreno-Hagelsieb,G., Vinuesa,P., Christen,J.A. and Collado-Vides,J. (2004) Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Mol. Biol. Evol.*, **21**, 1884–1894.
10. Tuller,T., Waldman,Y.Y., Kupiec,M. and Ruppin,E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl Acad. Sci. USA*, **107**, 3645–3650.
11. Beiko,R.G., Harlow,T.J. and Ragan,M.A. (2005) Highways of gene sharing in prokaryotes. *Proc. Natl Acad. Sci. USA*, **102**, 14332–14337.
12. Freilich,S., Kreimer,A., Meilijson,I., Gophna,U., Sharan,R. and Ruppin,E. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res.*, **38**, 3857–3868.
13. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
14. Corbin,R.W., Paliy,O., Yang,F., Shabanowitz,J., Platt,M., Lyons,C.E. Jr, Root,K., McAuliffe,J., Jordan,M.I., Kustu,S. *et al.* (2003) Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc. Natl Acad. Sci. USA*, **100**, 9232–9237.
15. Dagan,T. and Martin,W. (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl Acad. Sci. USA*, **104**, 870–875.
16. Neyman,J. (1971) Molecular studies of evolution: a source of novel statistical problems. In Gupta,S. and Jackel,Y. (eds), *Statistical Decision Theory and Related Topics*. Academic Press, New York, pp. 1–27.
17. Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–123.
18. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. BioSci.*, **13**, 555–556.
19. dos Reis,M., Savva,R. and Wernisch,L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, **32**, 5036–5044.
20. Vieira-Silva,S. and Rocha,E.P. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.*, **6**, e1000808.
21. Tuller,T., Kupiec,M. and Ruppin,E. (2007) Determinants of protein abundance and translation efficiency in *S. cerevisiae*. *PLoS Comput. Biol.*, **3**, e248.
22. Man,O. and Pilpel,Y. (2007) Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet.*, **39**, 415–421.
23. Drummond,D.A., Bloom,J.D., Adami,C., Wilke,C.O. and Arnold,F.H. (2005) Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA*, **102**, 14338–14343.
24. Newman,M.E. (2006) Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA*, **103**, 8577–8582.
25. Tuller,T., Kupiec,M. and Ruppin,E. (2009) Co-evolutionary networks of genes and cellular processes across fungal species. *Genome Biol.*, **10**, R48.
26. Waldman,Y.Y., Tuller,T., Shlomi,T., Sharan,R. and Ruppin,E. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Res.*, **38**, 2964–2974.
27. Percudani,R., Pavesi,A. and Ottonello,S. (1997) Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J. Mol. Biol.*, **268**, 322–330.
28. Kanaya,S., Yamada,Y., Kudo,Y. and Ikemura,T. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, **238**, 143–155.
29. Ikemura,T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, **151**, 389–409.
30. Dong,H., Nilsson,L. and Kurland,C.G. (1996) Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J. Mol. Biol.*, **260**, 649–663.
31. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
32. Ochman,H. and Lawrence,J.G. (1996) In Neidhardt,F.C. (ed.), *Escherichia coli and Salmonella typhimurium: Molecular and Cellular Biology*, 2nd edn. ASM Publications, Washington.
33. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
34. Kobayashi,I. (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.*, **29**, 3742–3756.
35. Stern,A., Mayrose,I., Penn,O., Shaul,S., Gophna,U. and Pupko,T. An evolutionary analysis of lateral gene transfer in thymidylate synthase enzymes. *Syst. Biol.*, **59**, 212–225.
36. Jain,R., Rivera,M.C. and Lake,J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA*, **96**, 3801–3806.
37. Tuller,T., Carmi,A., Vestsigian,K., Navon,S., Dorfan,Y., Zaborske,J., Pan,T., Dahan,O., Furman,I. and Pilpel,Y. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–354.
38. Foerstner,K.U., von Mering,C., Hooper,S.D. and Bork,P. (2005) Environments shape the nucleotide composition of genomes. *EMBO Rep.*, **6**, 1208–1213.
39. Lynch,M. (2007) *The Origins of Genome Architecture*. Sinauer Associates.
40. Warren,R.L., Freeman,J.D., Levesque,R.C., Smailus,D.E., Flibotte,S. and Holt,R.A. (2008) Transcription of foreign DNA in Escherichia coli. *Genome Res.*, **18**, 1798–1805.
41. Taoka,M., Yamauchi,Y., Shinkawa,T., Kaji,H., Motohashi,W., Nakayama,H., Takahashi,N. and Isobe,T. (2004) Only a small subset of the horizontally transferred chromosomal genes in Escherichia coli are translated into proteins. *Mol. Cell. Proteomics*, **3**, 780–787.
42. Wellner,A. and Gophna,U. (2008) Neutrality of foreign complex subunits in an experimental model of lateral gene transfer. *Mol. Biol. Evol.*, **25**, 1835–1840.
43. Novozhilov,A.S., Karev,G.P. and Koonin,E.V. (2005) Mathematical modeling of evolution of horizontally transferred genes. *Mol. Biol. Evol.*, **22**, 1721–1732.
44. Lawrence,J.G. and Ochman,H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, **44**, 383–397.
45. Garcia-Vallve,S., Romeu,A. and Palau,J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.*, **10**, 1719–1725.

46. Ragan,M.A., Harlow,T.J. and Beiko,R.G. (2006) Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends Microbiol.*, **14**, 4–8.
47. Gogarten,J.P., Doolittle,W.F. and Lawrence,J.G. (2002) Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.*, **19**, 2226–2238.
48. Ball,C.A., Awad,I.A., Demeter,J., Gollub,J., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Matese,J.C., Nitzberg,M., Wymore,F. *et al.* (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.*, **33**, D580–D582.
49. Birin,H., Gal-Or,Z., Elias,I. and Tuller,T. (2008) Inferring horizontal transfers in the presence of rearrangements by the minimum evolution criterion. *Bioinformatics*, **24**, 826–832.
50. Barker,D. and Pagel,M. (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput. Biol.*, **1**, e3.
51. Tuller,T., Birin,H., Gophna,U., Kupiec,M. and Ruppin,E. (2009) Reconstructing ancestral gene content by coevolution. *Genome Res.*, **20**, 122–132.