Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome

MIKE WITHERS, LORENZ WERNISCH, and MARIO DOS REIS

School of Crystallography, Birkbeck College, University of London, London WC1E 7HX, United Kingdom

ABSTRACT

Transfer RNA genes tend to be presented in multiple copies in the genomes of most organisms, from bacteria to eukaryotes. The evolution and genomic structure of tRNA genes has been a somewhat neglected area of molecular evolution. *Escherichia coli*, the first phylogenetic species for which more than two different strains have been sequenced, provides an invaluable framework to study the evolution of tRNA genes. In this work, a detailed analysis of the tRNA structure of the genomes of *Escherichia coli* strains K12, CFT073, and O157:H7, *Shigella flexneri* 2a 301, and *Salmonella typhimurium* LT2 was carried out. A phylogenetic analysis of these organisms was completed, and an archaeological map depicting the main events in the evolution of tRNA genes was drawn. It is shown that duplications, deletions, and horizontal gene transfers are the main factors driving tRNA evolution in these genomes. On average, 0.64 tRNA insertions/duplications occur every million years (Myr) per genome per lineage, while deletions occur at the slower rate of 0.30 per million years per genome per lineage. This work provides a first genomic glance at the problem of tRNA evolution as a repetitive process, and the relationship of this mechanism to genome evolution and codon usage is discussed.

Keywords: tRNA; *Escherichia coli*; codon usage; selfish gene

INTRODUCTION

Transfer RNAs are present in the genomes of all living organisms. They act as the adaptors that align amino acids on mRNA codons, allowing the formation of proteins as specified by the corresponding gene sequence. tRNAs bridge the gap between genetic information and protein expression. tRNA genes are normally present in multiple copies in most genomes, from prokaryotes to eukaryotes. These copies might be arranged as closely spaced clusters, as interspersed individual copies, or, in the case of bacteria, in polycistronic operons with other tRNA, rRNA, or protein genes (Inokuchi and Yamao 1995). Interestingly, tRNA gene species are represented unequally in most genomes, with some species being present in multiple gene copies, while others are present in single copies (Marck and Grosjean 2002). The factors that determine the distribution and number of copies of tRNAs within genomes seem to have been scarcely studied and focused on very particular

examples (e.g., Hosbach et al. 1980; Gonos and Goddard 1990).

It has been suggested that cellular tRNA content might be rate-limiting during protein translation (Kurland 1993). For example, in certain organisms like *Escherichia coli* or the baker's yeast, it has been observed that highly expressed genes use a subset of optimal codons that match the most abundant tRNAs within the cell (Ikemura 1981a; Bennetzen and Hall 1982). Furthermore, it has also been noticed that in these organisms amino acid usage within proteins is also correlated with the abundance of the respective aminoacyl-tRNAs (Yamao et al. 1991; Lobry and Gautier 1994). Since tRNA abundance and tRNA gene copy number are highly correlated (e.g., Percudani et al. 1997; Kanaya et al. 1999; Duret 2000), it is possible to think that the factors that determine the genomic structure of tRNA genes are also related to the evolution of codon and amino acid usage within genomes and ultimately with the evolution and regulation of protein expression. Despite the importance of tRNAs in protein translation, scant attention has been paid to the mechanisms of tRNA evolution, replication, and propagation within genomes. Understanding what evolutionary forces shape genomic tRNA structure is important in order to understand the fine details of protein expression. The advances of the post-genomic era and the improvements

Reprint requests to: Mario dos Reis, School of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX UK; e-mail: m.reis@mail.cryst.bbk.ac.uk; fax: +44-20-7631-6803.

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.2272306>.

TABLE 1. Genetic distances (K2P) for the genomes analyzed in this study

	<i>EcK12</i>	<i>EcO157</i>	<i>Sf2a301</i>	<i>EcCFT073</i>
<i>EcO157</i>	0.0280			
<i>Sf2a301</i>	0.0307	0.0322		
<i>EcCFT073</i>	0.0433	0.0444	0.0477	
<i>StLT2</i>	0.2452	0.2444	0.2433	0.2417

in the techniques for the detection of tRNA genes in diverse genomes can now be used to understand tRNA evolution in more detail.

In this work, the bacterium *E. coli* is used as a model organism to study tRNA evolution. Different strains of *E. coli* have already been sequenced (Blattner et al. 1997; Perna et al. 2001; Welch et al. 2002), including strains of the enteropathogenic *Shigella flexneri* (Jin et al. 2002), which has long been known to be a form of *E. coli* (Pupo et al. 2000). The availability of the genome sequence of these closely related organisms allows easy identification of sets of orthologous tRNA genes, their rate of divergence, and the identification of duplication, insertion, deletion, and inversion events during the evolution of tRNA genes. The analysis of the evolutionary dynamics of the genomic tRNA pool in these organisms can shed light onto the more general problem of tRNA evolution in other organisms and its relationship to protein expression. We aim to provide a first genomic picture of the evolution of tRNA genes as a repetitive process and its implications for the evolution of codon usage and ultimately gene expression.

RESULTS

The genomes studied in this work were those of *E. coli* strains K12, O157:H7 EDL933, and CFT073; *S. flexneri* 2a 301; and *Salmonella typhimurium* LT2 (used as an out-group). These organisms will be referred to here as *EcK12*, *EcO157*, *EcCFT073*, *Sf2a301*, and *StLT2*, respectively.

Phylogenetic analysis

The divergence time between the *Escherichia* and *Salmonella* lineages has been estimated to be ~100 million years (Myr) (Ochman and Wilson 1987; Doolittle et al. 1996). The average proportion of nucleotide substitutions between the *Escherichia*–*Shigella* group and *Salmonella* estimated for the whole genome in this work is 0.2452, with a standard deviation of only 0.0015, suggesting that these organisms have been diverging at a fairly constant rate. The genetic distances were calculated using Kimura's two-parameter model (K2P) (Kimura 1980; Table 1). Assuming a constant molecular clock, and utilizing the estimated genetic distances, an UPGMA tree was built that shows the estimated divergence time for the genomes analyzed (Fig. 1). Strains

FIGURE 1. UPGMA tree for the five genomes analyzed in this study. Divergence times are given in million years (Myr). The black circle indicates the point of the ancestral tRNA reconstruction.

EcK12, *EcO157*, and *Sf2a301* diverged roughly 12–13 million years ago (Mya) in a quasi-star phylogeny. This group diverged from *EcCFT073* 19 Mya. The close relationship of *Sf2a301* to the *Escherichia* strains agrees with previous studies (Pupo et al. 2000) that identify *Shigella* as a form of *E. coli*; however, our findings strongly disagree with previously estimated divergence times. We consider the implications of these findings in the discussion.

tRNA species and gene copy number

The genomes analyzed present between 85 and 99 tRNA genes. These genes comprise between 45 and 52 different tRNA species representing between 40 and 41 anticodons (Table 2). In this work, two tRNA genes are considered to be of the same species if they present the same anticodon and if they present no more than 0.10 substituted sites (K2P) in pairwise comparisons. Most tRNA gene species are presented in one copy within the chromosome, with fewer presented in two or more copies, with up to seven copies for the most abundant tRNA species in *EcO157*. Accordingly, the distribution of tRNA species versus gene copy number is long-tailed (Fig. 2).

Evolution of tRNA genes

We identified 78 tRNA genes in *EcK12* that have unambiguous orthologs in *StLT2*. Interestingly, these genes

TABLE 2. Number of tRNA genes, tRNA species, and anticodons present in each genome analyzed

	tRNA genes	tRNA species	Anticodons
<i>EcK12</i>	86	45	40
<i>EcO157</i>	99	52	41
<i>Sf2a301</i>	97	52	40
<i>EcCFT073</i>	88	48	41
<i>StLT2</i>	85	47	40

FIGURE 2. Distribution of number of tRNA species vs. gene copy number for each genome analyzed.

are highly conserved despite the fact that both organisms diverged ~ 100 Mya. Very few nucleotide substitutions were observed when the 78 genes in *EcK12* were aligned to their corresponding *StLT2* orthologs, implying a low substitution rate (Table 3). This rate contrasts sharply with the overall genomic substitution rate (Table 3). Similar results are obtained when a full comparison among all the strains is carried out (see supplemental material at <http://people.crysl.bbk.ac.uk/~fdosr01/trnas/>). Figure 3 depicts a detailed syntenic map showing the identified set of orthologous tRNAs across all of the strains analyzed. In general, a well-conserved tRNA backbone, which reflects the tRNA configuration of the last common ancestor of the *Escherichia-Shigella* clade, is present in all strains. With the exception of tRNA-rRNA operons, polycistronic tRNAs (Fig. 3) are well conserved across all strains including *SfLT2*, suggesting that they were already present in their modern configuration more than 100 Mya in the *Escherichia-Salmonella* ancestor.

Although tRNA genes are well conserved, their position and number of copies within the chromosome is not. We

identified 89 evolutionary events along the tree branches of the *Escherichia-Shigella* clade, which include 49 insertions, 23 deletions, three inversions and one translocation (Table 4). The most active genome seems to be *Sf2a301*, where the highest number of evolutionary events was detected, followed by that of *EcO157*. These two genomes contain large numbers of insertions that are probably related to horizontal transfer events. *EcK12* and *EcCFT073* presented the best conserved tRNA sets.

The approach applied in this work does not allow the identification of tRNA evolutionary events in *StLT2*, since this organism was used as an outgroup and ancestral character reconstruction at the most ancient node in the tree is uncertain (Cunningham et al. 1998). tRNA gene *gltT* can illustrate this point. This gene is present in *EcK12*, *EcO157*, *EcCFT073*, and *Sf2a301*, but it is absent in *StLT2* (Fig. 3). Two scenarios are possible: (1) *gltT* was not present in the *Escherichia-Salmonella* ancestor, and it was inserted in the genome of the *Escherichia* ancestor after divergence from *StLT2*; (2) *gltT* was indeed present in the last *Escherichia-Shigella* ancestor, but it was deleted somewhere in the *Salmonella* lineage. Both explanations are equally parsimonious, because both involve only one evolutionary event. With current data it is impossible to distinguish the most likely explanation. However, the resemblance of *StLT2* to the ancestral reconstruction suggests that its tRNA set has remained reasonably conserved.

Pseudo tRNAs

We identified nine putative pseudo-tRNAs in the *Escherichia-Shigella* clade (Table 5; see supplemental material). These genes can be distinguished by their unusually low-covariance model score (COVE score) (for details, see Materials and

TABLE 3. Substitution rates between *E. coli* and *S. typhimurium*

Type	Rate (per site per year)
Overall genome	1.2×10^{-9}
tRNA genes	
Functional	0.014×10^{-9}
Pseudo	3.8×10^{-9}
Protein-coding genes	
Synonymous	4.7×10^{-9} ^a
Nonsynonymous	0.22×10^{-9} ^a

^aAfter Sharp (1991).

The inclusion of other *Salmonella* strains into the phylogenetic tree might be needed in order to solve the issue.

Adaptation of codon usage to the ancestral tRNA set

It has previously been suggested that codon usage and tRNA genes coevolve in feedback fashion (Bulmer 1987): The most abundant tRNAs drive up the frequencies of their cognate codons, and certain codons already present in high frequencies can in turn drive up the intracellular levels of their respective cognate tRNAs. Since tRNA expression levels and tRNA gene copy number are correlated (Ikemura 1981a,b; Bennetzen and Hall 1982; Kurland 1993; Kanaya et al. 1999), we decided to investigate whether the variation in gene copy number observed for some tRNA genes along the lineages analyzed has had an effect on the codon usage of the genomes involved. In order to achieve this goal, the S-values (dos Reis et al. 2004) for the coadaptation between genomic tRNAs and codon usage was computed for each genome, and the process was carried out using modern and ancestral tRNA sets (see Materials and Methods). Table 6 shows the estimated S-values for the members of the *Escherichia-Shigella* clade. It can be seen that, with the exception of *EcK12*, the S-values estimated from the ancestral reconstruction are higher than those obtained from modern tRNAs. This indicates that the codon usage of the modern genomes is better adapted to the ancestral tRNA set than to the modern set.

DISCUSSION

Phylogenetic analysis

The phylogenetic analysis carried out in this work produced some interesting results. The tree topology obtained agrees with previous studies (Pupo et al. 2000; Reid et al. 2000), but the branch lengths estimated here disagree substantially with these reports. Pupo et al. (2000) suggested a divergence time of between 35,000 and 270,000 yr for the evolution of the main *Shigella* groups, but we find this estimate too low. Reid et al. (2000) suggested a divergence time for *EcK12* and *EcO157* of only 4.5 Myr, which contrasts with the 12 Myr estimated here. Both studies used previous estimates of synonymous substitution rates between *E. coli* and *S. typhimurium* (Guttman and Dykhuizen 1994; Whittam 1996) in order to calculate divergence

times. The average proportion of synonymous substituted sites between these organisms is very high ($K_s = 0.94$) (Sharp 1991), and clearly saturated for many genes, making the estimation of substitution rates unreliable. Furthermore, both studies based their analysis on housekeeping genes that present substantially biased codon usage and low synonymous substitution rates, probably leading to subestimation of divergence times. We consider our estimated branch lengths accurate, since the phylogenetic tree was built using whole-genome alignments. Trees built from a large collection of sequences have been shown to be very robust (Rokas et al. 2003). Furthermore, we used an overall estimate of the proportion of substituted sites between *Escherichia* clade and *Salmonella*, which is based on a whole-genome alignment with a value of 24.5%, far from saturation, and hence maintaining maximum phylogenetic signal.

The genomes analyzed present a chromosomal tRNA backbone that has been conserved for ~100 Myr

Considering the results obtained in this work, tRNA evolution can be described on three levels: (1) evolution of tRNA sequences themselves, (2) evolution of polycistronic tRNAs and the tRNA backbone, and (3) evolution of interspersed tRNA genes. The first level relates to the most fundamental mode of evolution, with the slow accumulation of nucleotide substitutions, slow divergence, and reflecting the action of natural selection on tRNA structure. The other two levels are related to the genomic organization of tRNA genes and probably reflect broader patterns of chromosome evolution in bacteria, like recombination and horizontal gene transfer.

Transfer RNA sequences evolve slowly, with tRNA orthologs in *Escherichia* spp and *Shigella* spp being nearly identical despite the fact that both groups diverged >100 Mya (Bachelier et al. 1996). We estimate that the substitution rate in functional tRNA genes is 0.012 times the average genomic rate, 0.03 times the rate for synonymous sites, and 0.065 times the average for nonsynonymous sites in protein-coding genes (Table 3). The genomic rate is an average that includes substitutions in coding and non-coding regions, and it reflects a balance between regions under strong selection and regions that accumulate substitutions freely. The genomic rate is hence a good baseline to which comparisons can be made. Synonymous substitutions are those that happen inside protein-coding genes but that do not change the encoded amino acid; these substitutions are relatively free from selective pressures. On the other hand, nonsynonymous mutations do change the encoded amino acid and hence are usually subjected to strong purifying selection (Sharp 1991). The fact that tRNA genes show such low

FIGURE 3. Orthologous tRNA gene sets for the five organisms analyzed. Blue and yellow boxes show polycistronic tRNA clusters in *EcK12*, and gray boxes show mixed rRNA-tRNA operons (Inokuchi and Yamao 1995). Purple boxes show inverted regions when compared with the ancestral reconstruction. Orange boxes show regions with a high number of insertions. The numbers on the left indicate the genomic coordinates of tRNA genes in *EcK12*. The numbers at the bottom indicate the total number of functional tRNA genes in each genome.

TABLE 4. Estimated number of tRNA gene evolutionary events along the *E. coli* clade

	<i>EcK12</i>	<i>EcO157</i>	<i>Sf2a301</i>	<i>EcCFT073</i>	Per lineage	Per lineage-Myr
Insertions/duplications	3	18	24	4	12.25	0.64
Deletions	4	5	12	2	5.75	0.30
Inversions	0	0	2	1	0.75	0.04
Translocations	0	0	1	0	0.25	0.01

substitution rates compared with nonsynonymous substitutions indicates that strong purifying selection acts on them in order to conserve their structure and identity (McClain 1995). In contrast, high rates of nucleotide substitution are observed during the death and generation of tRNA pseudogenes. We estimate that substitution rates in pseudogenes are more than three times the average genomic rate and more than 250 times that seen in functional tRNAs. This reflects the lack of function of these sequences, and their substitution rates probably reflect the underlying genomic mutational rate (Kimura 1983).

The ancestral tRNA set contained 86 tRNA genes. Seventy-eight genes from this set are still present and have remained nearly intact in the lineages leading to *EcK12* and *StLT2*. These 78 genes reflect a conserved tRNA backbone that has remained well conserved for >100 Myr and can be easily distinguished in all of the genomes analyzed. It is clear that strong selective pressure must be operating to maintain such a high degree of conservation after 100 Myr of divergence. It should be noticed that the conserved backbone is made up mostly of polycistronic tRNAs and mixed (tRNA and protein) operons (Inokuchi and Yamao 1995). How this set originated is an interesting question that needs to be addressed. A cluster-dispersion model of tRNA evolution has been proposed to account for the generation of new tRNA species by duplication and coadaptation to the genetic code in the ancient lineages of life (Xue et al. 2003). An extension of this model could perhaps be used to understand the modern tRNA structure of eubacteria; however, a more extensive analysis, including more distantly related bacteria, is needed in order to work out ancient evolutionary events that led to the formation of the modern tRNA clusters seen in *Escherichia* spp. and *Salmonella* spp.

Most evolutionary events observed in the genomes analyzed have involved interspersed tRNA genes. The genomes with the largest amounts of horizontally transferred DNA (*EcO157* and *Sf2a301*) showed the largest number of tRNA insertions. It has been suggested that tRNAs that code for codons that are rare in the bacterium chromosome may be needed to express some of the foreign protein genes (Hayasi et al. 2001). By comparison, *EcK12*, which shows the smallest genome and the best-conserved tRNA set, also presented the smallest number of tRNA-related events.

Evolution of tRNA genes as a repetitive process

Transfer RNA genes can be considered as a special type of repeated DNA (Bachellier et al. 1996), and their evolution has been previously labeled as a repetitive process (Jukes and Holmquist 1972). A lot of research has been carried out on the mechanisms of replication and propagation of various types of repeated sequences in prokaryotes (Bachellier et al. 1996), such as insertion sequences and transposons, but regrettably it seems that nearly no work has been done to understand how the evolutionary dynamics of repeated DNA can be used to understand the evolution, propagation, and maintenance of tRNA copies within genomes. An important question arises: What underlying mechanisms generate the distribution of tRNA gene copies seen in Figure 2? Do tRNA genes present in multiple copies convey a selective advantage to their host or are they just the product of selfish DNA like mechanisms of propagation? Interestingly, genome size and total genome tRNA gene number are highly correlated in both prokaryotes and eukaryotes (dos Reis et al. 2004), so perhaps the mechanisms that explain genome size evolution might be also taken into account to understand tRNA propagation within genomes. The study of genome size evolution has been very controversial (Petrov 2001), not least because of the iconoclastic ideas behind the selfish DNA hypothesis (Charlesworth et al. 1994). This hypothesis assumes that repetitive DNA is maintained and spread throughout genomes due to its inherent replicative properties and does not necessarily confer any selective advantage to its carriers (Orgel and Crick 1980). If tRNA genes can be considered to some extent to represent selfish genes, then the tRNA configuration of modern *Escherichia* strains simply reflects a series of selfish propagative events. From a selectionist point of view, it is interesting to notice that bacteria with small genomes tend to present a minimal set of tRNA genes that is nonredundant and that can adequately translate all codons (e.g., Muto et al. 1990). It is plausible that a small-sized bacterial ancestor suffered a series of genome expansions (e.g., through genome duplication, recombination, horizontal gene transfer, etc.) that led to an increased and redundant set of tRNA genes. The fact that many of the tRNA genes analyzed are presented as interspersed repeats associated to horizontal transfer events suggests that selfish-like mechanisms of propagation are operative. Furthermore, many of the well-conserved polycistronic tRNA

TABLE 5. Putative pseudo-genes identified with tRNAscan-SE

Genome	Type	Start ^a	End ^a	COVE	K2P ^b	Paralog
<i>EcK12</i>	Thr	296402	296478	42.36	—	<i>thrW</i>
	Undet	344558	344638	28.45	—	— ^c
	Arg	585242	585324	9.07	0.28	<i>argU</i>
<i>EcO157</i>	Thr	310574	310644	6.94	0.21	<i>thrW</i>
	Undet	402035	402115	31.25	0.01	— ^c
<i>Sf2a301</i>	Thr	328165	328094	10.56	0.27	<i>thrW</i>
	Thr	317706	317620	5.44	0.23	<i>thrW</i>
<i>EcCFT073</i>	Undet	434982	435062	28.45	0	— ^c
	Thr	1211231	1211306	14.13	0.20	<i>thrU</i>
<i>StLT2</i>	Arg	617984	618056	20.46	0.19	<i>argU</i>

^aStart and end of the genomic coordinates for each pseudogene.^bThe K2P distances were computed for each pseudogene aligned to their respective *EcK12* paralog.^cThese three RNAs are considered orthologous.

operons are structured as tandem arrays of repeated tRNA genes where recombination has been observed (Bachellier et al. 1996). Expansions (or reductions) of these repeats probably originated by unequal crossing over, a mechanism that accounts for the generation of large quantities of repetitive DNA. It is possible that the fundamental genomic tRNA backbone is now maintained by selection, but the highly dynamic nature of interspersed tRNAs simply reflects completely selfish processes.

Coevolution of transfer RNA genes and codon usage

Bulmer (1987) developed an elegant mathematical model that predicts that preferred codon frequencies in highly expressed genes coevolve with tRNA abundance within the cell. A critical point of this model is that mutations that produce small changes in the expression level of some

tRNA genes might be positively selected. In this case, the mutant tRNA presents a modified relative expression level that matches more closely the frequency of the respective codon. Although developed >18 yr ago, this model does not seem to have empirical confirmation yet.

tRNA expression levels within the cell are regulated by promoter sequences and tRNA gene copy number (Ikemura 1981a; Kurland 1993; Inokuchi and Yamao 1995; Kanaya et al. 1999). tRNA sequences presented in multiple copies within the bacterial chromosome tend to be expressed at higher levels than sequences with single copies. The number of copies of each tRNA species has been shown to be linearly and positively correlated to their respective expression levels. Bulmer's model does not take into account the dramatic changes in expression levels that might be brought about with the insertion or the deletion of tRNA sequences. It is also unclear how mutations in the promoter regions of polycistronic tRNAs might affect the model, since some of these clusters contain different tRNA species, and a mutation in the promoter would affect the expression levels of all of these tRNA species simultaneously.

In this work it was found that, for the majority of genomes analyzed, present codon-usage frequencies match more closely the ancestral set of tRNA genes than the modern ones. This has important implications for the understanding of codon usage and tRNA coevolution. The large number of tRNA insertions and deletions observed during the evolutionary history of *Escherichia* spp might have had

FIGURE 4. Kernel density estimate of COVE scores for all tRNA genes identified by tRNAscan-SE in the *Escherichia-Shigella* clade.**TABLE 6.** S-values for modern and ancestral tRNA sets

	Modern	Ancestral
<i>EcK12</i>	0.703	0.682
<i>EcO157</i>	0.581	0.636
<i>Sf2a301</i>	0.602	0.655
<i>EcCFT073</i>	0.528	0.576

profound effects on the relative expression levels of the different tRNA species. It is not surprising that modern codon usage is finely tuned to the ancestral, well-conserved tRNA backbone. Our data suggest that this conserved backbone arose >100 Mya, possibly the product of a series of tRNA duplications, deletions, and rearrangements. This backbone perhaps represents a frozen accident during the evolutionary history of the *Escherichia* lineage, an accident that shaped codon usage in these organisms. It is possible that minor tuning between promoter-regulated tRNA expression and codon usage happened later on in the evolutionary history of these organisms (Inokuchi and Yamao 1995) but with tRNA gene copy number being responsible for the broader pattern of codon usage seen in *E. coli*. However, our results indicate that in the modern *E. coli* genomes, selective pressure driving codon optimization toward the tRNA set has been reduced. The reasons for this are not clear. For example, in humans there is no evidence of codon-usage optimization (Urrutia and Hurst 2001), and codon usage does not reflect the tRNA composition of the genome. Humans present 501 tRNA gene copies that are undoubtedly important for our translational system (Lander et al. 2001). Why and when the tRNA genomic system is decoupled from the codon usage is an interesting question that needs to be addressed. A detailed analysis of codon-substitution patterns in protein-coding genes and their relationship to tRNA evolution is needed in order to clarify this issue.

MATERIALS AND METHODS

Phylogenetic analysis

The genomes studied in this work and their accession numbers: *Escherichia coli* strains K12 (NC_000913), O157:H7 EDL933 (NC_002655), CFT073 (NC_004431); *Shigella flexneri* 2a 301 (NC_004337), and *Salmonella typhimurium* LT2 (NC_003197). These organisms were referred to here as *EcK12*, *EcO157*, *EcCFT073*, *Sf2a301*, and *StLT2*, respectively. The five genomes were aligned using MultiPipMaker (Schwartz et al. 2000) with *EcK12* as the reference organism. Genetic distances (estimated number of nucleotide substitutions per site) were computed from the alignment using Kimura's two-parameter model (K2P) (Kimura 1980) and used to construct the UPGMA tree. The root of the tree was confirmed using *StLT2* as an outgroup.

The genomic substitution rate was estimated from the average K2P distances for the divergence between the *E. coli* and *Salmonella* groups and divided by the previously estimated divergence time of 100 Myr (Ochman and Wilson 1987; Doolittle et al. 1996).

Analysis of tRNA genes

Transfer RNA genes were identified using tRNAscan-SE (Lowe and Eddy 1997). Two low-sensitivity programs analyze genomic sequence in order to identify candidate tRNA genes; these filtered sequences are then analyzed by a highly selective tRNA covariance model (Eddy and Durbin 1994). The assigned COVE score is used

to score the probability of a sequence being a tRNA gene. A low COVE score simply means that the sequence in question shows distant similarity to the covariance tRNA model in question; this could mean that the tRNA is nonfunctional (pseudo-tRNA), that it is a special type of tRNA (viral, organellar, etc.), or that it is another sort of regulatory RNA or nonexpressed, RNA-like sequence present in the genome. The program was run with sensitive settings adjusted to identify putative ancient tRNAs (COVE cutoff score 1). The tRNAscan-SE predictions were compared for agreement with the annotated genomes. To identify sets of orthologous tRNA genes across lineages, distance matrices were constructed to compare all tRNA genes between any two given genomes. The tRNA gene sequences were aligned using clustalW and the genetic distances computed with the K2P method. The resulting matrices (see supplemental material) gave a first approximation of the orthologous sets of genes and their evolution. Discrepancies in determining detailed orthologous relationships were solved through genomic alignments of the problematic regions (pairwise Blast), and analyzing marker genes and sequences present upstream and downstream of the tRNAs. The ancestral set of tRNA genes for the *Escherichia-Shigella* clade was determined using the list of orthologous tRNAs and the phylogenetic tree obtained previously; this set was reconstructed as to minimize the number of tRNA gene insertions, deletions, translocations, and inversions along the phylogenetic tree.

Substitution rates in functional and pseudo-tRNA genes were calculated from their respective average K2P distances and using the estimated 100 Myr divergence time between the *Escherichia* and *Salmonella* groups. For comparison, synonymous and non-synonymous substitution rates in protein-coding genes for *Escherichia-Salmonella* were calculated from previous average estimates of the proportion of synonymous and nonsynonymous substitution in these organisms (Sharp 1991).

Adaptation of modern codon sequences to the ancestral tRNA set

The tRNA adaptation index (tAI) (dos Reis et al. 2004) is a measure that quantifies how well a protein-coding sequence is adapted to the tRNA pool of a given genome. This index takes into account the codon frequencies of the gene being analyzed, the frequencies of the tRNA genes in the genome, and how well those tRNAs recognize their respective codons. If a gene presents a high tAI value, it is assumed that the codon usage for this particular gene is finely tuned to match the tRNA composition of the genome. A test has been developed based on this index to assess how well the whole set of genes of a given genome is adapted to its tRNA pool, and this test has been successfully used to detect selection on codon usage in a wide variety of organisms (tAI) (dos Reis et al. 2004; but see Sharp et al. 2005). In order to carry out this test, tAI values are calculated for all genes in a genome, and a corrected version of Wright's effective number of codons (Wright 1990) is also calculated. The correlation between both measures, the *S*-value, gives an idea of how well codon usage and tRNA usage are coadapted in a given genome. In this work, *S*-values were calculated for *EcK12*, *EcO157*, *EcCFT073*, and *Sf2a301* using their respective tRNA gene pools, and the procedure was repeated for the same organisms but using the reconstructed ancestral set of tRNA genes. This gives a picture of how well "adapted"

the codon usage of modern protein-coding sequences is to the ancestral tRNA set and also sheds light onto proposed models of tRNA codon-usage coevolution (Bulmer 1987).

ACKNOWLEDGMENTS

M.d.R. is currently being supported by the Biosciences and Biotechnology Research Council, UK.

Received October 25, 2005; accepted March 2, 2006.

REFERENCES

- Bachellier, S., Gilson, E., Hofnung, M., and Hill, C.W. 1996. Repeated sequences. In *Escherichia coli and Salmonella: Cellular and molecular biology*. (ed. F.C. Neidhardt), pp. 2708–2720. ASM Press, Washington, DC.
- Bennetzen, J.L. and Hall, B.D. 1982. Codon selection in yeast. *J. Biol. Chem.* **257**: 3026–3031.
- Blattner, F.R., Plunkett 3rd, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**: 1453–1474.
- Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. *Nature* **325**: 728–730.
- Charlesworth, B., Sniegowski, P., and Stephan, W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**: 215–220.
- Cunningham, C.W., Omland, K.E., and Oakley, T.H. 1998. Reconstructing ancestral character states: A critical reappraisal. *Trends Ecol. Evol.* **13**: 361–366.
- Doolittle, R.F., Feng, D.F., Tsang, S., Cho, G., and Little, E. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271**: 470–477.
- 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.
- Duret, L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* **16**: 287–289.
- Eddy, S.R. and Durbin, R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22**: 2079–2088.
- Gonos, E.S. and Goddard, J.P. 1990. Human tRNA Glu genes: Their copy number and organisation. *FEBS Lett.* **276**: 138–142.
- Guttman, D.S. and Dykhuizen, D.E. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**: 1380–1383.
- Hayasi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., et al. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8**: 11–22.
- Hosbach, H.A., Silberklang, M., and McCarthy, B.J. 1980. Evolution of a *D. melanogaster* glutamate tRNA gene cluster. *Cell* **21**: 169–178.
- Ikemura, T. 1981a. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**: 1–21.
- . 1981b. 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.
- Inokuchi, H. and Yamao, F. 1995. Structure and expression of prokaryotic tRNA genes. In *tRNA: Structure, biosynthesis and function* (eds. D. Söll and U. RajBhandary), pp. 17–30. ASM Press, Washington, DC.
- Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J., Yang, F., et al. 2002. Genome sequence of *Shigella flexneri* 2a: Insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.* **30**: 4432–4441.
- Jukes, T.H. and Holmquist, R. 1972. Evolution of transfer RNA molecules as a repetitive process. *Biochem. Biophys. Res. Commun.* **49**: 212–226.
- 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.
- Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**: 111–120.
- . 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, UK.
- Kurland, C.G. 1993. Major codon preferences: Theme and variations. *Biochem. Soc. Trans.* **21**: 841–846.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., and Doyle, M. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- Lobry, J.R. and Gautier, C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res.* **22**: 3174–3180.
- 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.
- Marck, C. and Grosjean, H. 2002. tRNomics. Analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA* **8**: 1189–1232.
- McClain, W.H. 1995. The tRNA identity problem: Past, present and future. In *tRNA: Structure, biosynthesis and function* (eds. D. Söll and U. RajBhandary), pp. 335–347. ASM Press, Washington DC.
- Muto, A., Andachi, Y., Yuzawa, H., Yamao, F., and Osawa, S. 1990. The organization and evolution of transfer RNA genes in *Mycoplasma capricolum*. *Nucleic Acids Res.* **18**: 5037–5043.
- Ochman, H. and Wilson, A.C. 1987. Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26**: 74–86.
- Orgel, L.E. and Crick, F.H. 1980. Selfish DNA: The ultimate parasite. *Nature* **284**: 604–607.
- Percudani, R., Pavesi, A., and Ottonello, S. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **268**: 322–330.
- Perna, N.T., Plunkett 3rd, G., Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., and Kirkpatrick, H.A. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**: 529–533.
- Petrov, D.A. 2001. Evolution of genome size: New approaches to a new problem. *Trends Genet.* **17**: 23–28.
- Proudfoot, N. 1980. Pseudogenes. *Nature* **286**: 840–841.
- Pupo, G.M., Lang, R., and Reeves, P.R. 2000. Multiple independent origins of shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl. Acad. Sci.* **97**: 10567–10572.
- Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K., and Whittam, T.S. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**: 64–67.
- Rokas, A., Williams, B.L., King, N., and Carroll, S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**: 798–804.
- Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker—A web

- server for aligning two genomic DNA sequences. *Genome Res.* **10**: 577–586.
- Sharp, P.M. 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: Codon usage, map position, and concerted evolution. *J. Mol. Evol.* **33**: 23–33.
- Sharp, P.M., Bailes, E.R., Grocock, J., Peden, J.F., and Sockett, R.E. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33**: 1141–1153.
- Urrutia, A.O. and Hurst, L.D. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159**: 1191–1199.
- Welch, R.A., Burland, V., Plunkett 3rd, G., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci.* **99**: 17020–17024.
- Whittam, T.S. 1996. Genetic variation and evolutionary processes in natural populations of *Escherichia coli*. In *Escherichia coli and Salmonella: Cellular and molecular biology*. (ed. F.C. Neidhardt), pp. 2708–2720. ASM Press, Washington, DC.
- Wright, F. 1990. The ‘effective number of codons’ used in a gene. *Gene* **87**: 23–29.
- Xue, H., Tong, K.L., Marck, C., Grosjean, H., and Wong, J.T. 2003. Transfer RNA paralogs: Evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. *Gene* **310**: 59–66.
- Yamano, F., Andachi, Y., Muto, A., Ikemura, T., and Osawa, S. 1991. Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. *Nucleic Acids Res.* **19**: 6119–6122.

RNA

A PUBLICATION OF THE RNA SOCIETY

Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome

Mike Withers, Lorenz Wernisch and Mario Dos Reis

RNA 2006 12: 933-942

References

This article cites 42 articles, 10 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/12/6/933.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to RNA go to:
<http://rnajournal.cshlp.org/subscriptions>