12-16-2016

Selection and Robustness in Bacterial Genome Evolution

Seila Omer
University Of Connecticut, seila.omer@uconn.edu

Selection and Robustness in Bacterial Genome Evolution

Seila Omer, Ph.D.

University of Connecticut, 2016

The research presented in this thesis attempts to address research questions related to the role of natural selection in the evolution of bacterial genes not expressed for function and in building mutational tolerance to translational errors. Studies on evolution of protein coding DNA sequences have provided the evidence for a current paradigm in evolutionary biology: only functional genes are undergoing selection against the deleterious effects of allele variants (purifying selection). I provide evidence that similar footprints of selection can be detected in genes that are not normally expressed for function during the bacterial life cycle. Using simulations for DNA sequence evolution, I demonstrate statistically significant deviations from neutral evolution for the studied genes. I suggest that purifying selection affects both functional and non-functional genes. I propose this might be caused by the dominant toxic effects of low level translation of mutated products in bacteria, due to misfolding and misinteraction. Natural selection also acts to remove the effects of translational errors. Stop codon readthrough events are more likely to have major structural and functional effects than simple nucleotide changes. Recent research has shown that strength of selection experienced by protein-coding genes is positively correlated with the level of gene expression. Expression of 3' untranslated regions (3' UTRs) carries with it the influence of natural selection on elongated products. I show that, for the subset of highly expressed genes analyzed, 3' UTRs in *Escherichia coli* genomes display features normally associated with coding regions, indicating tolerance to effects of translational errors

Selection and Robustness in Bacterial Genome Evolution

Seila Omer

B.Sc., University of Bucharest, **2000**

M.Sc., University of Bucharest, **2002**

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Selection and Robustness in Bacterial Genome Evolution

Presented by

Seila Omer, B.Sc., M.Sc.

Major Advisor _____
Johann Peter Gogarten, Ph.D.

Associate Advisor _____
Paul Lewis, Ph.D.

Associate Advisor _____
Victoria Robinson, Ph.D.

Associate Advisor _____
Daniel Gage, Ph.D.

Associate Advisor _____
Joerg Graf, Ph.D.

University of Connecticut
2016

Acknowledgements

I would like to express my gratitude to my major advisor, Dr. Johann Peter Gogarten, for guiding my first steps in molecular evolution and for providing me with the inspiration and courage to dream big. I also want to thank him for his mentorship, his valuable support and advice in my research endeavors and for the opportunity to meet like-minded scientists.

I would also like to thank my Ph.D. committee members: Dr. Daniel Gage for his helpful insight in my research projects, Dr. Joerg Graf for his rigorous analysis of my work, Dr. Paul Lewis for introducing me to the world of maximum likelihood and Bayesian phylogenetic inference and his technical advice on my research projects and Dr. Victoria Robinson, for providing me with insight into protein structure and folding and all the way, unwavering moral support.

I am forever indebted to Timothy J. Harlow in Gogarten Lab without whom this research would not have been possible. Thank you for the patience in helping me find my way in the realm of computer programming. Many thanks to the rest of Gogarten lab (Matt, Shannon, Erika, Jeff, Ryan, Marlene and Josh) for the valuable scientific discussions and team spirit.My work would not have been possible without the support of my family (Neila, Sami, Neni, Leila) and close friends here, at University of Connecticut and elsewhere (Colleen, Anne, Pam and Nat, Stephanie, Dan and many others), who stood by me, comforted me and had faith in me all these years. To them, I will be eternally grateful.

List of Figures and Tables

TABLE OF CONTENTS

I. Chapter 1 – General Introduction

1.1. Natural Selection and Bacterial Genome Dynamics

In biological organisms, the heritable information of proteins is encoded by genes that organize the genetic message in sequences of triplets (combination of 3 bases) made of four DNA nucleotides (A, T, G and C). The genetic code defines the translational language of these triplets. The genetic code assigns 20 amino acids and translational stop signals to each triplet. For each amino acid, there are 1-3 tRNA molecules which ribosomes use to synthesize polypeptides.

Uncovering the evolutionary history of individual bacterial sequences within the complex, dynamic and self-governing genomic context is central to our understanding of the major evolutionary forces shaping bacterial genomes. The main factors thought to influence the evolution of bacterial genomes include mutation, recombination, natural selection and genetic drift [1]–[5].

Recombination and mutation represent main sources of genetic variation in bacterial genomes on which natural selection operates. Types of mutations include, but they are not limited to, single base substitutions, insertions, deletions, duplications, inversions and translocations. The most frequent type of mutation, base substitution, can change the nucleotide triplet with no effect on the amino acid sequence of the protein (synonymous change) or with the consequence of changing the encoded amino acid (nonsynonymous change). The rate of mutations in bacterial genomes is thought to be also responsible for rapid diversification and speciation within genera [5], [6] . This rate however is highly dependent on the genetic background of the organism (i.e. presence or absence of DNA repair genes, suppressor mutations) and physiological state of the bacterial cell (i.e. growth stage, sporulation) [5]. In addition, recombination via horizontal gene transfer has been repeatedly shown to greatly influence adaptation across bacterial phyla[7]–[9].

Genes with high adaptive value within a given environment have the potential to spread rapidly in closely bacterial populations or bacterial species sharing that environment.

Each mutation, gene gain or gene loss carries with it a fitness effect on the organism, smaller or larger[10]. From the existing pool of mutational and recombination events in a genome, natural selection, as a key mechanism for evolution of organisms, preserves the advantageous mutations and purges the detrimental ones. The traditional method for measuring direction and strength of natural selection consists in estimating both the rate of synonymous substitution, dS, and nonsynonymous substitution, dN, taking into consideration all possible substitutions on the protein coding sequence and the actual substitutions. Typically, a dN/dS value smaller than 1 indicates the presence of purifying selection whereas dN/dS values bigger than 1 may indicate nonsynonymous substitutions are favored by natural selection (positive selection). Neutral and nearly neutral evolution of protein coding sequences are indicated by the equal rates of fixation for synonymous and nonsynonymous changes (dN/dS = 1). Several methods for measuring dN/dS ratio have been developed [11]–[13], many taking into account the evolutionary history of the examined organisms (phylogenetic relationships). However, measuring dN/dS values in closely related lineages poses at least two challenges: presence of recombination, which tends to distort the vertical signal of inheritance described by phylogenetic trees, and the presence of a low number of substitutions. In Chapter 2, I propose a counting method to measuring dN/dS in closely related lineages while allowing for recombination.

According to Fisher's theory on natural selection [14], the level of standing variation is commensurate with the potential for gene heritability. The strength and type of selection typically determines the heritability value (proportion of phenotypic variance due to genotypic differences) and successful maintenance of gene variants in organismal lineages. Purifying

selection eliminates deleterious gene variants from populations and leads to a decrease in the number of sequence polymorphisms and consequently genetic variation. The strength of natural selection depends on the recombination frequency and is the lowest in organismswith small effective population sizes ($N_e = 10^3$-10^4 individuals) such as bacterial endosymbionts. In very large effective populations ($N_e = 10^7$-10^{13}), as it is the case with many free living, fast-growing bacterial species, purifying selection is the most effective [2], [10], [15]–[20].

Natural selection and genetic drift are largely responsible for the large variability in the genome sizes within bacterial lineages as well as within bacterial domain. These forces have significant effects on genome expansions (via duplication and horizontal gene transfer) and reductions (via deletion). In case of endosymbionts, the repeated occurrence of bottlenecks and the presence of a stable environment with very little fluctuations lead to massive loss of genes which products are already provided by the host cell or other symbiotic bacteria [21]–[23]. In contrast, the majority of free-living bacteria experience rapid turnover of non-essential genes (mostly mobile genetic elements)[24]–[26].

Natural selection is also responsible for the increased density of coding sequences in bacterial genomes and shrinkage of intergenic regions as opposed to eukaryotic genomes where non-coding sequences are overwhelmingly dominant. An example for the extreme effects natural selection has on free-living bacteria is *Prochlorococcus* spp [27]–[30]. These photosynthetic bacterial species generally display extreme genome reduction affecting coding and non-coding sequences (~1.65 Mb, ~1,500 genes) , small cell size (0.5-07 μm) and very large subpopulation sizes (~10^{13}). The nutritional scarcity, large population size and extensive phage predation underlines the increased level of selective pressures acting on its genome. Consequently, its

current pangenome is estimated to comprise ~ 57,000 genes, many horizontally transferred within populations and an abundant diversity [27], [29].

Considering these evolutionary aspects of bacterial genome dynamics, elucidating the causes of natural selection at gene, individual and population level at short timescales may provide significant insights into several aspects of bacterial adaptation. The research presented in Chapter 2 concerns the impact of natural selection on genes vertically inherited in closely related bacterial lineages which may have never functioned in their encoded capacity (for example, as enzymes, phage capsids). Under the assumption that natural selection operates to remove detrimental alleles from genomes, paradoxically, despite of the severely detrimental effects on organismal fitness, many genes (such as components of transposons, prophages and cryptic operons) are clonally inherited by bacteria over short timescales under purifying selection, with no obvious phenotypic effects, bringing into question the reasons for their maintenance. Several studies have attempted to explain this apparent contradiction by suggesting possible functions for the encoded products of these genes, either known [31], [32] or unknown [33], [34], for the benefit of element [35]–[37], the host organism by exaptation [38] or the group of organisms [39]. In these cases, the negative selective signatures (dN/dS values much smaller than 1), usually associated with bacterial genes with described impact on fitness [40], [41], are offered as supporting evidence for this interpretation. My hypothesis presented and discussed in Chapters 2 and 3 offers a radically different perspective on the evolution of phenotypically silent genes. I argue that the short term presence of these genes in bacteria can be explained by selective forces acting outside of the effect of function on fitness, mainly against genetic perturbations threatening protein structure stability and protein-protein interactions in the host organism. Because the traditional test for selection applied to protein coding genes, the dN/dS ratio, cannot

distinguish between the selection for function and selection for other causes (protein misfolding, misinteraction between proteins), I argue that it can no longer be used as the only supporting evidence to justify a role for a translation product in organismal fitness. As a corollary to this theory, decaying genes in bacterial genomes, accumulating mutations in their coding regions, which are still transcribed and translated are possibly maintained under purifying selection not because the mutations occurring their coding regions are beneficial but rather because the mutations themselves have a detrimental effect on the fitness of the organism.

Overall, the research presented in this thesis underlines the effect of natural selection in adaptation of closely related bacterial lineages and genomes to naturally occurring genetic perturbations (mutations) and reconsiders the role of selection tests in discriminating functional elements from nonfunctional ones in genomes.

1.2. The Role of Natural Selection in Shaping DNA Sequence Robustness

It is widely accepted in evolutionary biology that coding DNA sequences in organisms from bacteria to humans experience strong selective pressures which shape and maintain structure and function of their encoded macromolecules and molecular ensembles. As an essential biological process, translation, although usually carried out flawlessly, can be affected by errors. Generally, natural selection acts to remove the effects of translational errors detrimental to organismal fitness. Frameshifts and stop codon readthrough events, in particular, are more likely to have major structural and functional fitness effects including loss-of-function than point mutations [42], [43]. Each mutation has therefore a specific effect on fitness, measurable or not. This creates a distribution of fitness effects for mutations [44]–[47]. Replication errors [48], transcriptional errors [49], forward and backward ribosomal hopping [50], loss of translational fidelity [51], sense and non-sense recoding can lead to misincorporation of amino acids in a protein sequence. It is estimated that in *Escherichia coli*, misincorporation rate is $\sim 5 \times 10^{-4}$ per codon [52], [53].

It has been suggested that readthrough errors in *Saccharomyces cerevisiae* triggered by the presence of prion [PSI^{+}] may unlock the genetic variation hidden within adjacent regions to the open reading frames (i.e. 3' untranslated regions)[54], [55]. Studies on 3' UTRs in *Saccharomyces* spp. and *Drosophila* spp. suggest a source for hidden phenotypic variation in the gene products is provided by addition of these regions to the existing protein sequence length [56], [57]. Upon expression, these regions may generate phenotypic diversity, dependent on the sequence context around the stop codon and the coding potential of the downstream region. Extension of open reading frames in proteins into the non-coding sequences (such as 5' and 3'

untranslated regions- UTRs) also carries with it the influence of natural selection on frameshifted/elongated products. Subsequently, these non-coding sequences may gain coding potential and after a while, their sequences could become selectively neutral or beneficial, or in other words, robust. For example, following a mistranslation error such as a frameshift, the resulting expression product may not be detrimental [58].

In the context of translational errors, mutational robustness is defined as the capacity of protein coding DNA sequences under selection to accumulate multiple mutations as a result of mistranslation without deleterious phenotypic effects [59]. For example, a recent study has shown that, under stringent selection (high antibiotic concentration), mistranslating populations of *E.coli* expressing the antibiotic resistance gene TEM-1 mitigate the effects of translational errors by optimizing TEM-1 protein sequence. This optimization consists in the accumulation of stabilizing amino acid changes which ensure structure stability and prevent protein misfolding given the high number of translation events required for organism's survival. Also, at the same time, mutations resulting in amino acid changes that produce misfolded and nonfunctional molecules are removed from the same populations [60]. A mutationally robust gene sequence gradually becomes complex so that potential changes do not significantly change the impact the gene has on organismal fitness. Another example of mutational robustness is offered by the degeneracy of the genetic code where multiple triplets encode same amino acids or when replacement of triplets brings about conservative changes in the protein sequence [61]. On the other hand, a phenotypically robust sequence may give a structurally robust protein that tolerates presence of mutations as long as their hierarchically propagated effects are mitigated, avoiding propensity towards misfolding and aggregation or tendency to misinteract [60].

It is assumed that, in most bacteria, mutation is a random process and in general, A-T biased [3]. While most mutations in coding regions are deleterious, paradoxically, mutator alleles are favored by selection in bacterial lineages [62]–[64]. This can be explained as a "high risk, high payoff" strategy where large population sizes and small fitness differentials drive population expansion and adaptation to highly fluctuating environmental conditions or to new ecological niches [65]. While short-term, the consequences of high mutation rates may be severely detrimental, ultimately, this strategy may exhaustively explore the genotype and phenotype landscape for the optimum peaks.

The research presented in the Chapter 4 of this thesis investigates the role of natural selection in building mutational robustness to translation errors in the 3' UTR regions of highly expressed genes. In large populations of bacteria, such as *E.coli*, there may be a tradeoff between the costs of expressing potentially deleterious sequences and the growth requirements for a high translation rate [66]. My hypothesis puts forward that mitigation of translational errors such as stop codon readthrough and frameshifts occurs through expression of immediately adjacent sequence, rather than evolution of an error-free translation system. Using genome-based and pangenome-based approaches, I present evidence that 3' UTR regions of highly expressed genes in *E.coli* populations, when expressed, may confer neutral phenotypic effects to frameshifted or elongated translation products and therefore, may provide robustness to translation errors. The findings presented in Chapter 4 point towards a potential major role for natural selection in driving intra-species adaptation, by increasing tolerance to error-prone but essential biological mechanisms. Therefore, investigating robustness and understanding its effects on genomes may allow us to decipher the mechanisms underlying the adaptability of bacterial populations.

II. Chapter 2 – Investigation of natural selection in bacterial genes not expressed for function

2.1. Introduction

An essential aspect in deciphering the molecular mechanisms responsible for the evolution and diversification of organisms is the analysis of the selection signatures observed at the DNA level. A common assumption in studies of the evolution of protein-coding DNA sequences is the association between the detected patterns of selection against the deleterious effects of allelic variants (also known as purifying selection) and the functional status of the encoded protein. A popular approach to detect purifying selection in protein coding DNA sequences is to infer the excess in the rate of synonymous substitutions (dS) relative to the rate of non-synonymous (dN) substitutions in pairwise comparisons within a set of orthologous protein coding sequences. These findings are usually reported as $dN\text{-}dS < 0$ or $dN/dS < 1$. The latter is also known as the omega ratio (ω). dN estimates represent the number of inferred non-synonymous changes corrected over the total number of possible non-synonymous substitutions (about 75% of sites in codons assuming equal rates among base changes); dS values give the number of inferred synonymous changes over the total number of possible synonymous changes (about 25% of the possible substitutions) [67]. In evolutionary biology, it is generally acknowledged that the relative rate of amino acid replacements at protein sequence level, also known as dN, is heavily influenced by natural selection through the structural and functional constraints imposed on the protein sequence. A dN/dS ratio significantly smaller than 1 indicates that some non-synonymous substitutions were removed by selection. This observation implies that some mutations, those that were removed from the population, interfered with the function of the protein. Over the last decades several types of methods have been developed for determining the dN/dS ratio for a gene of interest: distance based methods [68], [69], maximum likelihood methods [67], [70] or Bayesian methods [12]. Maximum likelihood and Bayesian methods

typically require the presence of a phylogeny for inference of dN/dS values in an orthologous set of protein-coding sequences. Specifically, to accurately measure dN/dS, one assumes a bifurcating phylogenetic tree that depicts the evolutionary relationships among sequences and the amount of evolution undergone by each lineage. However, bacterial and archaeal genomes are known to experience recombination, especially among closely related organisms [7], [71]–[74]. These observed phenomena add a layer of complexity in reconstructing the evolutionary history of genes by obfuscating the vertical signal. Thus, gene phylogenies often appear star-like [75]. Measuring selective signatures in such genes becomes challenging and often, impossible. Recent studies have found hallmarks of purifying selection in some bacterial genes with limited distribution (ORFans- genes with no recognizable homologs in other genomes- and group specific genes [33], [34], Gene Transfer Agents-GTA [31]) and prophage genes [32].

During the lysogenic state, the prophage genes involved in viral particle production and host cell lysis (structural genes) are transcriptionally repressed through the intervention of genetic switches involving the expression of transcriptional regulators (regulatory genes). The prophage genome is replicated alongside the bacterial genome usually with no apparent negative effects for the host fitness and sometimes, with the suggested addition of competitive advantage in stressful environments [38]. When bacterial cells are exposed to environmental stress (starvation, UV radiation, chemical mutagens etc.), the viral regulatory processes remove the transcriptional repression of the structural genes and the phage enters the lytic cycle, resulting in release of viral particles and host lysis.

GTAs present similar features of phage particles [76]–[79] and are suggested to mediate the transfer of genetic material between bacterial populations [39]. The widespread distribution of GTA-like genes across alpha-proteobacterial phyla and the topological similarity of the capsid gene phylogeny with that of the 16S rRNA were interpreted as evidence of their ancient origin [80]. Several studies used the finding of purifying selection acting on GTA genes of *Rhodobacter capsulatus* and of the genus *Bartonella* to argue in favor of the functional benefit GTAs may provide to the host population or to the GTA genes themselves [80], [81].

Another example in which the same type of selective signature was interpreted to indicate functionality is given by the genes with no recognizable homologs in closely related species or groups (ORFans and group specific genes). An analysis of dN/dS in such orthologous genes spanning the *Escherichia coli* and *Salmonella enterica* clades has found values much lower than 1, interpreted to reflect the selective constraints associated with important, functional roles for cell fitness [33]. A more recent study of group specific genes within *E. coli* and *Shigella* spp. clades and more widely distributed non-ORFan genes revealed dN values lower than the dS estimates, observation which led the authors to suggest that most ORFans are, in fact, functional genes [34].

These findings of purifying selection were commonly interpreted to surmise that selection maintains these genes in bacterial genomes as a result of benefit to the gene (as a selfish element) [37], the host organism, or as a result of altruistic acts benefiting the related population (kin selection) [82], [83]. The debate between kin selection *versus* selfish gene hypotheses as explanation for apparent altruistic behavior at the organismal level is ongoing in the field of

microbial evolution [8]. For example, a theoretical study on DNA secretion in bacteria shows that gene-level selection can be responsible for maintaining a gene responsible for gene sharing in bacterial populations [84]. In contrast, several mechanisms involving altruism and cooperation have attempted to explain evolution of certain other bacterial traits [85]

It is clear that the presence of a dN/dS value much smaller than 1 in the above mentioned examples can be interpreted as either selection for a selfish genetic element, selection for an unknown function in the host organism or for an unknown function as part of group selection. From an evolutionary perspective, sequence conservation across evolutionary time is associated, in general, with selection on sequence function. I propose that signatures of purifying selection are not necessarily hallmarks of selection for function, neither at the gene, the organismal, nor the group level. Rather, genes not expressed for function but present in genomes may display signatures of negative selection against the detrimental effects of non-synonymous mutations on the structural stability of the protein and the consequential decrease of organismal fitness.. In the context of evolution, the presence of a gene expression product may be necessary and sufficient to generate a footprint of natural selection on the encoding gene sequence. Once expressed, a protein-coding gene has a specific but not necessarily quantifiable impact on the fitness of the organism [86]. Bacterial genomes generally exhibit a high-density of functional genes and just a small number of decaying coding sequences (pseudogenes) [87]. In the context of a deletion bias affecting bacterial genomes [1], [25], [26], however, some genes, detrimental to host fitness, are maintained in closely related strains in a repressed state with no apparent benefit for the organism harboring them. A large number of such genes include mobile genetic elements

(phages, transposases, etc.) and operons encoding toxic products with significant effects on bacterial fitness upon expression in certain environments [88].

I will provide analyses of several genes that appear to have been vertically inherited in the analyzed groups and that are candidates for genes that have not been expressed for function. One group of such genes is represented by prophages. Because the phage regulatory genes almost always have a direct impact on host fitness because their expression controls activation of the structural genes, I restricted the analyses to structural genes in prophages [40–42]. Within that scope, I included in these analyses defective prophages with different degrees of genome decay. Another group of genes with direct fitness effect on bacterial cell includes transposases. If a transposase is functional, the transposition process via a cut-and-paste mechanism often results in mutagenic effects at the DNA level. Because most mutations caused by transposition in coding regions of the host genome are deleterious to host fitness, they will be removed by natural selection. Additionally, regulation of existing transposases is selectively favored. For example, the inactivity of *Tn5* is caused by the inhibitory interaction between the N- and C-termini of the transposase preventing the mechanism of transposition to occur [92]. Additionally, to these types of repressed genes, assembly-line type of operons that synthesize toxic products, such as the malleilactone operon *Burkholderia pseudomallei,* with no detectable expression levels represent other candidates for studying the selection pressures on genes not expressed for function [93]. I discuss processes that might result in the signature of purifying selection. I conclude that a $dN/dS < 1$ may be insufficient to infer selection for function. I do not consider the mere fact of providing a template for transcription as providing a function; rather, I use the typical biological definition for function that implies a contribution to the fitness of the organism or to the gene

itself [32]. Couched in these terms, the findings presented in this chapter reveal that the "causal role" of being transcribed and translated at low levels is sufficient to create a signature of purifying selection; a selected effect that increases the gene's, host's, or group's fitness is not necessary. This hypothesis has been recently presented in an opinion article [95] which it has been included in Chapter 3.

2.2. Materials and Methods

2.2.1. Sequences

Nucleotide and amino acid sequences for the genes used in this study were collected from fully sequenced and draft genomes stored at the IMG DOE Joint Genome Institute [96]. I used the genome browser implemented at the JGI to visualize and identify orthologous gene neighborhoods.

I have identified and analyzed 10 structural gene sets found in the cryptic E14 prophage genome integrated in 6 *E. coli* strains. The prophage genome is flanked at the 3' end by the bacterial gene for isocitrate dehydrogenase (*idh*) and at 5' end by an iron transport operon (*sitABCD)* (Figure1A). Additionally, I have examined 13 structural and lysis genes sets from the conserved defective phage PBSX, found in 39 genomes of *Bacillus* genus (*subtilis, mojavensis*), flanked by altronate hydrolase (*uxaA*) and an inorganic phosphate transporter (*pit*) (Figure 1B), 4 structural genes from a prophage remnant, Lp3, located in 7 *Lactobacillus casei* genomes between a ribose phosphate pyrohosphokinase (*rpp*) and an amino acid permease (*aap*) (Figure 1C), a putative transposase present in a conserved gene neighborhood between aspartate 1-decarboxylase (*panD*) and panthotenate synthetase (*panC*) in 33 *E. coli* genomes (Figure 1D) and a cryptic operon containing 11 enzymes involved in the synthesis of malleilactone (*mal*), a polyketide synthase-derived cytotoxic siderophore, located between an auxin efflux transporter and a LuxR transcriptional regulator of 18 *Burkholderia pseudomallei* strains (Figure 1E). I have also identified 2 defective prophage genes in identical gene neighbourhoods in 12 *Anaplasma marginale* genomes, 1 *Anaplasma centrale* genome, 8 *Anaplasma phagocytophilum* genomes and 15 *Ehrlichia* spp. genomes, flanked downstream by a hemolysin and uroporphyrinogen

decarboxylase and upstream, by a methylase involved in ubiquinone/menaquinone biosynthesis. I have found another putative transposase, located between phospholipase D and peptidyl-prolyl cis-trans isomerase A, in 18 *Corynebacterium pseudotuberculosis* genomes. A complete list of gene sequences and accession numbers used in this study is available as an associated file (Associated File 1).

2.2.2. Alignment and tree building

Amino acid and nucleotide codon-based alignments were built using MUSCLE [97] as implemented in SeaView 4.2 [98]. Phylogenetic trees were constructed from nucleotide alignments by a maximum-likelihood method using PhyML [99], under a general time reversible model (GTR), with 4 gamma rate categories, fraction of invariant sites estimated from the data and 100 bootstrap replicates. Starting trees were generated by BioNJ algorithm, and tree search was carried out using the combination of the nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR) option.

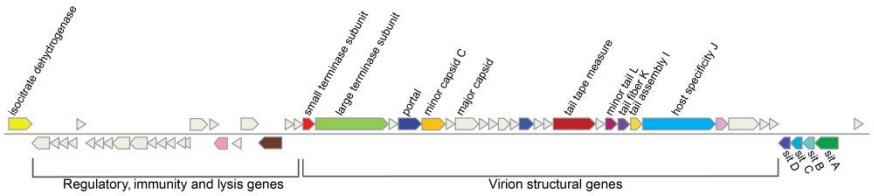
2.2.3. Counting the minimum number of substitutions

To determine the minimum number of synonymous and non-synonymous substitutions in a tree-independent way, *i.e.*, assuming that all homoplasies are due to recombination between sequences, I used a program written in Perl to count the number of observed nucleotide and

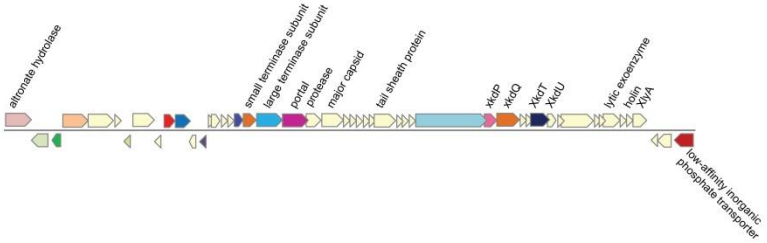
Figure 1. Generic genomic neighborhoods of the analyzed genes-

A. *E. coli* NC101 prophage E14**; B.** *B. subtilis* subsp. *subtilis* 168 prophage PBSX; **C.** *L. casei* W56 prophage Lp3**; D**. *Ehrlichia* spp. and *A. marginale/centrale/ phagocytophilum* putative phage; **E.** *E. coli* HS transposase; **F.** *C. pseudotuberculosis* putative transposase; **G.** *B. pseudomallei* K96243 chromosome 2 malleilactone cluster.

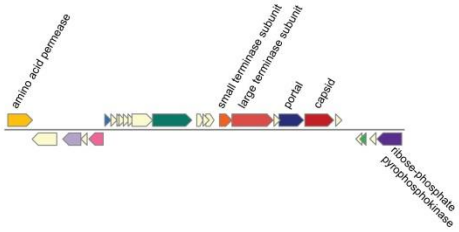
A *Escherichia coli* NC101 prophage E14



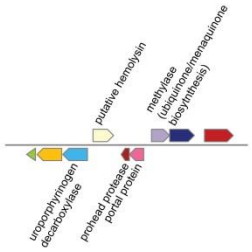
B *Bacillus subtilis subsp. subtilis* 168 prophage PBSX



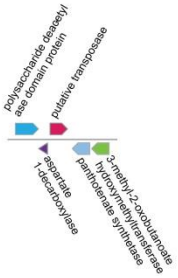
C *Lactobacillus casei* W56 prophage Lp3



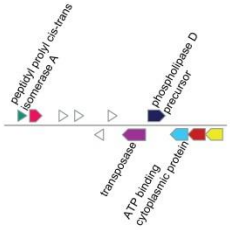
D *Ehrlichia sp.* and *Anaplasma marginale/ A. phagocytophilum/ A. centrale* putative phage



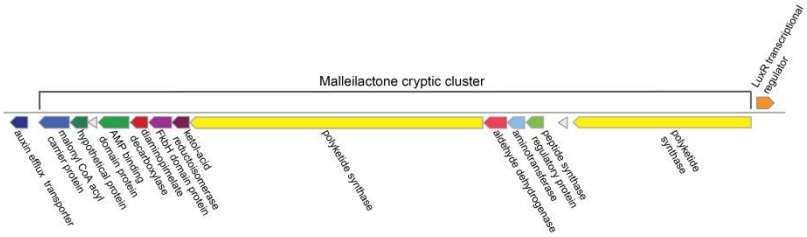
E *Escherichia coli* HS



F *Corynebacterium pseudotuberculosis* putative transposase



G *Burkholderia pseudomallei* K96243 chromosome 2



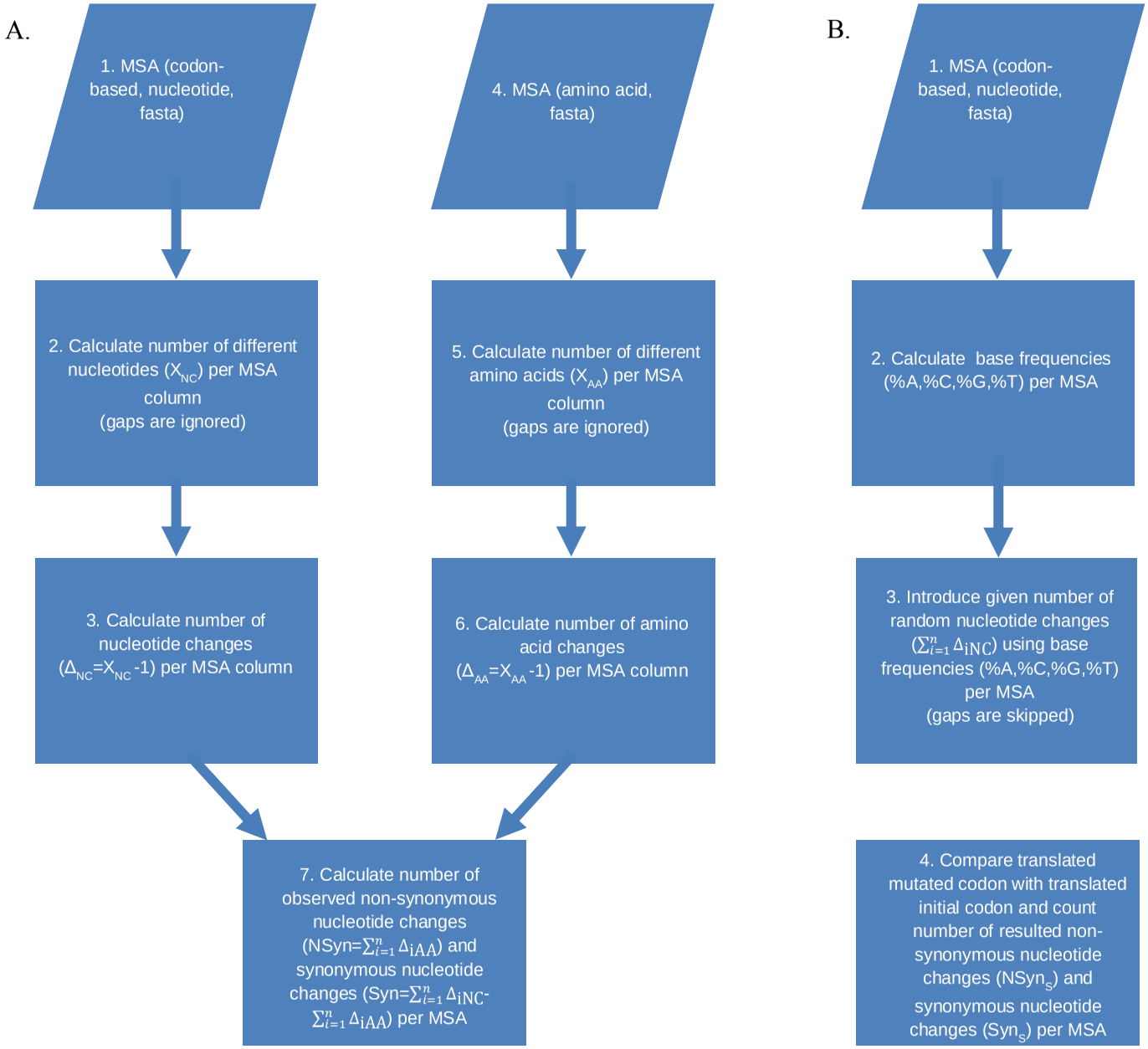
amino acid differences in each alignment. The algorithm is described in Figure 2A. Using a multiple sequence alignment (MSA) as input, I counted the number of different nucleotides (X_{NC}) in each column of the MSA, inferring the number of nucleotide changes (Δ_{NC}) as $\Delta_{NC}=X_{NC}-1$. This number is conservative as multiple parallel substitutions per site are not considered when making these calculations. Similarly, I counted the number of observed different amino acids (X_{AA}) and I calculated the observed number of amino acid differences (Δ_{AA}) as $\Delta_{AA}=X_{NC}-1$. I inferred the number of observed non-synonymous nucleotide changes (NSyn) as NSyn=$\sum_{i=1}^{n}\Delta_{iAA}$, where each non-synonymous change is assumed to be the result of only one observed nucleotide difference. This assumption may lead to slight underestimation of the non-synonymous changes, if divergent sequences are analyzed. The number of inferred synonymous changes (Syn) becomes Syn=$\sum_{i=1}^{n}\Delta_{iNC}$-NSyn. This approach assumes that in the evolutionary history of the sequences each change occurred only once and it was passed on to the sequences that contain this change through vertical inheritance or gene transfer followed by homologous recombination. The counting approach considers homoplasies found in the tree-based approach (below) as a result of homologous recombination between the sequences.

2.2.4. Simulations

Assuming that non-functional genes evolve neutrally, I replicated this process using a simulation program written in Perl (Figure 2B). This program uses as input the codon-based MSA for each gene. In our algorithm, the neutral mutational process consisted in placing the same number of nucleotide differences observed at the counting stage (X_{NC}) within the sequences of the MSA. The type of nucleotide substitutions was biased in terms of nucleotide

Figure 2. Diagrams depicting the algorithms used in the counting method for measuring dN/dS

A. Algorithm used to estimate the number of observed and simulated nucleotide changes (synonymous and non-synonymous) in each multiple sequence alignment (MSA). Syn- number of synonymous changes; NSyn- number of non-synonymous changes; NC- nucleotides; AA- amino acids; S-simulated; **B. Algorithm used to** replicate the pattern of substitutions under a neutral evolution scenario. A given number of random nucleotide substitutions was placed in each MSA using the inferred base frequencies and the resulting change in the corresponding codon was recorded accordingly as a synonymous or a non-synonymous change. The simulations were run 1 million times for each MSA under the given parameters.


A.
1. MSA (codon-based, nucleotide, fasta)
2. Calculate number of different nucleotides (X_{NC}) per MSA column (gaps are ignored)
3. Calculate number of nucleotide changes ($\Delta_{NC}=X_{NC}-1$) per MSA column
4. MSA (amino acid, fasta)
5. Calculate number of different amino acids (X_{AA}) per MSA column (gaps are ignored)
6. Calculate number of amino acid changes ($\Delta_{AA}=X_{AA}-1$) per MSA column
7. Calculate number of observed non-synonymous nucleotide changes (NSyn=$\sum_{i=1}^{n}\Delta_{iAA}$) and synonymous nucleotide changes (Syn=$\sum_{i=1}^{n}\Delta_{iNC}$-$\sum_{i=1}^{n}\Delta_{iAA}$) per MSA
B.
1. MSA (codon-based, nucleotide, fasta)
2. Calculate base frequencies (%A,%C,%G,%T) per MSA
3. Introduce given number of random nucleotide changes ($\sum_{i=1}^{n}\Delta_{iNC}$) using base frequencies (%A,%C,%G,%T) per MSA (gaps are skipped)
4. Compare translated mutated codon with translated initial codon and count number of resulted non-synonymous nucleotide changes ($NSyn_S$) and synonymous nucleotide changes (Syn_S) per MSA

content of the sequences in the MSA – the program calculated A, T, G and C percentages of the MSA and used them to randomly draw the nucleotide change. The mutational events were considered as single, independent, random with respect to the nucleotide position in the MSA and the nucleotide sequence accumulated substitutions at each position with the same probability. I simulated this substitution process for N=1,000,000 times. The outcome of each nucleotide substitution was recorded as a synonymous or non-synonymous nucleotide change. The output of each series of simulations resulted in count values used to build a distribution of observed synonymous changes when given X_{NC} as the total number of occurring substitutions. The results of simulations also gave me the frequencies of synonymous and non-synonymous changes (P_{NS}) and synonymous nucleotide changes (P_S) per MSA under the assumption of neutral evolution.

2.2.5. Statistical modeling

Under a neutral scenario, in the absence of any selection pressures a protein coding DNA sequence experiences and retains synonymous and non-synonymous changes with the same rate. The expected number of synonymous and non-synonymous substitutions under the null hypothesis of neutral evolution for DNA sequences can be modeled using a binomial distribution function. Under the neutral hypothesis, I can describe the p-value (P) as the cumulative probability of seeing the number of observed synonymous changes (k) or more given a certain number of random substitution events (n) and given the probability of observing a synonymous change (p) (Equation 1).

$P(k|n,p) = 1 - \sum_{i=0}^{|k-1|} \binom{n}{i} p^i (1-p)^{n-i}$ (Equation 1) for $k = 0, 1, 2, \ldots, n$,

Where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

2.2.6.dN/dS estimation

I used the calculated values for synonymous and non-synonymous changes and derived frequencies of synonymous and non-synonymous changes, F_{syn} and $F_{nsyn,}$ to calculate the dN/dS ratio values for each gene using the equations below (Equations 2-3).

$$dN = \frac{\text{Nsyn}}{P_{NS} \text{ x No.sites}} \quad \text{(Equation 2)}$$

$$dS = \frac{\text{Syn}}{P_S \text{ x No.sites}} \quad \text{(Equation 3)}$$

2.2.7. Tree based approach to assess dN/dS ratios

To consider the other extreme that all observed homoplasies are due to parallel independent substitution events, I used the maximum likelihood approach of Goldman and Yang [67] to calculate dN/dS ratios that is implemented in the program *codeml* from the PAML4 package [11]. I conducted likelihood ratio tests on maximum likelihood estimates of probable trees under three models for dN/dS ratios or ω (M0 with single, estimated ω ratio *versus* neutral M0 with ω=1; and neutral M0, ω=1 *versus* nearly neutral model M1a , ω_1=1, ω_2<1). The p-values for the likelihood ratio tests represent the right-tailed probability of the chi-square distribution function.

2.2.8.Parsimony

To estimate the number of homoplasies under a tree phylogeny of the analyzed datasets, I used the unrooted parsimony algorithm implemented in the program Dnapars from the PHYLIP package [100]–[102]. I designated gaps in our input alignments as unknown states and randomized input order 1000 times.

2.2.9. Homoplasy analysis and tree congruence testing

I used simulations of protein coding DNA sequence evolution using the Monte Carlo algorithm implemented in *evolver* from the PAML package [11] to gauge if homoplasies occurred with a significantly higher rate than expected under the assumption of a tree-like history without recombination. I generated 100 multiple sequence alignments (3507 sites) using codon frequencies and tree topology derived from the original codon-based alignment for *hsJ* gene from *E. coli* E14 prophage. The datasets were then employed to derive the lengths of the most parsimonious trees (as described in 2.2.8.) and then examined for differences in the homoplasy number when compared with the observed nucleotide changes in the original dataset. The significance was assessed by counting the occurrences of the same difference in the number of homoplasies as it was found in the original dataset. I also used the phylogenetic recombination detection algorithm implemented in GARD (HYPHY package) to test for recombination in our sequences [13]. Additionally, incompatibility between the trees of the tested genes with those of the neighboring bacterial genes was used infer the possibility of recombination events. To determine phylogenetic conflict between neighboring genes, I created sets of trees with similar

topology using maximum likelihood trees calculated from bootstrap replicates using PhyML. Site likelihoods were estimated with default settings by RaxML[103], and approximately unbiased (AU) [104]and Shimodaira-Hasegawa (SH) [105] tree congruence tests were performed as implemented in CONSEL [106].

2.3. Results

2.3.1. Counts and simulations

If genes that do not provide a function to the organism in which they evolve, or to themselves in case of selfish genetic elements or the group of organisms in case of group selection, do evolve neutrally, i.e. experience synonymous and non-synonymous substitutions with the same rate, then genes that are passed from parent to offspring without having functioned in their encoded capacity should evolve neutrally. To test this hypothesis, I identified and analyzed vertically inherited, repressed genes from several bacterial species (Figure 1). In my analyses, I used both tree based and tree independent approaches (Figure 2) [33]. A tree based approach strictly assumes a vertical signal where the apparent homoplasies (meaning identical nucleotide changes at the same sites in different taxa) are the result of multiple independent substitutions per site and involve two separate events. The tree independent approach, however, allows for a vertical inheritance signal and a horizontal transmission signal. In this latter case, apparent homoplasies are explained through recombination only, and are considered to be the consequence of a single change. The two approaches represent, therefore, the two extremes with respect to apparent homoplasies. The results from both approaches are very similar and summarized in Tables 1-9.

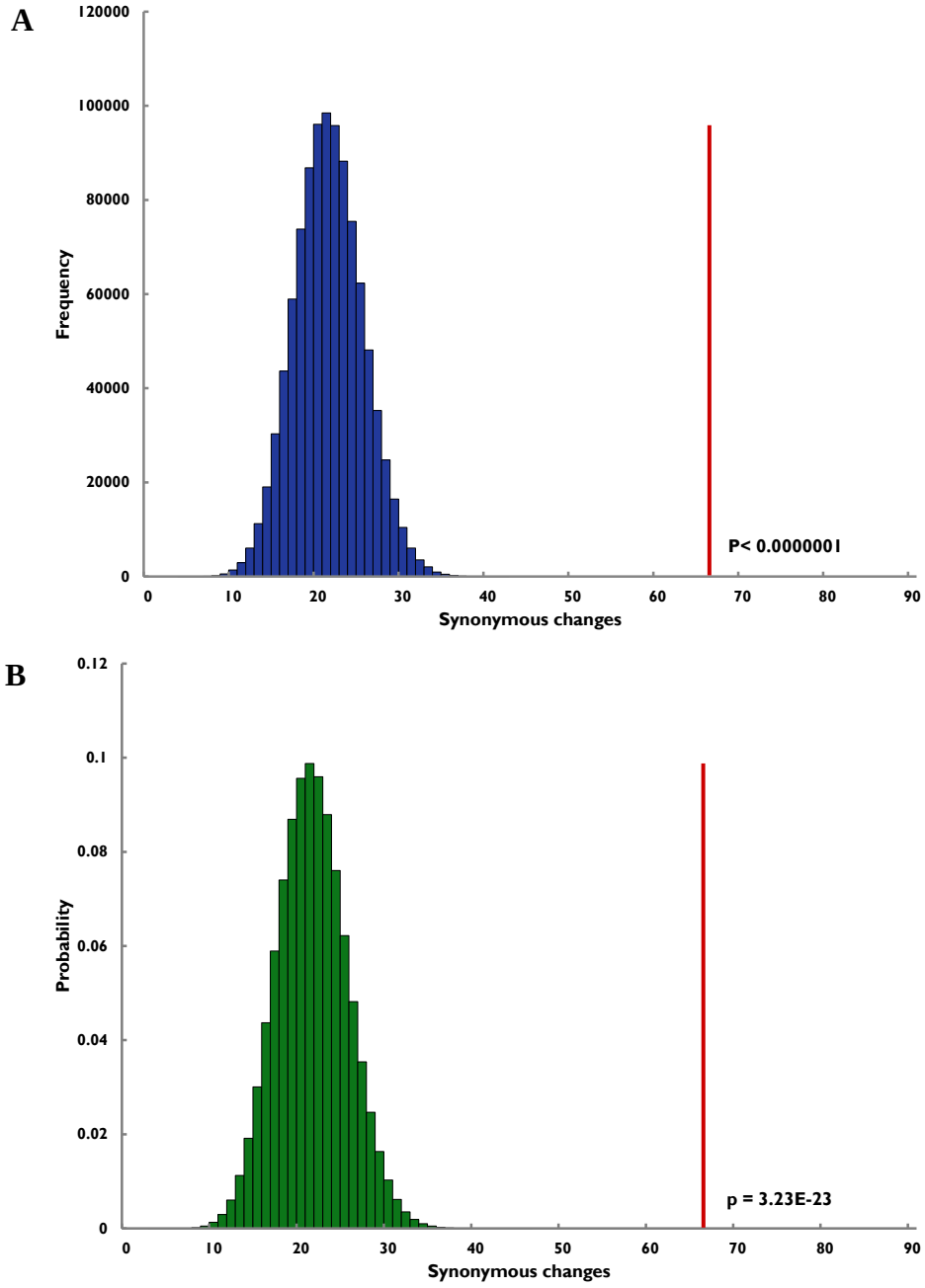
Our data do not distinguish mutations that exist as polymorphisms within a population from substitution events, i.e. mutations that were fixed in the population.

Under the hypothesis of neutral evolution, the expected ratio of non-synonymous to synonymous changes for either case is 1, and values significantly smaller than 1 reveal purifying selection removing some non-synonymous changes from the population and lowering their probability to become fixed in the lineage. The columns in Tables 1-5 giving the observed nucleotide differences, synonymous and non-synonymous changes, respectively, illustrate that, for the majority of the genes analyzed, the number of synonymous changes exceeds the number of non-synonymous changes as a proportion of observed nucleotide differences, contrary to the neutral expectation.

To test whether the observed number of non-synonymous and synonymous changes are significantly different from those expected for sequences undergoing neutral evolution, I carried out simulations in which the number of nucleotide substitutions placed match the value of observed nucleotide differences in each multiple sequence alignment (the Obs column in Tables 1-9). The frequency distribution of synonymous changes expected from simulations under the neutral model is illustrated in Figure 3A for the major capsid gene in *E. coli* E14 prophage. I can use this distribution to determine the empirical probability, *P*, of observing a given number of synonymous substitutions. I observed no occurrence of 66 or more synonymous changes in 10^6 simulations of 90 substitutions each. Thus, the distribution for the major capsid gene shows that this probability under the neutral hypothesis is less than 1 per 10^6. I found similar values for *P* in the majority of the genes I analyzed with the exception of *kar* and *lig* genes from *B. pseudomallei* malleilactone operon.

Figure 3. Distributions of occurring synonymous changes for major capsid gene from *E. coli* E14 prophage

A. Distribution of synonymous change counts resulted from simulations. Empirical probability (P) is the proportion of simulations (N=1,000,000) resulting in the inferred number of synonymous changes using our counting algorithm (red vertical bar); **B**. Binomial distribution of predicted probabilities for synonymous changes. Binomial p-value (p) is the probability of observing the same inferred number or more synonymous changes (red vertical bar).


A
120000
100000
80000
60000
40000
20000
0
Frequency
0
10
20
30
40
50
60
70
80
90
Synonymous changes
P< 0.0000001
B
0.12
0.1
0.08
0.06
0.04
0.02
0
Probability
0
10
20
30
40
50
60
70
80
90
Synonymous changes
p = 3.23E-23

2.3.2. Estimating probabilities from binomial distributions

If a nucleotide change has a fixed probability to result in a synonymous change, then the probability of observing values equal or larger than the observed can be calculated using the binomial distribution. Similarly to the simulation counts, I built a binomial probability distribution for the major capsid gene from the *E. coli* prophage (Figure 3B). To take codon composition into account, I used the probabilities for synonymous changes that resulted from simulations (Fsyn). Given the numbers of observed nucleotide differences, inferred number of synonymous changes and the probability for synonymous change, the distribution illustrates the statistically significant departure between the inferred synonymous count (the red vertical bar) and the predicted values of the binomial function (the grey distribution).

The results of this approach are summarized in Tables 1-9. The p-values for the observed fraction of synonymous changes are consistently very close to 0 in all the examined prophage genes from *E. coli, B. subtilis, L. casei*, *Anaplasma* spp. , *Ehrlichia* spp. and the putative transposase genes from *E. coli* and *C. pseudotuberculosis*. The p-values obtained when combining all parts of the syntenic regions into a single analysis can be confidently approximated to zero. The calculated p-values for the *B. pseudomallei* operon genes are higher, due the overall smaller number of changes included in the dataset. Three genes, *kar, lig* and *mlp*, show values close to 1. Nevertheless, the binomial p-value for the combined coding regions of the malleilactone operon is small, 1.26E-60, rejecting the neutral hypothesis for the combined syntenic region.

2.3.3. Determination of dN/dS ratio in sequences prone to recombination

Using the counts for synonymous and non-synonymous changes, their corresponding frequencies derived from simulations and the number of nucleotide sites in a gene, it becomes possible to calculate the rate of synonymous change (dS) and the rate of non-synonymous change (dN). I have thus estimated the dN/dS ratio in the studied genes without considering the phylogenetic relationships between the sequences, i.e., ignoring the possibility of independent parallel changes. Tables 1-9 and Supplemental Table S1 summarize the individual rates and overall dN/dS ratio for each gene. Prophage structural gene sequences in the examined bacterial genomes display a range of values, which may indicate variability in the strength of selection affecting them. Additionally, the *B. subtilis* dataset displays significantly lower dN/dS values than the ones calculated for *E. coli* and *L. casei* datasets. An analysis of individual dN and dS rates indicate that in fact the dS rates in *B. subtilis* are up to 10 times higher than in the other datasets. At short timescales, a larger number of non-synonymous changes is expected to be encountered in gene sequences because many non-synonymous differences may represent polymorphisms rather than fixed nucleotide differences, i.e. substitutions. Slightly deleterious mutations as result of non-synonymous substitutions are more likely to be removed by natural selection at longer timescales, thus it is more likely to observe an increased proportion of synonymous changes over longer time periods [34, 35]. Our observation is consistent with stronger purifying selection acting on synonymous sites at increasing divergence level in the case of *Bacillus* spp. PBSX and *Ehrlichia* spp. prophage gene sequences, which include not only closely related strains of *B. subtilis* but also more distant lineages. Because our datasets cover different time spans and therefore different ratios of polymorphisms to substitution events, the comparison of the magnitude of the dN/dS ratios between datasets is of limited value. To obtain

a better measure, I have also calculated the dN/dS ratios for neighboring functional genes from the same set of organisms. A comparison between dN/dS ratio values for each cluster and the dN/dS values corresponding to their flanking genes (listed in Table S1) reveals, in general, much lower values for the bacterial flanking genes indicative of stronger purifying selection operating on sequences likely to be functional.

2.3.4. Determination of dN/dS ratios ignoring recombination

Counting only the minimum number of synonymous and non-synonymous substitutions assumes that every substitution event occurred only once. I therefore compared the values obtained through the counting approach with those obtained through a tree-based approach that assumes that the sequences evolved on single tree without recombination, and that homoplasies are due to independent changes. Comparison of the maximum likelihood estimates of dN/dS with estimates inferred using the counting method reveals only small differences in values. The results of the likelihood ratio tests are shown in Tables 1-9 and in supplemental Table S2. These tests reject the hypothesis of neutral evolution at the 5% significance level for all genes, except for the *kar* gene from *B. pseudomallei*.

Table 1. Comparison of dN/dS estimates in *Escherichia coli* E14 prophage structural genes

					Simulations					Rec	ML[1]		Parsimony[2]
Gene	**Sites**	**Obs**	**Syn**	**NSyn**	**F_{syn}**	**F_{nsyn}**	**Binomial *P*-value**	**dN**	**dS**	**dN/dS**	**dN/dS**	**LRT p-value (ω=est. vs ω=1)**	**Steps**
mcps	1053	90	66	24	0.237	0.763	3.23E-23	0.0307	0.2727	0.1127	0.1314	2.01E-20	94
mcpsC	1329	81	59	22	0.250	0.750	1.98E-19	0.0221	0.1776	0.1243	0.1770	1.65E-14	84
ltsu	1926	75	67	8	0.224	0.776	7.10E-35	0.0054	0.1553	0.0345	0.0377	4.40E-17	82
stsu	546	100	66	34	0.232	0.768	1.07E-19	0.0816	0.5230	0.1560	0.2515	3.16E-09	101
ptl	1602	55	46	9	0.226	0.774	1.36E-21	0.0073	0.1270	0.0573	0.0808	7.17E-20	64[d]
mtL	699	30	26	4	0.253	0.747	2.72E-12	0.0077	0.1476	0.0521	0.0500	2.28E-12	30
hsJ	3507	723	498	225	0.232	0.768	4.72E-130	0.0870	0.5568	0.5568	0.1710	2.80E-112	773[d]
tfK	744	41	28	13	0.271	0.729	4.83E-08	0.0240	0.1394	0.1726	0.1835	9.87E-08	44
taI	672	39	33	6	0.267	0.733	6.62E-14	0.0122	0.1837	0.0663	0.0799	2.87E-12	40
ttmp	2580	375	243	132	0.237	0.763	7.92E-64	0.0671	0.3976	0.1689	0.2796	2.77E-29	378
Overall	14622	1609	1132	477	0.244*	0.756*	0.00E+00*						

*weighted; [1]- PAML; [2]- PHYLIP; [d]-significant difference when compared to **Obs**

Abbreviations: *mcps*- major capsid protein; *mcpsC*- minor capsid C protein; *ltsu*- large terminase subunit; *stsu*- small terminase subunit; *ptl*- portal protein; *mtL*- minor tail protein L; *hsJ*- host specifity protein J; *tfK*- tail fiber protein K; *taI*- tail assembly protein I; *ttmp*- tail tape measure protein; **Obs**- observed nucleotide differences; **Syn**- synonymous changes; **NSyn**-non-synonymous changes; **F_{syn}**- synonymous frequency; **F_{nsyn}**- non-synonymous frequency; **Rec**-our method, **ML**- maximum likelihood; **dN**: non-synonymous substitutions per non-synonymous site; **dS**: synonymous substitutions per synonymous site.

Table 2. Comparison of dN/dS estimates in *Lactobacillus casei* prophage structural genes

					Simulations					Rec	ML[1]		Parsimony[2]
Gene	**Sites**	**Obs**	**Syn**	**NSyn**	**F_{syn}**	**F_{nsyn}**	**Binomial *P*-value**	**dN**	**dS**	**dN/dS**	**dN/dS**	**LRT p-value (ω=est. vs ω=1)**	**Steps**
ltsu	1704	224	183	41	0.214	0.786	1.99E-82	0.0306	0.5019	0.0610	0.0722	1.78E-63	224
ptl	1185	72	51	21	0.220	0.780	1.18E-18	0.0271	0.1957	0.1160	0.1905	1.11E-10	72
cps	1584	263	206	57	0.225	0.775	4.38E-82	0.0464	0.5785	0.0802	0.1190	4.36E-53	263
stsu	471	98	70	28	0.218	0.782	1.37E-25	0.0760	0.6811	0.1116	0.1041	8.91E-20	98
Overall	4944	657	510	147	0.219*	0.746*	2.09E-202*						

*weighted; [1]- PAML estimation; [2]- PHYLIP estimation

Abbreviations: *ltsu* -large terminase subunit; *ptl* –portal protein; *cps*- capsid protein; *stsu*- small terminase subunit; Obs- observed nucleotide differences; **Syn**- synonymous changes; **NSyn**-non-synonymous changes; **F_{syn}**- synonymous frequency; **F_{nsyn}**- non-synonymous frequency; **Rec**-our method, **ML**- maximum likelihood; **dN**: non-synonymous substitutions per non-synonymous site; **dS**: synonymous substitutions per synonymous site.

Table 3. Comparison of dN/dS estimates in *Bacillus subtilis* PBSX prophage structural genes

					Simulations					Rec	ML[1]		Parsimony[2]
Gene	**Sites**	**Obs**	**Syn**	**NSyn**	**F_{syn}**	**F_{nsyn}**	**Binomial *P*-value**	**dN**	**dS**	**dN/dS**	**dN/dS**	**LRT p-value (ω=est. vs ω=1)**	**Steps**
cps	933	187	169	18	0.226	0.774	5.08E-87	0.0249	0.7999	0.0312	0.0258	0.028	265[d]
hn	261	91	61	30	0.222	0.778	7.56E-20	0.1478	1.0519	0.1405	0.1381	3.16E-24	111[d]
lex	837	335	222	113	0.230	0.770	6.81E-39	0.1754	1.1515	0.1523	0.1660	1.12E-75	458[d]
ltsu	1299	361	281	80	0.216	0.784	1.60E-114	0.0786	1.0011	0.0785	0.0645	2.07E-164	511[d]
ptl	1494	404	344	60	0.218	0.782	2.10E-162	0.0513	1.0574	0.0486	0.0466	6.22E-207	571[d]
pts	825	205	161	44	0.219	0.781	2.26E-66	0.0683	0.8892	0.0768	0.0837	2.14E-73	256[d]
stsu	795	206	154	52	0.227	0.773	1.76E-56	0.0846	0.8551	0.0989	0.0825	1.23E-74	280[d]
tsp	1407	409	362	47	0.234	0.766	4.70E-172	0.0436	1.0972	0.0398	0.0461	6.66E-191	529[d]
xkdP	705	263	191	72	0.209	0.791	3.36E-72	0.1291	1.2976	0.0995	0.0837	8.62E-94	357[d]
xkdQ	983	288	252	36	0.222	0.778	3.10E-123	0.0471	1.1529	0.0408	0.0320	4.94E-170	411[d]
xkdT	1044	322	264	58	0.228	0.772	3.60E-112	0.0719	1.1105	0.0648	0.0581	3.89E-167	492[d]
xkdU	576	160	130	30	0.229	0.771	7.60E-117	0.0675	0.9859	0.0685	0.0814	7.20E-98	264[d]
xlyA	903	314	247	67	0.247	0.753	5.65E-51	0.0985	1.1083	0.0889	0.0459	3.31E-131	475[d]
Overall	12062	3545	2838	707	0.225*	0.775*	0.00E+00*						

* weighted; [1]- PAML; [2]- PHYLIP; [d]-significant difference when compared to **Obs**

Abbreviations: *cps*- capsid protein; *hn*- holin; *lex*- lytic exoenzyme; *ltsu*- large terminase subunit; *ptl*- portal protein; *pts*- protease; *stsu*- small terminase subunit; *tsp*- tail sheath protein; *xkdP*- murein binding protein; *xkdQ*- tail protein; *xkdT*- putative base plate assembly protein; *xkdU*- hypothetical protein; *xlyA*- N-acetylmuramoyl-L-alanine amidase; **Obs**- observed nucleotide differences; **Syn**- synonymous changes; nsyn-non-synonymous changes; **F_{syn}**- synonymous frequency; **F_{nsyn}**- non-synonymous frequency; **Rec**-our method, **ML**- maximum likelihood; **dN**: non-synonymous substitutions per non-synonymous site; **dS**: synonymous substitutions per synonymous site.

Table 4. Comparison of dN/dS estimates in *Escherichia coli* putative transposase gene

					Simulations					Rec	ML[1]		Parsimony[2]
Gene	**Sites**	**Obs**	**Syn**	**NSyn**	**F_{syn}**	**F_{nsyn}**	**Binomial *P*-value**	**dN**	**dS**	**dN/dS**	**dN/dS**	**LRT p-value (ω=est. vs ω=1)**	**Steps**
tn	1029	185	125	60	0.215	0.785	4.31E-41	0.0742	0.5654	0.1313	0.1491	3.78E-55	319[d]

* weighted; [1]- PAML estimation; [2]- PHYLIP estimation; [d]-significant difference when compared to **Obs**

Abbreviations: *tn*-transposase; **Obs**- observed nucleotide differences; **Syn**- synonymous changes; **Nsyn**-non-synonymous changes; **F_{syn}**- synonymous frequency; **F_{nsyn}**- non-synonymous frequency; **Rec**-our method, **ML-** maximum likelihood; **dN:** non-synonymous substitutions per non-synonymous site; **dS:** synonymous substitutions per synonymous site

Table 5. Comparison of dN/dS estimates in *Burkholderia pseudomallei* malleilactone operon

					Simulations					Rec	ML[1]		Parsimony[2]
Gene	**Sites**	**Obs**	**Syn**	**NSyn**	**F_{syn}**	**F_{nsyn}**	**Binomial *P*-value**	**dN**	**dS**	**dN/dS**	**dN/dS**	**LRT p-value (ω=est. vs ω=1)**	**Steps**
adh	1547	18	13	5	0.259	0.741	5.90E-05	0.0044	0.0325	0.134	0.0366	4.03E-12	22
amt	1422	21	12	9	0.241	0.759	1.21E-03	0.0083	0.0350	0.239	0.0717	7.62E-09	34[d]
ddc	1262	19	14	5	0.245	0.755	8.89E-06	0.0052	0.0453	0.116	0.0367	3.45E-15	21
fas	1866	15	11	4	0.242	0.758	8.37E-05	0.0028	0.0244	0.116	0.019	8.40E-20	18
fp	1482	30	18	12	0.270	0.731	1.45E-04	0.0111	0.0451	0.246	0.3209	0.002	34
kar	1092	3	0	3	0.246	0.754	1.00E+00	0.0036	0.0000	<0.001	0.5418[§]	0.616	4
lig	1860	20	4	16	0.248	0.752	7.70E-01	0.0114	0.0087	1.322	0.379	0.032	27
mlp	432	2	1	1	0.242	0.759	4.25E-01	0.0031	0.0096	0.319	0.0320[§]	0.006	2
pks1	8547	343	203	140	0.256	0.744	1.29E-77	0.0220	0.0928	0.237	0.3129	1.78E-15	417[d]
pks2	12504	185	85	100	0.254	0.746	1.29E-09	0.0107	0.0267	0.401	0.1827	1.38E-46	322[d]
mta	900	8	6	2	0.249	0.751	4.12E-03	0.0030	0.0268	0.110	0.0191[§]	5.93E-11	8
Overall	32914	664	367	297	0.253*	0.746*	1.26E-60*						

* weighted; [1]- PAML ; [2]- PHYLIP; [d]-significant difference when compared to **Obs**

[§] standard error exceeds 5 fold the maximum-likelihood estimate

Abbreviations: *adh*-aldehyde dehydrogenase; *amt*-aminotransferase;*ddc*-diaminopimelate decarboxylase; *fas*-fatty acid synthetase; *fp*-fkbh-domain protein; *kar*-ketol acid reductoisomerase; *lig*-ligase; *mlp*- membrane lipoprotein; *pks1*- polyketide synthase 1; *pks2*-polyketide synthase 2; *mta*- malonyl transacylase; **Obs**- observed nucleotide differences; **Syn**- synonymous changes; **Nsyn**-non-synonymous changes; **F_{syn}**- synonymous frequency; **F_{nsyn}**- non-synonymous frequency; **Rec**-our method, **ML**- maximum likelihood; **dN:** non-synonymous substitutions per non-synonymous site; **dS:** synonymous substitutions per synonymous site.

Table 6. Comparison of dN/dS estimates in *Anaplasma marginale* prophage structural genes

					Simulations					Rec	ML[1]		Parsimony[2]
Gene	**Sites**	**Obs**	**Syn**	**NSyn**	**F_{syn}**	**F_{nsyn}**	**Binomial *P*-value**	**dN**	**dS**	**dN/dS**	**dN/dS**	**LRT p-value (ω=est. vs ω=1)**	**Steps**
ptl	1311	67	56	11	0.220	0.780	1.37E-26	0.0107	0.1938	0.0555	0.06759	1.11E-10	67
pts	519	50	18	32	0.233	0.767	3.20E-13	0.0452	0.2646	0.1709	0.3234	3.46E-04	50
Overall	1830	117	88	29	0.224*	0.776*	1.02E-33*						

*weighted; [1]- PAML estimation; [2]- PHYLIP estimation;

Abbreviations: *ptl* –portal protein; *pts*- protease; **Obs**- observed nucleotide differences; **Syn**- synonymous changes; **Nsyn**-non-synonymous changes; **F_{syn}**- synonymous frequency; **F_{nsyn}**- non-synonymous frequency; **Rec**-our method, **ML-** maximum likelihood; **dN:** non-synonymous substitutions per non-synonymous site; **dS:** synonymous substitutions per synonymous site.

Table 7. Comparison of dN/dS estimates in *Anaplasma phagocytophylum* prophage structural genes

					Simulations					Rec	ML[1]		Parsimony[2]
Gene	**Sites**	**Obs**	**Syn**	**NSyn**	**F_{syn}**	**F_{nsyn}**	**Binomial *P*-value**	**dN**	**dS**	**dN/dS**	**dN/dS**	**LRT p-value (ω=est. vs ω=1)**	**Steps**
ptl	1179	18	13	5	0.211	0.789	4.75E-06	0.0053	0.0522	0.1028	0.0877	1.49E-06	19
pts	471	2	2	0	0.217	0.783	4.72E-02	<0.0001	0.0195	<0.0001	0.0001[§]	3.59E-04	2
Overall	1650	20	15	5	0.213*	0.787*	4.26E-07*						

*weighted; [1]- PAML estimation; [2]- PHYLIP estimation; [§] standard error exceeds 5 fold the maximum-likelihood estimate

Abbreviations: *ptl* –portal protein; *pts*- protease; **Obs**- observed nucleotide differences; **Syn**- synonymous changes; **Nsyn**-non-synonymous changes; **F_{syn}**- synonymous frequency; **F_{nsyn}**- non-synonymous frequency; **Rec**-our method, **ML-** maximum likelihood; **dN:** non-synonymous substitutions per non-synonymous site; **dS:** synonymous substitutions per synonymous site

Table 8. Comparison of dN/dS estimates in *Ehrlichia* spp. prophage structural genes

					Simulations					Rec	ML[1]		Parsimony[2]
Gene	**Sites**	**Obs**	**Syn**	**NSyn**	**F_{syn}**	**F_{nsyn}**	**Binomial *P*-value**	**dN**	**dS**	**dN/dS**	**dN/dS**	**LRT p-value (ω=est. vs ω=1)**	**Steps**
ptl	1200	373	87	286	0.1962	0.8037	1.43E-124	0.0902	1.2145	0.0742	0.03095	1.39E-197	431[d]
pts	537	212	88	124	0.1953	0.8046	9.59E-36	0.2036	1.1820	0.1722	0.09573	1.06E-53	246[d]
Overall	1737	585	175	410	0.1959*	0.8040*	6E-154*						

*weighted; [1]- PAML estimation; [2]- PHYLIP estimation; [d]-significant difference when compared to **Obs**

Abbreviations: *ptl* –portal protein; *pts*- protease; **Obs**- observed nucleotide differences; **Syn**- synonymous changes; **Nsyn**-non-synonymous changes; **F_{syn}**- synonymous frequency; **F_{nsyn}**- non-synonymous frequency; **Rec**-our method, **ML-** maximum likelihood; **dN:** non-synonymous substitutions per non-synonymous site; **dS:** synonymous substitutions per synonymous site.

Table 9. Comparison of dN/dS estimates in *Corynebacterium pseudotuberculosis* putative transposase gene

					Simulations					Rec	ML[1]		Parsimony[2]
Gene	**Sites**	**Obs**	**Syn**	**NSyn**	**F_{syn}**	**F_{nsyn}**	**Binomial *P*-value**	**dN**	**dS**	**dN/dS**	**dN/dS**	**LRT p-value (ω=est. vs ω=1)**	**Steps**
tn	1041	24	11	13	0.214	0.785	6.63E-03	0.0159	0.0492	0.3225	0.3798	3.58E-02	24

*weighted; [1]- PAML estimation; [2]- PHYLIP estimation;

Abbreviations: *tn*-transposase; **Obs**- observed nucleotide differences; **Syn**- synonymous changes; **Nsyn**-non-synonymous changes; **F_{syn}**- synonymous frequency; **F_{nsyn}**- non-synonymous frequency; **Rec**-our method, **ML**- maximum likelihood; **dN:** non-synonymous substitutions per non-synonymous site; **dS:** synonymous substitutions per synonymous site.

Parsimony is a conservative approach to infer the minimum number of changes required under the assumption that sequences evolve on a phylogenetic tree. I employed Felsenstein's Dnapars program from the PHYLIP package [102] to calculate these changes for each gene. Comparison of our change counts with Dnapars values illustrates that, while for some genes the values are very close, for other genes there are larger discrepancies that appear unlikely to be due to multiple substitutions per site. An example of such gene is host specificity J gene (*hsJ*) from the *E. coli* prophage (Table 1).

To assess if the excess of parallel changes estimated by Dnapars is significantly higher than expected, under a tree like evolutionary history in the absence of recombination, I used the *evolver* program from the PAML package [11] to evolve *hsJ* gene sequences having the same parameters as estimated by *codeml* program (codon frequencies, transitions to transversions ratio, a single, estimated dN/dS ratio), the same number of sites and the same user tree (scaled up to reflect approximately the same number of nucleotide differences as in our data). I built a distribution of differences between the Dnapars estimates and our count estimates from 100 replicates, where each replicate represents the analysis of a multiple sequence alignment evolved under the specified parameters (supplemental Figure S3). Indeed, I observed no occurrence of homoplasy number seen in the *E. coli* prophage E14 *hsJ* gene compared to what would be expected according to the parametric bootstrap samples generated with *evolve*r ($p<0.01$). I suspected that this significant higher number of homoplasies could be the result of recombination. I therefore used GARD [13] to detect recombination signatures in this gene sequence. Two recombination breakpoints were identified at $p=0.01$ significance level, supporting the hypothesis that the increased number of homoplasies is probably due to recombination (supplemental Table S4).

To test if recombination events could potentially alter tree-based dN/dS estimates, I I ran tree congruence tests on neighboring gene trees and the datasets genes trees. The results of these tests show that there is no significant incongruence between trees. This can be explained by the presence of a vertical signal congruent with that of the host genes for the majority of gene trees within the same gene neighborhood. It is also possible that horizontal gene transfer events may occur frequently within closely related organisms without leading to significant phylogenetic conflict, especially if the genes are short and do not contain many substitutions. (supplemental Figure S5).

2.5. Discussion

This study presents support for the hypothesis that protein coding DNA sequences that are not expressed for function evolve under purifying selection before their pseudogenization, similar to functional genes.

2.5.1. Some genes with apparent detrimental effects upon full expression seem to be vertically inherited

The conserved gene neighborhoods in which I found the analyzed genes point to their clonal inheritance in bacterial lineages. I have found that this assumption was supported by the topology tests results on the majority of the syntenic genes (supplemental Figure S5). Co-evolution of putative prophage genes from *E. coli*, *L. casei*, *Anaplasma* spp., *Ehrlichia* spp and the transposase gene from *C. pseudotuberculosis* and the malleilactone operon genes with the host genomes is strongly supported by the AU and SH tests while some of the PBSX prophage gene phylogenies (*ptl*, *tsp*, *xkdP*, *xkdQ*, *xkdT*) and the *E.coli* transposase (*tn*) gene display significant incongruence with the phylogenies of the neighboring bacterial genes. This incongruence is probably due to recombination events, as suggested by the GARD results.

2.5.2. A counting method for measuring dN/dS in presence of recombination

Measurement of dN/dS ratio as a test for selection is traditionally carried out assuming the examined sequences are sufficiently divergent so that underlying phylogeny is reflecting synonymous and non-synonymous changes that were fixed in the respective populations. One of the limitations of using current methods such as dN/dS estimation by maximum-likelihood [68], [69] or by employing Bayesian approaches [12] that measure the strength and direction of natural selection is the assumption of vertical descent for the analyzed sequences. In this study, I used both tree-dependent and tree-independent method to measure the strength and direction of natural selection in closely related, orthologous and syntenic sequences. I have developed a tree-independent counting method applicable to closely related sequences, possibly prone to recombination.

A comparison of dN/dS estimates between our method and the ML approach shows that both methods yield very similar values. This is true for both closely related (*E. coli* E14, *A. marginale, A. phagocytophilum* and *L. casei* prophages, *E. coli* transposase, *B.pseudomallei* malleilactone operon) or more divergent sequences (*B.subtilis* PBSX and *Ehrlichia* spp. defective prophages), and with variable number of sequences.

Counting methods have been criticized for overestimation of non-synonymous changes and underestimation of synonymous ones as a consequence of the underestimation of substitution events [4]. The parsimony analysesunderline that, indeed, the values for the observed number of nucleotide substitutions might be underestimates of the real values. However, for my method, this potential downfall only reinforces the present findings as the probability for the observed excess in the number of synonymous changes to have occurred under the neutral hypothesis is already very close to 0 (Tables 1-5) .

Additionally, to obtain dN/dS values, I used the probability for synonymous and non-synonymous change estimated from their corresponding frequencies in simulations under the neutral hypothesis. The substitution process in these simulations was random with respect to the sequence space (number of taxa x length of alignment) and biased towards the inherent overall nucleotide frequencies of the analyzed sequences. Therefore, this approach partially compensated for the non-homogeneous rates of nucleotide exchange, codon bias and multiple substitutions per site. All of these parameters are simultaneously estimated in maximum likelihood methods. The employed tree based approaches allow for multiple parallel substitutions per site (homoplasies) but exclude recombination, whereas our tree-independent approach explains these homoplasies through recombination between the sequences. A good approximation for the number of possible multiple hits under the neutral hypothesis is included in the output of our simulation program that records the outcome of every mutation; however, for sequences with among side rate variation the number of homoplasies is expected to be much larger [48].

2.5.3. Purifying selection may be due to a function not yet recognized

Aside from the canonical mobile genetic elements (such as phages or transposable elements), restriction-modification systems [108], bacteriocins [109], [110] or toxin-antitoxin systems [111] are also suggested to represent cases of selfish entities.

The claim that a sequence not under selection for function nevertheless shows signs for purifying selection may seem radical, and alternative explanations have been proposed. For example, a

phage may be found in the same location, not because it was vertically inherited, but because the phage has a strong site preference. Even if a prophage or transposase was vertically inherited, it might have recombined with prophages or recombinases that actually were selected for function as part of their history [112]. Indeed, this scenario cannot be ruled out for our *E. coli* transposase example and *B.subtilis* PBSX genes where I detected presence of putative recombination breakpoints, although homologous recombination with similar non-functional prophages and transposases in the same genome location appears a more likely explanation. Another possible function selected for in the expressed prophage might be to destroy other bacteria, thereby creating new ecological niches for the host. Under this scenario the prophage is no longer propagating as a selfish mobile genetic element, rather the function of lysis would be under group selection, with the DNA and other cell constituents benefiting other members of the population [85]. Recently, the widespread signature of purifying selection detected in over 300 vertically inherited prophage sequences from *E. coli* and *S. enterica* strains (including structural and regulatory modules) was considered evidence for selection by the host for phage-encoded functions [32]. In each individual case it is impossible to exclude the possibility of recruitment by the host for a function not yet recognized by researchers, but seen by natural selection. However, the consistent detection of dN/dS ratios lower than one makes recruitment for function an unlikely explanation. Selection against detrimental effects of mutations cannot be ignored as cause for the observed purifying selection.

2.5.4. B. subtilis PBSX presence may be evolutionary favored

PBSX is a defective prophage, often described as a bacteriocin [113]–[118], which induced expression results in host lysis and the release of phage particles packing about 13 kb of random host DNA, similarly to gene transfer agents. The particles usually induce the lysis of nonlysogenic, non-cognate clones of *Bacillus subtilis* strains [115]. At present, there is no evidence for a *Bacillus* spp.strain naturally cured of PBSX prophage. Its presence in divergent strains under purifying selection, poses interesting evolutionary questions.

The finding that PBSX particles kill sensitive *Bacillus subtilis* strains not carrying the same phage variant and PBSX persistence in so many genomes might suggest a long term evolutionary advantage for the bacterial host strain. Persistence and evolution of bacteriocins in genomes have been discussed often in the context of kin recognition and kin selection, for example under the form of a poison-antidote mechanism [85]. Thus, kin selection may act on PBSX maintenance in *Bacillus* spp. populations. Namely, while the genealogy of sequences included in this study did not include genes that were possibly transcribed for function, the activation of the phage in a related member of the populations may have been recruited to benefit the group of clonal siblings and descendants, and thereby created a selection for function. It could be that this kin-selection pressure for maintenance contributed to the dN/dS ratio in our PBSX dataset being lower than in the other gene families studied. However, the increased divergence of the sequences, and consequently higher fraction of substitutions (*versus* polymorphisms) certainly is another, and possibly larger, factor lowering the dN/dS values for this dataset [41].

Alternatively, the inferred low dN/dS values coupled with observed recombination signatures in PBSX genes can be described as being the result of recombination either with fully functional prophages or with other defective viral sequences.

On the other hand, presence of a toxin-antitoxin system (TA), spoIISAB, adjacent to the PSBX lytic genes on the opposite strand in all *Bacillus* spp. strains examined, might suggest that this defective prophage is maintained in *Bacillus* lineages perhaps through an interaction with this system, similar with other mechanisms that involve TA systems [119], [120]. It is also possible that proximity of this selfish element to PBSX could insure maintenance of both, at long evolutionary time scales, through an intra-genomic conflict dynamics.

The diverse killing spectrum , the ubiquitous presence of PBSX and PBSX-like genes in all *Bacillus* spp. strains and divergence of tail specificity proteins may indicate that PBSX presence is selected because gene loss would trigger loss of immunity against bacteria carrying the same defective phage, similar to addiction modules [110]. In other words, the specific alleles present in each PBSX variant may select for maximizing their selfish transmission as part of the phage genome [36], either vertically, in the clonal bacterial population, or horizontally, by widespread, high-rate gene transfer and by removing potential competitor alleles present in other bacteria. This explanation for PBSX persistence does not exclude its co-option or domestication by *Bacillus* genomes, as bacteriocin, for their own benefit. In this perspective, PBSX function as a bacteriocin can be seen as a consequence rather than the cause of long term persistence and coexistence between the phage genome and bacterial genomes [121]. Certainly, whichever environmental conditions created the premises for this phage's persistence, more than one evolutionary scenario might explain its maintenance in *Bacillus* spp. including , possibly, some that were not discussed here.

2.5.5. Purifying selection may be targeting expressed genes regardless of their functional status

For the each of the cases presented above, selection for function, either at the organism level or at the population level (kin selection), cannot be certainly excluded as a possible cause for the observed level of sequence conservation. Nevertheless, given that most of the dN/dS values are below 0.5 similar to those observed in other studies concerning functional genes [122]–[125], other plausible interpretations should be considered.

The measurements of dN/dS values in phenotypically silent genes reveal that they may generally evolve under purifying selection at a short-time evolutionary scale. The binomial distributions derived from the frequency of synonymous and non-synonymous substitutions in simulations strongly support non-neutral evolution in the analyzed genes. The low dN/dS values illustrate that signatures of purifying selection might not be a good indicator for the functionality of genes. The findings presented here suggest that protein expression level could instead correlate with the strength of purifying selection experienced by genes [126].

A comparison between the dN/dS values for the analyzed genes and with those determined for native genes within the *E. coli* - *S. enterica* clade (dN/dS= 0.05 ± 0.001) reveals a higher dN rate at similar timescale in case of phenotypically silent genes than for known functional genes [33].

In contrast, an earlier study has linked similar dN/dS values to ours, in a genome-wide scan for prophage genes in *E.coli* genomes, with bacteria selecting for the functions encoded by the phage genes [32]. While I do not dispute these observations or the idea of domestication of some phage genes as stand-alone functional elements, I put forward that, alternatively, this process could be one of the consequences of long-term phage gene co-expression and co-existence with bacterial

host proteins within same environment (where interactions are inevitable) rather than being selected by bacteria. It is also possible that another scenario could explain the presence of purifying selection. For example, random protein-protein interactions may evolve quickly and may lead to a dependence of protein folding on a protein not expressed for any other function [127], [128] .

Initial studies by Jacob and Monod on the expression level of the *lac* operon [129] have shown that a very low level of the operon proteins exists even when the inducer, allolactose, is not present in the medium. Furthermore, a study analyzing the expression and fitness data for 3247 of the 4467 protein coding-genes in the *Shewanella oneidensis* MR-1 genome, has shown that, under 15 laboratory growth conditions tested (stress included), some genes with putative detrimental effects on fitness have significantly higher expression compared with other genes. [88]. These findings support the idea that regulation of gene expression via genome-wide and local repression mechanisms in bacteria might not be optimal given certain environmental conditions. This can imply that expression of detrimental genes may not be necessarily functionally relevant.

Significantly, most of the end products of bacterial gene expression share a common, confined environment [130]. A computational study of an *E. coli* cell modelling the Brownian dynamics of only 50 most abundant proteins suggests that macromolecular crowding has a considerable effect on protein folding and interactions. [131]. Within this small intracellular space, the newly synthesized polypeptides need to fold into their native state by overcoming free energy barriers, either independently or with the help of molecular chaperones [132].Therefore, the steric energy and hydrodynamic properties characterizing each folded/unfolded polypeptide may be essential

in driving the selection pressures that each protein-coding gene experiences during their evolutionary lifetime.

I speculate that this selection for molecular fitness may not be necessarily connected to the actual function of the polypeptide itself. The anti-correlation between the expression level of proteins in genomes and their evolutionary rate (measured in dN) reflects the central role of natural selection in shaping protein sequence upon gene expression [133].

Two main hypotheses have been proposed to explain this anti-correlation. Firstly, the protein-misfolding-avoidance hypothesis states that there may be selection pressures acting against error-free and error-induced protein misfolding as a result of the incurring fitness costs. This is supported by the finding of a positive correlation between the expression level and unfolding energy of the polypeptides [126], [134]. Secondly, the protein-misinteraction hypothesis proposes that natural selection also acts against errors in protein-protein interactions. This hypothesis is upheld by the studies in yeast and *E. coli* which suggest a positive correlation between the abundance of proteins and proportion of charged hydrophilic residues on the surface of the proteins [135], [136]. Additionally, differences in sequence conservation between functionally exposed regions of protein and strictly structure-related ones decrease with increasing level of protein expression [137].

Collectively, these findings constitute solid arguments in favor of natural selection targeting not only function but also protein folding and interactions, in the absence of function.

I propose that purifying selection signatures detected in genes not expressed for function can result from the potentially deleterious effects of mutations on protein folding and protein interactions within a crowded cellular environment. Because prophage structural proteins and

assembly-line enzymes tend to have multiple interacting partners (other proteins, DNA or RNA), I speculate that their evolution is constrained even when their expression level is low and functionally irrelevant. Additionally, I argue that this co-option of genes into bacterial genomes for adaptive purposes (termed domestication) might not be pervasive. In the context of bacterial genomes, horizontal gene transfers (HGTs) are responsible for considerable gene gain via mobile genetic elements (transposable elements, phages etc.). A significant proportion of these genes are however, expressed below currently detectable levels [138]. Acquisition of new HGT genes may be less favored by selection which may act on the expressed products, either functional or not [139]. In the absence of selection pressures acting when a gene product encounters the intracellular enviroonment, mutations at DNA level will accumulate, over evolutionary time, in the gene. Given the mutational target size, mutations that inactivate promoters and prevent transcription and translation may be less likely to occur and be fixed than mutations that directly affect the open reading frames. Regulatory mutations that prevent gene transcription and subsequent translation might trigger, at short evolutionary scales, rapid pseudogenization of genes. In comparison, mutations in coding regions bring structural and functional constraints on the expressed products and impact, more or less, the fitness of the cell in a given environmental setting. Gene inactivation, in this case, would probably take much longer or at least, until a promoter mutation occurs.

It has been recently claimed that bacteria often and pervasively domesticate prophages [32], demonstrated by the presence of a large number of prophage genomes within *E. coli* correlated with a number of diverse viral-related genetic elements (*e.g.* secretion systems) present in bacterial genomes and thought to originate via domestication. Gene loss in the case of these

mobile genetic elements is suggested to be explained by the replacement with analogous functions carried by new phages [32].

I submit that effects of these gene gains and losses may be combinatorial and variable in impact rather than anticipatory and purposeful. Their fixation in genomes depends heavily on the fluctuations of the intracellular (*e. g.* competition for binding, level of expression etc.) and extracellular environmental conditions (*e. g.* nutrient abundance, pH, temperature, aerobic/anaerobic, bacterial competition). Consistent with this view, a majority of the small prophages found in *E. coli* and *Salmonella* spp. strains seem to be the result of genome decay of much larger phages [32]. The prevalence of deletions among functional categories of genes with structural roles in host lysis and packaging may suggest that repressed genes incur inactivating mutations such as indels at a higher frequency. I posit this process happens because the residual expression products of these genes interfere with the structure and function of existing host encoded products. Because all macromolecules inadvertently experience nonspecific intermolecular interactions within a cell, the probability that newly made macromolecules entering an established system, as a result of HGT, will encounter very strong purifying selection, can be very high. This may mainly result in rapid gene inactivation and gene loss at short and long evolutionary timescales given the fluctuations of the intracellular and extracellular environmental conditions.

2.6. Conclusions

This work provides evidence that phenotypically silent genes maintained in bacterial lineages with no apparent functional role but some with detrimental effects upon expression are evolving under purifying selection. These results suggest that the mere presence of such selective signatures in the protein-coding DNA sequences cannot solely be used to indicate selection for function. Pseudogenization through mutations of the start codon, frameshift, or nonsense mutations may be the long-term fate of the ORFs that do not contribute towards function. However, along the route towards pseudogenization, the ORF remains subject to purifying selection, as long as it retains a minimal residual expression.

2.7. Supplemental Material

The following supplemental material is included in this reaearch study. Table S1 is a table listing the dN/dS estimates in flanking bacterial genes using our method and maximum-likelihood approach. Table S2 is a table listing the results of the likelihood ratio tests of maximum-likelihood estimates. Figure S3 is a histogram illustrating the inferred number of apparent homoplasies for host specificity J gene from *E. coli* E14 prophage. Table S4 is a table listing the results of the recombination tests. Table S5 is a table listing the results of topology tests conducted on the gene clusters studied and their flanking bacterial genes. Associated file 1 is a table listing the identification data of gene sequences used in the study. Appendix VI.1. is a Perl script for counting nucleotide and amino acid differences in multiple sequence alignments. Appendix VI.2. is a Perl script for random simulation of a given number of mutational events in a multiple sequence alignments.

Table S1. Comparison of dN/dS estimates in bacterial genes flanking analyzed genes

Abbreviations: *idh*-isocitrate dehydrogenase; *rpp*-ribose-phosphate pyrophosphokinase ;*pit*-inorganic phosphate transporter; *panC*-panthotenate synthetase; *luxR*- transcriptional regulator; *bioC*- putative biotin synthesis protein; f*agC*- ATP binding cytoplasmic membrane protein; **Obs**- observed nucleotide differences; **Syn**- synonymous changes; **Nsyn**-non-synonymous changes; $\mathbf{F_{syn}}$- synonymous frequency; $\mathbf{F_{nsyn}}$- non-synonymous frequency; **Rec**-our method, **ML**-maximum likelihood; **dN:** non-synonymous substitutions per non-synonymous site; **dS:** synonymous substitutions per synonymous site.

Gene/Genome	Sites	Obs	Syn	NSyn	Simulations F_{syn}	Simulations F_{nsyn}	dN	dS	Rec dN/dS	ML* dN/dS
idh/ E. coli	1248	40	40	0	0.231	0.769	0.0000	0.1389	<0.0001	0.0178§
rpp/ L.casei	942	11	9	2	0.231	0.769	0.0028	0.0413	0.0668	0.0783
pit/ B. subtilis	1002	222	195	27	0.246	0.754	0.0358	0.7898	0.0453	0.0526
panC/E. coli	852	152	135	17	0.239	0.761	0.0262	0.6626	0.0396	0.0296
luxR/ B. pseudomallei	717	4	3	1	0.229	0.771	0.0018	0.0183	0.0988	0.0001§
bioC/ A. marginale	801	108	67	41	0.220	0.780	0.0656	0.3808	0.1722	0.4415
bioC/ A. phagocytophylum	768	17	16	1	0.218	0.783	0.0017	0.0957	0.0174	0.4327
bioC/ Ehrlichia spp.	765	344	268	176	0.171	0.829	0.2777	2.0487	0.1355	0.1695
fagC/C. pseudotuberculosis	867	21	14	7	0.246	0.754	0.0107	0.0657	0.1628	0.2009

*- PAML; §- standard error exceeds 5 fold the maximum likelihood estimate;

Table S2. Summary of likelihood ratio tests of maximum-likelihood dN/dS estimates

Abbreviations: ***E. coli****:* Tn - transposase; E14: *mcps*- major capsid protein; *mcpsC*- minor capsid C protein; *ltsu*- large terminase subunit; *stsu*- small terminase subunit; *ptl*- portal protein; *mtL*- minor tail protein L; *hsJ*- host specifity protein J; *tfK*- tail fiber protein K; *taI*- tail assembly protein I; *ttmp*- tail tape measure protein; ***L.casei Lp3:*** *ltsu* -large terminase subunit; *ptl* –portal protein; *cps*- capsid protein; *stsu*- small terminase subunit; ***B. subtilis PBSX***: *cps*- capsid protein; *hn*- holin; *lex*- lytic exoenzyme; *ltsu*- large terminase subunit; *ptl*- portal protein; *pts*- protease; *stsu*- small terminase subunit; *tsp*- tail sheath protein; *xkdP*- murein binding protein; *xkdQ*- tail protein; *xkdT*- putative base plate assembly protein; *xkdU*- hypothetical protein; *xlyA*- N-acetylmuramoyl-L-alanine amidase; ***B. pseudomallei***: *adh*-aldehyde dehydrogenase; *amt*- aminotransferase;*ddc*-diaminopimelate decarboxylase; *fas*-fatty acid synthetase; *fp*- fkbh-domain protein;*kar*-ketol acid reductoisomerase; *lig*-ligase; *mlp*- membrane lipoprotein; *pks1*- polyketide synthase 1; *pks2*- polyketide synthase 2; *mta*- malonyl transacylase; ***A. marginale, A. phagocytophilum, Ehrlichia*** spp.***:*** *ptl*- portal protein; *pts*- protease protein; ***C. pseudotuberculosis:*** *tn*- transposase.

	LRT (ω=1 vs ω1=1, ω2<1)
Gene	p-value
B. pseudomallei maleilactone operon	
adh	3.57E-11
atf	7.99E-18
dc	2.56E-15
fas	9.70E-19
fkbh	3.58E-07
kar	2.74E-01
lig	6.20E-08
mp	1.50E-02
*pks1**	N/A
pks2	7.17E-106
ta	4.98E-10
B. subtilis PBSX prophage	
cps	1.55E-116
holin	2.32E-25
lex	3.50E-91
ltsu	8.38E-181
ptl	1.08E-224
pts	2.20E-80
stsu	6.93E-81
tsp	3.68E-197
xkdP	4.09E-99
xkdQ	2.56E-169
xkdT	4.93E-181
xkdU	8.61E-100
xlyA	2.35E-150
L.casei prophage	
cps	6.36E-60
*ptl**	N/A
ltsu	9.77E-66
stsu	4.25E-21

	LRT (ω=1 vs ω1=1, ω2<1)
Gene	p-value
E.coli E14 prophage	
hsJ	1.54E-161
ltsu	6.68E-33
mcpsC	1.84E-17
mcps	3.56E-24
ltsu	5.34E-39
mtL	2.04E-11
mtH	5.34E-39
protease	6.62E-08
ptl	5.11E-22
stsu	2.57E-15
taI	5.75E-12
tfK	2.03E-10
ttmp	1.26E-55
E. coli transposase	
tn	1.42E-82
A.marginale prophage	
ptl	1.88E-19
pts	1.73E-05
A. phagocytophylum prophage	
ptl	9.36E-06
pts	1.71E-03
Ehrlichia sp. prophage	
ptl	1.39E-197
pts	1.06E-53
C. pseudotuberculosis transposase	
tn	8.79E-02

* dN/dS estimated from pairwise comparison

Figure S1. Inferred number of homoplasies for host specificity J (*hsJ*) gene from *E. coli* E14 prophage

Histogram depicting the estimated number of apparent homoplasies detected when comparing parsimony steps with our substitution counts on simulated alignments (n=100) generated by *evolver* using same parameters estimated by *codeml* for host specificity J gene from *E. coli* E14 prophage. Red bar indicates the value of inferred apparent homoplasies in our alignment.

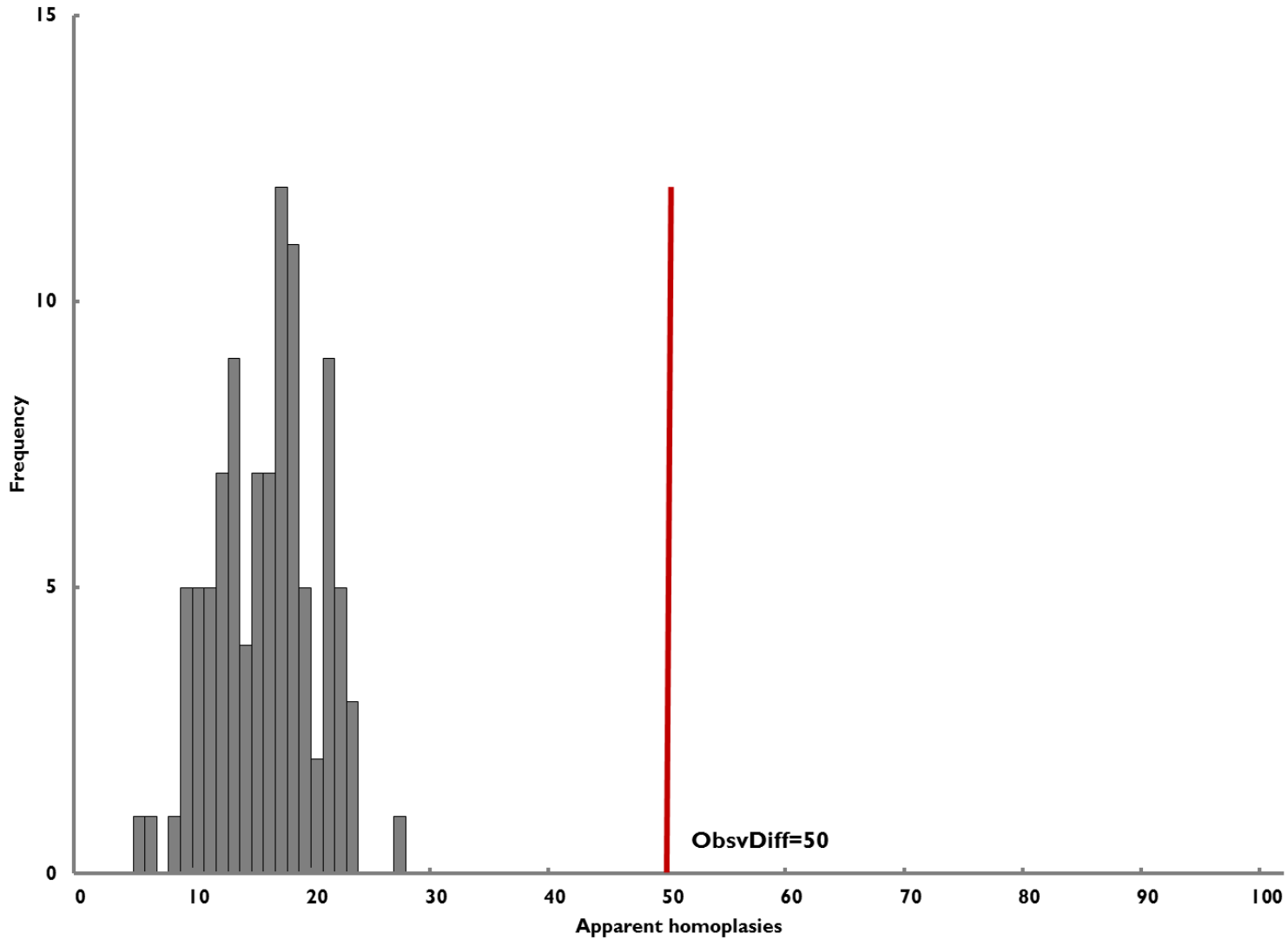

15
10
5
0
Frequency
0
10
20
30
40
50
60
70
80
90
100
Apparent homoplasies
ObsvDiff=50

Table S3. Recombination test results

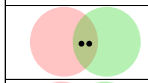
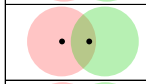
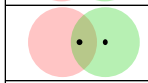
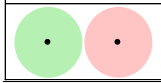
Recombination tests were carried out using HYPHY GARD recombination breakpoint genetic algorithm. The table lists the number of possible recombination breakpoints and their corresponding p-values for the genes shown.

Abbreviations: ***E. coli****: tn* - transposase; E14: *ptl*- portal protein; *hsJ*- host specifity protein J; ***B. subtilis PBSX***: *cps*- capsid protein; *hn*- holin; *lex*- lytic exoenzyme; *ltsu*- large terminase subunit; *ptl*- portal protein; *pts*- protease; *stsu*- small terminase subunit; *tsp*- tail sheath protein; *xkdP*- murein binding protein; *xkdQ*- tail protein; *xkdT*- putative base plate assembly protein; *xkdU*- hypothetical protein; *xlyA*- N-acetylmuramoyl-L-alanine amidase; ***B. pseudomallei***:; *amt*- aminotransferase; *pks1*- polyketide synthase 1; *pks2*- polyketide synthase 2; ***Ehrlichia*** spp.***:*** *ptl*- portal protein; *pts*- protease protein;

		Recombination breakpoints		
		p=0.01	*p=0.05*	*p=0.1*
E. coli	E14			
	ptl	0	0	0
	hsJ	0	0	1
	transposase			
	tn	2	2	2
B. subtilis	PBSX			
	cps	0	0	0
	hn	0	0	0
	lex	0	0	0
	ltsu	1	1	1
	ptl	1	1	1
	pts	0	0	0
	stsu	0	1	1
	tsp	0	0	0
	xkdP	0	1	1
	xkdQ	1	1	1
	xkdT	1	1	1
	xkdU	0	0	0
	xlyA	1	1	1
B. pseudomallei	malleilactone			
	amt	0	0	0
	pks1	1	1	2
	pks2	0	1	2
Ehrlichia sp.	prophage			
	ptl	0	0	0
	pts	0	0	0

Table S4. Tree congruence test results

Abbreviations: ***E. coli****:* Tn - transposase; E14: *mcps*- major capsid protein; *mcpsC*- minor capsid C protein; *ltsu*- large terminase subunit; *stsu*- small terminase subunit; *ptl*- portal protein; *mtL*- minor tail protein L; *hsJ*- host specifity protein J; *tfK*- tail fiber protein K; *taI*- tail assembly protein I; *ttmp*- tail tape measure protein; ***L.casei Lp3:*** *ltsu* -large terminase subunit; *ptl* –portal protein; *cps*- capsid protein; *stsu*- small terminase subunit; ***B. subtilis PBSX***: *cps*- capsid protein; *hn*- holin; *lex*- lytic exoenzyme; *ltsu*- large terminase subunit; *ptl*- portal protein; *pts*- protease; *stsu*- small terminase subunit; *tsp*- tail sheath protein; *xkdP*- murein binding protein; *xkdQ*- tail protein; *xkdT*- putative base plate assembly protein; *xkdU*- hypothetical protein; *xlyA*- N-acetylmuramoyl-L-alanine amidase; ***B. pseudomallei***: *adh*-aldehyde dehydrogenase; *amt*- aminotransferase;*ddc*-diaminopimelate decarboxylase; *fas*-fatty acid synthetase; *fp*- fkbh-domain protein;*kar*-ketol acid reductoisomerase; *lig*-ligase; *mlp*- membrane lipoprotein; *pks1*- polyketide synthase 1; *pks2*- polyketide synthase 2; *mta*- malonyl transacylase; ***A. marginale, A. phagocytophilum, Ehrlichia*** spp.***:*** *ptl*- portal protein; *pts*- protease protein; ***C. pseudotuberculosis:*** *tn*- transposase.

	E.coli											*L.casei*				*B. subtilis*												*B. pseudomallei*											
											Tn																											
	mcps	*mcpsC*	*mtL*	*ptl*	*ltsu*	*stsu*	*taI*	*tfK*	*hsJ*	*tmp*	*tn*	*cps*	*ltsu*	*ptl*	*stsu*	*cps*	*hn*	*lex*	*ltsu*	*ptl*	*pts*	*stsu*	*tsp*	*xkdP*	*xkdQ*	*xkdT*	*xkdU*	*xlyA*	*adh*	*atf*	*dc*	*fas*	*fkbh*	*kar*	*lig*	*mp*	*pks1*	*pks2*	*ta*
	•		•	•	•	•	•	•	•	•		•	•	•	•		•	•	•								•		•	•	•	•	•	•	•	•	•	•
		•														•					•	•					•											
																•					•	•					•											
											•									•			•	•	•	•												

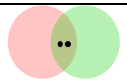
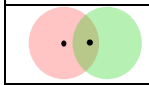
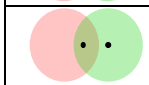
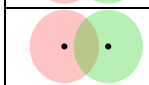
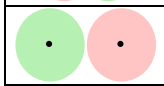
	A. marginale		*A. phagocytophyllum*		*Ehrlichia sp.*		*C. pseudotuberculosis*
							Tn
	ptl	*pts*	*ptl*	*pts*	*ptl*	*pts*	*tn*
			•	•	•	•	•
							
							
	•	•					
							

Figure key

Each circle represents a homogeneous set of topologies including 100 bootstrap replicates and the optimal ML tree (the black dot within it).

Light red- topology space for analyzed gene.

Light green –topology space for the bacterial gene flanking the analyzed syntenic region.

Intersection of the two circles represents either a common set of topologies or topologies that may possess common bipartitions, according to the AU and SH tests cutoff value ($p<0.05$).

III. Chapter 3 – Sequence Conservation and Selection for Function

The following chapter includes research published as an opinion article "Does Sequence Conservation Provide Evidence for Biological Function?" by Seila Omer, Timothy J. Harlow and Johann Peter Gogarten, published online in "Trends in Microbiology" on October 20, 2016 (in press)[95]. The contributions of each author to this chapter are listed in section **VII-Contributions**.

Trends in Microbiology

CellPress

Opinion

Does Sequence Conservation Provide Evidence for Biological Function?

Seila Omer,[1] Timothy J. Harlow,[1] and Johann Peter Gogarten[1,2,*]

Finding a signature of purifying selection in a gene is usually interpreted as evidence for the gene providing a function that is targeted by natural selection. This opinion offers a very different hypothesis: purifying selection may be due to removing harmful mutations from the population, that is, the gene and its encoded protein become harmful after a mutation occurred, possibly because the mutated protein interferes with the translation machinery, or because of toxicity of the misfolded protein. Finding a signature of purifying selection should not automatically be considered proof of the gene's selectable function.

Relationship between Sequence Conservation, Purifying Selection, and Biological Function

Natural selection is one of the main drivers of the adaptation and diversification of organisms. Two types of selection are commonly distinguished: positive selection, also known as directional or diversifying selection, acts on mutations that increase organismal fitness; purifying selection acts on mutations that decrease the organismal fitness. Here we challenge the often made assumption that evidence for purifying selection can be equated to the gene under purifying selection making a positive contribution to organismal fitness. We discuss the possibility of mutations leading to toxic or detrimental products that interfere with normal cellular functions. DNA sequence conservation indicates that natural selection operates against the deleterious effects of allelic variants (also known as purifying selection). A popular approach to detect purifying selection in protein-coding DNA sequences is to infer an excess in the rate of synonymous substitutions (dS) relative to the rate of nonsynonymous (dN) substitutions within a set of homologous and very similar protein-coding sequences (Box 1) [1]. It is generally assumed that the rate of amino acid replacements, measured as a low rate of substitutions that lead to changes at the protein sequence level (usually abbreviated as dN), reflects natural selection through the structural and functional constraints imposed on the protein sequence. That is, if a mutation changes an amino acid, and the protein after mutation does not function, or functions less efficiently, the fitness of organisms carrying the mutated gene generally will be lower, and selection will tend to eliminate the gene from the population. A dN/dS ratio estimate significantly smaller than 1 indicates that some nonsynonymous substitutions were removed by natural selection. Because this type of selection removes mutations from the population it is known as purifying selection. In the rare instance that changes in the amino acid sequence prove to be advantageous and are driven to fixation through the increased fitness of the carrier of the mutated alleles, the rate of nonsynonymous substitutions is higher than the rate of synonymous substitutions. For nearly all protein-coding genes, purifying selection is much stronger than positive selection (also known as directional selection), and the latter is usually restricted to individual sequence positions in a multiple sequence alignment. Thus, when present across closely related taxonomic groups, protein-coding DNA sequences appear nearly identical, that is, they are conserved.

Trends

The current accepted definitions of biological function for a gene assume an evolutionary history shaped by natural selection. This assumption is based on the observation that most genes, characterized genotypically and phenotypically, display in their DNA sequence a considerable excess of synonymous substitutions over what would be expected, if the substitution process were random.

Many genes potentially detrimental for bacterial fitness, lacking functionally relevant expression, share a common evolutionary record with their host organisms through vertical descent. Their DNA sequences present the same type of selective footprints found in neighboring functional genes.

The observation that natural selection operates on both functional and putative nonfunctional genes challenges the default connection between sequence conservation and biological function.

[1]Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Storrs, CT 06269-3125, USA
[2]Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA

*Correspondence: gogarten@uconn.edu (J.P. Gogarten).

Box 1. dN/dS (a.k.a. Ka/Ks or ω) as a Test for Selection

$$dN - \frac{\text{number of nonsynonymous substitutions}}{\text{number of possible nonsynonymous substitutions}}$$

$$dS - \frac{\text{number of synonymous substitutions}}{\text{number of possible synonymous substitutions}}$$

Nonsynonymous (dN) and synonymous (dS) rates of substitutions are defined as the actual number of synonymous and nonsynonymous substitutions, respectively, relative to the number of all possible synonymous or nonsynonymous substitutions. The ratio of the two rates (dN/dS) measures the strength and direction of natural selection experienced by protein-coding DNA sequences. If mutations have no effect on fitness, the nonsynonymous rate is expected to be the same as the synonymous rate (dN/dS = 1). This neutral rate is commonly used as a null hypothesis against which the presence of selection is tested. If organisms with nonsynonymous mutations in a site are less fit than organisms without the mutation, purifying selection will purge them from lineages so that dN/dS <1. If nonsynonymous mutations are favored by selection, they will be fixed at a higher rate than synonymous mutations resulting in dN/dS >1. The latter is a rare occurrence and is known as positive, directional, or diversifying selection. Averaged over all possible amino acid encoding codons, a random mutation is about three times more likely to be nonsynonymous than synonymous. The redundancy of the genetic code is unevenly distributed over the three codon positions (depicted in Figure I). When considering all codons, a change in the 3rd codon position frequently results in synonymous substitutions, whereas changes in the 2nd codon position always are nonsynonymous. Figure I gives the estimated proportion of possible synonymous and nonsynonymous substitutions for the TTT triplet (UUU codon in the messenger RNA).

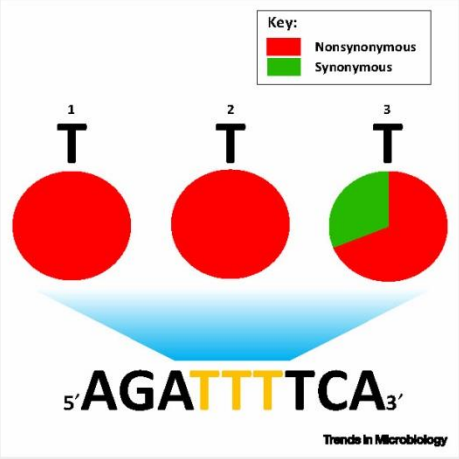


Figure I. Fractions of synonymous and nonsynonymous substitutions for the TTT triplet.

Recent studies have found patterns of purifying selection in bacterial genes with limited distribution for which the positive contribution to the organismal or the gene's fitness was not obvious: ORFans (genes with no recognizable homologs in other genomes) and group-specific genes [2,3], gene transfer agents (GTAs) [4], transposase [5], and prophage genes [6]. Analysis of dN/dS in orthologous genes (i.e., homologs that diverged together with the lineages) spanning the *Escherichia coli* and *Salmonella enterica* clades has found values much lower than 1, interpreted to reflect the selective constraints associated with important, functional roles for cell fitness [3]. A more recent study of group-specific genes within *E. coli* and *Shigella* clades and more widely distributed non-ORFan genes revealed dN values lower than the dS estimates, which led the authors to suggest that most ORFans are, in fact, functional genes [2]. Several studies [7,8] proposed that the observed purifying selection acting on GTA genes in *Rhodobacter capsulatus* and in the genus *Bartonella* argues for functional benefits that GTAs provide to

Table 1. dN/dS Estimates for Genes Not Expressed for Function

Gene/Organism[a]	dN/dS	Refs
ORFans/*Escherichia coli-Salmonella enterica*	0.19 ± 0.030 (average)	[3]
Prophage genes/*E. coli*	0.22 (median)	[6]
orfB IS *6110*/*Mycobacterium tuberculosis*	0.58168	[4]
ORFans/*E. coli, Shigella, Salmonella*	<1[b]	[2]
tn/*E. coli*	0.1491	NCBI Gene ID: 7152325 + 28 syntenic homologs [c]
ltsu/*Lactobacillus casei*	0.0722	NCBI Gene ID: 6406903 + 6 syntenic homologs[c]
stsu/*L. casei*	0.1041	NCBI Gene ID: 6405444 + 6 syntenic homologs[c]
pks1/*Burkholderia pseudomallei*	0.3129	NCBI Gene ID: 4906294 + 17 syntenic homologs [c]

[a]Abbreviations: *tn*, transposase; *ltsu*, prophage large terminase subunit; *stsu*, prophage small terminase subunit; *pks1*, polyketide synthase (part of the malleilactone cryptic operon).
[b]The authors [2] calculated and reported the difference (dN-dS) not the ratio of dN and dS.
[c]Syntenic homologs were retrieved from Integrated Microbial Genomes (IMG) at Joint Genome Institute (JGI) https://img.jgi.doe.gov/ [36].

the host population or to the GTA genes themselves. More recently, the widespread signature of purifying selection detected in over 300 vertically inherited prophage sequences from *E. coli* and *S. enterica* strains (including structural and regulatory modules) was considered evidence for selection by the host or host population for phage-encoded functions [6]. Table 1 summarizes some of the evidence on the ubiquity of the phenomenon of purifying selection acting on genes without an obvious function beneficial to the gene (in case of selfish genetic elements) [9], the host, or the host population (as an altruistic act benefiting relatives known as kin selection) [10,11].

In this Opinion article, we challenge the default connection between sequence conservation and the functional status of the gene. We argue that low-level expression of a gene may be sufficient to generate a selective footprint on the encoding DNA sequence, that is, a dN/dS value smaller than 1. No additional selection related to the gene product's function may be necessary. In case of dN/dS values not much smaller than 1, assigning biological function to conserved genes must involve additional corroborating pieces of evidence linking the genes to particular traits.

Selection for Function: Other Explanations for Purifying Selection

The *E.coli* genome is estimated to contain upwards of 5% of nonfunctional elements (pseudogenes, defective mobile genetic elements). This dispensable fraction varies in bacterial genomes and experiences a rapid turnover across evolutionary timescales, being prone to deletions as a result of natural selection [12–14]. A recent study finds the energetic cost associated with replication of a DNA segment sufficient to trigger the action of natural selection in a large bacterial population [15]. Despite deletions occurring in nonfunctional genes, some genes are found in closely related strains in a repressed state with no apparent benefit for the organism harboring them, similar to the build-up of items in a car trunk (Figure 1) – these genes might fall victim to deletions in the future. A large number of such genes include mobile genetic elements (phages, transposases, etc.) and operons encoding toxic products with significant effects on bacterial fitness upon expression in certain environments [16]. Under stressful conditions, many dormant prophage genes responsible for phage packaging and lysis become expressed, leading to the demise of the bacterial host. Similar fitness effects are observed in case of transposases. If a transposase is functional, the transposition process via a cut-and-paste mechanism often results in mutagenic effects at the DNA level. Because most mutations caused by transposition in coding regions of the host genome are deleterious to host fitness,

Trends in Microbiology

Figure 1. Car Trunk Analogy for How Purifying Selection Can Act on Bacterial Genes. In many cars the trunk accumulates items most of which are selectively nearly neutral, such as old papers or broken CDs; a few items are useful under certain conditions (e.g., spare tires, tennis racquet, toolbox, map). These items correspond to genes that are neutral or under purifying selection, respectively. Functional genes (the useful items) experience strong selective pressures against mutations that perturb protein stability, lead to a loss of function and overall loss of organismal fitness (purifying selection). Consequently, many mutations that change the amino acid sequence are not kept in populations (analogous with removing a broken tennis racquet, now unusable). However, another reason for the presence of purifying selection is that mutations may lead to a product whose presence is detrimental (e.g., protein toxicity) to the organism. In the trunk of a car analogy, a piece of cheese left in the trunk will rot, producing an obnoxious smell, which will lead to its removal from the trunk. Similarly, a sack of fertilizer might turn into an explosive, prompting fast elimination. In this case, purifying selection occurs because the gene product after the mutation is detrimental, not because the gene before mutation was beneficial.

they will be removed by natural selection. The persistence of mobile genetic elements in bacterial lineages by natural selection raises the question of their biological function either for the benefit of the element itself or for the benefit of the host organism. This question is particularly puzzling for selfish genes that are retained in the same genomic location and were passed on vertically from parent to daughter cells. The debate between kin selection *versus* selfish gene hypotheses as an explanation for apparent altruistic behavior at the organismal level is ongoing in the field of microbial evolution [17]. Several studies involving altruism and cooperation have attempted to explain evolution of certain bacterial traits [18]. By contrast, a theoretical study on DNA secretion in bacteria shows that gene-level selection can be responsible for maintaining a gene promoting gene sharing in bacterial populations [19].

For each individual case of dN/dS values smaller than 1 that is discussed in this article one can invoke complicated scenarios on how selection for function could be acting on the gene in question. For example, a phage may be found in the same location, not because it was vertically inherited, but because the phage has a strong site preference. Even a prophage or transposase that was vertically inherited might have recombined with prophages or recombinases that actually were selected for function as part of their history [20]. Persistence and evolution of

prophage and elements of defective phage, bacteriocins, have been discussed in the context of kin recognition and kin selection, for example under the form of a poison–antidote mechanism [18]. The prophage's function might be to destroy other, less related bacteria after entering the lytic part of its life cycle, thereby creating new ecological niches for the host. Under this scenario the prophage remnant is no longer propagating as a phage, rather the function of lysis would be under group selection, with the cell whose prophage has been activated benefiting other members of the population. In addition, it is impossible to exclude the possibility that the gene in question has acquired a function not yet recognized by researchers, but seen by natural selection; for example, a gene expressed in a lineage for a long time will experience random interactions with other cell constituents, and these may evolve into required interactions through constructive neutral evolution [21], resulting in requirement for protein homeostasis [22] that may be the cause for purifying selection. However, the consistent detection of dN/dS ratios lower than 1 make unknown function or group selection an unlikely explanation, especially, if the gene has not traveled in a lineage for long periods of time.

Expressed Genes May Experience Purifying Selection Regardless of Their Functional Status

Recent studies have advanced the idea that the level of expression may be one of the main determinants of sequence conservation. In support of this assertion, highly expressed genes tend to evolve slower than lowly expressed genes [23,24]. However, a low level of gene expression may not necessarily imply functional relevance. A study analyzing the expression and fitness data for 3247 of the 4467 protein-coding genes in the *Shewanella oneidensis* MR-1 genome has shown that, under 15 laboratory growth conditions tested (stress included), some genes with putative detrimental effects on fitness have significantly higher expression than expected when compared with other genes [16]. These findings support the idea that gene expression regulation via genome-wide and local repression mechanisms in bacteria might not be optimal under some environmental conditions.

Furthermore, the expression levels of proteins in genomes have been shown to be anti-correlated with their evolutionary rate (measured in dN) [24]. Two main hypotheses have been proposed to explain this anti-correlation. First, the protein-misfolding-avoidance hypothesis states that there may be selection pressures acting against error-free and error-induced protein misfolding as a result of the incurring fitness costs. This is supported by the finding of a positive correlation between the expression level and unfolding energy of the polypeptides [23,25]. Second, the protein-misinteraction hypothesis proposes that natural selection also acts against errors in protein–protein interactions. This hypothesis is upheld by studies in yeast and *E. coli*, which suggest a positive correlation between the abundance of proteins and the proportion of charged residues on the surface of the proteins [26,27]. Additionally, differences in sequence conservation between functionally exposed regions of proteins and strictly structure-related regions decrease with increasing level of protein expression [28].

Protein synthesis, folding, and interactions occur in a busy, confined environment [29]. A computational study of an *E. coli* cell modelling the Brownian dynamics of only the 50 most abundant proteins suggests that macromolecular crowding has a considerable effect on protein folding and interactions [30]. Within this small intracellular space, the newly synthesized polypeptides need to fold into their native state, either independently or with the help of molecular chaperones. Because the compartment is very small relative to the size and number of the unfolded polypeptides, aggregation in toxic products is possible [31]. The toxicity of these aggregates is manifested as a 'gain-of-toxic-function' through interactions with many non-cognate partners [32]. In the car trunk analogy, a sack of forgotten fertilizer could become unstable and explode, becoming detrimental for the car's integrity.

Key Figure

Mutations Affecting Protein Structure May Be Subject to Purifying Selection.

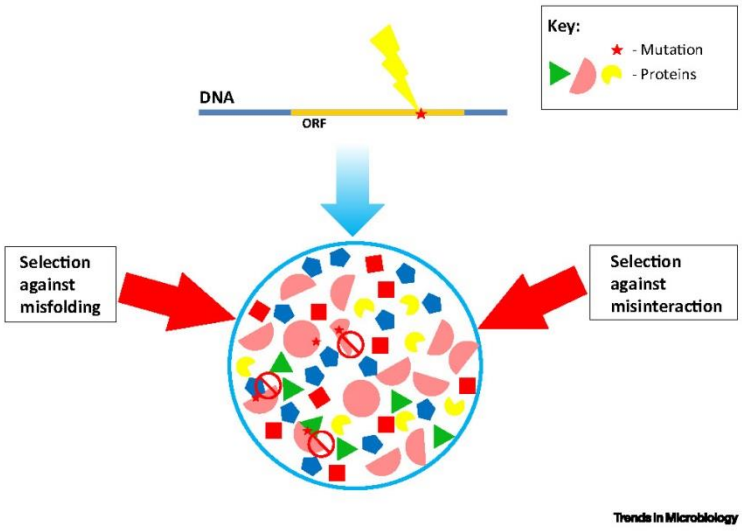


Figure 2. Mutations at the level of DNA may result in changes in the primary structure of proteins. Most changes propagate hierarchically in the newly made macromolecules, to their secondary and tertiary structures, affecting protein stability and interaction with cognate partners. Purifying election acts to remove such mutations, even in cases when protein expression is minimal. Therefore, no selective effects for protein function by the host organism are required to explain signatures of purifying selection in protein-coding DNA sequences.

Mutations can also alter the impact of an encoded gene product by affecting the actual process of protein synthesis. For instance, mutations that create rare codons alter the level of protein expression and ribosome turnover rates by causing ribosome stalling [33]. Furthermore, mutations that alter the secondary structure of an mRNA molecule might impact movement of the ribosomes and translation efficiency [34].

Collectively, these findings constitute arguments for natural selection targeting protein folding and interactions even in the absence of protein function (Figure 2, Key Figure).

We propose that purifying selection signatures detected in genes not expressed for function can result from the potentially deleterious effects of mutations on protein folding and protein interactions. Because prophage structural proteins, transposases, and assembly-line enzymes tend to have multiple interacting partners (other proteins, DNA or RNA), we speculate that their evolution is constrained even when their expression level is low and functionally irrelevant.

Given the mutational target size in bacteria, mutations that inactivate promoters and prevent transcription and translation may be less frequent than mutations that directly affect the open

reading frames (ORFs). Regulatory mutations that prevent transcription can drive rapid inactivation (pseudogenization) of genes not expressed for function. In comparison, mutations in coding regions bring structural and functional constraints on the expressed products, even when the expression level is very low. For example, missense mutations that alter initiation codons or nonsense mutations anywhere within ORFs can result in shorter translation products. These truncated ORFs will still be subject to purifying selection as long as occasional transcriptional and translational events happen.

A comparison between the dN/dS values for the genes we have included in this article and those determined for native genes within the *E. coli*–*S. enterica* clade (dN/dS = 0.05 ± 0.001) reveals a higher dN rate at similar timescales in case of phenotypically silent genes than for known functional genes [3]. By contrast, a study on *E. coli* prophage genes has advanced the idea that moderately low dN/dS values reflect selection by bacteria for the functions encoded by the phage genes [6]. While we do not dispute that domestication of some phage genes as standalone functional elements can occur, we put forward that, alternatively, this process could be one of the consequences of phage gene coexpression and coexistence with bacterial host proteins within same environment (where interactions are inevitable) rather than resulting from a fitness increase of the bacterial population.

We do not consider the mere fact of providing a template for transcription as providing a function; rather, we use the typical biological definition for function that implies a contribution to the fitness of the organism or to the gene itself [35]. Couched in these terms, our hypothesis is that a gene being transcribed and translated at low levels is sufficient to create a signature of purifying selection; a positive contribution of the gene to the gene's, host's, or group's fitness is not necessary.

Concluding Remarks

There is growing evidence that phenotypically silent genes maintained in bacterial lineages with no apparent functional role, but with possible detrimental effects upon expression, are evolving under purifying selection. Therefore, the mere presence of such selective signatures in the protein-coding DNA sequences should not be used as the only argument to indicate function of the encoded protein. Pseudogenization through mutations of the start codon, frameshift, or nonsense mutations are the likely long-term fate of ORFs that do not contribute towards function. However, along the route towards pseudogenization, the ORF may remain subject to purifying selection, as long as it retains a minimal residual expression. A documented dN/dS <1 is necessary but not sufficient to prove biological function for a gene within an organism. In the context of high-throughput sequencing and genome analysis, it is important to develop clear criteria for determining the functional importance from substitution rates. To prove the positive contribution of a gene to the fitness of the organism (or to the gene itself in case of a selfish genetic element), either dN/dS values much smaller than the ones reported for genes without apparent function, or a clear understanding of the encoded product and data that connect the genotype to the phenotype, may be needed (see Outstanding Questions).

Outstanding Questions

How could we design a selection test using known protein structure data which can differentiate between the effect of mutations on structure and effects on function?

What types of evidence, beyond the evolutionary argument, should be included to assign potential biological functions to genes?

What fraction of the bacterial protein-coding genes is 'junk-in-the-trunk' and what fraction is actually under functional constraints?

References

1. Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–736
2. Yu, G. and Stoltzfus, A. (2012) Population diversity of ORFan genes in *Escherichia coli*. *Genome Biol. Evol.* 4, 1176–1187
3. Daubin, V. and Ochman, H. (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* 14, 1036–1042
4. Lang, A.S. *et al.* (2012) Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.* 10, 472–482
5. Thabet, S. *et al.* (2015) Evolutionary trends of the transposase-encoding open reading frames A and B (*orfA* and *orfB*) of the mycobacterial IS6110 insertion sequence. *PLoS ONE* 10, e0130161
6. Bobay, L-M. *et al.* (2014) Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12127–12132
7. Guy, L. *et al.* (2013) A gene transfer agent and a dynamic repertoire of secretion systems hold the keys to the explosive radiation of the emerging pathogen *Bartonella*. *PLoS Genet.* 9, e1003393

8. Lang, A.S. and Beatty, J.T. (2007) Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* 15, 54–62
9. Dawkins, R. (1989) *The Selfish Gene, Oxford.*
10. Hamilton, W.D.D. (1964) The genetical evolution of social behaviour. I. *J. Theor. Biol.* 7, 1–16
11. Hamilton, W.D. (1964) The genetical evolution of social behaviour. II. *J. Theor. Biol.* 7, 17–52
12. Ochman, H. and Davalos, L.M. (2006) The nature and dynamics of bacterial genomes. *Science* 311, 1730–1733
13. Kuo, C.H. and Ochman, H. (2010) The extinction dynamics of bacterial pseudogenes. *PLoS Genet.* 6, e1001050
14. Koskiniemi, S. *et al.* (2012) Selection-driven gene loss in bacteria. *PLoS Genet.* 8, e1002787
15. Lynch, M. and Marinov, G.K. (2015) The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci.* 112, 201514974
16. Price, M.N. *et al.* (2013) Indirect and suboptimal control of gene expression is widespread in bacteria. *Mol. Syst. Biol.* 9, 660
17. Olendzenski, L. and Gogarten, J.P. (2009) Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer. *Ann. N.Y. Acad. Sci.* 1178, 137–145
18. Strassmann, J.E. *et al.* (2011) Kin discrimination and cooperation in microbes. *Annu. Rev. Microbiol.* 65, 349–367
19. Draghi, J.a. and Turner, P.E. (2006) DNA secretion and gene-level selection in bacteria. *Microbiol. Read. Engl.* 152, 2683–2688
20. Castillo-Ramírez, S. *et al.* (2011) The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathog.* 7, e1002129
21. Gray, M.W. *et al.* (2010) Irremediable complexity? *Science* 330, 920–921
22. Bershtein, S. *et al.* (2015) Protein homeostasis imposes a barrier on functional integration of horizontally transferred genes in bacteria. *PLoS Genet.* 11, e1005612
23. Drummond, D.A. *et al.* (2005) Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14338–14343
24. Pál, C. *et al.* (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158, 927–931
25. Geller-Samarotte, K.a. *et al.* (2011) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 108, 680–685
26. Yang, J-R. *et al.* (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* 109, E831–E840
27. Plata, G. *et al.* (2010) The rate of the molecular clock and the cost of gratuitous protein synthesis. *Genome Biol.* 11, R98
28. Eames, M. and Kortemme, T. (2007) Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. *Struct. Lond. Engl.* 15, 1442–1451
29. Zhou, H-X. *et al.* (2008) Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* 37, 375–397
30. McGuffee, S.R. and Elcock, A.H. (2010) Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.* 6, e1000694
31. Hartl, F.U. *et al.* (2011) Molecular chaperones in protein folding and proteostasis. *Nature* 475, 324–332
32. Oikawa, T. *et al.* (2016) α-Synuclein fibrils exhibit gain-of-toxic function, promoting tau aggregation and inhibiting microtubule assembly. *J. Biol. Chem.* 291, 15046–15056
33. Keiler, K.C. (2015) Mechanisms of ribosome rescue in bacteria. *Nat. Rev. Microbiol.* 13, 285–297
34. Wachter, A. (2014) Gene regulation by structured mRNA elements. *Trends Genet.* 30, 172–181
35. Doolittle, W.F. *et al.* (2014) Distinguishing between "function" and "effect" in genome biology. *Genome Biol. Evol.* 6, 1234–1237
36. Markowitz, V.M. *et al.* (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* 42, D560–567

IV. Chapter 4 –Robustness and Coding Potential in 3’ Untranslated Regions of Highly Expressed Genes

4.1.Introduction

The latest research on protein synthesis in organisms spanning species from all domains of life has revealed translationis carried out with a high error rate. The three stop codons, UAA, UGA and UAG typically terminate the process of translation by the ribosome through recruitment of release factors at the site of translation. Sometimes the ribosome fails to recognize the stop signal and incorporates an amino acid instead, continuing translation until it encounters a new stop codon. This phenomenon, called nonsense suppression, may occur in case of UGA recoding for tryptophan, a stop suppressor tRNA for UGA coding selenocysteine [140] or a stop suppressor tRNA for UAG coding pyrrolysine [141]. A similar outcome of bypassing the existing stop codon is obtained when insertions and deletions in DNA sequence upstream in the coding sequence produce a frameshift. A few studies suggested that off-frame, ambush or hidden stop codons (OSCs) present in +1 and +2 reading frames of genes are frequently encountered because of the cost associated with expressing a longer sequence which could be potentially cytotoxic[142], [143]. The translational error rate in *Bacillus subtilis*, a fast-growing organism, isestimated to be 1 in every 200 codons [58] with variations between different physiological states (stationary phase, stress response and exponential growth). Among the most often encountered errors are amino acid misincorporation, recoding, frameshifting and stop codon read-through. Expression of GFP fusion proteins in *B. subtilis* [58] and mass spectrometry, nucleotide k-mer composition, synonymous SNP bias, periodicity of secondary structure pairing frequency and GFP fusion protein expression in *Drosophila melanogaster* [144] have revealed considerable stop codon readthrough. The high error rate detected in *B. subtilis* suggests there are

molecular mechanisms which counteract the deleterious effects of DNA mutations and mistranslation. Immediate mechanisms involved ribosome stalling and ribosome rescue [145].

Transfer-messenger RNA-ssrA and ArfA systems rescue bacterial ribosomes stalled on messenger RNA molecules lacking a stop codon [146]–[149]. The tmRNA-ssrA system tags nascent peptides for destruction by ClpXP and other proteases. In eukaryotes, ribosome profiling data has revealed that Dom34 rescues ribosomes stalled in 3' untranslated regions of many mRNA molecules. Additionally, recent research on translation termination has unveiled the physical presence of ribosomes in 3' untranslated regions of more than 10% of transcriptionally active yeast genes [150].

In a theoretical study on the emergence of translational robustness in organisms, Rajon and Masel put forth the hypothesis thatin large populations ($>10^8$ individuals), protein sequence robustness to the effects that minimizes the errors might be the local solution to the problem of expressing potentially deleterious sequences [66]. In contrast, in small populations ($<10^3$ individuals), protein sequences may display low robustness as result of increased rate of genetic drift[151], and, according to Rajon and Masel's hypothesis [66], the detrimental effect of errors is minimized through selection favoring a low error rate.. In this context, robustness is defined as a system's (protein sequence) persistence to external and internal perturbations (errors and mutations) while retaining a potential for change (evolvability) [152], [153]. Overall, these findings suggest that failure of recognizing translational stop signal might constitute an important source of phenotypic variation as result of natural selection in populations., and that, at least in

organisms that live in large populations, the sequence encoded by the 3'UTR following readthrough might be under selection to minimize damage following translation errors.

One of the areas of great interest in evolutionary biology involves the mechanisms underlying gene innovation and their consequences for genome evolution. One source of gene innovation is represented by co-option (inclusion) of the non-coding regions into the coding frames of existing genes [154]. An example of co-option has been suggested to occur in many organisms from yeast to mammals through incorporation of 3' untranslated regions due to stop codon readthrough [144], [150], [155]–[158]. This conversion of non-coding sequences into coding regions has been suggested to contribute to innovation of phenotypic variation and consequently, the evolvability of genomes [57], [66]. A similar suggestion regarding new gene evolution through co-option of 3' UTRs has been proposed for genes in many bacterial species [159]. The present research attempts to address the question whether stop codon readthrough leads to conservation of regions immediately following the stop codons. High translational error rate including read-through in bacteria has been shown to occur with variations throughout their life cycle [58]. Because there is a considerable fitness cost associated with expressing potentially deleterious DNA sequences, selection acts to remove most mutations, including stop codon readthrough, whenever the newly recruited regions become expressed. The main three evolutionary forces directing the protein sequence evolution are natural selection, mutation and genetic drift. The tradeoff between maintaining protein function and preserving the capacity for change of the coding sequence is primarily influenced byof the action of natural selection [160]. The strength of this selection varies with expression level, effective population size and mutation rates, between free living organisms (e.g. *E .coli, Haemophilus influenzae, Vibrio cholerae, Clostridium perfringens, B.*

subtilis) and endosymbionts (*Buchnera aphidicola, Wolbachia* spp.)). The estimated values for the population sizes in bacteria range from approximately 10^3 individuals for *Buchnera* spp. [161] to $5x10^7$ for *E. coli* populations [162] and 10^{11} individuals for *Prochlorococcus* spp. [163]. Given the readthrough error rates, the probability of these mutations occurring is higher for highly expressed genes (HEG) due to increased number of translational events. I hypothesize that the signature of natural selection affecting elongating sequences may be detected in 3' UTR sequences in bacteria. I comparatively measured the evolutionary rates of highly and lowly expressed genes (LEG) in 62 *E. coli* genomes and their corresponding 3' UTR sequences. I found similar signatures of evolutionary processes affecting both the open reading frames and the 3'UTR sequences of highly expressed genes. Additionally, bootstrapping analyses have shown these rates are statistically significant. 3mer decomposition analysis of 3' UTR sequences has revealed significant deviations from expectation under the null hypothesis of randomDNA sequences of same nucleotide content.. The deviation from the expectation under the null hypothesis is larger for HEG than for LEG, indicating the presence of selection in 3' UTR sequences.

Collectively, these findings suggest that 3' UTR sequences of HEG may be co-opted in the translation products with potentially neutral effects on the fitness of *E. coli* cells.

4.2. Materials and Methods

4.2.1. Sequences and alignments

Nucleotide and amino acid sequences for 62 *E. coli* fully sequenced reference genomes used in this study were downloaded from the NCBI Genome database (http://www.ncbi.nlm.nih.gov) [164]. The list of genomes used in this study are included in Appendix 7.4.. A list of 253 putative highly expressed genes (HEG) for *E. coli* K12 was downloaded from Highly Expressed Gene Database (HEG-DB) [165] and were included in the initial highly expressed genes dataset. I used the Codon Adaptation Index (CAI) calculator (CAIcal_ECAI_v1.4.pl) implemented by Puigbo *et al.* [166] to calculate CAI for all protein-coding sequences from *E. coli* K12 MG1655. 399 genes with the lowest CAI values were selected and included in the initial putative lowly expressed genes (LEG) dataset. Genome feature files were used to extract non-overlapping 3' untranslated regions (Figure S2). I performed amino acid Usearch [167] reciprocal best match searches against *E. coli K12* with a cutoff E-value of 1e-10 to assemble orthologous gene datasets with representation in all 62 genomes included in this study. Dataset assembly at this stage yielded 392 lowly expressed gene datasets and 253 highly expressed gene datasets. Genome coordinates for protein-coding and RNA genes were used to extract non-overlapping 3' untranslated region sequences (UTRs) containing the first 30 bases after the stop codon for all genomes. Genome- specific 3' UTR databases including these sequences were created. This step yielded 3' UTR sequences for 141 HEG and 223 LEG in *E. coli K12* genome. The gene identification numbers for the HEG and LEG orthologous ORF sets were used to assemble orthologous HEG and LEG 3' UTR datasets from the existing 3' UTR databases. I carried out clustering with UClust [167] of these orthologous 3' UTR datasets with a threshold of 75%

nucleotide identity, selecting only single-clustered datasets with representation in all 62 genomes. This filtering step makes 3' UTR sequences amenable to downstream analyses. Following this step, I were left with ORF and corresponding 3' UTR datasets for 62 LEG and 59 HEG which entered the phylogenetic analysis.

4.2.2. Alignments, putative bootstrapping and tree building

Nucleotide codon-based alignments for the open reading frame sequences were built using MACSE [168]. Following the clustering step, no alignment of 3' UTR datasets was necessary. I performed concatenation of ORF and corresponding 3' UTR nucleotide alignments for the 59 HEG and 62 LEG using FASconCAT-G [169]. I used custom in-house Perl scripts to generate 1000 putative bootstrap samples, 30 bases in length, from each of the 121 concatenated alignments. For each gene, the bootstrapping process assembled 30 random positions within the concatenated multiple sequence alignment in a randomized order. To measure phylogenetic distance (tree length), phylogenetic trees were constructed from nucleotide alignments (ORF and 3' UTR datasets) and bootstrap samples by a maximum-likelihood method using RAxML [103], using default settings under a general time reversible model (GTR), with 4 gamma rate categories, fraction of invariant sites estimated from the data and 100 bootstrap replicates. Additionally, we performed parsimony analysis on the 3' UTR datasets and bootstrap samples using PAUP [170] under default settings, heuristic search with random addition of sequences, 10 replicates and a maximum of 5000 trees saved.

4.2.3. N-mer decomposition and analysis

I generated 1000 random nucleotide datasets based on the nucleotide frequencies determined for 141 HEG 3' UTR sequences and 141 LEG 3' UTR sequences using the random DNA sequence generator GenRGenS [171]. I decomposed each random DNA sequence within a dataset in constituting n-mers (where n=2, 3) and computed counts for each 2mer (e.g. from sequence AGTCTA I get 5 2mers AG, GT,TC,CT,TA) and for each 3mer (e.g. from sequence AGTCTA I get 4 3mers AGT, GTC, TCT, CTA). In addition, I calculated the expected probability of each 3mer from individual base frequencies and the conditional probability of each 3mer from dimer frequencies. The probability of a 3mer from individual base frequencies is calculated by multiplying the actual frequencies of the bases in each position. The expected number of counts of a 3mer is determined by the expected probability multiplied by the total number of 3mers. The conditional probability of a 3mer ABC (where A, B and C each specify one of the four nucleotides) is calculated as the product between the expected probability of a 2mer AB (P(AB)) and the conditional probability of the other composing 2mer, BC, given the 2mer AB that occupies already the second position within the 3mer (P(BC|AB)). The expected probability of each 2mer is the frequency of that 2mer in the 141 3' UT sequences.

4.2.4. Statistical modeling

Under a scenario in which I expect to see a given number of events occurring, I can describe the p-value (P) as the cumulative probability of seeing the number of observed events (k) or more given a certain number of trials (n) and given the probability of observing a desired outcome (p) (Equation 2).

$P(k|n,p) = 1 - \sum_{i=0}^{\lfloor k-1 \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$ (Equation 2) for $k = 0, 1, 2, \ldots, n$,

Where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Binomial P-values determined for the 3mer decomposition analysis were submitted to a Holm-Bonferroni correction [172].

4.3. Results

4.3.1. Measurement of evolutionary rates in HEG and LEG ORF and 3' UTR

The selective pressure operating on 3' UTR sequences varies with the translational error rate. I argue that purifying selection acts upon the 3' UTR-based amino acid sequences because mistranslation events produce elongated protein products with organismal fitness effects. These elongated gene products are likely to display both genotypic and phenotypic robustness as result of selective pressures. I hypothesize that 3' UTR DNA sequences are optimized by natural selection for phenotypic robustness. If highly expressed genes encounter natural selection in the bacterial cell more often as result of increased translation rate when compared with lowly expressed genes, then probability of translation errors via frameshift and readthrough is also likely to increase. As a result, I I would expect to see similar signatures of selection both in the ORF and 3' UTR of highly expressed genes. As the highly expression genes experience more translational events than lowly expressed genes, I expect differences in the strength of selection affecting 3' UTR sequences.

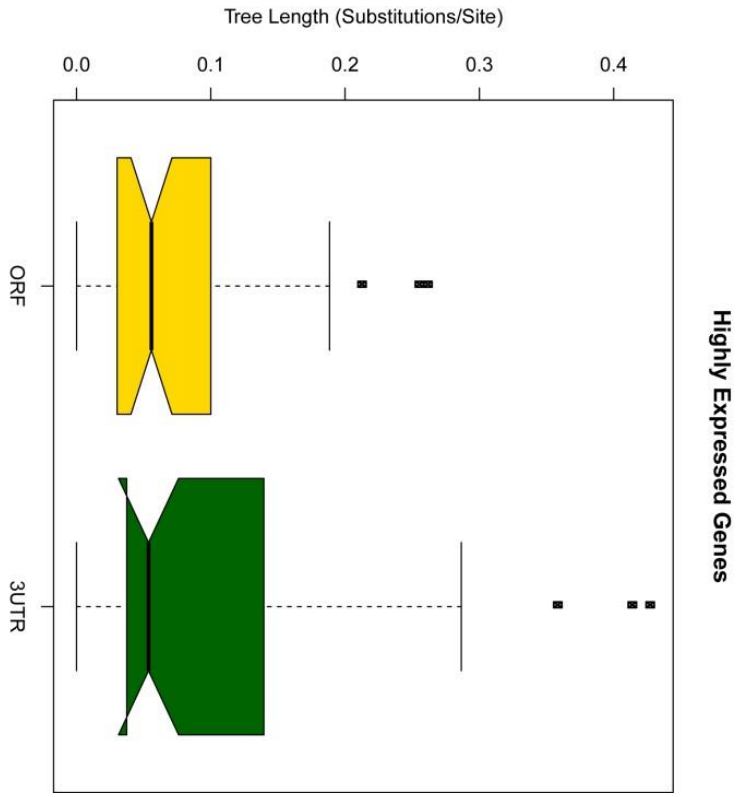
In order to evaluate the amount of selective pressure operating in 3' UTR regions of highly expressed genes, I initially measured the evolutionary rates of these sequences using the maximum likelihood method implemented in RAxML. I included in this phylogenetic analysis 59 HEG (ORF and 3' UTR) and 62 LEG (ORF and 3' UTR). I used tree length expressed in substitutions per site as a tool to quantify the amount of evolution each sequence has experienced. The results of these analyses are shown in Figure 4. While the tree length medians for the HEG ORF and HEG 3' UTR are apparently similar, in contrast there is a slight difference in the tree length medians for the LEG ORF and LEG 3' UTR.

To assess whether the differences observed (or lack thereof) is attributable potentially to a signature of natural selection operating in the 3' UTR sequences, I carried out non-parametric statistical tests to evaluate the significance of these comparisons. Table 10 summarizes the results of the Kolmogorov-Smirnov and Wilcoxon Rank Sum tests. The comparison between the distribution of tree lengths for HEG ORF and the distribution for HEG 3' UTR yields a KS p-value of 0.07216. In contrast, the difference between LEG ORF tree length distribution and LEG 3' UTR tree length distribution has a p-value of 0.0261. Under the assumption that HEG ORF evolve at a different rate when compared with LEG ORF the p-value of the comparison is 0.00054. Surprisingly, the Wilcoxon Rank sum test produced p-values above the significance value (0.05) in contrast with KS test results for the ORF versus 3' UTR comparisons (HEG and LEG). I investigated the discrepancy by performing tests for the homogeneity of variances (a common assumption made with statistical tests). The results of 3 tests (Levene's, valid for normal distributions, Bartlett's and Fligner-Killeen's for non-normal distributions) are presented in Table S5.

In order to determine whether the length of the analyzed 3’ UTR (30 bases) influences the outcome of the tree length measurement, I developed a bootstrapping method. I generated 30 base long 1000 bootstraps from the concatenated multiple sequence alignments of ORF and 3’ UTR regions for each HEG and LEG. I determined their tree

Figure 4. Distributions of RAxML tree length values for HEG and LEG ORF and 3' UTR

Coding (ORF) and corresponding non-coding sequences (3'UTR) for 59 HEG and 62 LEG genes were analyzed by maximum likelihood and tree length values were plotted using R package ggplot2.


Tree Length (Substitutions/Site)
0.0
0.1
0.2
0.3
0.4
ORF
3UTR
Highly Expressed Genes

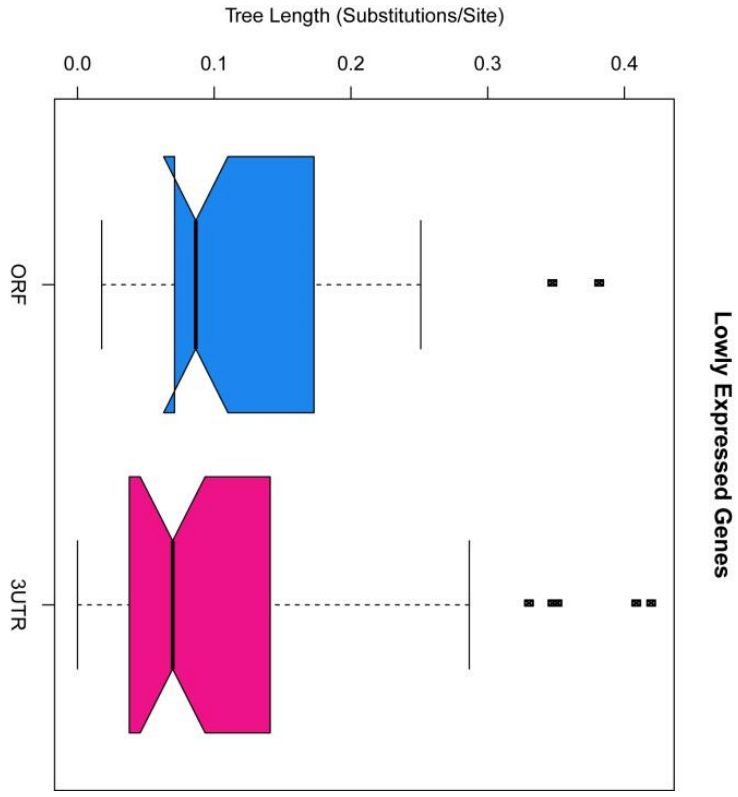

Tree Length (Substitutions/Site)
0.0
0.1
0.2
0.3
0.4
ORF
3UTR
Lowly Expressed Genes

Table 10. Statistical analysis on the tree length values measured by maximum likelihood

Tree Length Value Comparison	Kolmogorov-Smirnov		Wilcoxon Rank Sum	
	D	***P-value***	***W***	***P-value***
HEG ORF vs 3'UTR	0.23729	0.07216	1620	0.5184
LEG ORF vs 3' UTR	0.27119	0.02610	1741	0.3670
LEG ORF vs HEG ORF	0.37288	0.00054	1062	0.0002
LEG 3' UTR vs HEG 3' UTR	0.20339	0.17410	1417	0.08211

Abbreviations: LEG- Lowly Expressed Genes, HEG- Highly Expressed Genes, ORF- Open Reading Frame, 3' UTR- 3' Untranslated Region

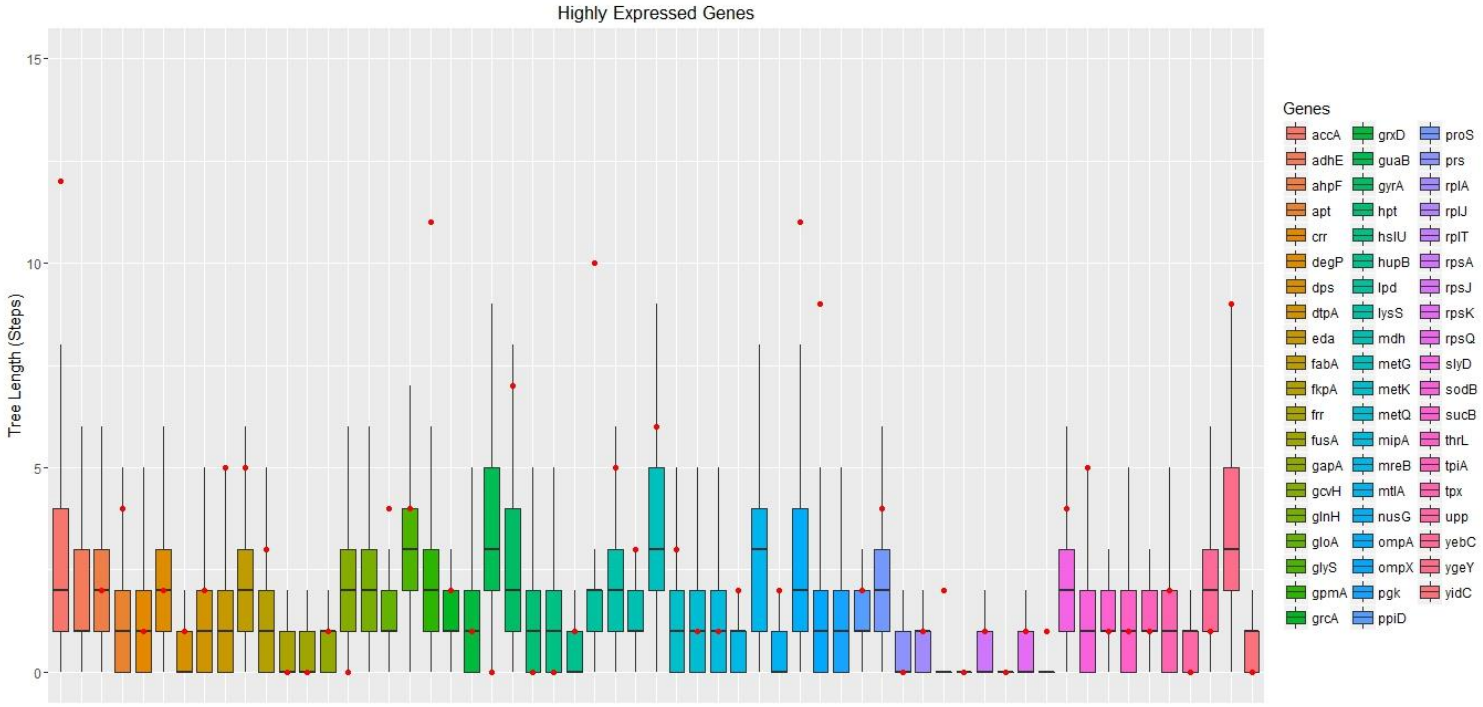
lengths by two methods: maximum likelihood and parsimony. Firstly, I measured the tree lengths expressed in substitutions per site using maximum likelihood and I compared them with the values measured for 3' UTR regions. The results of this comparison are shown in the Figure 5. The distributions of tree length values for bootstrap samples across HEG and LEG genes show significant variation. I found 15 of 59 (25.4%) HEG 3'UTR tree length values grouping outside 95% of bootstrap tree length value distribution and 23 of 62 (37.1%) LEG 3'UTR tree length values grouping outside 95% of bootstrap data. In this case, I calculated the binomial probability of observing 15 out of 59 HEG 3'UTR tree length values given the probability of 0.37 (according to the null hypothesis of LEG 3' UTR not encountering selection as consequence of readthrough), to be 0.019 below 0.05 significance level.

Secondly, I carried out measurement of tree lengths of HEG and LEG 3' UTR and their corresponding bootstrap datasets, expressed as minimum number of steps (changes) describing the phylogenetic relationships among the 62 taxa, using the parsimony method implemented in Paup. The results of the parsimony analysis are illustrated in the Figure 6. I encountered 20 of 59 (33.9%) HEG 3'UTR tree length values outside 95% of bootstrap tree length values and 24 of 62 (38.7%) LEG 3'UTR tree length values grouping outside 95% of bootstrap tree length data. I calculated the binomial probability of observing 20 out of 59 HEG 3'UTR tree length values given the probability of 0.387 (according to the null hypothesis of LEG 3' UTR not encountering selection), to be 0.081 above 0.05 significance level.

Figure 5. Distributions of tree lengths (substitutions/site) using maximum likelihood analysis of evolutionary rates for putative bootstrap replicates

A. Highly Expressed Genes and **B**. Lowly Expressed Genes. A number of 1000 bootstrapped sequences (30 bases) were generated from concatenated sequences comprising open reading frames and their adjacent 3' untranslated regions (30 bases long) for 59 highly expressed genes and 62 lowly expressed genes. Tree lengths for these sequences and (30 bases) were determined used RAxML. The red dots on the graphs show tree length values of the actual 3' UTR sequences.

A.


Highly Expressed Genes
Tree Length (Steps)
15
10
5
0
Genes
accA
adhE
ahpF
apt
crr
degP
dps
dtpA
eda
fabA
fkpA
frr
fusA
gapA
gcvH
glnH
gloA
glyS
gpmA
grcA
grxD
guaB
gyrA
hpt
hslU
hupB
lpd
lysS
mdh
metG
metK
metQ
mipA
mreB
mtlA
nusG
ompA
ompX
pgk
ppiD
proS
prs
rplA
rplJ
rplT
rpsA
rpsJ
rpsK
rpsQ
slyD
sodB
sucB
thrL
tpiA
tpx
upp
yebC
ygeY
yidC

B.

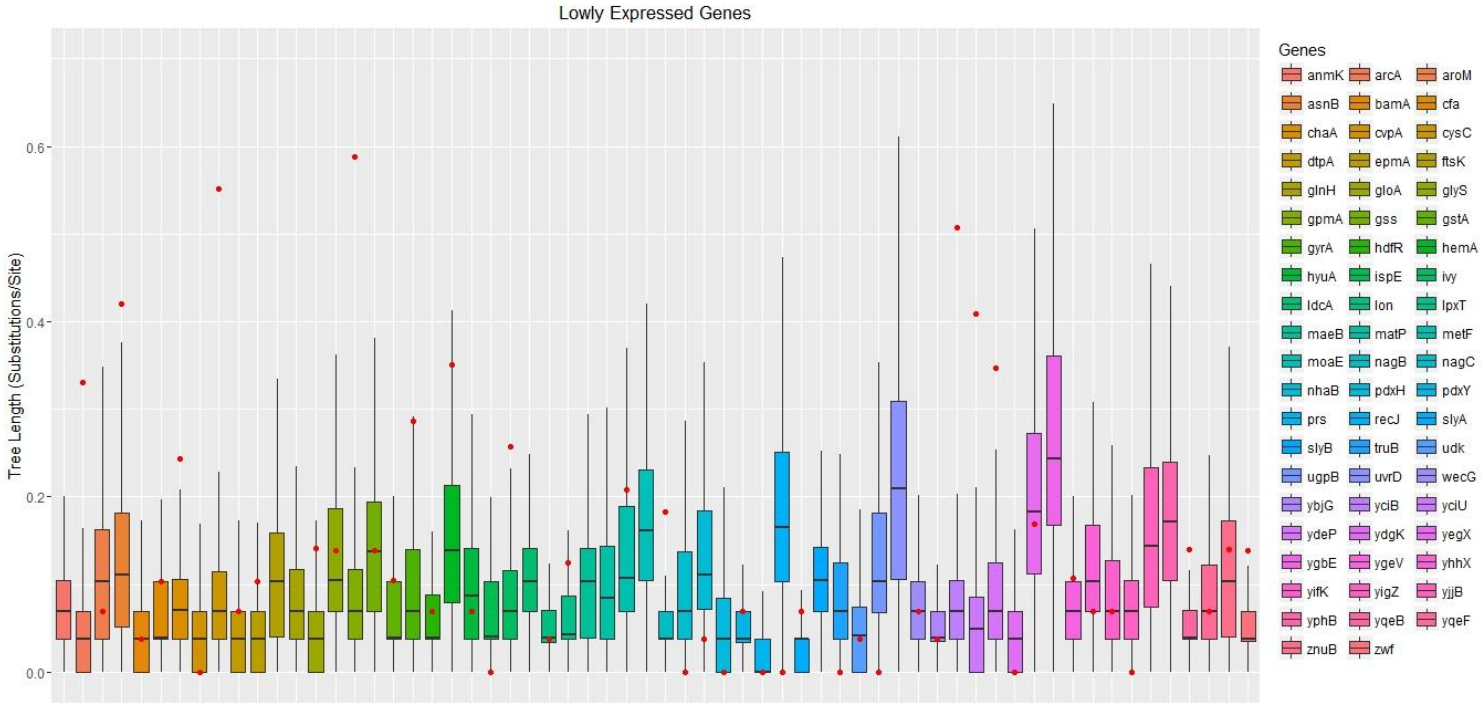
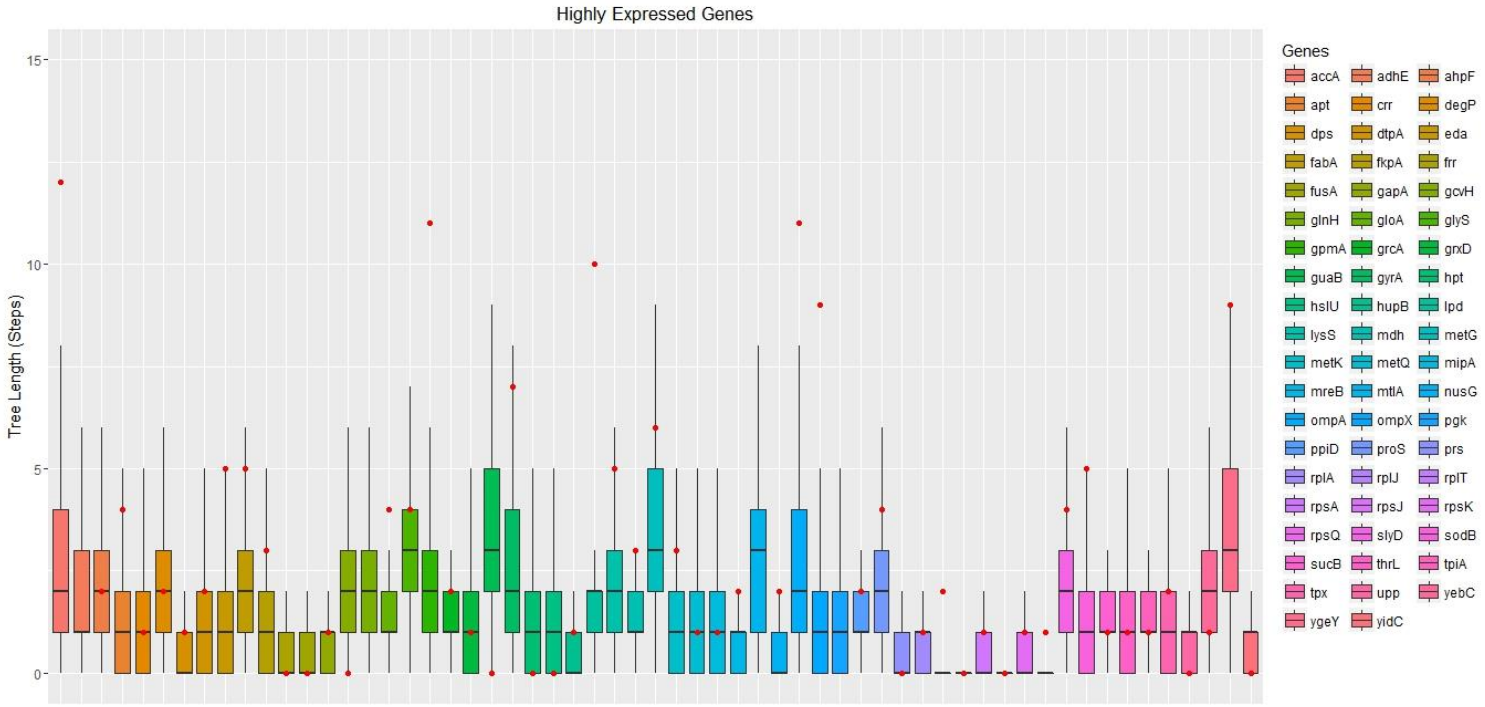

Lowly Expressed Genes
Tree Length (Substitutions/Site)
0.0
0.2
0.4
0.6
Genes
anmK arcA aroM
asnB bamA cfa
chaA cvpA cysC
dtpA epmA ftsK
glnH gloA glyS
gpmA gss gstA
gyrA hdfR hemA
hyuA ispE ivy
ldcA lon lpxT
maeB matP metF
moaE nagB nagC
nhaB pdxH pdxY
prs recJ slyA
slyB truB udk
ugpB uvrD wecG
ybjG yciB yciU
ydeP ydgK yegX
ygbE ygeV yhhX
yifK yigZ yjjB
yphB yqeB yqeF
znuB zwf

Figure 6. Distributions of tree lengths (steps) using parsimony analysis of evolutionary rates for putative bootstrap replicates

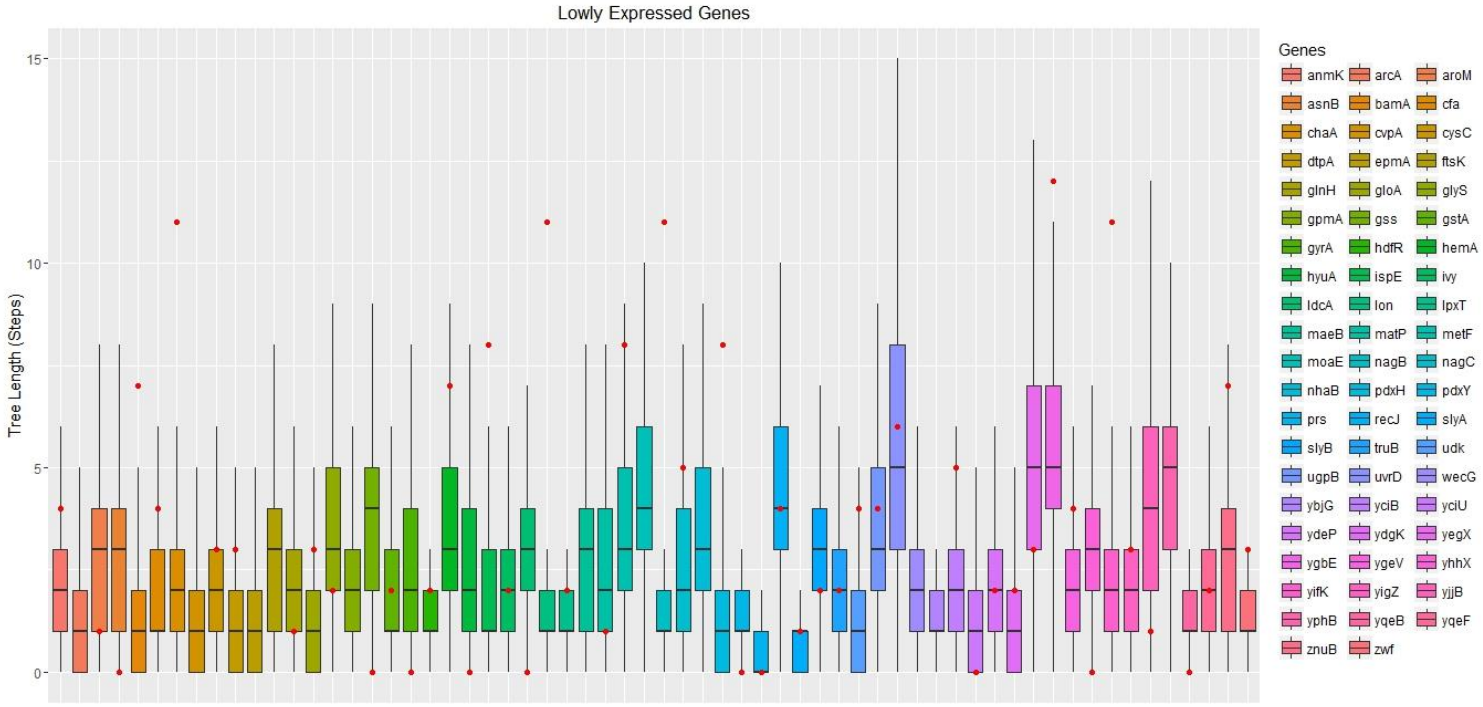
A. Highly Expressed Genes and **B**. Lowly Expressed Genes. A number of 1000 bootstrapped sequences (30 bases) were generated from concatenated sequences comprising open reading frames and their adjacent 3′ untranslated regions (30 bases long) for 59 highly expressed genes and 62 lowly expressed genes. Tree lengths for these sequences and (30 bases) were determined using **Paup***. The red dots on the graphs show tree length values of the actual 3′ UTR sequences corresponding to the genes shown.

A.


Highly Expressed Genes
Tree Length (Steps)
15
10
5
0
Genes
accA
adhE
ahpF
apt
crr
degP
dps
dtpA
eda
fabA
fkpA
frr
fusA
gapA
gcvH
glnH
gloA
glyS
gpmA
grcA
gnxD
guaB
gyrA
hpt
hslU
hupB
lpd
lysS
mdh
metG
metK
metQ
mipA
mreB
mtlA
nusG
ompA
ompX
pgk
ppiD
proS
prs
rplA
rplJ
rplT
rpsA
rpsJ
rpsK
rpsQ
slyD
sodB
sucB
thrL
tpiA
tpx
upp
yebC
ygeY
yidC

97

B.


Lowly Expressed Genes
Tree Length (Steps)
15
10
5
0
Genes
anmK arcA aroM
asnB bamA cfa
chaA cvpA cysC
dtpA epmA ftsK
glnH gloA glyS
gpmA gss gstA
gyrA hdfR hemA
hyuA ispE ivy
ldcA lon lpxT
maeB matP metF
moaE nagB nagC
nhaB pdxH pdxY
prs recJ slyA
slyB truB udk
ugpB uvrD wecG
ybjG yciB yciU
ydeP ydgK yegX
ygbE ygeV yhhX
yifK yigZ yjjB
yphB yqeB yqeF
znuB zwf

98

4.3.2. 3mer decomposition and analysis

Under the null hypothesis describing the absence of expression-based selection of 3' UTR sequences, a random DNA sequence of the same length (30 bases) and the same nucleotide content as a 3' UTR sequence should display the same frequencies of composing 3mers as the ones observed or frequencies related to the presence of regulatory elements in these noncoding sequences (e.h. Rho terminator utilization sites, intrinsic transcription terminators).

With this purpose, I generated 1000 random DNA sequence datasets, each containing 141 sequences 30 bases long, using the nucleotide frequencies found in 141 HEG 3' UTR sequences. For comparison, I applied the same procedure to 141 LEG 3' UTR sequences. I decomposed each sequence from the 3'UTR and random DNA dataset in 3mers and I determined the counts for each 3mer per each dataset. I then compared the individual 3mer distribution for the random DNA with the corresponding HEG or LEG 3' UTR 3mer values. Because any given 3mer is represented in the genetic code, I grouped 3mers function of the potential amino acid it might encode for. The results of this analysis are shown in Figures 7and 8 and Figures S3-S21. Comparative 3mer analysis has revealed 18 overrepresented and 12 underrepresented 3mers out of possible 64 in HEG 3'UTRs (at 0.05 significance level) while for the LEG 3'UTR I found 17 underrepresented and 11 overrepresented 3mers. The 3mer TGG encoding tryptophan (Trp), the most conserved amino acid in protein sequences, is found significantly underrepresented in both HEG and LEG 3' UTR (Figure 8).

Assuming that 3'UTR regions lack codon-specfic information across the 3 reading frames (0,+1,+2) it is expected that each individual 3mer frequency encountered is no better than the frequency expected given the individual base frequencies found in these regions. I calculated the expected frequency of a 3mer given the base frequencies for A,C,G and T and I included the

probability of seeing this 3mer in 3948 total 3mers from 141 3' UTR sequences (HEG or LEG) in calculation of the binomial probability of the observed counts (Tables 11 and 12). I then applied the Holm-Bonferroni correction [172] for multiple comparisons. I observed 6 overrepresented 3mers (AAA, CGC, GGC, TAA, TGC, TTT) in HEG 3' UTR sequences with corrected p-values less than 0.05 significance level. In comparison, in LEG 3' UTR sequences I found 7 overrepresented 3mers (AAA, CCG, CGG, CGC, GCC, GCG, TTT) with corrected p-values less than 0.05.

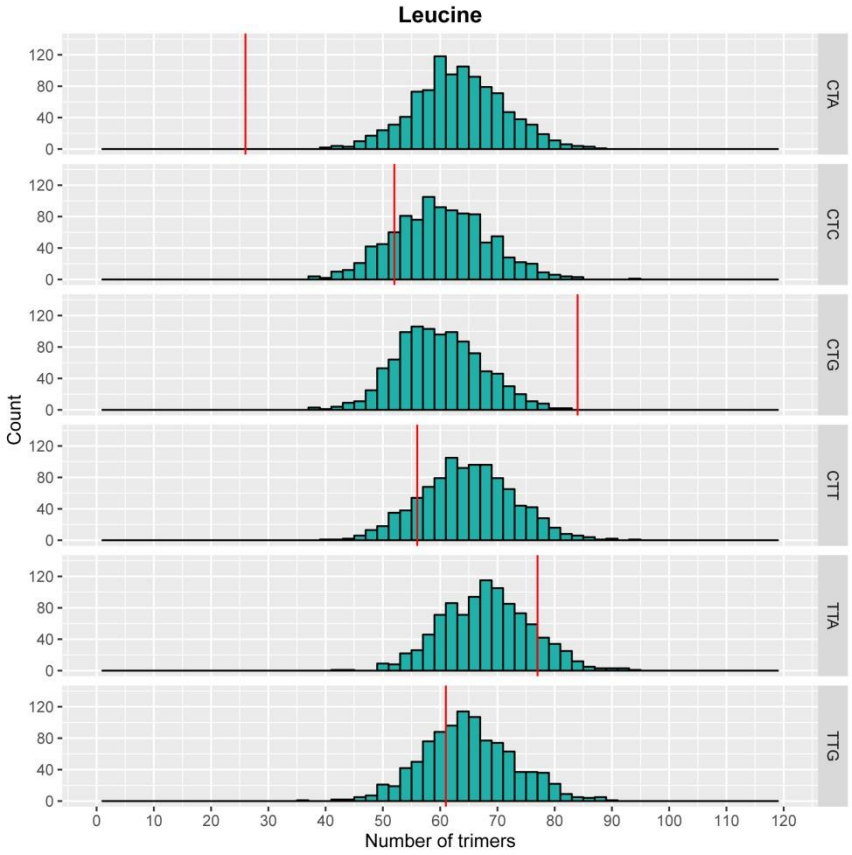
Assuming no impact of natural selection on the 3' UTR sequence then a biased 2mer pool should produce a pool of 3mers that reflect the frequency of its composing 2mers.

I estimated the conditional probability of observing a 3mer given the observed frequency of its composing dimers. I employed this probability to calculate binomial probability of observing the counts of 3mers in the HEG or LEG 3' UTR sequences, adjusting the p-values for multiple comparisons by Holm-Bonferroni method. This analysis yielded 2 overrepresented 3mers (CAG, CTG) for HEG 3' UTR and 1 overrepresented 3mer (TCA) for LEG 3' UTR with a corrected p-value under 0.05.

Figure 7. Distributions of trimer counts encoding Leucine

A. Highly Expressed Genes ; **B.** Lowly Expressed Genes.

A.


Leucine
CTA
CTC
CTG
CTT
TTA
TTG
Count
Number of trimers

B.

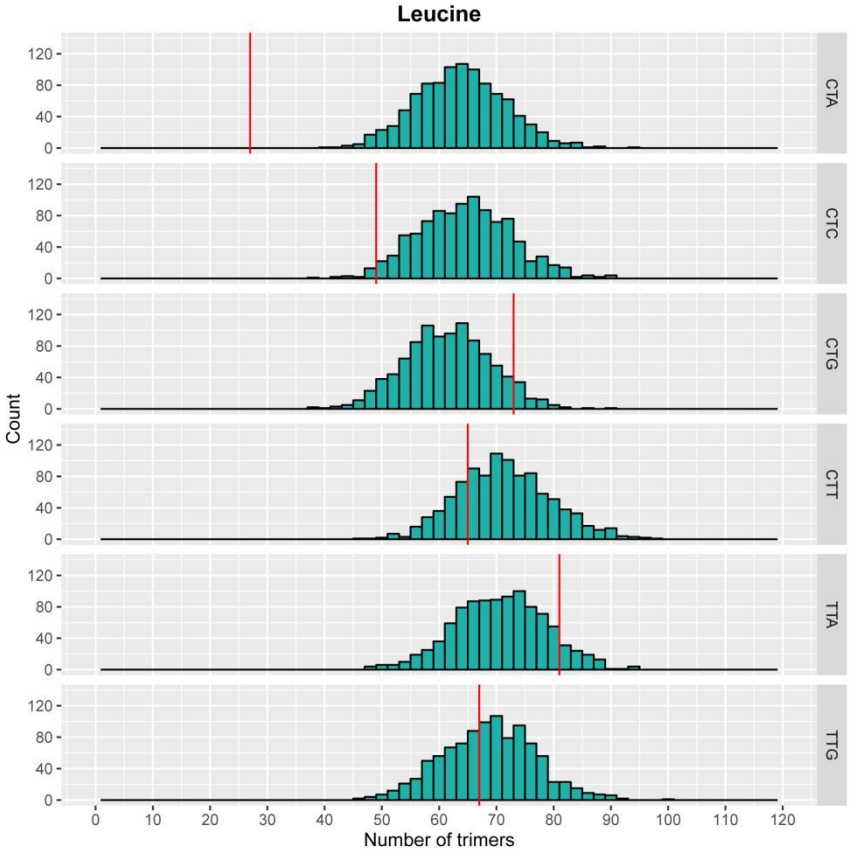
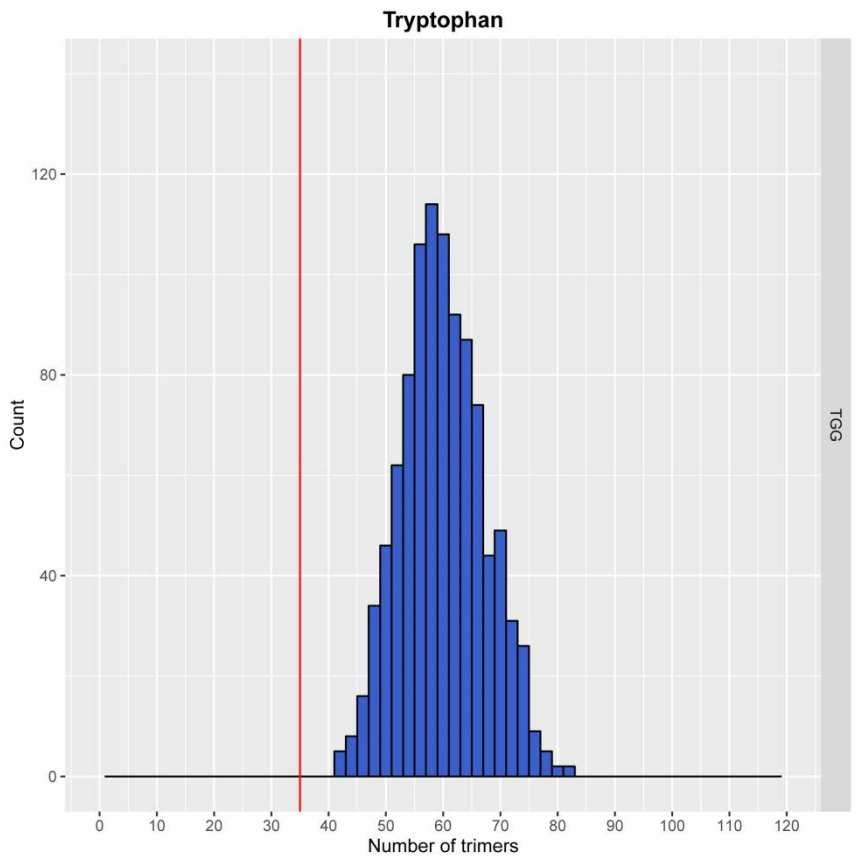

Leucine
CTA
CTC
CTG
CTT
TTA
TTG
Count
Number of trimers

Figure 8. Distributions of trimer counts encoding tryptophan

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Tryptophan
Count
Number of trimers
TGG
0
40
80
120
0 10 20 30 40 50 60 70 80 90 100 110 120

B.

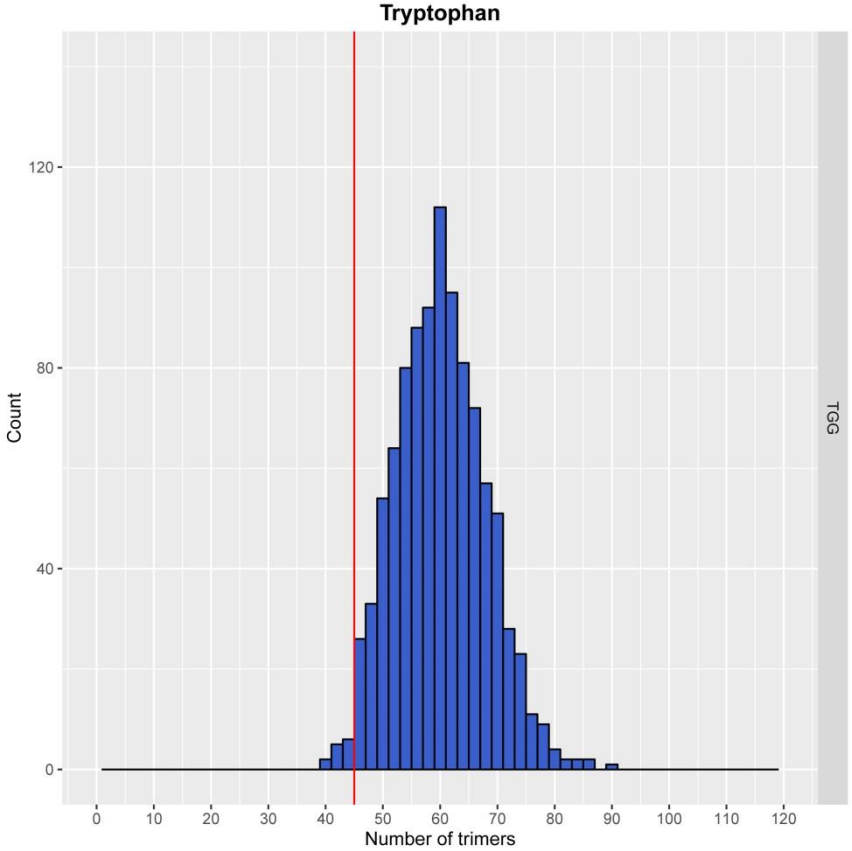

Tryptophan
Count
Number of trimers
TGG
0
40
80
120
0 10 20 30 40 50 60 70 80 90 100 110 120

Table 11. Trimer composition analysis of 3' untranslated regions of highly expressed genes

Trimer	Actual Counts	Expected Counts[1]	Probability Expected[1]	Binomial *P-value*[1]	*P-value*[1] (Holm-Bonferroni)	Expected Counts[2]	Probability Expected[2,3]	Binomial *P-value*[2]	*P-value*[2] (Holm-Bonferroni)
AAA	166	64.620	0.016	0.000	0.000	144.158	0.037	0.031	1.000
AAC	69	61.853	0.016	0.163	1.000	71.166	0.018	0.572	1.000
AAG	74	61.552	0.016	0.051	1.000	75.181	0.019	0.524	1.000
AAT	79	66.485	0.017	0.057	1.000	90.874	0.023	0.888	1.000
ACA	56	61.853	0.016	0.750	1.000	44.345	0.011	0.037	1.000
ACC	42	59.203	0.015	0.989	1.000	45.683	0.012	0.675	1.000
ACG	51	58.915	0.015	0.835	1.000	55.240	0.014	0.688	1.000
ACT	40	63.638	0.016	0.999	1.000	43.007	0.011	0.642	1.000
AGA	48	61.552	0.016	0.957	1.000	45.775	0.012	0.335	1.000
AGC	51	58.915	0.015	0.835	1.000	65.016	0.016	0.958	1.000
AGG	64	58.629	0.015	0.217	1.000	47.395	0.012	0.008	0.510
AGT	35	63.328	0.016	1.000	1.000	40.711	0.010	0.792	1.000
ATA	53	66.485	0.017	0.950	1.000	48.217	0.012	0.219	1.000
ATC	59	63.638	0.016	0.694	1.000	59.601	0.015	0.497	1.000
ATG	56	63.328	0.016	0.805	1.000	58.485	0.015	0.595	1.000
ATT	71	68.405	0.017	0.347	1.000	74.111	0.019	0.614	1.000
CAA	55	61.853	0.016	0.790	1.000	84.670	0.021	1.000	1.000
CAC	41	59.203	0.015	0.992	1.000	41.799	0.011	0.508	1.000
CAG	66	58.915	0.015	0.160	1.000	44.157	0.011	0.001	0.048
CAT	60	63.638	0.016	0.648	1.000	53.374	0.014	0.163	1.000
CCA	38	59.203	0.015	0.998	1.000	54.351	0.014	0.988	1.000
CCC	61	56.668	0.014	0.255	1.000	55.991	0.014	0.226	1.000
CCG	79	56.392	0.014	0.002	0.090	67.705	0.017	0.077	1.000
CCT	53	60.912	0.015	0.830	1.000	52.711	0.013	0.448	1.000
CGA	46	58.915	0.015	0.953	1.000	64.218	0.016	0.990	1.000
CGC	84	56.392	0.014	0.000	0.012	91.212	0.023	0.759	1.000
CGG	80	56.118	0.014	0.001	0.054	66.491	0.017	0.045	1.000
CGT	61	60.616	0.015	0.446	1.000	57.114	0.014	0.275	1.000
CTA	26	63.638	0.016	1.000	1.000	43.569	0.011	0.997	1.000
CTC	52	60.912	0.015	0.862	1.000	53.856	0.014	0.565	1.000
CTG	84	60.616	0.015	0.002	0.090	52.848	0.013	0.000	0.002
CTT	56	65.475	0.017	0.870	1.000	66.968	0.017	0.904	1.000
GAA	70	61.552	0.016	0.126	1.000	82.480	0.021	0.911	1.000
GAC	41	58.915	0.015	0.992	1.000	40.718	0.010	0.441	1.000
GAG	40	58.629	0.015	0.994	1.000	43.015	0.011	0.642	1.000
GAT	67	63.328	0.016	0.294	1.000	51.994	0.013	0.018	1.000
GCA	66	58.915	0.015	0.160	1.000	72.999	0.018	0.776	1.000
GCC	86	56.392	0.014	0.000	0.005	75.202	0.019	0.096	1.000
GCG	90	56.118	0.014	0.000	0.001	90.934	0.023	0.512	1.000
GCT	61	60.616	0.015	0.446	1.000	70.796	0.018	0.869	1.000
GGA	43	58.629	0.015	0.980	1.000	51.996	0.013	0.884	1.000
GGC	83	56.118	0.014	0.000	0.016	73.853	0.019	0.130	1.000
GGG	51	55.845	0.014	0.716	1.000	53.837	0.014	0.618	1.000
GGT	47	60.321	0.015	0.956	1.000	46.245	0.012	0.417	1.000
GTA	44	63.328	0.016	0.994	1.000	38.922	0.010	0.183	1.000
GTC	41	60.616	0.015	0.995	1.000	48.112	0.012	0.831	1.000
GTG	50	60.321	0.015	0.901	1.000	47.211	0.012	0.309	1.000
GTT	62	65.156	0.017	0.623	1.000	59.824	0.015	0.357	1.000
TAA	100	66.485	0.017	0.000	0.003	78.831	0.020	0.009	0.515
TAC	44	63.638	0.016	0.994	1.000	38.916	0.010	0.183	1.000
TAG	25	63.328	0.016	1.000	1.000	41.112	0.010	0.995	1.000
TAT	43	68.405	0.017	0.999	1.000	49.693	0.013	0.810	1.000
TCA	72	63.638	0.016	0.132	1.000	60.719	0.015	0.067	1.000
TCC	49	60.912	0.015	0.933	1.000	62.551	0.016	0.956	1.000
TCG	66	60.616	0.015	0.220	1.000	75.637	0.019	0.856	1.000
TCT	69	65.475	0.017	0.303	1.000	58.887	0.015	0.084	1.000
TGA	82	63.328	0.016	0.010	0.510	58.218	0.015	0.001	0.074
TGC	92	60.616	0.015	0.000	0.004	82.690	0.021	0.138	1.000
TGG	35	60.321	0.015	1.000	1.000	60.279	0.015	1.000	1.000
TGT	48	65.156	0.017	0.985	1.000	51.778	0.013	0.670	1.000
TTA	77	68.405	0.017	0.134	1.000	64.289	0.016	0.052	1.000
TTC	80	65.475	0.017	0.034	1.000	79.468	0.020	0.446	1.000
TTG	61	65.156	0.017	0.670	1.000	77.980	0.020	0.974	1.000
TTT	107	70.379	0.018	0.000	0.001	98.814	0.025	0.187	1.000

[1]Derived from individual base frequencies; [2]Derived from dimer frequencies; [3]Conditional probability

Table 12. Trimer composition analysis of 3' untranslated regions of lowly expressed genes

Trimer	Actual Counts	Expected Counts[1]	Probability Expected[1]	Binomial *P-value*[1]	*P-value*[1] (Holm-Bonferroni)	Expected Counts[2]	Probability Expected[2,3]	Binomial *P-value*[2]	*P-value*[2] (Holm-Bonferroni)
AAA	108	57.833	0.015	0.000	0.000	104.830	0.027	0.353	1.000
AAC	70	58.280	0.015	0.057	1.000	61.628	0.016	0.128	1.000
AAG	60	56.269	0.014	0.280	1.000	60.357	0.015	0.484	1.000
AAT	81	63.980	0.016	0.016	0.877	91.806	0.023	0.863	1.000
ACA	51	58.280	0.015	0.814	1.000	44.019	0.011	0.130	1.000
ACC	50	58.731	0.015	0.861	1.000	49.913	0.013	0.458	1.000
ACG	45	56.704	0.014	0.937	1.000	52.123	0.013	0.821	1.000
ACT	40	64.474	0.016	0.999	1.000	41.256	0.010	0.537	1.000
AGA	37	56.269	0.014	0.996	1.000	37.674	0.010	0.501	1.000
AGC	61	56.704	0.014	0.256	1.000	59.635	0.015	0.396	1.000
AGG	49	54.746	0.014	0.759	1.000	43.922	0.011	0.197	1.000
AGT	34	62.249	0.016	1.000	1.000	42.218	0.011	0.886	1.000
ATA	66	63.980	0.016	0.368	1.000	60.627	0.015	0.221	1.000
ATC	68	64.474	0.016	0.301	1.000	59.358	0.015	0.117	1.000
ATG	72	62.249	0.016	0.097	1.000	69.759	0.018	0.364	1.000
ATT	69	70.779	0.018	0.554	1.000	89.291	0.023	0.986	1.000
CAA	67	58.280	0.015	0.114	1.000	75.923	0.019	0.835	1.000
CAC	51	58.731	0.015	0.829	1.000	44.633	0.011	0.151	1.000
CAG	59	56.704	0.014	0.347	1.000	43.713	0.011	0.011	0.644
CAT	58	64.474	0.016	0.771	1.000	66.490	0.017	0.838	1.000
CCA	48	58.731	0.015	0.914	1.000	61.490	0.016	0.957	1.000
CCC	61	59.185	0.015	0.374	1.000	69.723	0.018	0.840	1.000
CCG	91	57.142	0.014	0.000	0.001	72.811	0.018	0.016	0.943
CCT	65	64.973	0.016	0.466	1.000	57.631	0.015	0.148	1.000
CGA	34	56.704	0.014	0.999	1.000	56.115	0.014	0.999	1.000
CGC	83	57.142	0.014	0.000	0.027	88.825	0.022	0.712	1.000
CGG	91	55.170	0.014	0.000	0.000	65.420	0.017	0.001	0.064
CGT	62	62.730	0.016	0.503	1.000	62.882	0.016	0.511	1.000
CTA	27	64.474	0.016	1.000	1.000	46.991	0.012	0.999	1.000
CTC	49	64.973	0.016	0.977	1.000	46.008	0.012	0.296	1.000
CTG	73	62.730	0.016	0.088	1.000	54.069	0.014	0.005	0.339
CTT	65	71.327	0.018	0.754	1.000	69.208	0.018	0.668	1.000
GAA	61	56.269	0.014	0.238	1.000	63.216	0.016	0.578	1.000
GAC	25	56.704	0.014	1.000	1.000	37.163	0.009	0.978	1.000
GAG	35	54.746	0.014	0.997	1.000	36.397	0.009	0.549	1.000
GAT	70	62.249	0.016	0.146	1.000	55.362	0.014	0.023	1.000
GCA	56	56.704	0.014	0.502	1.000	71.474	0.018	0.967	1.000
GCC	100	57.142	0.014	0.000	0.000	81.044	0.021	0.017	0.979
GCG	92	55.170	0.014	0.000	0.000	84.632	0.021	0.192	1.000
GCT	61	62.730	0.016	0.554	1.000	66.988	0.017	0.747	1.000
GGA	52	54.746	0.014	0.612	1.000	46.002	0.012	0.167	1.000
GGC	78	55.170	0.014	0.001	0.078	72.817	0.018	0.247	1.000
GGG	41	53.265	0.013	0.952	1.000	53.631	0.014	0.957	1.000
GGT	54	60.565	0.015	0.782	1.000	51.550	0.013	0.333	1.000
GTA	54	62.249	0.016	0.839	1.000	46.781	0.012	0.129	1.000
GTC	39	62.730	0.016	0.999	1.000	45.802	0.012	0.825	1.000
GTG	52	60.565	0.015	0.852	1.000	53.828	0.014	0.564	1.000
GTT	70	68.865	0.017	0.414	1.000	68.899	0.017	0.416	1.000
TAA	86	63.980	0.016	0.003	0.183	75.923	0.019	0.112	1.000
TAC	45	64.474	0.016	0.994	1.000	44.633	0.011	0.439	1.000
TAG	32	62.249	0.016	1.000	1.000	43.713	0.011	0.961	1.000
TAT	67	70.779	0.018	0.647	1.000	66.490	0.017	0.443	1.000
TCA	78	64.474	0.016	0.042	1.000	53.095	0.013	0.000	0.031
TCC	49	64.973	0.016	0.977	1.000	60.204	0.015	0.921	1.000
TCG	48	62.730	0.016	0.969	1.000	62.870	0.016	0.970	1.000
TCT	53	71.327	0.018	0.986	1.000	49.763	0.013	0.291	1.000
TGA	72	62.249	0.016	0.097	1.000	54.528	0.014	0.009	0.566
TGC	86	62.730	0.016	0.002	0.111	86.314	0.022	0.485	1.000
TGG	45	60.565	0.015	0.978	1.000	63.571	0.016	0.991	1.000
TGT	63	68.865	0.017	0.739	1.000	61.105	0.015	0.372	1.000
TTA	81	70.779	0.018	0.101	1.000	73.843	0.019	0.183	1.000
TTC	63	71.327	0.018	0.825	1.000	72.298	0.018	0.852	1.000
TTG	67	68.865	0.017	0.558	1.000	84.966	0.022	0.975	1.000
TTT	128	78.302	0.020	0.000	0.000	108.756	0.028	0.030	1.000

[1]Derived from individual base frequencies; [2]Derived from dimer frequencies; [3]Conditional probability

4.4. Discussion

4.4.1 3' UTR regions of a subset of E. coli HEG evolve at a similar rate as ORF regions

A typical protein-coding gene in any given genome consists in an open reading frame and its associated regulatory sequences. Upon transcription, the resulting messenger RNA transcript will contain a 5' untranslated region, the coding region (open reading frame) and 3' untranslated region. The traditional, textbook view of 5' and 3' untranslated regions as being non-coding regions is currently challenged by several studies in the field of evolutionary biology [58], [144], [156], [157], [173]. A recent study suggests the 3' UTR sequence might be modulating movement of ribosomes on the mRNA molecules in yeast [150]. Conservation of these sequences in eukaryotic 3' UTRs indicates the presence of purifying selection at lower levels than the normal open reading frames [174].

Other studies have shown that highly expressed genes in bacteria evolve at the slower rate and their translation is strongly influenced by selection [135]. The consequences of DNA mutations and translational errors are reflected in the structural and functional stability of the expressed products. Because most highly expressed genes play a central role in performing cellular processes, any change in their nucleotide and amino acid sequence can have profound effects on their expression level and function. Frameshifts and stop codon readthrough events are most likely to impact the overall fitness of organisms. Therefore, the DNA sequence immediately following the open reading frame of a gene, the 3' untranslated region (3' UTR), is likely to be expressed during mistranslation events. Upon expression, these 3' UTR sequences will be subjected to the influence of natural selection proportional with the rate of translation of the

mRNA transcript involved. Consequently, non-overlapping 3' UTR regions of highly expressed genes are more likely to encounter selective pressures than it would be expected for the 3' UTRs of lowly expressed genes.

To investigate the role of natural selection in sequence modulation of 3' UTR regions, I firstly used a phylogenetic approach. Assuming that HEG 3' UTR sequences are often expressed, their evolutionary rates should be comparatively closer to the rates seen in the ORFs than in the case of LEG 3'UTR. I carried out phylogenetic reconstruction analyses by maximum likelihood and parsimony to determine these evolutionary rates as overall tree length values in HEG and LEG 3' UTR regions and ORF. The boxplots in Figure 4 illustrate that the distributions for HEG 3' UTR and HEG ORF tree length values, while displaying different variances, have similar medians suggesting similar evolutionary rates. The values for HEG and LEG 3'UTR displayed a larger variance than the corresponding ORF values, probably because the number of phylogenetically informative sites was much smaller (about 1000 bases in length for an average ORF and 30 bases for the 3'UTR) . In contrast, there is a marked difference in the medians for the tree length distribution of LEG ORF and 3'UTR. A slight shift towards higher evolutionary rates of LEG 3'UTR distribution when compared to ORF values, seem to support the idea that for LEG, ORF evolve at a different rate than the 3' UTR. I carried out statistical tests to evaluate the significance of these observations. Because I made no assumptions about the shape of the distributions tested, I applied two non-parametric tests: Kolmogorov-Smirnov and Wilcoxon Rank Sum. The KS test result comparing HEG 3'UTR and ORF tree length values cannot support rejection the null hypothesis of the two sets of values coming the same distribution possibly for lack of sufficient data ($p>0.05$). The null hypothesis has been rejected in case of

LEG ORF versus 3' UTR ($p<0.05$). To verify whether the HEG and LEG evolve at different rates as it has been previously suggested [126], I carried out KS test only on the ORF of HEG and LEG. The p-value of the test was highly significant ($p<0.01$) demonstrating that indeed LEG sequences evolve at a different rate than HEG sequences. Additionally, I carried out the KS test separately for the HEG and LEG 3' UTR. The high p-value of the test underlines that the HEG 3'UTR and LEG 3'UTR tree length values distributions might be quite similar ($p>0.05$).

Surprisingly, I performed the Wilcoxon Rank Sum test on the same data, all p-values were higher than 0.05 with the exception of HEG ORF versus LEG ORF comparison ($p<0.01$).This would suggest that when comparing both LEG and HEG subsets, 3' UTR and ORF tree length values may be part of the same distribution or because, in this case, the test lacks power. Because variances play an important role in determining the shape of the distribution and ultimately the fate of the statistical test, I examined the discrepancy between the KS and Wilcoxon tests by carrying out analyses on the homogeneity of variances (Supplemental Table 1). I employed 3 tests: Levene's- robust to low deviations from normal distributions, Bartlett's parametric test-applicable to normal distributions and Fligner-Killeen's non-parametric test applicable to non-normal distributions. The Fligner-Killeen's test suggests that the variances of HEG ORF and HEG 3'UTR values are not significantly different (p-value > 0.05). I saw a similar outcome of the test for LEG ORF and LEG 3'UTR values (p-value >0.05).

Taking into consideration these observations, I can conclude that the variance of the distributions plays no role in the lack of congruity between KS and Wilcoxon Rank Sum tests results. As KS test is more sensitive to any discrepancies between two distributions and Wilcoxon Rank Sum is

especially sensitive to the differences in the medians of the distributions, it is mybelief that the KS test outcome represents a clearer picture of the differences between distributions.

Collectively, these findings suggest that HEG 3' UTR for the 59 genes analyzed display significantly a similar level of sequence conservation that is found in the ORF in the 62 *E. coli* species. Comparatively, the 62 LEG 3' UTR regions seem to evolve at a different, higher rate than I found for the corresponding ORF.

Separately, in order to test whether 3'UTR regions evolve at the same or different rates than their corresponding ORF in HEG and LEG subsets, I generated 1000 putative bootstrap samples 30 bases long from each concatenated alignment including the ORF and the adjacent 3' UTR. I then measured their tree lengths generated by two methods (maximum likelihood and parsimony) and compared them with the values of the HEG and LEG 3' UTR. The results of the maximum likelihood analysis of these bootstrap samples show that a smaller percentage of 3'UTR values of the HEG subset (25.4%) group significantly outside of 95% bootstrap values compared to the LEG subset (37.1%). This would argue that HEG 3'UTR have evolutionary rates closer to their cognate ORF than seen in the case of LEG. The parsimony analysis however yielded 33.9% of 3'UTR tree length values (measured in steps) grouping outside of 95% of bootstrap data for the HEG subset and 38.7% of 3'UTR values for the LEG subset. Because this difference was supported by the binomial probability, I assume the results of the parsimony analysis were inconclusive.

4.4.2. HEG and LEG 3'UTR sequences show enrichment in preferred codons

Each coding sequence in bacterial genomes is usually organized in contiguous nucleotide triplets (or 3mers) which, according to the genetic code, will determine initiation or termination of translation as well as insertion of cognate amino acids. The distribution of individual nucleotide frequencies for a coding sequence is thus expected to follow a nonrandom pattern as result of natural selection operating on the amino acid sequence. Some biological sequences display serial correlations for oligonucleotides with increasing hierarchical order (e.g. single nucleotide, 2mer, 3mer). Decoding preference for certain synonymous triplets was partially explained in the early studies of oligonucleotide composition of DNA sequences by the occurring nucleotide and doublet (2mer) frequencies in the sequences [175], [176]. Additionally, the asymmetries in frequency patterns and variation of certain dinucleotides were explained by nearest neighbor preferences [176], structural constraints of the DNA packaging [177], In contrast, assuming that 3'UTR sequences are evolving under no selective constraints or selective constraints different than the ones experienced by coding sequences, it is expected that their 3mer composition and frequencies should be no different than the ones determined for random DNA sequences of the same nucleotide content. I have generated 1000 sets of 141 random DNA sequences 30 bases long with the same nucleotide content either to LEG or HEG 141 3' UTR sequences. By using the occurring nucleotide and 2mer frequencies to determine the expected probability of encountering each of the 64 possible 3mers in the analyzed sequences, I introduced a Markovian background to account for the first order and second order dependencies among these composing oligonucleotides. Based on the counts distribution comparisons, I found that HEG 3' UTR regions show significant underrepresentation for 18 3mers (28% of 3mers) and overrepresentation for 12 3mers (18%). The LEG 3'UTR regions display 17 underrepresented

(26%) and 11 overrepresented 3mers (17%). For the 3mer encoding Tryptophan (TGG) both 3mer count distributions for HEG and LEG 3' UTR show underrepresentation. As tryptophan is one of the most conserved amino acids, its lower count in 3' UTR might argue that its presence in that region might be deleterious.

Under the hypothesis that 3'UTR regions contain no coding information, the presence of this number of count deviations demonstrate the 3'UTR sequences undergo coding sequence optimization as result of expression. I determined the actual and expected counts for all 64 possible 3mers found in the 3' UTR. I derived the expected probability for each 3mer based on individual nucleotide and 2mer composition of each set of 141 3'UTR sequences (HEG or LEG). I then calculated the binomial probability of observing each 3mer given the expected probabilities. I have found significant overrepresentation based on individual nucleotide frequencies for 6 3mers (AAA, CGC, GGC, TAA, TGC, TTT) encoding lysine, arginine, glycine, cysteine, phenylalanine and stop of translation for the HEG 3' UTR dataset. All 6 3mers are preferably decoded by the tRNAs during protein synthesis in *E. coli* K12 [178], [179]. Presence of lysine, arginine, cysteine and glycine in 3' UTR sequences may favored as selection against "stickiness" (hydrophobicity) of residues in coding sequences has been shown to be inversely correlated with level of expression [135]. Additionally, 2 more 3mers (CAG, CTG) encoding glutamine and leucine, have shown to be significantly overrepresented based on the conditional probability of composing 2mers and not based on the nucleotide frequencies. They are also preferentially decoded by tRNAs. These observations are consistent with the overrepresentation detected in the comparison with random DNA with the exception of CAG (glutamine).

By comparison, for the LEG 3' UTR dataset I determined significant binomial p-values from individual nucleotide frequencies for 7 3mers (AAA, CCG, CGG, CGC, GCC, GCG, TTT) encoding lysine, proline, arginine, alanine and phenylalanine. All 3mers, with the exception of CGG, are preferred codons. An additional 3mer (TCA) encoding serine showed significant overrepresentation based on the expected probabilities for dimers but not when using the individual nucleotide frequencies.

The presence of a significant deviation from background of 3mers TTT and AAA both in HEG and LEG 3' UTR sequences may indicate, possibly, the presence of regulatory sequences rich in A and T bases (for example, transcription terminators) [180]. Additionally, I cannot exclude the possibility of stochastic effects due to short length of the analyzed 3' UTR sequences.

The lack of concordance between the 3mers significantly overrepresented based on composing 2mer frequencies and the 3mers signficantly overrepresented based on individual nucleotide probabilities may be caused by the difference of information content provided by nucleotide frequencies versus 2mer frequencies. Comparison of the expected probabilities for the possible 16 2mers based on nucleotide frequencies with the actual occurences has revealed no significant overrepresentation or underrepresentation (corrected binomial p-value > 0.05, data not shown). However, when calculating the expected conditional probability for a 3mer using the conditional probabilities of co-occuring 2mers, position information is also taken into consideration rather than just the simple, independent co-occurrence of individual nucleotides. The 3mers significantly overrepresented based on 2mer frequencies in HEG (CAG, CTG) and LEG (TCA) may indicate selection for the presence of these 3mers in the 3' UTR sequences as result of 3'

UTR expression. *I.e.* The amino acids encoded may be selected for their properties in these regions of the genome.

Overall, the presence of significantly overrepresented 3mers in the 3' UTR sequences of HEG and LEG may suggest that selection may operate in these sequences, by possibly favoring coding sequences with neutral fitness effects. Further compositional analysis of these sequences taking into consideration other types of oligonucleotides (4mer, 5mer or 6 mer) may reveal more significant discrepancies between the expected and actual values, similar to the findings of Volinia et al [181]. The study has shown significant overrepresentation and underrepresentation from expectations for 4mers, 5mers and 6mers in both coding and noncoding sequences. Additionally, patterns of missing DNA oligonucleotides may also provide information on the level of selection occurring in 3' UTR sequences.

Furthermore, in the analysis of the composition of DNA sequences, the level of information content can be used as a measure of non-randomness. Another direction of research could investigate the level of information stored in 3' UTR sequences by using information theory techniques such as Shannon's information entropy measure [182]. Entropy measures have attempted to differentiate between coding from noncoding sequences with no conclusive results. However, significant differences in entropy values were found between normal reading frames and frameshifted sequences [183], indicating a possible use in determining the strength of selection between the two types of sequences in 3' UTR regions.

4.5. Conclusions

Collectively, these findings and observations presented in this exploratory study suggest that gene innovation through stop codon read-through and possibly frameshift events might be pervasive in *E. coli* and, possibly, in other bacterial genomes. Elucidating this mechanism of gene innovation will have a major impact on our understanding of the emergence of new genes and/or functions as well as genome evolvability through the interplay between selection and mutation.

4.6. Supplemental Data

Figure S2. Sequence dataset assembly pipeline

A set of 253 of putative highly expressed genes from Highly Expressed Gene Database (HEG-DB) and a set of 453 of putative lowly expressed genes determined using codon usage analysis (cal-CAI) were used in this study. Gene Reference genome sequence information (.ptt, .rnt, .faa, .fna) was used to extract sequences for open reading frames and corresponding 3' untranslated regions.

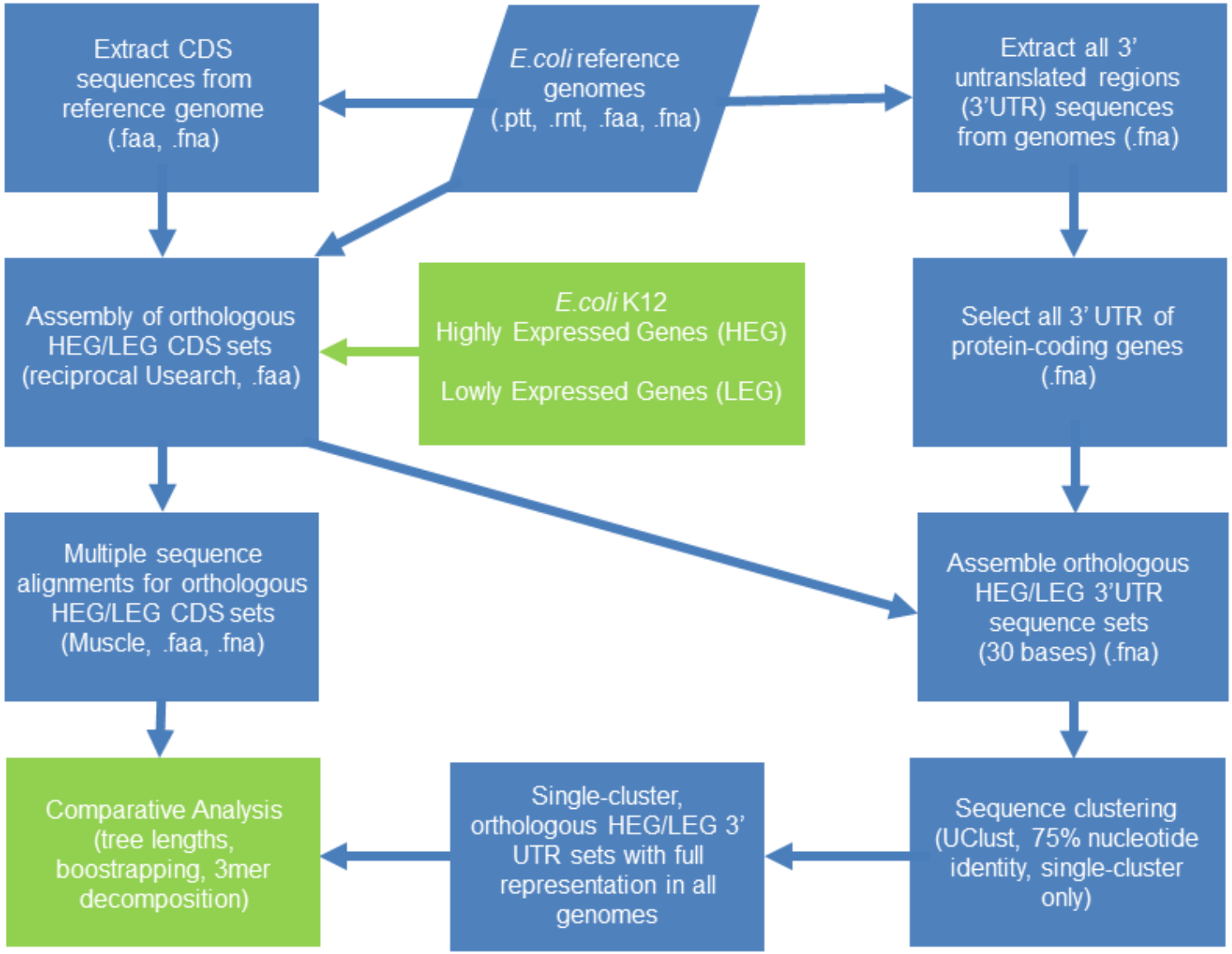

Extract CDS sequences from reference genome (.faa, .fna)
E.coli reference genomes (.ptt, .rnt, .faa, .fna)
Extract all 3' untranslated regions (3'UTR) sequences from genomes (.fna)
Assembly of orthologous HEG/LEG CDS sets (reciprocal Usearch, .faa)
E.coli K12
Highly Expressed Genes (HEG)
Lowly Expressed Genes (LEG)
Select all 3' UTR of protein-coding genes (.fna)
Multiple sequence alignments for orthologous HEG/LEG CDS sets (Muscle, .faa, .fna)
Assemble orthologous HEG/LEG 3'UTR sequence sets (30 bases) (.fna)
Comparative Analysis (tree lengths, boostrapping, 3mer decomposition)
Single-cluster, orthologous HEG/LEG 3' UTR sets with full representation in all genomes
Sequence clustering (UClust, 75% nucleotide identity, single-cluster only)

Table S5. Homogeneity of variances in HEG and LEG ORF and 3' UTR tree length datasets

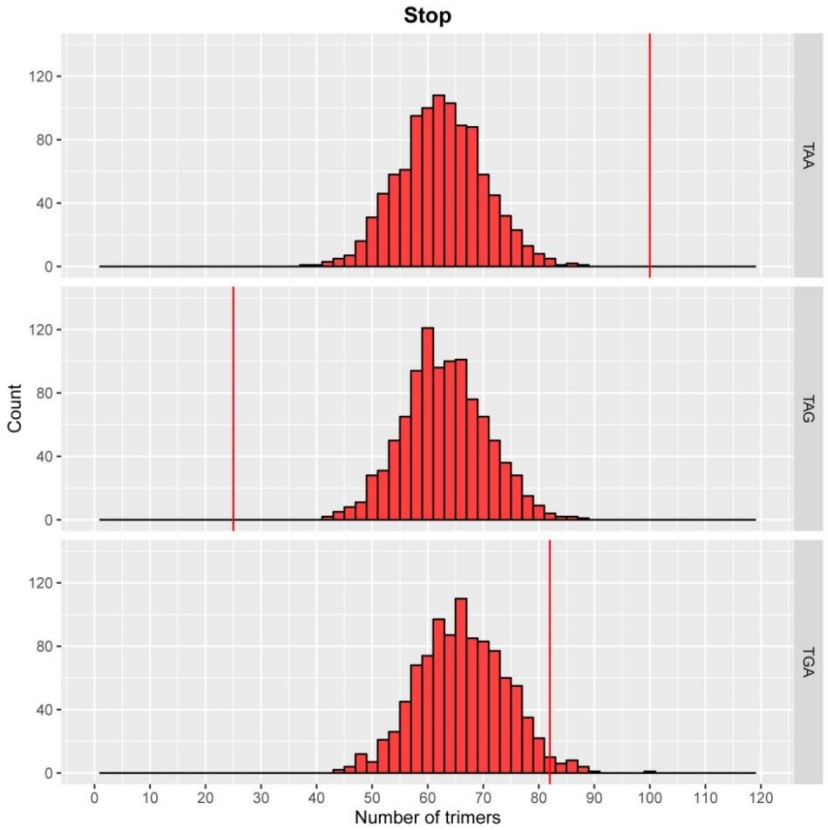
Tree Length Value Comparison	Levene's Test		Bartlett's Test		Fligner-Killeen's Test	
	F-value	Pr(>F)	K-squared	p-value	Med chi-squared	p-value
LEG ORF vs 3UTR	3.6391	0.05956	5.7135	0.01683	1.5658	0.2108
HEG ORF vs 3UTR	6.3504	0.01329	12.3519	0.0004405	2.8741	0.09001

Abbreviations: LEG- Lowly Expressed Genes, HEG- Highly Expressed Genes, ORF- Open Reading Frame, 3' UTR- 3' Untranslated Region

Figure S3. Distributions of trimer counts encoding Stop

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Stop
Count
Number of trimers
TAA
TAG
TGA
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

B.

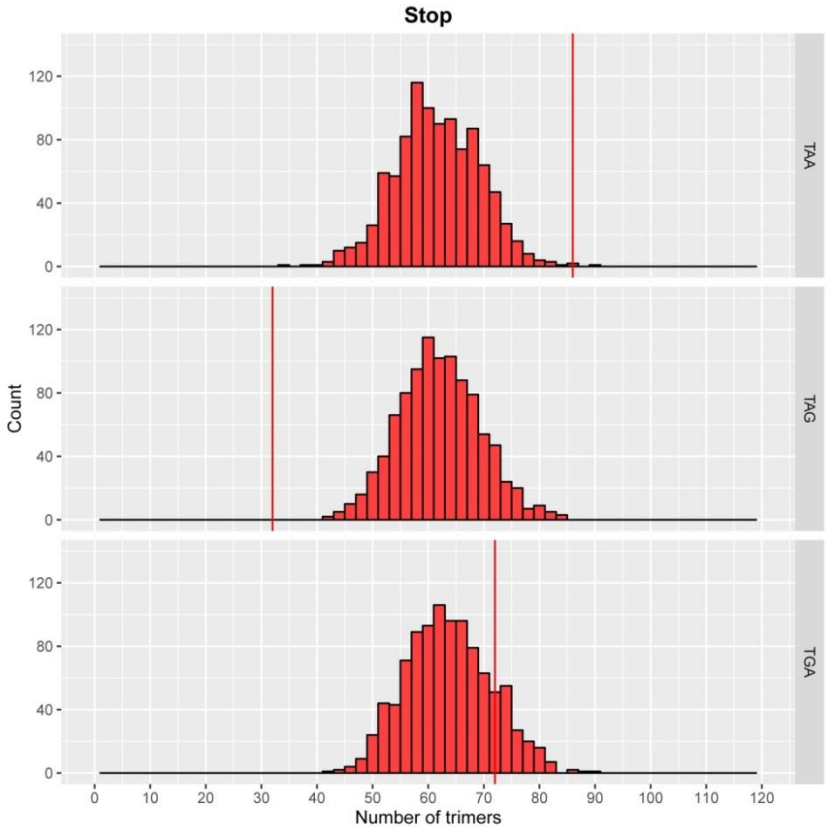
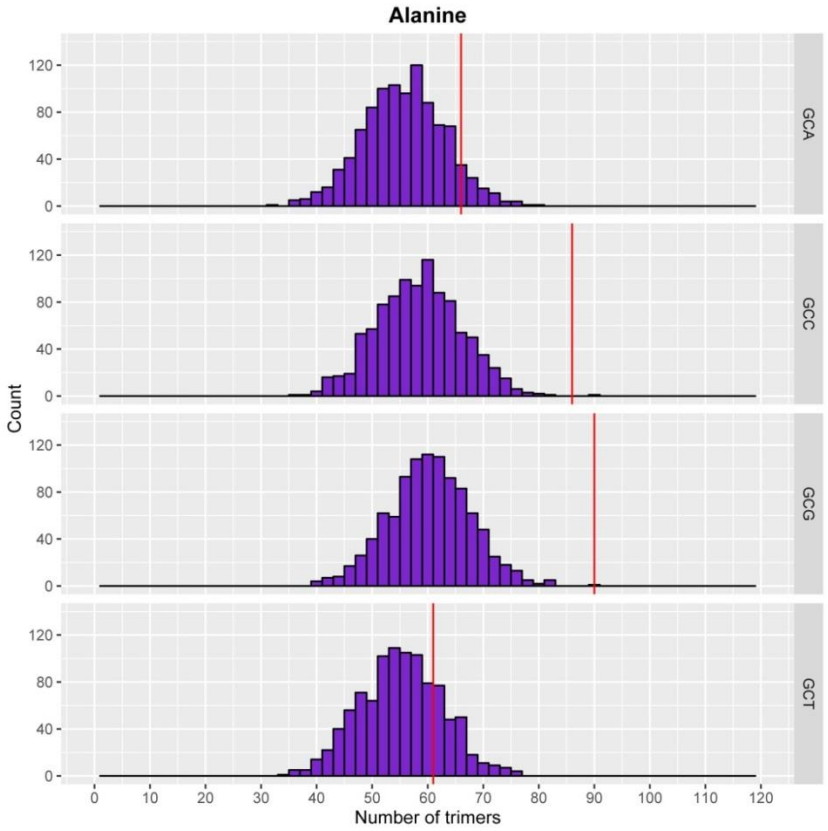

Stop
Count
Number of trimers
TAA
TAG
TGA
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

Figure S4. Distributions of trimer counts encoding Alanine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Alanine
GCA
GCC
GCG
GCT
Count
Number of trimers
0
10
20
30
40
50
60
70
80
90
100
110
120
0
40
80
120

B.

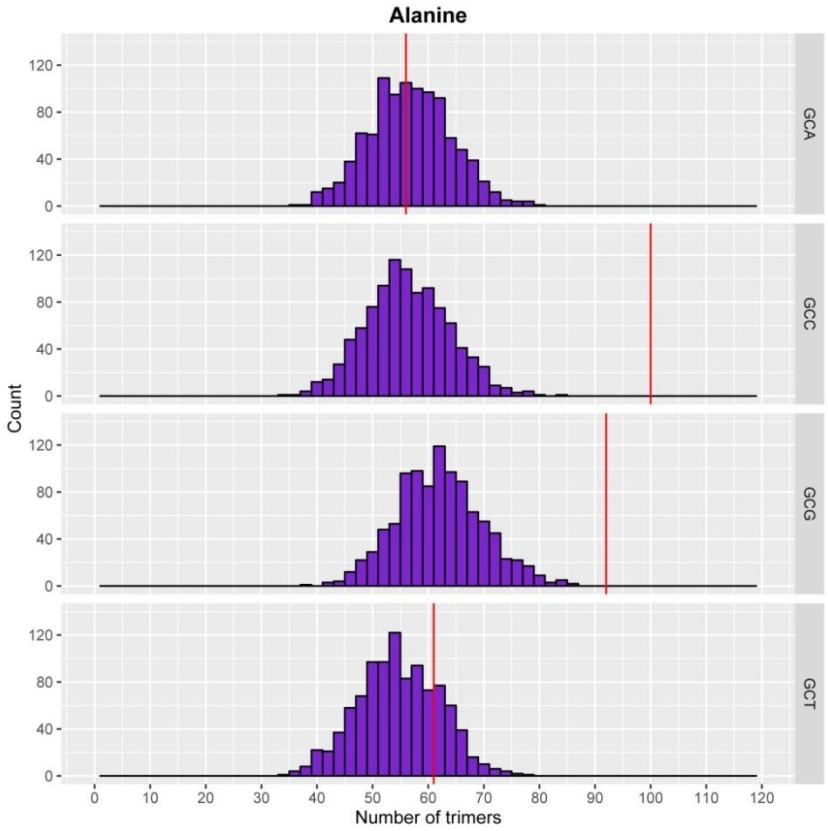
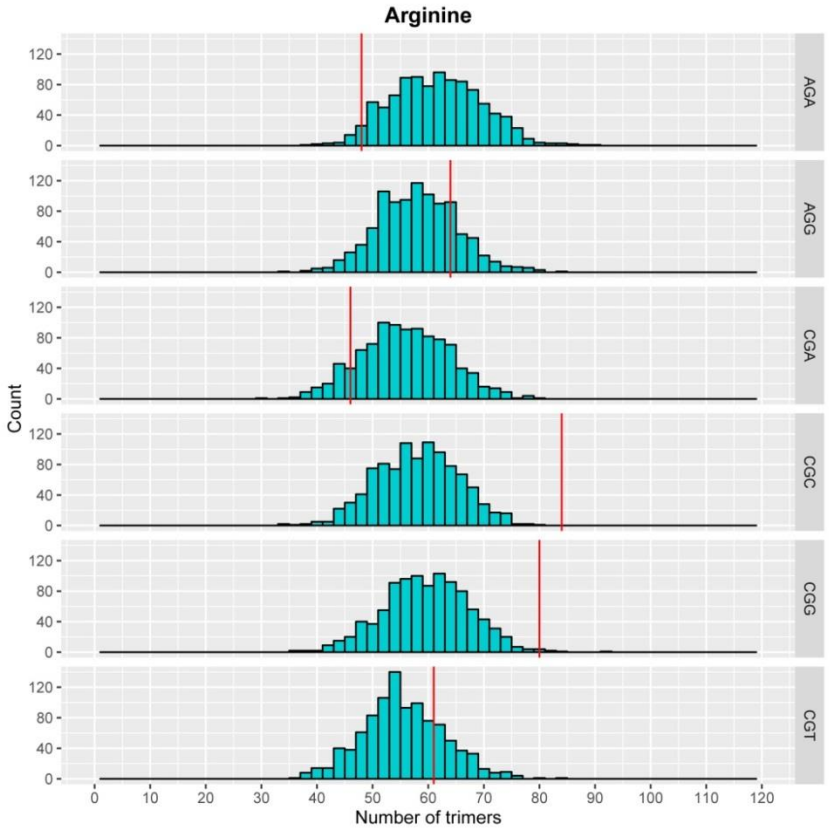

Alanine
GCA
GCC
GCG
GCT
Count
Number of trimers
0
10
20
30
40
50
60
70
80
90
100
110
120
0
40
80
120

Figure S5. Distributions of trimer counts encoding Arginine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Arginine
AGA
AGG
CGA
CGC
CGG
CGT
Count
Number of trimers

B.

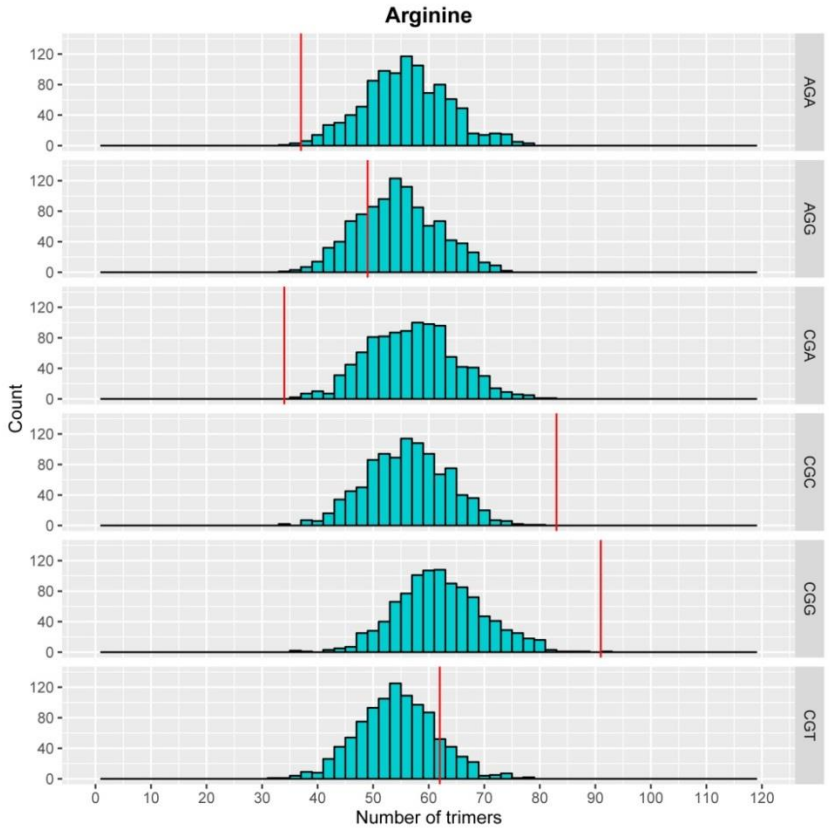
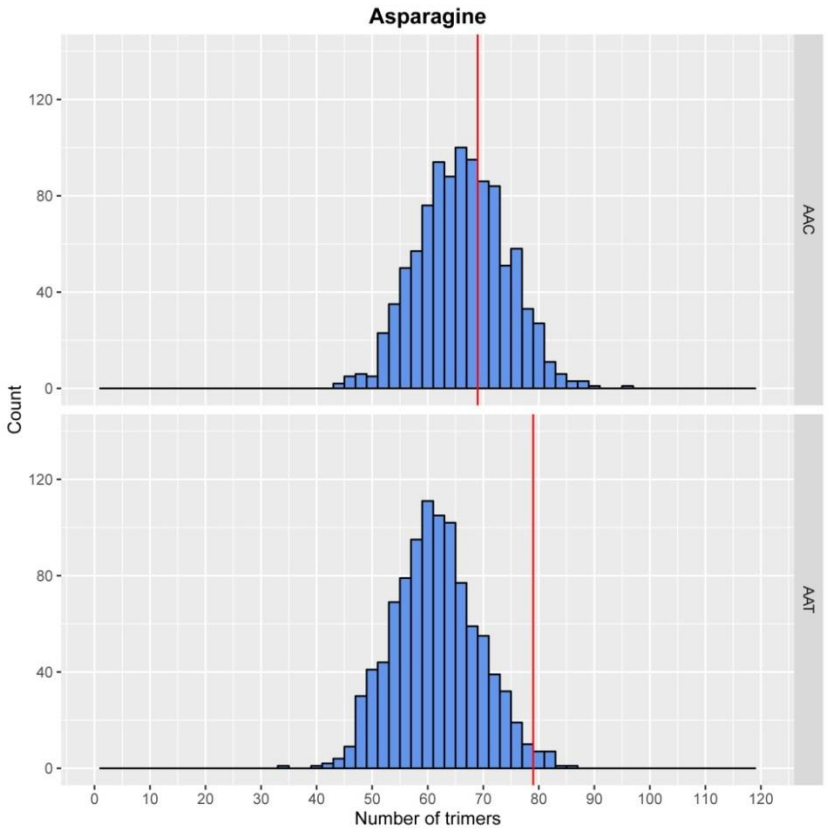

Arginine
AGA
AGG
CGA
CGC
CGG
CGT
Count
Number of trimers

Figure S6. Distributions of trimer counts encoding Asparagine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Asparagine
Count
AAC
AAT
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120
Number of trimers

B.

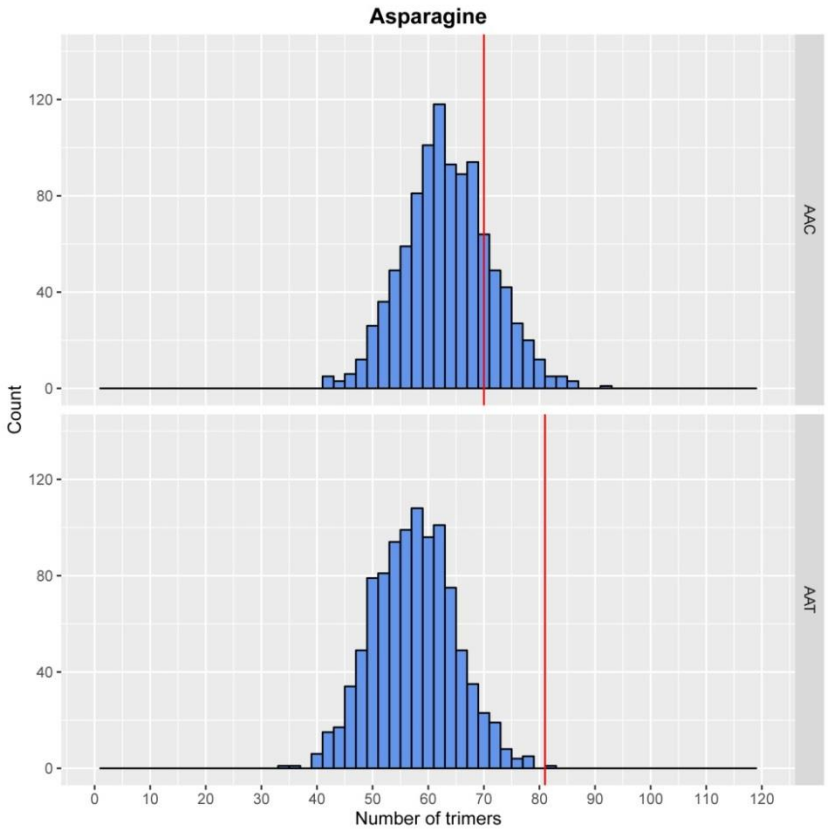
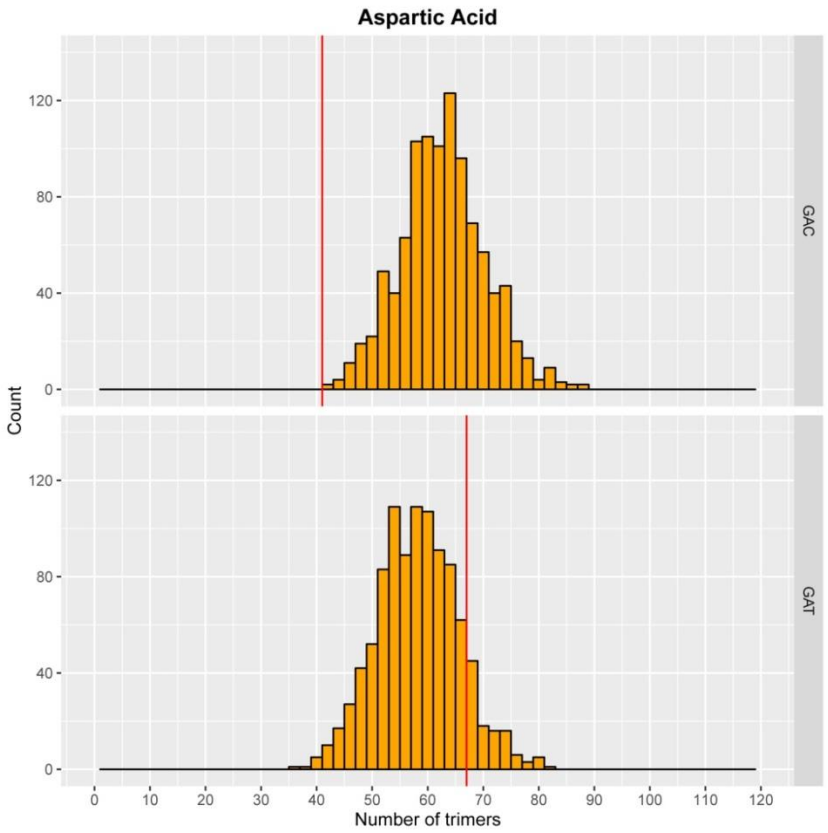

Asparagine
Count
AAC
AAT
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120
Number of trimers

Figure S7. Distributions of trimer counts encoding Aspartic Acid

A. Highly Expressed Genes and **B.** Lowly Expressed Genes.

A.


Aspartic Acid
Count
GAC
GAT
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120
Number of trimers

B.

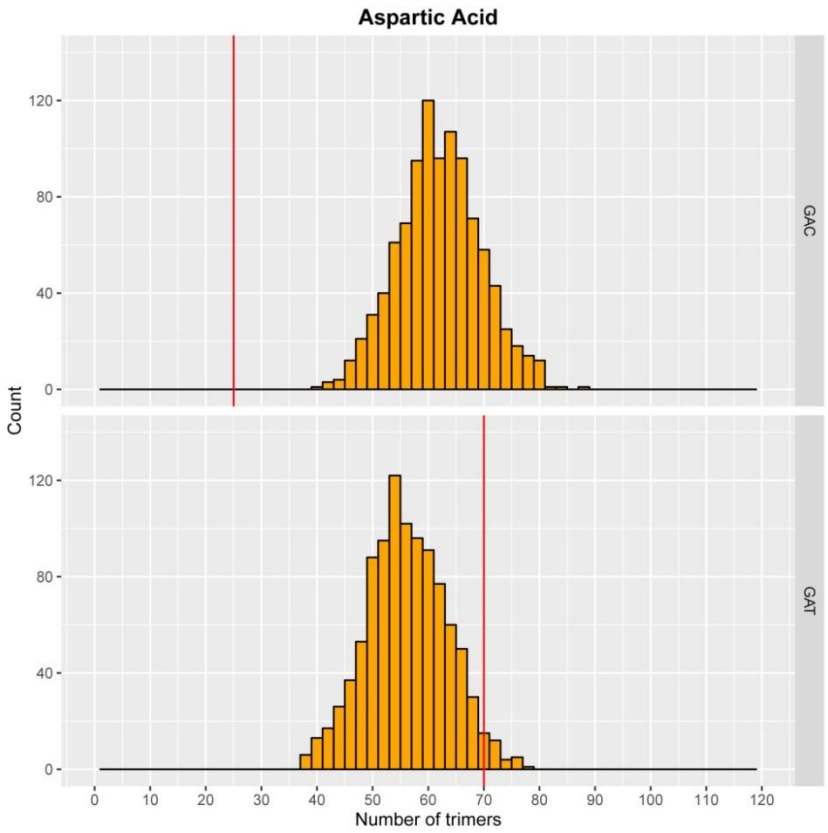
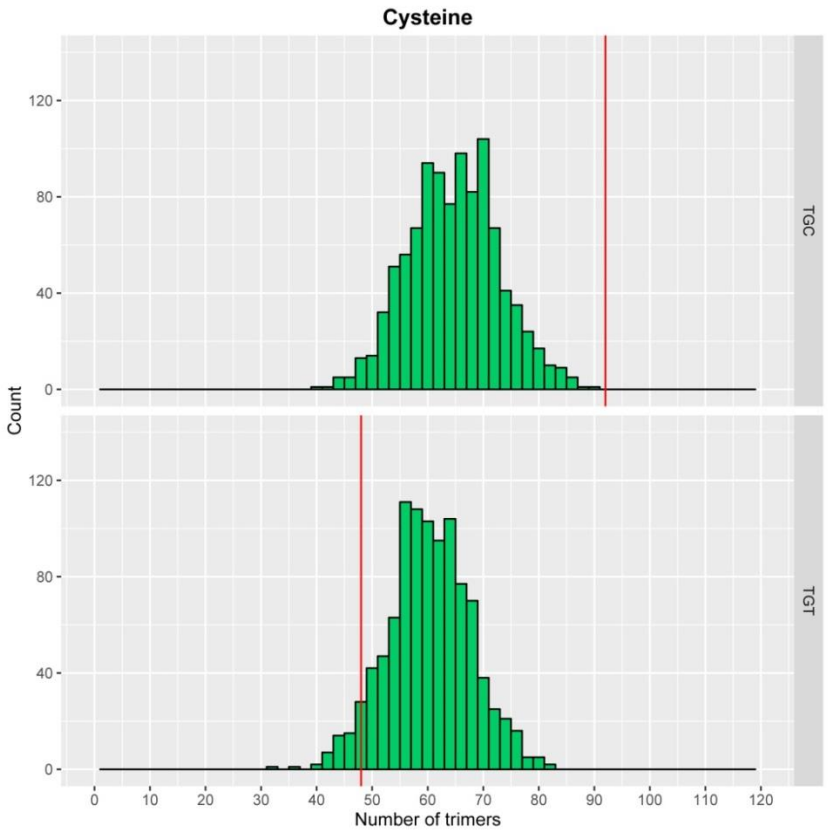

Aspartic Acid
Count
GAC
GAT
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120
Number of trimers

Figure S8. Distributions of trimer counts encoding Cysteine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Cysteine
TGC
TGT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

B.

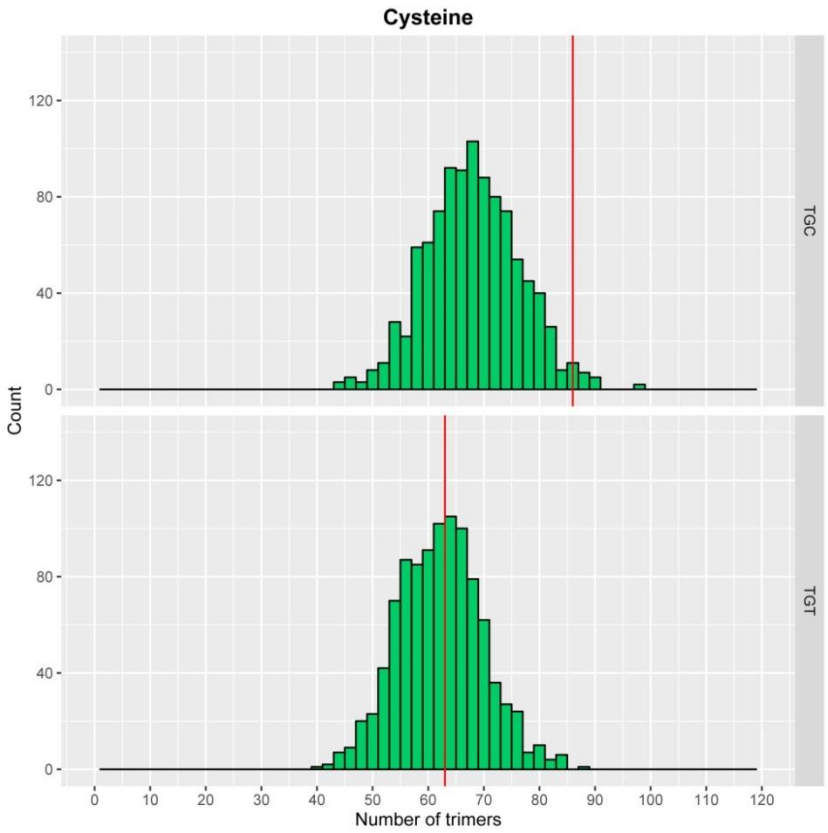
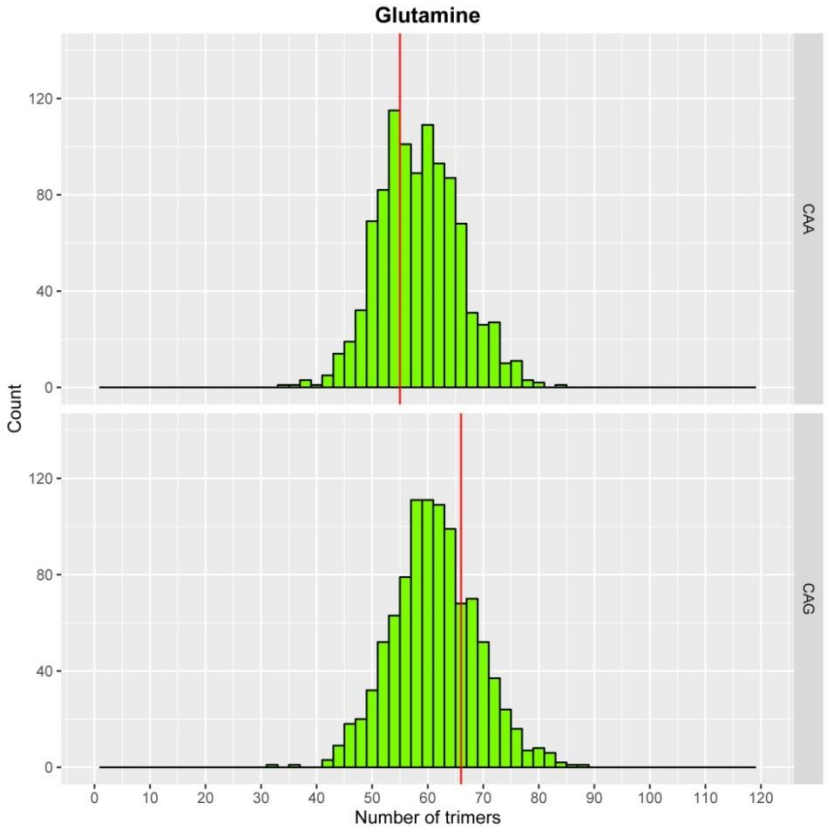

Cysteine
TGC
TGT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

Figure S9. Distributions of trimer counts encoding Glutamine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Glutamine
Count
CAA
CAG
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120
Number of trimers

B.

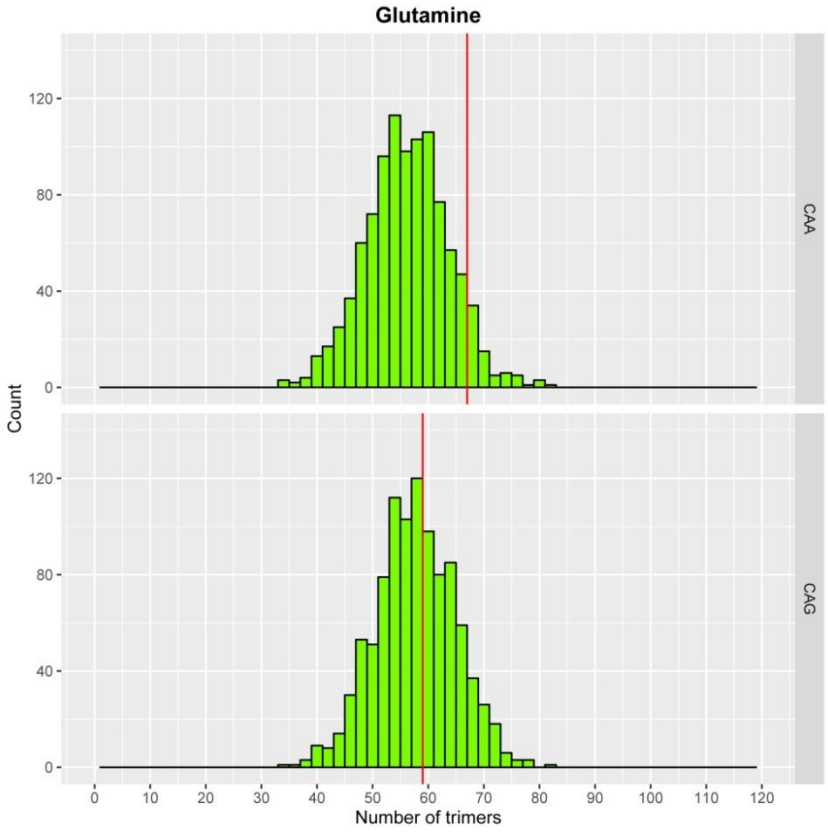
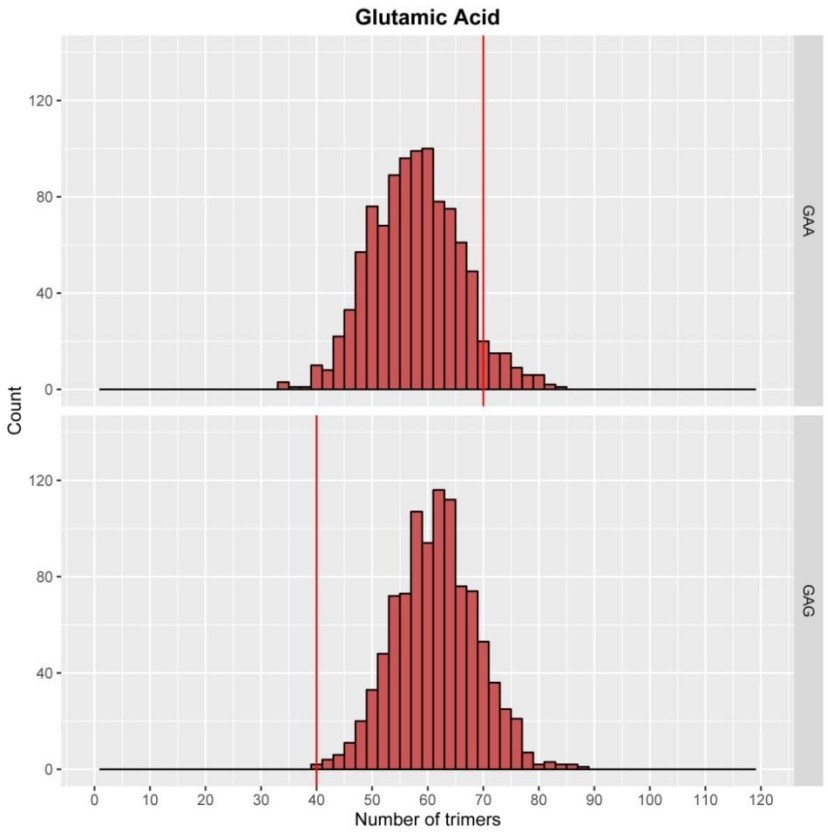

Glutamine
Count
CAA
CAG
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120
Number of trimers

Figure S10. Distributions of trimer counts encoding Glutamic Acid

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Glutamic Acid
Count
Number of trimers
GAA
GAG
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

B.

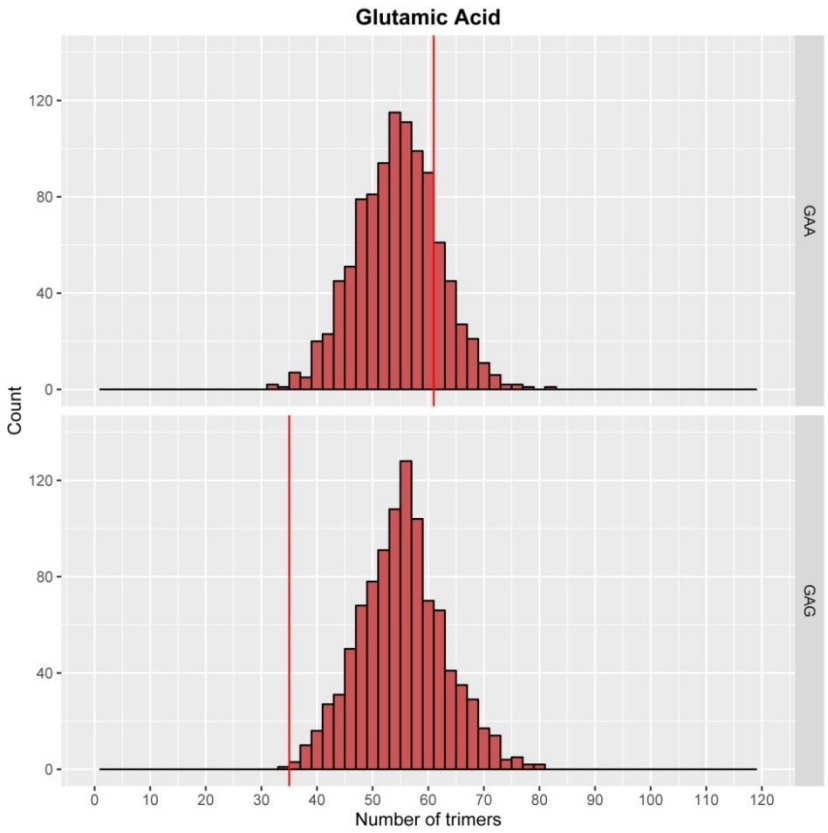
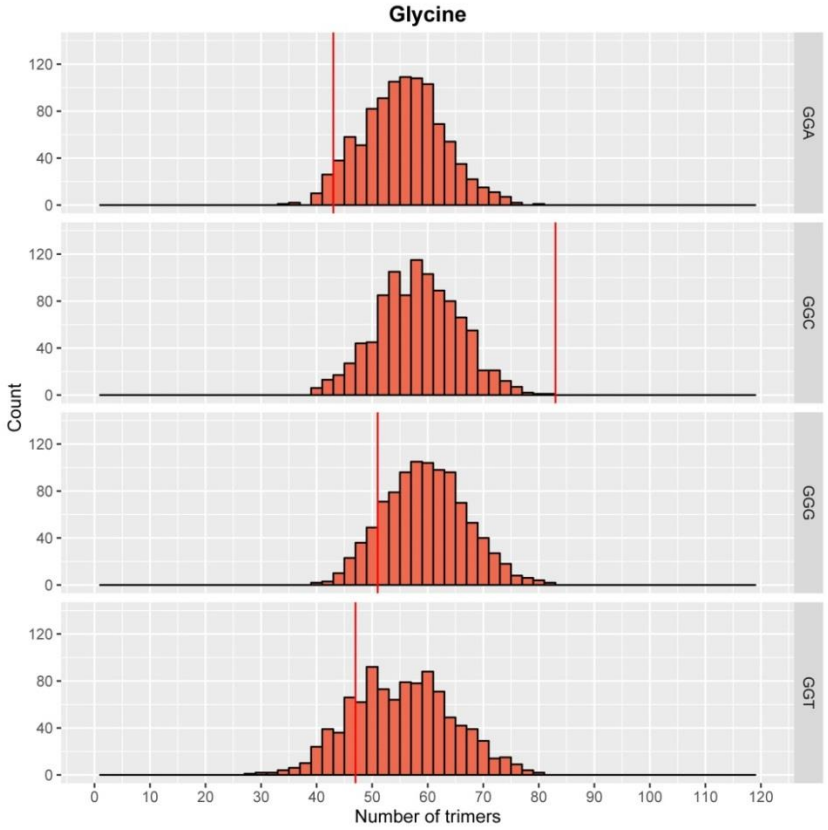

Glutamic Acid
Count
Number of trimers
GAA
GAG
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

Figure S11. Distributions of trimer counts encoding Glycine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Glycine
GGA
GGC
GGG
GGT
Count
Number of trimers

B.

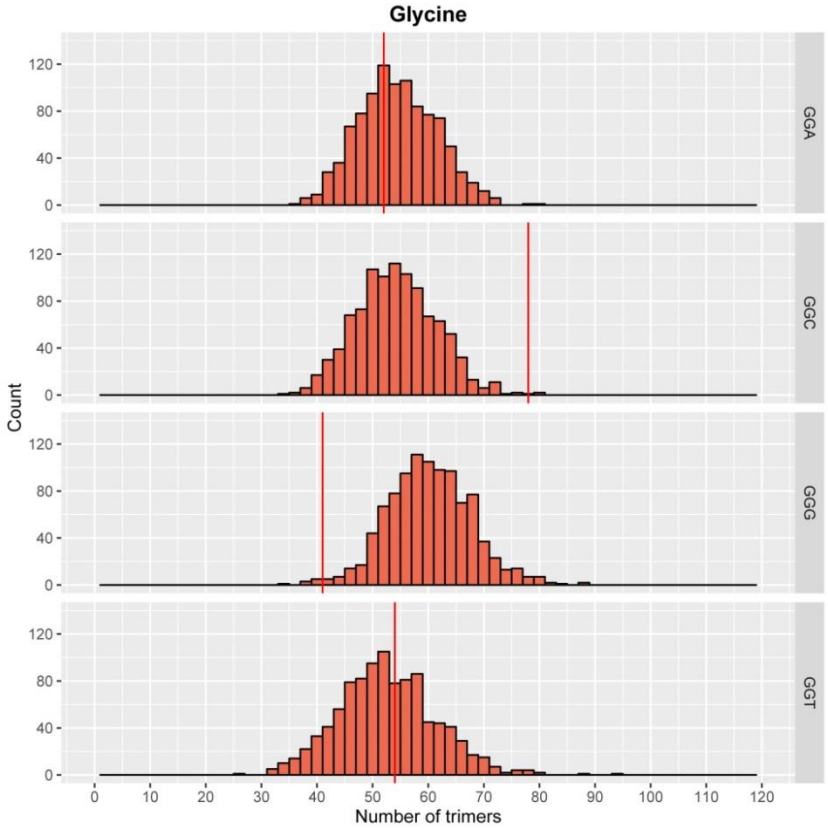
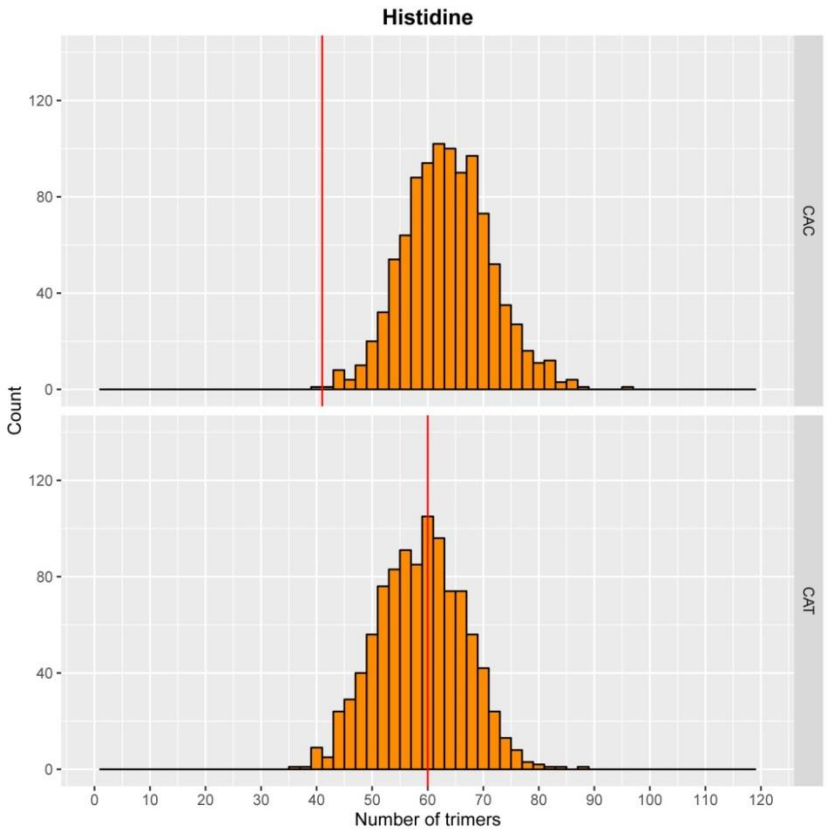

Glycine
GGA
GGC
GGG
GGT
Count
Number of trimers

Figure S12. Distributions of trimer counts encoding Histidine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Histidine
CAC
CAT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

B.

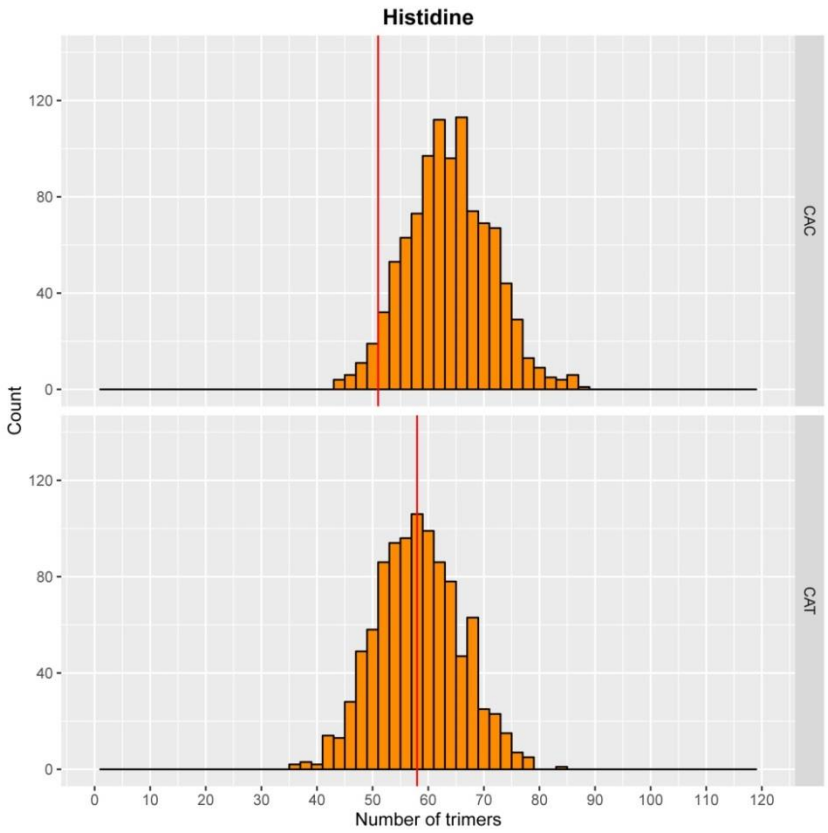
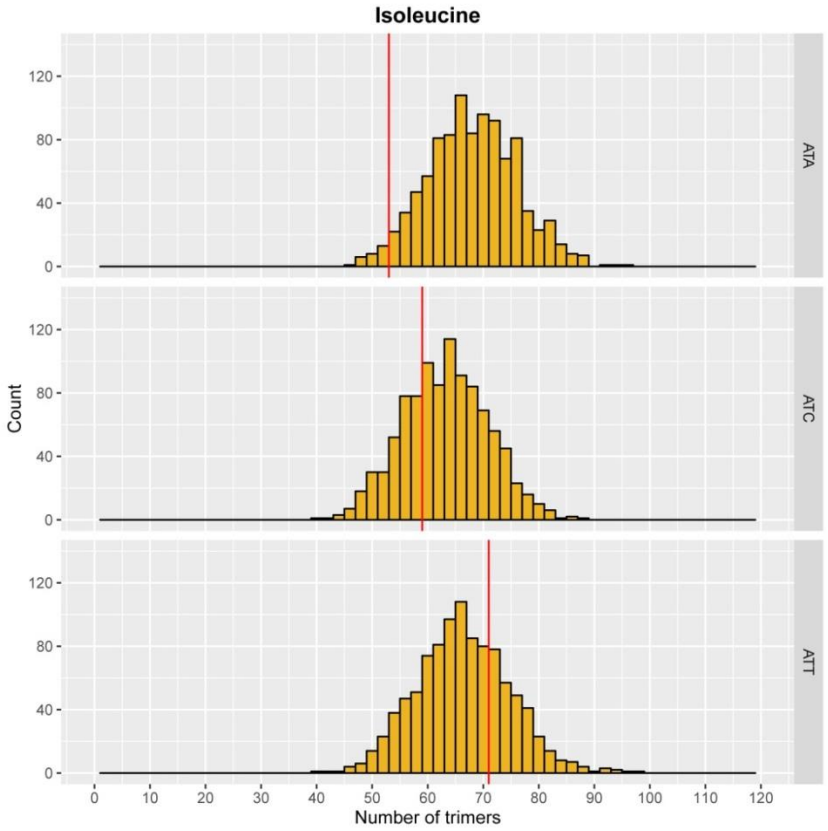

Histidine
CAC
CAT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

Figure S13. Distributions of trimer counts encoding Isoleucine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Isoleucine
ATA
ATC
ATT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

B.

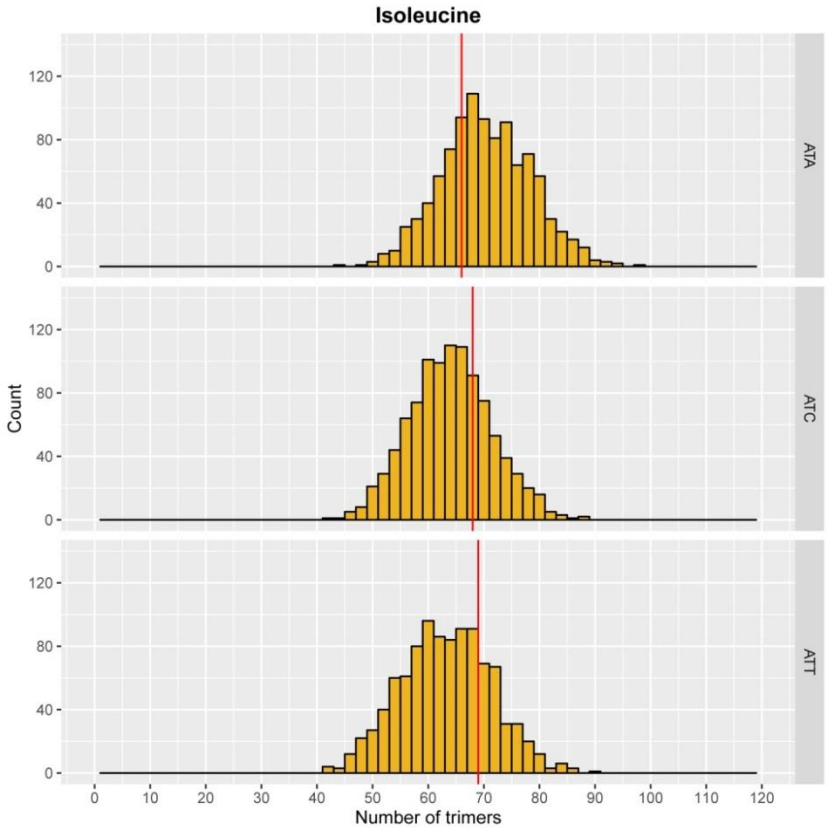
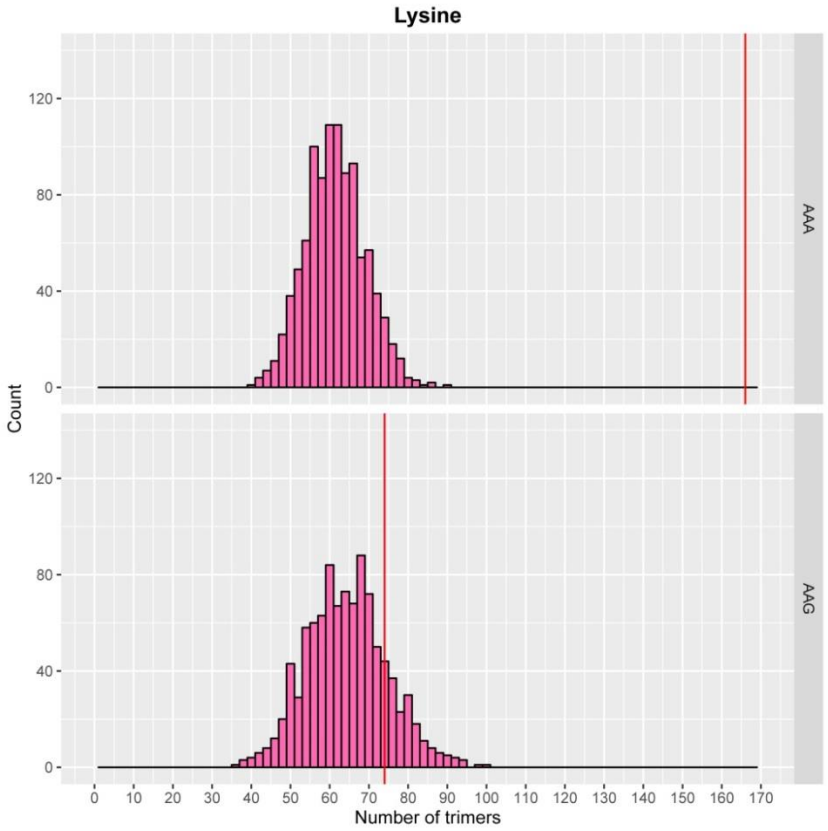

Isoleucine
ATA
ATC
ATT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

Figure S14. Distributions of trimer counts encoding Lysine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Lysine
Count
AAA
AAG
0
40
80
120
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170
Number of trimers

B.

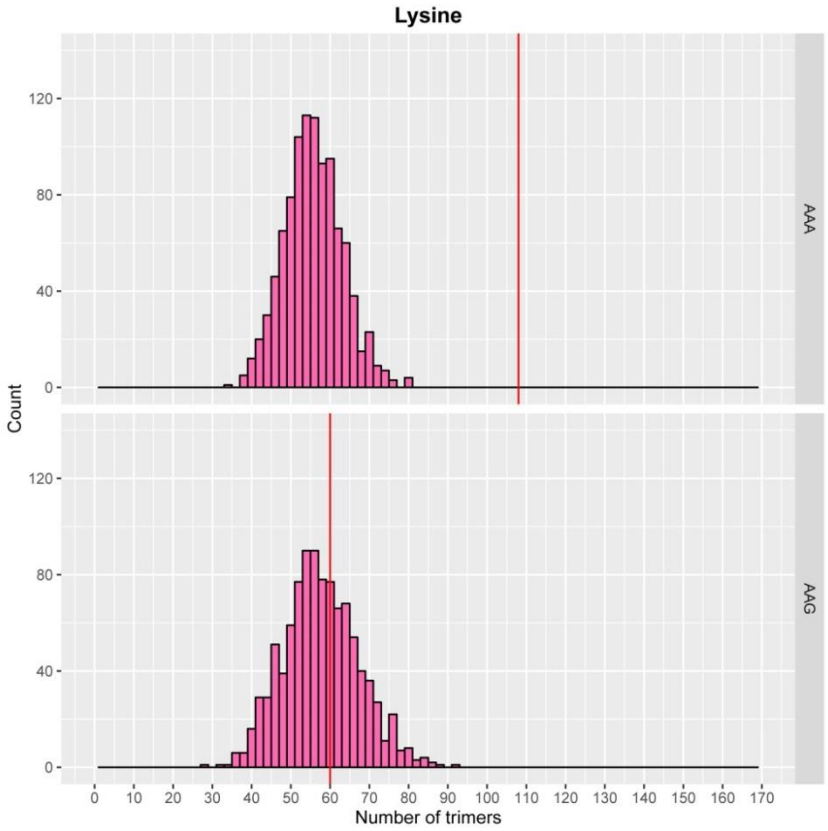
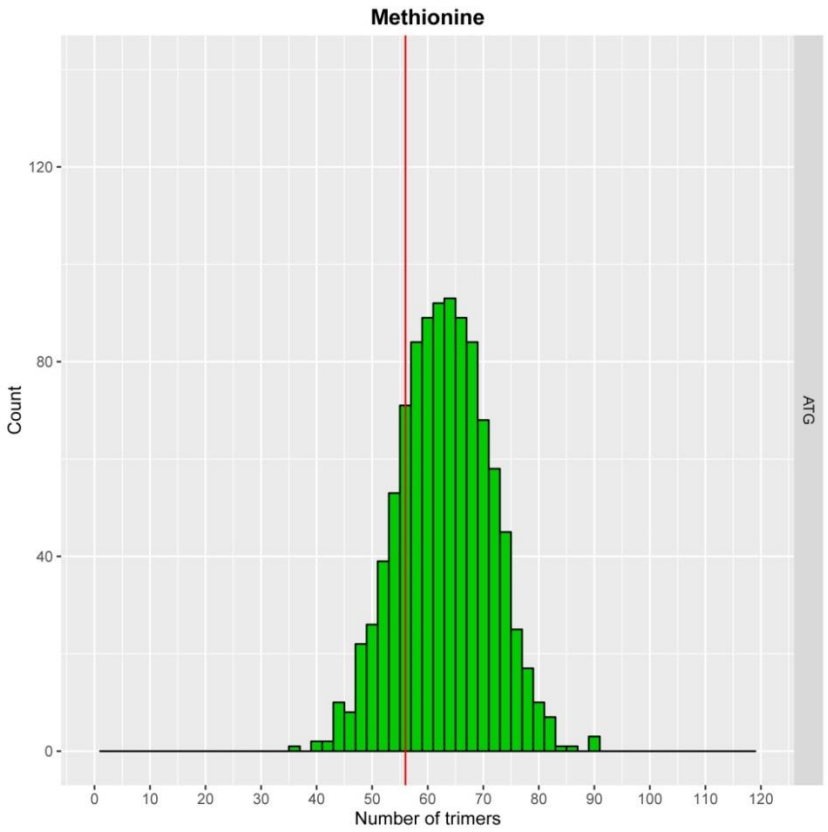

Lysine
Count
AAA
AAG
0
40
80
120
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170
Number of trimers

Figure S15. Distributions of trimer counts encoding Methionine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Methionine
Count
120
80
40
0
0 10 20 30 40 50 60 70 80 90 100 110 120
Number of trimers
ATG

B.

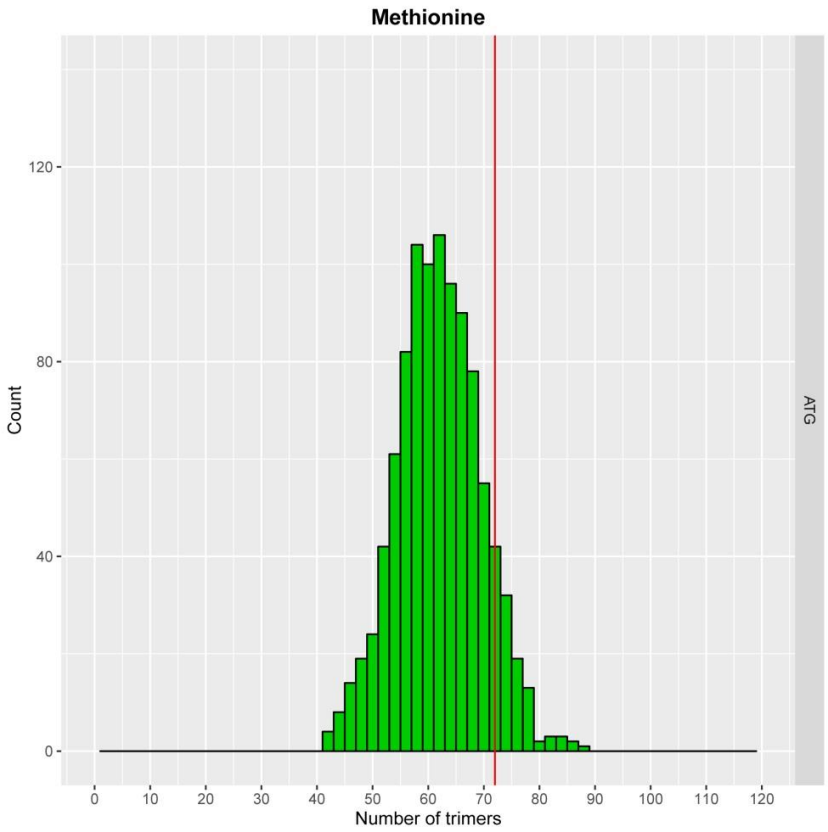
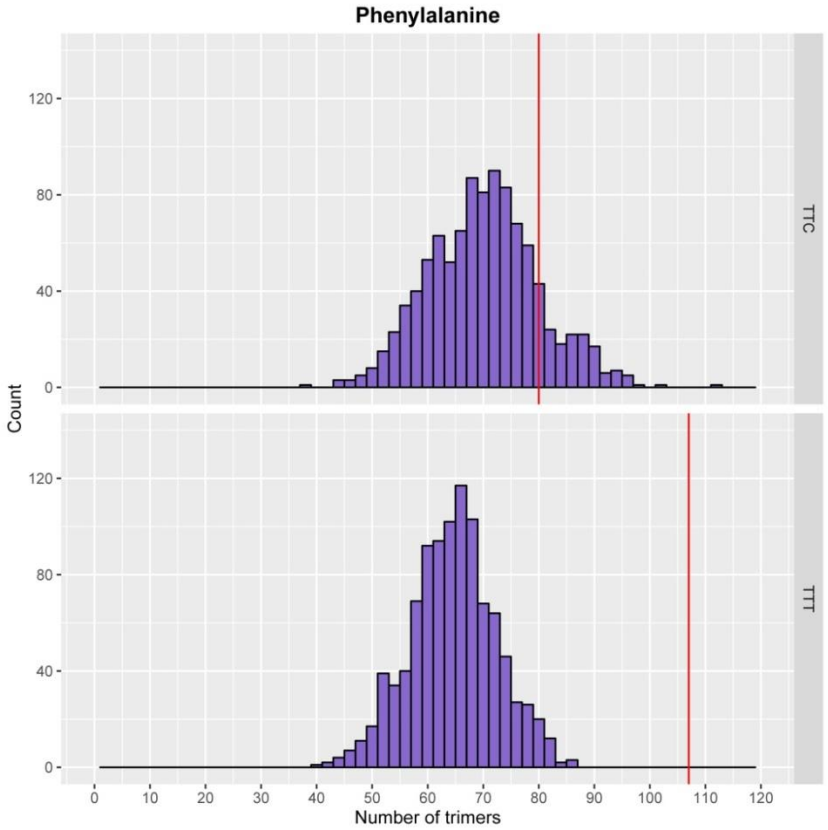

Methionine
Count
120
80
40
0
0 10 20 30 40 50 60 70 80 90 100 110 120
Number of trimers
ATG

Figure S16. Distributions of trimer counts encoding Phenylalanine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Phenylalanine
TTC
TTT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

B.

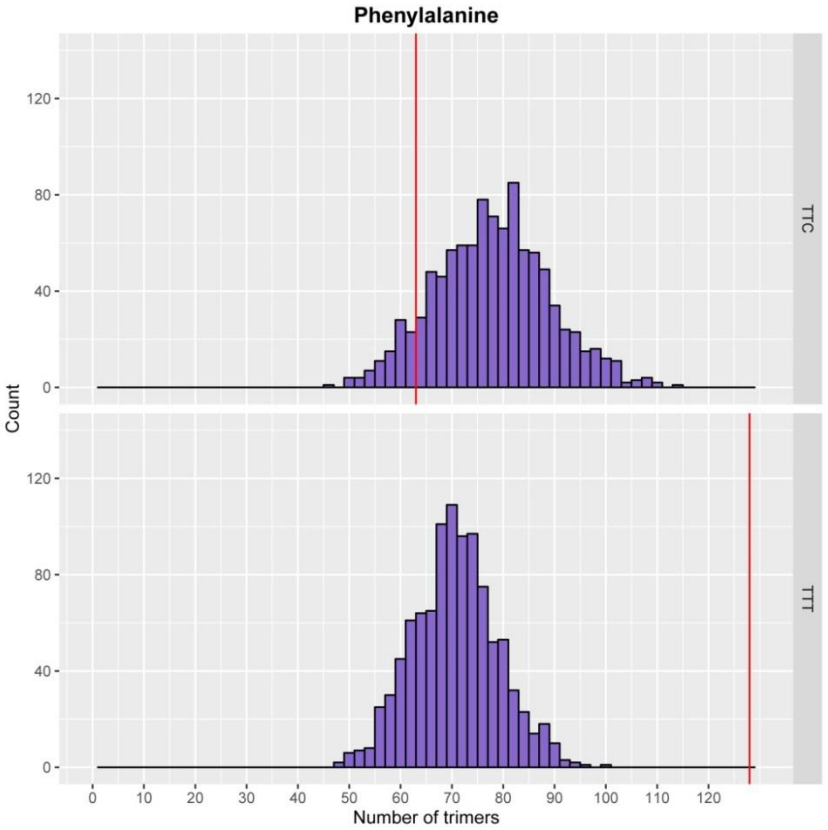
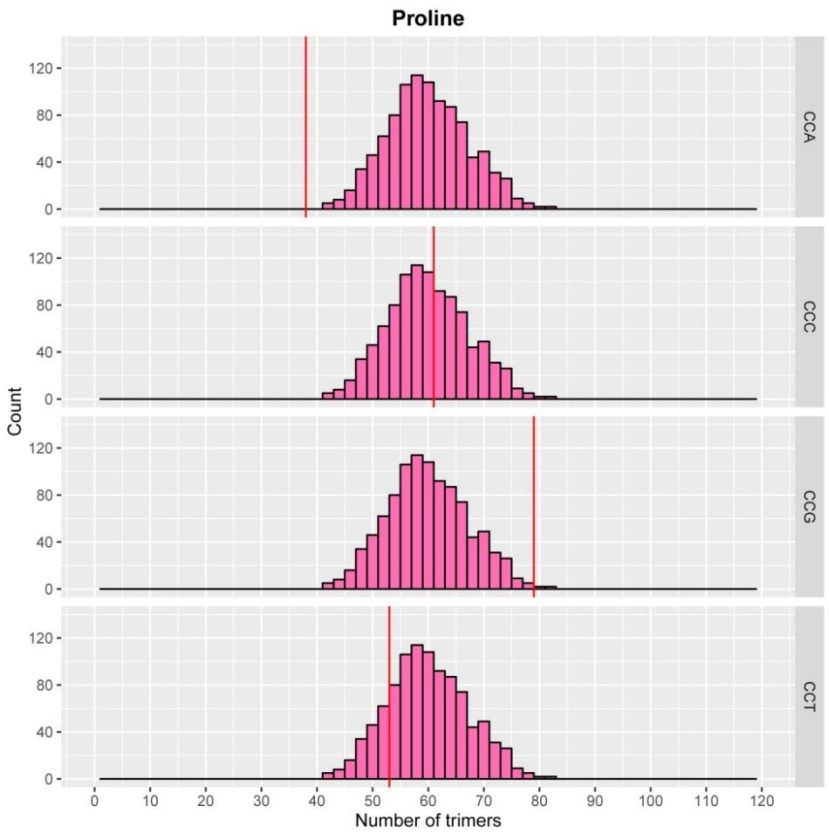

Phenylalanine
TTC
TTT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

Figure S17. Distributions of trimer counts encoding Proline

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Proline
CCA
CCC
CCG
CCT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

B.

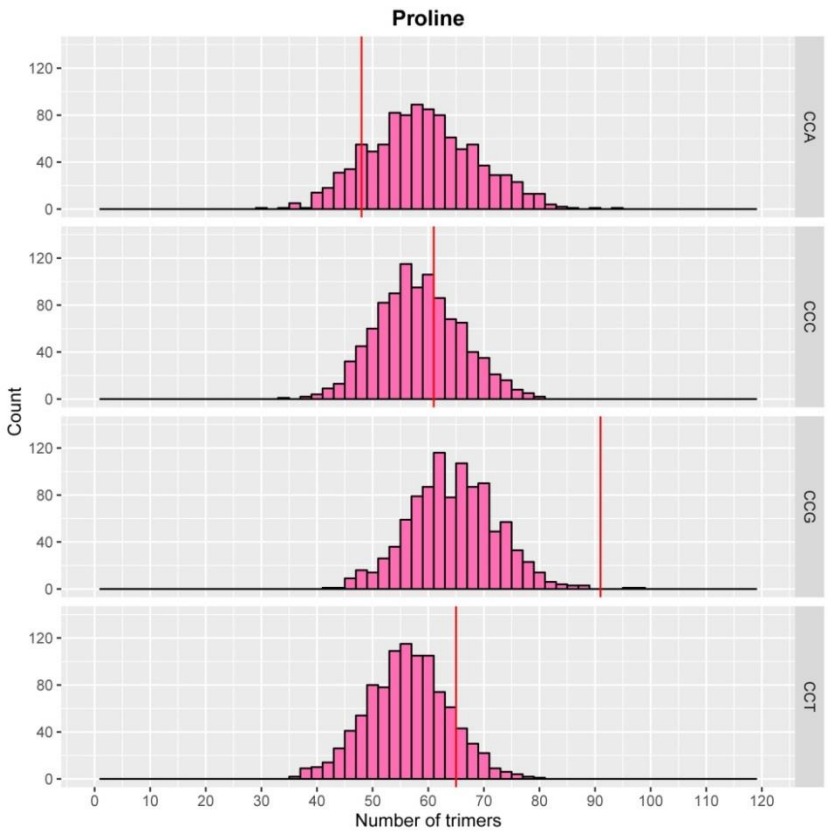
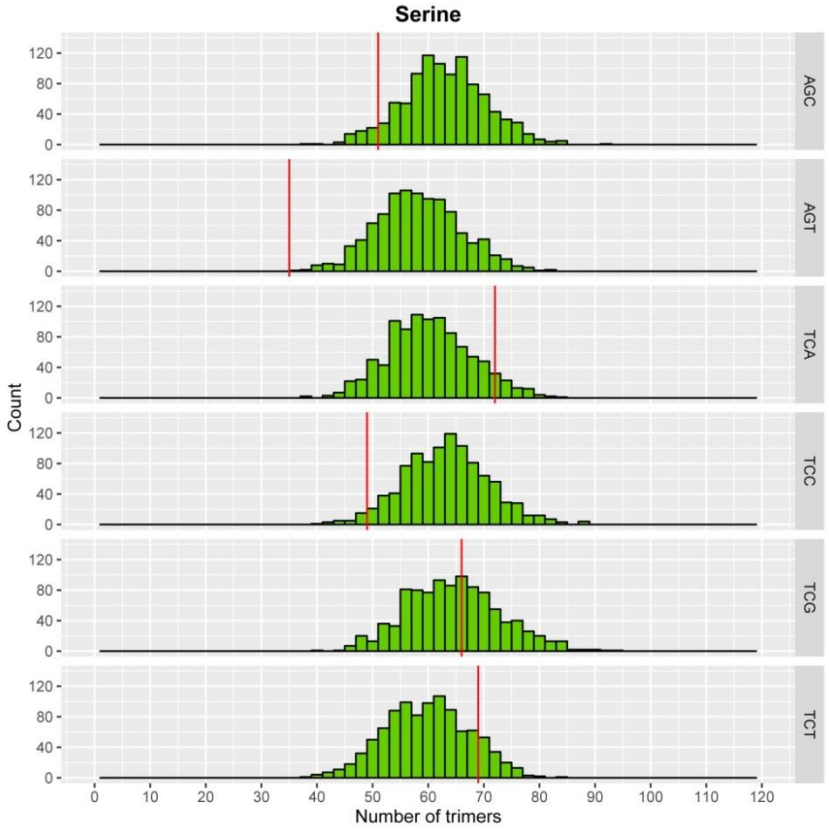

Proline
CCA
CCC
CCG
CCT
Count
Number of trimers
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120

Figure S18. Distributions of trimer counts encoding Serine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Serine
AGC
AGT
TCA
TCC
TCG
TCT
Count
Number of trimers

B.

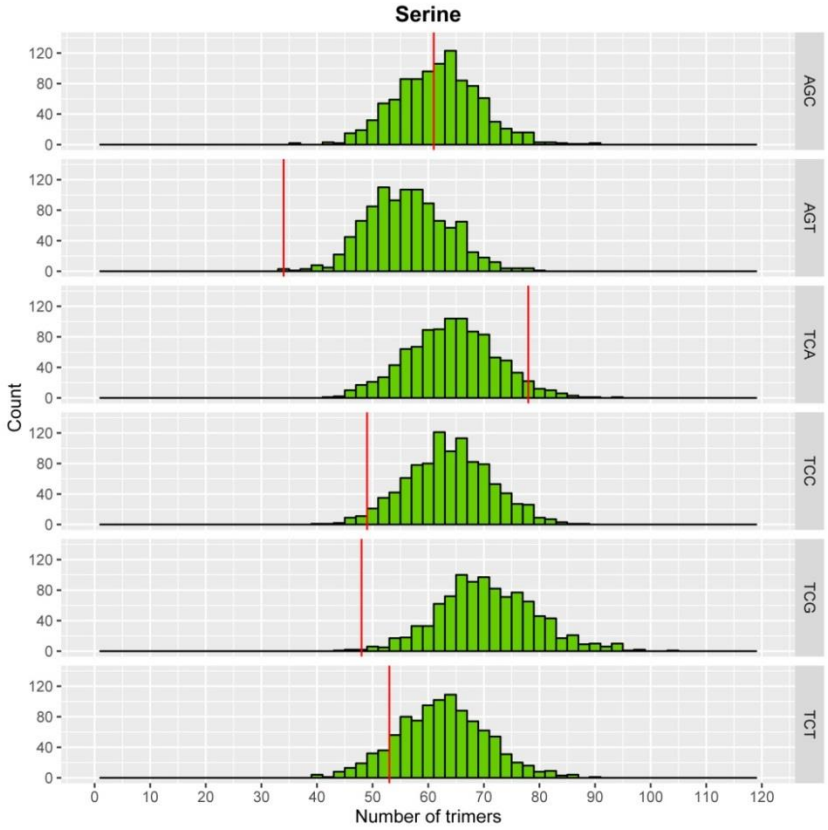
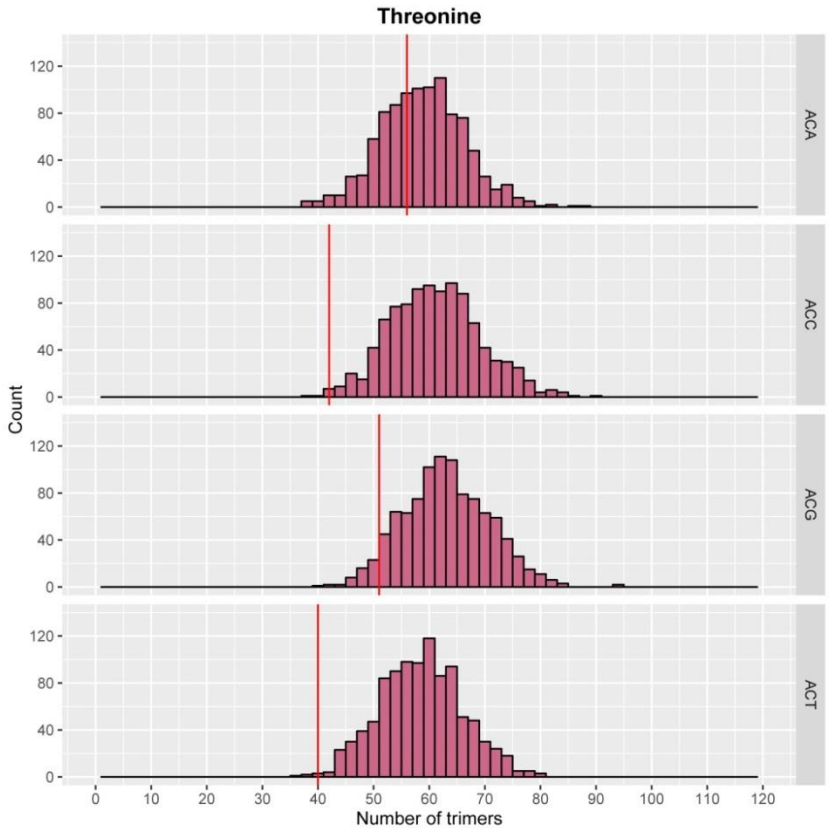

Serine
AGC
AGT
TCA
TCC
TCG
TCT
Count
Number of trimers

Figure S19. Distributions of trimer counts encoding Threonine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Threonine
ACA
ACC
ACG
ACT
Count
Number of trimers

B.

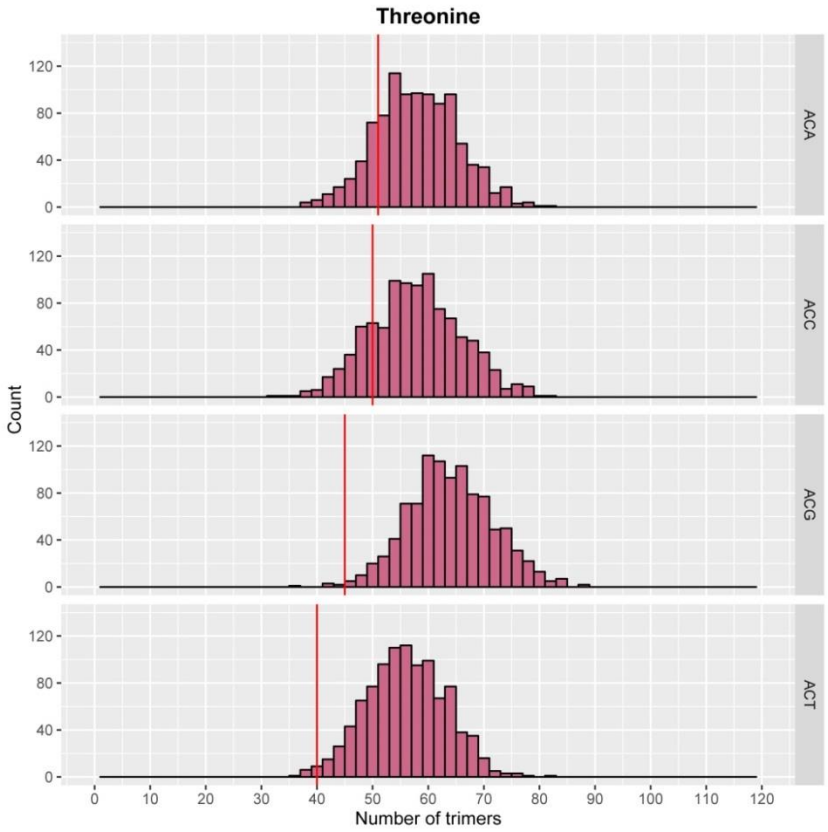
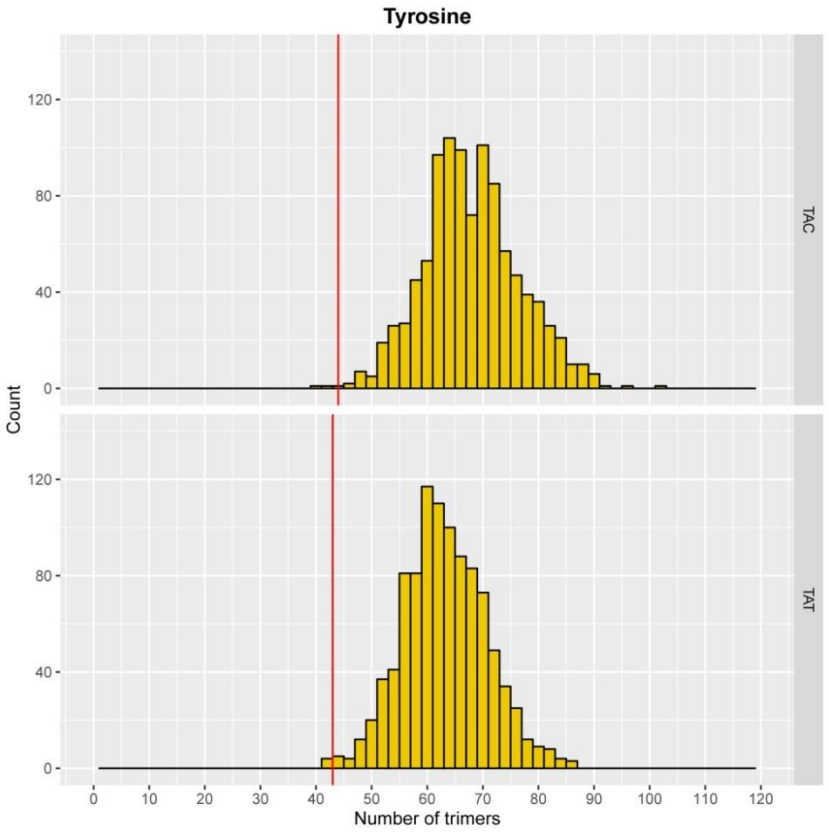

Threonine
ACA
ACC
ACG
ACT
Count
Number of trimers

Figure S20. Distributions of trimer counts encoding Tyrosine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Tyrosine
Count
TAC
TAT
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120
Number of trimers

B.

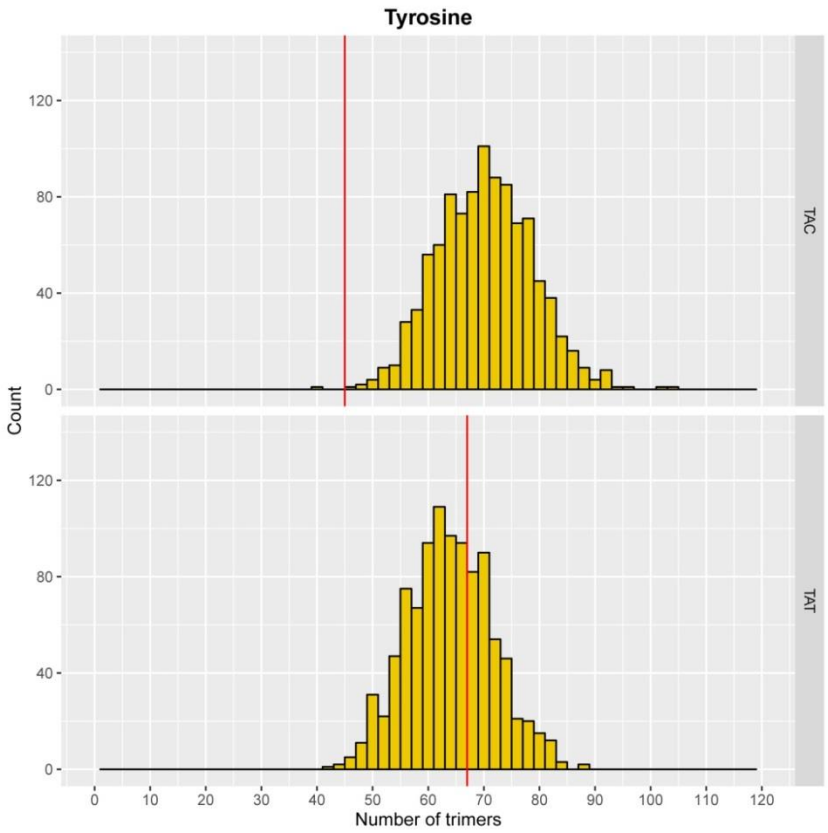
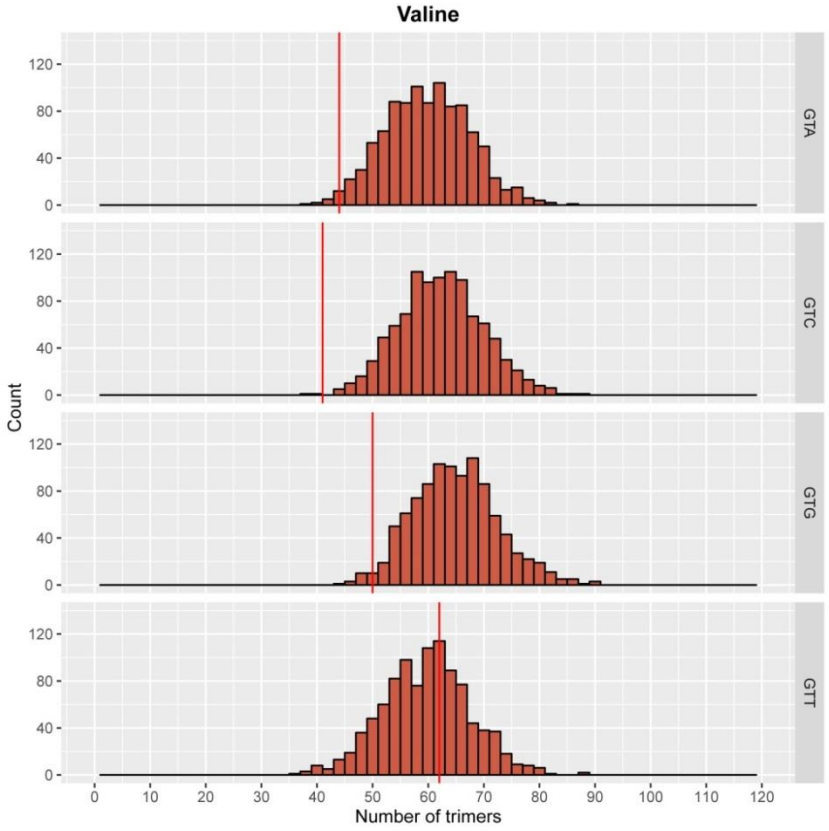

Tyrosine
Count
TAC
TAT
0
40
80
120
0
10
20
30
40
50
60
70
80
90
100
110
120
Number of trimers

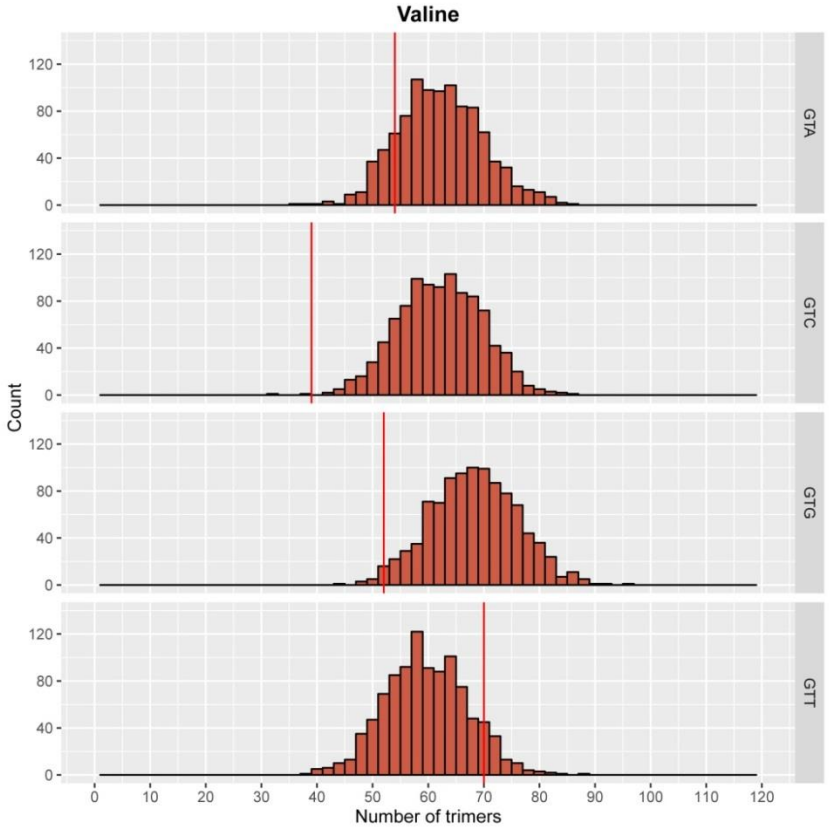
Figure S21. Distributions of trimer counts encoding Valine

A. Highly Expressed Genes; **B.** Lowly Expressed Genes.

A.


Valine
GTA
GTC
GTG
GTT
Count
Number of trimers

B.


Valine
GTA
GTC
GTG
GTT
Count
Number of trimers

V. Chapter 5 - Future Perspectives

5.1. Evaluate the impact on organismal and population fitness of co-option of 3' untranslated regions into translated bacterial gene products

As suggested by the research results presented in Chapter 4, the study of evolution of 3' UTRs in bacterial genomes from a translational perspective may hold great insights into the evolution of non-coding components of genomes and the proteome repertoire across the three domains of life. This potential research project challenges the traditional view of 3' UTRs as non-coding sequences limited to their regulatory roles at DNA and RNA level. A possible strategy for mitigating the effects of translational errors such as readthrough would consist in the additional expression of 3' untranslated regions in bacterial genomes. Across bacterial populations with large number of individuals, immediately adjacent 3' UTR sequences may encode amino acids with marginal or absent phenotypical effects upon error-prone expression of genes. It can be hypothesized that this strategy may be widespread in bacterial lineages and it may represent a mechanism of coding sequence evolution.

5.1.1. Rationale

Mutagenesis studies involving proteins have shown that, in some cases, despite the severity of mutations, protein function is preserved[58]. In bacteria, during the stationary phase, the cells experience an increase in the error rate of translation resulting in variant proteins with possible impaired structures and functions. As a consequence, the vast majority of mutations, for

example, result in non-synonymous changes that alter protein's primary sequence and therefore, its structure.

The neutral theory argues that, under the influence of natural selection, the existent genotypic variability mostly harbors neutral fitness changes while only a small proportion of changes are in fact beneficial [184]. In large populations, where the impact of natural selection on the genome sequence is fairly considerable, this would predict that the majority of existing sequence variants are in fact phenotypically neutral (they have neutral fitness effects). I predict that 3' UTR sequences when added to the original reading frames of proteins in bacteria with large (*Escherichia coli*) and very large populations (*Prochlorococcu*s spp.) will display neutral phenotypes when compared to the normal length proteins. I believe small phenotypic differentials at molecular level will not affect the overall the organismal fitness. To test this hypothesis, BipA protein in *Escherichia coli* can be used as a test case.

BipA is a 67.4 kDA protein with GTPase activity, closely related with EF-G, EF-Tu and LepA elongation factors. BipA contains a unique C terminal domain with an important role in binding the ribosome and is conserved across bacterial species. It is a regulator of several cellular processes in bacteria such as antimicrobial resistance [185], pathogenicity [186], motility [187], capsule formation [188], symbiosis [189] and growth at low temperature [190].

BipA makes an excellent candidate for testing effects of 3' UTR additions on the protein at molecular level as the *Escherichia coli* deletion mutants for BipA display no phenotype in normal growth conditions in rich medium [189], [190] despite the fact that BipA has been

predicted to be a highly expressed gene in given conditions based on preferred codon analysis [165], [191]. Also, any additions to the C terminal domain of BipA are likely to impact its binding to the ribosome.

5.1.2. Experimental Plan

To test the impact of incorporating adjacent 3' UTR encoded amino acid sequences into a protein sequence, a computational search similar to the one described in Chapter 4 can be carried out in *E.coli* and *Prochlorococcus* spp. reference genomes. Each 3' UTR sequence variant found in *E.coli* reference genomes (corresponding for *bipA* gene) may consist, for example, in a 10 amino acid sequence encoded by the first 30 nucleotides after the stop codon. BipA is a conserved protein across prokaryotic phyla, chimaeric polypeptides using the *E.coli bipA* reading frame and *Prochlorococcus* spp. orthologous 3' UTRs. The wild-type 3' UTR sequences can be cloned in frame with the BipA protein sequence at its C terminus in an expression plasmid. To test successful cloning, the plasmids can be sequenced using the Illumina Sequencer. The knock-out BipA *E.coli* BL21 (DE3) mutant cells can be transformed with expression plasmids carrying BipA variants with N-terminal tags suitable for assessing expression levels. Because BipA expression has no phenotypical effects in normal growth conditions but it does generate a phenotype during stressful conditions, low temperature can be used as a stressor. In this way, the growth curves of *E.coli* cells can be used to assess fitness of the bacterial populations.

To evaluate the direct fitness effects produced by inclusion of 3' UTR sequences in the BipA protein at the individual level, mutant cells containing the chimaeric polypeptides may be grown

in rich, liquid medium at low temperatures to stationary phase. The typical optical density at 600 nm for the stationary phase on the bacterial log growth curve is expected to be around 1-1.2. Upon collection of samples at the selected time points and desired optical density, viability of cells in the medium can be tested rapidly and reliably by using fluorescence based viability assay and flow cytometry. A type of viability assay can consist in differential nucleic acid staining of live and dead cells by Syto9 and propidium iodide dyes [192]–[194]. Syto9 is able to penetrate the cell membrane and to bind the nucleic acids in all cells (live and dead) while propidium iodide dislocates Syto 9 in dead cells as a result of damaged membranes. To this purpose, the LIVE/DEAD BacLight bacterial viability and counting kit combined with a flow cytometer of choice can be used to estimate the number of bacterial cells either dead or alive. These counts may provide a direct assessment of the fitness impact of elongated BipA protein variants.

An additional measure of the organismal fitness to be considered in this project is the growth rate of the bacterial populations. Cells containing chimaeric constructs can be grown in Luria-Bertani liquid culture at low temperatures for 8 hours. Samples can be taken at several time points during growth and the optical density at 600 nm of each culture can be measured.

Because organismal fitness is strongly influenced at low temperatures by the level of expression of BipA, another approach can measure the expression level of *bipA* in mutant cells. With that purpose, each protein variant can be tagged at N-terminus, for example, with a SNAP tag [195]. The small size of the tag will prevent misfolding and mislocalization of the fusion protein. By correlating the expression level of the protein variants with the measurements of the organismal fitness it may become possible to determine the level of phenotypic robustness conferred by the addition of 3’ UTRs to the protein amino acid sequence.

VI. Contributions

The following contributions from the authors listed below were included in this thesis.

Chapter 2. I carried out the data collection, the phylogenetic, computational and statistical analyses, participated in the experimental design, coordination and data interpretation and drafted the chapter manuscript. Timothy J. Harlow participated in the experimental design and helped to draft the chapter manuscript. Johann Peter Gogarten conceived the study, participated in its design and coordination and helped to draft the chapter manuscript. All authors read and approved the final chapter manuscript.

Chapter 3. I carried out the data collection computational and statistical analyses, participated in the coordination and data interpretation and drafted the chapter manuscript. Timothy J. Harlow helped to draft the chapter manuscript. Johann Peter Gogarten helped to draft the chapter manuscript. All authors read and approved the final chapter manuscript.

Chapter 4. I carried out the data collection, the phylogenetic, computational and statistical analyses, participated in the experimental design, coordination and data interpretation and drafted the chapter manuscript. Timothy J. Harlow participated in the data collection, experimental design and helped to draft the chapter manuscript. Johann Peter Gogarten participated in the design and coordination of the research and helped draft the chapter manuscript. All authors read and approved the final chapter manuscript.

VII. Appendices

7.1.. Perl Script for Counting Changes in a DNA Multiple Sequence Alignment


```
#!/usr/perl env -w
use strict;

unless(@ARGV == 1) {die "informative.pl <alignment>\n";}
my $alignment_file = $ARGV[0];
my $max_sequence_number = 0;
my @sequence = ();              #the sequence

open (IN, "< $alignment_file") || die "can't open $alignment_file\n";

while (<IN>) {
   chomp;

   if (/^>/) {
      #this is the ">..." line
      $max_sequence_number++;
                                 $sequence[$max_sequence_number] = "";
   }
   else {
      #this is part of the sequence
      $sequence[$max_sequence_number] .= $_;
   }
}
close (IN);

my $j;
my $i;
my $aa;
```

```
my %seen = ();
my $count;

for ($j = 0; $j < length($sequence[1]); $j++) {                    #loop over all columns of the alignment
   for ($i = 1; $i <= $max_sequence_number; $i++) {          #sequence numbering is from 1 to n
                                    $aa = substr($sequence[$i], $j, 1);                    #extract j-th amino acid of i-th sequence
                                    if ($aa ne "-") {                                    #ignore gaps
         $seen{$j}{$aa} = 1;                                  #remember the amino acids in this column
                                    }
   }
}
print "column\tunique\tcount\tcount-1\n";

for ($j = 0; $j < length($sequence[1]); $j++) {
   print $j+1;                        #0 in the array is actually the 1st column, and so on
   print "\t";
   foreach $aa (sort keys %{$seen{$j}}) {
      print $aa;
   }
   print "\t";
   $count = keys %{$seen{$j}};
   print $count;
   print "\t";
   print $count - 1;
   print "\n";
}
```

7.2. Perl Script For Simulating Random Changes in a DNA Multiple Sequence Alignment


```
#!/bin/env perl
use strict;

my $trial = 0;

open (OUT, "> sims.txt"); #output to file
my %countsyn = ();

while (1) {

$trial++;
my $stop = 0;
my $synonymous = 0;
my $nonsynonymous = 0;
my $nochange = 0;

# This program takes a DNA multiple sequence alignment as input.

open (IN, "< Eco_cps_aln.fna");        #the input alignment

#Step 1: read the alignment file into a hash
#The index of the hash is the sequence name in the file.
#Sequences are numbered as they appear in the file, starting at 1.

my %seq = ();                    #hash containing the sequences
```

```
my $seqcount = 0;            #to count the sequences
my %seqname = ();

while (<IN>) {                                  #while there are lines in the file, read one at a time
   chomp;                                       #eat newlines at the end of each line

   if (/>(\w+)/) {                          #if this is a sequence name line
      $seqcount++;
      $seqname{$seqcount} = $1;             #store the name of each sequence
   }
   else {                                       #else it must be the sequence itself
      $seq{$seqname{$seqcount}} .= $_;      #sequence may span multiple lines
   }
}

#Step 2:

my %basefreq = ();
my %basefreq123 = ();
my %mutantbasefreq = ();
my $basetotal = 0;
my %basetotal123 = 0;
my $basenumber = 0;

for (my $i = 1; $i <= $seqcount; $i++) {
   my @base = split //, $seq{$seqname{$i}};

   for (my $j = 1; $j <= @base; $j++) {
      my $b = $base[$j - 1];

                              if ($b ne "-") {
         $basefreq{$b}++;
                                 $basetotal++;
```

```
        $basenumber = $j % 3;

        if ($basenumber == 0) {
          $basenumber = 3;
        }
                                $basetotal123{$basenumber}++;
                                $basefreq123{$basenumber}{$b}++;
                            }
    }
}

foreach my $b (keys %basefreq) {
   $basefreq{$b} = $basefreq{$b} / $basetotal;
}
my $alnlength = length($seq{$seqname{1}});

my $randomseq = 0;
my $codon;
my $thebase;
my $newbase;
my $newcodon;

########for ($col = 1; $col <= $alnlength; $col++) {          #go through each column of the alignment
for (my $col = 1; $col <= 20; $col++) {         #go through each column of the alignment;input number of given mutations
   my $realcol = int(rand($alnlength)) + 1;
   do {                                                     #do this until we select something that isn't a gap
      $randomseq = int(rand($seqcount)) + 1;            #random number between 1 and seqcount, inclusive
      $basenumber = $realcol % 3;                       #which base of the codon is it? 0 (last), 1 (first), or 2 (middle)

      if ($basenumber == 0) {                   #renumber last position from 0 to 3
        $basenumber = 3;
      }
      $codon = substr($seq{$seqname{$randomseq}}, $realcol - $basenumber, 3);
```

```
                                                    #get the entire codon that this column is passing through
                                                    #  for the chosen sequence
                                                    #  if that happens to be a gap, we'll try again
    } until ($codon ne "---");

    my $origbase = substr($codon, $basenumber - 1, 1);

    do {
       my $randombase = rand(1);                    #a real number from 0 to 1

       foreach $thebase (keys %basefreq) {          #keys are each of the bases AGCT
          if ($randombase < $basefreq{$thebase}) {  #choose this base, depending on how likely the background is
             $newbase = $thebase;
                                  last;                                 #we've chosen the mutant base identity at this point, so exit out of
                                                                              #  this loop
                                  }
                                  else {
                                     $randombase -= $basefreq{$thebase};      #we didn't choose this base, so instead of the random
number being
                                  }
       }
    } until ($origbase ne $newbase);

    $newcodon = $codon;
    substr $newcodon, $basenumber - 1, 1, $newbase;
                                                    #substitute in the mutant base at the appropriate position
    $mutantbasefreq{$newbase}++;            #keep a count of how many times each of the four bases was chosen
                                                    #  as the mutant

    #The Bacterial, Archaeal and Plant Plastid Code (transl_table=11)
    my $AAs    = "FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG";
    my $Starts = "---M---------------M------------MMMM---------------M------------";
    my $Base1  = "TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG";
```

```
my $Base2  = "TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG";
my $Base3  = "TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG";

my @aasarray = split //, $AAs;
my @base1array = split //, $Base1;
my @base2array = split //, $Base2;
my @base3array = split //, $Base3;

my %codontable = ();
my $c;

for (my $i = 1; $i <= 64; $i++) {
                        $c = $base1array[$i - 1].$base2array[$i - 1].$base3array[$i - 1];
                        $codontable{$c} = $aasarray[$i - 1];
}
#print "$realcol\t$randomseq\t$basenumber";
#print "\t$codon\t$codontable{$codon}";
#print "\t$newcodon\t$codontable{$newcodon}\t";

if ($codon eq $newcodon) {
   $nochange++;
                        #print "Same\n";
}
elsif ($codontable{$codon} eq $codontable{$newcodon}) {
                        $synonymous++;
                        #print "S\n";
}
elsif ($codontable{$newcodon} eq "*") {
   $stop++;
   #print "Ter\n";
}
else {
   $nonsynonymous++;
```

```
                                    #print "N\n";
    }
}
#print "--\n";
my $SandNandTer = $synonymous + $nonsynonymous + $stop;
my $Nt = $nonsynonymous + $stop;

#foreach $b (keys %basefreq) {
#    print "$b\t$basefreq{$b}\n";
#}

foreach $b (keys %{$basefreq123{1}}) {
    $basefreq123{1}{$b} = $basefreq123{1}{$b} / $basetotal123{1};
}

foreach $b (keys %{$basefreq123{2}}) {
    $basefreq123{2}{$b} = $basefreq123{2}{$b} / $basetotal123{2};
}

foreach $b (keys %{$basefreq123{3}}) {
    $basefreq123{3}{$b} = $basefreq123{3}{$b} / $basetotal123{3};
}

foreach $b (keys %mutantbasefreq) {
    $mutantbasefreq{$b} = $mutantbasefreq{$b} / $alnlength;
}

#foreach $seqnumber (sort {$a <=> $b} keys %seqname) {
#    print "$seqnumber\t$seqname{$seqnumber}\n";
#}

print OUT "$trial\t$synonymous\n";
$countsyn{$synonymous}++;
```

```
if ($trial % 100 == 0) {
   print "after $trial trials...\n";
   print "syn\tcount\n";

   foreach my $key (sort {$a <=> $b} keys %countsyn) {
      print "$key\t";
                                 print $countsyn{$key};
                                 print "\n";
   }

}

}
```

7.3. Article Permission Use (pertaining to Chapter 3)

Article permission use

Jones, Jennifer (ELS-OXF) <J.Jones@elsevier.com> 6 December 2016 at 07:00
To: Seila Omer <seila.omer@uconn.edu>
Cc: JohannPeter Gogarten <gogarten@uconn.edu>

Dear Seila Omer

Thank you for your email. As author of the requested article, you do not need to seek Elsevier's permission to include it / material from it in your thesis as it is part of the rights you retain as an Elsevier journal author. And if your thesis is being posted online, then the article should be embedded in the thesis and not as a standalone item.

For further information on the rights you retain as an Elsevier journal author, please visit our web page http://www.elsevier.com/about/company-information/policies/copyright.

Yours sincerely
Jennifer Jones
Permissions Specialist
Global Rights Department

Elsevier Ltd
PO Box 800
Oxford OX5 1GB
UK

Elsevier Limited, a company registered in England and Wales with company number 1982084, whose registered office is The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, United Kingdom.

From: On Behalf Of Seila Omer
Sent: 01 December 2016 01:28
To: Rights and Permissions (ELS)
Cc: JohannPeter Gogarten
Subject: Article permission use

Hello!

My name is Seila Omer and I am the first author of a paper recently published online in "Trends in Microbiology"- "Does Sequence Conservation Provide Evidence for Biological Function?" DOI: http://dx.doi.org/10.1016/j.tim.2016.09.010

I would like to include the article in my PhD thesis and I would like your permission to do so. The thesis will be included in the institutional digital repository at University of Connecticut (http://digitalcommons.uconn.edu/). The thesis can be placed under embargo.

Please let me know.

Thank you!

Sincerely,

Seila Omer

Journal author rights

In order for Elsevier to publish and disseminate research articles, we need publishing rights. This is determined by a publishing agreement between the author and Elsevier. This agreement deals with the transfer or license of the copyright to Elsevier and authors retain significant rights to use and share their own published articles. Elsevier supports the need for authors to share, disseminate and maximize the impact of their research and these rights, in Elsevier proprietary journals* are defined below:

For subscription articles	For open access articles
Authors transfer copyright to the publisher as part of a journal publishing agreement, but have the right to: Share their article for Personal Use, Internal Institutional Use and Scholarly Sharingpurposes, with a DOI link to the version of record on ScienceDirect (and with the Creative Commons CC-BY-NC- ND license for author manuscript versions) Retain patent, trademark and other intellectual property rights (including research data). Proper attribution and credit for the published work.	Authors sign an exclusive license agreement, where authors have copyright but license exclusive rights in their article to the publisher**. In this case authors have the right to: Share their article in the same ways permitted to third parties under the relevant user license (together with Personal Use rights) so long as it contains a CrossMark logo, the end user license, and a DOI link to the version of record on ScienceDirect. Retain patent, trademark and other intellectual property rights (including research data). Proper attribution and credit for the published work.

*Please note that society or third party owned journals may have different publishing agreements. Please see the journal's guide for authors for journal specific copyright information.

**This includes the right for the publisher to make and authorize commercial use, please see "Rights granted to Elsevier" for more details.

7.4. List of *E. coli* genomes (Chapter 4)

NCBI Genome ID	Genome Name
387605479	Escherichia_coli_042_uid161985
110640213	Escherichia_coli_536_uid58531
218693476	Escherichia_coli_55989_uid59383
253771435	Escherichia_coli__BL21_Gold_DE3_pLysS_AG__uid59245
386632422	Escherichia_coli__clone_D_i14__uid162049
386627502	Escherichia_coli__clone_D_i2__uid162047
386637352	Escherichia_coli_ABU_83972_uid161975
117622295	Escherichia_coli_APEC_O1_uid58623
443615330	Escherichia_coli_APEC_O78_uid187277
170018061	Escherichia_coli_ATCC_8739_uid58783
254160123	Escherichia_coli_B_REL606_uid58803
387825439	Escherichia_coli_BL21_DE3__uid161947
387823261	Escherichia_coli_BL21_DE3__uid161949
238899406	Escherichia_coli_BW2952_uid59391
26245917	Escherichia_coli_CFT073_uid57915
386593590	Escherichia_coli_DH1_uid161951
387619774	Escherichia_coli_DH1_uid162051
157154711	Escherichia_coli_E24377A_uid58395
218687878	Escherichia_coli_ED1a_uid59379
387610477	Escherichia_coli_ETEC_H10407_uid161993
157159467	Escherichia_coli_HS_uid58393
218552585	Escherichia_coli_IAI1_uid59377
218698419	Escherichia_coli_IAI39_uid59381
386597751	Escherichia_coli_IHE3034_uid162007
556550243	Escherichia_coli_JJ1886_uid226103
170079663	Escherichia_coli_K_12_substr__DH10B_uid58979
471332236	Escherichia_coli_K_12_substr__MDS42_uid193705
556503834	Escherichia_coli_K_12_substr__MG1655_uid57779
388476123	Escherichia_coli_K_12_substr__W3110_uid161931
386698504	Escherichia_coli_KO11FL_uid162099
378710836	Escherichia_coli_KO11FL_uid52593
222154829	Escherichia_coli_LF82_uid161965
544388862	Escherichia_coli_LY180_uid219461
386617516	Escherichia_coli_NA114_uid162139
260842239	Escherichia_coli_O103_H2_12009_uid41013
410480139	Escherichia_coli_O104_H4_2009EL_2050_uid175905
407466711	Escherichia_coli_O104_H4_2009EL_2071_uid176128
407479587	Escherichia_coli_O104_H4_2011C_3493_uid176127
260866153	Escherichia_coli_O111_H__11128_uid41023
215485161	Escherichia_coli_O127_H6_E2348_69_uid59343
209395693	Escherichia_coli_O157_H7_EC4115_uid59091
16445223	Escherichia_coli_O157_H7_EDL933_uid57831
254791136	Escherichia_coli_O157_H7_TW14359_uid59235
15829254	Escherichia_coli_O157_H7_uid57781
260853213	Escherichia_coli_O26_H11_11368_uid41021
291280824	Escherichia_coli_O55_H7_CB9615_uid46655
387504934	Escherichia_coli_O55_H7_RM12579_uid162153
386622414	Escherichia_coli_O7_K1_CE10_uid162115
387615344	Escherichia_coli_O83_H1_NRG_857C_uid161987
386703215	Escherichia_coli_P12b_uid162061
544574430	Escherichia_coli_PMV_1_uid219679
218556939	Escherichia_coli_S88_uid62979
209917191	Escherichia_coli_SE11_uid59425
387828053	Escherichia_coli_SE15_uid161939
170679574	Escherichia_coli_SMS_3_5_uid58919
386602643	Escherichia_coli_UM146_uid162043
218703261	Escherichia_coli_UMN026_uid62981
386612163	Escherichia_coli_UMNK88_uid161991
91209055	Escherichia_coli_UTI89_uid58541
386607309	Escherichia_coli_W_uid162011
386707734	Escherichia_coli_W_uid162101
387880559	Escherichia_coli_Xuzhou21_uid163995

VIII. Bibliography

[1] S. Koskiniemi, S. Sun, O. G. Berg, and D. I. Andersson, “Selection-driven gene loss in bacteria,” *PLoS Genet*, vol. 8, no. 6, p. e1002787, Jun. 2012.

[2] C.-H. Kuo, N. A. Moran, and H. Ochman, “The consequences of genetic drift for bacterial genome complexity.,” *Genome Res.*, vol. 19, no. 8, pp. 1450–4, 2009.

[3] R. Hershberg and D. A. Petrov, “Evidence that mutation is universally biased towards AT in bacteria.,” *PLoS Genet.*, vol. 6, no. 9, p. e1001115, Sep. 2010.

[4] B. R. Levin and O. E. Cornejo, “The population and evolutionary dynamics of homologous gene recombination in bacteria,” *PLoS Genet*, vol. 5, no. 8, p. e1000601, Aug. 2009.

[5] E. Denamur and I. Matic, “Evolution of mutation rates in bacteria,” *Mol. Microbiol.*, vol. 60, no. 4, pp. 820–827, 2006.

[6] J. W. Drake, “A constant rate of spontaneous mutation in DNA-based microbes.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 88, no. 16, pp. 7160–7164, 1991.

[7] C. P. Andam, D. Williams, and J. P. Gogarten, “Biased gene transfer mimics patterns created through shared ancestry.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 23, pp. 10679–84, Jun. 2010.

[8] L. Olendzenski and J. P. Gogarten, “Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer.,” *Ann. N. Y. Acad. Sci.*, vol. 1178, pp. 137–45, Oct. 2009.

[9] G. Bertani, “Transduction-like gene transfer in the methanogen Methanococcus voltae,” *J. Bacteriol.*, vol. 181, no. 10, pp. 2992–3002, May 1999.

[10] I. Gordo, L. Perfeito, and A. Sousa, “Fitness effects of mutations in bacteria.,” *J. Mol. Microbiol. Biotechnol.*, vol. 21, no. 1–2, pp. 20–35, Jan. 2011.

[11] Z. Yang, “PAML 4: phylogenetic analysis by maximum likelihood,” *Mol. Biol. Evol.*, vol. 24, no. 8, pp. 1586–1591, 2007.

[12] J. P. Huelsenbeck and F. Ronquist, “MRBAYES: Bayesian inference of phylogenetic trees,” *Bioinforma.* , vol. 17, no. 8, pp. 754–755, Aug. 2001.

[13] S. L. K. Pond, S. D. W. Frost, and S. V Muse, “HyPhy: hypothesis testing using phylogenies,” *Bioinformatics*, vol. 21, no. 5, pp. 676–679, 2005.

[14] R. A. Fisher, *The genetical theory of natural selection.* Oxford, England: Clarendon Press, Oxford, England, 1930.

[15] H. A. Orr *et al.*, “The distribution of fitness effects among beneficial mutations.,” *Genetics*, vol. 163, no. 4, pp. 1519–26, 2003.

[16] P. D. Keightley, “Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: A simulation study,” *Genetics*, vol. 150, no. 3, pp. 1283–1293, 1998.

[17] J. E. Barrick, M. R. Kauth, C. C. Strelioff, and R. E. Lenski, “Escherichia coli rpoB mutants have increased evolvability in proportion to their fitness defects.,” *Mol. Biol. Evol.*, vol. 27, no. 6, pp. 1338–47, Jun. 2010.

[18] V. Mozhayskiy and I. Tagkopoulos, “Horizontal gene transfer dynamics and distribution of fitness effects during microbial in silico evolution,” *BMC Bioinformatics*, vol. 13, no. Suppl 10, p. S13, Jan. 2012.

[19] X. Jiang, B. Mu, Z. Huang, M. Zhang, X. Wang, and S. Tao, “Impacts of mutation effects and population size on mutation rate in asexual populations: a simulation study,” *BMC Evol. Biol.*, vol. 10, no. 1, p. 298, 2010.

[20] C. S. Pepperell *et al.*, “The role of selection in shaping diversity of natural M. tuberculosis populations.,” *PLoS Pathog.*, vol. 9, no. 8, p. e1003543, Aug. 2013.

[21] J. J. Wernegreen, “Reduced selective constraint in endosymbionts: elevation in radical amino acid replacements occurs genome-wide.,” *PLoS One*, vol. 6, no. 12, p. e28905, Jan. 2011.

[22] P. Remigi *et al.*, “Transient hypermutagenesis accelerates the evolution of legume endosymbionts following horizontal gene transfer.,” *PLoS Biol.*, vol. 12, no. 9, p. e1001942, Sep. 2014.

[23] B. Batut, C. Knibbe, G. Marais, and V. Daubin, “Reductive genome evolution at both ends of the bacterial population size spectrum.,” *Nat. Rev. Microbiol.*, vol. 12, no. 12, pp. 841–50, Dec. 2014.

[24] J. O. Andersson and S. G. Andersson, “Pseudogenes, junk DNA, and the dynamics of Rickettsia genomes.,” *Mol. Biol. Evol.*, vol. 18, no. 5, pp. 829–839, 2001.

[25] C.-H. Kuo and H. Ochman, “The extinction dynamic of bacterial pseudogenes,” *PLoS Genet.*, vol. 6, no. 8, p. e1001050, Aug. 2010.

[26] H. Ochman and L. M. Davalos, “The nature and dynamics of bacterial genomes.,” *Science*, vol. 311, no. 5768, pp. 1730–3, Mar. 2006.

[27] S. J. Biller, P. M. Berube, D. Lindell, and S. W. Chisholm, “Prochlorococcus: the

structure and function of collective diversity," *Nat. Rev. Microbiol.*, vol. 13, no. 1, pp. 13–27, Dec. 2014.

[28] Z. Sun and J. L. Blanchard, "Strong genome-wide selection early in the evolution of Prochlorococcus resulted in a reduced genome through the loss of a large number of small effect genes.," *PLoS One*, vol. 9, no. 3, p. e88837, Jan. 2014.

[29] F. Partensky and L. Garczarek, "Prochlorococcus: advantages and limits of minimalism.," *Ann. Rev. Mar. Sci.*, vol. 2, pp. 305–31, Jan. 2010.

[30] N. Kashtan *et al.*, "Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus," *Science*, vol. 344, no. 6182, pp. 416–420, Apr. 2014.

[31] A. S. Lang, O. Zhaxybayeva, and J. T. Beatty, "Gene transfer agents: phage-like elements of genetic exchange," *Nat. Rev. Microbiol.*, vol. 10, no. 7, pp. 472–482, 2012.

[32] L.-M. Bobay, M. Touchon, and E. P. C. Rocha, "Pervasive domestication of defective prophages by bacteria," *Proc. Natl. Acad. Sci.*, vol. 111, no. 33, pp. 12127–12132, 2014.

[33] V. Daubin and H. Ochman, "Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli.," *Genome Res.*, vol. 14, no. 6, pp. 1036–42, Jun. 2004.

[34] G. Yu and A. Stoltzfus, "Population diversity of ORFan genes in Escherichia coli.," *Genome Biol. Evol.*, vol. 4, no. 11, pp. 1176–87, Jan. 2012.

[35] L. E. Orgel and F. H. Crick, "Selfish DNA: the ultimate parasite.," *Nature*, vol. 284, no. 5757, pp. 604–607, 1980.

[36] J. Lawrence, "Selfish operons: the evolutionary impact of gene clustering in prokaryotes

and eukaryotes," *Curr. Opin. Genet. Dev.*, vol. 9, no. 6, pp. 642–648, Dec. 1999.

[37] R. Dawkins, *The Selfish Gene*. Oxford, England: Oxford University Press, Oxford, England, 1976.

[38] X. Wang *et al.*, "Cryptic prophages help bacteria cope with adverse environments.," *Nat. Commun.*, vol. 1, no. 9, p. 147, Jan. 2010.

[39] L. D. McDaniel, E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, and J. H. Paul, "High frequency of horizontal gene transfer in the oceans.," *Science*, vol. 330, p. 50, 2010.

[40] P. S. Novichkov, Y. I. Wolf, I. Dubchak, and E. V Koonin, "Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes," *J. Bacteriol.*, vol. 191, no. 1, pp. 65–73, Jan. 2009.

[41] E. P. C. Rocha *et al.*, "Comparisons of dN/dS are time dependent for closely related bacterial genomes," *J. Theor. Biol.*, vol. 239, no. 2, pp. 226–235, 2006.

[42] Y. Huang, E. V Koonin, D. J. Lipman, and T. M. Przytycka, "Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage.," *Nucleic Acids Res.*, vol. 37, no. 20, pp. 6799–810, Nov. 2009.

[43] T. Warnecke, Y. Huang, T. M. Przytycka, and L. D. Hurst, "Unique cost dynamics elucidate the role of frameshifting errors in promoting translational robustness," *Genome Biol. Evol.*, vol. 2, pp. 636–645, Jan. 2010.

[44] C. Zeyl and J. A. DeVisser, "Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae.," *Genetics*, vol. 157, no. 1, pp. 53–61, 2001.

[45] R. A. Goldstein, “Population size dependence of fitness effect distribution and substitution rate probed by biophysical model of protein thermostability.,” *Genome Biol. Evol.*, vol. 5, no. 9, pp. 1584–93, Jan. 2013.

[46] G. Faure and E. V. Koonin, “Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins.,” *Phys. Biol.*, vol. 12, no. 3, p. 35001, May 2015.

[47] N. Tokuriki, F. Stricher, J. Schymkowitz, L. Serrano, and D. S. Tawfik, “The stability effects of protein mutations appear to be universally distributed.,” *J. Mol. Biol.*, vol. 369, no. 5, pp. 1318–32, Jun. 2007.

[48] T. A. Kunkel, “DNA Replication Fidelity,” *J. Biol. Chem.* , vol. 279, no. 17, pp. 16895–16898, Apr. 2004.

[49] C. C. Traverse and H. Ochman, “Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 12, pp. 3311–3316, Mar. 2016.

[50] K. L. Herbst, L. M. Nichols, R. F. Gesteland, and R. B. Weiss, “A mutation in ribosomal protein L9 affects ribosomal hopping during translation of gene 60 from bacteriophage T4,” *Proc. Natl. Acad. Sci.* , vol. 91, no. 26, pp. 12525–12529, Dec. 1994.

[51] H. S. Zaher and R. Green, “Fidelity at the molecular level: lessons from protein synthesis,” *Cell*, vol. 136, no. 4, pp. 746–762, Feb. 2009.

[52] D. A. Drummond and C. O. Wilke, “The evolutionary consequences of erroneous protein

synthesis.," *Nat. Rev. Genet.*, vol. 10, no. 10, pp. 715–24, Oct. 2009.

[53] E. B. Kramer and P. J. Farabaugh, "The frequency of translational misreading errors in E. coli is largely determined by tRNA competition," *RNA*, vol. 13, no. 1, pp. 87–96, 2007.

[54] H. L. True and S. L. Lindquist, "A yeast prion provides a mechanism for genetic variation and phenotypic diversity.," *Nature*, vol. 407, no. 6803, pp. 477–483, 2000.

[55] R. Halfmann, S. Alberti, and S. Lindquist, "Prions, protein homeostasis, and phenotypic diversity.," *Trends Cell Biol.*, vol. 20, no. 3, pp. 125–33, Mar. 2010.

[56] J. Masel, "Cryptic genetic variation is enriched for potential adaptations," *Genetics*, vol. 172, no. 3, pp. 1985–1991, Mar. 2006.

[57] N. Torabi and L. Kruglyak, "Genetic basis of hidden phenotypic variation revealed by increased translational readthrough in yeast.," *PLoS Genet.*, vol. 8, no. 3, p. e1002546, Jan. 2012.

[58] M. Meyerovich, G. Mamou, and S. Ben-Yehuda, "Visualizing high error levels during gene expression in living bacterial cells.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 25, pp. 11543–8, Jun. 2010.

[59] M. A. Fares, "The origins of mutational robustness," *Trends Genet.*, vol. 31, no. 7, pp. 373–381, Jul. 2015.

[60] S. Bratulic, F. Gerber, and A. Wagner, "Mistranslation drives the evolution of robustness in TEM-1 β-lactamase," *Proc. Natl. Acad. Sci.*, vol. 112, no. 41, p. 201510071, Oct. 2015.

[61] S. J. Freeland and L. D. Hurst, "The genetic code is one in a million," *J. Mol. Evol.*, vol.

47, no. 3, pp. 238–248, 1998.

[62] F. Taddei, M. Radman, J. Maynard-Smith, B. Toupance, P. H. Gouyon, and B. Godelle, "Role of mutator alleles in adaptive evolution," *Nature*, vol. 387, no. 6634, pp. 700–702, Jun. 1997.

[63] P. D. Sniegowski, P. J. Gerrish, and R. E. Lenski, "Evolution of high mutation rates in experimental populations of E. coli," *Nature*, vol. 387, no. 6634, pp. 703–705, Jun. 1997.

[64] P. D. Sniegowski, P. J. Gerrish, T. Johnson, and A. Shaver, "The evolution of mutation rates: separating causes from consequences," *BioEssays*, vol. 22, no. 12, pp. 1057–1066, 2000.

[65] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami, "Evolution of digital organisms at high mutation rates leads to survival of the flattest," *Nature*, vol. 412, no. 6844, pp. 331–333, Jul. 2001.

[66] E. Rajon and J. Masel, "Evolution of molecular error rates and the consequences for evolvability.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 3, pp. 1082–1087, 2011.

[67] N. Goldman and Z. Yang, "A codon-based model of nucleotide substitution for protein-coding DNA sequences," *Mol. Biol. Evol.*, vol. 11, no. 5, pp. 725–736, 1994.

[68] M. Nei and T. Gojobori, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.," *Mol. Biol. Evol.* , vol. 3, no. 5, pp. 418–426, Sep. 1986.

[69] Z. Yang and R. Nielsen, "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models," *Mol. Biol. Evol.*, vol. 17, no. 1, pp. 32–43, Jan. 2000.

[70] S. V Muse and B. S. Gaut, “A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.,” *Mol. Biol. Evol.*, vol. 11, no. 5, pp. 715–724, 1994.

[71] M. F. Polz, E. J. Alm, and W. P. Hanage, “Horizontal gene transfer and the evolution of bacterial and archaeal population structure.,” *Trends Genet.*, vol. 29, pp. 170–5, 2013.

[72] G. Schönknecht *et al.*, “Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote.,” *Science*, vol. 339, pp. 1207–10, 2013.

[73] S. K. Sheppard, N. D. McCarthy, D. Falush, and M. C. J. Maiden, “Convergence of Campylobacter species: implications for bacterial evolution.,” *Science*, vol. 320, pp. 237–239, 2008.

[74] E. F. Mongodin *et al.*, “The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, pp. 18147–18152, 2005.

[75] T. Dagan, Y. Artzy-Randrup, and W. Martin, “Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, pp. 10039–10044, 2008.

[76] B. Marrs, “Genetic recombination in Rhodopseudomonas capsulata.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 71, pp. 971–973, 1974.

[77] B. J. Rapp and J. D. Wall, “Genetic transfer in Desulfovibrio desulfuricans.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 84, no. 24, pp. 9128–30, Dec. 1987.

[78] S. B. Humphrey, T. B. Stanton, N. S. Jensen, and R. L. Zuerner, “Purification and

characterization of VSH-1, a generalized transducing bacteriophage of Serpulina hyodysenteriae," *J. Bacteriol.*, vol. 179, no. 2, pp. 323–329, 1997.

[79] G. Bertani, "Transduction-like gene transfer in the methanogen Methanococcus voltae," *J Bacteriol*, vol. 181, no. 10, pp. 2992–3002, 1999.

[80] A. S. Lang and J. T. Beatty, "Importance of widespread gene transfer agent genes in alpha-proteobacteria," *Trends Microbiol.*, vol. 15, no. 2, pp. 54–62, 2007.

[81] L. Guy *et al.*, "A gene transfer agent and a dynamic repertoire of secretion systems hold the keys to the explosive radiation of the emerging pathogen Bartonella.," *PLoS Genet.*, vol. 9, no. 3, p. e1003393, Mar. 2013.

[82] W. D. D. Hamilton, "The genetical evolution of social behaviour. I," *J. Theor. Biol.*, vol. 7, no. 1, pp. 1–16, 1964.

[83] W. D. Hamilton, "The genetical evolution of social behaviour. II.," *J. Theor. Biol.*, vol. 7, no. 1, pp. 17–52, 1964.

[84] J. A. Draghi and P. E. Turner, "DNA secretion and gene-level selection in bacteria," *Microbiology*, vol. 152, no. 9, pp. 2683–2688, Sep. 2006.

[85] J. E. Strassmann, O. M. Gilbert, and D. C. Queller, "Kin discrimination and cooperation in microbes.," *Annu. Rev. Microbiol.*, vol. 65, pp. 349–67, Jan. 2011.

[86] W. F. Doolittle and C. Sapienza, "Selfish genes, the phenotype paradigm and genome evolution," *Nature*, vol. 284, no. 5757, pp. 601–603, Apr. 1980.

[87] Y. Liu, P. M. Harrison, V. Kunin, and M. Gerstein, "Comprehensive analysis of

pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes," *Genome Biol.*, vol. 5, no. 9, p. R64, 2004.

[88] M. N. Price *et al.*, "Indirect and suboptimal control of gene expression is widespread in bacteria.," *Mol. Syst. Biol.*, vol. 9, no. 660, p. 660, Jan. 2013.

[89] C. Canchaya, C. Proux, G. Fournous, A. Bruttin, and H. Brüssow, "Prophage genomics.," *Microbiol. Mol. Biol. Rev.*, vol. 67, no. 2, pp. 238–276, 2003.

[90] A. D. Johnson, A. R. Poteete, G. Lauer, R. T. Sauer, G. K. Ackers, and M. Ptashne, "Lambda repressor and cro--components of an efficient molecular switch.," *Nature*, vol. 294, no. 5838, pp. 217–223, 1981.

[91] R. Menouni, S. Champ, L. Espinosa, M. Boudvillain, and M. Ansaldi, "Transcription termination controls prophage maintenance in Escherichia coli genomes," *Proc. Natl. Acad. Sci.*, vol. 110, no. 35, pp. 14414–14419, 2013.

[92] C. Sasakawa, J. B. Lowe, L. McDivitt, and D. E. Berg, "Control of transposon Tn5 transposition in Escherichia coli," *Proc Natl Acad Sci U S A*, vol. 79, no. 23, pp. 7450–7454, 1982.

[93] J. B. Biggins, M. a Ternei, and S. F. Brady, "Malleilactone, a polyketide synthase-derived virulence factor encoded by the cryptic secondary metabolome of Burkholderia pseudomallei group pathogens.," *J. Am. Chem. Soc.*, vol. 134, no. 32, pp. 13192–5, Aug. 2012.

[94] W. F. Doolittle, T. D. P. Brunet, S. Linquist, and T. R. Gregory, "Distinguishing between 'function' and 'effect' in genome biology," *Genome Biol. Evol.*, vol. 6, pp. 1234–1237,

May 2014.

[95] S. Omer, T. J. Harlow, and J. P. Gogarten, “Does Sequence Conservation Provide Evidence for Biological Function?,” *Trends Microbiol.*, vol. (in press), 2016.

[96] V. M. Markowitz *et al.*, “IMG: the Integrated Microbial Genomes database and comparative analysis system.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D115-22, Jan. 2012.

[97] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004.

[98] M. Gouy, S. Guindon, and O. Gascuel, “SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building,” *Mol. Biol. Evol.*, vol. 27, no. 2, pp. 221–224, 2010.

[99] S. Guindon and O. Gascuel, “A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood,” *Syst. Biol.*, vol. 52, no. 5, pp. 696–704, Oct. 2003.

[100] R. V. Eck and M. O. Dayhoff, *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Springs, Maryland, 1966.

[101] A. G. Kluge and J. S. Farris, “Quantitative phyletics and the evolution of anurans,” *Syst. Zool.*, vol. 18, no. 1, pp. 1–32, 1969.

[102] J. Felsenstein, “PHYLIP - Phylogeny inference package - v3.2,” *Cladistics*. pp. 164–166, 1989.

[103] A. Stamatakis, “RAxML version 8: a tool for phylogenetic analysis and post-analysis of

large phylogenies," *Bioinforma.* , vol. 30, no. 9, pp. 1312–1313, May 2014.

[104] H. Shimodaira, "An approximately unbiased test of phylogenetic tree selection.," *Syst. Biol.*, vol. 51, no. 3, pp. 492–508, Jun. 2002.

[105] H. Shimodaira and M. Hasegawa, "Multiple comparisons of log-likelihoods with applications to phylogenetic inference," *Mol. Biol. Evol*, vol. 16, no. 8, pp. 1114–1116, 1999.

[106] H. Shimodaira and M. Hasegawa, "CONSEL: for assessing the confidence of phylogenetic tree selection.," *Bioinformatics*, vol. 17, pp. 1246–1247, 2001.

[107] S. Kryazhimskiy and J. B. Plotkin, "The population genetics of dN/dS," *PLoS Genet*, vol. 4, no. 12, p. e1000304, 2008.

[108] I. Kobayashi, "Restriction-modification systems as minimal forms of life," in *Nucleic Acids and Molecular Biology*, vol. 14, 2004, pp. 4–6.

[109] M. A. Riley and J. E. Wertz, "Bacteriocins: evolution, ecology, and application.," *Annu. Rev. Microbiol.*, vol. 56, pp. 117–37, Jan. 2002.

[110] R. F. Inglis, B. Bayramoglu, O. Gillor, and M. Ackermann, "The role of bacteriocins as selfish genetic elements.," *Biol. Lett.*, vol. 9, no. 3, p. 20121173, Apr. 2013.

[111] R. L. Dy, R. Przybilski, K. Semeijn, G. P. C. Salmond, and P. C. Fineran, "A widespread bacteriophage abortive infection system functions through a Type IV toxin-antitoxin mechanism.," *Nucleic Acids Res.*, vol. 42, no. 7, pp. 4590–605, Apr. 2014.

[112] S. Castillo-Ramírez *et al.*, "The impact of recombination on dN/dS within recently

emerged bacterial clones.," *PLoS Pathog.*, vol. 7, no. 7, p. e1002129, Jul. 2011.

[113] S. Krogh, M. O'Reilly, N. Nolan, and K. M. Devine, "The phage-like element PBSX and part of the skin element, which are resident at different locations on the Bacillus subtilis chromosome, are highly homologous," *Microbiology*, vol. 142, no. 8, pp. 2031–2040, 1996.

[114] S. Krogh, S. Jørgensen, and K. M. Devine, "Lysis genes of the Bacillus subtilis defective prophage PBSX," *J. Bacteriol.*, vol. 180, no. 8, pp. 2110–2117, 1998.

[115] G. E. McDonnell, H. Wood, K. M. Devine, and D. J. McConnell, "Genetic control of bacterial suicide: regulation of the induction of PBSX in Bacillus subtilis.," *J. Bacteriol.*, vol. 176, no. 18, pp. 5820–30, Sep. 1994.

[116] H. E. Wood, M. T. Dawson, K. M. Devine, and D. J. McConnell, "Characterization of PBSX, a defective prophage of Bacillus subtilis," *J. Bacteriol.*, vol. 172, no. 5, pp. 2667–2674, 1990.

[117] R. Shingaki, Y. Kasahara, T. Inoue, S. Kokeguchi, and K. Fukui, "Chromosome DNA fragmentation and excretion caused by defective prophage gene expression in the early-exponential-phase culture of Bacillus subtilis.," *Can. J. Microbiol.*, vol. 49, no. 5, pp. 313–25, May 2003.

[118] T. Jin *et al.*, "Biological and genomic analysis of a PBSX-like defective phage induced from Bacillus pumilus AB94180.," *Arch. Virol.*, vol. 159, no. 4, pp. 739–52, Apr. 2014.

[119] V. Kasari, T. Mets, T. Tenson, and N. Kaldalu, "Transcriptional cross-activation between toxin-antitoxin systems of Escherichia coli.," *BMC Microbiol.*, vol. 13, no. 1, p. 45, Jan.

2013.

[120] X. Wang and T. K. Wood, “Toxin-antitoxin systems influence biofilm and persister cell formation and the general stress response.,” *Appl. Environ. Microbiol.*, vol. 77, no. 16, pp. 5577–83, Aug. 2011.

[121] J. H. Werren, “Selfish genetic elements, genetic conflict, and evolutionary innovation,” *Proc. Natl. Acad. Sci. ,* vol. 108, no. Supplement 2, pp. 10863–10870, Jun. 2011.

[122] C. Milheiriço, A. Portelinha, L. Krippahl, H. de Lencastre, and D. C. Oliveira, “Evidence for a purifying selection acting on the β-lactamase locus in epidemic clones of methicillin-resistant Staphylococcus aureus.,” *BMC Microbiol.*, vol. 11, p. 76, 2011.

[123] N. E. Tunstall, T. Sirey, R. D. Newcomb, and C. G. Warr, “Selective pressures on Drosophila chemosensory receptor genes,” *J. Mol. Evol.*, vol. 64, no. 6, pp. 628–636, 2007.

[124] G. Jerzak, K. a. Bernard, L. D. Kramer, and G. D. Ebel, “Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection,” *J. Gen. Virol.*, vol. 86, no. 8, pp. 2175–2183, 2005.

[125] J. Bohlin, O. B. Brynildsrud, C. Sekse, and L. Snipen, “An evolutionary analysis of genome expansion and pathogenicity in Escherichia coli,” *BMC Genomics*, vol. 15, no. 1, p. 882, 2014.

[126] D. A. Drummond, J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold, “Why highly expressed proteins evolve slowly.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 40, pp. 14338–43, Oct. 2005.

[127] M. W. Gray, J. Lukeš, and J. M. Archibald, “Irremediable complexity?,” *Science (80-.).*, vol. 330, no. 6006, pp. 920–921, 2010.

[128] S. Bershtein *et al.*, “Protein homeostasis imposes a barrier on functional integration of horizontally transferred genes in bacteria,” *PLoS Genet.*, vol. 11, no. 10, p. e1005612, 2015.

[129] F. Jacob and J. Monod, “Genetic Regulatory Mechanisms in the Syntheis of Proteins,” *J. Mol. Biol.*, vol. 3, pp. 318–356, 1961.

[130] H.-X. Zhou, G. Rivas, and A. P. Minton, “Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences,” *Annu. Rev. Biophys.*, vol. 37, no. 1, pp. 375–397, May 2008.

[131] S. R. McGuffee and A. H. Elcock, “Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm.,” *PLoS Comput. Biol.*, vol. 6, no. 3, p. e1000694, Mar. 2010.

[132] F. U. Hartl, A. Bracher, and M. Hayer-Hartl, “Molecular chaperones in protein folding and proteostasis.,” *Nature*, vol. 475, no. 7356, pp. 324–32, Jul. 2011.

[133] C. Pál, B. Papp, and L. D. Hurst, “Highly expressed genes in yeast evolve slowly,” *Genetics*, vol. 158, no. 2, pp. 927–931, Jun. 2001.

[134] K. A. Geiler-Samerotte, M. F. Dion, B. A. Budnik, S. M. Wang, D. L. Hartl, and D. A. Drummond, “Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 2, pp. 680–685, Jan. 2010.

[135] J.-R. Yang, B.-Y. Liao, S.-M. Zhuang, and J. Zhang, “Protein misinteraction avoidance causes highly expressed proteins to evolve slowly.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 14, pp. E831-40, Apr. 2012.

[136] G. Plata, M. Gottesman, and D. Vitkup, “The rate of the molecular clock and the cost of gratuitous protein synthesis,” *Genome Biol.*, vol. 11, no. 9, p. R98, 2010.

[137] M. Eames and T. Kortemme, “Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance.,” *Structure*, vol. 15, no. 11, pp. 1442–51, Nov. 2007.

[138] M. Taoka *et al.*, “Only a small subset of the horizontally transferred chromosomal genes in Escherichia coli are translated into proteins.,” *Mol. Cell. Proteomics*, vol. 3, no. 8, pp. 780–787, 2004.

[139] M. J. Lercher and C. Pál, “Integration of horizontally transferred genes into regulatory interaction networks takes many million years,” *Mol. Biol. Evol.*, vol. 25, no. 3, pp. 559–567, 2008.

[140] Y. Zhang, H. Romero, G. Salinas, and V. N. Gladyshev, “Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues.,” *Genome Biol.*, vol. 7, no. 10, p. R94, Jan. 2006.

[141] D. G. Longstaff *et al.*, “A natural genetic code expansion cassette enables transmissible biosynthesis and genetic encoding of pyrrolysine.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 3, pp. 1021–6, Jan. 2007.

[142] H. Seligmann and D. Pollock, “The ambush hypothesis : hidden stop codons prevent off-frame gene reading,” *DNA Cell Biol.*, vol. 23, no. 10, pp. 701–705, 2004.

[143] T. R. Singh and K. R. Pardasani, “Ambush hypothesis revisited: Evidences for phylogenetic trends.,” *Comput. Biol. Chem.*, vol. 33, no. 3, pp. 239–44, Jun. 2009.

[144] I. Jungreis *et al.*, “Evidence of abundant stop codon readthrough in Drosophila and other metazoa,” *Genome Res.*, vol. 21, no. 12, pp. 2096–2113, 2011.

[145] B. D. Janssen and C. S. Hayes, “Kinetics of paused ribosome recycling in Escherichia coli,” *J. Mol. Biol.*, vol. 394, no. 2, pp. 251–267, 2009.

[146] F. Garza-Sánchez, R. E. Schaub, B. D. Janssen, and C. S. Hayes, “tmRNA regulates synthesis of the ArfA ribosome rescue factor.,” *Mol. Microbiol.*, vol. 80, no. 5, pp. 1204–19, Jun. 2011.

[147] J. S. Seidman, B. D. Janssen, and C. S. Hayes, “Alternative fates of paused ribosomes during translation termination.,” *J. Biol. Chem.*, vol. 286, no. 36, pp. 31105–12, Sep. 2011.

[148] R. E. Schaub, S. J. Poole, F. Garza-Sánchez, S. Benbow, and C. S. Hayes, “Proteobacterial ArfA peptides are synthesized from non-stop messenger RNAs.,” *J. Biol. Chem.*, vol. 287, no. 35, pp. 29765–75, Aug. 2012.

[149] M. G. Gagnon, S. V Seetharaman, D. Bulkley, and T. a Steitz, “Structural basis for the rescue of stalled ribosomes: structure of YaeJ bound to the ribosome.,” *Science*, vol. 335, no. 6074, pp. 1370–2, Mar. 2012.

[150] N. R. Guydosh and R. Green, “Dom34 rescues ribosomes in 3’ untranslated regions.,”

Cell, vol. 156, no. 5, pp. 950–62, Feb. 2014.

[151] A. Wagner, "The molecular origins of evolutionary innovations.," *Trends Genet.*, vol. 27, no. 10, pp. 397–410, Oct. 2011.

[152] J. A. Draghi, T. L. Parsons, G. P. Wagner, and J. B. Plotkin, "Mutational robustness can facilitate adaptation.," *Nature*, vol. 463, no. 7279, pp. 353–5, Jan. 2010.

[153] E. Rajon and J. Masel, "Compensatory evolution and the origins of innovations.," *Genetics*, vol. 193, no. 4, pp. 1209–20, Apr. 2013.

[154] S. Chen, B. H. Krinsky, and M. Long, "New genes as drivers of phenotypic evolution.," *Nat. Rev. Genet.*, vol. 14, no. 9, pp. 645–60, Aug. 2013.

[155] M. G. Giacomelli, A. S. Hancock, and J. Masel, "The conversion of 3' UTRs into coding regions.," *Mol. Biol. Evol.*, vol. 24, no. 2, pp. 457–64, Feb. 2007.

[156] O. Namy, G. Duchateau-Nguyen, and J.-P. Rousset, "Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae.," *Mol. Microbiol.*, vol. 43, no. 3, pp. 641–52, Feb. 2002.

[157] O. Namy, G. Duchateau-nguyen, I. Hatin, S. H. Denmat, M. Termier, and J. Rousset, "Identification of stop codon readthrough genes in Saccharomyces cerevisiae," *Nucleic Acids Res.*, vol. 31, no. 9, pp. 2289–2296, May 2003.

[158] N. T. Ingolia *et al.*, "Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes," *Cell Rep.*, vol. 8, no. 5, pp. 1365–1379, 2014.

[159] A. A. Vakhrusheva, M. D. Kazanov, A. A. Mironov, and G. A. Bazykin, "Evolution of

prokaryotic genes by shift of stop codons.," *J. Mol. Evol.*, vol. 72, no. 2, pp. 138–46, Feb. 2011.

[160] U. Bergthorsson, D. I. Andersson, and J. R. Roth, "Ohno's dilemma: evolution of new genes under continuous selection.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 43, pp. 17004–9, Oct. 2007.

[161] A. Mira and N. A. Moran, "Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria," *Microb. Ecol.*, vol. 44, no. 2, pp. 137–143, 2002.

[162] J. Charlesworth and A. Eyre-Walker, "The rate of adaptive evolution in enteric bacteria," *Mol. Biol. Evol.*, vol. 23, no. 7, pp. 1348–1356, Jul. 2006.

[163] F. Baumdicker, W. R. Hess, and P. Pfaffelhuber, "The infinitely many genes model for the distributed genome of bacteria," *Genome Biol. Evol.*, vol. 4, no. 4, pp. 443–456, Feb. 2012.

[164] N. A. O'Leary *et al.*, "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation," *Nucleic Acids Res.*, vol. 44, no. D1, pp. 733–745, Jan. 2016.

[165] P. Puigbò, A. Romeu, and S. Garcia-Vallvé, "HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. 524–527, Jan. 2008.

[166] P. Puigbò, I. G. Bravo, and S. Garcia-Vallvé, "E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI)," *BMC Bioinformatics*, vol. 9, no. 1,

pp. 1–7, 2008.

[167] R. C. Edgar, “Search and clustering orders of magnitude faster than BLAST,” *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Oct. 2010.

[168] V. Ranwez, S. Harispe, F. Delsuc, and E. J. P. Douzery, “MACSE: Multiple alignment of coding sequences accounting for frameshifts and stop codons,” *PLoS One*, vol. 6, no. 9, p. e22594, 2011.

[169] P. Kück and G. C. Longo, “FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies,” *Front. Zool.*, vol. 11, no. 1, p. 81, 2014.

[170] D. L. Swofford, “PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods).,” *Sinauer Assoc. Sunderland, Massachusetts.*, pp. 1–142, 2002.

[171] Y. Ponty, M. Termier, and A. Denise, “GenRGenS: Software for generating random genomic sequences and structures,” *Bioinformatics*, vol. 22, no. 12, pp. 1534–1535, 2006.

[172] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.

[173] J. G. Dunn, C. K. Foo, N. G. Belletier, E. R. Gavis, and J. S. Weissman, “Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster.,” *Elife*, vol. 2, p. e01179, Jan. 2013.

[174] S. A. Shabalina, A. Y. Ogurtsov, N. A. Spiridonov, and E. V Koonin, “Evolution at protein ends: major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals.,” *Nucleic Acids Res.*, vol. 42, no.

11, pp. 7132–44, Jan. 2014.

[175] R. Nussinov, “The universal dinucleotide asymmetry rules in DNA and the amino acid codon choice,” *J. Mol. Evol.*, vol. 17, no. 4, pp. 237–244, 1981.

[176] R. Nussinov, “Nearest neighbor nucleotide patterns. Structural and biological implications.,” *J. Biol. Chem.* , vol. 256, no. 16, pp. 8458–8462, Aug. 1981.

[177] R. Nussinov, “Strong doublet preferences in nucleotide sequences and DNA geometry,” *J. Mol. Evol.*, vol. 20, no. 2, pp. 111–119, 1984.

[178] Y. Nakamura, T. Gojobori, and T. Ikemura, “Codon usage tabulated from international DNA sequence databases: status for the year 2000.,” *Nucleic Acids Res.*, vol. 28, no. 1, p. 292, Jan. 2000.

[179] P. M. Sharp and W. H. Li, “The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.,” *Nucleic Acids Res.*, vol. 15, no. 3, pp. 1281–1295, 1987.

[180] J. M. Peters, A. D. Vangeloff, and R. Landick, “Bacterial Transcription Terminators: The RNA 3′-End Chronicles,” *J. Mol. Biol.*, vol. 412, no. 5, pp. 793–813, Oct. 2011.

[181] S. Volinia, R. Gambari, F. Bernardi, and I. Barrai, “The frequency of oligonucleotides in mammalian genic regions,” *Comput. Appl. Biosci. CABIOS* , vol. 5, no. 1, pp. 33–40, Feb. 1989.

[182] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[183] G. Lauc, I. Ilić, and M. Heffer-Lauc, “Entropies of coding and noncoding sequences of DNA and proteins,” *Biophys. Chem.*, vol. 42, no. 1, pp. 7–11, 1992.

[184] M. Kimura, *The Neutral Theory of Molecular Evolution*. Cambridge University Press, New York, 1983.

[185] H. C. Barker, N. Kinsella, A. Jaspe, T. Friedrich, and C. D. O’Connor, “Formate protects stationary-phase Escherichia coli and Salmonella cells from killing by a cationic antimicrobial peptide,” *Mol. Microbiol.*, vol. 35, no. 6, pp. 1518–1529, Jan. 2002.

[186] A. J. Grant, M. Farris, P. Alefounder, P. H. Williams, M. J. Woodward, and C. D. O’Connor, “Co-ordination of pathogenicity island expression by the BipA GTPase in enteropathogenic Escherichia coli (EPEC),” *Mol. Microbiol.*, vol. 48, no. 2, pp. 507–521, Apr. 2003.

[187] M. Farris, A. Grant, T. B. Richardson, and C. D. O’Connor, “BipA: a tyrosine-phosphorylated GTPase that mediates interactions between enteropathogenic Escherichia coli (EPEC) and epithelial cells,” *Mol. Microbiol.*, vol. 28, no. 2, pp. 265–279, Apr. 1998.

[188] S. Rowe, N. Hodson, G. Griffiths, and I. S. Roberts, “Regulation of the Escherichia coli K5 capsule gene cluster: Evidence for the roles of H-NS, BipA, and integration host factor in regulation of group 2 capsule gene clusters in pathogenic E. coli,” *J. Bacteriol.*, vol. 182, no. 10, pp. 2741–2745, 2000.

[189] E. Kiss, T. Huguet, V. Poinsot, and J. Batut, “The typA gene is required for stress adaptation as well as for symbiosis of Sinorhizobium meliloti 1021 with certain Medicago truncatula lines,” *Mol. Plant-Microbe Interact.*, vol. 17, no. 3, pp. 235–244, Mar. 2004.

[190] P. L. Pfennig and A. M. Flower, “BipA is required for growth of Escherichia coli K12 at low temperature,” *Mol. Genet. Genomics*, vol. 266, no. 2, pp. 313–317, Oct. 2001.

[191] S. Karlin, J. Mrázek, A. Campbell, D. Kaiser, and J. A. N. Mra, “Characterizations of highly expressed genes of four fast-growing bacteria.,” *J. Bacteriol.*, vol. 183, no. 17, pp. 5025–5040, 2001.

[192] L. Boulos, M. Prévost, B. Barbeau, J. Coallier, and R. Desjardins, “LIVE/DEAD® BacLight™: application of a new rapid staining method for direct enumeration of viable and total bacteria in drinking water,” *J. Microbiol. Methods*, vol. 37, no. 1, pp. 77–86, Jul. 1999.

[193] M. Berney, F. Hammes, F. Bosshard, H.-U. Weilenmann, and T. Egli, “Assessment and interpretation of bacterial viability by using the LIVE/DEAD BacLight Kit in combination with flow cytometry,” *Appl. Environ. Microbiol.*, vol. 73, no. 10, pp. 3283–3290, May 2007.

[194] P. Stiefel, S. Schmidt-Emrich, K. Maniura-Weber, and Q. Ren, “Critical aspects of using bacterial cell viability assays with the fluorophores SYTO9 and propidium iodide,” *BMC Microbiol.*, vol. 15, p. 36, Feb. 2015.

[195] G. Crivat and J. W. Taraska, “Imaging proteins inside cells with fluorescent tags,” *Trends Biotechnol.*, vol. 30, no. 1, pp. 8–16, Jan. 2012.