PLOS ONE

RESEARCH ARTICLE

Codon Pairs are Phylogenetically Conserved: A comprehensive analysis of codon pairing conservation across the Tree of Life

Justin B. Miller[1☯], Lauren M. McKinnon[1☯], Michael F. Whiting[1,2], John S. K. Kauwe[1], Perry G. Ridge[1]*

1 Department of Biology, Brigham Young University, Provo, UT, United States of America, **2** M.L. Bean Museum, Brigham Young University, Provo, UT, United States of America

☯ These authors contributed equally to this work.
* perry.ridge@byu.edu

Citation: Miller JB, McKinnon LM, Whiting MF, Kauwe JSK, Ridge PG (2020) Codon Pairs are Phylogenetically Conserved: A comprehensive analysis of codon pairing conservation across the Tree of Life. PLoS ONE 15(5): e0232260. https://doi.org/10.1371/journal.pone.0232260

Editor: Marc Robinson-Rechavi, Universite de Lausanne Faculte de biologie et medecine, SWITZERLAND

Received: August 28, 2019

Accepted: April 10, 2020

Published: May 13, 2020

Data Availability Statement: All algorithms used to recover and compare phylogenies, including documentation and test files, are freely available on GitHub at https://github.com/ridgelab/codon_pairing.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Identical codon pairing and co-tRNA codon pairing increase translational efficiency within genes when two codons that encode the same amino acid are translated by the same tRNA before it diffuses from the ribosome. We examine the phylogenetic signal in both identical and co-tRNA codon pairing across 23 428 species using alignment-free and parsimony methods. We determined that conserved codon pairing typically has a smaller window size than the length of a ribosome, and codon pairing tracks phylogenies across various taxonomic groups. We report a comprehensive analysis of codon pairing, including the extent to which each codon pairs. Our parsimony method generally recovers phylogenies that are more congruent with the established phylogenies than our alignment-free method. However, four of the ten taxonomic groups did not have sufficient orthologous codon pairings and were therefore analyzed using only the alignment-free methods. Since the recovered phylogenies using only codon pairing largely match phylogenies from the Open Tree of Life and the NCBI taxonomy, and are comparable to trees recovered by other algorithms, we propose that codon pairing biases are phylogenetically conserved and should be considered in conjunction with other phylogenomic techniques.

Introduction

Phylogenies allow biologists to infer similar characteristics of closely related species and provide an evolutionary framework for analyzing biological patterns [1]. Phylogenies are statements of homology, and represent a continuity of biological information [2]. Although genetic data facilitate the analysis of diverse species, molecular data typically require data cleaning (e.g., alignment, annotation, and ortholog identification) before they become useful [3]. Furthermore, contaminations and deep unrecognized paralogy often cause single-gene trees and species trees to be incongruent [3]. However, when these issues are properly handled and orthologs are identified, phylogenies can be recovered through parsimony [4, 5], maximum

likelihood [6], Bayesian inference [7], or distance-based techniques such as neighbor-joining [8].

Alignment-free methods have recently gained traction because they do not require a sequence alignment, which allows species with unannotated genes to be placed on the species tree. Furthermore, the alignment-free algorithms are usually computationally inexpensive, which allows for more species to be compared together. Proponents of alignment-free techniques claim that they are resistant to shuffling and recombination events and are not affected by assumptions regarding a high correlation between sequence changes and evolutionary time [9]. Alignment-free techniques typically use Chaos Theory to calculate distances of basic genomic features (e.g., GC content, oligomer frequency, etc.) that are then used to recover the phylogeny [10, 11]. Since genomic features are compared between species instead of sequence homology, these genomic features can be located on different chromosomes or encompass the whole genome. More recently, another technique limits the alignment-free search space to all genic regions within a species, comparing species-wide patterns of codon aversion and amino acid aversion independent of gene annotations [12]. Generally, alignment-free approaches can be grouped into three main types. The first group determines the frequency of words of a certain length (e.g., FFP [13, 14] and CVTree [15]). The second group finds match lengths between sequences (e.g., ACS [16], KMACS [17], and Kr [18]). The last group calculates informational content between sequences (e.g., Co-phylog [19], FSWM [20], andi [21], and CAM [12]).

Although sequence alignments are typically used in parsimony, other features, such as the aversion to certain codons, can also be used to recover phylogenies [22, 23]. In those analyses, each ortholog was encoded with 64 characters, one for each codon. Codons were given a binary representation of '1' if the codon was used within the ortholog and '0' if the codon was not used, each character was added to a matrix, and phylogenies were recovered using parsimony and alignment-free methods.

Codons are sequences of three consecutive nucleotides of coding DNA that are transcribed into mRNA, mRNA is translated into amino acids, and amino acids form proteins [24]. The 20 canonical amino acids are formed from 61 codons, with the other three codons encoding the stop signal [25]. Although multiple codons encode the same amino acid, both mutation and selection can cause an unequal distribution of synonymous codons to occur within species [26], suggesting that synonymous codons might play different roles in species fitness [27]. An unequal distribution of tRNA anticodons directly coupling codons led to the wobble hypothesis: tRNA anticodons do not need to bind to all three codon nucleotides for translation [28]. Codon usage is also highly associated with the most abundant tRNA present in the cell [29] and codon usage patterns affect gene expression [30].

Recharging a tRNA while the tRNA is still attached to the ribosome increases translational efficiency and decreases overall resource utilization. This process occurs when codons encoding the same amino acid are located in close proximity to each other on the mRNA strand [31]. Co-tRNA codon pairing is when two non-identical codons that encode the same amino acid are near each other in a gene and the tRNA is recharged to translate both codons before the tRNA diffuses. Similarly, identical codon pairing occurs when identical codons are near each other in a gene and the tRNA is recharged to translate both codons before the tRNA diffuses. Co-tRNA and identical codon pairing conserve resources and increase translational efficiency by approximately 30% [31]. Co-tRNA codon pairing has previously been reported as more prominent in eukaryotes, while identical codon pairing has been reported in eukaryotes, bacteria [32], and archaea [33].

Here, we present two novel approaches that capitalize on biases in codon pairing to determine species relationships using either a parsimony or alignment-free method. We perform a

Table 1. Number of species passing preprocessing filters and analyzed by each algorithm.

Taxonomic Group	Alignment-free	Parsimony	Maximum Likelihood	NCBI Taxonomy	OTL
All	23 428	0	0	22 794	12 337
Archaea	418	100	418	416	362
Bacteria*	15 068	0	0	14 612	11 227
Fungi	234	0	58	234	214
Invertebrates	149	57	57	149	147
Plants	89	61	60	89	87
Protozoa	75	15	24	75	75
Mammals	107	97	100	107	105
Other vertebrates	123	114	118	123	120
Viruses*	7 233	0	0	7 045	0

The alignment-free methods did not require any preprocessing of the coding sequences. Parsimony used a stricter preprocessing cutoff than maximum likelihood, and therefore used fewer species. The NCBI taxonomy includes viruses and more species than the OTL. We did not run ortholog-based phylogenetic methods on taxonomic groups when fewer than 5% of the total species remained after initial filtering.

*Sixty-eight bacteria and viruses overlap.

https://doi.org/10.1371/journal.pone.0232260.t001

comprehensive analysis of codon pairing across the Tree of Life and explain how codon pairing can be implemented in a phylogenomic framework.

Results

Table 1 report the number of species that were included in each analysis after the preprocessing filters were applied (e.g., each species in the parsimony analysis included at least 5% of the parsimony-informative characters). In total, we included 23 428 species, with each species generally containing thousands of genes. S1–S3 Tables in S1 File report the number of species that were included for each ribosomal window size in each of the three parsimony analyses: identical codon pairing, co-tRNA codon pairing, and a combined approach. The alignment-free methods analyzed all species because the methods are not affected by missing ortholog calls.

We used reference phylogenies from the National Center for Biotechnology Information (NCBI) Taxonomy Browser [34–37] and the Open Tree of Life (OTL) [38]. The NCBI taxonomy contains more species than the OTL, and the OTL does not contain any viruses. The species trees vary between the OTL and the NCBI taxonomy by 1–9%, with the mammal phylogenies being the most similar and the fungi phylogenies being the least similar. Descriptions on filters applied to parsimony and maximum likelihood are found in S1 Text in S1 File.

After filtering for parsimony-informative codons, we used parsimony to recover phylogenies with the highest percent overlap based on binary representations of codon pairings (i.e., if the codon occurred in a gene, it was encoded as '1' and if it did not occur, it was encoded as '0'). We opted to use the branch congruence (i.e., percentage of edge similarity) metric to compare our recovered phylogenies because it is less sensitive to polytomies, small changes in leaf nodes, unrooted trees, and large phylogenies such as the trees recovered in this study. Branch congruence was also used by Miller, McKinnon [12], which facilitates the direct comparison of our results with the results from their study. The number of parsimony-informative codon pairings used in each comparison are reported in S4–S6 Tables in S1 File, with mammals typically having the most parsimony-informative characters and invertebrates typically having the fewest parsimony-informative characters.

Fig 1 shows the percent overlap of the unrooted trees recovered using the six codon pairing methods (identical codon pairing, co-tRNA codon pairing, and combined codon pairing for

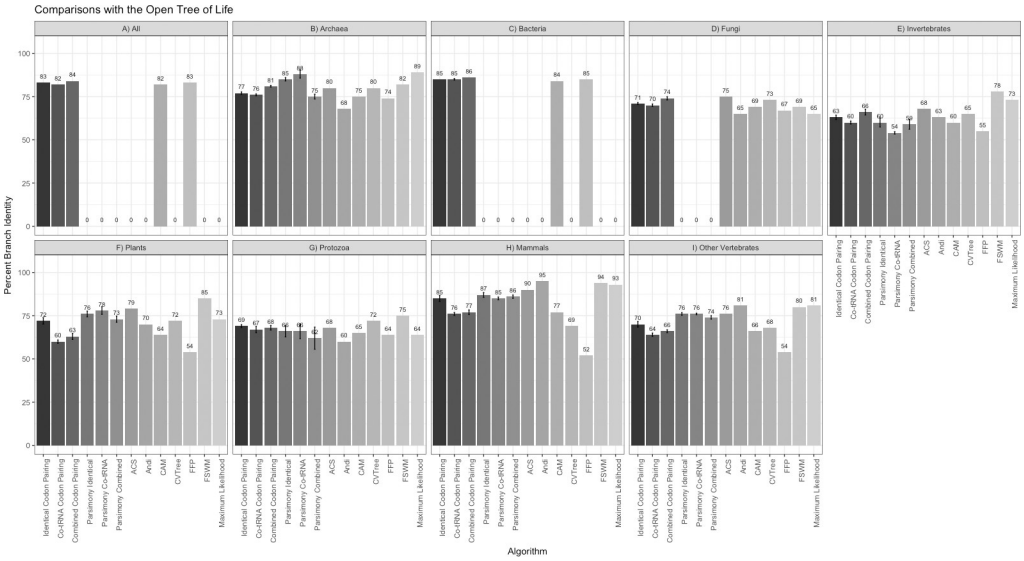

Fig 1. Percent edge overlap for comparisons of each algorithm against the OTL. The alignment-free and parsimony codon pairing methods report the mean percent edge overlap with the OTL based on using different ribosome windows from 2–11. Error bars are reported for the codon pairing methods, signifying one standard deviation from the mean. The other methods were previously reported in Miller, McKinnon [12] and are used for comparison.

https://doi.org/10.1371/journal.pone.0232260.g001

parsimony and alignment-free) compared to the OTL. For comparison, trees recovered from other alignment-free methods (CAM, FFP, CVTree, ACS, Andi, and FSWM) and maximum likelihood are also compared to the OTL in Fig 1. Fig 2 shows unrooted tree comparisons for each method compared to the NCBI taxonomy.

The alignment-free and parsimony codon pairing methods recovered phylogenetic relationships that are highly congruent with both the OTL and the NCBI taxonomy. The alignment-free method had a branch percent identity ranging from 62% to 86% with the OTL taxonomy, and a range of 68.4% to 92.6% for the NCBI taxonomy. The parsimony pairing method performed slightly better with branch percent identities ranging from 64% to 90% with the OTL, and 63.5% to 94% with the NCBI taxonomy. The alignment-free method outperformed parsimony when compared to the OTL only for invertebrates, which also had the fewest parsimony-informative characters and is paraphyletic.

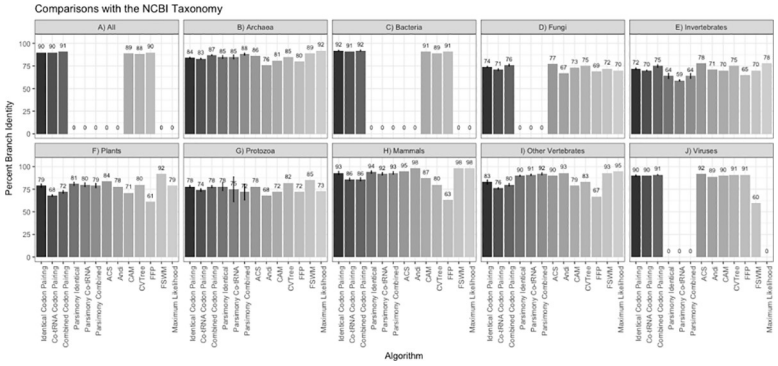

Fig 2. Percent edge overlap for comparisons of each algorithm against the NCBI taxonomy. The alignment-free and parsimony codon pairing methods report the mean percent edge overlap with the NCBI taxonomy based on using different ribosome windows from 2–11. Error bars are reported for the codon pairing methods, signifying one standard deviation from the mean. The other methods were previously reported in Miller, McKinnon [12] and are used for comparison.

https://doi.org/10.1371/journal.pone.0232260.g002

The comparisons of the codon pairing algorithms to maximum likelihood and other alignment-free algorithms show that none of the phylogenetic algorithms consistently recovered phylogenies with the highest percent edge similarity with the OTL or the NCBI taxonomy. However, the codon pairing algorithms consistently outperformed CAM and FFP. The codon pairing methods were comparable to maximum likelihood, CVTree, ACS, Andi, and FSWM in all taxonomic groups.

S7 Table in S1 File shows the optimal window sizes and the method (identical, co-tRNA, or combined codon pairing) that recovered the most congruent tree with the established phylogenies, with a description of the results in S2 Text in S1 File. S8-S19 Tables in S1 File report the percent edge overlap for identical, co-tRNA, and combined codon pairing compared to the OTL and the NCBI taxonomy for both the alignment-free and parsimony approaches at each ribosome window size from 2–11 codons. For both the alignment-free and parsimony approaches, combining co-tRNA codon pairing with identical codon pairing produced the most congruent tree with the OTL and the NCBI Taxonomy in most taxonomic groups.

We also compared the codon pairing motifs (i.e., the set of codons that paired within a gene) across each taxonomic group. For example, a gene that has identical codon pairing for AAA and AAT would have a motif of {AAA, AAT}, regardless of how many times AAA or AAT paired. We found that fewer than 10% of codon pairing motifs were identified in multiple species in most taxonomic groups (see S1–S10 Figs in S1 File). Bacteria had the most repeated codon pairing motifs (13.7%) and fungi had the fewest repeated motifs (0.7%).

S3 Text in S1 File describes the frequency of codon pairing in each taxonomic group, with S11–S19 Figs in S1 File showing boxplots of the percentage of genes that have codon pairing for each codon. We further analyzed the number of codons that paired within each gene. We counted the number of codon pairing motifs that included 1, 2, 3,. . ., 61 codons and report the distribution for each taxonomic group in S20–S29 Figs in S1 File. In most taxonomic groups, each motif contains ~10–40 codons. However, bacteria, archaea, and viruses are more likely to have fewer codons in each motif, while vertebrates typically have more codons in each motif.

We quantified the frequency of repeated motifs by counting the number of times each motif was used in each taxonomic group. S30–S39 Figs in S1 File show the distribution of repeated motif frequencies in each taxonomic group. In most taxonomic groups, most repeated motifs are repeated 1–20 times with a steep decreasing slope as the motif is repeated more frequently. However, in archaea, the number of times a motif repeats quickly decreases between 1–30 and then the slope increases until 61 before sharply dropping to near zero. We also found that although the total number of codon pairings is highly correlated with gene length (R-squared = 0.973–1.0; see S40–S48 Figs in S1 File), the number of codons that pair at least once in each gene is not correlated with gene length (R-squared = 0.0–0.08; see S49–S57 Figs in S1 File). Therefore, the binary encoding of codon pairing is not driven by gene length. The scripts we used to create each supplementary Fig can be found at https://github.com/ridgelab/codon_pairing/supplementary_graphs.

Since we used a distance-based approach for the alignment-free method, we tested for saturation in each of the taxonomic groups by graphing the computed distance against the taxonomic distance of the compared species. As the taxonomic distance increases, we expect the computed distance to also increase. However, computed distances of 1.0 indicate that full saturation has occurred for the metric and additional species cannot be differentiated using that method. S58–S66 Figs in S1 File show saturation often occurred for identical codon pairing, but the distances for co-tRNA and combined methods rarely fully saturated. Therefore, the co-tRNA and combined alignment-free methods have sufficient diversity of codon pairing motifs to recover the large phylogenies in our analyses, but the identical codon pairing alignment-free method may suffer from long branch attraction.

Additionally, we calculated the retention index of codon pairing for the parsimony method. A retention index of 1.0 indicates that the recovered tree contains no reversals, parallel gains, or parallel losses. We conducted 1 000 random permutations of species placement on the OTL tree topology, calculated the retention index of codon pairing for each of the random permutations, and then plotted the average codon pairing retention index from the OTL against the distribution of the average retention index of the permutated trees for each taxonomic group in S67–S72 Figs in S1 File. If codon pairing has a strong phylogenetic signal, we would expect the observed average retention index to exceed the average from all permutated trees. The observed average retention index was significantly higher than the permutations in archaea, invertebrates, mammals, other vertebrates, and plants. The paraphyletic group, protozoa, was the only taxonomic group where a few permutations had slightly higher average retention indices. These results explain why the trees recovered using parsimony were largely congruent with the OTL.

Discussion

We show that both identical and co-tRNA codon pairing are phylogenetically conserved across all domains of life. We further illustrate that combining identical and co-tRNA codon pairing improves the concordance of recovered phylogenies with the NCBI taxonomy and the OTL in most taxonomic groups. This comprehensive analysis shows that codon pairing is a novel phylogenetic character state that can be used in conjunction with other phylogenomic methods. Additionally, we provide tools for quickly analyzing thousands of species using our provided framework.

As opposed to common ortholog-based techniques that use shared character states to infer phylogenies, identical and co-tRNA codon pairing analyze sequence features that are associated with gene expression. Since gene expression plays a crucial role in adaptive divergence and ecological speciation [39], and codon pairing affects gene expression, we propose that detectable patterns in codon pairing not only inform phylogeny, but may track phenotypic variation between species. We show that codon pairing alone can recover phylogenies that are comparable to other alignment-free or maximum likelihood approaches, and patterns in codon pairing are largely conserved between species.

In some instances, codon pairing (or lack of codon pairing) might be due to protein structure instead of translational efficiency. Arginine (Arg) is very positively charged and highly repulsive to other like-charged amino acids. Although rarely pairing compared to other amino acid residues, arginine pairing is essential to some protein-protein interactions and occurs more frequently than expected by random chance [40]. In protein folding, coiled-coil interfaces often make asparagine (Asn)-Asn conformations that face away from the hydrophobic core [41]. Since coiled-coil proteins are present throughout all domains of life and occur in upwards of 10% of the proteome [42], it is likely that they play a nontrivial role in affecting codon pairing. Our analysis of codon pairing confirms that asparagine pairing occurs much more frequently than arginine pairing. These interactions suggest that asparagine and arginine pairing conservation might be based on structure instead of codon translational efficiency. However, the structural implications of codon usage do not rule out the additional effects of tRNA composition. Interestingly, although many species use only tRNA GTT [43],which may contribute to the bias for AAC codons and AAC codon pairing, AAT actually pairs more frequently in invertebrates, plants, protozoa, and viruses (see S11–S19 Figs in S1 File), suggesting that structural implications may drive codon pairing in addition to tRNA composition.

Leucine zipper T cell receptors have the highest expression values [44]. Furthermore, the leucine zipper is a 60–80 amino acid protein domain that allows for faster gene expression,

sequence-specific DNA-binding, and dimerization [45]. Our results show that leucin-encoding codons are among the most commonly paired codons. However, leucine-encoding CTA pairs significantly less frequently than other leucine-encoding codons. Further exploration into CTA interactions with other leucine-encoding codons may help determine why CTA pairs much less frequently.

Although co-tRNA codon pairing is less prominent in prokaryotes than in eukaryotes [26, 32, 33], we show that identical codon pairing and co-tRNA codon pairing are both phylogenetically conserved in all domains of life. However, we also show that using an alignment-free framework, the most congruent vertebrate and plant phylogenies are generally recovered using only identical codon pairing. Similarly, the parsimony method recovered the most congruent mammal phylogeny using only identical codon pairing. However, parsimony used only co-tRNA codon pairing in plants and the combined approach in non-mammalian vertebrates to recover the most congruent phylogenies. We show that although identical and co-tRNA codon pairing do not occur in equal frequencies, they are both phylogenetically conserved. We also show that combining identical and co-tRNA codon pairing recovers phylogenies that most support established phylogenies in seven out of ten taxonomic groups.

We recognize that systematic biases likely exist in RefSeq and may affect the results of our analyses. However, all algorithms are subject to the same limitations, and codon pairing performs comparably to these other algorithms. Furthermore, our filtering criteria for parsimony requires species to contain at least 5% of orthologs and orthologs to be called in at least 5% of the species to limit the effects of missing data. The number of species and parsimony-informative characters remaining after filtering are shown in S1-S6 Tables in S1 File. We opted to use this Big Data approach because we were interested in macro trends of codon pairing that we were able to identify through this comprehensive analysis. Furthermore, although codon pairing affects translational efficiency, the underlying mechanism governing codon pairing biases, like other codon usage biases, could be mutational biases or selection. Codon aversion biases, codon composition biases, mutational biases, and tRNA abundance might be the main source of the observed phylogenetic signal, although codon pairing appears to track that signal.

We used codon pairing to assess the controversial placement of falcons or pigeons as sister taxa to Neoaves, which was least supported by Shen, Hittinger [46]. The two phylogenies that we tested are found in S1 and S2 Phylogenies. We analyzed codon pairings using the conditional probability of the observed character states tracking a phylogeny, as established by Miller, McKinnon [23]. Our analyses add additional support to the current placement of pigeons on the Open Tree of Life as sister taxa to Neoaves because the conditional probability of codon pairings tracking the Open Tree of Life are higher than the probability of codon pairings tracking the alternative placement of falcons (see S20 Table in S1 File). Therefore, we propose keeping the current placement of falcons on the Open Tree of Life.

In taxonomic groups that have well-documented orthologous relationships, we show that codon pairing recovers parsimony trees that are largely congruent with the OTL and the NCBI taxonomy. Since maximum likelihood has been widely used to establish the reference phylogenies that we used for our comparisons, it is unsurprising that in the most established taxonomic groups, such as vertebrates, maximum likelihood recovers trees that are most congruent with the references. However, in plants and protozoa, the parsimony analysis elucidates a phylogenetic signal using only codon pairing that is sufficient to recover more congruent trees with the OTL and the NCBI taxonomy than maximum likelihood. Given the high degree of congruence between the established phylogenies, phylogenies recovered using other techniques, and the trees recovered using only codon pairing, we propose that codon pairing should be considered in future phylogenomic analyses.

Materials and methods

Data collection and processing

We downloaded all reference genomes and annotations from NCBI [47–49] in September, 2017. Reference genomes were used because they represent the most commonly accepted nucleotides in each species[48, 49]. We used the coding sequences (CDS) from the longest isoform of each gene, and we removed genes with previously-annotated exceptions (e.g., translational exception, unclassified transcription discrepancy, suspected errors, partial genes, etc.). A total of 23 428 species were divided into the following taxonomic groups based on RefSeq annotations, with some overlap between bacteria and viruses: 418 archaea, 15 063 bacteria, 234 fungi, 149 invertebrates, 89 plants, 75 protozoa, 107 mammalian vertebrates, 123 other vertebrates, and 7 233 viruses. While invertebrates, other vertebrates, and protozoa are paraphyletic outgroups, we opted to maintain these species classifications to facilitate analyses between different studies that use RefSeq. Furthermore, all algorithms will be subject to the same potential biases associated with analyzing paraphyletic groups.

Accounting for differences in ribosomal footprint

Estimates of the ribosome footprint vary drastically and can range from 15 nucleotides (5 codons) to about 45 nucleotides (15 codons) with a commonly accepted length of 28 nucleotides (about nine codons) [50]. Since codon pairing requires at least two codons, we examined pairing lengths (i.e., a sliding window) of 2–11 codons. This technique allows for variations in the ribosomal footprint among different taxonomic groups and can determine if codon pairing is dispersed throughout the ribosomal footprint or is more phylogenetically conserved at a smaller window size.

Calculating identical and co-tRNA codon pairing

For both the parsimony and alignment-free methods, we encoded identical codon pairings, co-tRNA codon pairings, and either identical and co-tRNA codon pairings with a binary representation (i.e., if a codon paired within a gene, it was given a value of '1' regardless of the number of times the pairing occurred). We determined which codons used identical codon pairing for each gene by adding each codon that occurred multiple times within the sliding window to a set of codons for that gene. Similarly, we created a set of amino acids for co-tRNA codon pairings for each gene by adding the amino acid product of the paired non-identical codons that encode that amino acid to the ordered set. Since the combined approach can use either identical or co-tRNA codon pairing, we calculated combined pairing by translating the gene sequence and identifying amino acids that paired within the ribosome window, adding each residue occurring multiple times in the sliding window to a set.

Alignment-free codon pairing calculation

We present three alignment-free methods to calculate a distance matrix: 1) based on identical codon pairing, 2) based on co-tRNA codon pairing, and 3) based on a combination of either identical or co-tRNA codon pairing. Although genes must be assembled, orthologous relationships are not required or used in the distance matrix calculation. All three methods use a binary (occurs or does not occur) representation of codon pairing within a gene. First, if identical codon pairing occurs anywhere within a gene, the codons are added to a set for that gene. If co-tRNA codon pairing or the combined approach is selected, then amino acids are added to a set if they occur two or more times within the ribosomal footprint anywhere in the gene. Next, the sets are alphabetized and converted to a tuple (immutable list) so they can be added

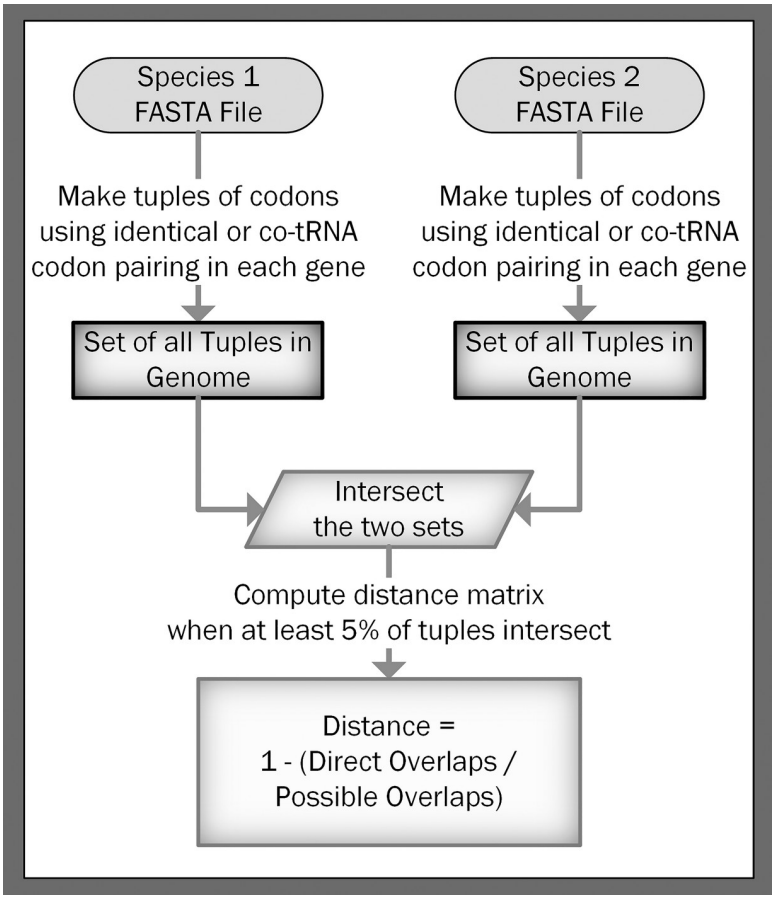

Fig 3. Process to calculate the distance matrix. Starting with the coding sequences of each gene in a species (FASTA file), codons that use codon pairing within the ribosomal footprint are included in a tuple that is then added to a set for that species. Sets of tuples are intersected to calculate the distance between species. These distances are then added to a distance matrix that can be used to recover phylogenies.

https://doi.org/10.1371/journal.pone.0232260.g003

to a set for the entire species. This process is repeated for each gene within a species until all gene pairings have been made into tuples and added to a set for the species. We repeat this process for each species until all species have a set of tuples representing the codons (or amino acids) that are pairing within at least one gene. Finally, we calculate the distances between each species in a pairwise manner. This process is depicted in Fig 3.

Similar to the method used by Miller, McKinnon [12], the pairwise distance between two species, *A* and *B*, is calculated as one minus the relative similarity of the species. The relative similarity of the species is the number of overlapping tuples between the sets of tuples, *a* and *b*, from both species divided by the total number of tuples from *a* or *b* with the fewest number of tuples. This distance is given in Eq 1:

$$Dist(A, B) = 1 - \frac{|a \cap b|}{\min(|a|, |b|)} \quad \text{Eq1}$$

If the ratio of tuples in *a* and *b* does not exceed 5%, the species are assigned the maximum distance of 1.0. This filter limits small genome bias (e.g., without this cutoff, if one gene from a virus with two genes has the same codon pairing profile as a gene in a vertebrate with 20 000 genes, then the distance between the virus and the vertebrate would be 0.5). This process

allows us to calculate a distance, with a maximum of 1.0, where more closely related species have a smaller distance because their overall codon pairing biases across all genes are more similar. A summary of alignment-free options is found in S4 Text in S1 File.

Parsimony analysis

We used the NCBI gene annotations for our parsimony analysis, which includes annotations from species-specific nomenclature committees, NCBI staff curations, and the NCBI annotation pipeline. We used Python 3.5 to implement `parsimony_pairing.py` to create a character matrix of parsimony-informative codon pairings from a directory of FASTA files containing gene sequences for each species, one file per species. Each row in the matrix contains a record for a different species. Each column in the matrix represents a parsimony-informative codon (or amino acid) within a specific ortholog. For each species, each codon (or amino acid) in each ortholog is labelled '0' if it does not pair within a ribosomal window, '1' if it does pair, or '?' if the ortholog call is unavailable for that species.

To be considered parsimony-informative, each included ortholog was present in at least four species, each codon (or amino acid) paired in at least one species, and each codon (or amino acid) did not pair in at least one species. We further required all species to contain at least 5% of all the parsimony-informative codons (or amino acids) to limit the effect of missing data. We created this character matrix and a key file containing an ordered list of each parsimony-informative codon (or amino acid) that was included in the matrix in a single step at runtime (see Fig 4). The following command demonstrates typical usage for identical codon pairing, where `$[51]` is the path to a directory containing one FASTA file per species, `${MATRIX}` is the path to the output matrix, and `${KEYS}` is the path to the output key file containing the ordered list of parsimony-informative codons.

```
python getPairingMatrix.py -id $[51] -o ${MATRIX} -oc ${KEYS}
```

A summary of program options is found in S5 Text in S1 File.

Constructing phylogenetic trees using parsimony

We used Tree Analysis Using New Technology (TNT) [52] to recover phylogenetic trees using parsimony. We selected TNT based on its ability to handle large datasets and its fast tree-searching algorithms. We found up to 100 most parsimonious trees, saving multiple trees recovered using tree bisection reconnection (tbr) branch swapping [53]. A discussion on the maximum number of species that can be placed on a tree using this method is found in S6 Text in S1 File.

Reference phylogenies

We inferred subtrees of each taxonomic group from both the OTL and the NCBI Taxonomy Browser for each taxonomic group. The OTL combines phylogenetic relationships reported in primary literature and contains a web application programming interface (API) that allows for querying the OTL database. Although the NCBI Taxonomy Browser gathers information from a variety of sources and is therefore not considered a primary source for taxonomic relationships, it contains more species than the OTL and provides added insights into our analyses. We use both phylogenies as reference trees to compare the alignment-free and parsimony trees obtained from codon pairing.

Open Tree of Life

We used `getOTLtree.py` [12] to obtain reference trees for each taxonomic group from the OTL in a single step at runtime. This program utilizes the OTL application programming

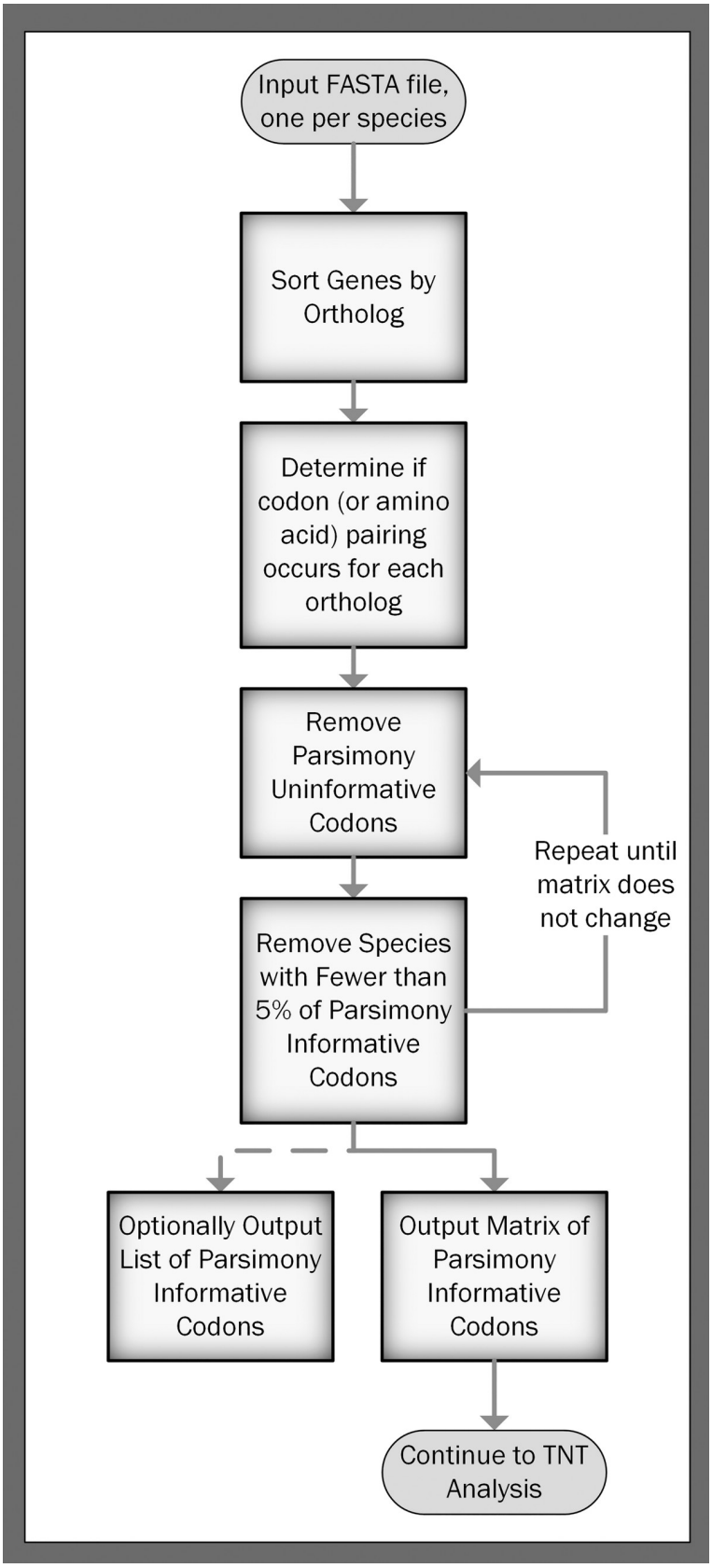
Input FASTA file, one per species
Sort Genes by Ortholog
Determine if codon (or amino acid) pairing occurs for each ortholog
Remove Parsimony Uninformative Codons
Repeat until matrix does not change
Remove Species with Fewer than 5% of Parsimony Informative Codons
Optionally Output List of Parsimony Informative Codons
Output Matrix of Parsimony Informative Codons
Continue to TNT Analysis

Fig 4. Flow chart for the parsimony analysis. We start with input FASTA files, one for each species. For each codon (or amino acid) within each ortholog, we assign a binary value of '0', '1', or '?' depending on if codon pairing for that codon (or amino acid) occurs. We then remove parsimony-uninformative characters. Next, we remove any species that do not contain at least 5% of the parsimony-informative codons, and we conduct the analysis only if at least 5% of the species pass the filter. Finally, we output the parsimony-informative character matrix for each codon (or amino acid) pairing to be used in a TNT analysis and an optional list of parsimony-informative characters.

https://doi.org/10.1371/journal.pone.0232260.g004

interface (API) to programmatically query the OTL database to first obtain OTL taxonomy identifiers (OTT ids) for each species and then query the OTL database to retrieve the reference tree for the species found. The program also allows users to select the correct domain of life when multiple OTT ids are found for a species (e.g., *Nannospalax galili* is currently listed in the OTL database as both a eukaryote and a bacterium). The output file contains the inferred reference tree from the OTL and a list of any species that the OTL did not include in the tree. We ran this program using the following command, where `${INPUT}` is a list of species, and `${OUTPUT}` is the output file:

```
python getOTLtree.py -i ${INPUT} -o ${OUTPUT}
```

NCBI taxonomy browser

We used the NCBI Taxonomy Browser (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) to download the taxonomical relationships of each taxonomic group in PHYLIP [54] format. We included unranked taxa to maximize the number of included species for each taxonomic group.

Tree comparison

We assessed the accuracy of our identical, co-tRNA, and combined codon pairing methods by comparing the trees we recovered to the reference trees from the OTL and the NCBI taxonomy. While both the OTL and the NCBI taxonomy combine phylogenetic trees presented by a variety of sources that may include source-specific biases, both phylogenies facilitate large-scale analyses across large and diverse taxonomic groups. We determined the similarity between trees by using the ete-compare module from the Environment for Tree Exploration toolkit (ETE3) [55], which computes the percentage of branch similarity between two trees. A higher percentage of branch similarity indicates higher congruence between trees. The branch similarity method has a relatively low computational cost for large datasets, and it allows for unrooted tree comparisons and comparisons of trees with polytomies. For the parsimony analysis, if any taxonomic comparison produced more than one equally parsimonious tree, we computed the percentage of edge similarity between each generated tree and the reference tree. We then reported the average percent overlap of all comparisons. Descriptions of the other methods used in our comparisons are found in S7 Text in S1 File.

Supporting information

S1 File.
(DOCX)

Acknowledgments

We appreciate the support of Brigham Young University and the technical assistance of the Fulton Supercomputing Laboratory staff.

Author Contributions

Conceptualization: Justin B. Miller, Perry G. Ridge.

Data curation: Justin B. Miller.

Formal analysis: Justin B. Miller, Lauren M. McKinnon.

Investigation: Justin B. Miller, Lauren M. McKinnon.

Methodology: Justin B. Miller, Lauren M. McKinnon, Michael F. Whiting, John S. K. Kauwe, Perry G. Ridge.

Project administration: Perry G. Ridge.

Resources: Perry G. Ridge.

Software: Justin B. Miller.

Supervision: Perry G. Ridge.

Validation: Justin B. Miller, Lauren M. McKinnon, Michael F. Whiting.

Writing – original draft: Justin B. Miller, Lauren M. McKinnon, Michael F. Whiting, Perry G. Ridge.

Writing – review & editing: Justin B. Miller, Lauren M. McKinnon, Michael F. Whiting, John S. K. Kauwe, Perry G. Ridge.

References

1. Soltis DE, Soltis PS. The Role of Phylogenetics in Comparative Genetics. Plant Physiology. 2003; 132 (4):1790–800. https://doi.org/10.1104/pp.103.022509 PMC526274. PMID: 12913137
2. Haszprunar G. The types of homology and their significance for evolutionary biology and phylogenetics. Journal of Evolutionary Biology. 1992; 5(1):13–24. https://doi.org/10.1046/j.1420-9101.1992.5010013.x
3. Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, et al. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biology. 2011; 9(3):e1000602. https://doi.org/10.1371/journal.pbio.1000602 PMID: 21423652
4. Wilgenbusch JC, Swofford D. Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics. 2003;Chapter 6:Unit 6 4. https://doi.org/10.1002/0471250953.bi0604s00 PMID: 18428704.
5. Farris J. The logical basis of phylogenetic analysis. 1983.
6. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981; 17(6):368–76. Epub 1981/01/01. https://doi.org/10.1007/BF01734359 PMID: 7288891.
7. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. 2012; 13(5):303–14. https://doi.org/10.1038/nrg3186 PMID: 22456349
8. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4(4):406–25. Epub 1987/07/01. https://doi.org/10.1093/oxfordjournals.molbev.a040454 PMID: 3447015.
9. Bonham-Carter O, Steele J, Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. Briefings in Bioinformatics. 2014; 15(6):890–905. https://doi.org/10.1093/bib/bbt052 WOS:000345386100002. PMID: 23904502
10. Vinga S, Almeida J. Alignment-free sequence comparison—a review. Bioinformatics. 2003; 19(4):513–23. https://doi.org/10.1093/bioinformatics/btg005 PMID: 12611807
11. Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. Inferring phylogenies of evolving sequences without multiple sequence alignment. Sci Rep. 2014; 4:6504. https://doi.org/10.1038/srep06504 PMID: 25266120; PubMed Central PMCID: PMC4179140.
12. Miller JB, McKinnon LM, Whiting MF, Ridge PG. CAM: An alignment-free method to recover phylogenies using codon aversion motifs. PeerJ Preprints. 2019; 7:e27756v1. https://doi.org/10.7287/peerj.preprints.27756v1

13. Jun S-R, Sims GE, Wu GA, Kim S-H. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proceedings of the National Academy of Sciences. 2010; 107(1):133–8. https://doi.org/10.1073/pnas.0913033107 PMID: 20018669

14. Sims GE, Jun SR, Wu GA, Kim SH. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci U S A. 2009; 106(8):2677–82. Epub 2009/02/04. https://doi.org/10.1073/pnas.0813249106 PMID: 19188606; PubMed Central PMCID: PMC2634796.

15. Zuo G, Hao B. CVTree3 Web Server for Whole-genome-based and Alignment-free Prokaryotic Phylogeny and Taxonomy. Genomics Proteomics Bioinformatics. 2015; 13(5):321–31. Epub 2015/11/14. https://doi.org/10.1016/j.gpb.2015.08.004 PMID: 26563468; PubMed Central PMCID: PMC4678791.

16. Ulitsky I, Burstein D, Tuller T, Chor B. The average common substring approach to phylogenomic reconstruction. J Comput Biol. 2006; 13(2):336–50. Epub 2006/04/07. https://doi.org/10.1089/cmb.2006.13.336 PMID: 16597244.

17. Leimeister CA, Morgenstern B. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. Bioinformatics. 2014; 30(14):2000–8. Epub 2014/05/16. https://doi.org/10.1093/bioinformatics/btu331 PMID: 24828656; PubMed Central PMCID: PMC4080746.

18. Haubold B, Pfaffelhuber P, Domazet-Loso M, Wiehe T. Estimating mutation distances from unaligned genomes. J Comput Biol. 2009; 16(10):1487–500. Epub 2009/10/07. https://doi.org/10.1089/cmb.2009.0106 PMID: 19803738.

19. Yi H, Jin L. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. Nucleic Acids Res. 2013; 41(7):e75. Epub 2013/01/22. https://doi.org/10.1093/nar/gkt003 PMID: 23335788; PubMed Central PMCID: PMC3627563.

20. Leimeister CA, Sohrabi-Jahromi S, Morgenstern B. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. Bioinformatics. 2017; 33(7):971–9. Epub 2017/01/12. https://doi.org/10.1093/bioinformatics/btw776 PMID: 28073754; PubMed Central PMCID: PMC5409309.

21. Haubold B, Klotzl F, Pfaffelhuber P. andi: fast and accurate estimation of evolutionary distances between closely related genomes. Bioinformatics. 2015; 31(8):1169–75. Epub 2014/12/17. https://doi.org/10.1093/bioinformatics/btu815 PMID: 25504847.

22. Miller JB, Hippen AA, Belyeu JR, Whiting MF, Ridge PG. Missing something? Codon aversion as a new character system in phylogenetics. Cladistics. 2017:545–56. https://doi.org/10.1111/cla.12183

23. Miller JB, McKinnon LM, Whiting MF, Ridge PG. Codon use and aversion is largely phylogenetically conserved across the tree of life. Molecular Phylogenetics and Evolution. 2020; 144:106697. https://doi.org/10.1016/j.ympev.2019.106697 PMID: 31805345

24. Crick F. Central dogma of molecular biology. Nature. 1970; 227(5258):561–3. https://doi.org/10.1038/227561a0 PMID: 4913914.

25. Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. General nature of the genetic code for proteins. Nature. 1961; 192:1227–32. https://doi.org/10.1038/1921227a0 PMID: 13882203.

26. Quax TE, Claassens NJ, Soll D, van der Oost J. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell. 2015; 59(2):149–61. https://doi.org/10.1016/j.molcel.2015.05.035 PMID: 26186290; PubMed Central PMCID: PMC4794256.

27. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 1986; 24(1–2):28–38. https://doi.org/10.1007/BF02099948 PMID: 3104616.

28. Crick FH. Codon—anticodon pairing: the wobble hypothesis. J Mol Biol. 1966; 19(2):548–55. https://doi.org/10.1016/s0022-2836(66)80022-0 PMID: 5969078.

29. Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proc Natl Acad Sci U S A. 1979; 76(4):1697–701. https://doi.org/10.1073/pnas.76.4.1697 PMID: 377281; PubMed Central PMCID: PMC383457.

30. Gutman GA, Hatfield GW. Nonrandom utilization of codon pairs in Escherichia coli. Proc Natl Acad Sci U S A. 1989; 86(10):3699–703. https://doi.org/10.1073/pnas.86.10.3699 PMID: 2657727; PubMed Central PMCID: PMC287207.

31. Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, et al. A role for codon order in translation dynamics. Cell. 2010; 141(2):355–67. https://doi.org/10.1016/j.cell.2010.02.036 PMID: 20403329.

32. Shao ZQ, Zhang YM, Feng XY, Wang B, Chen JQ. Synonymous codon ordering: a subtle but prevalent strategy of bacteria to improve translational efficiency. PLoS One. 2012; 7(3):e33547. https://doi.org/10.1371/journal.pone.0033547 PMID: 22432034; PubMed Central PMCID: PMC3303843.

33. Zhang YM, Shao ZQ, Yang LT, Sun XQ, Mao YF, Chen JQ, et al. Non-random arrangement of synonymous codons in archaea coding sequences. Genomics. 2013; 101(6):362–7. https://doi.org/10.1016/j.ygeno.2013.04.008 PMID: 23603537.

34. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2012; 40(Database issue):D13–25. Epub 2011/12/06. https://doi.org/10.1093/nar/gkr1184 PMID: 22140104; PubMed Central PMCID: PMC3245031.

35. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2011; 39(Database issue):D38–51. Epub 2010/11/26. https://doi.org/10.1093/nar/gkq1172 PMID: 21097890; PubMed Central PMCID: PMC3013733.

36. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2010; 38(Database issue):D5–16. Epub 2009/11/17. https://doi.org/10.1093/nar/gkp967 PMID: 19910364; PubMed Central PMCID: PMC2808881.

37. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2009; 37(Database issue):D5–15. Epub 2008/10/23. https://doi.org/10.1093/nar/gkn741 PMID: 18940862; PubMed Central PMCID: PMC2686545.

38. Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci U S A. 2015; 112(41):12764–9. https://doi.org/10.1073/pnas.1423041112 PMID: 26385966; PubMed Central PMCID: PMC4611642.

39. Pavey SA, Collin H, Nosil P, Rogers SM. The role of gene expression in ecological speciation. Ann N Y Acad Sci. 2010; 1206:110–29. Epub 2010/09/24. https://doi.org/10.1111/j.1749-6632.2010.05765.x PMID: 20860685; PubMed Central PMCID: PMC3066407.

40. Lee D, Lee J, Seok C. What stabilizes close arginine pairing in proteins? Phys Chem Chem Phys. 2013; 15(16):5844–53. https://doi.org/10.1039/c3cp00160a WOS:000316803500014. PMID: 23486862

41. Thomas F, Niitsu A, Oregioni A, Bartlett GJ, Woolfson DN. Conformational Dynamics of Asparagine at Coiled-Coil Interfaces. Biochemistry. 2017; 56(50):6544–54. https://doi.org/10.1021/acs.biochem.7b00848 PMID: 29166010

42. Truebestein L, Leonard TA. Coiled-coils: The long and short of it. Bioessays. 2016; 38(9):903–16. Epub 2016/08/05. https://doi.org/10.1002/bies.201600062 PMID: 27492088.

43. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004; 32(17):5036–44. https://doi.org/10.1093/nar/gkh834 PMID: 15448185; PubMed Central PMCID: PMC521650.

44. Foley KC, Spear TT, Murray DC, Nagato K, Garrett-Mayer E, Nishimura MI. HCV T Cell Receptor Chain Modifications to Enhance Expression, Pairing, and Antigen Recognition in T Cells for Adoptive Transfer. Molecular Therapy Oncolytics. 2017; 5:105–15. https://doi.org/10.1016/j.omto.2017.05.004 PMC5447397. PMID: 28573185

45. Ellenberger T. Getting a grip on DNA recognition: structures of the basic region leucine zipper, and the basic region helix-loop-helix DNA-binding domains. Current Opinion in Structural Biology. 1994; 4 (1):12–21. https://doi.org/10.1016/S0959-440X(94)90054-X.

46. Shen X-X, Hittinger CT, Rokas A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecology & Evolution. 2017; 1(5):0126. https://doi.org/10.1038/s41559-017-0126 PMID: 28812701

47. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014; 42(Database issue):D756–63. https://doi.org/10.1093/nar/gkt1114 PMID: 24259432; PubMed Central PMCID: PMC3965018.

48. Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. 2000; 16(1):44–7. Epub 2000/01/19. https://doi.org/10.1016/s0168-9525(99)01882-x PMID: 10637631.

49. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2007; 35(Database issue):D5–12. https://doi.org/10.1093/nar/gkl1031 PMID: 17170002; PubMed Central PMCID: PMC1781113.

50. Martens AT, Taylor J, Hilser VJ. Ribosome A and P sites revealed by length analysis of ribosome profiling data. Nucleic Acids Res. 2015; 43(7):3680–7. https://doi.org/10.1093/nar/gkv200 PMID: 25805170; PubMed Central PMCID: PMC4402525.

51. Tan L, Wang H-F, Tan M-S, Tan C-C, Zhu X-C, Miao D, et al. Effect of CLU genetic variants on cerebrospinal fluid and neuroimaging markers in healthy, mild cognitive impairment and Alzheimer's disease cohorts. Scientific Reports. 2016; 6(1):26027. https://doi.org/10.1038/srep26027 PMID: 27229352

52. Goloboff PA, Farris JS, Nixon KC. TNT: Tree Analysis Using New Technology. 2005; 54:176–8. https://doi.org/10.1080/10635150590905830

53. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018; 35(6):1547–9. Epub 2018/05/04. https://doi.org/10.1093/molbev/msy096 PMID: 29722887; PubMed Central PMCID: PMC5967553.

54. Felsenstein J. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics. 1989; 5:164–6. citeulike-article-id:2344765.

55. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016; 33(6):1635–8. Epub 2016/02/28. https://doi.org/10.1093/molbev/msw046 PMID: 26921390; PubMed Central PMCID: PMC4868116.