D210–D218 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 18 October 2016*
doi: 10.1093/nar/gkw934

ATGC database and ATGC-COGs: an updated resource for micro- and macro-evolutionary studies of prokaryotic genomes and protein family annotation

David M. Kristensen[1,2,*], Yuri I. Wolf[2] and Eugene V. Koonin[2]

[1]Department of Biomedical Engineering, University of Iowa, Iowa City, IA 52242, USA and [2]National Center for Biotechnology Information, National Library of Medicine, National Institute of Health, Bethesda, MD 20894, USA

Received August 26, 2016; Revised October 05, 2016; Accepted October 12, 2016

ABSTRACT

The Alignable Tight Genomic Clusters (ATGCs) database is a collection of closely related bacterial and archaeal genomes that provides several tools to aid research into evolutionary processes in the microbial world. Each ATGC is a taxonomy-independent cluster of 2 or more completely sequenced genomes that meet the objective criteria of a high degree of local gene order (synteny) and a small number of synonymous substitutions in the protein-coding genes. As such, each ATGC is suited for analysis of microevolutionary variations within a cohesive group of organisms (e.g. species), whereas the entire collection of ATGCs is useful for macroevolutionary studies. The ATGC database includes many forms of pre-computed data, in particular ATGC-COGs (Clusters of Orthologous Genes), multiple sequence alignments, a set of 'index' orthologs representing the most well-conserved members of each ATGC-COG, the phylogenetic tree of the organisms within each ATGC, etc. Although the ATGC database contains several million proteins from thousands of genomes organized into hundreds of clusters (roughly a 4-fold increase since the last version of the ATGC database), it is now built with completely automated methods and will be regularly updated following new releases of the NCBI RefSeq database. The ATGC database is hosted jointly at the University of Iowa at dmk-brain.ecn.uiowa.edu/ATGC/ and the NCBI at ftp.ncbi.nlm.nih.gov/pub/kristensen/ATGC/atgc_home.html.

INTRODUCTION

As genome sequencing continues at its inexorably rapid pace, the resulting information provides unprecedented glimpses into the diversity of life in our biosphere. Numerous novel organisms, with new metabolic pathways and unique nanostructural designs continue to be discovered, primarily through metagenomics and single cell genomics. As the depth of coverage increases and multiple genome assemblies become available for a greater number of organisms, pangenomic analysis provides for in-depth study of microevolutionary changes. The Aligned Tight Genomic Clusters (ATGC) database is designed to aid in macro- and micro-evolutionary research by providing groups of organisms that meet the following criteria:

- *Alignable* – their genomes share synteny over ≥85% of their lengths
- *Tight* – synonymous substitution rate of ≤1.5 (indicating that the great majority of genes have a rate below saturation)
- *Genomic* – contains only completely-sequenced genome assemblies, allowing for maximally accurate determination of orthology and paralogy
- *Clusters* – groups of closely related organisms

One of the primary goals of the ATGCs resource is to provide a substantial amount of pre-calculated data that can be readily in used large-scale studies on evolution of bacteria and archaea (1,2). These data sets include clusters of orthologous genes (ATGC-COGs) that are constructed using the COGs approach (3–5) that, despite its simplicity, is widely considered to be one of the most accurate methods for orthology identification in prokaryotes (6–8). In comparison, other resources for pan-genomic studies (9), tend to use pairwise methods (such as USEARCH (10)) that are faster but less accurate than COGs. The ATGC-COGs are supplemented by multiple alignments of orthologous protein and nucleotide sequences and by lists of 'index orthologs', i.e. the most well-conserved members of an orthologous family from each of the included genomes. Additionally, each ATGC is accompanied by the phylogenetic tree of the constituent organisms and matrices of pairwise intergenomic distances based on genome-wide analysis of synonymous and non-synonymous substitution rates.

*To whom correspondence should be addressed. Tel: +1 319 335 5241; Email: david-kristensen@uiowa.edu

The ATGCs have been used to perform a large-scale study into the genome dynamics of prokaryotic supergenomes (quantifying rates of gene gains/losses, family expansions/reductions, etc.) (11); to measure the nature and intensity of selection pressure on CRISPR-associated genes (12); to analyze defense islands (13), toxin–antitoxin systems and related mobile stress response systems in prokaryotes (14); and for several other projects on specific aspects of microbial evolution (15–22).

The current release of the ATGCs database consists of 410 groups that jointly encompass >3700 genomes encoding 12.4 million proteins that are classified into >50 000 ATGC-COGs. This represents a roughly 4-fold increase in the size of the ATGC database since the previous version (1) with 104 ATGCs covering only hundreds of genomes. The extremely large size of the data set necessitated the development of automated processing to keep pace and to facilitate more regular updates. The ATGC data set itself, as well as ATGC-COGs and a substantial amount of pre-computed data, are made freely available at dmk-brain.ecn.uiowa.edu/ATGC/ and ftp.ncbi.nlm.nih.gov/pub/kristensen/ATGC/atgc_home.html.

DATABASE CONTENT

The ATGCs database currently consists of 410 groups of closely-related genomes of bacteria and archaea. Altogether these groups encompass >3700 complete genome assemblies that represent 74% of the prokaryotic genomes available in the NCBI's RefSeq database (23) as of June 2016 (Table 1). Reflecting the overall distribution of taxa among the complete genome assemblies in RefSeq, the majority of ATGCs represent the best-studied groups of bacteria and archaea (Figure 1A). For instance, among archaea, *Euryarchaeota* are covered much wider than *Crenarchaeota*, and among Bacteria, the best-represented phyla are *Proteobacteria* (e.g. *Escherichia*, *Pseudomonas*), *Firmicutes* (e.g. *Streptococcus*, *Staphylococcus*) and *Actinobacteria* (e.g. *Mycobacterium*), with several other phyla represented at lower coverage, such as *Tenericutes* (e.g. *Mycoplasma*), *Chlamydiae*, *Cyanobacteria*, etc. The ATGCs do not represent taxa that lack sufficient sampling depth, i.e. where no clusters of genomes are related closely enough to meet the inclusion criteria. Examples include the sparsely sampled archaeal phyla *Korarchaeota* and *Thaumarchaeota*, the proteobacterial class *Zetaproteobacteria* with no complete genomes available, and the firmicute class *Tissierellia*, with only four complete genome assemblies from four different genera available. Overall, the ATGCs cover the entire spectrum of characteristics represented among the complete prokaryotic genomes in RefSeq, such as genome size (from <1 Mb in bacterial symbionts with drastically reduced genome sizes, up to >10 MB in spore-producing soil bacteria such as *Streptomyces*), GC content (from <25% in *Mycoplasma*, up to 75% in *Anaeromyxobacter*) and optimal growth temperature (from psychrophilic bacteria to hyperthermophilic archaea and bacteria).

Most of the ATGCs consist of only a few closely-related genome assemblies but several large groups include many tens, and the largest ATGCs even include hundreds of genomes (Figure 1B). The majority of the ATGCs (96%) contain only members from the same genus or species, with the exceptions representing groups of bacteria that exchange many genes with one another in their common environment (e.g. the intestinal *Enterobacteriaceae*) or genera that are imprecisely defined (such as *Escherichia* versus *Shigella* (24), or *Vibrio* versus *Aliivibrio* (25)). This grouping of genomes that formally belong to different genera is allowed because the ATGCs were built to reflect objective criteria of genome relatedness, in order to provide data sets that are useful for evolutionary studies. In contrast, some genera and even species are split between multiple ATGCs (such as *Prochlorococcus marinus*, which forms 4 groups) when a group reaches a level of evolutionary divergence exceeding the ATGC criteria.

Of the 12.4 million proteins encoded in the >3700 genomes in the ATGCs, the overwhelming majority (~97%) are shared between two or more organisms. Collectively, these proteins form >50 000 ATGC-COGs.

DATABASE ACCESS

The ATGCs database is accessible through the University of Iowa at http://dmk-brain.ecn.uiowa.edu/ATGC/, ftp://dmk-brain.ecn.uiowa.edu/ATGC/atgc_home.html, or the NCBI at ftp://ftp.ncbi.nlm.nih.gov/pub/kristensen/ATGC/atgc_home.html. The homepage provides links to the main ATGC list page, bulk downloads and a help page. The main ATGC list page contains all pertinent information needed to search for a particular ATGC, such as the names of the genus and species it contains (Figure 2). This page also allows one to choose ATGCs that meet various pertinent criteria such as minimum and/or maximum synonymous substitution rate, GC content or genome size. Links are also provided to the most current COG pages, in cases when the species present in an ATGC overlap with the COG database (26).

Clicking on a particular ATGC takes the user to a page that gives in-depth information about that group (Figure 3). This page provides descriptive collective genome statistics, such as the number of genome assemblies, alongside a phylogenetic tree. Information on each of the genome assemblies in the given ATGC including the total number of genes, the number of proteins that are shared across multiple genomes, and the number of genome-specific proteins is given in the form of a table and a histogram. Finally, the page includes a download section where the pre-computed data for each ATGC is made available including ATGC-COGs, multiple sequence alignments, the high-resolution phylogenetic tree and matrices of pairwise intergenomic distances.

Clicking on a particular genome assembly takes the user further to a detailed page containing all protein-coding genes from that genome. For each gene, information is provided such as its ATGC-COG membership, quantitative measure of synteny support, genomic coordinates, product name, start and stop codon and various external links (gene name, protein accession, etc.). Unlike the COGs database, individual html pages are not provided for each of the >50 000 ATGC-COGs although addition of this feature is planned for a future release of the ATGCs database. This information is available instead via a page listing all of the

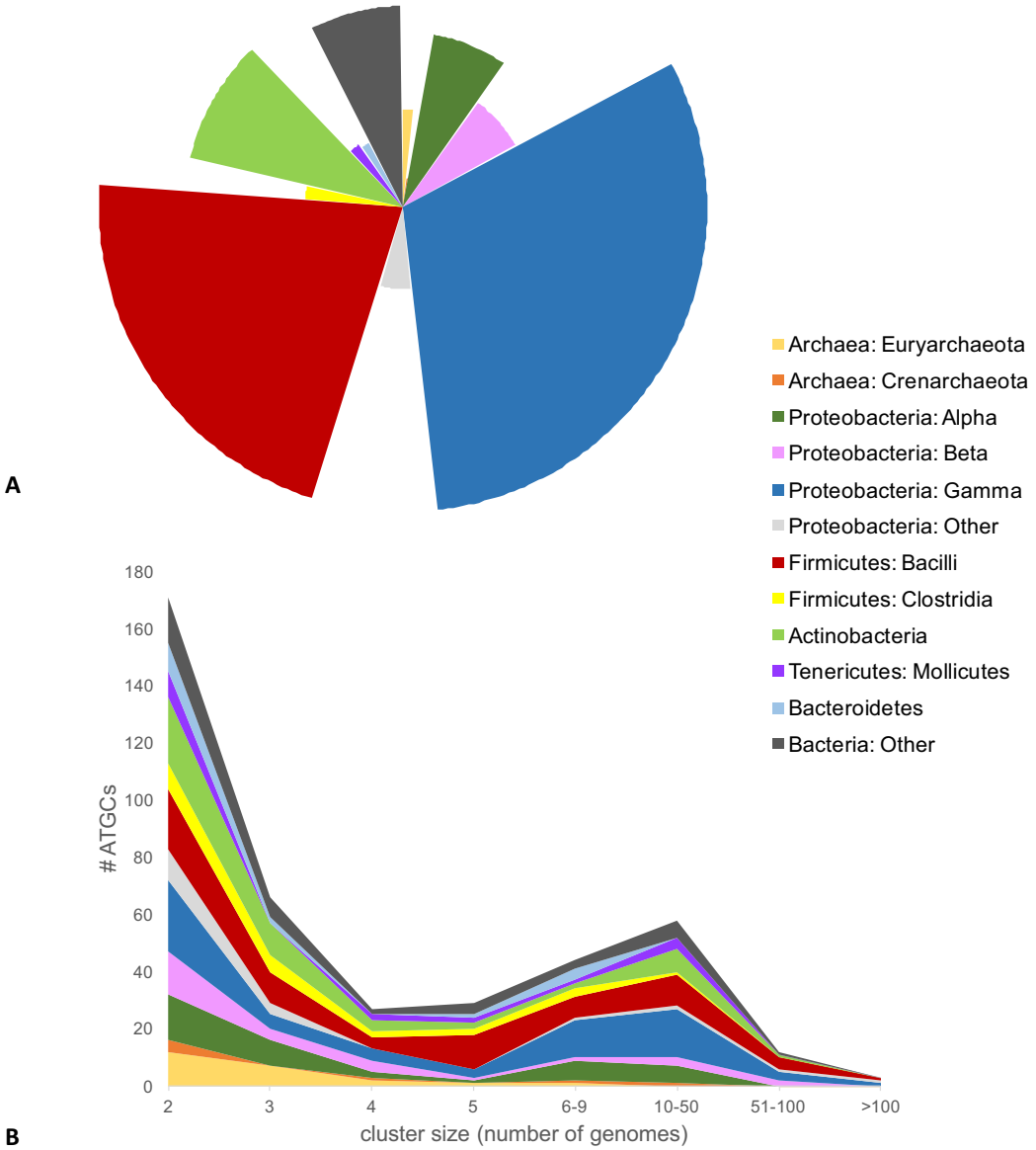
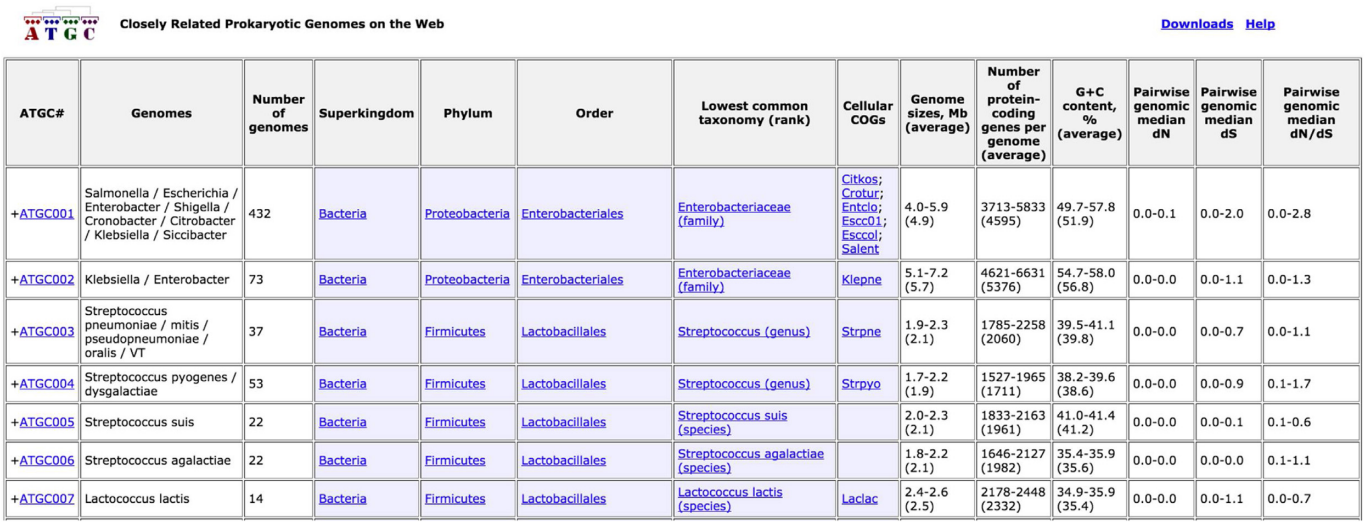


Figure 1. Distribution of archaeal and bacterial taxa in the Aligned Tight Genomic Clusters (ATGCs). (**A**) multidimensional chart showing taxa with more members as wider slices and those represented in more ATGCs as greater distance from the center. (**B**) Stacked area plot showing ATGC size.

Table 1. Taxonomic distribution of genomes.

Taxonomic group	NCBI Genomes		genomes in ATGCs		number of ATGCs	
Archaea: Euryarchaeota	146	3%	64	2%	23	6%
Archaea: Crenarchaeota	56	1%	38	1%	7	2%
Archaea: Other	6	0%	0	0%	0	0%
Proteobacteria: Alpha	391	8%	262	7%	41	10%
Proteobacteria: Beta	373	7%	273	7%	30	7%
Proteobacteria: Gamma	1384	28%	1163	31%	71	17%
Proteobacteria: Other	310	6%	236	6%	19	5%
Firmicutes: Bacilli	886	18%	805	22%	71	17%
Firmicutes: Clostridia	159	3%	91	2%	23	6%
Actinobacteria	523	10%	345	9%	51	12%
Tenericutes: Mollicutes	137	3%	101	3%	18	4%
Bacteroidetes	161	3%	63	2%	17	4%
Bacteria: Other	462	9%	262	7%	39	10%
Totals:	4 994		3 703		410	



ATGC Closely Related Prokaryotic Genomes on the Web

Downloads Help

ATGC#	Genomes	Number of genomes	Superkingdom	Phylum	Order	Lowest common taxonomy (rank)	Cellular COGs	Genome sizes, Mb (average)	Number of protein-coding genes per genome (average)	G+C content, % (average)	Pairwise genomic median dN	Pairwise genomic median dS	Pairwise genomic median dN/dS
+ATGC001	Salmonella / Escherichia / Enterobacter / Shigella / Cronobacter / Citrobacter / Klebsiella / Siccibacter	432	Bacteria	Proteobacteria	Enterobacteriales	Enterobacteriaceae (family)	Citkos; Crotur; Entclo; Escc01; Esccol; Salent	4.0-5.9 (4.9)	3713-5833 (4595)	49.7-57.8 (51.9)	0.0-0.1	0.0-2.0	0.0-2.8
+ATGC002	Klebsiella / Enterobacter	73	Bacteria	Proteobacteria	Enterobacteriales	Enterobacteriaceae (family)	Klepne	5.1-7.2 (5.7)	4621-6631 (5376)	54.7-58.0 (56.8)	0.0-0.0	0.0-1.1	0.0-1.3
+ATGC003	Streptococcus pneumoniae / mitis / pseudopneumoniae / oralis / VT	37	Bacteria	Firmicutes	Lactobacillales	Streptococcus (genus)	Strpne	1.9-2.3 (2.1)	1785-2258 (2060)	39.5-41.1 (39.8)	0.0-0.0	0.0-0.7	0.0-1.1
+ATGC004	Streptococcus pyogenes / dysgalactiae	53	Bacteria	Firmicutes	Lactobacillales	Streptococcus (genus)	Strpyo	1.7-2.2 (1.9)	1527-1965 (1711)	38.2-39.6 (38.6)	0.0-0.0	0.0-0.9	0.1-1.7
+ATGC005	Streptococcus suis	22	Bacteria	Firmicutes	Lactobacillales	Streptococcus suis (species)		2.0-2.3 (2.1)	1833-2163 (1961)	41.0-41.4 (41.2)	0.0-0.0	0.0-0.1	0.1-0.6
+ATGC006	Streptococcus agalactiae	22	Bacteria	Firmicutes	Lactobacillales	Streptococcus agalactiae (species)		1.8-2.2 (2.1)	1646-2127 (1982)	35.4-35.9 (35.6)	0.0-0.0	0.0-0.0	0.1-1.1
+ATGC007	Lactococcus lactis	14	Bacteria	Firmicutes	Lactobacillales	Lactococcus lactis (species)	Laclac	2.4-2.6 (2.5)	2178-2448 (2332)	34.9-35.9 (35.4)	0.0-0.0	0.0-1.1	0.0-0.7

Figure 2. The ATGC data set webpage. A user can choose an individual ATGC to find more information about it (link in left-most column), or use the provided information to choose a set of ATGCs for large-scale analyses. The color scheme was chosen to match that of the most current version of the COG database (blue, bacteria; orange, archaea), with links provided to organisms appearing in both databases in the 'COGs' column (middle of table).

ATGC-COGs and all of their member proteins within a given ATGC (accessible by clicking on the histogram chart or the link below it). However, for larger ATGCs containing several tens to hundreds of genome assemblies, it is recommended instead to simply access this information via the text files provided in the download section of an individual ATGC page. For example, the ATGC containing the *Escherichia coli* species (ATGC001, with 432 genome assemblies) represents nearly 2 million proteins which are collected into >17 700 ATGC-COGs that are shared between multiple genomes, plus an additional 8900 genome-specific genes. The next largest cluster containing *Staphylococcus aureus* species (ATGC052, with 109 genome assemblies) represents nearly 300 000 proteins, which are collected into almost 4000 ATGC-COGs, plus an additional 600 genome-specific genes. Although a web browser may be able to display all this information, it is recommended to use another form of access that can more readily handle large volumes of data.

All of the information in the ATGC database is also accessible from the bulk downloads page containing raw data files, e.g. a user can view the list of ATGCs in an external program of their choice such as Microsoft Excel, the Unix 'grep' command, or a computer programming language. Both the Iowa and NCBI sites provide completely free and anonymous FTP access with no need for a password or to create a login account.

DATABASE CONSTRUCTION

The automated protocol to construct ATGCs includes three major steps. Prior to these, all RefSeq complete genome assemblies were downloaded and pre-processed (concatenate genomic partitions, extract nucleotide sequence of protein-coding genes, format per-genome BLAST databases and several additional steps), followed by three rounds of progressive genome clustering. Finally, a post-processing step calculates data derived from the resulting ATGCs, such as ATGC-COGs and matrices of pairwise genomic distances.

Step 1: initial tree clustering

Starting with all complete genome assemblies in RefSeq (currently, nearly 5000 entries), the data set is first divided into clusters of more manageable size. Initial clusters are constructed by collecting all RefSeq complete genomic assemblies belonging to genera that appear in a phylogenetic tree below a depth threshold of <7% root-to-tip distance. This threshold was arbitrarily chosen to produce clusters slightly broader than needed for ATGC construction so as to avoid many unnecessary pairwise comparisons in the later refinement steps. To obtain the phylogenetic tree, we used the pre-computed species tree of life (sTOL) constructed from the data in the SUPFAM database, which is updated on a daily basis (27). During this process, clusters are allowed to merge in their entirety (but not partially) if the maximum distance to their leaves is below the threshold, which allows for some non-monophyletic groups to be formed. For example, a branch of fast-evolving parasites can form a separate ATGC from one containing several clades of free-living microbes that also descended from their common ancestor.

The result of this initial tree clustering step is a set of complete genome assemblies of bacteria and archaea that are more closely related than the Family level but broader than the Genus level. These initial clusters are then refined by the subsequent steps.

Step 2: cluster refinement by synteny

The initial tree clusters that are represented by 2 or more completely-sequenced genomes from RefSeq are next subject to refinement by synteny. Putative orthologous genes conserved between each pair of genomes are identified by BLAST (28) Bidirectional Best Hits (BBH), and considered

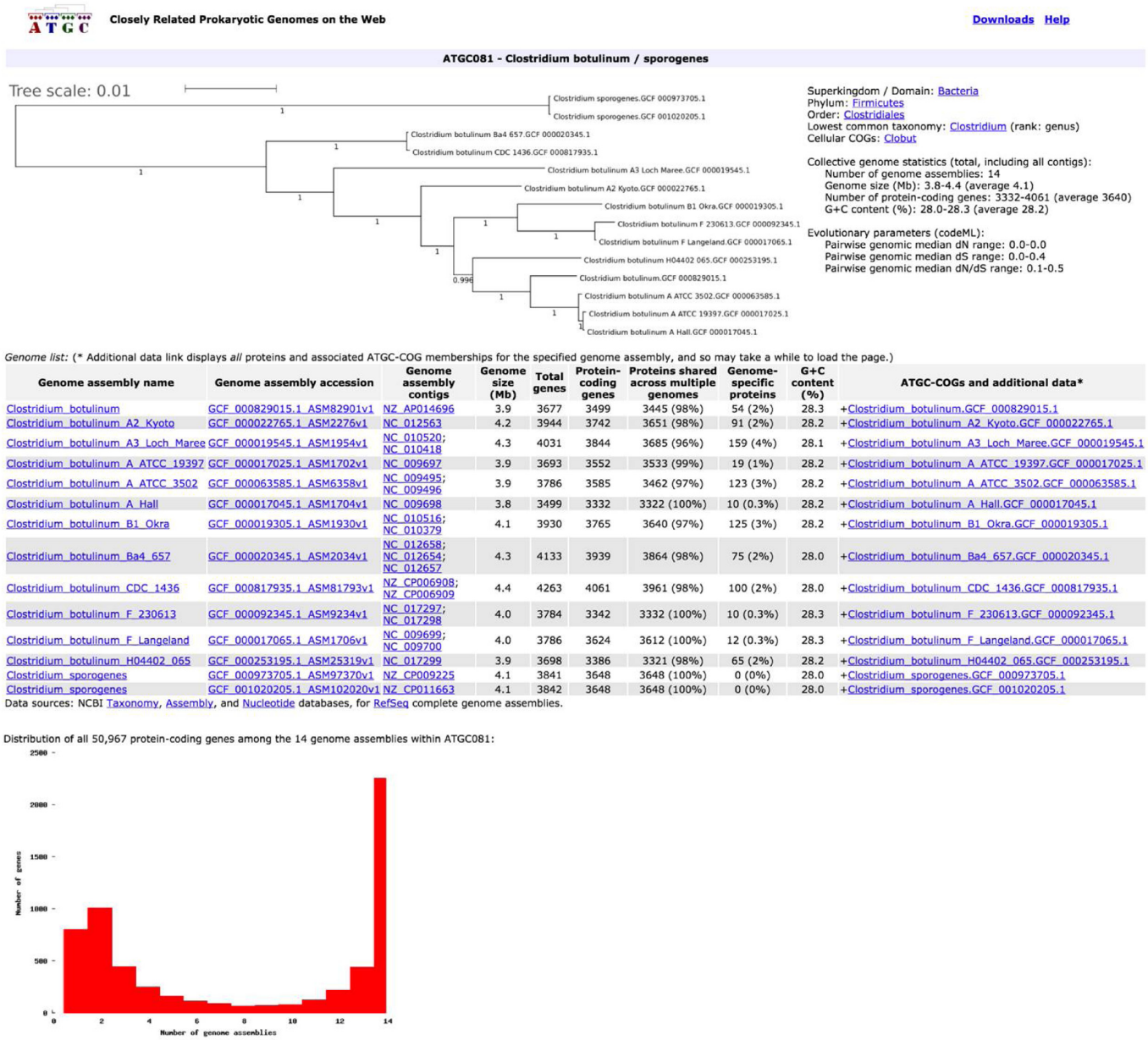


Figure 3. An individual ATGC webpage. The page provides descriptive information about the ATGC, as well as about each genomic assembly in the group and the pangenomic distribution of genes shared across multiple genomes. Clicking the name of an individual genome assembly (right-most column of the table) takes the user to a page containing all of the proteins in that genome and additional data for each, such as the ATGC-COG membership, genomic coordinates, various known gene names and symbols and other information.

supported by synteny if the majority of genes surrounding the query also appear in the target genome within a fixed window of 7 genes (initial BBH + 3 upstream + 3 downstream). Mathematically, a 'rearrangement distance' is calculated as DY = (Nb−Ns)/Nb, where Ns is the number of BBHs supported by synteny and Nb is the total number of BBHs, and single-linking clustering of genomes is performed with DY $\leq$ 0.15. This cutoff is based on manual inspection of the results, corresponding to 85% of putative orthologs having conserved local gene order (1). This criterion allows for small-scale (sub-operon level) rearrangements of gene order, whereas larger-scale rearrangements serve as an evolutionary distance measure (29). The use of single-linkage clustering allows for more permissive inclusion of genomes with large-scale rearrangements with DY $>$ 0.15, so long as they are alignable with DY $\leq$ 0.15 to at least one other genome in the same ATGC.

The order of cluster refinement does not affect the final outcome given that the resulting ATGCs will eventually have to meet the requirements of both synteny and synonymous substitution rate (see below). However, the examination of local gene order in step 2 synteny refinement is a much less computationally expensive process than the Maximum Likelihood calculations of synonymous substi-

tution rates, and so performing this step first helps reduce the overall number of required pairwise computations. In addition, the use of synteny-supported BBHs provides a more accurate estimation of putative orthologs (until the full ATGC-COG membership will be calculated in the final post-processing step).

Step 3: cluster refinement by synonymous substitution rate

Synteny-supported clusters are further refined using the rate of synonymous substitutions (dS = number of synonymous substitutions per synonymous site) in the synteny-supported BBHs. For each of the latter, the protein sequences are aligned with MUSCLE, the coding nucleotide sequence are threaded onto this alignment, and dS is estimated by a Maximum-Likelihood approach using the CODEML program of the PAML package (30,31). Clusters are refined by splitting groups until those remaining contain only genomes with median dS over all synteny-supported BBHs $\leq$1.5 (implemented by using an ultrametric tree constructed with the KITSCH program of the Phylip package (32). This cutoff was chosen to obtain clusters that present a substantial amount of evolutionary divergence while keeping the rate below saturation for the majority of orthologous genes.

Step 4: pre-calculated data available for each ATGC

Upon the completion of the cluster refinement steps, with the resulting ATGCs (Clusters of complete Genomes now meeting the criteria of Alignability (DY$\leq$0.15) and Tightness (dS $\leq$ 1.5)), a substantial amount of data was calculated and made available to the user.

ATGC-COGs. The original COGs method was designed to find orthologs among large numbers of distantly-related organisms (4). In contrast, the ATGCs often contain far fewer genomes that are closely related and thus contain more in-paralogs that can be difficult to distinguish from one another due to high sequence identity, the method was adapted for closely related organisms. First, 'seed' COGs were built with conservative parameters using the standard COGtriangles algorithm (3). Next, using an updated version of the COGNITOR algorithm (33) (PSI-COGNITOR), these seed clusters were extended with PSI-BLAST profiles (28), adding new members to the best-matched seed COG. Each protein in a given genome should in theory belong to a single ATGC-COG, so the cases where this did not appear to be the case (due to domain rearrangement, unresolved paralogy and other anomalies) were resolved by assigning the protein to the best-matched COG (such cases were reported as warnings to alert the user of their occurrence).

Index orthologs. To facilitate the use of ATGCs for genome evolution studies, a set of 'index' orthologs representing the most well-conserved members of each ATGC-COG in each genome were calculated (34). For each genome, a single protein was chosen to be the best representative of each ATGC-COG by virtue of being most similar to the other members in the group, first in terms of synteny support (see below), with ties broken by choosing the member with the lowest mean dS.

Multiple sequence alignments. The protein sequences from each ATGC-COG were aligned with MUSCLE, and the coding nucleotide sequences were again threaded onto this alignment. This procedure is similar to the pairwise alignments performed in Step 3, but at this stage representing multiple alignments of the entire ATGC-COG. Index orthologs were separately aligned with MUSCLE, and coding nucleotide sequences once again threaded onto this alignment.

Synteny conservation. A quantitative measure of synteny conservation was calculated for each ATGC-COG and for each gene within an ATGC-COG. First, a collective gene neighborhood representation was assembled by counting the number of ATGC-COGs appearing in the set of genomes in the given ATGC and present within a local gene neighborhood window of six genes surrounding the query ATGC-COG (three upstream and three downstream) in each genome. Next, the synteny conservation score of each gene in each ATGC-COG was measured as follows:

$$S_{\mathrm{p}} = \frac{\sum_{\mathrm{g}=1}^{G} \quad \sum_{\mathrm{w}=-3}^{+3} C_{\mathrm{g,w}}}{G * U}$$

where S_{p} is the synteny conservation score for protein p, $C_{\mathrm{g,w}}$ is the number of genes belonging to the same COG as p, across each genome g in that COG and across each local gene neighborhood window offset w, G is the total number of genomes in the ATGC and U is the total number of unique COGs appearing in the collective neighborhood representation, across all genomes and across the neighborhood window of six genes in each. Thus, a protein whose local gene neighborhood perfectly matches the respective gene neighborhoods of all other members of the ATGC-COG would be assigned a perfect score of 1. Negative deviations from that value reflect a less well-conserved neighborhood (by increasing U), whereas less common positive deviations can occur via tandem duplications (by decreasing U below 6). Next, after S_{p} was used to determine index orthologs, the collective gene neighborhood was re-assembled to include only index orthologs and the synteny conservation of index orthologs was re-calculated with respect to this new, better conserved neighborhood. This procedure prevents disruption of a high level of local synteny conservation in situations such as horizontal gene transfer that increase the copy number of a gene that already has an existing member in the genome. Then, rather than discard the scores for non-index orthologs, these are kept so as to provide the level of similarity to the overall neighborhood surrounding all copies, including the new and pre-existing copy. Finally, the overall synteny conservation score of each ATGC-COG S_{c} was also calculated with respect to the index orthologs as the mean of the index ortholog synteny conservation scores S_{p} over all of its member genomes G (for which there are exactly one protein per genome).

$$S_{\mathrm{c}} = \sum_{\mathrm{g}\,=\,1}^{G} S_{\mathrm{p}} / G$$

Phylogenetic tree. A high-resolution phylogenetic tree was constructed from the conserved gene core of each ATGC. This set was constrained to include only ATGC-COGs that met the following criteria: (i) universal conservation, i.e. represented in every organism within the given ATGC; (ii) present only in a single copy in each organism, with no families that contain paralogs allowed; (iii) a high level of synteny support, with the threshold of ≥75% chosen by manual inspection; (iv) complete coding sequence readily available for all members, i.e. no ribosomal frame-shifts or gene disruption for any other reasons. Despite the strictness of these criteria, the resulting core gene sets typically include thousands of ATGC-COGs. On average, 2200 ATGC-COGs meet the criteria, varying from a minimum of 300 genes in the small (∼1 Mb) genomes of *Rickettsia*, up to >9000 in the large genomes of *Amycolatopsis mediterranei* (>10 MB). The multiple sequence alignments of the core ATGC-COGs were then concatenated into a single nucleotide coding sequence alignment, and the FastTree program with a generalized time-reversible model used to build the phylogenetic tree (35) (to our knowledge, FastTree is the only phylogenetic tree construction program that can handle this amount of data, at virtually no loss in accuracy compared to traditional maximum-likelihood methods such as RAxML (36). The resulting tree was then rerooted and ladderized, with the final graphic made by the iTOL website (37).

Matrices of pairwise intergenomic distances. In addition to the phylogenetic tree, another measure of intergenomic distance between members of an ATGC is made available in the form of matrices. These include the synonymous (*dS*) and non-synonymous (*dN*) substitution rates of the genomic median value among all pairs of genes calculated in Step 3 with CODEML (38).

CHANGES FROM PREVIOUS ATGCS

Compared to the previous dedicated publication on the ATGC database (1), the coverage of genomes in RefSeq has increased over 8-fold (from 446 to >3700 genomes), and many other substantial changes have occurred. Whereas the prior release was built mostly using manual procedures, all steps are completely automated in the current release, which will help keep the resource up to date in the future. The content of ATGCs has also changed slightly to no longer include incomplete genomes, which increases the quality of the data set. Within each genome, plasmids, transposons and pseudogenes are no longer manually removed from the data set. Length mismatches between the nucleotide and protein sequences of a gene (mostly ribosomal slippage events in transposases) are now automatically removed. Unusual start or stop codons are reported but the genes they occur in are not automatically removed, leaving this quality-control decision up to the end-user.

Major changes were made to each individual step of ATGC construction although each affected only the computational cost rather than the final outcome. For instance, rather than incurring the expense of building a high-quality phylogenetic tree from single-copy COGs, the daily-updated sTOL tree was used instead (down to the genus level). Also, initial tree clusters are no longer required to be monophyletic which allows splitting clusters much earlier on in the pipeline. The ordering of cluster refinement steps was altered: in the new pipeline, synteny refinement precedes the *dS* refinement because the former step is much less computationally expensive and reduces the amount of computations performed during the latter step. Despite these efforts at optimizations, the computational requirements of constructing the entire database were enormous and so parallelization was implemented wherever possible.

One important difference between the current and the previous releases of the ATGCs is the implementation of ATGC-COGs using the more robust COGs approach of 3-way BBHs (39). This procedure is less prone to many types of errors that single-linkage clustering of synteny-supported BBHs are subject to (variations in gene copy number, displacement of an existing gene with a foreign copy, etc.). The introduction of PSI-COGNITOR also increases the sensitivity of this approach, and allows complete accounting for every protein with respect to the conservation within the ATGC.

Several major differences made by RefSeq since the last release of the ATGCs database prompted corresponding changes to the current ATGCs. For instance, due to the lack of scalability of manual curation processes, the source of manually-curated names of bacterial organisms is no longer available, and so automated names are used instead for the >3700 complete genome assemblies (which themselves are only a small fraction of the 55 000 total entries in RefSeq release 71 (23)). Another change in RefSeq involves the use of multispecies accession codes, which combined with the retirement of the GI number system necessitated the creation of unique protein IDs for the ATGCs that were generated automatically.

Because the JavaScript graphical interface to the ATGC webpage used in the previous release prevented access to large-scale data downloads, the current webpage was redesigned as a set of simple HTML pages. Text files are also provided that are suitable for working with in Microsoft Excel, OpenOffice, parsing with a programmatic language and other purposes. These files are accessible via both an individual ATGC page and the bulk download page. The input data used to build the ATGCs—raw genome assembly data and pairwise genome comparison values for dY, dN and dS—are also provided in the bulk downloads page.

AVAILABILITY

The new version of the ATGC database, including ATGC-COGs and other pre-computed data, is available at the University of Iowa at http://dmk-brain.ecn.uiowa.edu/ATGC/ and ftp://dmk-brain.ecn.uiowa.edu/ATGC/atgc_home.html, and at the NCBI at ftp://ftp.ncbi.nlm.nih.gov/pub/kristensen/ATGC/atgc_home.html. Both locations will be updated as changes are made in the future, although the FTP sites will host only simple webpages devoid of active content such as JavaScript and/or CGI while the HTTP server will have active content added as it is developed. All queries and comments regarding the ATGC database should be directed to DMK.

FUTURE DEVELOPMENTS

In response to changes by RefSeq, some genomic assemblies that were previously classified as 'complete genomes' are no longer recognized as such. As a result, some organisms that formerly formed an ATGC no longer have sufficient sampling density to qualify for ATGC membership. These are currently denoted on the ATGC webpage as 'missing' entries, and the plan is to manually curate at least these relevant genomic records to salvage the missing ATGCs.

In the automated construction pipeline for ATGC construction, parallelization was used whenever possible. Nevertheless, at several points in the pipeline bottlenecks exist that proved challenging for thousands of proteins. In the future these bottlenecks may not only cause the pipeline to run significantly slower, but could cause it to cease functioning entirely. Notable examples include alignment of a single protein family with MUSCLE, and building COG families in each ATGC. These processes proved intractable on a machine with 50 GB of memory, and succeeded only after several weeks of calculations on a machine with a Terabyte of memory. In the future, it will be necessary to replace these tools with ones capable of handling even tens of thousands of genes (40,41). Unless another major change in RefSeq necessitates complete re-building of all ATGCs, an iterative approach can be used to assign new members to existing ATGCs, leaving a much smaller data set of genomes for which the full ATGC construction pipeline would need to be run. Similarly, new orthologs can be added to existing ATGC-COG gene families without having to run the entire all-against-all protein comparison for several million proteins.

FUNDING

Intramural Research Program of the U.S. National Institutes of Health at the National Library of Medicine (to D.M.K. and E.V.K.); Department of Biomedical Engineering at the University of Iowa (Iowa City, USA) (to D.M.K.). Funding for open access charge: NIH.
Conflict of interest statement. None declared.

REFERENCES

1. Novichkov,P.S., Ratnere,I., Wolf,Y.I., Koonin,E.V. and Dubchak,I. (2009) ATGC: A database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. *Nucleic Acids Res.*, **37**, D448–D454.
2. Novichkov,P.S., Wolf,Y.I., Dubchak,I. and Koonin,E.V. (2009) Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *J. Bacteriol.*, **191**, 65–73.
3. Kristensen,D.M., Kannan,L., Coleman,M.K., Wolf,Y.I., Sorokin,A., Koonin,E.V. and Mushegian,A. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, **26**, 1481–1487.
4. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
5. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
6. Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.
7. Chen,F., Mackey,A.J., Vermunt,J.K. and Roos,D.S. (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, **2**, e383.
8. Hulsen,T., Huynen,M.A., de Vlieg,J. and Groenen,P.M. (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.*, **7**, R31.
9. Chaudhari,N.M., Gupta,V.K. and Dutta,C. (2016) BPGA- an ultra-fast pan-genome analysis pipeline. *Sci. Rep.*, **6**, 24373.
10. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
11. Puigbo,P., Lobkovsky,A.E., Kristensen,D.M., Wolf,Y.I. and Koonin,E.V. (2014) Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.*, **12**, 66.
12. Takeuchi,N., Wolf,Y.I., Makarova,K.S. and Koonin,E.V. (2012) Nature and intensity of selection pressure on CRISPR-associated genes. *J. Bacteriol.*, **194**, 1216–1225.
13. Makarova,K.S., Wolf,Y.I., Snir,S. and Koonin,E.V. (2011) Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.*, **193**, 6039–6056.
14. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2009) Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol. Direct*, **4**, 19.
15. Lobkovsky,A.E., Wolf,Y.I. and Koonin,E.V. (2016) Evolvability of an optimal recombination rate. *Genome Biol. Evol.*, **8**, 70–77.
16. Faure,G. and Koonin,E.V. (2015) 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.*, **12**, 035001.
17. Ran,W., Kristensen,D.M. and Koonin,E.V. (2014) Coupling between protein level selection and codon usage optimization in the evolution of bacteria and archaea. *MBio*, **5**, doi:10.1128/mBio.00956-14.
18. Busby,B., Kristensen,D.M. and Koonin,E.V. (2013) Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. *Environ. Microbiol.*, **15**, 307–312.
19. Gophna,U., Kristensen,D.M., Wolf,Y.I., Popa,O., Drevet,C. and Koonin,E.V. (2015) No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. *ISME J.*, **9**, 2021–2027.
20. Ish-Am,O., Kristensen,D.M. and Ruppin,E. (2015) Evolutionary Conservation of Bacterial Essential Metabolic Genes across All Bacterial Culture Media. *PLoS One*, **10**, e0123785.
21. Povolotskaya,I.S., Kondrashov,F.A., Ledda,A. and Vlasov,P.K. (2012) Stop codons in bacteria are not selectively equivalent. *Biol. Direct*, **7**, 30.
22. Povolotskaya,I.S. and Kondrashov,F.A. (2010) Sequence space and the ongoing expansion of the protein universe. *Nature*, **465**, 922–926.
23. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
24. Lan,R. and Reeves,P.R. (2002) Escherichia coli in disguise: Molecular origins of Shigella. *Microbes Infect.*, **4**, 1125–1132.
25. Urbanczyk,H., Ast,J.C., Higgins,M.J., Carson,J. and Dunlap,P.V. (2007) Reclassification of Vibrio fischeri, Vibrio logei, Vibrio salmonicida and Vibrio wodanis as Aliivibrio fischeri gen. nov., comb. nov., Aliivibrio logei comb. nov., Aliivibrio salmonicida comb. nov. and Aliivibrio wodanis comb. nov. *Int. J. Syst. Evol. Microbiol.*, **57**, 2823–2829.
26. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
27. Fang,H., Oates,M.E., Pethica,R.B., Greenwood,J.M., Sardar,A.J., Rackham,O.J., Donoghue,P.C., Stamatakis,A., de Lima Morais,D.A. and Gough,J. (2013) A daily-updated tree of (sequenced) life as a reference for genome research. *Sci. Rep.*, **3**, 2015.
28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
29. Koonin,E.V. and Wolf,Y.I. (2010) Constraints and plasticity in genome and molecular-phenome evolution. *Nat. Rev. Genet.*, **11**, 487–498.
30. Yang,Z. (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.

31. Yang,Z. and Bielawski,J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.*, **15**, 496–503.
32. Felsenstein,J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author*. Department of Genome Sciences, University of Washington, Seattle.
33. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
34. Jordan,I.K., Wolf,Y.I. and Koonin,E.V. (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.*, **4**, 22.
35. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
36. Liu,K., Linder,C.R. and Warnow,T. (2011) RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One*, **6**, e27731.
37. Letunic,I. and Bork,P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W245.
38. Wolf,Y.I., Snir,S. and Koonin,E.V. (2013) Stability along with extreme variability in core genome evolution. *Genome Biol. Evol.*, **5**, 1393–1402.
39. Kristensen,D.M., Wolf,Y.I., Mushegian,A.R. and Koonin,E.V. (2011) Computational methods for Gene Orthology inference. *Brief. Bioinform.*, **12**, 379–391.
40. Liu,K., Linder,C.R. and Warnow,T. (2010) Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLoS Curr.*, **2**, RRN1198.
41. Mirarab,S., Nguyen,N., Guo,S., Wang,L.S., Kim,J. and Warnow,T. (2015) PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J. Comput. Biol.*, **22**, 377–386.