Towards automatic reconstruction of a highly resolved tree of life.

Francesca D. Ciccarelli[1,2,3*], Tobias Doerks[1*], Christian von Mering[1], Christopher J. Creevey[1], Berend Snel[4] and Peer Bork[1,5].

* These authors contributed equally to the work.

1 European Molecular Biology Laboratory, Meyerhofstr. 1, 69012 Heidelberg, Germany.

2 Department of Experimental Oncology, European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy.

3 FIRC Institute of Molecular Oncology Foundation, Via Adamello 16, 20139 Milan, Italy.

4 Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Center, Toernooiveld 1, 6525 ED, Nijmegen, The Netherlands.

5 Max-Delbrück-Centrum, Berlin-Buch, Germany.

Corresponding author:

Peer Bork
EMBL
Meyerhofstr.1
69012 Heidelberg (Germany)
E-mail: peer.bork@embl.de
Tel. +49 6221 387 526
Fax +49 6221 387 517

Manuscript information: Text pages: 15; Figures: 3 (submitted as separate files); Table: 1.

Word and character counts: Words in the abstract: 99; Words in the text (comprising references, figure and table captions): 2592.

Abstract

We have developed an automatable procedure for reconstructing the tree of life with branch lengths comparable across all three domains on the basis of a concatenation of 31 orthologs occurring in 191 species with sequenced genomes. The tree revealed inter-domain discrepancies in taxonomic classification. Systematic detection and subsequent exclusion of products of horizontal gene transfer increased phylogenetic resolution, allowing us to confirm accepted relationships and resolve disputed and preliminary classifications. For example, we place the previously defined Acidobacteria phylum as a sister group of δ-proteobacteria, support a gram-positive origin of Bacteria, and suggest a thermophilic last universal common ancestor.

Reconstructing the phylogenetic relationships amongst all living organisms is one of the fundamental challenges in biology. Numerous attempts to derive a tree of life through various methods have been published (for a review see (1)), and its principal existence has been questioned recently (2, 3). Moreover, even under the assumption of a tree of life, numerous groupings and taxonomic entities still remain heavily debated and the advent of molecular and genomic data has increased the variety of classifications rather than reducing the problem (1). Theoretical and practical limits to the reconstruction of a tree of life have been put forward, such as the insufficient amount of discriminating characters available, even in information-rich genomic datasets (4), and the computing resources required to cope with large numbers of species (1). Furthermore, there are factors that hamper the accurate reconstruction of phylogenetic trees regardless of the methods used, such as the sampling bias of the species included (5) and the dilution of phylogenetic signal by horizontal gene transfer (HGT) (6), the extent of which is still extremely controversial (2, 3, 7). In addition to these difficulties, different datasets have been used with a variety of methods and parameter settings, making it almost impossible to quantitatively compare the proposed results. Hence, there exists the challenge and requirement for a reproducible and updatable pipeline to reconstruct the tree of life by means of a commonly available dataset such as completely sequenced genomes. Here, we demonstrate the feasibility of the tree construction and present a phylogeny based on an alignment of sufficient length and resolution to accurately calculate comparable branch lengths across all three domains of life. We have created for this purpose a supermatrix of 31 concatenated, universally occurring genes with indisputable orthology in 191 species with completely annotated genomes (Fig. 1, Table S1). Although the initial identification

and analysis of these genes required considerable manual effort (8), the inclusion of additional species with completely annotated genomes has pipeline character (Fig. 1). As the 31 universal genes are all involved in translation, we applied the same tree-building procedure to independent sets of domain-specific non-translational genes (8).

For the tree reconstruction we mostly used standard approaches (Fig. 1), with the exception of a procedure for the detection and selective exclusion of HGTs, which turned out to be essential for obtaining a highly resolved tree. We started with 36 genes universally present in all 191 species for which orthologs could be unambiguously identified (8) and eliminated five of them from the analysis (mostly t-RNA synthetases) because they have undergone multiple horizontal transfers and/or were difficult to align (Fig.1, Table S1). Although the 31 remaining genes are unlikely to be subjected to lateral transfers as they mainly encode for ribosomal proteins (9), we systematically tested them for any HGT event not yet identified. We randomly allocated the 31 gene products into 4 groups and for each group we derived the corresponding subsets of trees where each protein was in turn missing from the alignment (resampling with displacement). We subsequently checked for topological incongruence within each subset of trees and further tested candidate HGTs by two other independent measures (8). If at least one of these two measures could confirm the jack-knife indication, the gene was considered horizontally transferred and removed from the corresponding alignment. (Fig. 1, Table S2).

Our approach (confirmed by single tree analysis (8)) detected a total of 7 HGT candidates (i.e. orthologous gene displacements (10)) among 31 orthologs from 191 species, with some species being involved in more than one HGT event (Table S2). Three

out of the four aminoacyl-tRNA synthetases (aa-RSs) used in this analysis have undergone HGT, including Valyl-RS (COG0525), which had been reported before (11) thus confirming the mobility of these enzymes (12). Clostridia is the only class that acquired ribosomal proteins by lateral transfer, likely in a single ancient event, as the displaced orthologs are present in all sequenced Clostridia (Table S2). To our knowledge, only one other horizontal transfer of ribosomal proteins has been reported so far (13). However, the discovery of seven HGTs in the 31 translation-related genes compares favourably with 30 (ten per domain) lateral transfers detected in domain-specific trees from 24 non-translational genes (8).

The species-specific exclusion of HGTs and concatenation of all gene product alignments resulted in a supermatrix of 8090 positions for 191 species. This supermatrix was subsequently used to reconstruct the tree of life shown in Fig. 2 (for a detailed version, see http://www.bork.embl.de/tree_of_life/).

The global tree topology was supported by two independent measurements: Firstly, using domain-specific subtrees from non-translational genes we could confirm the monophyly of all major divisions and reproduce most of their branching orders (8), albeit with weaker statistical support. This is due to lower sequence coverage and/or conservation as well as a higher number of excluded characters because of the higher incidence of HGTs (8). Secondly, three independent tests carried out on the individual gene trees revealed that although they are not identical, they share similarities with both the obtained tree of life and with each other (8). While it may be possible to reject the null hypothesis of each of these tests without much difficulty, their combined evidence suggests that the gene trees have a cohesive phylogenetic signal.

Within the tree of life as many as 65% of the branches are supported by a bootstrap proportion (BP) of 100%, and 81.7% have more than 80% BP support, enabling us to propose resolutions to debated classifications at both the root and the tips of the tree (Table 1). Even though in Prokaryotes the statistical support of the early branches is generally weaker than that of the recent ones, it is noteworthy that within the Bacteria, the Firmicutes appear to comprise the earliest branching phylum, in agreement with a proposed gram-positive ancestor for all Bacteria (14) (Fig. 2, Table 1). In our tree the Firmicutes are placed at the earliest division of Eubacteria, with 66% BP support and 33% of remaining BP show at least a subclade of Firmicutes at the earliest division. This placement and the fact that the 15 slowest-evolving taxa of the Bacteria are all gram-positive (8) support the theory of a gram-positive origin of Bacteria. Furthermore, the thermophilic Firmicute *Thermoanaerobacter tengcongensis* is the taxon with the shortest overall phylogenetic distance to the root of Bacteria (Fig. 2), and as such is most likely to have retained ancestral states (15). Together with the fact that slowest-evolving, ancestral Archaea (Table S7) are also (hyper)thermophilic (8), this lends support to the hypothesis that the last universal common ancestor (LUCA) was living at high temperatures.

At the base of the Proteobacteria, the monophyletic Acidobacteria appear as a sister-group to the δ-proteobacteria (Fig. 2). The 64% BP support for this relationship, indicates that the Acidobacteria may be a sixth divergent class within Proteobacteria. The monophyly of the Proteobacterial/Acidobacterial clade is supported with a BP of 98%, further raising the question whether Acidobacteria should indeed be an independent phylum (16).

Towards the tips of the tree, within the Cyanobacteria, *Synechococcus* (sp.WH8102) groups with the Prochloroccales, and *Nostoc* groups with *Synechocystis*, a result that has been supported by some rRNA studies (17) and challenges the classical order Chroococcales (firstly defined in 1849 and based upon morphological features (18)).

Within the Archaea, the position of Nanoarchaeota remains debated (see e.g. (19)). We find (with 100% BP support) that they are a sister group of Crenarchaeota, without an indication of reported HGTs from Crenarchaeota (19) in all the core genes studied.

Within the Eukaryotes, our tree gives clear support for the classical Coelomata hypothesis that groups the Arthropods with Deuterostomia (chordates) in a monophyletic clade. This is in contrast to the "new animal phylogeny" that groups nematodes and arthropods into the monophyletic Ecdysozoa (20, 21). The Ecdysozoan clade has been supported by SSU-rRNA and single-gene phylogenies ((22) and references therein) but has been rejected by a number of recent studies based on genomics features and whole genome phylogenies ((23) and references therein). The current sampling bias and accelerated evolution of sequenced representatives of certain Metazoan lineages (e.g. arthropods and nematodes; see Fig. 2) may factor in these results. This highlights the need for the sequencing of slow evolving species (15), which may resolve such controversies in the tree.

Despite a highly resolved and robust tree, we cannot exclude a few uncertainties in tree topology due to biased species sampling or long branch attraction (LBA) (24). For example, the close grouping of Thermotoga and Aquifex in our and other trees might be

partially caused by their common thermophilic lifestyle (25), while LBA might account for the placement of parabasalia (*G.lamblia*) as the most basal eukaryal taxon (Table 1).

The use of a common protein set across all three domains of life also ensures that the observed branch lengths are comparable across the entire tree. This enables, for example, an objective, quantitative analysis of the consistency of traditional taxonomic groupings (Fig. 3). As expected, the hierarchy of taxonomic groups correlates with the phylogenetic diversity measured between and within them (e.g. species belonging to the same family have a shorter branch length distance than species belonging only to the same phylum). Within each taxonomic level, the branch lengths distances vary considerably, apparently owing to factors that influence substitution rates such as differences in lifestyle or population size. However, even when taking this effect into account, we observe a strong discrepancy between taxonomic divisions in Eukaryotes and Prokaryotes (Fig. 3A). Organisms that have been assigned to separate phyla in Eukaryotes would clearly belong to the same phylum in the prokaryotic classification. Historically, Eukaryotes have obviously been given more taxonomic resolution than Prokaryotes – a testament to their greater morphological diversity.

Another universal trend is that smaller genomes evolve faster (i.e. have longer branch lengths, Fig. 3B). This is easily explained for pathogenic or endosymbiontic organisms with reduced genomes, which often have only limited capabilities to remove mutations by means of recombination or DNA repair (26). However, we observe this trend also for genomes of larger sizes, including free-living Prokaryotes, and Eukaryotes. Intriguingly, there is not a single organism sequenced that is both fast evolving and has a large genome (Fig. 3B). This suggests that the coupled processes of genome reduction

and evolutionary speedup may be irreversible: genomes apparently do not grow again after a prolonged phase of genome reduction.

The pan-domain phylogeny that resulted from the procedure presented here will increase in resolution with more species being sequenced. This updatable reference phylogeny of completely sequenced species allows accurate comparisons of branch lengths across domains. The resulting tree of life will be an invaluable tool in many areas of biological research ranging from classical taxonomy, via studies on the rate of evolution, to environmental genomics where DNA fragments of unknown phylogenetic origin need to be assigned.

Figure 1. Overview of the procedure.

The white boxes represent the major steps in the process of building the pan-domain phylogeny presented here. The steps in grey represent automatable parts of the procedure that need to be carried out for the inclusion of further species. For the 31 COGs used in the analysis, we manually derived 1:1 orthologs by removing mitochondrial and chloroplast paralogs from the corresponding multiple alignments. We built domain-specific alignments using the corresponding proteins encoded by the 31 orthologs and aligned the resulting profiles. With this procedure we maximized the number of positions of the global alignment and reduced the number of misaligned residues. For a detailed description of the methods see (8). COG = Cluster of Orthologous Groups; HGT = Horizontal Gene Transfer; MSA = Multiple Sequence Alignment.

Figure 2. Global phylogeny of fully sequenced organisms.

The phylogenetic tree is based on a cleaned and concatenated alignment of 31 universal protein families, and covers 191 species whose genomes have been fully sequenced. A detailed version of this tree, including bootstrap support values and branch lengths, is available at http://www.bork.embl.de/tree_of_life/. Green section: Archaea, red: Eukaryotes, blue: Bacteria. Labels and color shadings indicate various frequently used subdivisions. The branch separating Eukaryotes and Archaea from Bacteria in this unrooted tree has been shortened for display purposes.

Figure 3. Global analysis of branch length information.

A) Average sequence divergence within taxonomic classification units. Each data point denotes a pairwise comparison of two taxa, relating their inter-taxa branch length distance (i.e. sequence divergence) with their level of relatedness according to the NCBI taxonomy ('taxonomic distance'). Horizontal bars denote 95% intervals and medians of the data. Some minor taxonomy hierarchy levels have been omitted. Marked items: (a) *Homo sapiens* vs. *Pan troglodytes*. The sequence divergence between human and chimp is low; they most likely would have been assigned the same genus if they had been Prokaryotes (see also (27) for a proposed revision). (b) *Synechococcus* (sp. WH8102) vs. *Prochlorococcus marinus 9313*. The two species are annotated as distinct orders but they appear quite closely related, challenging the classical order of Chroococcales (see text).

B) Evolutionary speed and genome size. For each taxon, the cumulative branch length from the tip to the root is plotted against the genome size (measured here as number of genes).

Table 1. Noteworthy Topological Features Of The Tree Of Life.

Domain	Topological Feature	BP (%)
Eukaryota	Coelomata hypothesis	100
	Amoebozoa related to Opistokonta	41
	Deep branching of Parabasalia	100
Eubacteria	**Relationships within phyla**	
	Separation between β and γ-proteobacteria	100
	Disruption of Chroococcales monophyly	100
	Disruption of Actinomycetales monophyly	100
	Acidobacteria/Proteobacteria clade	98
	Cluster of *F.succinogenes* next to the Chlorobium/Bacterioidales (Sphingobacteria hypothesis)	62
	Cluster of *F.nucleatum* with hyperthermophylic Bacteria	36
	Relationships between phyla	
	Grouping of Chlamydiae, Spirochetes, Actinobacteria and Bacteriodales/Chlorobi	67
	Grouping of Cyanobacteria, hyperthermophylic and Deinococcales/Chloroflexi	51
	Relationships between super-phyla	
	Grouping of Proteobacteria with Cyanobacteria, hyperthermophylic and Deinococcales/Chloroflexi	74
	Deep branching of Firmicutes	66
Archaeabacteria	**Relationships within phyla**	
	A. fulgidus with halobacterium and methanosarcina	99
	Relationships between phyla	
	Nanoarchaea as a sister branch of Crenarchaea	100

Legend to Table 1:

Selected features of the phylogeny that are novel, debated or difficult to reproduce according to current literature. An extended version of the table is available as Table S6. In the case of Firmicutes as the earliest branching bacterial phylum, it is noteworthing that the remaining 33% of the bootstrap proportion show at least a subclade of the Firmicutes at the earliest division.

References and Notes

1. F. Delsuc *et al.*, *Nat. Rev. Genet.* **6**, 361 (2005).
2. W. F. Doolittle, *Science* **284**, 2124 (1999).
3. J. P. Gogarten *et al.*, *Mol. Biol. Evol.* **19**, 2226 (2002).
4. J. R. Brown *et al.*, *Nat. Genet.* (2001).
5. D. M. Hillis *et al.*, *Syst. Biol.* **52**, 124 (2003).
6. H. Philippe, C. J. Douady, *Curr. Opin. Microbiol.* **6**, 498 (2003).
7. V. Daubin *et al.*, *Science* **301**, 829 (2003).
8. Supporting Online Material is available on Science online.
9. R. Jain *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3801 (1999).
10. E. V. Koonin *et al.*, *Trends Genet.* **12**, 334 (1996).
11. C. G. Kurland, S. G. E. Andersson, *Microbiol. Mol. Biol. Rev.* **64**, 786 (2000).
12. Y. I. Wolf *et al.*, *Genome Res.* **9**, 689 (1999).
13. C. Brochier *et al.*, *Trends Genet.* **16**, 529 (2000).
14. A. L. Koch, *Trends Microbiol.* **11**, 166 (2003).
15. F. Raible *et al.*, *Science* **310**, 1325 (2005).
16. P. Hugenholtz *et al.*, *J. Bacteriol.* **180**, 4765 (1998).
17. D. Honda *et al.*, *J. Mol. Evol.* **48**, 723 (1999).
18. Nägeli, *Allg. Schweiz. Ges. Gesammten Naturwiss.* **10**, 45 (1849).
19. C. Brochier *et al.*, *Genome Biology* **6**, R42 (2005).
20. A. Adoutte *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 4453 (2000).
21. K. M. Halanych, *Annu. Rev. Ecol. Evol. Syst.* **35**, 229 (2004).
22. H. Philippe *et al.*, *Mol. Biol. Evol.* (2005).
23. G. K. Philip *et al.*, *Mol. Biol. Evol.* **22**, 1175 (2005).
24. J. Felsenstein, *Syst. Zool.* **27**, 401 (1978).
25. D. P. Kreil, C. A. Ouzounis, *Nucl. Acids Res.* **29**, 1608 (2001).
26. N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 2873 (1996).
27. D. E. Wildman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7181 (2003).

28. **Acknowledgements:** We thank Sean Hooper for help in designing trees and are grateful to members of the Bork group as well as Toby Gibson, Manolo Gouy, Martijn Huynen, Aiden Budd and Matthias Wolf for stimulating discussions. This work was supported in part by BMBF (NGFN grant 01GR0454 to PB) and by the NWO.

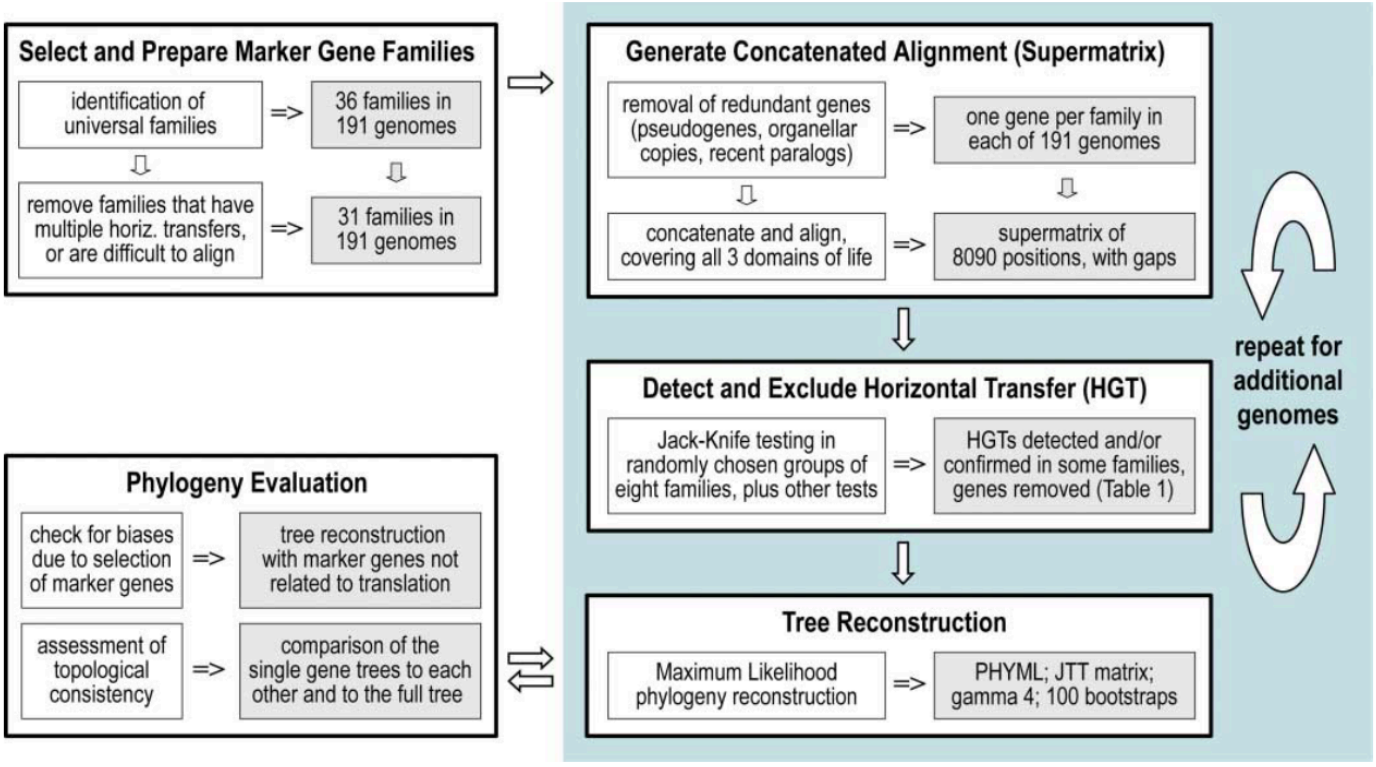

Select and Prepare Marker Gene Families
identification of universal families
=>
36 families in 191 genomes
remove families that have multiple horiz. transfers, or are difficult to align
=>
31 families in 191 genomes
Generate Concatenated Alignment (Supermatrix)
removal of redundant genes (pseudogenes, organellar copies, recent paralogs)
=>
one gene per family in each of 191 genomes
concatenate and align, covering all 3 domains of life
=>
supermatrix of 8090 positions, with gaps
Detect and Exclude Horizontal Transfer (HGT)
Jack-Knife testing in randomly chosen groups of eight families, plus other tests
=>
HGTs detected and/or confirmed in some families, genes removed (Table 1)
Tree Reconstruction
Maximum Likelihood phylogeny reconstruction
=>
PHYML; JTT matrix; gamma 4; 100 bootstraps
repeat for additional genomes
Phylogeny Evaluation
check for biases due to selection of marker genes
=>
tree reconstruction with marker genes not related to translation
assessment of topological consistency
=>
comparison of the single gene trees to each other and to the full tree

Figure 1

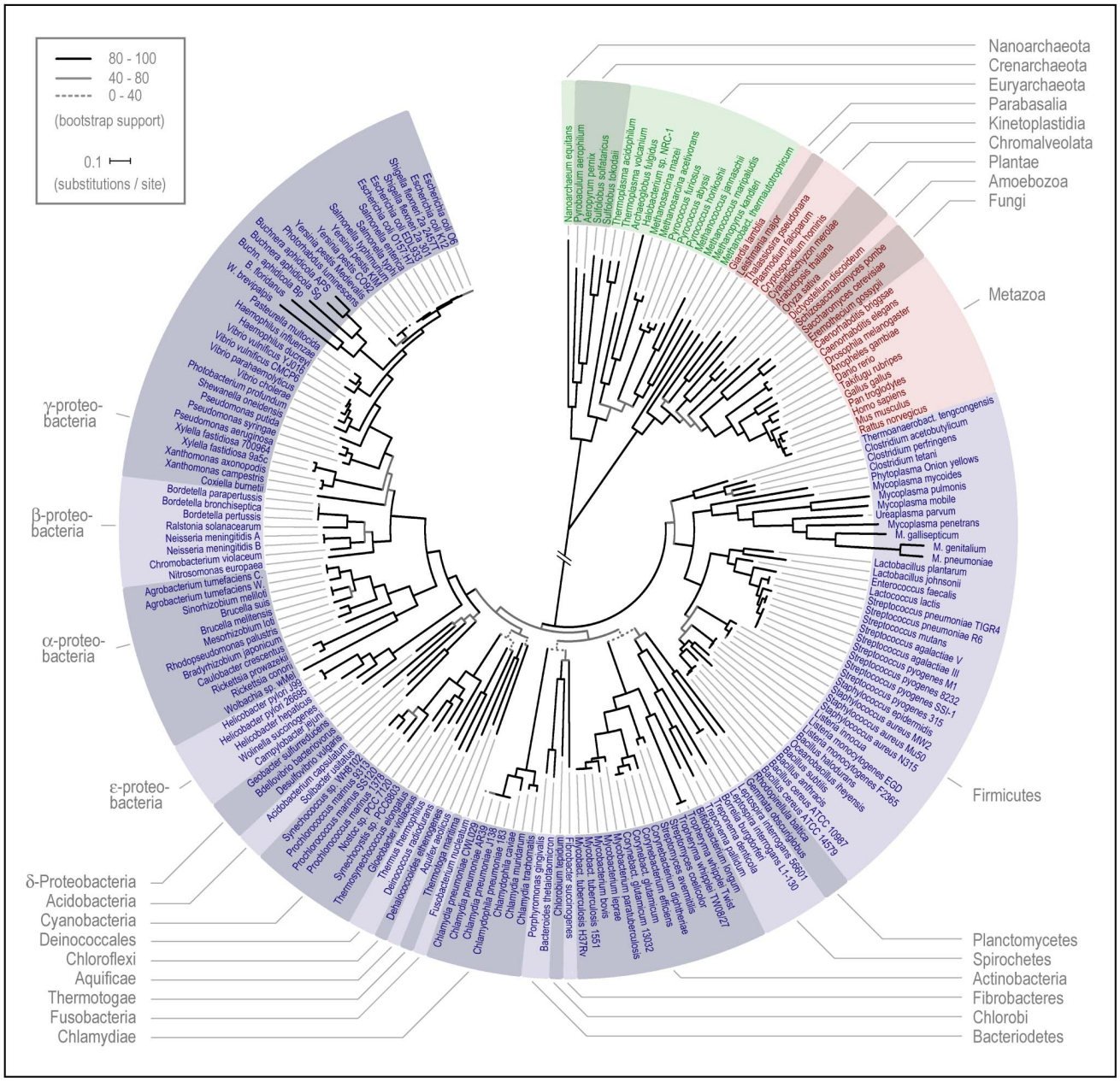
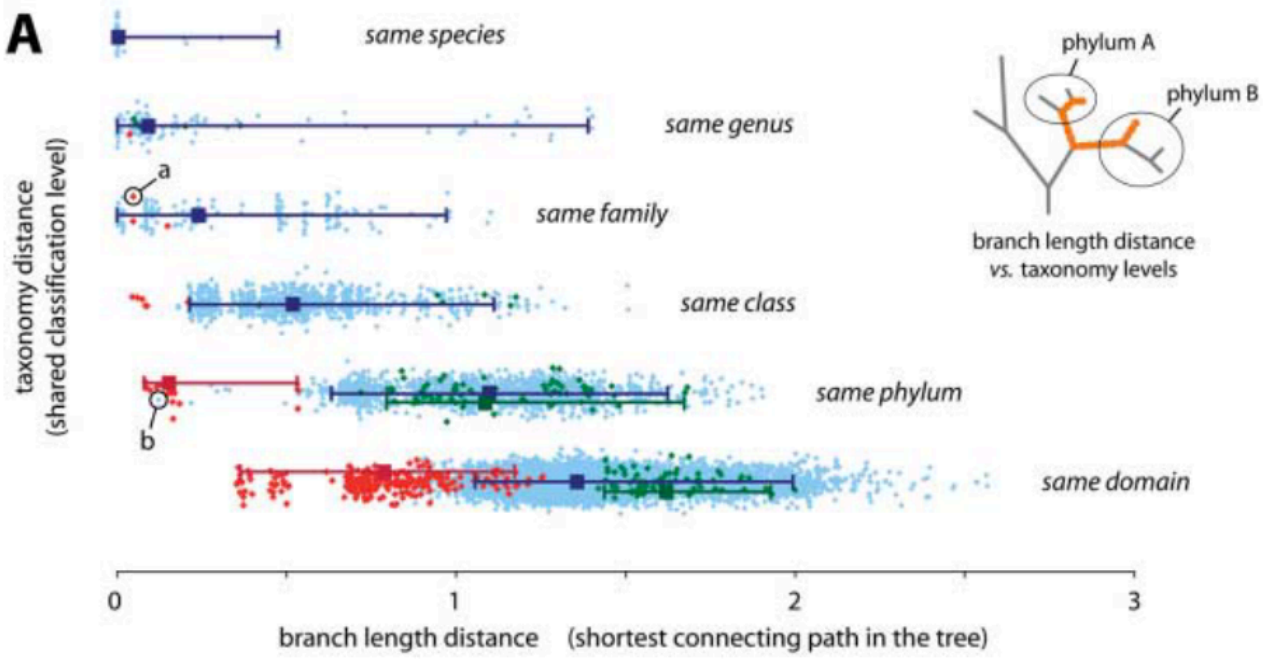

80 - 100
40 - 80
0 - 40
(bootstrap support)
0.1
(substitutions / site)
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Parabasalia
Kinetoplastidia
Chromalveolata
Plantae
Amoebozoa
Fungi
Metazoa
Firmicutes
Planctomycetes
Spirochetes
Actinobacteria
Fibrobacteres
Chlorobi
Bacteriodetes
Chlamydiae
Fusobacteria
Thermotogae
Aquificae
Chloroflexi
Deinococcales
Cyanobacteria
Acidobacteria
δ-Proteobacteria
ε-proteo-bacteria
α-proteo-bacteria
β-proteo-bacteria
γ-proteo-bacteria

Figure 2


A
same species
same genus
same family
same class
same phylum
same domain
taxonomy distance
(shared classification level)
a
b
0
1
2
3
branch length distance (shortest connecting path in the tree)
phylum A
phylum B
branch length distance
vs. taxonomy levels

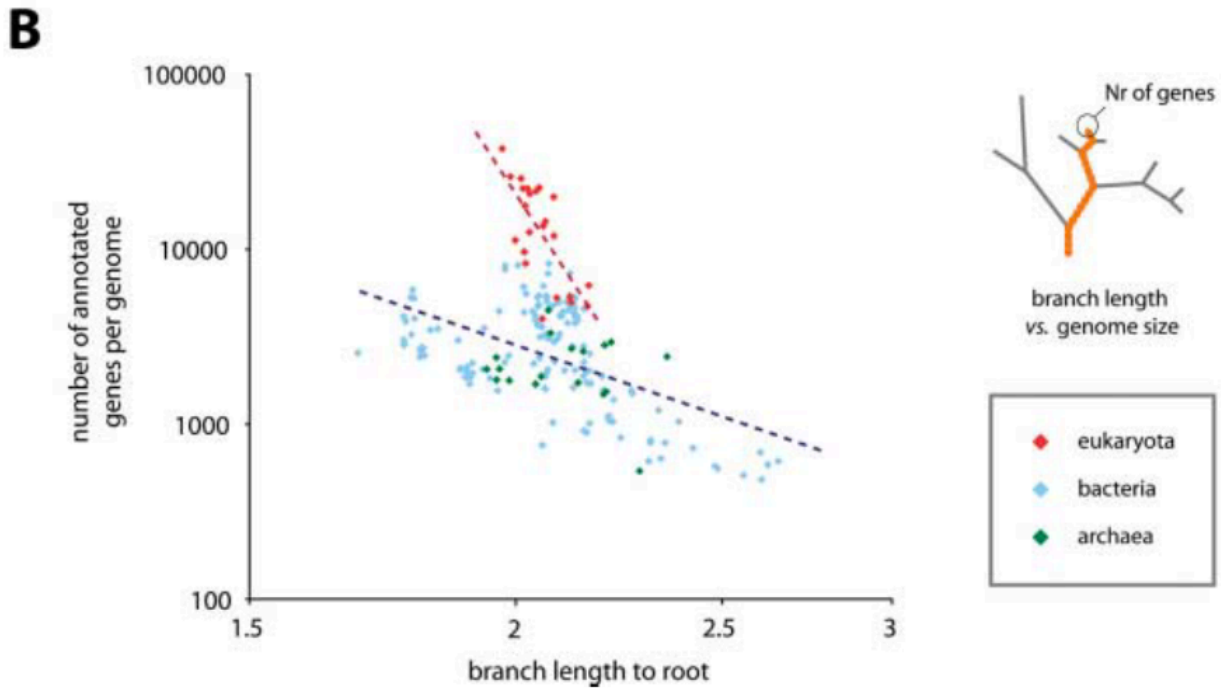

B
100000
10000
1000
100
number of annotated
genes per genome
1.5
2
2.5
3
branch length to root
Nr of genes
branch length
vs. genome size
eukaryota
bacteria
archaea

Figure 3