Evolutionary Bioinformatics

Libertas Academica
FREEDOM TO RESEARCH

ORIGINAL RESEARCH

Meta-Analysis of General Bacterial Subclades in Whole-Genome Phylogenies Using Tree Topology Profiling

Thomas Meinel[1,2] and Antje Krause[3]

[1]Charité—University Medicine Berlin, Institute for Physiology, Structural Bioinformatics Group, Thielallee 71, 14195 Berlin, Germany. [2]Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany. [3]FH Bingen University of Applied Sciences, Bioinformatics, Berlinstrasse 109, 55411 Bingen am Rhein, Germany.
Corresponding author email: thomas.meinel@charite.de

Abstract: In the last two decades, a large number of whole-genome phylogenies have been inferred to reconstruct the Tree of Life (ToL). Underlying data models range from gene or functionality content in species to phylogenetic gene family trees and multiple sequence alignments of concatenated protein sequences. Diversity in data models together with the use of different tree reconstruction techniques, disruptive biological effects and the steadily increasing number of genomes have led to a huge diversity in published phylogenies. Comparison of those and, moreover, identification of the impact of inference properties (underlying data model, inference technique) on particular reconstructions is almost impossible. In this work, we introduce tree topology profiling as a method to compare already published whole-genome phylogenies. This method requires visual determination of the particular topology in a drawn whole-genome phylogeny for a set of particular bacterial clans. For each clan, neighborhoods to other bacteria are collected into a catalogue of generalized alternative topologies. Particular topology alternatives found for an ordered list of bacterial clans reveal a topology profile that represents the analyzed phylogeny. To simulate the inhomogeneity of published gene content phylogenies we generate a set of seven phylogenies using different inference techniques and the SYSTERS-PhyloMatrix data model. After tree topology profiling on in total 54 selected published and newly inferred phylogenies, we separate artefactual from biologically meaningful phylogenies and associate particular inference results (phylogenies) with inference background (inference techniques as well as data models). Topological relationships of particular bacterial species groups are presented. With this work we introduce tree topology profiling into the scientific field of comparative phylogenomics.

Keywords: comparative phylogenomics, gene content phylogeny, inner tree topology, SYSTERS protein family set, tree topology profiling, whole-genome phylogeny

Evolutionary Bioinformatics 2012:8 489–525

doi: 10.4137/EBO.S9642

This article is available from http://www.la-press.com.

Background

The first representation of evolution of life on Earth in the form of a tree is the drawing headed ‘I think’ by Charles Darwin,[1] and the concept of a Tree of Life (ToL) has been continuously challenged and developed. The most valuable input, however, has come from the field of molecular biology. Due to the steadily increasing number of completely sequenced genomes[2] over the last two decades as well as inventions and refinements of algorithmic methodologies to computationally infer whole-genome phylogenies, a large number of strategies for utilizing molecular data have been developed. Consequently, a large number of ToL inferences have been published. The most robust kind of characterization is the gene content tree (GCT)—also known as the genomic tree, whole-genome tree, or genome tree—, which is based on the representation of gene family members in completely sequenced species and therefore is based on sequence homology. Published GCTs contain from tens to hundreds of analyzed species. Several approaches are attempted to optimize or to vary the underlying data, to compensate disturbances from evolutionary events. Other attempts have been dominated by optimization of the algorithmic methods and data processing pipelines. As a consequence, the complexity of whole-genome phylogeny inferences is overwhelming. Or, as House stated: ‘There is not a single gene content method. Rather, this is a broad category of genomic analysis that includes a wide variety of implemented methods, each with its own individual assumptions, strengths, and weaknesses’.[3]

In order to study the inference results for the ToL—in the last two decades represented by a large number of extremely different whole-genome phylogenies—and especially for the study of GCTs, it is essential to understand the background of the phylogeny inferences. Therefore we perform in the following chapter an in-depth analysis of the related literature recording data model, inference method and the purpose of attempted data conditioning. Reviewed literature on inferred whole-genome phylogenies is also part of Table 1.

Many of published whole-genome phylogenies are based on gene content data. However, published phylogenies are difficult to trace and, moreover, are often not reproducible because underlying data such as raw data or gene content meta-data are publicly not available. Gene content meta-data are representing gene families across a fixed set of species in a binary event matrix. The presence—absence representation of a particular gene family across a set of (completely sequenced) species in defined order, known as a phylogenetic profile,[4] ‘phyletic pattern’[5] or ‘conservation profile’,[6] can be used for the comparison of gene families. A phylogenetic profile is the first dimension of this presence—absence event matrix. The second dimension is orthogonal to a phylogenetic profile. It is a list in defined order across all gene families found in at least one of the species in the species set. It characterizes each single species, thereby. This particular view on the matrix basically allows for direct comparisons of the species. There exist some repositories such as the SYSTERS database[7] containing the PhyloMatrix as event matrix that offer the access to such meta-data.

The lack of methods to systematically or automatically evaluate respective whole-genome phylogenies is caused by heterogeneity in inference background and numbers of included species as well as by unavailability of raw or meta-data. With this paper we present a strategy to compare such phylogenies. It concentrates on tree topologies of drawn phylogenies. The manual analysis, by visual inspection of topologies of selected species groups (taxa), focuses on the determination of the two neighbors of a taxon in a bifurcating tree (most of published phylogenies are bifurcating). For each taxon, a row of topology alternatives (sets with different neighbors) can be found in the literature of the last two decades. All findings are, after generalization, compiled into a catalogue of alternative topologies and translated into digital states. These states enable a semi-quantitative validation by clustering. The advantage is that the topology characteristics are independent from the kind of phylogeny and species numbers in the trees. We define this method as ‘tree topology profiling’. To simulate the inhomogeneity in published phylogenies we, moreover, utilize the SYSTERS-PhyloMatrix[7] gene content data as standardized data model to apply seven frequently used phylogeny inference methods as testing variation. Such phylogenies can be mixed with published phylogenies and analyzed analogously using tree topology profiling. An overview of our overall strategy is presented in Figure 1. Applying tree topology profiling on 47 published and seven new inferences, the overall aim of this study is to provide general insights into the inference results of whole-genome phylogenies and the respective inference background.

Whole-Genome Phylogenies Inferences in the Literature

A large fraction of published whole-genome phylogenies are gene content phylogenies. Generally, gene content phylogeny inference is based on four operational steps: (i) the generation of the basic *data model*—comprising sets of homologous or orthologous gene groups in the majority of cases—, (ii) the *compensation* of disturbances by evolutionary events in the data by a conditioning approach, (iii) the construction of the event matrix of gene families across species—which is frequently binary—and (iv) using multiple *tree inference methodologies* that utilize the event matrix to generate the phylogenies in form of bifurcating trees. If distance-based algorithms are applied for tree inferences the metadata level of the binary event matrix must be translated into a set of distance matrices in forehand. Alternative concepts for whole-genome phylogeny inference require other data models and, depending from those, other inference algorithms.

A gene content data model and, subsequently, the quality of the resulting whole-genome phylogeny is determined by the accurate discovery of the relationship of evolutionarily dependent genes in different species. Relationships are presented as protein families or homologous or orthologous groups, depending on the basic aim of the procedure. First, exhaustive sequence similarity searches are limited by the comparability of protein sequences. The correct association of evolutionarily related genes to a shared gene family is then judged by separation criteria for similar genes from all other genes; such criteria characterize the family inference method; any kind of gene family inference requires these two essential steps.

A large fraction of GCTs analyzed in this study, Table 1, was inferred from the Clusters of Orthologous Groups (COGs),[8,9] a set of protein families found from completely sequenced prokaryotes (and a few eukaryotes). COGs are generated by a number of automatic and supervised processing steps. Other approaches consist of fully un-supervised data pipelines such as TRIBES[10] or SYSTERS.[11] Several GCT-constructing studies used own approaches to control the inference of orthologous groups (or other sets) or to provide the option for improvement (see later); other studies have incorporated annotation, eg, enzyme functionality of gene families.[12] Gene content phylogenies can be based on different levels of homology. The evolutionary objectives range from more focused (the orthologues) to a broader view (the homologues including the paralogues). Several inference attempts have provided alternatives for the homology background,[13–15] in particular, gene family inferences based on e-value variation.[16] COG-like inferences or inferences retrieved from reciprocal best matches on ORFs.[6,17–19] COGs are exploited in a large set of publications for the inference of gene content phylogenies.[14,20–28] Content data have also been published on the basis of functionality and enzyme content.[12,29–31] Moreover, protein domain content[32] and fold occurrence,[20,33] as well as gene order in COGs[14] have been exploited.

Alternatives to content data concepts are available: Super-alignments have been performed using COGs housekeeping genes[34,35] or marker gene families.[27,36] Resulting data were compiled within a database for orthologous groups including the COGs, known as the 'evolutionary genealogy of genes: nonsupervised orthologous groups' (eggNOG)[27] that has recently been extended.[37] Such integration across (concatenated sequences of multiple housekeeping) genes is an efficient substitute for any single gene phylogeny because 'no single gene (family) can serve as a proxy for the tree of life'.[38] Another data concept is the super-tree[39] built from phylogenetic trees of single gene families. A single gene family that has already been used to infer the ToL is the ubiquitous 16S rRNA family that is often denoted as 'the gold standard for an inference based on a phylogenetic tree'.[38] An example of such a phylogeny is given by Gevers et al (2004)[29] in combination with a paralogy analysis. A 16S rRNA phylogeny was also used for the reciprocal illumination of GCT inferences using the 'corroboration metric'.[16] Here, the authors inferred more than hundred GCTs based on the content of homologous genes based on COGs in order to find the optimal tree.

Drastic changes in genomes occur as particular evolutionary events. This can be attributed to gain and loss of sets of genes. As an example of gene loss, parasitic organisms partially utilize the genome of their host species and synchronously reduce their own genome.[40] The term 'reductive evolution' characterizes

Table 1. Characterization of published studies analyzed in this paper.

ID	Study	Figure	Tree size	Aim of the study variation in …		Inference background
				Data model	**Tree inference**	**Data model**
Tree 1	Deeds et al[33]	Fig. 4	50	no	✓	SCOP
Tree 2	Lin and Gerstein[20]	Fig. 2A	11	✓	no	COGs
Tree 3	Deeds et al[33]	Fig. 6	50	no	✓	SCOP
Tree 4	Snel et al[17]	Fig. 2A	13	no	no	COG-like
Tree 5	Grishin et al[70]	Fig. 3	19	no	no	Large protein domain families; COG-like; ird
Tree 6	Ciccarelli et al[34]	Fig. 2	191	no	no	Concatenated alignment of 31 COGs
Tree 7	Daubin et al[39]	Fig 2A	45	✓	✓	SuperTree of 730 phylogenetic trees
Tree 8	Brown et al[35]	Fig. 2	45	✓	no	14 concatenated proteins; minus HGT
Tree 9	Brown et al[35]	Fig. 1	45	✓	no	23 concatenated proteins
Tree 10	Ma and Zeng[12]	Fig. 1B	82	(✓)	✓	Enzyme content in metabolic network
Tree 11	Gevers et al[29]	Fig. 1	106	no	no	16S rRNA with functional annotation in the set of paralogs
Tree 12	Yang et al[32]	Fig. 3	174	no	no	SCOP; separation according to the kingdoms of life
Tree 13	Muller et al[27]	Fig. 1	630	no	no	COGs, KOGs and other OGs, particular gene families
Tree 14	Moran et al[43]	Fig. 1	72	–	–	Widely supported findings from different studies for symbionts
Tree 15	Wu and Eisen[36]	Fig. 2	578	no	no	Concatenated alignment of 31 housekeeping genes
Tree 16	Gophna et al[19]	Fig. 5	147	✓	no	ORFs, reciprocal best match; prevalence
Tree 17	Dutilh et al[23]	Fig. 4	89	no	✓	COGs, presence/absence profiles, weighted characters
Tree 18	Korbel et al[14]	Fig. 2	50	✓	no	COGs; gene order
Tree 19	Ge et al[56]	Fig. 2	40	✓	no	COGs
Tree 20	Clarke et al[18]	Fig. 5	37	✓	no	ORFs; after elimination of discordants
Tree 21	Clarke et al[18]	Fig. 2	37	✓	no	ORFs; before elimination of discordants
Tree 22	Wolf et al[21]	Fig. 5	40	✓	✓	probable orthologs
Tree 23	Sangaralingam et al[28]	Fig. 2	50	no	✓	COGs
Tree 24	Gophna et al[19]	Fig. 1	147	✓	no	ORFs, reciprocal best match
Tree 25	Korbel et al[14]	Fig. 1	50	✓	no	COGs; gene content
Tree 26	Dutilh et al[23]	Fig. 3	89	no	✓	COGs, presence/absence profiles
Tree 27	Daubin et al[39]	Fig. 2B	45	✓	✓	SuperTree of 730 phylogenetic Trees
Tree 28	Henz et al[69]	Fig. 2	91	no	✓	HSPs; matched distance; GBDP
Tree 29	Tekaia et al[6]	Fig. S2	99	✓	no	ORF products: orthologs
Tree 30	Wolf et al[21]	Fig. 4	40	✓	✓	COGs; gene pairs
Tree 31	Lienau et al[16]	Fig. 6	166	✓	no	SLC, conditioned reconstruction
Tree 32	Hughes et al[15]	Fig. 3	99	✓	no	SLC, e-value 10-6, similarity 60/80
Tree 33	Tekaia et al[92]	Fig. 2A	23	✓	no	ORF products
Tree 34	Hughes et al[15]	Fig. 2	99	✓	no	SLC, e-value 10-6, similarity 30/50
Tree 35	Ma and Zeng[12]	Fig. 1A	82	(✓)	✓	Enzyme content in metabolic network
Tree 36	Tekaia et al[6]	Fig. S3	99	✓	no	ORF products: homologs; ancestral duplications and weighted conservation
Tree 37	Tekaia et al[6]	Fig. S1	99	✓	no	ORF products; minimal profiles

		[all]	[ana]	[16S]	Tree type
Inference approach	Distance metric				
Dollo		4	2	✓	Domain content ?
Kitch	Hamming	12	1	✓	GCT
NJ		4	2	✓	Domain content ?
NJ	Simpson	2	1	✓	GCT
FM		1	1		GCT
ML					MSA-ML
BIONJ	gcd	4	2		SuperTree
MP		2	2		MSA-MP
MP		2	2		MSA-MP
NJ	Korbel	2 (+4)	2	(✓)	ECT
NJ		1	1	✓	16S rRNA
NJ		4	1		Domain content ?
ML					MSA-ML
					Integrative
ML					MSA-ML
FM	Weighted gene content	6	2		GCT
NJ	Korbel	2	2	(✓)	GCT
NJ	Korbel	3	2	✓	Gene order
NJ	PAM	1	1		GCT
FM		4	2	✓	GCT
FM		4	2	✓	GCT
NJ	Median of the percent identity distribution	4	3	✓	GCT
Non-phylogenetic model	Conditioned logdet distances	3	3		GCT
FM	Weighted gene content	6	2		GCT
NJ	Korbel	3	2	✓	GCT
NJ	Korbel	2	2	(✓)	GCT
ML		4	2		SuperTree
BIONJ		4	1		GCT
CA, HC	Jaccard	4	4		GCT
Dollo		4	3		GCT
Parsimony		1 (+7)	1	(✓)	GCT
Strict consensus tree of 6 MP Trees		3	2		SuperTree
CA; HC		3	1		GCT
Single MP		3	2		GCT ?
NJ	Jaccard	2 (+4)	2	(✓)	ECT
CA, HC	Jaccard	4	4		GCT
CA, HC	Jaccard	4	4		GCT

(*Continued*)

Table 1. (*Continued*)

ID	Study	Figure	Tree size	Aim of the study variation in …		Inference background
				Data model	Tree inference	Data model
Tree 38	Sangaralingam et al[28]	Fig. 1	49	no	✓	COGs
Tree 39	Wolf et al[22]	Fig. 1	59	no	no	COGs
Tree 40	Sangaralingam et al[28]	Fig. 3	50	no	✓	COGs
Tree 41	Spencer et al[25]	Fig. 4	66	✓	no	COGs; birth-death model
Tree 42	Spencer et al[25]	Fig. 5	66	✓	no	COGs; blocks model
Tree 43	Spencer et al[26]	Fig. 3	50	no	✓	COGs
Tree 44	Gu and Zhang[24]	Fig. 3	35	no	no	COGs
Tree 45	Tekaia et al[6]	Fig. 4	99	✓	no	ORF products: profiles
Tree 46	Hong et al[30]	Fig. 2A	42	no	no	Metabolic pathway content matrix
Tree 47	Wolf et al[21]	Fig. 3	40	✓	✓	COGs; gene content

this historical process. Species that have experienced this phenomenon can be found within several phyla or classes of the Bacteria, namely the Chlamydiae, the alpha-Proteobacteria,[41] the gamma-Proteobacteria,[42] the Actinobacteria, the Mollicutes and the Spirochaetes. As a consequence, the topologies of symbiotic[43] and obligate parasitic species are often incorrectly arranged in whole-genome phylogenies,[26] based on insights from taxonomy or detailed knowledge from molecular biology. Findings from molecular biology support the correct placement of the Mollicutes, *Rickettsia*, *Buchnera* and Leptospiraceae at the periphery of the ToL. The topology of the genus *Buchnera* is well supported within the gamma-Proteobacteria, the class Mollicutes within the Firmicutes and the genus *Rickettsia* within the alpha-Proteobacteria: These topologies have been confirmed using independent techniques such as conserved gene order for the genera *Buchnera*[42] and *Rickettsia*,[44] and using significant phylogenetic marker genes other than rRNA, such as *PGK*, for Mollicutes.[45] For the Leptospiraceae, phylogenetic analyses of the unusual, spirochete-specific 16S rRNA[46] as well as chemotaxonomy studies[47] confirmed the detailed topology within the phylum Spirochaetes. For *Rickettsia*, particular protein signatures[48] or a group of functionally related genes in a protein complex were used,[49] as well as a combination of gene content, gene order and gene conservation.[50]

Such information helps to identify a species group for that an ambiguous topology can be identified as correct or incorrect (such topology is 'confirmable'). Otherwise, such a decision is impossible ('non-confirmable' topology; no support by other experimental findings). Taxa to be analyzed are often branching deep in the inner tree; such topologies are seldom confirmable. Confirmability of a particular topology is an essential argumentation feature in this paper; it also can be a feature of a character state in the topology analysis.

Another major evolutionary event is horizontal gene transfer (HGT), a process that enables organisms to acquire genetic material from one another. HGT has been demonstrated for a set of genes in particular species groups, for example in the Cyanobacteria[51]

Inference approach	Distance metric	[all]	[ana]	[16S]	Tree type
Modified BIONJ	Conditioned logdet distances	3	3		GCT
FM	Jaccard	1	1		GCT
Phylogenetic model	Conditioned logdet distances	3	3		GCT
Least squares, inverse square weighting					GCT
Least squares, inverse square weighting					GCT
Modified BIONJ	Conditioned logdet distances	3	3	✓	GCT
NJ	ggd	1	1		GCT
CA, HC	Jaccard	4	4		GCT
Complete linkage clustering		2	1	✓	ECT
Dollo		4	3		GCT

Notes: Publications are ordered according to the heatmap Figure 2. Given are numbers for [all] and in this study [ana]-lyzed Trees and indication if a [16S] rRNA Tree is used or represented by the authors. Further study parameters are extracted: aim, background information, Tree size (number of species), and the internal Tree ID which gives orientation in the text of our study. Tree type is determined from the inference background of the phylogeny.

Abbreviations: BIONJ[68]; CA, correspondence analysis; Dice, Dice distance; Dollo, Dollo parsimony algorithm; FM, Fitch-Margoliash algorithm; GBDP, genome blast distance phylogeny; gcd, gamma corrected distance; ggd, general genome distance; HC, hierarchical clustering; -HGT, (without) horizontal gene transfer; HSPs, high-scoring (sequence) segment pairs; ird, inter-protein rate distribution; Jaccard, Jaccard distance; Kitch, Kitch algorithm; Korbel, distance given in Korbel; ML, maximum likelihood; MP, maximum parsimony; NJ, Neighbor Joining algorithm; SCOP, Structural Classification of Proteins[93]; COGs, Clusters of Orthologous Groups[8]; SLC, single linkage clustering; ORFs, open reading frame; Simpson, Simpson distance.

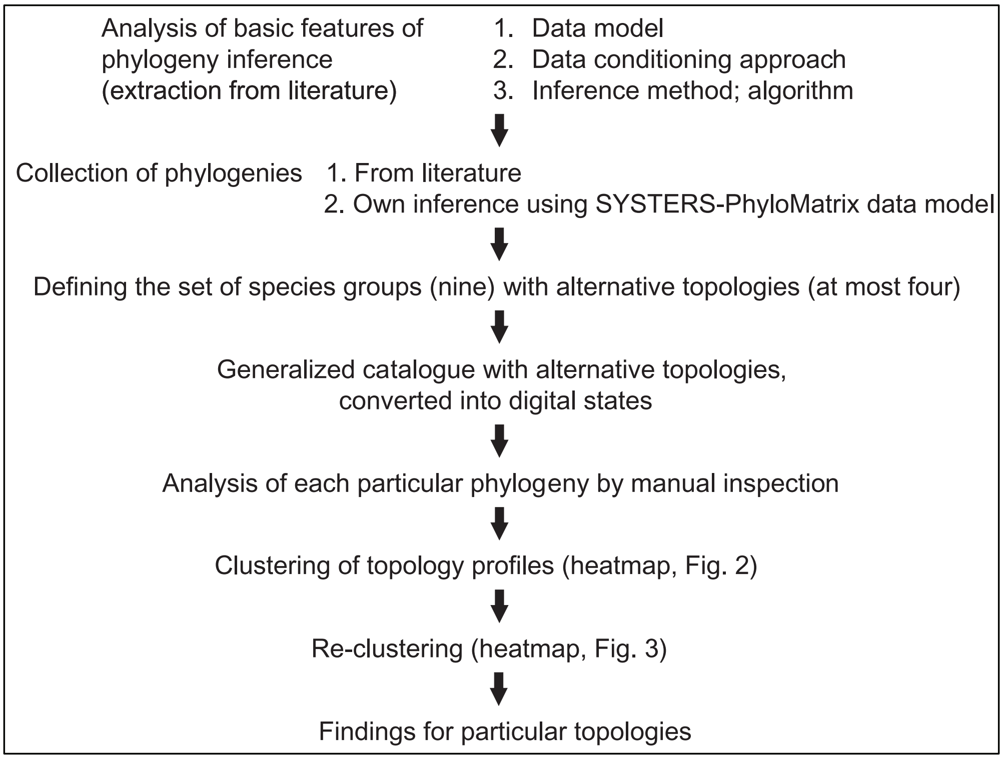

Analysis of basic features of phylogeny inference (extraction from literature)
1. Data model
2. Data conditioning approach
3. Inference method; algorithm
Collection of phylogenies
1. From literature
2. Own inference using SYSTERS-PhyloMatrix data model
Defining the set of species groups (nine) with alternative topologies (at most four)
Generalized catalogue with alternative topologies, converted into digital states
Analysis of each particular phylogeny by manual inspection
Clustering of topology profiles (heatmap, Fig. 2)
Re-clustering (heatmap, Fig. 3)
Findings for particular topologies

Figure 1. Workflow applied in this study.
Note: See details in the sections Material and Methods and Results.

and the Chlamydiae,[52] or for single genes such as the *GAPDH* gene in the Spirochaetes,[53] chitinases in the Actinobacteria[54] and the ubiquitous 16S rRNA gene family. The latter revealed how life's early history can be depicted in the context of HGT.[38] For two gene families, GTPase and dimethyladenosine transferase, re-ordering of the respective gene phylogenies was demonstrated by Koonin and Wolf.[2] The next paragraph discusses strategies that are also used to find solutions to the problem of HGT compensation by data conditioning.

To compensate such disruptive biological effects on phylogeny inferences, several attempts were initiated to model related evolutionary events, the gain and loss of genes. This is achieved with conditioning of data,[55] the successive reduction of discordant homologues,[18] the reduction of noise to achieve consistent signals,[23] the weighting of trees based on prevalence and concordance,[19] or balancing single disturbing events at genome-scale dimensions.[56] Stochastic mapping[57] involves varying the rates for particular gene families. Conditioned reconstruction was found to work well even in the presence of HGT;[58] however, the interpretation of phylogenies varies according to the method used to construct them. Modeling of loss and gain has been performed using 'blocks' of genes rather than single genes.[24,25] As mentioned in Cohen and Pupko,[57] there are no tests for HGT inference based on probabilistic-evolutionary models. But HGT was modeled by among-gene-family-rate variations[59,60] or as a phylogenetic mixture model[26] in about 50 species. Here, 11 out of the 12 parasites within 50 analyzed bacterial species clustered in a single, monophyletic clade. Therefore, gene loss has recently been modeled by the same researcher group[28] using algorithms such as phylogenetic mixture models in conditioned log-det phylogenies. This study questioned whether a bias in the three underlying biological data models or the tree inference methods caused the strong convergent signal for parasites in a whole-genome phylogeny, the 'artefactual parasitic eubacteria clan'.[28] The authors concentrated on the method, they found that 'the most successful methods for estimating a reliable phylogenetic tree for parasitic and endosymbiotic eubacteria from gene content data are still ad-hoc approaches such as the SHOT[14] distance method'.

Gene content data can be exploited by parsimony or heuristic, distance-based algorithms. An early approach that exploits the presence—absence status of genes is the Dollo algorithm.[61,62] This parsimony method was used in several investigations, often for comparison with other inference techniques.[21,33] It allows a single invention of a particular gene family (at one point in time; as contribute to one particular edge of the rooted Dollo tree) and minimizes the number of gene losses across all species. The Dollo parsimony was also connected with the convergent signal for parasites in a whole-genome phylogeny to be placed a single, monophyletic clade[21,63] but it is mentioned that there is a 'potential of simpler methods to cope better with the issue of genome size echoes'.[63] Other maximum parsimony methods, such as the Wagner parsimony of the 'mix' program in the Phylip suite[64] or similar implementations, were applied to sets based on sequence data[35] or homologous family inference settings.[15] A large number of whole-genome trees, however, are generated using distance-based algorithms such as Neighbor Joining (NJ),[65] the Fitch-Margoliash algorithm (FM),[66] the Unweighted Pair Group Method with Arithmetic Mean (UPGMA),[67] or the KITCH algorithm.[64] The 'Shared Ortholog and Gene Order Tree Reconstruction Tool' (SHOT)[14] enables not only the comparison of different homology approaches but also the option for both the NJ and the FM algorithm in combination with different distance metrics. Moreover, NJ is a frequently used algorithm[12,13,17,21,23,24,32,56] with convenient features in terms of computing performance. It allows variability in the distance metrics and larger numbers of species. A derivate of the NJ algorithm, established in BIONJ,[68] was used to study the difference between a distance-based tree and two super-trees,[39] that were inferred either with a Maximum Likelihood (ML) method or are based on a set of phylogenetic gene family trees. NJ was also compared with UPGMA and totally different concepts such as split graphs.[69] The FM algorithm, a method that uses a weighted least squares method for clustering and often synonymously referred to as 'weighted least squares', was used to compare variations in weighted gene content,[19] the elimination of discordant genes[18] and a COG-like approach for protein domain families that studies substitution rates between proteins.[70] FM was also used to compare the resulting tree with earlier results inferred by Dollo parsimony phylogenies.[21,22] Here, the KITCH algorithm together with the Hamming distance (genomic data are from COGs families) was studied with regard to differences between parsimony trees based on structural or rRNA features, but only using an extremely small number of species early on in the study of whole-genome phylogenies.[20]

Distance-based algorithms have been used with a broad arsenal of standard metrics such as the Hamming, the Jaccard, or the Simpson metrics; for a review of the mathematical background see Cheetham and Hazel.[71] Another metric, based on the geometric and arithmetic mean, was introduced by Korbel et al[14] with better inference results;[23] this metric has been extensively compared with the Simpson and Jaccard metrics.[12] Other authors introduced their own metrics such as the median of percent identity[21] or 'gamma corrected distance'[24] to infer adaptations by refining parameterization or a 'general genome distance'.[39]

Methods

Set of nine species groups (Taxa) with alternative topologies

For our meta-analysis we evaluated published topologies and topologies from own whole-genome phylogeny inferences according to particular bacterial subclades or their representatives. Our selection of relevant species groups (taxa) followed existing studies on gene loss, as reported elsewhere.[26,28] We undertook these 12 species that are organized in six species groups of diverse taxonomic ranks (ie, genus, family, class or phylum): *Mycobacterium leprae* (Actinobacteria, rank: phylum); *Chlamydia trachomatis*, *Chlamydia pneumoniae CWL*029 (Chlamydiae, rank: phylum); *Buchnera sp. APS* (*Buchnera*, rank: genus); *Treponema pallidum*, *Borrelia burgdorferi* (Spirochaetes, rank: phylum); *Rickettsia prowazekii*, *Rickettsia conorii* (*Rickettsia*, rank: genus); *Ureaplasma urealyticum*, *Mycoplasma pulmonis*, *Mycoplasma pneumoniae* and *Mycoplasma genitalium* (Mollicutes, rank: class). We extended this selection by eight more species (for details see supplemental data, Table S1) belonging to the six groups and one species, *Leptospira interrogans*, belonging to another parasitic group (Leptospiraceae, rank: family). We included a further four species belonging to two additional groups, Cyanobacteria (rank: phylum) and Chlorobi (rank: phylum), to cover gene gain events by particular gene families such as those for the photosynthetic apparatus. Firmicutes and Proteobacteria were the anchor clades in this study and were therefore not regarded as characters in topology analyses. In sum, we collected nine species groups for topology analyses.

Catalogue of alternative topologies for nine taxa—classification rules

A set of up to four most common topology alternatives for each species group was obtained from the respective literature by analyzing drawn phylogenies. In order to computationally treat found topology descriptions, particular verbal descriptions are translated into states (scores and colors). The result is compiled in Table 2 as a catalogue, after generalization across all available topologies for four groups of species with confirmable (ie, supported by additional literature) and five species groups with non-confirmable topologies (ie, topologies that can be not supported by respective literature). Several of the evaluated species are parasites and are frequently observed to appear together in a shared sub-clade. If such topology is observed, we denote this topology state as a '*(shared) parasitic subclade*'.

For confirmable topologies, there are only two alternatives observed, the true, *ergo* the confirmed, topology [+2] or the association to the parasitic subclade [−2], in particular:

- ForBuchnera,Rickettsia,Mollicutes,Leptospiraceae, the status is +2 if respective taxa are within gamma-Proteobacteria, alpha-Proteobacteria, Firmicutes, Spirochaetes.

We defined the character sets with more than two alternatives for the following five species groups with non-confirmable topologies in the order of decision:

- Actinobacteria: if near Eukaryota/Archaea and Cyanobacteria, status is +1; if near Eukaryota/Archaea and other, status is +2; if within the parasitic subclade, status is −2; otherwise, status is −1.
- Cyanobacteria: if near Actinobacteria, status is −1; if near Chlorobi, status is +1; if within the parasitic subclade, status is −2; otherwise, is +2.
- Chlamydiae: if within the parasitic subclade, status is −2; if near Spirochaetes only, status is −1; otherwise, status is +2.
- Chlorobi: if near Cyanobacteria, status is +2; if within the parasitic subclade, status is −2; otherwise, status is −1.

Table 2. Generalized topology alternatives found in published phylogenies of selected nine species groups.

Species group	Topology	Topology alternative			
Leptospiraceae	Confirmable	Not within Spirochaetes	–	–	Within Spirochaetes
Buchnera	Confirmable	Among parasites	–	–	Within γ-Proteobacteria
Rickettsia	Confirmable	Among parasites	–	–	Within α-Proteobacteria
Mollicutes	Confirmable	Among parasites	–	–	Near other Firmicutes
Spirochaetes	Non-confirmable	Among parasites	Between Chlamydiae and (Proteobacteria or E A)	Between Firmicutes and (A or Proteobacteria or Cyanobacteria or Actinobacteria)	Between Chlamydiae and Firmicutes
Chlorobi	Non-confirmable	Among parasites	Other	–	Near Cyanobacteria
Chlamydiae	Non-confirmable	Among parasites	Near Spirochaetes only	–	Other
Cyanobacteria	Non-confirmable	Among parasites	Near Actinobacteria	Near Chlorobi	Other
Actinobacteria	Non-confirmable	Among parasites	Other	Near or between Cyanobacteria and E,A	Near or between Cyanobacteria and (Firmicutes or Proteobacteria)
Score/character status		−2	−1	1	2
Color in heat map, Figure 3		Red	Orange	Turquoise	Blue

Notes: Description is given with respective metrics (+2, +1, −1, or −2) and color coding. The term 'within' indicates that the species group 'is monophyletic with' the respective topology alternative, ie, the species are contained in the subclade. Otherwise, the species group is paraphyletic, ie, is a sister group of the indicated subclade(s).
Abbreviations: E, Eukaryota; A, Archaea (if not ignored in phylogeny).

– Spirochaetes: if within the parasitic subclade, status is −2; if between the Chlamydiae and Proteobacteria or Eukaryota/Archaea, status is −1; if between the Chlamydiae and Firmicutes, status is +2; otherwise, status is +1.

Analysis of published phylogenies using the catalogue of alternative topologies—event matrix

Each published phylogeny tree was manually analyzed by visual inspection and selecting the best fitting state from the general catalogue (Table 2). Observable events in each tree were determined using the exclusion principle (see sub-section above). If the appropriate characteristic was based on a missing feature we used the best alternative. In recent phylogenies that contain hundreds of species, the original set of topology alternatives may be obvious because the analyzed phylogeny includes newer species clades. In such cases, the subjectively best approximation was chosen for validation. The complete inspection result can be found as event matrices (for published phylogenies as Table S2 in the Supplement and for SYSTERS-PhyloMatrix results in Table 3) and in the translated visualization given by Figure 2.

Clustering of the event matrix

We used the R package[72] to generate a heatmap (pheatmap library) from the topology event matrix with default settings such as the hierarchical clustering for rows (with phylogeny annotation) and average linkage clustering with Euclidean distance.

Selection of particular taxa and phylogenies (division 1 and 2)

We introduced two general divisions of the heatmap according to the following decisions. The first division separated four taxa with a confirmable topology (Leptospiraceae (within Spirochaetes), *Buchnera* (within gamma-Proteobacteria), *Rickettsia* (within alpha-Proteobacteria) and Mollicutes (within Firmicutes)) from the remaining set, including the parasites Spirochaetes, Chlamydiae and Actinobacteria. As the second general division, correctly inferred phylogenies were separated from wrong phylogenies with respect to the correct status of the four confirmable topologies.

Table 3. Topology profiles for seven SYSTERS-PhyloMatrix gene content trees across the nine species groups in the topology catalogue: event matrix; score definitions can be found in Table 2 and phylogenies in Figures S1 to S7.

Applied algorithm	Distance metric	Leptospiraceae	Buchnera	Rickettsia	Mollicutes	Spirochetes	Chlamydiae	Actinobacteria	Cyanobacteria	Chlorobi
NJ (Neighbor Joining)	Korbel	2	2	2	2	–1	–1	–1	–1	–1
NJ	Simpson	2	2	2	2	1	1	1	–2	–1
Dollo parsimony		–2	–2	–2	–2	–2	–2	–1	2	–1
NJ	Hamming	–2	–2	–2	–2	–2	–2	–1	1	2
Wagner parsimony		–2	–2	–2	–2	–2	–2	2	–1	–1
NJ	Dice	–2	–2	–2	–2	–2	–2	1	–1	–1
NJ	Jaccard	–2	–2	–2	–2	–2	–2	2	–1	2

Three situations arose: possessed the correct topology (all four with +2 status), or possessed partially correct topologies (mixed status [+2; –2]) or incorrect topologies (all four with –2 status). The latter typically included the shared parasitic subclade. For simplification, only the second step is shown in Figure 2.

Re-clustering of part of the heatmap

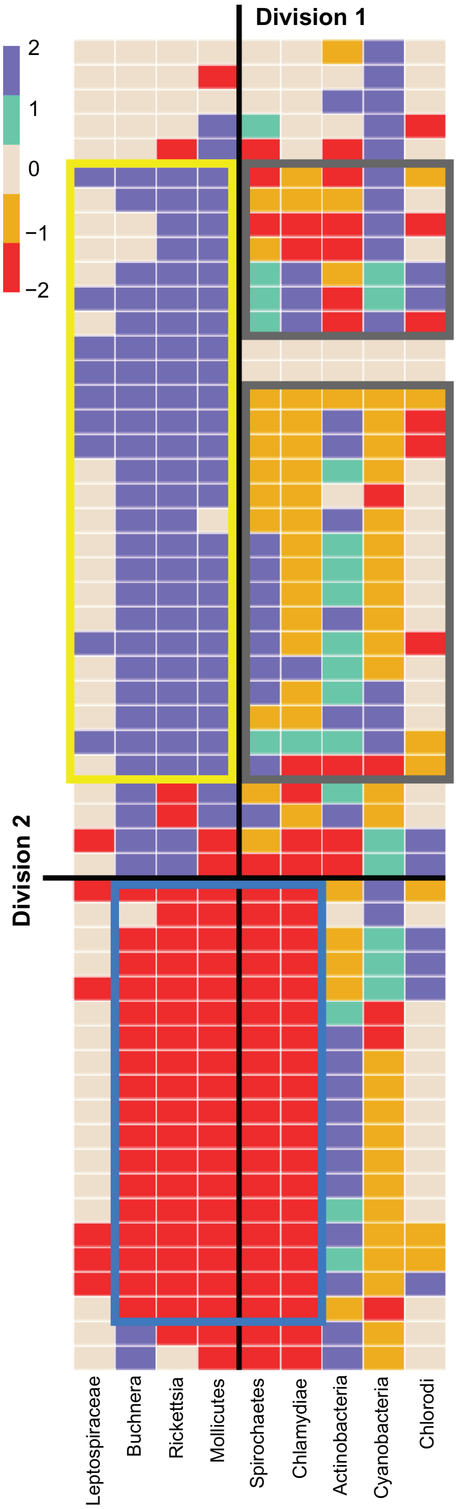
Further analysis was based on the intersection of the two selection divisions. However, several phylogenies were excluded from further analysis because they lacked a considerable number of interesting species groups. This is because whole-genome research ignored parasitic species for particular inferences, very early phylogenies were too small, or the study focused on only a single bacterial clade. Phylogenies with unresolved topologies were also excluded (ie, no bifurcations in the inner tree). The resulting subset comprised (a) data (phylogenies) with the correct topology for all confirmable taxa and (b) taxa that have no confirmable topology. The relevant part (see subcluster of selected phylogenies and taxa in the gray rectangle in Fig. 2) of the heatmap was re-clustered with the same clustering parameters.

Use of SYSTERS-PhyloMatrix data model for phylogeny inference

We retrieved the PhyloMatrix data[7] from SYSTERS release 4 as a binary presence–absence matrix of 19374 protein families from 106 completely sequenced species. The SYSTERS protein family data[11] are based on the fully automated inference of families in two hierarchies by protein sequence similarity; SYSTERS does not explicitly infer families at a particular level of homology. A detailed list of the 106 species is provided in the supplemental data, Table S3. The PhyloMatrix data model (http://systers.molgen.mpg.de/cgi-bin/info.pl) is generated under the condition that at least three completely sequenced species are represented in a single family.

Gene content tree inference and algorithms

We used the Phylip package[64] to generate GCTs. The binary matrix (PhyloMatrix) was translated into Phylip-compatible distance matrices between all pairs of the 106 species using our own Perl scripts according to the five distance or similarity metrics (Hamming, Jaccard, Dice, Korbel and Simpson). With the exception of the Hamming distance, metric distances d were calculated from similarity s according to the formula $d = 1 - s$. These distance metrics are described in more detail elsewhere,[71] as is the Korbel metric.[14] The 'neighbor' program (Neighbor Joining) was used as a distance-based algorithm with standard settings. Bootstrapping with 100 replicates was conducted for distance-based trees using the 'seqboot' and 'consense' program. For parsimony trees, we used the 'mix' program with Wagner parsimony, as well as the 'dollop' program for Dollo trees with standard settings. The Dollo phylogeny is a rooted tree in contrast to all other phylogenomic inferences.



Deeds et al 2005	Fig. 4	Tree 1	Domain content	Improvement
Lin and gerstein 2000	Fig. 2A	Tree 2	GCT	Improved vs. 16S rRNA
Deeds et al 2005	Fig. 6	Tree 3	Domain content	Improvement
Snel et al 1999	Fig. 2a	Tree 1	GCT	Improved vs. 16S rRNA
Grishin et al 2000	Fig. 3	Tree 5	GCT	Single phylogeny
Ciccarelli et al 2006	Fig. 2	Tree 6	MSA-ML	Single phylogeny
Daubin et al 2002	Fig. 2A	Tree 7	Supertree	Reference for own work
Brown et al 2002	Fig. 2	Tree 8	MSA-MP	Improvement
Brown et al 2001	Fig. 1	Tree 9	MSA-MP	Reference for own work
Ma and zeng 2004	Fig. 1B	Tree 10	ECT	Improvement
Gevers et al 2004	Fig. 1	Tree 11	16S rRNA	Single phylogeny
Yang et al 2005	Fig. 3	Tree 12	Domain content	Single phylogeny
Muller et al 2010	Fig. 1	Tree 13	MSA-ML	Single phylogeny
Moran et al 2008	Fig. 1	Tree 14	Integrative	Integrative phylogeny
Phylomatrix/NJ/Korbel				
Wu and Eisen 2008	Fig. 2	Tree 15	MSA-ML	Single phylo.geny
Gophna et al 2005	Fig. 5	Tree 16	GCT	Improvement
Dutilh et al 2004	Fig. 4	Tree 17	GCT	Improvement
Korbel et al 2002	Fig. 2	Tree 18	Gene order	Improved vs. 16S rRNA
Ge et al 2005	Fig. 2	Tree 19	GCT	Single phylogeny
Clarke et al 2002	Fig. 5	Tree 20	GCT	Improved vs. 16S rRNA
Clarke et al 2002	Fig. 2	Tree 21	GCT	Improved vs. 16S rRNA
Wolf et al 2001	Fig. 5	Tree 22	GCT	Improvement
Sangaralingam et al 2010	Fig. 2	Tree 23	GCT	Improvement
Gophna et al 2005	Fig. 1	Tree 24	GCT	Reference for own work
Korbel et al 2002	Fig. 1	Tree 25	GCT	Improved vs. 16S rRNA
Dutilh et al 2004	Fig. 3	Tree 26	GCT	Reference for own work
Daubin et al 2002	Fig. 2B	Tree 27	Supertree	Improvement
Phylomatrix/NJ/Simpson				
Henz et al 2005	Fig. 2	Tree 28	GCT	Improvement
Tekaia et al 2005	Fig. S2	Tree 29	GCT	Improvement
Wolf et al 2001	Fig. 4	Tree 30	GCT	Improvement
Lienau et al 2006	Fig. 6	Tree 31	GCT	Optimized data model
Hughes et al 2005	Fig. 3	Tree 32	Supertree	Improvement
PhyloMatrix/Dollo parsimony				
Tekaia et al 1999	Fig. 2a	Tree 33	GCT	Reference for own work
Hughes et al 2005	Fig. 2	Tree 34	GCT ?	Reference for own work
Ma and zeng 2004	Fig. 1A	Tree 35	ECT	Reference for own work
PhyloMatrix/NJ/Hamming				
Tekaia et al 2005	Fig. S3	Tree 36	GCT	Improvement
Tekaia et al 2005	Fig. S1	Tree 37	GCT	Improvement
Sangaralingam et al 2010	Fig. 1	Tree 38	GCT	Reference for own work
Wolf et al 2002	Fig. 1	Tree 39	GCT	Improvement
Sangaralingam et al 2010	Fig. 3	Tree 40	GCT	Improvement
Spencer et al 2006	Fig. 4	Tree 41	GCT	Reference for own work
Spencer et al 2006	Fig. 5	Tree 42	GCT	Improvement
Spencer et al 2009	Fig. 3	Tree 43	GCT	Improvement
Gu and Zhang 2004	Fig. 3	Tree 44	GCT	Single phylogeny
Phylomatrix/Wagner parsimony				
Phylomatrix/NJ/Dice				
Phylomatrix/NJ/Jaccard				
Tekaia et al 2005	Fig. 4	Tree 45	GCT	Improvement
Hong et al 2004	Fig. 2A	Tree 46	ECT	Improved vs 16S rRNA
Wolf et al 2001	Fig. 3	Tree 47	GCT	Reference for owm work

Figure 2. Topology profiles across nine species groups derived from 54 phylogenies (seven SYSTERS-PhyloMatrix gene content trees and 47 whole-genome phylogenies, Table 1).

Notes: Annotations in the table: citation and figure index in the respective publication, tree ID in this study, data model, and approach in the author's opinion. Heatmap color definitions for up to four topology alternatives of respective taxa are given in Table 2 (light gray: species not regarded in the respective publication). Division 1 separates confirmable topologies from non-confirmable. Division 2 occurs several times, shown is here only one division that mainly separates the parasitic subclade from the rest; see Methods section. Particular topology states (in event matrices) are given in Table 3 (SYSTERS-PhyloMatrix trees) and in the supplemental Table S2.

For comparison, the branches in all other trees were swapped using NJPLOT[73] with the goal of placing the *Buchnera* at the top and the Eukaryota at the bottom of the respective tree.

Results

Literature analysis

The first part of our study comprised an in-depth analysis of published whole-genome phylogenies, see Table 1. We extracted the underlying biological data model and the tree inference methodology as the two main background features in phylogeny inferences. Based on these two features, we classified the tree type in a separate column. This classification was essential for the subsequent topology analysis. We also give the aim of each particular study as well as statistics such as the number of species included, the number of trees inferred, the number of trees used for analysis in this study, and whether a 16S rRNA reference phylogeny was presented. The order of the particular phylogenies is oriented on the clustering result later. The classification focused on three issues:

- Underlying data models for the generation of whole-genome phylogenies
- Data conditioning, modeling of evolutionary events
- Methodologies used to infer whole-genome phylogenies

Table 1 includes a large number of existing modeling approaches. We decided that modeling does not require further extensions by an own work. Moreover, a comparative and comprehensive analysis of existing topologies remained to be carried out: we developed tree topology profiling for this purpose. We designed the workflow that is sketched in Figure 1.

Topologies found in published phylogenies

Topology analysis

The catalogue of generalized alternative topologies, as shown in Table 2, was derived from published whole-genome phylogenies with organisms from all super-kingdoms of life, the Bacteria, Archaea and Eukaryota. By inspecting 47 whole-genome phylogenies from 30 literature references, we determined the existing topologies for the nine selected groups of species. Applying the catalogue of characteristic topology alternatives led to an event matrix for all referenced phylogenies. The resulting event matrix can be traced in the supplemental data, Table S2. The order of the phylogenies in Table S2 (as well as in the descriptive Table 1) is derived from the clustering shown in Figure 2, which is already extended by the seven new and later described phylogeny inferences.

Topology of parasitic species and shared subclades

It has often been reported that the parasitic species occur together in a monophyletic subclade. Within the 47 analyzed whole-genome phylogenies we found a considerable number of that quality. Using several data models and inference methods. The parasitic subclade includes three species groups with confirmable topologies (*Buchnera*, *Rickettsia* and Mollicutes) and, of those with no confirmable topology, the *Chlamydiae* and *Spirochaetes*. Interestingly, in most cases, these five species groups in the topology profile share the parasitic subclade if such is formed. After clustering, the corresponding subcluster (status [–2]; see Table 2) is located in the lower part to the left of the event matrix and of Figure 2 (blue rectangle).

Parasitic monophyly has been reported for Dollo parsimony [tree 47][21] but also for ML trees of single gene families, eg, a 16S rRNA phylogeny [this tree was not analyzed in this work].[26] The parasitic monophyly also results from inferences using correspondence analysis (CA) [tree 29; tree 36; tree 37; tree 45],[6] inferred with distance methods that use the Hamming or Jaccard or other distances [tree 2; tree 35; tree 38; tree 39].[12,20,22,28]

Only seven phylogenies include a partial parasitic subclade that is constructed from only some of the five parasitic species groups [tree 5; tree 29; tree 30; tree 31; tree 32; tree 46; tree 47].

Confirmable topologies of particular groups of parasitic species

As the alternative to the shared subclade of parasites, well-confirmed topologies are known for the three species groups *Buchnera*, *Rickettsia* and Mollicutes; correct placement is supported by findings from molecular biology. No other topology alternatives as the respective correct topologies or parasitic monophyly were observed for these three species groups. Topology information for the Leptospiraceae is only available from seven more recent publications. They are correctly located within the Spirochaetes in nearly all phylogenies. In one of the exceptions [tree 31],[16] Leptospiraceae are neither found in the parasitic subclade nor in the correct placement near the Spirochaetes. The placement of all four species groups (if present in the phylogeny) in their respective confirmed subclade occurs in 23 published phylogenies. This situation corresponds to a respective topology subcluster and can be found in the upper part to the left of the event matrix (Table S2; yellow rectangle in Fig. 2).

The types of whole-genome phylogenies with correct topologies for the confirmable parasites were either super-trees [tree 7; tree 27], MSA-ML trees [tree 6; tree 13; tree 15] or MSA-MP trees [tree 8; tree 9]. A more recent MSA-ML tree [tree 13] and the literature consensus tree for symbionts [tree 14] were only resolved for these four taxa and were included in the analysis. The majority of the correct trees, however, are (eleven) GCTs inferred using distance methods [tree 4; tree 16; tree 17; tree 19; tree 20; tree 21; tree 22; tree 24; tree 25; tree 26; tree 28]. Interestingly, three gene content phylogenies [tree 17; tree 25; tree 26], one enzyme content phylogeny [tree 10] and one *gene order* phylogeny [tree 18] were inferred with NJ using the Korbel distance. The underlying data models, however, vary widely, from phylogenetic trees to MSAs, gene order, enzyme content and gene content.

Species groups with non-confirmable inner tree topologies

Five out of the nine species groups in our profile are regarded as having non-confirmable topologies, since these subclades branch deep within the tree. These groups are the parasites Chlamydiae and Spirochaetes, the Actinobacteria, the Cyanobacteria and Chlorobi. Therefore, topologies for each of the five species groups are, as observed, more heterogeneous (in comparison to the findings for the confirmable species groups Leptospiraceae, Buchnera, Rickettsia and Mollicutes) which led to extensions of the topology catalogue in Table 2. Interestingly, the parasitic Actinobacteria behave in a different way, and do not share a subclade with other parasites. Instead, within the section of phylogenies with parasitic monophyly, in the lower part of Figure 2, three out of the four possible topology states for the *Actinobacteria* are observed. Furthermore, all four topology states for the Actinobacteria are observed within the phylogeny inferences with correct topologies for the Leptospiraceae, Mollicutes, *Rickettsia* and *Buchnera* (see column for Actinobacteria in the grey bordered rectangles, Figure 2).

Topology relationships

The two parasite groups Chlamydiae and Spirochaetes do not belong to species groups with a confirmable topology. If the parasitic subclade is formed, both taxa are included in this subclade in most cases. If not, a conserved topology can be observed for these two species groups. They are placed together as sister clades. The respective topologies are supported by seven trees [tree 15; tree 16; tree 17; tree 18; tree 19; tree 27; tree 7]. According to our catalogue of topology alternatives, the Spirochaetes lie between the Chlamydiae and the Proteobacteria (alternatively Eukaryota and Archaea), and the Chlamydiae are, reciprocally, placed near the Spirochaetes.

The general relationship between two other species groups, the Actinobacteria and Cyanobacteria, is found in more than 50% of all phylogenies of the analyzed literature. The proximity of the Actinobacteria to the Cyanobacteria is observed in 26 phylogenies (status +1 or +2); this correlates with the reciprocal finding for the Cyanobacteria near the Actinobacteria (21 occurrences; status −1).

Conserved placements can also be found for both reciprocal relationships. Some phylogenies show conserved proximity for the Cyanobacteria to the Actinobacteria, together with conserved proximity for the Chlamydiae to the Spirochaetes. This intersect situation is apparent in nine phylogenies [tree 15; tree 16; tree 17; tree 19; tree 20; tree 21; tree 22; tree 23; tree 24]. For this particular combination, reciprocal relationships are found in four phylogenies as proximity of Spirochaetes to the Chlamydiae (overlap with the general observation above) and as proximity of the Actinobacteria to the Cyanobacteria, [tree 15; tree 16; tree 17; tree 19].

Ignored phylogenies

The following publications were excluded from further analysis. A single phylogenomic inference approach did generally not consider parasitic bacteria[33] [tree 1; tree 3]. Two early phylogenies [tree 2; tree 4] were too small, containing only 11 or 13 species.[17,20] Two recent publications considered inner tree uncertainty with non-resolved topologies for Spirochaetes, Chlamydiae, Actinobacteria, Cyanobacteria and Chlorobi,[27,37] [tree 13]. A single integrative study presented undecided topologies for bacterial symbionts from different sources [tree 14];[43] another paper considered only proteobacteria with 329 species.[74]

Some of the more recent phylogenies analyzed in this study, however, focused on a modeling of gene loss in a single super-kingdom of life, the bacteria [tree 14; tree 15; tree 23; tree 38; tree 40; tree 41; tree 42; tree 43].[25,26,28,36] Respective sparse phylogenies were not ignored because of their importance for our analysis.

SYSTERS-PhyloMatrix GCT inferences—topology profiles and event matrix

We derived a set of seven GCTs using the data model SYSTERS-PhyloMatrix.[7] The respective trees are available in the supplementary files as Figures S1 to S7. For the topology analysis, we used the catalogue of topologies, Table 2. The resulting topologies for the seven SYSTERS-PhyloMatrix GCTs across the nine taxa are presented in Table 3 as an event matrix; each of the seven rows is the respective topology profile to the tree. Hence, the seven tree topology profiles were constructed in the same way as that of published phylogenies.

The seven topology profiles fall into two groups. The first group of similar profiles was generated using NJ with the two distance metrics Simpson and Korbel. In contrast to the other five algorithmic approaches, the four confirmable species groups Mollicutes, *Rickettsia*, *Buchnera* and Leptospiraceae are correctly associated with their respective higher taxonomic ranks. However, the two profiles differ in four of the five characters in the section including the non-confirmable topologies, namely the Spirochaetes, Chlamydiae, Cyanobacteria and Actinobacteria. The topology of the Chlorobi is not clearly defined (Korbel) or they are placed within the Proteobacteria (Simpson).

The NJ algorithm with the Korbel distance (supplementary files, Fig. S1) supports the topology of the Chlamydiae, the proximity to the Spirochaetes, the reciprocal proximity of the Spirochaetes to the Chlamydiae, and the proximity of the Cyanobacteria to the Eukaryota and the Actinobacteria. This is very frequent in the majority of all published phylogenies. However, the Actinobacteria are observed in a seldom found topology state (‘other’).

The NJ algorithm with the Simpson metric (Fig. S2) supports the proximity of the Actinobacteria to the Cyanobacteria and the Eukaryota (and the Chlamydiae). Reciprocally, the Cyanobacteria are close to the Eukaryota (state +2). However, three cases of incongruence occur: Spirochaetes and Chlamydiae possess a topology that is seldom reproduced elsewhere in published phylogenies.

The next five SYSTERS-PhyloMatrix inferences possess a more or less shared topology for the five parasites (*Rickettsia*, *Buchnera*, Mollicutes, Spirochaetes, Chlamydiae). The Leptospiraceae are associated with species groups other than the Spirochaetes, even those that are parasites. We found them, moreover, in proximity to the Chlorobi, although this placement is not scientifically supported. All five SYSTERS-PhyloMatrix phylogenies should therefore be ignored. However, it is interesting that similar topology profiles can be observed for inferences after applying Wagner parsimony (Fig. S5), NJ with Dice metric (Fig. S6) or NJ with Jaccard metric (Fig. S7). According to the topology events, these three inferences support the strong proximity of Cyanobacteria and Actinobacteria reciprocally to each other, which is consistent with findings in the literature. The Dollo parsimony (Fig. S3) and NJ with Hamming distance (Fig. S4) place the Actinobacteria between the Firmicutes and Proteobacteria. Here, the Cyanobacteria are close to the Chlorobi (NJ with Hamming) or near the Eukaryota (Dollo). Such topologies are seldom or not supported in published phylogenies.

Combination of SYSTERS-PhyloMatrix GCTs and published phylogenies

Tree topology profiling

Figure 2 integrates the 54 phylogenies from the two sources: seven phylogenies were newly generated from the SYSTERS-PhyloMatrix data model and 47 were retrieved from the published literature, as described in Table 1. Clustering of the tree topology profiles orders the phylogenies in vertical direction and results in the presented heatmap picture. Reference data, numeric topology profiles and event matrices, were given in Table 2 and Table S2.

Subclusters of particular taxa and phylogenies after divisions

Three major subclusters were found in the heatmap after applying the two principal divisions. Vertical division 1 separates the species groups with confirmable topologies from those with non-confirmable. Division 2, applied several times, separates phylogenies that can be excluded (those with partially wrong or completely wrong topologies, eg, such with a shared parasitic monophyletic subclade).

Three subclusters reveal indicated by colored rectangles. The blue-bordered rectangle was correlated with the parasitic subclade in thirteen published and five of our own whole-genome inferences. However, division 1 did not divide the parasites since the Spirochaetes and Chlamydiae shared almost always the parasitic monophyletic subclade together with the other parasites with confirmable topology. In contrast, the Leptospiraceae were not observed in the parasitic subclade. Moreover, proximity to the Chlorobi (also to *Aquifex*) was suggested if they were not found in their correct monophyly with the Spirochaetes. This situation occurred in the five not well-supported SYSTERS-PhyloMatrix trees as well as in [tree 31].

The yellow-bordered rectangle in the heatmap comprises confirmable topologies that are correctly inferred for all four species groups together (if present in the tree). The corresponding phylogenies are of further interest in the section for non-confirmable topologies. This section is depicted as the gray-bordered rectangle; it is the complementary area to the yellow-bordered. Two phylogenies [tree 13; tree 14] were excluded since there was no information in that sector for non-confirmable topologies. The resulting set consisted of 21 published plus two SYSTERS-PhyloMatrix whole-genome phylogenies (NJ with Korbel and Simpson metrics). In contrast to the statement that corresponding topologies are not confirmable (and even if there were strong arguments for inner tree uncertainty) this subcluster is characterized by resolved topologies. To determine the inference background of these phylogenies, we performed a further analysis of the remaining 23 phylogenies.

Re-clustering

Re-clustering of the 21 published phylogenies and two SYSTERS-PhyloMatrix phylogenies across the five species groups (Spirochaetes, Chlamydiae, Cyanobacteria, Actinobacteria, Chlorobi; gray rectangles in Fig. 2) revealed a second heatmap, Figure 3. This consists, according to the dendrogram, of five subgroups that are indicated using numbers and colored circles.

Subgroup 1 (blue bar in Fig. 3) consists of the SYSTERS-PhyloMatrix GCT inferred with NJ and the Simpson metric (Fig. S2). It is accompanied by another GCT [tree 26],[23] based on COGs and inferred with NJ and the Korbel distance metric, and a supertree [tree 27].[39]

Subgroup 2 (green) consists of the SYSTERS-PhyloMatrix GCT inferred with NJ and the Korbel distance (Fig. S1) and three further GCTs [tree 16; tree 17; tree 19],[19,23,56] a gene order tree [tree 18],[14] and a MSA-ML tree [tree 15].[36] The underlying data models for the respective phylogenies were derived from the COGs, ORF-based reciprocal best hits, or sequences of housekeeping genes (MSA). The algorithms are all distance-based heuristics, two of them are using the Korbel distance metric as in the SYSTERS-PhyloMatrix GCT in this subcluster. This subcluster supports the reciprocal topology annotation of the two pairs of sister clades, the connections of the Spirochaetes to the Chlamydiae and the Actinobacteria to the Cyanobacteria.

The other three subgroups (3 to 5) comprise GCTs inferred using distance methods on the basis of the COGs or separate ORF inferences [tree 20; tree 21; tree 22; tree 23; tree 24; tree 25] (subgroup 3, orange). Here the topologies of the Actinobacteria and Cyanobacteria are reciprocally adjacent, unlike those of the Spirochaetes and Chlamydiae. Subgroup 4 (red) combines one super-tree and three MSA-ML/MP phylogenies [tree 6; tree 7; tree 8; tree 9]; subgroup 5 (pink) consists of inferences by distance-based heuristics from totally different data models [tree 10; tree 11; tree 12]. The topology relationships in the latter two subgroups are not frequently observed.

The following general trends were observed:

i. In Figure 2: the subcluster of 21 published +2 SYSTERS-PhyloMatrix phylogenies, shown within gray-bordered rectangles, includes many inference results that were regarded by the authors as improvements compared with initial set-ups. (An improvement can also be a phylogeny in comparison to a given 16S rRNA tree.)

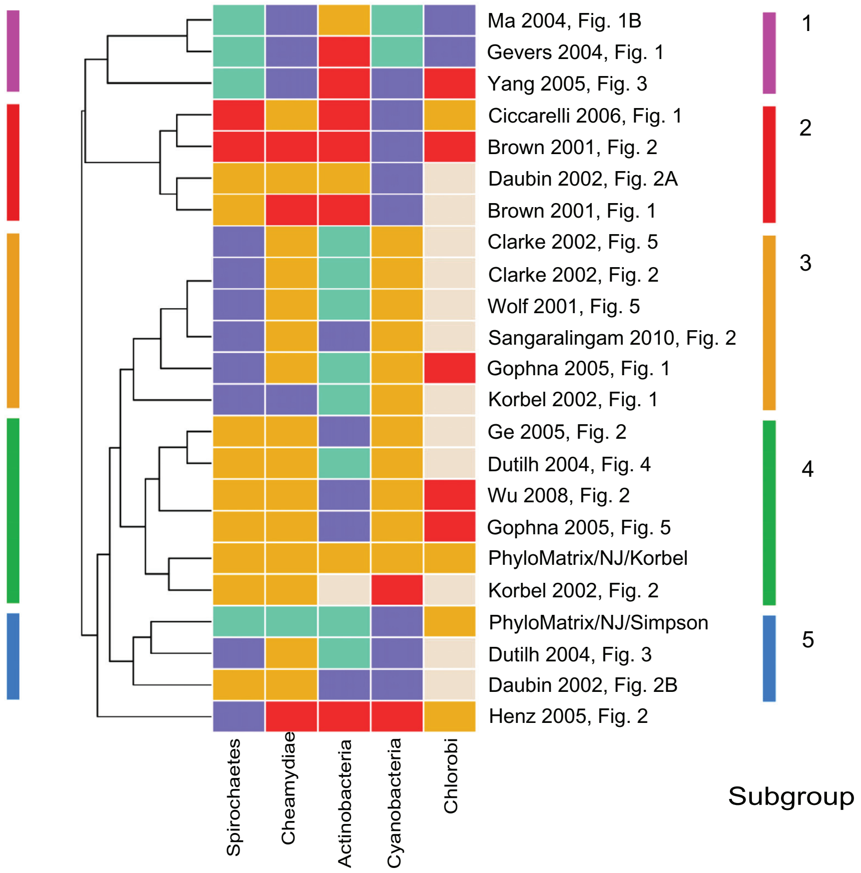


Figure 3. Subcluster from Figure 2 of well supported topology alternatives.

Notes: Five subgroups (1 to 5, colored bars) result from the clustering according to the dendrogram. Heatmap color definitions for up to four topology alternatives of respective taxa are given in Table 2 (light gray: species not regarded in respective publication). Particular topology states (in event matrices) are given in Table 3 and supplemental data Table S2.

ii. In Figure 3: all MSA-ML or MSA-MP phylogenies are included in the set of 23 reliable results. One of the subclusters consists exclusively of these phylogenies.

iii. Three other subclusters comprise phylogenies inferred with distance-based methods. Here, the most frequently applied metric is the Korbel metric.

iv. The Jaccard metric was frequently used for distance-based inferences that led to (not always completely) wrong topologies for the four confirmable parasite groups.

v. Reciprocally adjacent topologies are apparent for Actinobacteria and Cyanobacteria (frequently in subgroup 2 and significant in subgroup 3, Fig. 3 and even in many phylogenies with a shared subclade for all parasites in Fig. 2) and for Spirochaetes and Chlamydiae (significant in subgroup 2, Fig. 3).

Discussion

A new methodology for comparing phylogenies with different inference backgrounds

The topology of a species group (taxon) in a phylogeny can be described by proximities to other taxa. Topology alternatives occur if different topologies for a single taxon are determined in multiple publications. For a given set of taxa we denote the set of particular topology alternatives as ‘topology profile’ and introduce ‘tree topology profiling’ as a comparative method based on such profiles. These profiles are the translation of verbal and topological descriptions into a scoring. Formally similar to phylogenetic profiling, tree topology profiling enables the computational processing and visualization of complex phylogenomic contexts. The use of topology alternatives does not require a ‘true topology’ assumption—which could be determined by an acknowledged taxonomy or any scientifically supported knowledge such as that obtained via molecular biology. However, it is meaningful and useful for further interpretation if a particular state represents a scientifically accepted topology.

In general, the character set in a topology profile allows the independent and unbiased comparison of any whole-genome phylogeny from any source. This is more important than the potential disadvantage stemming from the empirical decision to use the 'most plausible' character status set. The transcription into digital states (scores), furthermore, enables semi-quantitative referral of the phylogeny inference results to the data background, the data model and the inference methodologies. It is clear to us that the chosen status index set across the entire topology profile is empirical. The clustering result depends, of course, on the number of characters in the topology profile (which is relatively small with a profile length of nine characters, depending on the data available in the literature) and on the appropriate choice of character states. For this reason, different index sets were compared with the inference results beforehand (data not shown). The presented approach, however, is accurate enough to highlight and elucidate the situation in published inner tree topologies of the ToL.

A topology catalogue for further general bacterial subclades can be created analogously to the presented set of taxa, followed by a similar analysis. More species groups exist in the inner tree, such as the Aquificae, Deinococci, Fusobacteria and Thermotogae. Extending the topology profile by appending these taxa to the character set would be conducive to understanding early evolution. However, like the Leptospiraceae and the Chlorobi, most of them are missing from the majority of the analyzed literature, especially in early publications. These species were therefore ignored here. Greater success would be expected if the fine-grained topology of a particular subclade at the periphery of the ToL, such as the Firmicutes or the Proteobacteria, were to be analyzed using tree topology profiling. Here, success will depend on the quantity of data (numbers of species and published trees) and variability in the existing literature. Our use of topology alternatives can be regarded as a semi-quantitative quality assessment of whole-genome phylogenies. This study thus has two main results, first, the general introduction of tree topology profiling and second, specific evidence of topologies in the inner ToL.

Progress in ToL inferences

Progress in recent whole-genome phylogeny inferences can be attributed to several aspects, for phylogenies based on gene content as well as on other data models. The first is the dramatic increase in the number of completely sequenced species represented in a ToL—the most recent dataset in this review contained more than 1100 species.[37] This challenges inference methods and computational performance. Second, the methodology for appropriate inferences of whole-genome phylogenies is moving away from GCTs to MSA-ML trees. Some GCT inference approaches can keep up with more recent techniques. For MSA-ML phylogenies, the overwhelming increase in the amount of whole-genomes data is obviously compensated by reduction in numbers of considered gene families; the key here is the selection of representative, eg, housekeeping genes. The third aspect is that recent inner trees, in contrast to earlier trees, are displayed as non-bifurcating topologies. This is due to the trend for not all gene families to be included in tree inferences as well as the widely discussed tree-unlikeness (which reflects the early bacterial evolution obviously better, as already discussed in the literature). All newer inferences have enhanced the resolution of whole-genome phylogenies at the periphery of the ToL; however, the inner tree topology remains an open question.

The notion of the 'Tree' of Life has been widely questioned in the light of rampant HGT[75] and gene loss, despite the fact that a tree can be derived in the presence of HGT by conditioned reconstruction.[55] As already mentioned, HGT events have a natural impact on GCT phylogenies. Focusing GCT inference on the Eukaryotes led to considerations on how these species arose from their prokaryotic ancestors and further to the frequently discussed 'ring of life'[76,77] or the 'network of life'.[78] Inconsistencies and methodological limitations have been reviewed and discussed.[79] An estimate was made of how much branch attraction, which depends on tree inference techniques, affects the accuracy of the inner tree.[80,81] As a result, and after introducing appropriate models,[82] the root of the tree of life[83,84] remains under debate, including the monophyly of particular clades of prokaryotes.[85]

Automatically inferred whole-genome phylogenies exhibit sparsely supported inner trees. In more recently published phylogenies, such uncertainty in near-root topologies is indicated by short branch lengths, dashed inner-tree edges, low bootstrap supports or lack of resolution (avoiding bifurcations). This is consistent with further analyses of our own data. When applying the average bootstrap support method,[86] the most well-accepted, confirmable topologies showed the lowest support (data not shown). For published and our own phylogenies, we could not find a larger conserved and generally admitted topology arrangement for taxa with non-confirmable topologies. This observation was the cause of the discussion in the literature[76,78,82,85] regarding tree-unlikeness and reliability. We hence denoted such topologies as 'not confirmable'.

Species with confirmable topologies

Inner tree topologies frequently show the subclade shared by the majority of parasitic species, which is clearly wrong at least for some of them. Such topologies were also observed in more recent publications that were intended to model the effects of gene loss[25,26,28] in the Leptospiraceae, *Buchnera*, *Rickettsia* and Mollicutes. In accordance with the supporting literature, correct topologies were reproduced by phylogenies inferred with more recent techniques, including particular (gene) content approaches using data models such as the COGs or SYSTERS-PhyloMatrix. Here, distance methods using the Korbel or Simpson metric were found to be successful.

Species with non-confirmable topologies

The trend in newer publications towards presenting non-bifurcations for the major bacterial subclades in the inner ToL is less informative but obviously more correct. The respective five inner tree species groups that belong to our analysis are the parasites Chlamydiae, Spirochetes and Actinobacteria as well as the bacterial subclades of the Chlorobi and Cyanobacteria. In contrast to non-bifurcating (unresolved) paraphyly for these five bacterial subclades, for example [tree 13],[27] most of the phylogenies analyzed here explicitly show bifurcations in the inner tree and, therefore, suggest obvious resolution in that region of the ToL. This contrast was one of the factors stimulating the study presented here.

Our findings regarding the proximity between two taxa (Actinobacteria and Cyanobacteria; Spirochetes and Chlamydiae) is shown. Here, the high likeliness of true description of evolution should not be seen under the aspect of majority (which is suggested by a heatmap picture like the presented) but more under the aspect of congruence of similar results coming from several well-performing and accepted methods.

Uncertain near-root topology for the three subclades of the Cyanobacteria, the Chlorobi and the Eukaryota does not generally rule out their possible proximity to each other. Uncertainty is in line with the report that the reliable placement of the Cyanobacteria in the whole-genome topology is 'somewhat difficult'.[33] Proximity of the Cyanobacteria to the Eukaryota is not a separate character state in our catalogue (Table 2). It is indirectly included in the dedicated states. We found this situation ten times among the 21 well-supported parasites phylogenies in the literature (in state −1 between Actinobacteria and Eukaryota and also in state +2 between Proteobacteria and Eukaryota). The fact that *Chlorobium tepidum*, the only evaluable representative of the Chlorobi, is a green sulfur bacterium could be connected with the evolutionary proximity to other photosynthetically active species, as reported elsewhere.[87] The proteome of *Chlorobium* comprises constituents of the photosynthetic apparatus similar to the proteomes of the Cyanobacteria and plants.[88,89] However, there is little evidence for its proximity to Cyanobacteria. Only fourteen newer phylogenies include the Chlorobi. The respective reciprocal events (Chlorobi: near Cyanobacteria, status +2; Cyanobacteria: near Chlorobi, status +1; according to our catalogue) occur six times in total and only twice in reliable phylogenies with correct topologies of the confirmable parasites. The Chlorobi are never observed in the proximity of the Eukaryota. As a consequence, any shared functionality, here in the form of genes with photosynthetic function, is not the driving force for the suggested proximity of the Chlorobi, Eukaryotes or Cyanobacteria.

GCT inference alternatives and the ToL

Researchers have made many efforts to optimize the combination of data model, data conditioning and inference methods. We show that tree topology profiling on confirmable and non-confirmable topologies can provide additional insights.

Our search of the relevant literature revealed that acceptable phylogenies are only produced by particular settings. These consist of the main data models such as MSAs (in combination with ML or MP) or phylogenetic trees as the basis for super-trees and also gene content data. For the latter, however, only a fraction of phylogenies was successfully inferred. The fact that GCTs derived from homology data models (including SYSTERS-PhyloMatrix) are present in both main parts of the heatmap, (ie, phylogenies with gray-bordered subcluster versus phylogenies with blue bordered subcluster; as referred to Fig. 2), indicates the tolerance of the content data model in general. We hypothesize therefore that the data model is not the critical factor for whole-genome phylogeny inferences.

As a consequence of this hypothesis, the methodology should appear to limit the inference of gene content phylogenies. We can exclude a number of inference methods connected with content data as the cause of obviously wrong results. Along with this, we can conclude that phylogenies comprising a subclade shared by all parasites—which is clearly wrong—are based on an artifact that is caused by the only alternative, the inference methodology. The method variation based on the SYSTERS-PhyloMatrix data model comprises also several algorithms and distance metrics that reveal wrong results: our parsimony approaches do not give reliable results, here; some of our own distance-based phylogenies (NJ with Jaccard, Dice or Hamming metric) are similarly un-reliable, like those results that were inferred with other content data such as the COGs.

In this context it should be noted that gene content calculation means counting shared gene families. Shared status numbers have to be normalized with the genome sizes of the two genomes considered. For a discussion of the latter issue, see the overview of the influence of several similarity metrics on calculated similarity by Cheetham and Hazel.[71] Similarity metrics possess the function as a normalization factor to compare the shared families in a standardized manner. Normalization is, mathematically, performed by the denominator in the similarity calculation. In most similarity metrics, the denominator includes both genome sizes, which means that the denominator is not a constant comparing eg, a small genome with several larger genomes. However, there are similarity metrics that tend to be a constant factor for large genome size differences. The denominator of the Simpson metric is *per se* a constant since it is the size of the smaller genome. The Korbel metric begins to behave in that way if one genome is about three times larger (or more) than the other genome; for large size differences, such normalization behaves like the Simpson metric. The opposite effect occurs if a similarity metric retains the mathematical dependency from the two genome sizes. Such behavior is covered by the Jaccard, the Dice and the Hamming metrics, which, using tree topology profiling in the presented meta-analysis, were shown to infer wrong topologies. Thus, the absence of genes after reductive evolution is an important factor for similarity calculations (which led in fact to a reduction in genome size). This is especially observed when comparing a small (eg, a parasitic) genome with a larger one. For distance-based methods, we therefore simulated the behavior of several distance metrics depending on the ratio of the sizes of two genomes (data not shown). Observations here suggest that the Simpson and Korbel metrics have a balancing effect at higher genome size differences, in contrast to the Jaccard, Dice and Hamming metrics. The balancing effect is consistent with our observations from the clustering result that particular gene content inferences are similar to inferences from more sophisitcated methodological approaches. Here, two combinations lead to successfully inferred SYSTERS-PhyloMatrix phylogenies: NJ with the Simpson metric and NJ with the Korbel metric. Especially the latter has a good reputation in other published phylogeny inferences.

The topologies of *Leptospira* and the parasitic Actinobacteria are different in comparison to the other parasites since they are not found in a shared parasitic subclade. This can be explained by their larger genome sizes (~1000 genes and more) versus those of the other parasitic species groups (~300 to 500 genes; see supplemental data Table S3). Consultation of a number of published phylogenies (including our own) reveals that the Simpson and Korbel metrics have the balancing effect; since the genome sizes differ by a factor of about 3 or more, and the respective inference results are reliable.

Distance-based approaches validate gene quantities according to gene presence and, indirectly, gene absence. The main difference to all other inference techniques is that these involve quantitative analyses of the properties of genes that are always present. Here, recent improvements in inference methodologies, for instance phylogenies of concatenated protein sequence alignments [tree 6; tree 15],[34,36] produce similar results that make concordant gene content phylogenies reliable. However, differences in particular non-confirmable topologies remain.

Beyond the here presented phylogenies, several phylogeny inferences in the literature are related to data models that have varying definitions of orthology or are restricted to a particular, eg, the eukaryotic, subclade. For instance, the fungal phylogeny was assessed using a range of methodological approaches,[90] and an optimum was reported for the super-alignment approach combined with restrictive orthology as data model. Another study, restricted to the taxonomic subclade of the *Archaea*, revealed the systematic bias of a conditioned reconstruction method compared to other frequently used approaches like super-trees, concatenated alignments, 16S rRNA trees or distance-based methods[58] with the result 'that genome phylogenies need to be interpreted differently, depending on the method used to construct them'.

GCTs based on gene content data models

A large number of the whole-genome phylogenies analyzed in this study were derived from content data. Depending on the inference method, all major data model approaches led to successful results, in particular the COGs with its derivates and the SYSTERS-PhyloMatrix as an un-modified data model.

The SYSTERS-PhyloMatrix is a homology-based data model analogous to the widely accepted COGs. Using SYSTERS as the underlying protein family set resulting from a dynamic hierarchical clustering[11] indirectly considers the heterotachy of protein families. This is an advantage that probably compensates for iterative optimization through more than 100 tree inferences, as excessively done and validated by 'reciprocal illumination' elsewhere.[16] From our inferences of gene content phylogenies and the presented analysis we conclude that both data models, the COGs and SYSTERS-PhyloMatrix, are directly comparable in terms of phylogeny inference results.

Using the SYSTERS-PhyloMatrix has the following advantages and consequences. First, SYSTERS-PhyloMatrix phylogenies are traceable in terms of the data model and inference methods. Using the SYSTERS-PhyloMatrix for whole-genome phylogeny inference, moreover, ensures that the (transformed) molecular sequence data basis is identical for all inference variations. Second, the SYSTERS-PhyloMatrix inference results are found within two subgroups of the subcluster in Figure 3, with well-supported whole-genome phylogeny inference results. Thereby, it is shown that the SYSTERS-PhyloMatrix data model is useful for the inference of a biologically meaningful ToL. Third, even the underlying sequence data basis itself, the SYSTERS homology inference model and the resulting protein family set, appears to be biologically meaningful.

Conclusion

With this study, we introduced a strategy for the comparison of published whole-genome phylogenies that differ extremely in tree size and inference background. Respective raw or meta-data are not or only seldom available for retracing such phylogenies by a controlled and computationally supported comparison approach. Therefore we developed a strategy for the manual validation of drawn phylogenies: tree topology profiling. In particular, we analyzed general bacterial subclades in 47 whole-genome phylogenies from 30 publications with respect to inner tree topologies. We enlarged the analysis with findings from seven own gene content phylogeny inferences based on the SYSTERS-PhyloMatrix data model. In total, more than 400 topology alternatives for nine analyzed species groups and more than 50 phylogenies are presented in tree topology profiles.

The presented tree topology profiling was applied on bacterial clans that were found to branch in the inner tree. Using this approach we first of all can separate the widely-spread artifact in published phylogenies, the shared subclade of parasitic bacteria, from topologies supported by other sophisticated methods. True topologies in a whole-genome phylogeny had been substantially shown for the eubacteria Buchnera, Rickettsia and Mollicutes. Artificial topologies for these bacteria as well as for Chlamydiae and Spirochaetes were predominantly found for particular gene content phylogenies generated with distance-based heuristics. Especially, distance metrics have more or less sensitive influence on the inference result. We give hints that small genome sizes are evident for this behavior.

Our own data model as well as other gene content data models produced good inference results with appropriate inference methods such as Neighbor Joining with Simpson or Korbel metrics. Our findings showed that the SYSTERS-PhyloMatrix gene content trees, along with the SYSTERS protein family set, are biologically meaningful. Other approaches such as gene order or multiple sequence alignments exploit, notably, the quality of present molecular information. In contrast, gene content methods validate also information of gene absence, which might cause the observed aberrances. The re-analysis revealed evidence for two general findings across a number of different phylogeny inference methods, the reciprocal proximity of the Chlamydiae to the Spirochaetes, and the reciprocal proximity of the Actinobacteria to the Cyanobacteria.

With this paper we were demonstrating the connection between topology information and particular inference parameters. The presented strategy, tree topology profiling, can moreover be used as a template for analogous studies in the scientific field of comparative phylogenomics.

Supplementary Data

Table S1: Set of 25 species that are known for gain and loss of gene families. Table S2: Event matrix for the validated gene content trees from literature for nine species groups. Table S3: Set of 106 species in SYSTERS-PhyloMatrix ordered by protein family size. Figures S1 to S7: Seven SYSTERS-PhyloMatrix gene content phylogenies.

Supplementary Files

Figures S1 to S7: Seven SYSTERS-PhyloMatrix gene content phylogenies in Nexus file format.

Author Contributions

Conceived and designed the experiments: TM, AK. Analyzed the data: TM. Wrote the first draft of the manuscript: TM. Contributed to the writing of the manuscript: TM. Agree with manuscript results and conclusions: TM, AK. Jointly developed the structure and arguments for the paper: TM, AK. Made critical revisions and approved final version: TM, AK. All authors reviewed and approved of the final manuscript.

Abbreviations

CA, Correspondence Analysis; COGs, Clusters of Orthologous Groups of proteins; eggNOG, evolutionary genealogy of genes: nonsupervised orthologous groups; FM, Fitch-Margoliash (algorithm); GCT, gene content tree; HGT, horizontal (syn. lateral) gene transfer; ML, maximum likelihood; MP, maximum parsimony; MSA, multiple sequence alignment; NJ, Neighbor Joining (algorithm); ORF, open reading frame; rRNA, ribosomal ribonucleic acid; SHOT, Shared Ortholog and Gene Order Tree Reconstruction Tool; SYSTERS, SYSTEmatic Re-Searching (algorithm); ToL, Tree of Life; UPGMA, Unweighted Pair Group Method with Arithmetic Mean.

Acknowledgements

The authors wish to express their gratitude to Hannes Luz for many fruitful discussions in the early phase of the study. Our colleague and friend passed away very suddenly in June 2011. The authors wish to thank Eike Staub for valuable hints and initial suggestions as well as Martin Vingron for provision of the computer capacities in the Computational Molecular Biology Department at the Max Planck Institute in Berlin-Dahlem. Initial parts of this work have been presented on the conference ‘Celebrating Darwin: From The Origin of Species to Deep Metazoan Phylogeny’, DMP 2009, Berlin, Germany.[91]

Funding

TM’s grant was supported by the BMBF (Federal Ministry of Education and Research of Germany), MedSys project ‘Therapeutic Systems Immunology’ (0315450A).

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Darwin CR. Notebook B. Transmutation of Species (1837–1838). http://darwin-online.org.uk/content/frameset?viewtype=side&itemID=CUL-DAR121.-&pageseq=38. Updated Jun 24, 2011. Accessed Jan 10, 2012.
2. Koonin EV, Wolf YI. Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World. *Nucleic Acids Res*. 2008;36(21):6688–719.
3. House CH. The Tree of Life Viewed through the Contents of Genomes. In: Boekels Gogarten M, Gogarten JP, Olendzenski LC, editors. *Horizontal Gene Transfer: Genomes in Flux*. Humana Press; 2009;141–61.
4. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*. 1999;96(8):4285–8.
5. Gaasterland T, Ragan MA. Constructing multigenome views of whole microbial genomes. *Microb Comp Genomics*. 1998;3(3):177–92.
6. Tekaia F, Yeramian E. Genome trees from conservation profiles. *PLoS Comput Biol*. 2005;1(7):e75.
7. Meinel T, Krause A, Luz H, Vingron M, Staub E. The SYSTERS protein family database in 2005. *Nucleic Acids Res*. 2005;33(Database issue):D226–9.
8. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science*. 1997;278(5338):631–7.
9. Tatusov RL, Fedorova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 2003;4:41.
10. Enright AJ, Kunin V, Ouzounis CA. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res*. 2003;31(15):4632–8.
11. Krause A, Stoye J, Vingron M. Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics*. 2005;6:15.
12. Ma HW, Zeng AP. Phylogenetic comparison of metabolic capacities of organisms at genome level. *Mol Phylogenet Evol*. 2004;31(1):204–13.
13. Fitz-Gibbon ST, House CH. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res*. 1999;27(21):4218–22.
14. Korbel JO, Snel B, Huynen MA, Bork P. SHOT: a web server for the construction of genome phylogenies. *Trends Genet*. 2002;18(3):158–62.
15. Hughes AL, Ekollu V, Friedman R, Rose JR. Gene family content-based phylogeny of prokaryotes: The effect of criteria for inferring homology. *Syst Biol*. 2005;54(2):268–76.
16. Lienau EK, DeSalle R, Rosenfeld JA, Planet PJ. Reciprocal illumination in the gene content tree of life. *Syst Biol*. 2006;55(3):441–53.
17. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. *Nat Genet*. 1999;21(1):108–10.
18. Clarke GD, Beiko RG, Ragan MA, Charlebois RL. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized blastp scores. *J Bacteriol*. 2002;184(8):2072–80.
19. Gophna U, Doolittle WF, Charlebois RL. Weighted genome trees: refinements and applications. *J Bacteriol*. 2005;187(4):1305–16.
20. Lin J, Gerstein M. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res*. 2000;10(6):808–18.
21. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol*. 2001;1:8.
22. Wolf YI, Rogozin IB, Grishin NV, Koonin EV. Genome trees and the tree of life. *Trends Genet*. 2002;18(9):472–9.
23. Dutilh BE, Huynen MA, Bruno WJ, Snel B. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J Mol Evol*. 2004;58(5):527–39.
24. Gu X, Zhang H. Genome phylogenetic analysis based on extended gene contents. *Mol Biol Evol*. 2004;21(7):1401–8.
25. Spencer M, Susko E, Roger AJ. Modelling prokaryote gene content. *Evol Bioinform Online*. 2006;2:157–78.
26. Spencer M, Sangaralingam A. A phylogenetic mixture model for gene family loss in parasitic bacteria. *Mol Biol Evol*. 2009;26(8):1901–8.
27. Muller J, Szklarczyk D, Julien P, et al. eggNOG V2.0: Extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res*. 2010; 38(Database issue): D190–5.
28. Sangaralingam A, Susko E, Bryant D, Spencer M. On the artefactual parasitic eubacteria clan in conditioned logdet phylogenies: heterotachy and ortholog identification artefacts as explanations. *BMC Evol Biol*. 2010; 10:343.
29. Gevers D, Vandepoele K, Simillon C, Van de Peer Y. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol*. 2004;12(4):148–54.
30. Hong SH, Kim TY, Lee SY. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Appl Microbiol Biotechnol*. 2004;65(2):203–10.
31. McInerney JO, Wilkinson M. New methods ring changes for the tree of life. *Trends Ecol Evol*. 2005;20(3):105–107.
32. Yang S, Doolittle RF, Bourne PE. Phylogeny determined by protein domain content. *Proc Natl Acad Sci U S A*. 2005;102(2):373–8.
33. Deeds EJ, Hennessey H, Shakhnovich EI. Prokaryotic phylogenies inferred from protein structural domains. *Genome Res*. 2005;15(3):393–402.
34. Ciccarelli FD, Doerks T, von Mering C, et al. Toward automatic reconstruction of a highly resolved tree of life. *Science*. 2006;311(5765):1283–7.
35. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. Universal trees based on large combined protein sequence data sets. *Nat Genet*. 2001; 28(3):281–5.
36. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*. 2008;9(10):R151.
37. Powell S, Szklarczyk D, Trachana K, et al. eggNOG V3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res*. 2012;40(D1):D284–9.
38. Gogarten JP, Fournier G, Zhaxybayeva O. Gene transfer and the reconstruction of life's early history from genomic data. *Space Sci Rev*. 2008;135:115–31.
39. Daubin V, Gouy M, Perriere G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res*. 2002;12(7):1080–90.
40. Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. *Trends Genet*. 2001;17(10):589–96.
41. Blanc G, Ogata H, Robert C, et al. Reductive genome evolution from the mother of rickettsia. *PLoS Genet*. 2007;3(1):e14.
42. Moran NA, Mira A. The Process of genome shrinkage in the obligate symbiont buchnera aphidicola. *Genome Biol*. 2001;2(12):RESEARCH0054.
43. Moran NA, McCutcheon JP, Nakabachi A. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet*. 2008;42:165–90.
44. Zakharov IA, Markov AV. Gene orders in genomes of alpha-proteobacteria: similarity and evolution. *Genetika*. 2005;41(12):1624–33.

45. Wolf M, Muller T, Dandekar T, Pollack JD. Phylogeny of firmicutes with special reference to mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. *Int J Syst Evol Microbiol*. 2004; 54(Pt 3):871–5.
46. Paster BJ, Dewhirst FE. Phylogenetic foundation of spirochetes. *J Mol Microbiol Biotechnol*. 2000;2(4):341–4.
47. Olsen I, Paster BJ, Dewhirst FE. Taxonomy of spirochetes. *Anaerobe*. 2000; 6(1):39–57.
48. Gupta RS. Protein signatures distinctive of alpha proteobacteria and its subgroups and a model for alpha-proteobacterial evolution. *Crit Rev Microbiol*. 2005;31(2):101–35.
49. Gillespie JJ, Brayton KA, Williams KP, et al. Phylogenomics reveals a diverse rickettsiales type iv secretion system. *Infect Immun*. 2010;78(5):1809–23.
50. Kunin V, Ahren D, Goldovsky L, Janssen P, Ouzounis CA. Measuring genome conservation across taxa: divided strains and United Kingdoms. *Nucleic Acids Res*. 2005;33(2):616–21.
51. Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Res*. 2006;16(9):1099–1108.
52. Wolf YI, Aravind L, Koonin EV. Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. *Trends Genet*. 1999;15(5): 173–5.
53. Figge RM, Cerff R. GAPDH gene diversity in spirochetes: a paradigm for genetic promiscuity. *Mol Biol Evol*. 2001;18(12):2240–9.
54. Kawase T, Saito A, Sato T, et al. Distribution and phylogenetic analysis of family 19 chitinases in actinobacteria. *Appl Environ Microbiol*. 2004;70(2): 1135–44.
55. Lake JA, Rivera MC. Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned reconstruction. *Mol Biol Evol*. 2004; 21(4):681–90.
56. Ge F, Wang LS, Kim J. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol*. 2005;3(10):e316.
57. Cohen O, Pupko T. Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol Biol Evol*. 2010;27(3): 703–13.
58. McCann A, Cotton JA, McInerney JO. The tree of genomes: An empirical comparison of genome-phylogeny reconstruction methods. *BMC Evol Biol*. 2008;8:312.
59. Hao W, Golding GB. The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res*. 2006;16(5):636–43.
60. Hao W, Golding GB. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics*. 2008;9:235.
61. Le Quesne WJ. The uniquely evolved character concept and its cladistic application. *Syst Biol*. 1974;23(4):513–7.
62. Farris JS. Phylogenetic analysis under Dollo's law. *Syst Zool*. 1977;26(1): 77–88
63. Snel B, Huynen MA, Dutilh BE. Genome trees and the nature of genome evolution. *Annu Rev Microbiol*. 2005;59:191–209.
64. Felsenstein J. Phylip—phylogeny inference package. *Cladistics*. 1989;5: 164–6.
65. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4(4):406–25.
66. Fitch WM, Margoliash E. Construction of phylogenetic trees. *Science*. 1967;155:279–84.
67. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*. 1958;38:1409–38.
68. Gascuel O. BIONJ: An improved version of the nj algorithm based on a simple model of sequence data. *Mol Biol Evol*. 1997;14(7):685–95.
69. Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC. Whole-genome prokaryotic phylogeny. *Bioinformatics*. 2005;21(10):2329–35.
70. Grishin NV, Wolf YI, Koonin EV. From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res*. 2000;10(7):991–1000.
71. Cheetham AH, Hazel JE. Binary (Presence-absence) similarity coefficients. *Journal of Paleontology*. 1969;43(5):1130–6.
72. Gentleman RC, Carey VJ, Bates DM et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80.
73. Perriere G, Gouy M. WWW-query: An on-line retrieval system for biological sequence banks. *Biochimie*. 1996;78(5):364–9.
74. Kloesges T, Popa O, Martin W, Dagan T. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol Biol Evol*. 2011;28(2):1057–74.
75. Doolittle WF. Phylogenetic classification and the universal tree. *Science*. 1999;284(5423):2124–9.
76. Rivera MC, Lake JA. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*. 2004;431(7005):152–5.
77. Simonson AB, Servin JA, Skophammer RG, et al. Decoding the genomic tree of life. *Proc Natl Acad Sci U S A*. 2005;102 Suppl 1:6608–13.
78. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. The net of life: reconstructing the microbial phylogenetic network. *Genome Res*. 2005;15(7): 954–9.
79. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*. 2005;6(5):361–75.
80. Philippe H, Germot A. Phylogeny of eukaryotes based on ribosomal Rna: long-branch attraction and models of sequence evolution. *Mol Biol Evol*. 2000;17(5):830–4.
81. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol*. 2005;54(5):743–57.
82. Lopez P, Forterre P, Philippe H. The root of the tree of life in the light of the covarion model. *J Mol Evol*. 1999;49(4):496–508.
83. Philippe H, Forterre P. The rooting of the universal tree of life is not reliable. *J Mol Evol*. 1999;49(4):509–23.
84. Forterre P, Philippe H. Where Is the root of the universal tree of life? *Bioessays*. 1999;21(10):871–9.
85. McInerney JO, Cotton JA, Pisani D. The prokaryotic tree of life: past, present... And future? *Trends Ecol Evol*. 2008;23(5):276–81.
86. Huson DH, Steel M. Phylogenetic trees based on gene content. *Bioinformatics*. 2004;20(13):2044–9.
87. Rujan T, Martin W. How many genes in arabidopsis come from cyanobacteria? An estimate from 386 protein phylogenies. *Trends Genet*. 2001;17(3): 113–20.
88. Xiong J, Fischer WM, Inoue K, Nakahara M, Bauer CE. Molecular evidence for the early evolution of photosynthesis. *Science*. 2000;289(5485): 1724–30.
89. Mulkidjanian AY, Koonin EV, Makarova KS, et al. The cyanobacterial genome core and the origin of photosynthesis. *Proc Natl Acad Sci U S A*. 2006;103(35):13126–31.
90. Dutilh BE, van Noort V, van der Heijden RT, et al. Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics*. 2007;23(7):815–24.
91. Loesel R. 150 Years Beyond Darwin's Origin of Species: Finding new approaches to reconstruct early animal phylogeny. *Biol Lett*. 2009;5(4): 436–8.
92. Tekaia F, Lazcano A, Dujon B. The genomic tree as revealed from whole proteome comparisons. *Genome Res*. 1999;9(6):550–7.
93. Andreeva A, Howorth D, Brenner SE, et al. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*. 2004;32(Database issue):D226–9.

Supplementary Data

Table S1. Set of 25 species that are known for gain and loss of gene families. Identification by the NCBI Taxonomy and the UniProt HAMAP systematic; provided is also information about gene loss analyses elsewhere.

Species	Grouping to the taxonomic ranks of genus, family, class, or phylum	NCBI TaxID	UniProt code	Feature analyzed in literature*
Bifidobacterium longum	Actinobacteria	216816	BIFLO	
Corynebacterium efficiens	Actinobacteria	152794	COREF	
Corynebacterium glutamicum	Actinobacteria	1718	CORGL	
Mycobacterium leprae	Actinobacteria	1769	MYCLE	Gene loss
Mycobacterium tuberculosis	Actinobacteria	1773	MYCTU	
Streptomyces coelicolor	Actinobacteria	1902	STRCO	
Buchnera aphidicola (Acyrthosiphon pisum)	Buchnera	118099	BUCAI	Gene loss
Buchnera aphidicola (Schizaphis graminum)	Buchnera	98794	BUCAP	
Chlamydia muridarum	Chlamydia	83560	CHLMU	
Chlamydia trachomatis	Chlamydia	813	CHLTR	Gene loss
Chlamydophila pneumoniae	Chlamydia	83558	CHLPN	Gene loss
Chlorobium tepidum	Chlorobia	1097	CHLTE	
Nostoc sp. PCC 7120	Cyanobacteria	103690	ANASP	
Synechococcus elongatus	Cyanobacteria	32046	SYNEL	
Synechocystis sp. PCC 6803	Cyanobacteria	1148	SYNY3	
Leptospira interrogans	Leptospiraceae	173	LEPIN	
Mycoplasma genitalium	Mollicutes	2097	MYCGE	Gene loss
Mycoplasma penetrans	Mollicutes	28227	MYCPE	
Mycoplasma pneumoniae	Mollicutes	2104	MYCPN	Gene loss
Mycoplasma pulmonis	Mollicutes	2107	MYCPU	Gene loss
Ureaplasma parvum	Mollicutes	134821	UREPA	Gene loss
Rickettsia conorii	Rickettsia	781	RICCN	Gene loss
Rickettsia prowazekii	Rickettsia	782	RICPR	Gene loss
Borrelia burgdorferi	Spirochaetes	139	BORBU	Gene loss
Treponema pallidum	Spirochaetes	160	TREPA	Gene loss

Note: *Spencer M and Sangaralingam A.[26]

Table S2. Event matrix for the validated gene content trees from literature for nine species groups.

Literature	Figure	Tree in Table 1	Leptospiraceae	Buchnera	Rickettsia	Mollicutes	Spirochetes	Chlamydiae	Actinobacteria	Cyanobacteria	Chlorobi
Deeds et al 2005*****	Fig. 4	Tree 1	0	0	0	0	0	0	−2	1	0
Lin and Gerstein 2000	Fig. 2A	Tree 2	0	0	0	−1	0	0	0	1	0
Deeds et al 2005*****	Fig. 6	Tree 3	0	0	0	0	0	0	1	1	0
Snel et al 1999	Fig. 2A	Tree 4	0	0	0	1	2	0	0	1	−1
Grishin et al 2000	Fig. 3	Tree 5	0	0	−1	1	−1	0	−1	1	0
Ciccarelli et al 2006	Fig. 2	Tree 6	1	1	1	1	−1	−2	−1	1	−2
Daubin et al 2002	Fig 2A	Tree 7	0	1	1	1	−2	−2	−2	−1	0
Brown et al 2001	Fig. 2	Tree 8	0	0	1	1	−1	−1	−1	1	−1
Brown et al 2001	Fig. 1	Tree 9	0	0	1	1	−2	−1	−1	1	0
Ma and Zeng 2004	Fig. 1B	Tree 10	0	1	1	1	2	1	−2	2	1
Gevers et al 2004*,**	Fig. 1	Tree 11	1	1	1	1	2	1	−1	2	1
Yang et al 2005*	Fig. 3	Tree 12	0	1	1	1	2	1	−1	1	−1
Muller et al 2010	Fig. 1#	Tree 13	1	1	1	1					
Moran et al 2008*,****	Fig. 1	Tree 14	1	1	1	1					
Wu and Eisen 2008*	Fig. 2	Tree 15	1	1	1	1	−2	−2	1	−2	−1
Gophna et al 2005	Fig. 5	Tree 16	1	1	1	1	−2	−2	1	−2	−1
Dutilh et al 2004	Fig. 4	Tree 17	0	1	1	1	−2	−2	2	−2	0
Korbel et al 2002	Fig. 2	Tree 18	0	1	1	1	−2	−2	0	−1	0
Ge et al 2005	Fig. 2	Tree 19	0	1	1	0	−2	−2	1	−2	0
Clarke et al 2002	Fig. 5	Tree 20	0	1	1	1	1	−2	2	−2	0
Clarke et al 2002	Fig. 2	Tree 21	0	1	1	1	1	−2	2	−2	0
Wolf et al 2001	Fig. 5	Tree 22	0	1	1	1	1	−2	2	−2	0
Sangaralingam et al 2010*	Fig. 2	Tree 23	0	1	1	1	1	−2	1	−2	0
Gophna et al 2005	Fig. 1	Tree 24	1	1	1	1	1	−2	2	−2	−1
Korbel et al 2002	Fig. 1	Tree 25	0	1	1	1	1	1	2	−2	0
Dutilh et al 2004	Fig. 3	Tree 26	0	1	1	1	1	−2	2	1	0
Daubin et al 2002	Fig 2B	Tree 27	0	1	1	1	−2	−2	1	−1	0
Henz et al 2005	Fig. 2	Tree 28	0	1	1	1	1	−1	−1	−1	−2
Tekaia et al 2005	Fig. S2	Tree 29	0	1	−1	1	−2	−1	2	−2	0
Wolf et al 2001	Fig. 4	Tree 30	0	1	−1	1	1	−2	1	−2	0
Lienau et al 2006	Fig. 6	Tree 31	−1	1	1	−1	−2	−1	−1	2	1
Hughes et al 2005	Fig. 3	Tree 32	0	1	1	−1	−1	−1	−1	2	1
Tekaia et al 1999	Fig. 2A	Tree 33	0	0	−1	−1	−1	−1	0	1	0
Hughes et al 2005	Fig. 2	Tree 34	0	−1	−1	−1	−1	−1	−2	2	1
Ma and Zeng 2004	Fig. 1A	Tree 35	0	−1	−1	−1	−1	−1	−2	2	1
Tekaia et al 2005	Fig. S3	Tree 36	0	−1	−1	−1	−1	−1	2	−1	0
Tekaia et al 2005	Fig. S1	Tree 37	0	−1	−1	−1	−1	−1	1	−1	0
Sangaralingam et al 2010*	Fig. 1	Tree 38	0	−1	−1	−1	−1	−1	1	−2	0
Wolf et al 2002	Fig. 1	Tree 39	0	−1	−1	−1	−1	−1	1	−2	0
Sangaralingam et al 2010*	Fig. 3	Tree 40	0	−1	−1	−1	−1	−1	1	−2	0
Spencer et al 2006	Fig. 4	Tree 41	0	−1	−1	−1	−1	−1	1	−2	0
Spencer et al 2006	Fig. 5	Tree 42	0	−1	−1	−1	−1	−1	1	−2	0
Spencer et al 2009*	Fig. 3	Tree 43	0	−1	−1	−1	−1	−1	1	−2	0
Gu and Zhang 2004	Fig. 3	Tree 44	0	−1	−1	−1	−1	−1	2	−2	0
Tekaia et al 2005	Fig. 4	Tree 45	0	−1	−1	−1	−1	−1	−2	−1	0

(*Continued*)

Table S2. (*Continued*)

Literature	Figure	Tree in Table 1	Leptospiraceae	Buchnera	Rickettsia	Mollicutes	Spirochetes	Chlamydiae	Actinobacteria	Cyanobacteria	Chlorobi
Hong et al 2004	Fig. 2A	Tree 46	0	1	–1	–1	–1	–1	1	–2	0
Wolf et al 2001	Fig. 3	Tree 47	0	1	0	–1	–1	–1	1	–2	0

Notes: Abbreviation and definitions for the values used in the matrix and in the clustering can be found in Table 2. Publications are ordered according to the heat map in Figure 2. #Interactive see also, http://eggnog.embl.de/cgi_bin/stats.pl; *Eukaryota and Archaea are ignored; **16S rRNA Tree; ***proteobacteria only; ****symbionts only; *****parasites excluded.

Table S3. 106 completely sequenced species as used for SYSTERS-PhyloMatrix GCT inferences ordered by protein family size.

Species*	Belongs to the parasites	NCBI TaxID	UniProt code	SYSTERS-PhyloMatrix protein family size**
Guillardia theta		55529	GUITH	300
Mycoplasma genitalium	Mollicutes	2097	MYCGE	329
Ureaplasma parvum	Mollicutes	134821	UREPA	337
Mycoplasma pneumoniae	Mollicutes	2104	MYCPN	351
Mycoplasma pulmonis	Mollicutes	2107	MYCPU	377
Buchnera aphidicola (Schizaphis graminum)	Buchnera	98794	BUCAP	470
Buchnera aphidicola (Acyrthosiphon pisum)	Buchnera	118099	BUCAI	476
Mycoplasma penetrans	Mollicutes	28227	MYCPE	478
Borrelia burgdorferi	Spirochaetae	139	BORBU	491
Wigglesworthia glossinidia endosymbiont of glossina brevipalpis		36870	WIGBR	515
Treponema pallidum	Spirochaetae	160	TREPA	524
Rickettsia prowazekii	Rickettsia	782	RICPR	553
Rickettsia conorii	Rickettsia	781	RICCN	630
Encephalitozoon cuniculi		6035	ENCCU	644
Chlamydia trachomatis	Chlamydia	813	CHLTR	724
Chlamydia muridarum	Chlamydia	83560	CHLMU	726
Chlamydophila pneumoniae	Chlamydia	83558	CHLPN	740
Thermoplasma volcanium		50339	THEVO	810
Thermoplasma acidophilum		2303	THEAC	815
Aeropyrum pernix		56636	AERPE	868
Aquifex aeolicus		63363	AQUAE	890
Helicobacter pylori J99		85963	HELPJ	891
Helicobacter pylori		210	HELPY	900
Methanopyrus kandleri		2320	METKA	901
Bifidobacterium longum	Actinobacteria	216816	BIFLO	942
Pyrobaculum aerophilum		13773	PYRAE	943
Halobacterium sp. NRC-1		64091	HALN1	988
Methanothermobacter thermautotrophicus str. Delta H		187420	METTH	1035
Fusobacterium nucleatum subsp. nucleatum		76856	FUSNN	1036
Methanocaldococcus jannaschii		2190	METJA	1037
Thermotoga maritima		2336	THEMA	1039
Chlorobium tepidum	CHLTE	1097	CHLTE	1043
Mycobacterium leprae	Actinobacteria	1769	MYCLE	1057
Campylobacter jejuni		197	CAMJE	1076
Sulfolobus tokodaii		111955	SULTO	1085
Sulfolobus solfataricus		2287	SULSO	1108
Neisseria meningitidis serogroup B		491	NEIMB	1164
Archaeoglobus fulgidus		2234	ARCFU	1172
Neisseria meningitidis serogroup A		65699	NEIMA	1188
Streptococcus pneumoniae R6		171101	STRR6	1201
Leptospira interrogans	LEPIN	173	LEPIN	1220
Pyrococcus horikoshii		53953	PYRHO	1221
Streptococcus mutans		1309	STRMU	1229
Lactococcus lactis subsp. lactis		1360	LACLA	1235
Pyrococcus abyssi		29292	PYRAB	1250

(*Continued*)

Table S3. (*Continued*)

Species*	Belongs to the parasites	NCBI TaxID	UniProt code	SYSTERS-PhyloMatrix protein family size**
Haemophilus influenzae		727	HAEIN	1251
Streptococcus pneumoniae		1313	STRPN	1273
Streptococcus pyogenes MGAS8232		186103	STRP8	1282
Streptococcus pyogenes MGAS315		198466	STRP3	1283
Pyrococcus furiosus		2261	PYRFU	1297
Streptococcus pyogenes		1314	STRPY	1298
Streptococcus agalactiae serogroup III		216495	STRA3	1302
Streptococcus agalactiae serogroup V		216466	STRA5	1347
Deinococcus radiodurans		1299	DEIRA	1360
Thermoanaerobacter tengcongensis		119072	THETN	1362
Pasteurella multocida		747	PASMU	1383
Clostridium perfringens		1502	CLOPE	1404
Methanosarcina mazei		2209	METMA	1408
Xylella fastidiosa		2371	XYLFA	1414
Synechococcus elongatus	Cyanobacteria	32046	SYNEL	1474
Corynebacterium efficiens	Actinobacteria	152794	COREF	1490
Methanosarcina acetivorans		2214	METAC	1513
Corynebacterium glutamicum	Actinobacteria	1718	CORGL	1530
Staphylococcus epidermidis		1282	STAEP	1551
Listeria monocytogenes		1639	LISMO	1588
Listeria innocua		1642	LISIN	1620
Mycobacterium tuberculosis	Actinobacteria	1773	MYCTU	1638
Clostridium acetobutylicum		1488	CLOAB	1657
Synechocystis sp. PCC 6803	Cyanobacteria	1148	SYNY3	1695
Staphylococcus aureus subsp. aureus N315		158879	STAAN	1794
Staphylococcus aureus subsp. aureus MW2		196620	STAAW	1808
Staphylococcus aureus subsp. aureus Mu50		158878	STAAM	1825
Saccharomyces cerevisiae		4932	YEAST	1848
Caulobacter vibrioides		155892	CAUCR	1859
Oceanobacillus iheyensis		182710	OCEIH	1882
Brucella melitensis biovar Suis		29461	BRUSU	1887
Brucella melitensis		29459	BRUME	1925
Bacillus halodurans		86665	BACHD	1980
Bacillus subtilis		1423	BACSU	2032
Schizosaccharomyces pombe		4896	SCHPO	2048
Vibrio cholerae		666	VIBCH	2090
Nostoc sp. PCC 7120	Cyanobacteria	103690	ANASP	2110
Shewanella oneidensis		70863	SHEON	2184
Ralstonia solanacearum		305	RALSO	2208
Streptomyces coelicolor	Actinobacteria	1902	STRCO	2225
Yersinia pestis		632	YERPE	2229
Vibrio vulnificus		672	VIBVU	2283
Xanthomonas campestris pv. campestris		340	XANCP	2307

(*Continued*)

Table S3. (*Continued*)

Species*	**Belongs to the parasites**	**NCBI TaxID**	**UniProt code**	**SYSTERS-PhyloMatrix protein family size****
Xanthomonas axonopodis pv. citri		92829	XANAC	2365
Agrobacterium tumefaciens str. C58		176299	AGRT5	2561
Pseudomonas aeruginosa		287	PSEAE	2652
Salmonella typhi		601	SALTI	2686
Sinorhizobium meliloti		382	RHIME	2686
Escherichia coli O6		217992	ECOL6	2715
Salmonella typhimurium		602	SALTY	2771
Mesorhizobium loti		381	RHILO	2800
Arabidopsis thaliana		3702	ARATH	2836
Escherichia coli O157:H7		83334	ECO57	2876
Escherichia coli		562	ECOLI	3046
Caenorhabditis briggsae		6238	CAEBR	3143
Caenorhabditis elegans		6239	CAEEL	3353
Drosophila melanogaster		7227	DROME	4252
Anopheles gambiae		7165	ANOGA	4514
Takifugu rubripes		31033	FUGRU	6460
Mus musculus		10090	MOUSE	6649
Homo sapiens		9606	HUMAN	6655

Notes: Identification of the NCBI Taxonomy and the UniProt HAMAP systematic is provided. The protein family size is the number of SYSTERS families that are present in the PhyloMatrix data set. *Explore the taxonomic tree for all 106 species in the PhyloMatrix data set using the 'taxonomic tree' link at http://systers.molgen.mpg.de/PhyloMatrix/; **explore the full set of SYSTERS protein families for a species at http://systers.molgen.mpg.de/cgi-bin/selecttaxon.pl; find the respective PhyloMatrix protein family subset by copy and paste using http://systers.molgen.mpg.de/PhyloMatrix/.

Figures S1 to S7: Seven SYSTERS-PhyloMatrix gene content phylogenies

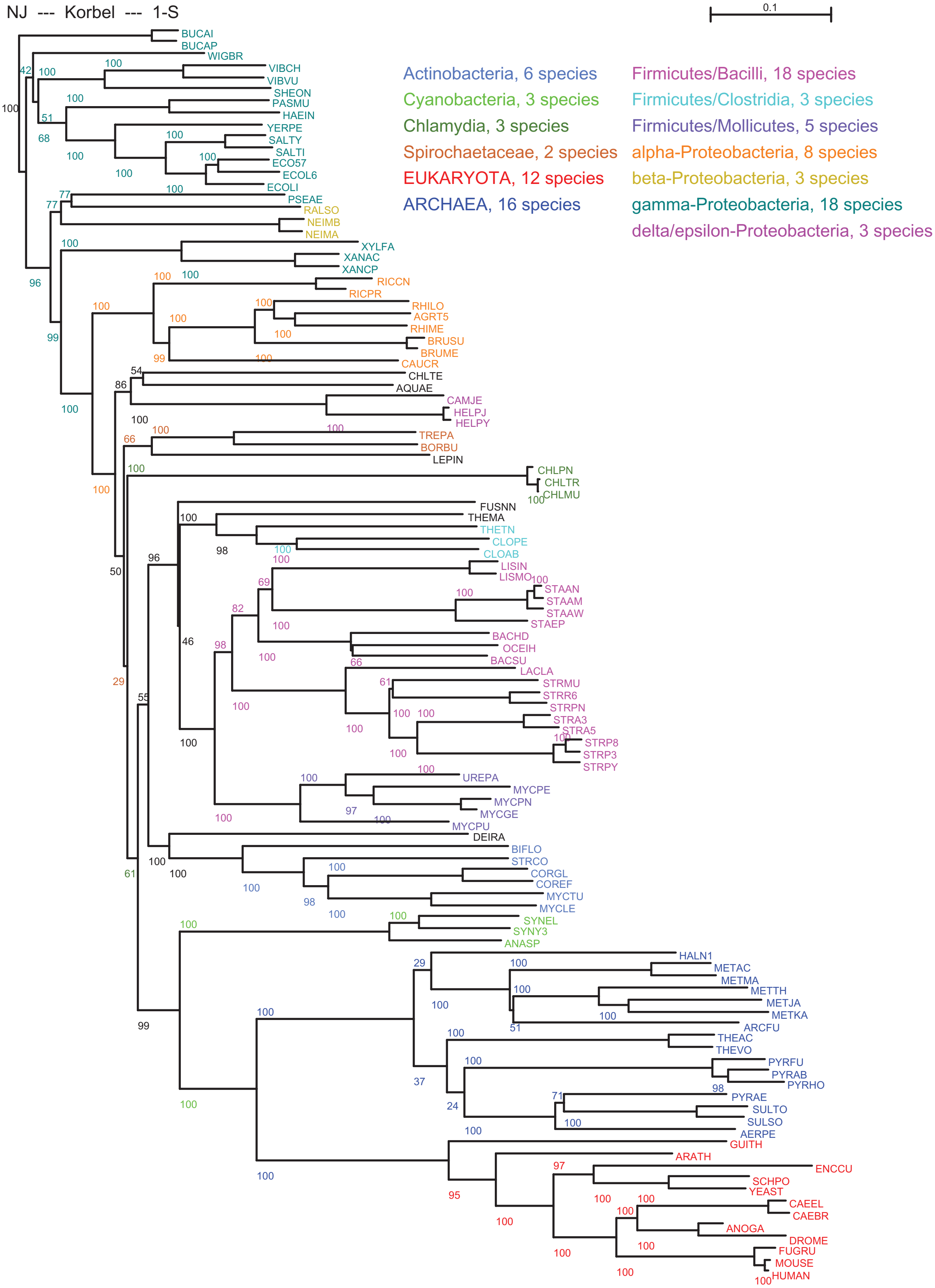


Figure S1. Trees are inferred with Neighbor Joining (NJ) and Korbel distance metric.

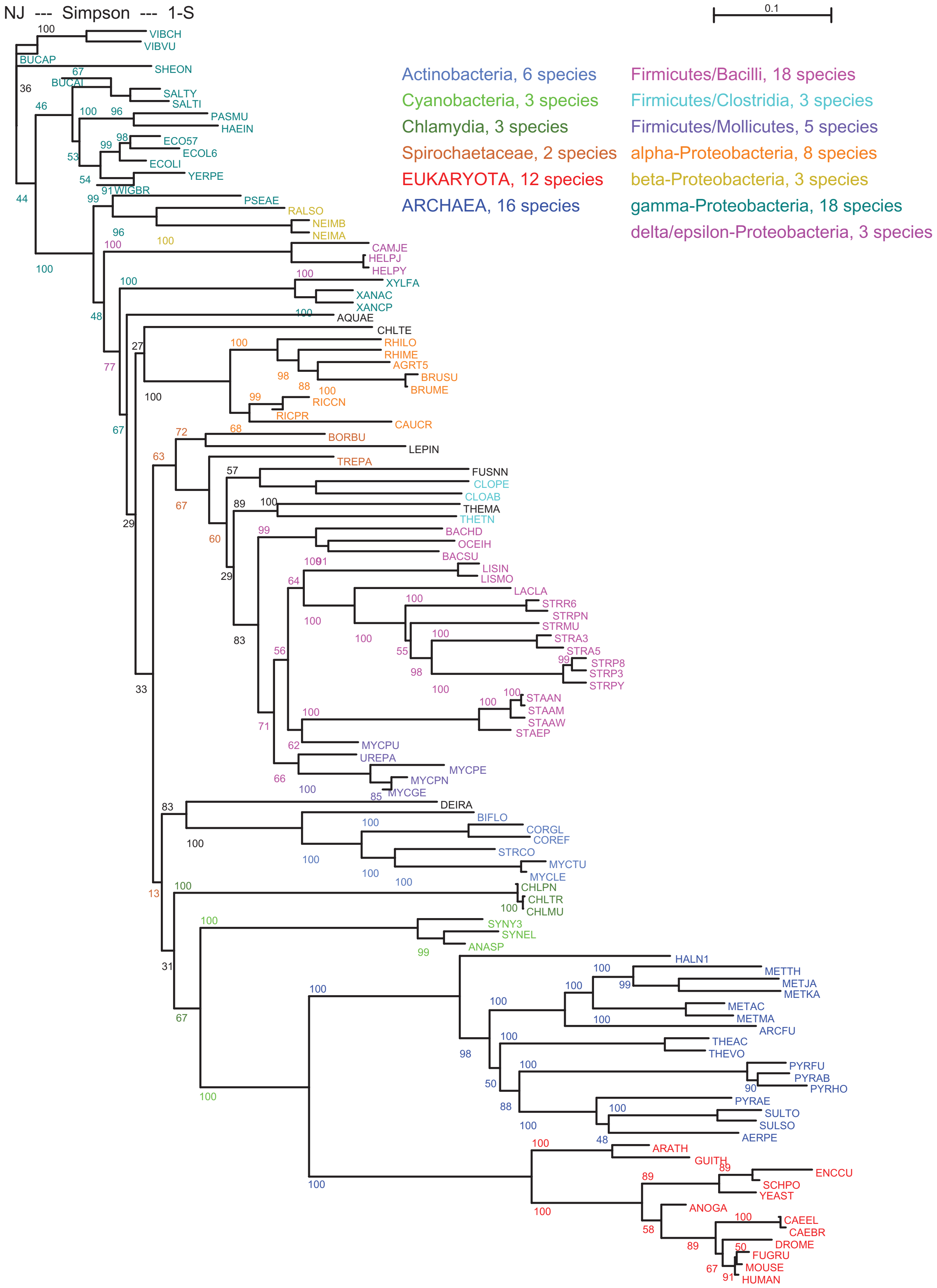


Figure S2. NJ with Simpson metric.

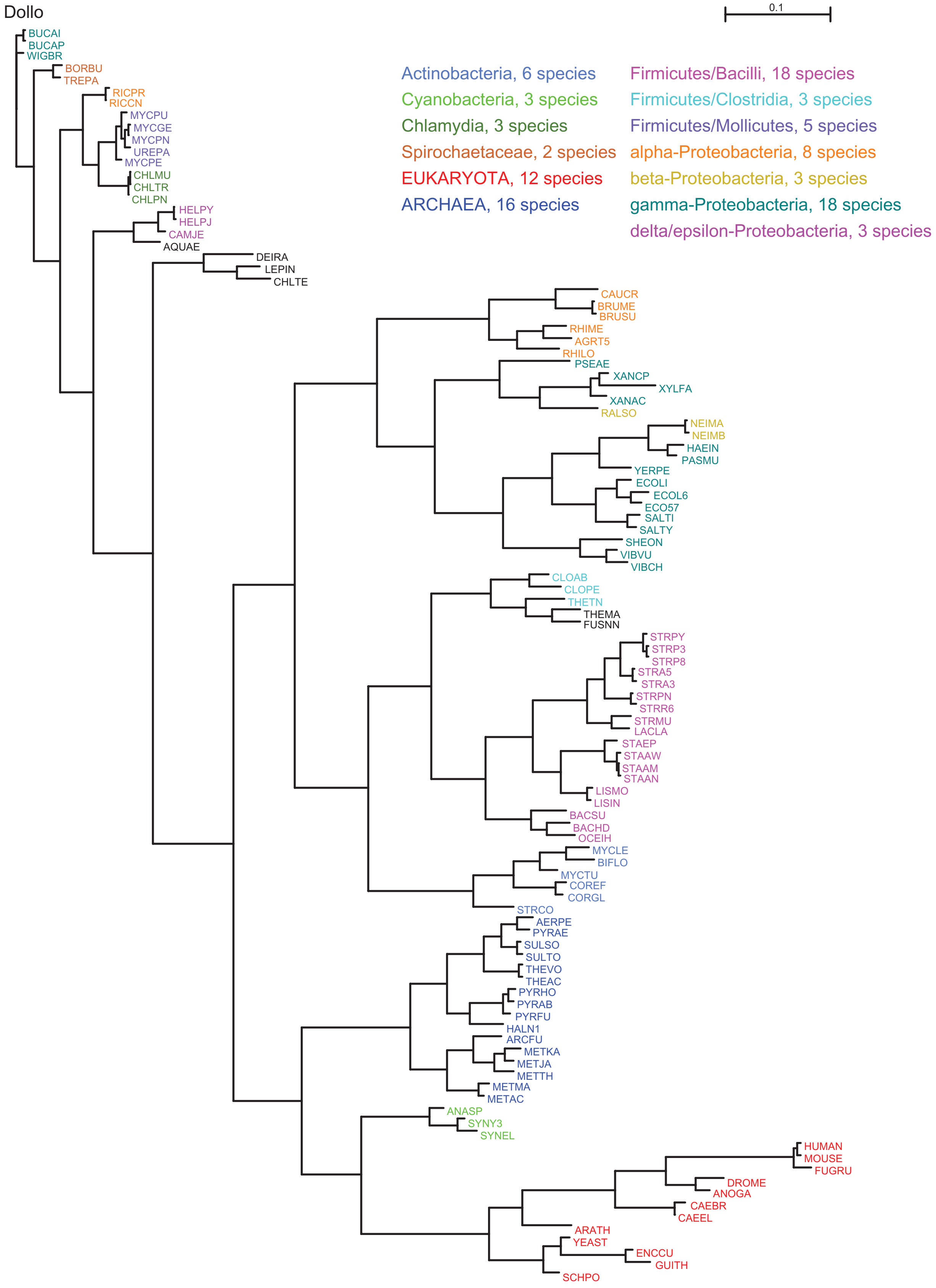


Figure S3. Dollo parsimony.

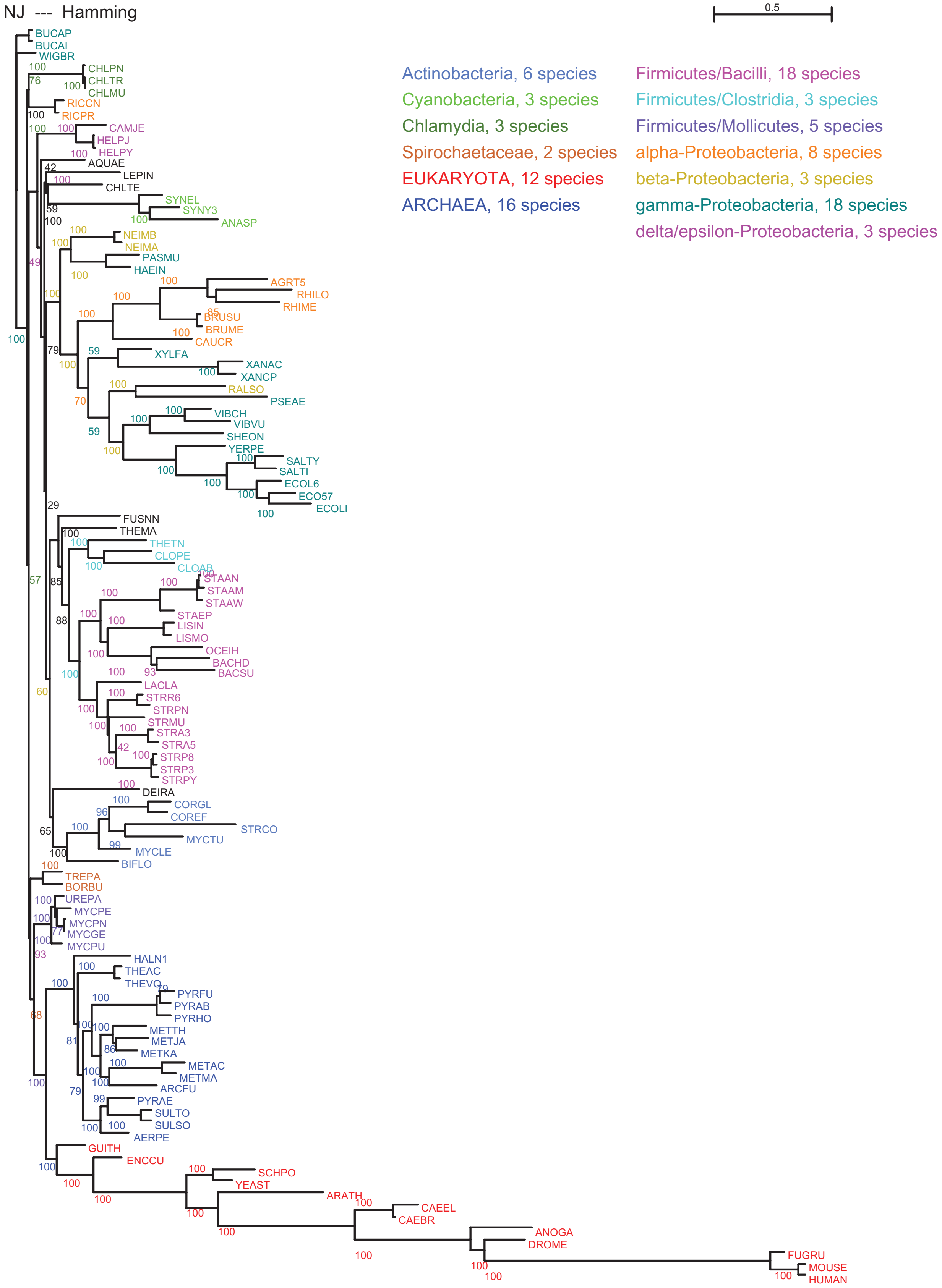


Figure S4. NJ with Hamming distance metric.

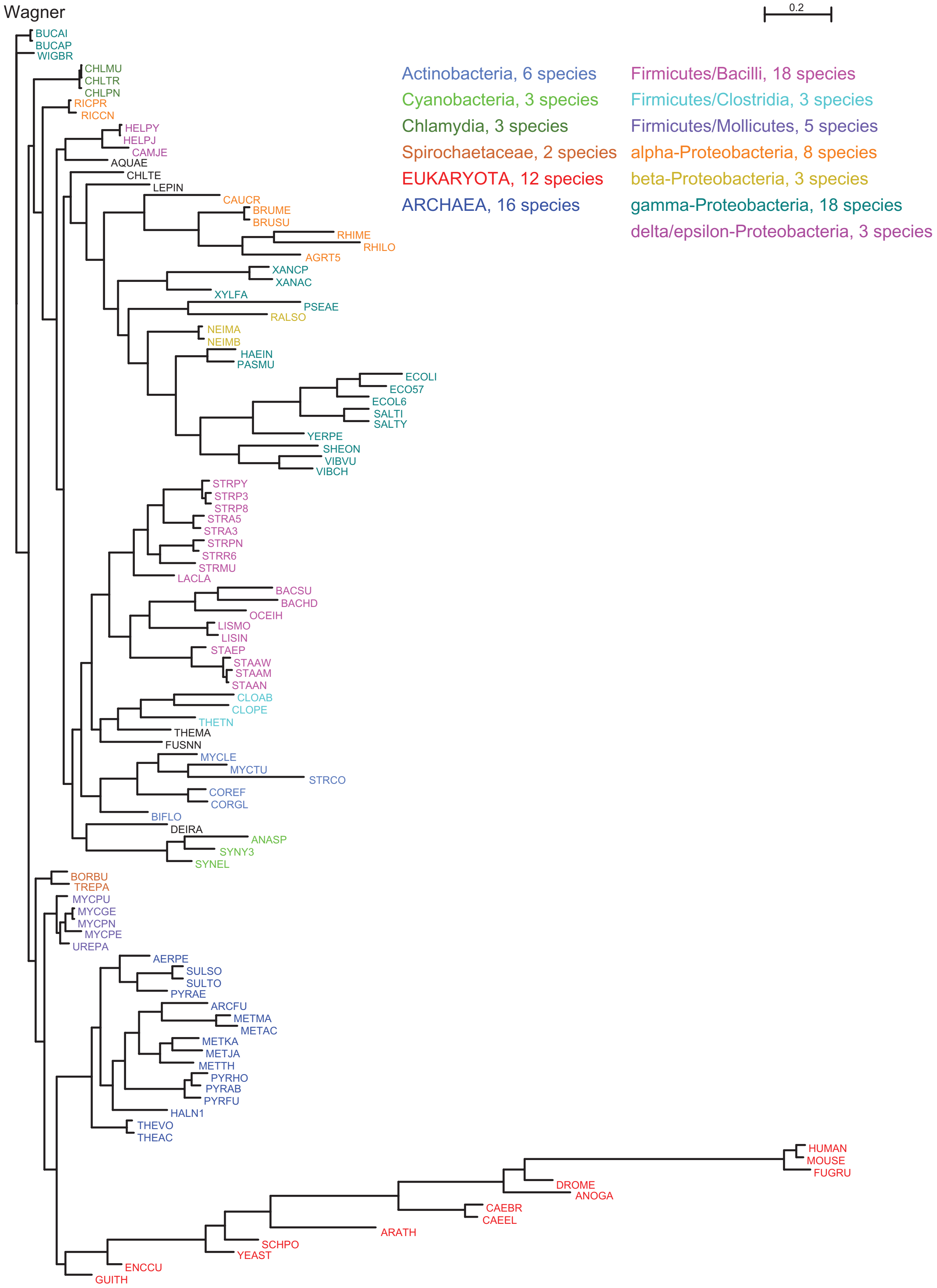


Figure S5. Wagner parsimony.

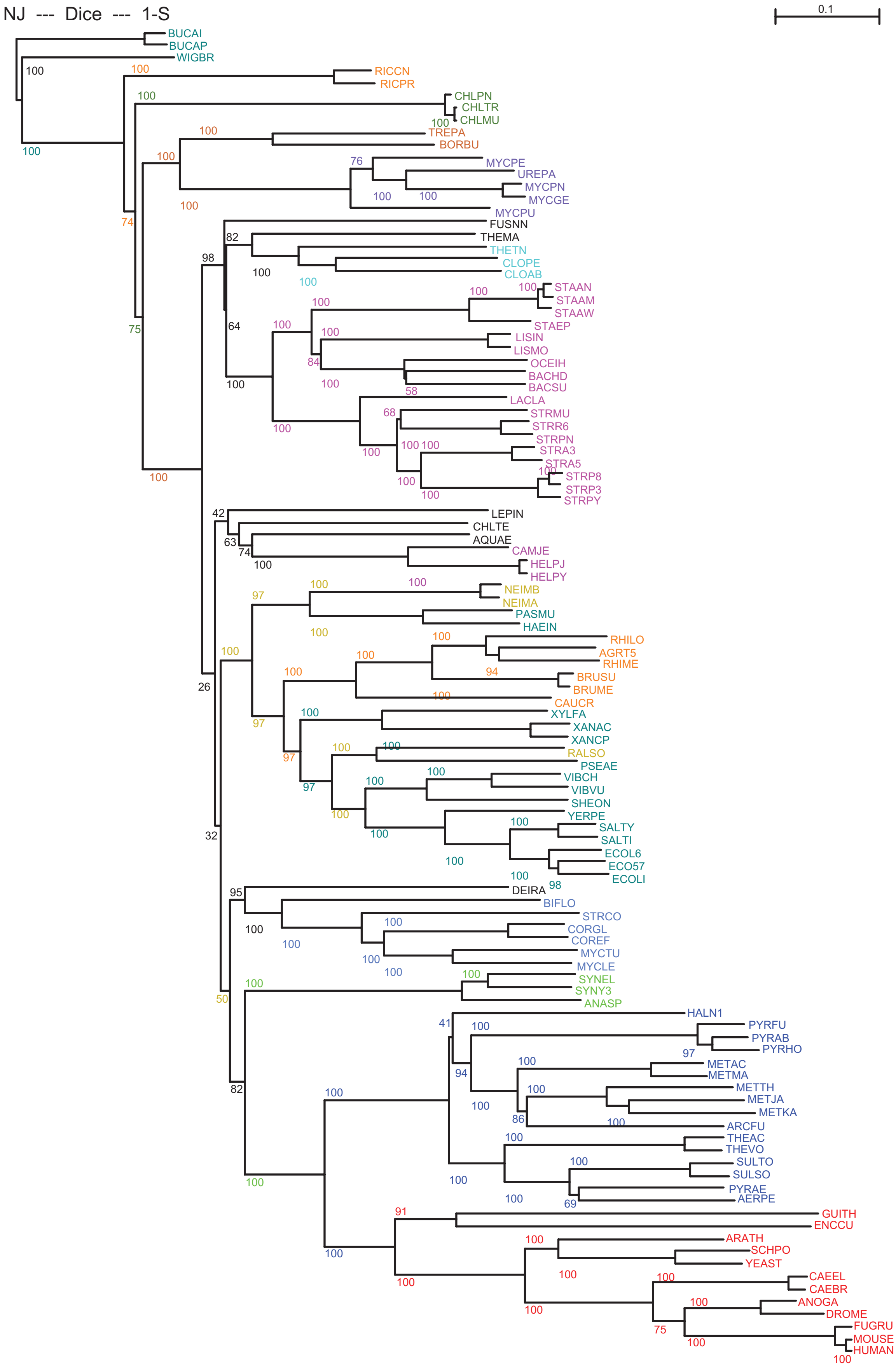


Figure S6. NJ with Dice metric.

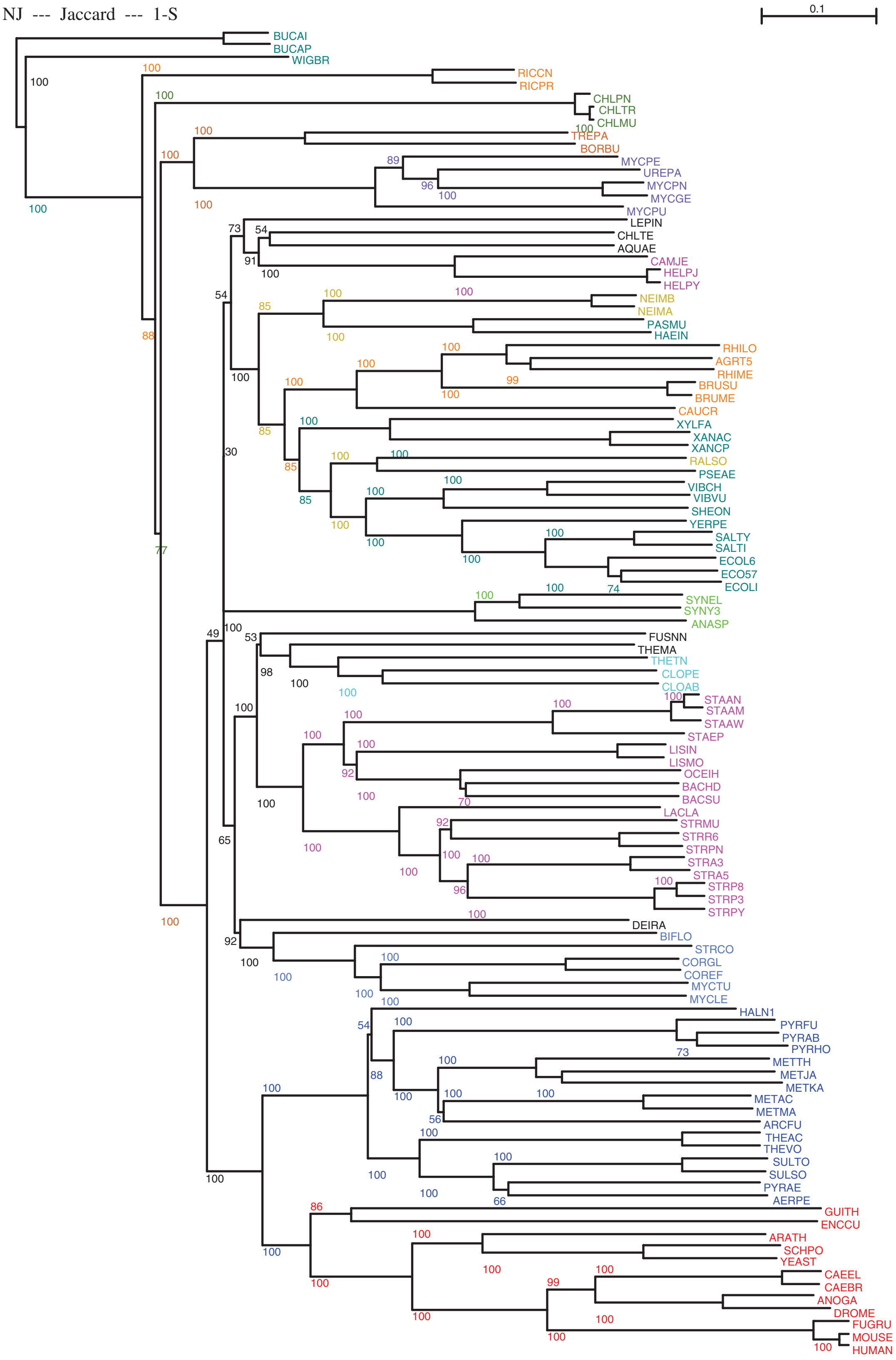


Figure S7. NJ with Jaccard metric.