

The mega-matrix tree of life: using genome-scale horizontal gene transfer and sequence evolution data as information about the vertical history of life

E. Kurt Lienau^{a,b,c,*}, Rob DeSalle^a, Marc Allard^c, Eric W. Brown^c, David Swofford^d,
Jeffrey A. Rosenfeld^b, Indra N. Sarkar^e and Paul J. Planet^{a,f}

^aSackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West at 79th St, New York, NY 10024, USA;

^bDepartment of Biology, Graduate School of Arts and Science, New York University, 100 Washington Square East, New York, NY 10003, USA;

^cDivision of Microbiology, Center for Food Safety and Nutrition, Food and Drug Administration, 5100 Paint Branch Parkway, College Park, MD 20740, USA; ^dDuke Institute for Genomes and Science Policy, 366 BioSci, Duke University, Durham, NC 27708, USA; ^eMarine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA; ^fDepartment of Pediatrics, Children's Hospital of New York, Columbia University, College of Physicians and Surgeons, New York, NY 10032, USA

Accepted 26 June 2010

Abstract

Because horizontal gene transfer can confound the recovery of the largely prokaryotic tree of life (ToL), most genome-based techniques seek to eliminate horizontal signal from ToL analyses, commonly by sieving out incongruent genes and data. This approach greatly limits the number of gene families analysed to a subset thought to be representative of vertical evolutionary history. However, formalized tests have not been performed to determine whether combining the massive amounts of information available in fully sequenced genomes can recover a reasonable ToL. Consequently, we used empirically defined gene homology definitions from a previous study that delineate xenologous gene families (gene families derived from a common transfer event) to generate a massively concatenated, combined-data ToL matrix derived from 323 404 translated open reading frames arranged into 12 381 gene homologue groups coded as amino acid data and 63 336, 64 105, 65 153, 66 922 and 67 109 gene homologue groups coded as gene presence/absence data for 166 fully sequenced genomes. This whole-genome gene presence/absence and amino acid sequence ToL data matrix is composed of 4867 184 characters (a combined data-type mega-matrix). Phylogenetic analysis of this mega-matrix yielded a fully resolved ToL that classifies all three commonly accepted domains of life as monophyletic and groups most taxa in traditionally recognized locations with high support. Most importantly, these results corroborate the existence of a common evolutionary history for these taxa present in both data types that is evident only when these data are analysed in combination.

© The Willi Hennig Society 2010.

A tangled web

Tree of life (ToL) systematists can be sure that a great proportion of the taxa we analyse are to some degree the products of horizontal gene transfer (HGT) from distantly related organisms. Consequently, techniques have been developed that attempt to recover a ToL by excluding as much potentially misleading signal as possible (data sieving). Multiple techniques for data

sieving have been proposed. Some techniques avoid using gene sequence data in favour of using other types of characters such as functional characteristics (Briones et al., 2005), similarity of metabolic networks (Oh et al., 2006), protein conservation profiles (Tekaia and Yeramian, 2005), gene order (Kunisawa, 2006), or gene network composition (Ding et al., 2008). Most analyses seek to eliminate horizontal signal from ToL analyses by limiting the gene families used to those thought to approximate the vertical history of life based upon a variety of criteria (Table 1). Recently, the implied histories of gene families identified as not subject to

*Corresponding author:

E-mail addresses: kurt.lienau@FDA.HHS.gov; klienau@gmail.com

Table 1
Methods of approximation of vertical history in TOL analyses

Method	Reference
COG reciprocal best hits test	Ge et al. (2005)
Inclusion in a single, pre-determined genome	Lake and Rivera (2004)
The super-tree approach	Daubin et al. (2002), Beiko et al. (2005), Pisani et al. (2007)
The complexity hypothesis	Rivera et al. (1998), Jain et al. (1999), Brochier and Philippe (2002)
Wide, balanced distribution of gene presence	Charlebois and Doolittle (2004)
Ubiquitous gene presence in the ToL	Lerat et al. (2003), Ciccarelli et al. (2006)
Consistency with small subunit DNA (SSUrDNA)	Harris et al. (2003)
Congruence with other putative non-HGT gene families	Brochier et al. (2005), Ciccarelli et al. (2006)

HGT were found to be so different as to preclude the ability to recover anything other than a “semi-rake” of a tree (Baptiste et al., 2008). Because of the unresolved and sometimes bizarre inferences of relationship made for the prokaryotic ToL, as well as the propensity of phylogenetic techniques that use the presence and absence of genes as characters to group genomes of unusual size (GOUS) based on size (Big Genome Attraction, or BGA; Lake and Rivera, 2004), serious questions have arisen about the feasibility of constructing a vertical ToL, or whether Darwin’s tree hypothesis even applies to the evolution of prokaryotic organisms (Doolittle and Baptiste, 2007; Koonin and Wolf, 2009; Mcinerney et al., 2008; Puigbo et al., 2009). The purpose of the present study was to explore the possibility that a resolved and consistent vertical ToL can be obtained by using as much genomic information as possible. We used empirically defined similarity values to delineate gene homologue groups out of the whole genome translated ORF SWISSPROT database (sensu Lienau et al., 2006; see Methods) that are representative of xenologue families (Lienau et al., 2006 and unpublished data), or gene groups that are descended from a common gene transfer event, and analysed them both as whole-genome amino acid sequence data (AD) and whole-genome gene presence/absence data (PD) in a massively concatenated, combined data (CD) ToL matrix. Phylogenetic analysis of this CD matrix yields a single, highly supported and corroborated ToL that reveals a strong phylogenetic signal common to both the AD and PD that was not apparent during separate analyses of these two kinds of data.

Methods

Three massively concatenated matrices

PD matrix. For the PD matrix, we used the gene presence/absence data of Lienau et al. (2006), a study that accounted for the uncertainty as to what similarity threshold to use to arrange gene clusters to test relationships in the ToL by testing the phylogenetic

behaviour of PD matrices constructed at a range of similarity thresholds for determining gene homology definitions. Lienau et al. (2006) used 111 different BLAT (Kent, 2002) *e*-value cutoffs in a single linkage sequence-clustering (SLC) algorithm to define gene homology groups out of 323 404 translated open reading frames (ORFs) for 166 fully sequenced genomes. An SLC algorithm includes any sequence that is similar above a defined similarity threshold to at least one member of a gene cluster in that homology group. In Lienau et al. (2006), the homologue group definitions derived from the SLC algorithm at 111 different *e*-value similarity cutoffs were used to construct 111 gene PD matrices. Phylogenetic analysis of these matrices yielded 111 whole-genome gene PD phylogenetic trees. The study then examined the strict consensus tree resolution and character consistency of each phylogenetic tree made from each translated ORF presence/absence matrix defined by each separate *e*-value. Specifically, we used the combined corroboration metric (CCM) defined as the product of the rescaled consistency index (RCI; Farris, 1989) and the Rohlf consensus index 1 (Rohlf, 1982) as an indication of the quality of the gene homology definitions for use in phylogenetic analysis. The results showed that there were several *e*-values that gave rise to data sets that yielded ToLs with nearly identically high CCM scores, each separated by ranges of *e*-values that gave data sets and ToLs with less optimal CCM scores. Because of this, and to account for the likelihood that no one *e*-value is suitable for gene homology definition for the entirety of the ToL, Lienau et al. (2006) selected the top five PD data matrices (those derived from the *e*-values: 10^{-88} , 10^{-85} , 10^{-81} , 10^{-73} and 10^{-72} , and defining 63 336, 64 105, 65 153, 66 922 and 67 109 gene clusters, respectively) for use in a concatenated gene presence/absence matrix. We used this gene presence/absence matrix as the PD in this study.

AD matrix. We used the best *e*-value (10^{-88}) found in Lienau et al. (2006) to define gene homologue groups in the amino acid sequence matrix. Because multiple gene sequences in a taxon belonging to a single gene family

can be counted as a single “present” in a presence/absence matrix, the Lienau et al. (2006) publication did not need to address the problem of paralogy; however, we needed to address this problem when using amino acid sequences in the AD and CD analyses. To remove all paralogous sequences, we searched for and removed any gene sequences that were present in two or more copies for any taxon in the analysis. This yielded a data set of 12 381 homologous gene groups for 165 taxa; we aligned these groups with muscle (Edgar, 2004) and concatenated the alignments to yield a multilocus translated ORF matrix that contained 4 540 579 characters. We customized a version of PAUP*4.0 so that it was able to import the 4450 579 characters in the amino acid matrix. We then output the 846 999 parsimony-informative amino acid sequence characters to a parsimony-informative only AD matrix (see Table 2).

CD mega-matrix. We concatenated the gene PD matrix from Lienau et al. (2006) (see Table 2) with the AD matrix to create a combined data (gene gain and loss and amino acid sequence evolution) ToL mega-matrix (CD matrix—see Table 2) comprising 166 taxa and 4867 184 characters, 1173 599 of which are parsimony-informative (846 999 from the concatenated AD matrix and 326 605 from the PD matrix); one eukaryote, *Homo sapiens*, had no amino acid data due to the exclusion of paralogous sequences and therefore its placement was based only on information from the PD matrix.

Tree search

We searched for the optimal topology for all data sets using the parsimony ratchet technique (Nixon, 1999). We used PAUPRat OS X (Sikes and Lewis, 2001) to generate command files for implementation in paup*4b10-ppc-macosx (Swofford, 2000) and did performed runs of 200 iterations each (three times with 15, 17 and 21% character perturbation per iteration using

TBR). After the ratchet we continued the tree search using TBR and the “Mul-trees” option in paup*4b10-ppc-macosx (Swofford, 2000) using the most optimal tree(s) from the ratchet search as starting topologies. Gaps were treated as missing data. For final parsimony calculations, all characters and state transformations were given equal weight.

Data-type parsimony ratchet

To further explore the tree space in our heuristic tree search for the combined matrix, we designed a strategy based on the ratchet method of Nixon (1999) that differentially weighted the characters from each data type (i.e. PD and AD matrices) as opposed to randomly selecting characters to up-weight as in the original parsimony ratchet technique. For this “data-type ratchet” search we increased the number of PD characters by three, to provide equal chance that a character from each data-type would be chosen for up-weighting in the parsimony ratchet procedure. To test the effectiveness of this method, we used the ratchet procedure as implemented above (random selection of characters to up-weight during that part of the technique, which favoured the selection of AD characters, which occurs by default as this partition was larger). The data-type ratchet technique found a shorter tree (calculated using the equally weighted matrix) than the regular parsimony ratchet analysis. We therefore used this approach when calculating Bremer support measures, but not bootstrap support measures (see below). For all analyses, gaps in the AD matrix were treated as missing data.

Support calculations

Bootstrap support was calculated using the bootstrap search in paup*4b10-ppc-macosx (Swofford, 2000) at 100 iterations with the un-weighted CD matrix. We used Auto Decay (Eriksson, 2001) to generate Bremer

Table 2
Matrix construction

Matrix	Homologue groups	Taxa	Characters
Amino acid (AD)	12 381 homologue groups	165 (three domains—missing <i>Homo sapiens</i>)	4540 579 amino acid sequence characters (846 999 informative)
Gene presence/absence (PD)	63 336, 64 105, 65 153, 66 922 and 67 109 homologue groups	166 (three domains—includes <i>Homo sapiens</i>)	326 605 informative gene gain and loss characters (five matrices constructed at five <i>e</i> -value cutoffs from the same protein set; Lienau et al., 2006)
Combined AD + PD (the mega-matrix, CD)	63 336, 64 105, 65 153, 66 922 and 67 109 homologue groups coded as gene presence/absence, 12 381 of those homologue groups coded as amino acid sequence data	166 (three domains—includes <i>Homo sapiens</i>)	4867 184 characters; 1173 604 informative (the combination of the AD and PD data matrices)

support indices (BSIs; Bremer, 1994). Global hidden support for each partition on the combined, total evidence tree was calculated after the technique of Baker and Desalle (1997) by measuring the difference in length of the partition on its own most optimal tree from that of its length on the total evidence tree (CDToL). We used Auto decay (Eriksson, 2001) to generate partition Bremer support indices (pBSIs) for all three data matrices (PD, AD and CD) on all nodes of the CDToL according to Baker et al. (1998).

We used a new measure, the local hidden support index (lhBSI) to measure the contribution of support of each partition to each node of the CDToL. We calculated the contribution of support due to the combination of data types for each node in the CDToL by subtracting the sum of the BSI calculated for each individual partition (PD and AD) on each node of the CDToL from the BSI calculated for the combined CD matrix on each node of the CDToL. This operation can be summarized by the following formula:

lhBSI for each node on the CDToL:

$$\text{lhBSI (CDToL)} = [\text{CDToL BSI} - (\text{PD CDToL BSI} + \text{AD CDToL BSI})],$$

which can be read, “the local hidden Bremer support index for a node on the CDToL equals the Bremer support index for that node on the CDToL calculated using the CD matrix minus the sum of the Bremer support index for that node on the CDToL calculated using the PD matrix plus the Bremer support index calculated for that node using the AD matrix.”

By measuring the difference between the sum of the BSIs calculated on the CDToL using the PD and AD separately from the BSI calculated using the CDToL using the CD together, this calculation gives a measure of the support for each node in the CDToL that is due to simultaneous analysis of the two partitions together.

Comparison of the mega matrix ToL to the Ciccarelli et al. (2006) ToL

We compared the topologies generated by the Ciccarelli et al. (2006) data set (a matrix constructed using a set of 31 universal gene families that showed little reticulate evolution and were readily aligned for 191 genomes from all domains of life) and the CDToL by measuring the resolution of the strict consensus tree of the two trees generated by each data set with the Rohlf consensus index 1 (Rohlf, 1982). We also compared the average bootstrap support indices for each hierarchical level of the trees generated by the Ciccarelli et al. (2006) data set and the CDToL as well as the distribution of Bremer support by node for the CD ToL and the tree generated by the Ciccarelli et al. (2006) data set. To compare the support for each

matrix (that of Ciccarelli et al. (2006) and the CD matrix) under the same conditions we followed the identical procedure we used for tree searching and support calculations we used for the mega-matrix on the Ciccarelli et al. (2006) dataset.

Results

A combined data-type (CD) mega-matrix

We generated a translated ORF sequence super-matrix (amino acid alignment) based on the xenologue clusters, or gene presences, implied in the optimal presence/absence matrix from Lienau et al. (2006) (e -value 10^{-88}). After removing all sequences that were paralogous (that had two or more copies in a genome; see Methods), this matrix contained 4540 579 (846 999 parsimony-informative) amino acid characters corresponding to 12 381 xenologue families and 165 taxa from all three domains of life. Phylogenetic analysis of the concatenated amino acid character matrix yielded 50 most-parsimonious trees, the strict consensus of which failed to resolve any domain of life as monophyletic (Fig. 1). Even so, compared with the phylogeny derived from the concatenated five best e -value gene presence/absence matrices reported in Lienau et al. (2006), this amino-acid-based phylogeny was more successful in grouping many of the small genome bacteria such as *Rickettsia*, *Wolbachia*, *Buchnera* and *Mycoplasmales* in (or near in the case of *Mycoplasmales*) well-accepted taxonomic groupings for these enigmatic taxa, indicating that amino acid data are not as vulnerable to BGA (Lake and Rivera, 2004) as are gene content data. A potential explanation for the ambiguous results of the concatenated amino acid analysis is the lack of data pertaining to the acquisition and loss of genes that is contained in the presence/absence data.

We hypothesized that the addition of these data to our amino acid sequence data might help to resolve many of the more ancient relationships in the ToL by providing another test of history based on a separate biological process. To test this hypothesis we combined the AD data set with the five most optimal gene content data sets from Lienau et al. (2006)—the PD dataset—to construct a combined, gene presence/absence and concatenated amino acid sequence CDToL mega-matrix. Derived from 166 fully sequenced genomes that span all three domains of life, this combined data set is composed of 4867 184 characters, 1173 599 of which are parsimony-informative (846 999 from the concatenated AD matrix and 326 605 from the PD matrix), representing 323 404 translated genes. This body of data is observed from six “vantage points” of test (sensu Lienau and Desalle, 2008), meaning that this analysis

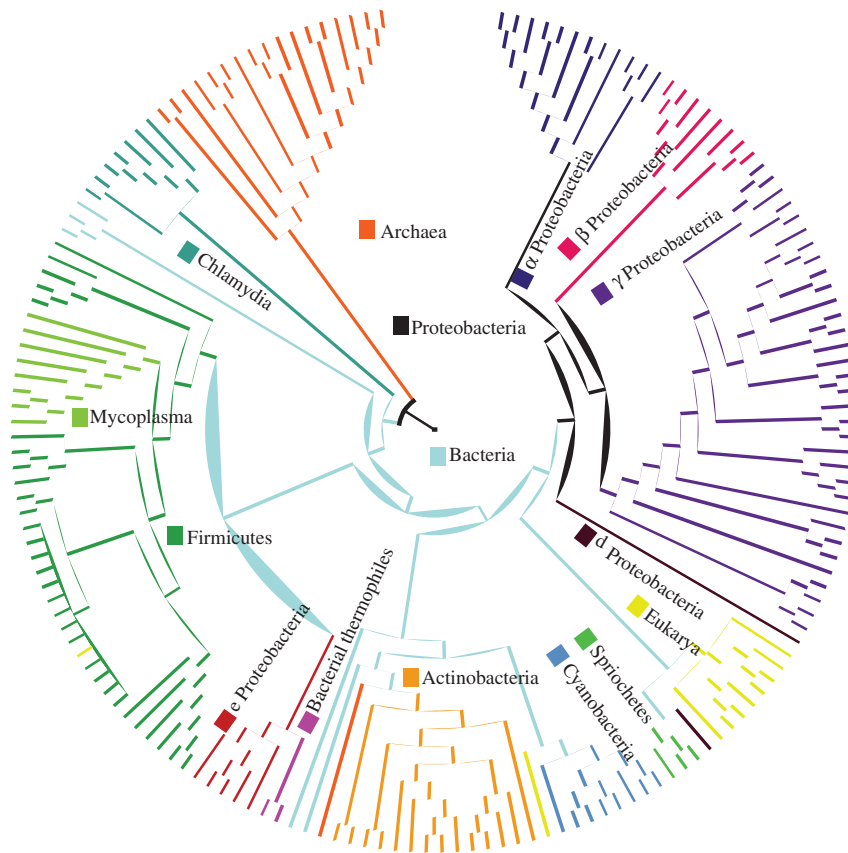


Fig. 1. The strict consensus of 50 most-parsimonious trees derived from the concatenated amino acid sequence matrix all of lengths 2 302 639, consistency indices (CI) of 0.742, homoplasy indices (HI) of 0.258, retention indices (RI) of 0.795, rescaled consistency indices (RC) of 0.589 and a CCM (Lienau et al., 2006) of 0.571. Major taxonomic groups are labelled in colours as follows: Firmicutes (green), Cyanobacteria (darker blue), Actinobacteria (lighter orange), Mycoplasmales (yellow-green), Chlamydiales (blue-green). Taxa belonging to the Archaea are coloured darker orange; Eukarya, yellow; Bacteria, pale blue. Taxa belonging to the α Proteobacteria are navy; β , mauve; γ , purple; δ , burgundy; ϵ , red. Bootstrap and Bremer support indices were low for basal nodes (data not shown). Note both the paraphyly of all three domains Archaea, Eukarya and Bacteria as well as the general lack of resolution in the tree. Also note the placement of small genomes, such as of *Mycoplasma*, and HGT promiscuous genomes, such as of *Fusobacterium*, bacteria in or close to traditional locations.

tests the hypotheses of relationship of the 166 taxa using five different PD data sets derived from the e -values 10^{-88} , 10^{-85} , 10^{-81} , 10^{-73} and 10^{-72} , and defining 63 336, 64 105, 65 153, 66 922 and 67 109 xenologue families, respectively, and one amino acid sequence data set derived from the e -value 10^{-88} corresponding to 12 381 of those xenologue families. Data-type parsimony ratchet analysis (see Methods) after Nixon (1999) with varied character perturbations (see Methods) of this CD mega-matrix yielded a single most-parsimonious ToL with an RCI of 0.613, and given that the tree is totally resolved, a CCM of 0.613, indicating highly consistent and congruent phylogenetic signal. In other words, these data contain tree-like structure. Greater than 94% of the nodes on this tree have 100% bootstrap support and the BSIs range from 11 to 146 667. All three domains of life (Woese, 1987; Woese et al., 1990; Pace et al., 1986) are monophyletic with BSI > 108. We rooted our tree of life (Fig. 2) with the traditional division between the

Bacteria and Archaea/Eukarya (Iwabe et al., 1989; Gogarten et al., 1989).

Character interaction in the mega-matrix ToL (CDToL)

Our combined gene PD and AD data set (the mega matrix; CD) produced a fully resolved, highly corroborated ToL hypothesis (the CDToL) while the concatenated translated gene sequence data set alone produced an unresolved tree which was unable to classify any domain of life as monophyletic. To investigate this phenomenon, we examined the contribution of support from each data type using the partition Bremer support method of Baker and Desalle (1997) and a new, related method we call the lhBSI method. We found that the CD mega-matrix showed significant Bremer support for all nodes of the CDToL and that separate analysis of the AD matrix and the PD matrix alone gave patchy support (Fig. 3a). Thus, the combination of these two types of

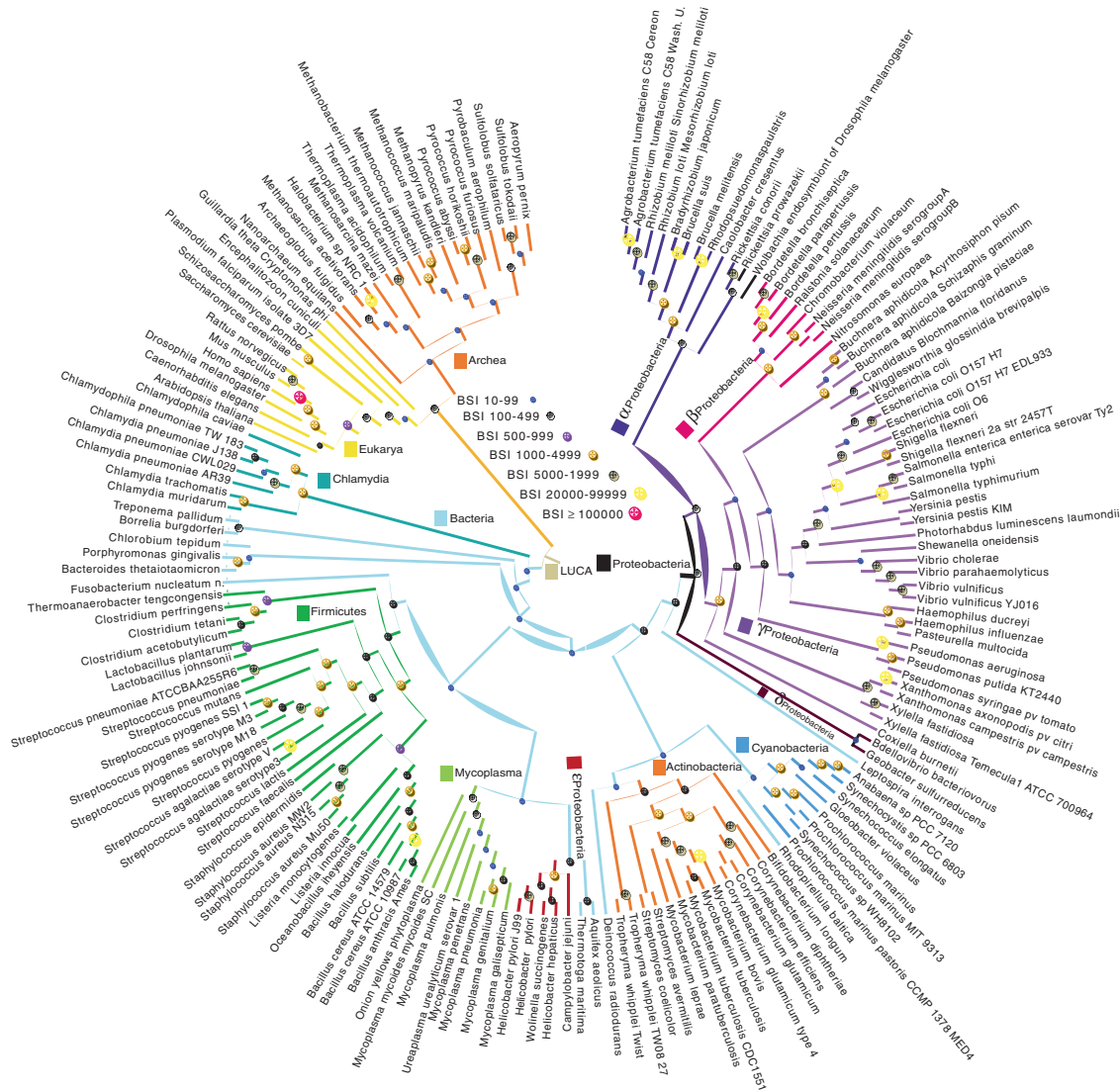


Fig. 2. The combined data mega-matrix tree of life (CDToL) is a single most-parsimonious tree with a tree length of 2 710 148, consistency index (CI) of 0.751, homoplasy index (HI) of 0.249, retention index (RI) of 0.816, a rescaled consistency index (RC) of 0.613 and therefore a CCM of 0.613. Bremer support indices are indicated above each node by differently coloured and sized circles. Bootstrap values are not listed (see Fig. 4); 94.5% of nodes had bootstrap support of 100%. Major taxonomic groups are labelled with boxes of the colours used in Fig. 1. All Proteobacteria are in traditional locations (including the small genome *Rickettsiales* in the α proteobacterial clade, and the small genome *Buchnera* within the γ proteobacteria) with the exception of the ϵ Proteobacteria, which resolve with the thermophilic bacteria as sister to Mycoplasma. This mycoplasma/ ϵ proteobacteria clade is sister to the traditional relatives of Mycoplasmales, the Firmicutes.

data in a simultaneous analysis revealed a clear phylogenetic signal that was common to both types of data but not strong enough in either matrix alone to support the combined tree; this type of phylogenetic signal is the hidden support of Gatesy et al. (1999, 2002) (Fig. 3b).

Comparison of the CDToL to a recent ToL analysis

We compared the topologies and support measures of the CDToL with the results of Ciccarelli et al.

(2006). Overall, the tree topologies (pruned to contain the same taxa) showed remarkable congruence, as indicated by the resolution of the strict consensus trees made from both topologies (data not shown). We compared the average bootstrap support and Bremer support measures for the nodes for both matrices and found that the bootstrap measures were generally comparable, but the CD mega-matrix showed much higher and more evenly distributed Bremer support for all nodes (Fig. 4).

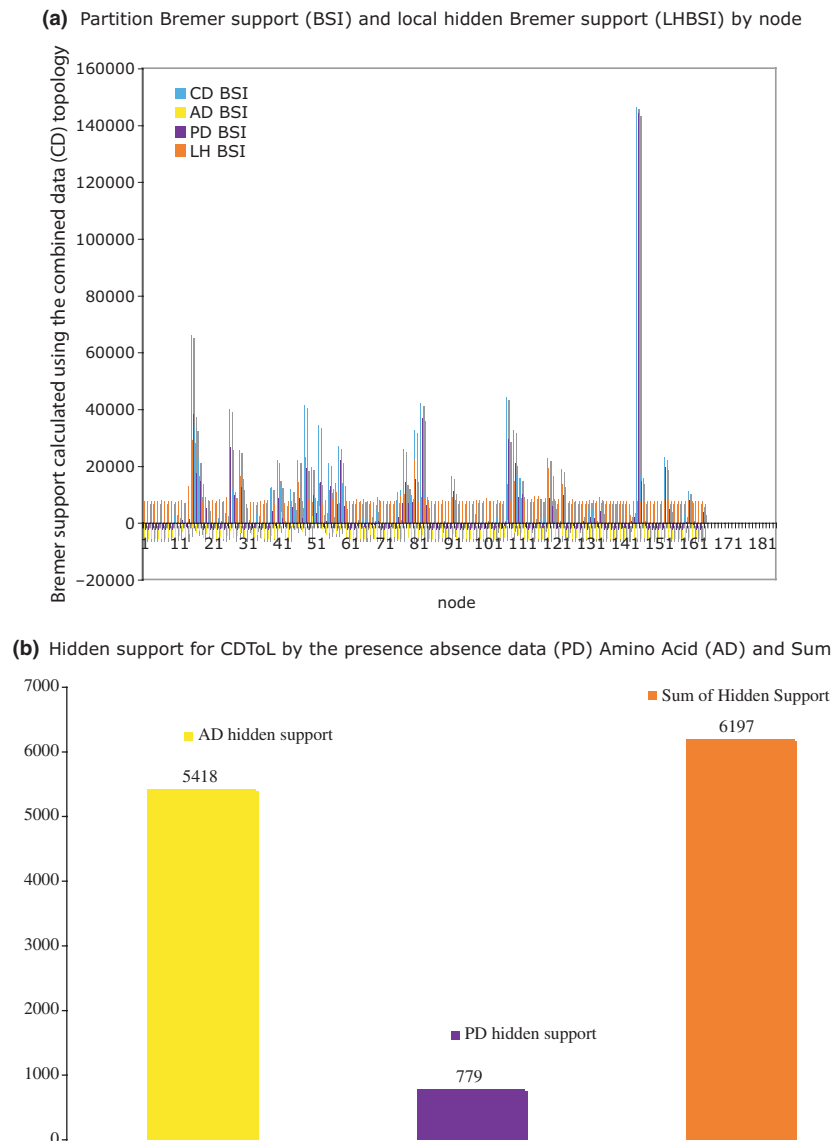


Fig. 3. Partitioned and hidden support in the CDTOL. (a) Bremer support indices are indicated for each node of the CDTOL as calculated for CD, the combined matrix (blue), AD, the amino-acid matrix (yellow), and PD, the presence/absence matrix (purple). The zero line on the graph corresponds to the number of steps required by each matrix to explain the data on the combined mega-matrix ToL. Positive Bremer support values indicate that the data partition under study favours the existence of the node; negative values indicate that the data partition under study is explained more parsimoniously on tree topologies that lack that node. Local hidden Bremer support (lhBSI)—or the difference between the sum of supports lent to the CDTOL by both the amino acid sequence and gene presence/absence matrices when analysed separately and the support given to the CDTOL by the simultaneous analysis of both data types for each node—is coloured orange. (b) Hidden support (HS)—or the difference in length of a partition when measured on its own most-parsimonious tree from the length of that partition on the total evidence tree—is shown for the PD partition (purple), the AD partition (yellow) and the sum of both partitions, the CD mega-matrix (orange). The existence of hidden support shows that the data, when analysed in concert, supports a different hypothesis of evolutionary history than either does when analysed separately.

Discussion

Concatenation of presence/absence (PD) data and amino acid (AD) sequences delivers a highly resolved, highly consistent, well-supported ToL

Although some scientists assert that the recovery of a vertical tree of life is unfeasible (e.g. Rokas and

Carroll, 2006; Baptiste et al., 2008) or that the ToL does not exist in any meaningful sense (Doolittle and Baptiste, 2007), it is necessary to test this hypothesis. The simplest and most relevant tests have, until now, not been done on large genomic-scale data sets as it has been assumed that data sieving is a necessary prerequisite to obtaining a reasonable, well-supported, and resolved ToL. The data-rich, concatenated CD

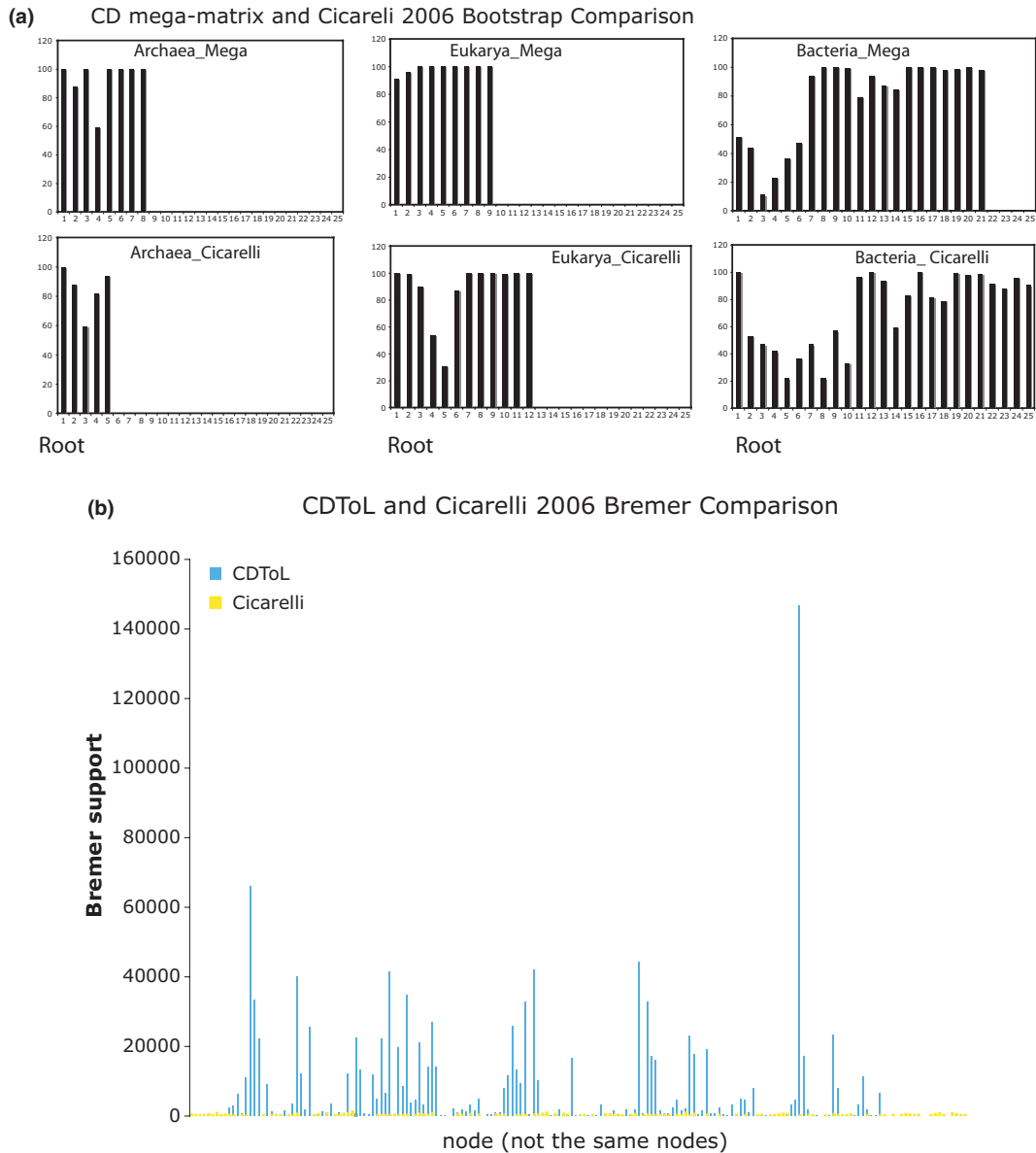


Fig. 4. Support measure comparison of CDToL with Ciccarelli et al. (2006). (a) Average bootstrap for each hierarchical level for both the CDToL and Ciccarelli et al. (2006). For each domain, we calculated the average bootstrap value for all nodes occupying the same tree height on the CDToL (Mega) and the Ciccarelli et al. (2006) ToL to give an overall indication of support for different hierarchical levels in each ToL analysis. (b) The distribution of BSI for each node of the CDToL (grey) and the Ciccarelli et al. (2006) ToL (yellow). It is important to note that, because there are more taxa in the Ciccarelli et al. ToL than the CDToL, the nodes are not depicted one to one; the presentation is meant to show overall trends of support for both ToLs.

mega-matrix constructed for this study gives a resolved, very consistent and highly supported ToL. The decision as to whether the tree represents reality at all nodes is, like any phylogenetic hypothesis, up to interpretation and further test. However, given that the CDToL hypothesis has been tested by the information imbedded in over 300 000 gene sequences, we suggest that it represents the most corroborated hypothesis of the evolutionary history of life on Earth to date. We examine some of the more interesting relationships

postulated by the CDToL below to highlight them as targets for further test.

External corroboration: congruence of the CDToL with existing bacterial classification systems

Most relationships that can be inferred from the CDToL (Fig. 2) are congruent with current ideas about bacterial evolution and systematics. Some of these relationships attest to the effectiveness of data

combination at resolving previously ambiguous relationships. For instance, genomes of unusually small sizes that are prone to BGA bias (Jain et al., 2002) such as *Rickettsia* (Alphaproteobacteria), *Buchnera* (Deltaproteobacteria) and *Mycoplasma* (Firmicutes) are classified in (and/or near in the case of *mycoplasma*) traditional locations in the CDTOL. The observation from the CDTOL that all three domains of life are monophyletic with high support shows that the concatenated matrix approach does not suffer from the same inability to recover ancient relationships as the concatenated amino acid data (Fig. 1).

An illustration of the CDTOL's ability to unambiguously classify organisms that show high levels of gene transfer is the robustly supported placement of *Fusobacterium*, the genome of which shows evidence of massive HGT from all three domains (Mira et al., 2004). The CDTOL resolves *Fusobacteria* as a basal Firmicutes with high support (100% BS, 301 BSI), a result consistent with recent analyses based on ribosomal 16S and 23S gene sequences (Mira et al., 2004), phosphoglycerate kinase (Pkg) amino acid sequences (Wolf et al., 2004), gene order (Kunisawa, 2006), and, to a lesser extent, indel analysis (Gupta and Griffiths, 2002).

Other relationships in the CDTOL are of particular note to the study of bacterial evolution. First, the thermophiles *Aquifex aeolicus* and *Thermotoga maritima* group with the Epsilonproteobacteria with high support, nested together with Mycoplasmales sister to the rest of the Firmicutes (Fig. 2). This result is inconsistent with the hypothesis, based on rDNA and concatenated protein sequence phylogenetic analyses (Brown et al., 2001) and ancestral reconstruction of amino acid composition (Di Giulio, 2003), that thermophiles are the last living universal common ancestors (LUCAs) of all life, but is consistent with the hypothesis that the Epsilonproteobacteria are not true proteobacteria (Dutilh et al., 2004; House and Fitz-Gibbon, 2002; Tekaia et al., 1999; Wolf et al., 2001, 2002; Gu and Zhang, 2004; Yang et al., 2005; Hughes et al., 2005; Lienau et al., 2006). Indeed, there has been much disagreement as to the proper classification of bacterial thermophiles in the ToL. A super-tree analysis, (Daubin et al., 2002) placed *Thermotoga maritima* and *Aquifex aeolicus* together on a long branch just basal to the Firmicutes (sometimes with the Epsilonproteobacteria) and a genome-wide scan classified these thermophiles as Firmicutes (Zhaxybayeva et al., 2009), results consistent with the hypothesis that the thermophiles are ancient, but not the *most* ancient, bacteria. Ribosomal RNA and PGK gene sequence analyses favour the hypothesis that *Thermotoga maritima* is just basal to the Firmicutes (Wolf et al., 2004), while indel parsimony analysis (Gupta and Griffiths, 2002), gene order analysis (Kunisawa, 2006), analysis of ribosomal protein sequences

(Klenk et al., 1999; Teeling et al., 2004), gene content (Wolf et al., 2001; Teeling et al., 2004) and RNA polymerase subunits (Teeling et al., 2004) place *Aquifex aeolicus* and or *Thermotoga maritima* just basal to the Proteobacteria, consistent with the hypothesis that thermophiles are basal Proteobacteria.

Our tree offers a reconciliatory explanation for these seemingly incompatible results in postulating that bacterial thermophiles are not living LUCAs and the Epsilonproteobacteria are not true Proteobacteria, but both are members of, or close relatives of, the Firmicutes. Still, given the multitude of inconsistent evidence, more information in the form of increased taxon sampling and character delineation will be necessary to better resolve these relationships.

The most basal group within the Bacteria clade (BS 100%, BSI 108) in this total evidence ToL is *Chlamydia* (100% BS, BSI 4838). The hypothesis that *Chlamydia* is an ancient bacterial lineage is consistent with analysis of bacterial SET domain architecture (Alvarez-Venegas et al., 2007), concatenated ribosomal proteins (Teeling et al., 2004; Wolf et al., 2001), concatenated protein sequences (Wolf et al., 2001), and gene presence/absence (Wolf et al., 2001; Lienau et al., 2006) but is inconsistent with indel parsimony analysis (Gupta and Griffiths, 2002) (although one indel out of 18 supported a grouping with the Archaea).

Comparing the CDTOL with other recent ToL analyses

The CD mega-matrix ToL and the Ciccarelli et al. (2006) ToL give similar hypotheses of the evolutionary history for life on Earth, each with high support for most groups, and both classify the three accepted domains of life as monophyletic. The remarkable agreement in the arrangement of the deepest relationships for these two ToLs derived from different data adds further credence, through corroboration of independent test results, to the accuracy of each hypothesis about the evolutionary history of life. In addition, the CD mega-matrix showed much higher and more even support throughout the tree than did the Ciccarelli et al. (2006) matrix (Fig. 4), indicating that the addition of data can help to bolster support for seemingly enigmatic relationships.

The CDTOL: reciprocal illumination of the ToL through data combination

When analysed both as presence/absence data (PD) and translated amino acid data (AD) in the combined (CD) mega-matrix the 67 109 empirically defined homologous gene groups used in this study yielded a totally resolved, highly consistent and well-supported hypothesis of the evolution of life using the information contained in 323 404 translated ORFs. This combined

data-type ToL (the CDToL) was more consistent, better supported and more congruent with the results of other, external analyses than the ToLs found using either data-type (AD or PD; Lienau et al., 2006) alone. In fact, pBSIs and lhBSI showed that a great proportion of the support for the CDToL was due to the simultaneous analysis of the two data types.

Taken together, these results indicate the existence of a strong phylogenetic signal that is common to all genes in this analysis that was revealed only by levying two types of phylogenetic test, those of gene loss and gain (HGT or duplication) and of amino acid sequence change (base change or insertion deletion), on competing ToL topologies. The revelation of this hidden signal through the simultaneous analysis of different data types can be thought of as reciprocal illumination (Hennig, 1966) not just among individual characters with independent histories, but also between different character classes with different modes of inheritance. In the absence of a better explanation for the existence of a hidden, highly corroborated signal that is common to all ORFs in this analysis, we suggest that these data corroborate Darwin and Wallace's ToL hypothesis of evolutionary history for the Eukarya, Archaea and Bacteria.

Acknowledgements

We thank Al Phillips Mark Siddall and Rich Baker for insightful discussion; Francisca Almeida, George Amato, Angelica Cibiran Jaramillio, Sergios-Orestis Kolokotronis, Matt Leslie, Cristina Pomilla and Ilya Temkin for reviewing an earlier version of this paper. We further thank three anonymous reviewers for their valuable criticism. This work was funded by grants from the US National Institutes of Health (to David Figurski, R.D. and P.J.P.). E.K.L. and J.A.R. were partially supported by a training grant from the US National Institutes of Health to New York University and E.K.L. and J.A.R. were further supported by a McKracken Fellowship of NYU. Our group also acknowledges the support of the Sackler Institute for Comparative Genomics and the Lewis and Dorothy Cullman Program in Molecular Systematics at the American Museum of Natural History. E.K.L. is now supported by a fellowship from the Oak Ridge Institute for Science Education to the Food and Drug Administration of the USA.

References

Alvarez-Venegas, R., Sadder, M., Tikhonov, A., Avramova, Z., 2007. Origin of the bacterial SET domain genes: vertical or horizontal? *Mol. Biol. Evol.* 24, 482–497.

- Baker, R.H., Desalle, R., 1997. Multiple sources of character information and the phylogeny of Hawaiian *Drosophilids*. *Syst. Biol.* 46, 654–667.
- Baker, R.H., Yu, X., Desalle, R., 1998. Assessing the relative contribution of molecular and morphological characters in simultaneous analysis trees. *Mol. Phylogenet. Evol.* 9, 427–436.
- Bapteste, E., Susko, E., Leigh, J., Ruiz-Trillo, I., Bucknam, J., Doolittle, W.F., 2008. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the Prokaryotic phylogeny. *Mol. Biol. Evol.* 25, 83–91.
- Beiko, R.G., Harlow, T.J., Ragan, M.A., 2005. Highways of gene sharing in prokaryotes. *Proc. Natl Acad. Sci. USA* 102, 14332–14337.
- Bremer, K., 1994. Branch support and tree stability. *Cladistics* 10, 295–304.
- Briones, C., Manrubia, S.C., Lazaro, E., Lazcano, A., Amils, R., 2005. Reconstructing evolutionary relationships from functional data: a consistent classification of organisms based on translation inhibition response. *Mol. Phylogenet. Evol.* 34, 371–381.
- Brochier, C., Philippe, H., 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* 417, 244.
- Brochier, C., Gribaldo, S., Zivanovic, Y., Confalonieri, F., Forterre, P., 2005. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol.* 6, R42.
- Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J., 2001. Universal trees based on large combined protein sequence data sets. *Nat. Genet.* 28, 281–285.
- Charlebois, R.L., Doolittle, W.F., 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* 14, 2469–2477.
- Ciccarelli, F.D., Doerks, T., Von Mering, C., Creevey, C.J., Snel, B., Bork, P., 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283–1287.
- Daubin, V., Gouy, M., Perriere, G., 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* 12, 1080–1090.
- Di Giulio, M., 2003. The universal ancestor and the ancestor of bacteria were hyperthermophiles. *J. Mol. Evol.* 57, 721–730.
- Ding, G., Yu, Z., Zhao, J., Wang, Z., Li, Y., Xing, X., Wang, C., Liu, L., Li, Y., 2008. Tree of life based on genome context networks. *PLoS ONE* 3, e3357.
- Doolittle, W.F., Bapteste, E., 2007. Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl Acad. Sci. USA* 104, 2043–2049.
- Dutilh, B.E., Huynen, M.A., Bruno, W.J., Snel, B., 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J. Mol. Evol.* 58, 527–539.
- Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.
- Eriksson, T., 2001. AutoDecay ver. 5.0. Bergius Foundation, Royal Swedish Academy of Sciences, Stockholm.
- Farris, J., 1989. The retention index and the rescaled consistency index. *Cladistics* 5, 417–419.
- Gatesy, J., 2002. Relative quality of different systematic datasets for cetartiodactyl mammals: assessments within a combined analysis framework. *EXS* 2002(92), 45–67.
- Gatesy, J., O'grady, P., Baker, R.H., 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics* 15, 271–313.
- Ge, F., Wang, L.-S., Kim, J., 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol.* 3, 1709–1718.
- Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T., 1989. Evolution of the vacuolar H⁺-ATPase: implications

- for the origin of eukaryotes. *Proc. Natl Acad. Sci. USA* 86, 6661–6665.
- Gu, X., Zhang, H., 2004. Genome phylogenetic analysis based on extended gene contents. *Mol. Biol. Evol.* 21, 1401–1408.
- Gupta, R.S., Griffiths, E., 2002. Critical issues in bacterial phylogeny. *Theor. Popul. Biol.* 61, 423–434.
- Harris, J.K., Kelley, S.T., Spiegelman, G.B., Pace, N.R., 2003. The genetic core of the universal ancestor. *Genome Res.* 13, 407–412.
- Hennig, W., 1966. *Phylogenetic Systematics* (translated by D. Dwight Davis and Rainer Zangerl). University of Illinois Press, Urbana, IL.
- House, C.H., Fitz-Gibbon, S.T., 2002. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J. Mol. Evol.* 54, 539–547.
- Hughes, A., Ekollu, V., Friedman, R., Rose, J., 2005. Gene family content-based phylogeny of prokaryotes: the effect of criteria for inferring homology. *Syst. Biol.* 54, 268–276.
- Iwabe, N., Kuma, K.I., Hagesawa, M., Osawa, S., Miyata, T., 1989. Evolutionary relationship of archaeobacteria, eubacteria, and the eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl Acad. Sci. USA* 86, 9355–9359.
- Jain, R., Rivera, M.C., Lake, J.A., 2002. Horizontal gene transfer among genomes: the complexity hypothesis. *PNAS* 96(7), 3801–3806.
- Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664.
- Klenk, H.P., Meier, T.D., Durovic, P., Schwass, V., Lottspeich, F., Dennis, P.P., Zillig, W., 1999. RNA polymerase of *Aquifex pyrophilus*: implications for the evolution of the bacterial rpoBC operon and extremely thermophilic bacteria. *J. Mol. Evol.* 48, 528–541.
- Koonin, E.V., Wolf, Y.I., 2009. The fundamental units, processes and patterns of evolution, and the tree of life conundrum. *Biol. Direct* 4, 33.
- Kunisawa, T., 2006. Dichotomy of major bacterial phyla inferred from gene arrangement comparisons. *J. Theor. Biol.* 239, 367–375.
- Lake, J.A., Rivera, M.C., 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol. Biol. Evol.* 21, 681–690.
- Lerat, E., Daubin, V., Moran, A., 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol.* 1, E19.
- Lienau, E.K., Desalle, R., 2008. Evidence content and corroboration in the tree of life. *ACTA Biotheor.* 57, 187–199.
- Lienau, E.K., Desalle, R., Rosenfeld, J.A., Planet, P.J., 2006. Reciprocal illumination in the gene content tree of life. *Syst. Biol.* 55, 441–453.
- Mcinerney, J.O., Cotton, J.A., Pisani, D., 2008. The prokaryotic tree of life: past, present... and future? *Trends Ecol. Evol.* 23, 276–281.
- Mira, A., Pushker, R., Legault, B.A., Moreira, D., Rodriguez-Valera, F., 2004. Evolutionary relationships of *Fusobacterium nucleatum* based on phylogenetic analysis and comparative genomics. *BMC Evol. Biol.* 4, 50.
- Mirkin, B.G., Fenner, T.I., Galperin, M.Y., Koonin, E.V., 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* 3, 2.
- Nixon, K.C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15, 407–414.
- Oh, S.J., Joung, J.G., Chang, J.H., Zhang, B.T., 2006. Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. *BMC Bioinformatics* 7, 284.
- Pace, N.R., Olsen, G.J., Woese, C.R., 1986. Ribosomal RNA phylogeny and the primary lines of evolutionary descent. *Cell* 45, 325–326.
- Pisani, D., Cotton, J.A., Mcinerney, J.O., 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol. Biol. Evol.* 24, 1752–1760.
- Puigbo, P., Wolf, Y.I., Koonin, E.V., 2009. Search for a ‘Tree of Life’ in the thicket of the phylogenetic forest. *J. Biol.* 8, 59.
- Rivera, M.C., Jain, R., Moore, J.E., Lake, J.A., 1998. Genomic evidence for two functionally distinct gene classes. *Proc. Natl Acad. Sci. USA* 95, 6239–6244.
- Rohlf, F., 1982. Consensus indices for comparing classifications. *Math. Biosci.* 59, 131–144.
- Rokas, A., Carroll, S.B., 2006. Bushes in the tree of life. *PLoS Biol* 4, e352.
- Sikes, D.S., Lewis, P.O., 2001. PAUPRat: PAUP* implementation of the parsimony ratchet, version 1. Available from the authors at http://users.iab.uaf.edu/~derek_sikes/software2.htm.
- Swofford, D., 2000. PAUP*: Phylogenetic Analysis Using Parsimony (* and other methods). Version 4.0 b10. 4.0b10 ed. Sinauer Associates, Sunderland, MA.
- Teeling, H., Lombardot, T., Bauer, M., Ludwig, W., Glockner, F.O., 2004. Evaluation of the phylogenetic position of the planctomycete ‘*Rhodopirellula baltica*’ SH 1 by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees. *Int. J. Syst. Evol. Microbiol.* 54, 791–801.
- Tekaia, F., Yeramian, E., 2005. Genome trees from conservation profiles. *PLoS Comput. Biol.* 1, e75.
- Tekaia, F., Lazzcano, A., Dujon, B., 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* 9, 550–557.
- Woese, C.R., 1987. Bacterial evolution. *Microbiol. Rev.* 51, 221–271.
- Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA* 87, 4576–4579.
- Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L., Koonin, E.V., 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* 1, 8.
- Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Koonin, E.V., 2002. Genome trees and the tree of life. *Trends Genet.* 18, 472–479.
- Wolf, M., Muller, T., Dandekar, T., Pollack, J.D., 2004. Phylogeny of Firmicutes with special reference to *Mycoplasmata* (*Mollicutes*) as inferred from phosphoglycerate kinase amino acid sequence data. *Int. J. Syst. Evol. Microbiol.* 54, 871–875.
- Yang, S., Doolittle, R.F., Bourne, P.E., 2005. Phylogeny determined by protein domain content. *Proc. Natl Acad. Sci. USA* 102, 373–378.
- Zhaxybayeva, O., Swithers, K.S., Lapierre, P., Fournier, G.P., Bickhart, D.M., Deboy, R.T., Nelson, K.E., Nesbo, C.L., Doolittle, W.F., Gogarten, J.P., Noll, K.M., 2009. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proc. Nat. Acad. Sci.* 106, 5865–5870.