BMC Genomics

BioMed Central

Methodology article

Open Access

An improved probability mapping approach to assess genome mosaicism

Olga Zhaxybayeva and J Peter Gogarten*

Address: Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Storrs, CT, 06269-3125, USA

Email: Olga Zhaxybayeva - olga@carrot.mcb.uconn.edu; J Peter Gogarten* - gogarten@uconn.edu

* Corresponding author

Published: 15 September 2003

BMC Genomics 2003, **4**:37

Received: 19 June 2003

Accepted: 15 September 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/37

Abstract

Background: Maximum likelihood and posterior probability mapping are useful visualization techniques that are used to ascertain the mosaic nature of prokaryotic genomes. However, posterior probabilities, especially when calculated for four-taxon cases, tend to overestimate the support for tree topologies. Furthermore, because of poor taxon sampling four-taxon analyses suffer from sensitivity to the long branch attraction artifact. Here we extend the probability mapping approach by improving taxon sampling of the analyzed datasets, and by using bootstrap support values, a more conservative tool to assess reliability.

Results: Quartets of orthologous proteins were complemented with homologs from selected reference genomes. The mapping of bootstrap support values from these extended datasets gives results similar to the original maximum likelihood and posterior probability mapping. The more conservative nature of the plotted support values allows to focus further analyses on those protein families that strongly disagree with the majority or plurality of genes present in the analyzed genomes.

Conclusion: Posterior probability is a non-conservative measure for support, and posterior probability mapping only provides a quick estimation of phylogenetic information content of four genomes. This approach can be utilized as a pre-screen to select genes that might have been horizontally transferred. Better taxon sampling combined with subtree analyses prevents the inconsistencies associated with four-taxon analyses, but retains the power of visual representation. Nevertheless, a case-by-case inspection of individual multi-taxon phylogenies remains necessary to differentiate unrecognized paralogy and shared phylogenetic reconstruction artifacts from horizontal gene transfer events.

Background

The analysis of four-taxon trees promises to provide valuable insight and visual documentation of genome mosaicism [1–5]. However, like other four-taxon analyses, our probability mapping approach for comparative genome analyses [4] is vulnerable to the long branch attraction (LBA) artifact because it analyzes datasets consisting of only four sequences. LBA is a well-known phylogenetic artifact [6]. It is especially well studied for the case of four-taxon trees (e.g., see [7–11]). In short, regardless of the

reconstruction method and model used, if the branches are long enough, the reconstructed tree might be affected by LBA although to different degrees. Furthermore, four-taxon analyses were shown to be instable and misleading under some circumstances [12,13]. Addition of more taxa can break up the long branches and increases reliability. Simulation studies have shown that increase of the size of a dataset by introducing additional homologous sequences improves the accuracy of the reconstruction [14] (see [15] and [16] for the recent discussion). An increase in the sequence lengths of the analyzed data also can improve the reliability of phylogenetic reconstruction [16], but lumping different putative orthologs into a single dataset would defeat the purpose of the probability mapping approach, i.e., the detection of genes that have incompatible evolutionary histories. Merging proteins with different histories into concatenated datasets would not help to resolve their phylogenies.

Here we report an extension of probability mapping that increases the number of homologous sequences per dataset, throughout the rest of the article referred as Operational Taxonomic Unit (OTU) sampling, but retains the power to visualize genomic mosaicism from the original approach. A quartet of orthologous proteins (QuartOP) is defined as four homologs from four genomes that pick each other as top-scoring reciprocal hits in BLAST searches of the respective genomes (for more details see [4]). For each QuartOP detected in a genome quartet we add homologous sequences and evaluate the branching order of the QuartOP in 100 bootstrap samples. The bootstrap support values then are mapped into a barycentric coordinate system. We compare the mapping results with previously reported ones [4], and give examples that illustrate the utility of this approach in detecting horizontally transferred genes.

Results and Discussion

Interdomain Genome Quartets

In [4] we described the analyses of several interdomain genome quartets. Some of the analyses were performed using a posterior probability mapping approach referred to as Maximum Likelihood (ML) mapping, a name that was coined in the original description of this approach [17]. We will use this term throughout the manuscript. In ML mapping posterior probabilities are calculated from the maximum likelihood values (see [17] and [4] for the details). One noteworthy finding was that in the genome quartet including *Synechocystis* sp., *Halobacterium* sp., *Aquifex aeolicus* and *Thermotoga maritima* the grouping of *Halobacterium* sp. with *Synechocystis* sp. was recovered by many more QuartOPs than the grouping expected following 16S rRNA phylogeny (see Fig. 1A). Note that throughout the manuscript we refer to a particular tree by mentioning two species out of four (e.g., in this case grouping of *Halobacterium* sp. with *Synechocystis* sp.); however, the trees are unrooted and therefore grouping of the other two taxa is implied. To test if this association was specific for *Synechocystis* sp., we had repeated the analyses replacing *Synechocystis* sp. with *Bacillus subtilis*. The results were qualitatively the same (data not shown). To test for the possibility that LBA [6] might be the reason for the strong support of *Halobacterium* sp. grouping with *Synechocystis* sp., we had repeated the analyses replacing the *Halobacterium* sp. genome with that from *Archaeoglobus fulgidus*, another archaeon. The majority of QuartOPs supported the grouping of the thermophilic archaeon *Archaeoglobus* with the thermophilic bacteria *Aquifex* and *Thermotoga* (see Fig. 1B). In this study, we reanalyzed the above-mentioned genome quartets by adding homologous sequences from sixty reference genomes to each QuartOP creating what we call "extended datasets". The dataflow is depicted in Figure 2. For each extended dataset we obtained bootstrap support values for each of the three four-taxon "subtrees" and we plotted the bootstrap support values into barycentric coordinates. Throughout this manuscript we use a graph theory definition of a subtree, i.e. "A tree G' whose graph vertices and graph edges form subsets of the graph vertices and graph edges of a given tree G" [18]. In particular, sequences (OTUs) included in the subtree are not required to be neighbors in the original tree. Subtrees defined according to these rules are different from subclades (see figure 2 for an illustration). For example, if the topology ((A,D),(B,C)) is supported by a given bootstrap sample, this means that in the tree calculated from this sample the sequence from genome A groups closer to the one from D than to the one from B or C (figure 2).

Maps of bootstrap support values calculated from the extended datasets are shown in figure 3. The results are similar to the analyses using ML mapping (see [4] and figure 1). At every level of support the plurality consensus groups the mesophilic archaeon with the mesophilic bacterium and the two thermophilic bacteria with one another (figure 1 and 3, panels A). The plurality support changes in favor of the archaeon – *Aquifex* grouping, when the genome of a mesophilic archaeon is replaced with that of an extremely thermophilic archaeon *Archaeoglobus fulgidus* (figure 1 and 3, panels B). The main difference between ML maps and bootstrap support maps from extended datasets is that the confidence values are much lower for the extended datasets evaluated with bootstrap. The cause for these lower support values is discussed below.

Interdomain transfer, interphylum transfer, or shared artifact?

The relation between the different bacterial phyla, and the placement of the bacterial root remains uncertain (e.g.,

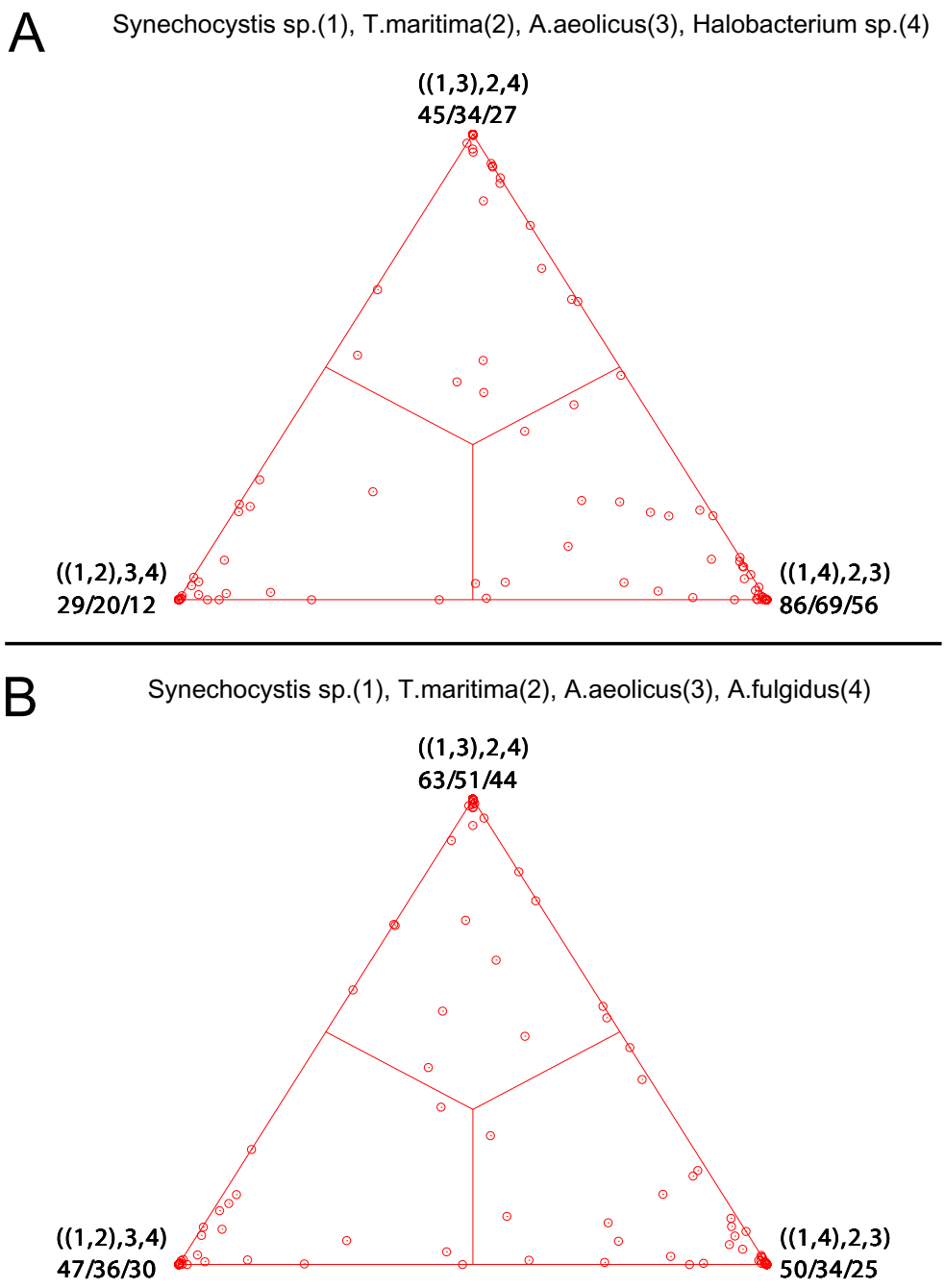

Figure 1

Posterior probability maps of genome quartets containing *Synechocystis* sp. Posterior probabilities were calculated according the maximum likelihood mapping approach described in [4,17]. Tree topologies assigned to the vertices are depicted in New Hampshire tree format near the corresponding vertex of the triangle and they should be considered as unrooted tree topologies. The three numbers associated with each tree topology indicate how many QuartOPs fall into each of the three zones: "total" (i.e. posterior probability for the tree topology is larger than posterior probabilities for the other two topologies), 90% and 99% posterior probability respectively. **A)** Genome quartet consisting of *Synechocystis* sp., *Halobacterium* sp., *Aquifex aeolicus* and *Thermotoga maritima*. The majority of the QuartOPs support the grouping of the *Halobacterium* sp. with *Synechocystis* sp. **B)** Genome quartet consisting of *Synechocystis* sp., *Archaeoglobus fulgidus*, *Aquifex aeolicus* and *Thermotoga maritima*. The archaeon – *Synechocystis* sp. grouping is supported by fewer QuartOPs than in panel A.

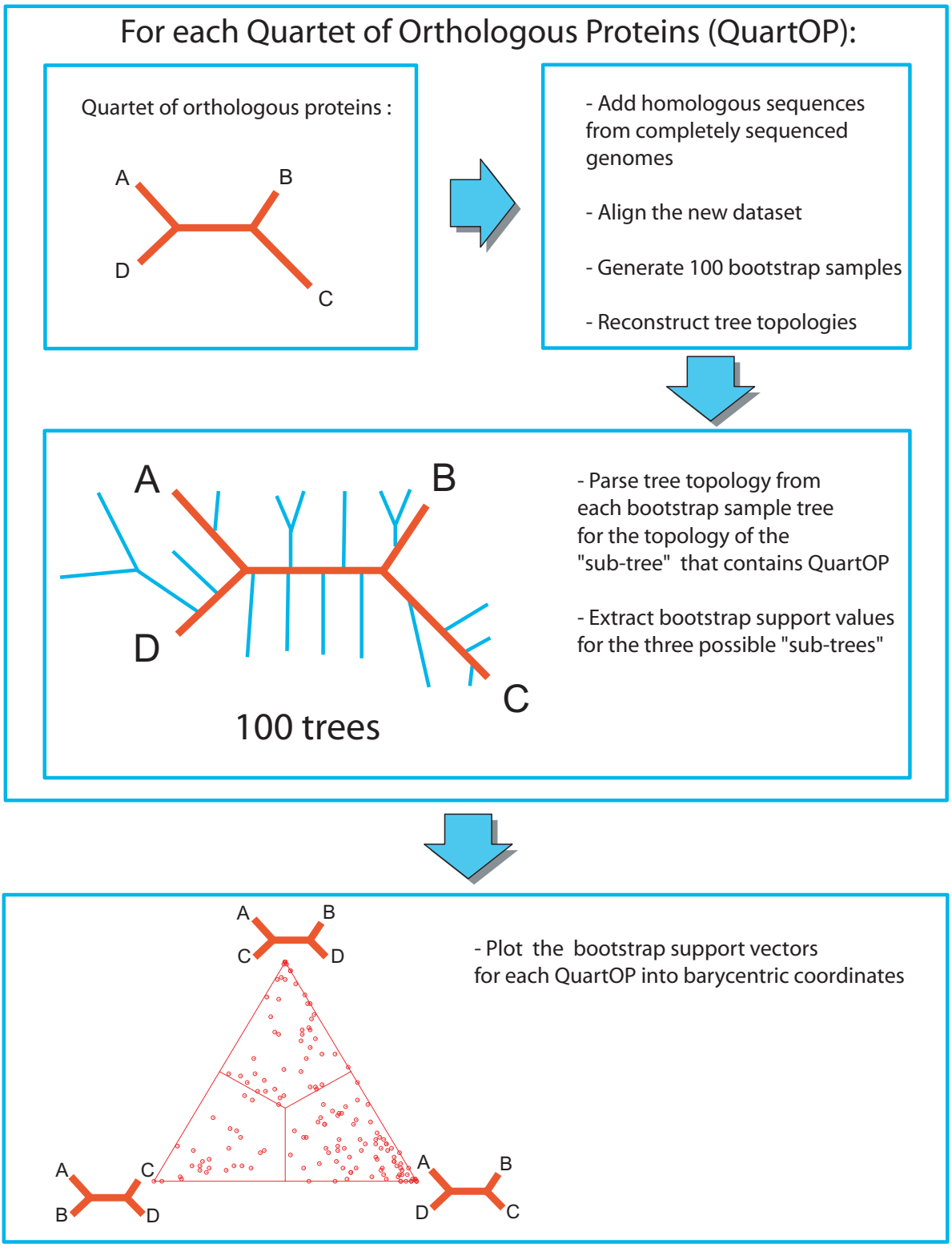

Figure 2
Dataflow for construction and mapping of extended datasets for all QuartOPs in a genome quartet. See Materials and Methods for details.

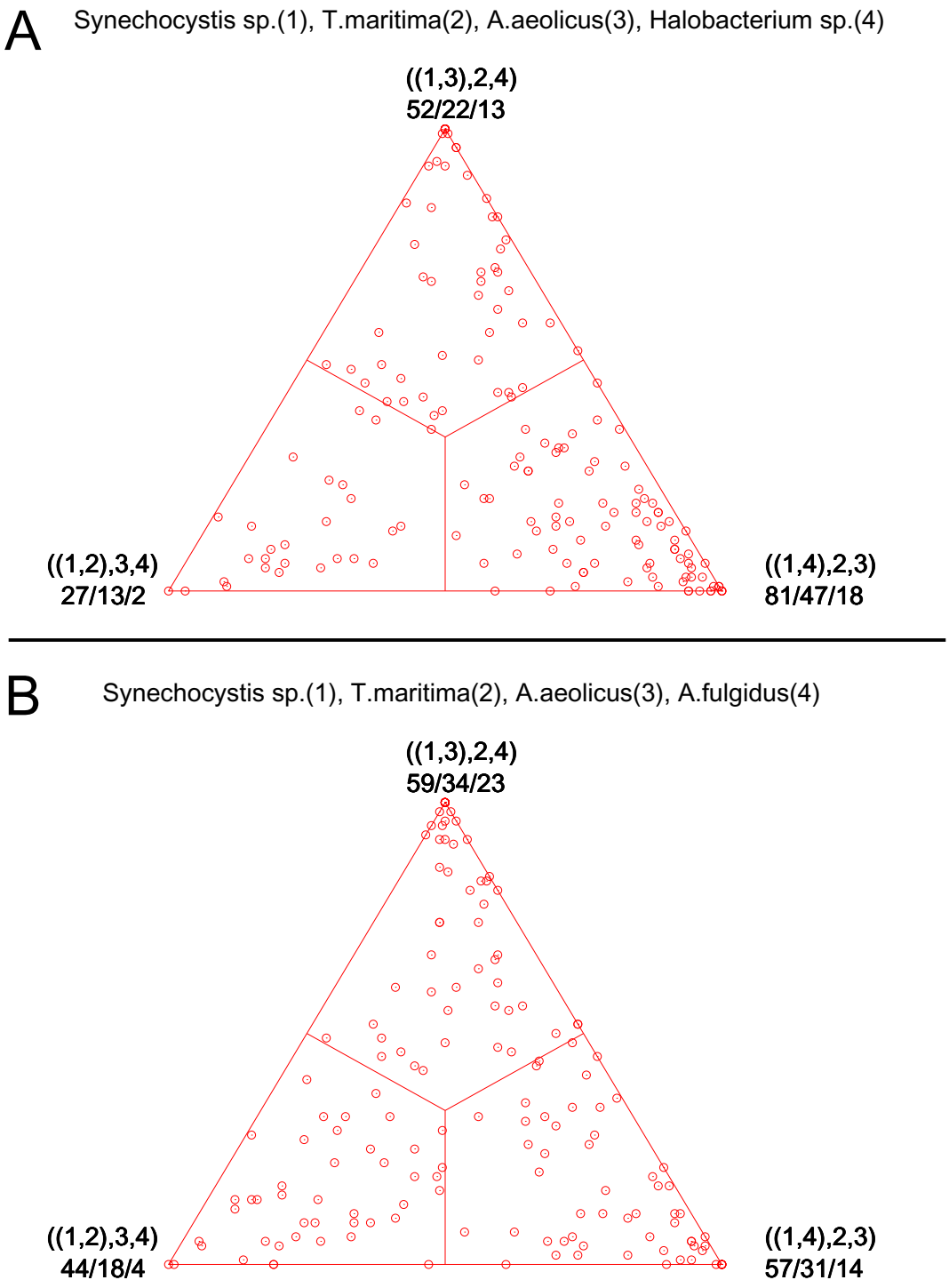

Figure 3
Maps of bootstrap support from extended datasets for genome quartets containing *Synechocystis* sp. The three numbers associated with each tree topology indicate how many QuartOPs fall into each of the three zones: "total", 70% and 90% bootstrap support respectively. For other figure notations see figure 1. **A)** Genome quartet consisting of *Synechocystis* sp., *Halobacterium* sp., *Aquifex aeolicus* and *Thermotoga maritima*. **B)** Genome quartet consisting of *Synechocystis* sp., *Archaeoglobus fulgidus*, *Aquifex aeolicus* and *Thermotoga maritima*. These maps are similar to the ML maps depicted in figure 1.

see [19]). Therefore, it is not clear which of the three unrooted trees for the genome quartet in question represents the true organismal phylogeny of the four genomes. The lack of phylogenetic signal alone should result in QuartOPs that map to the center of the triangle. However, we observe many genes that prefer one topology to the other two. The genes in figures 1A and 3A that group the ortholog from *Halobacterium* with its putative ortholog from *Synechocystis* might do so for a variety of different reasons:

A) Horizontal gene transfer

A_1) between a mesophilic bacterium and a mesophilic archaeon

A_2) between the extremely mesophilic bacteria *Aquifex* and *Thermotoga*

B) Phylogenetic reconstruction artifacts

B_1) due to long branch attraction

B_2) due to compositional bias

B_3) due to the lack of phylogenetic signal

C) Unrecognized paralogy

D) This grouping reflects organismal evolutionary history.

It is important to realize that any of the processes listed under A through C, alone or in combination, might result in well supported QuartOPs that group the halobacterial homolog with the cyanobacterial counterpart. Some of these possibilities can be distinguished in the individual phylogenies of the extended datasets. Table 1 summarizes these findings. There are many datasets which don't conform to the 16S rRNA based expectation in more than one respect: *Aquifex* and *Thermotoga* are recovered as sister groups, and either the cyanobacterial sequences groups within a cluster of Archaeal sequences, or the halobacterial sequence groups within a well-supported cluster of Bacterial homologs. We considered these unexpected phylogenies as supported, if the branch that separated this group from the other homologs had at least 70% bootstrap support (note: in this case we used the bootstrap support for individual branches, not the support for subtrees). At first sight this might appear as a rather low level of support; however, adding more sequences tends to shorten the internal branches and thus lowers their individual bootstrap support value (for discussion see [20]). The support for the subtree topology (see figure 2) is larger than 90% for all entries. There is only one case where the observed subtree ((*Aquifex*, *Thermotoga*), (*Halobacterium*, *Synechocystis*)) appears to be due to unrecognized paralogy. In four instances a well-supported branch suggests an interdomain gene transfer and in 11 instances an exchange between *Thermotoga* and *Aquifex*. These findings apparently support the notion that genes are frequently transferred between divergent prokaryotes; however, *Halobacterium* has an amino acid composition that often deviates significantly from that of the other homologs. The instances where both the halobacterial sequence and its phylogenetic neighbor failed a test for homogeneous composition are indicated in table 1. This leaves eight well-supported instances of at least one horizontal gene transfer event out of the twelve datasets without shared compositional biases.

The analyses depicted in figure 3 and table 1 demonstrate that bootstrap support value mapping in general, and support value mapping using extended datasets in particular, are useful in screening for genes that were transferred between divergent organisms. Replacing the genome from a mesophilic archaeon with that from an extremely thermophilic one, changes the topology of the subtree that has plurality support. This observation is in agreement with the hypothesis that genes are more frequently shared between organisms that live in similar environments [21]. However, given that the *Halobacterium* genome is renowned for its large number of genes with bacterial character [22,23], the total number of genes identified in this study as putatively transferred between the mesophilic bacteria and the halobacteria is very small. There are several reasons for this observation. Useful phylogenetic information retained in molecular sequences is constantly overwritten by more recent substitution events. The more divergent the analyzed genomes are, the more QuartOPs will be undecided about the most supported topology. Furthermore, support value mapping can only identify gene transfers that resulted in orthologous replacement. Last but not least, the applied approach to assemble QuartOPs is overly restrictive. Lineage specific duplications result in two orthologs being present in a single genome. These genes are paralogs of one another, but both are orthologs to the gene present in the genomes that branch off before the lineage specific duplication [24]. Despite these shortcomings, support value mapping, especially when using extended datasets, provides a quick method to appraise the extent of genomic mosaicism, to delineate preliminarily the major flows of genes in microbial evolution (plurality or majority consensus), and to find subsets of potentially transferred genes.

Screen for more recent interphylum transfers

To assess the utility of probability and bootstrap support values mapping for detecting more recent interphylum gene transfer events, we calculated extended datasets for the genome quartet of *Synechocystis* sp., *Chlorobium*

Table 1: Analyses of the tree topologies for 18 candidates for horizontal gene transfer. Support for putative transfers between Bacteria and Archaea is shown in the B&A column, and between *Aquifex* and *Thermotoga* is shown in the A&T column. The compositional bias is listed as "strong", if both the halobacterial sequence and its nearest phylogenetic neighbor failed the test for homogeneous composition. See Materials and Methods for details on performed analyses.

ID	Functional Assignment	B&A	A&T	Comments	Compositional bias
008	thymidylate kinase	weak	strong	-	-
020	seryl-tRNA synthetase	strong	strong	-	-
054	valyl-tRNA synthetase	none	strong	-	Strong
062	excision nuclease chain A	none	strong	-	Strong
072	chromosome segregation SMC protein	strong	None	Cyanobacteria, *Rickettsia* and *Aquifex* group within Archaea	Strong
076	hypothetical protein	weak	strong	*Halobacterium* groups within Bacteria	Strong
080	prolyl-tRNA synthetase	none	strong	Archaeal type homologs are found in some bacteria	-
100	hypothetical protein	-	-	Uninterpretable: no resemblance with assumed organismal phylogeny	-
105	DNA gyrase, subunit B	none	strong	Archaea do not form a group	-
106	arginyl-tRNA synthetase	strong	weak	-	Strong
107	DNA gyrase, subunit A	none	weak	Both *Thermotoga* and *Aquifex* group with Archaea	-
110	cysteinyl-tRNA synthetase	strong	strong	Both *Thermotoga* and *Aquifex* group within Archaea	-
113	hypothetical protein	weak	none	Both *Thermotoga* and *Aquifex* group within Archaea	Strong
121	chorismate synthase	weak	none	-	-
122	3-phosphoshikimate-1-carboxyvinyltransferase	none	strong	-	-
134	histidinol dehydrogenase	weak	weak	Putative paralogs	-
144	50S ribosomal protein L2	none	strong	-	-
145	30S ribosomal protein S19	weak	strong	-	-

tepidum, *Rhodobacter capsulatus* and *Rhodopseudomonas palustris* (see figure 4 and [3]). This genome quartet has a strong phylogenetic signal grouping together the two alpha proteobacteria *R. capsulatus* and *R. palustris*. This example had previously been utilized to demonstrate the validity of ML mapping [3] showing that the vast majority of QuartOPs group the proteins from the two more closely related organisms together. However, there are 14 QuartOPs that support the two alternative topologies with 99% posterior probability. These have to be regarded as candidates for horizontal gene transfer. Analysis of this genome quartet using extended datasets shows that some of these 14 QuartOPs are also supported by high bootstrap support values (above 90%). Figures 5 through 8 provide further analysis of the extended datasets for these QuartOPs. The cases of the cation-transporting ATPases (figure 5) and the hypothetical proteins depicted in figure 6 probably represent unrecognized paralogies. The proteins from *R. palustris* and *R. capsulatus* each group together with homologs from other alpha proteobacteria, and in some instances a single genome encodes both paralogs (*Bradyrhizobium japonicum* in case of hypothetical protein family and *Sinorhizobium meliloti* in case of the cation-transporting ATPases). It appears likely that *R. palustris* has lost one and *R. capsulatus* the other paralog. In these two instances the unexpected behavior of the QuartOPs is due to failure of the strategy to select orthologous genes. In contrast, the cases of the water channel protein family and the methionyl-tRNA synthetases are best explained by horizontal gene transfer. None of the reference genomes contains paralogs whose differential loss might explain the observed phylogenies (figures 7 and 8).

The higher frequency of unrecognized paralogs among the putatively horizontally transferred genes is due to the much larger number of QuartOPs analyzed. The detected number of unrecognized paralogs corresponds to less than 1% of the QuartOPs that contain sufficient phylogenetic information to support a topology with more than 90% bootstrap support (figure 4C). Every instance of unrecognized paralogy will result in a QuartOP deviating from the majority consensus revealed in this analysis.

Loss of support strength: due to conservative measure or taxon sampling?

It is difficult to compare posterior probabilities of QuartOPs directly with bootstrap support values of much larger datasets. Empirical studies [4] as well as the simula-

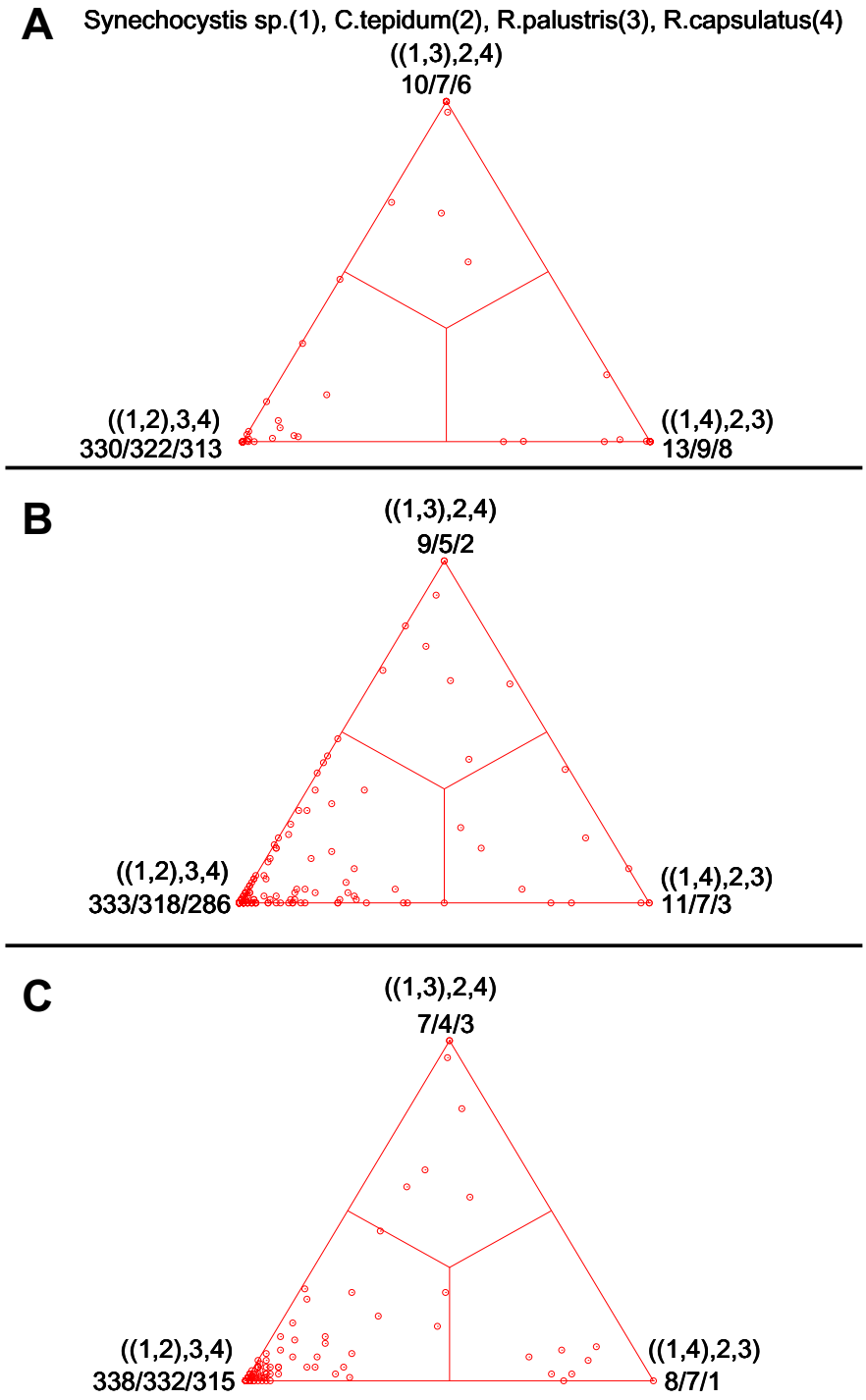

Figure 4
Genome quartet of *Synechocystis* sp., *Chlorobium tepidum*, *Rhodobacter capsulatus* and *Rhodopseudomonas palustris*. A) Posterior probability map calculated using probability mapping as described in [4,17]. **B)** Bootstrap support map (see [4] for methodology of bootstrap support map reconstruction). Only the four putatively orthologous sequences were utilized in the analyses. **C)** Bootstrap support map from extended datasets. For details on the figure notations see legends for figures 1 and 3. The majority of QuartOPs support one tree topology grouping two alpha proteobacteria together. The QuartOPs located in the two other corners of the triangle are candidates for horizontal gene transfer.

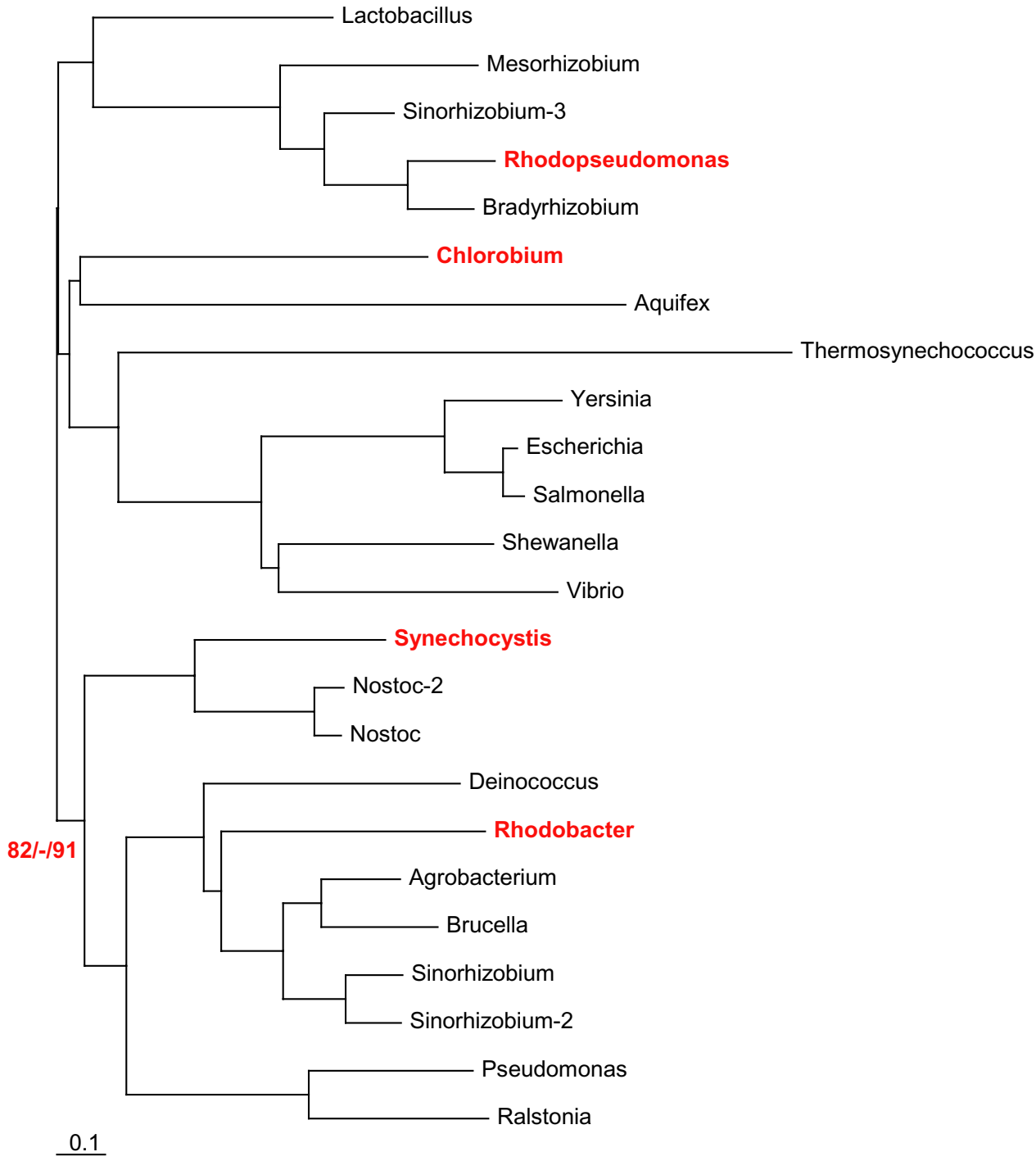

Figure 5
Phylogeny of homologs of cation-transporting ATPases. Members of the QuartOP are highlighted in red. The three support values indicated on a branch are bootstrap support values from Neighbor-Joining trees based on TREE-PUZZLE distances, bootstrap support values from parsimony analyses and posterior probabilities from Bayesian analyses respectively. See Materials and Methods for details on phylogenetic reconstruction. The finding of other homologs from alpha proteobacteria grouping with the *Rhodopseudomonas* and *Rhodobacter* sequences, and the finding of both homologs coexisting in the same genome (*Sinorhizobium*) suggests that this QuartOP represents a case of unrecognized paralogy with differential gene loss.

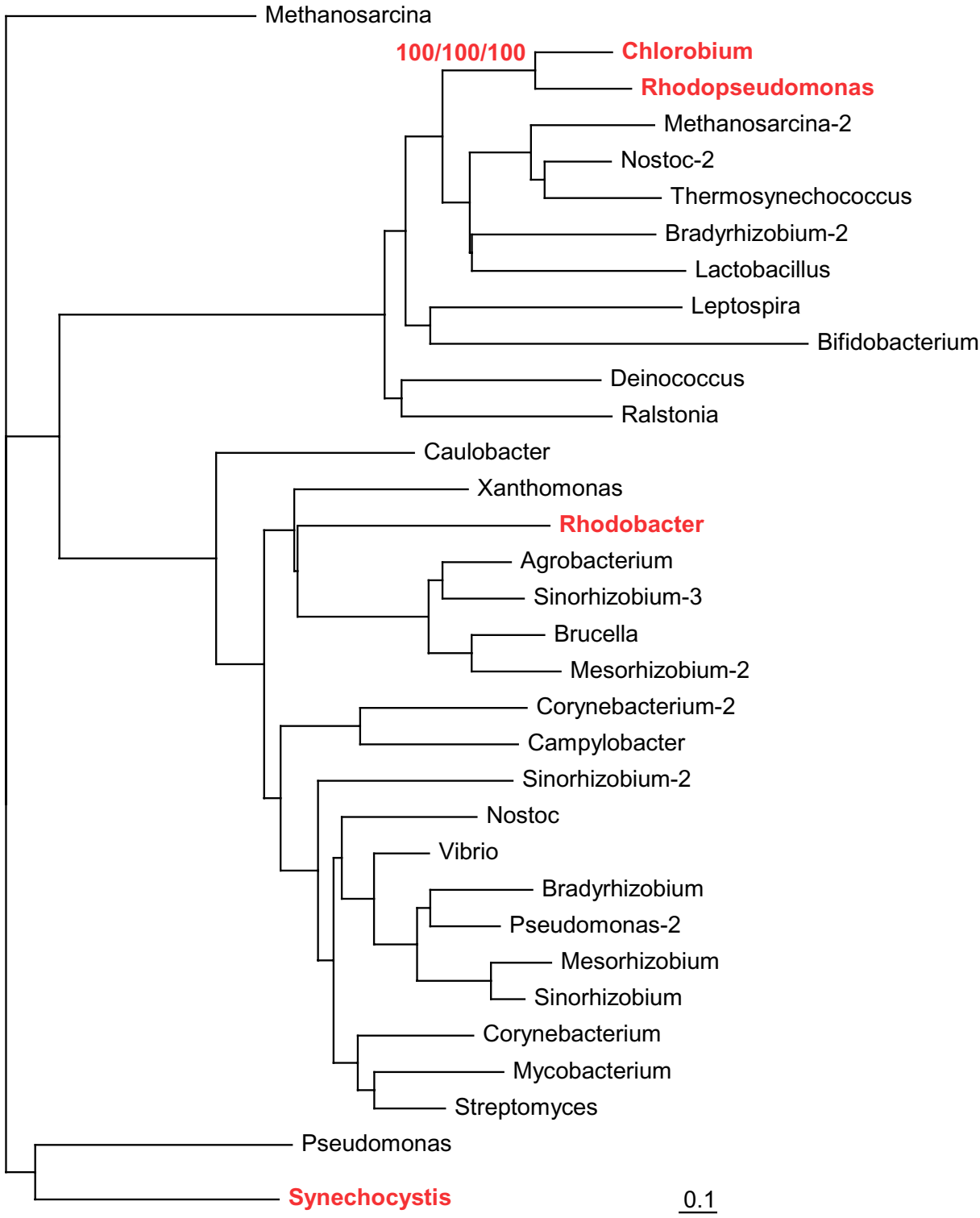

Figure 6
Phylogeny of hypothetical protein homologs. See figure 5 for notations and Materials and Methods for details on phylogenetic reconstruction. This QuartOP represents another likely example of unrecognized paralogy. See text and figure 5 for discussion.

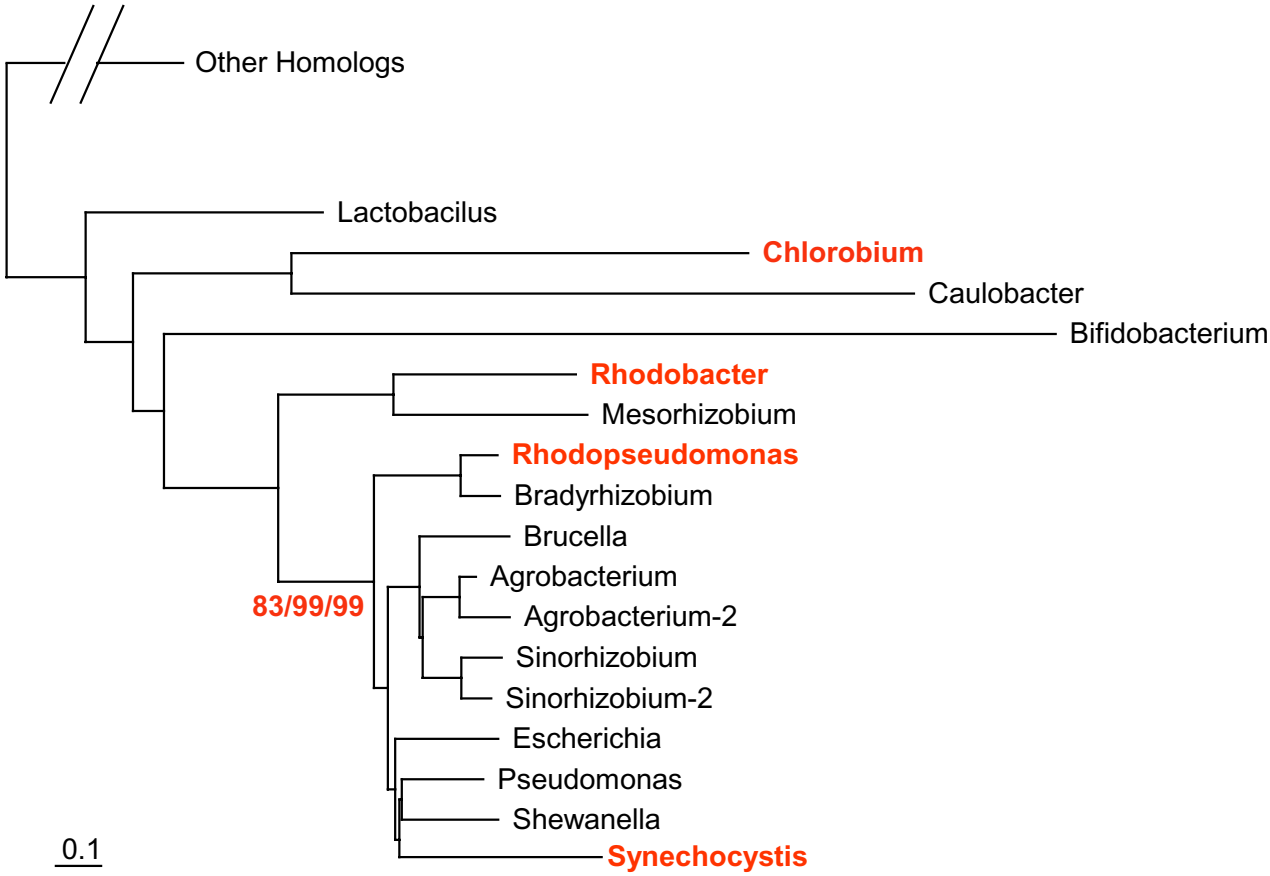

Figure 7
Phylogeny of water channel proteins family. See figure 5 for notations and Materials and Methods for details on phylogenetic reconstruction. The two homologs from *Rhodobacter* and *Rhodopseudomonas* are separated by a well-supported branch. The homologs grouping with *Rhodobacter* do not have a paralog among the homologs grouping with *Rhodopseudomonas* and vice versa. This QuartOP remains a strong candidate for horizontal gene transfer.

tion studies [25,26] indicate that bootstrap measures are much more conservative than Bayesian posterior probabilities. In the four-taxon cases analyzed in [4] a posterior probability of 0.99 calculated according to Strimmer and von Haeseler [17] was found to correspond to only 70% bootstrap support calculated from non-extended datasets.

To demonstrate that the observed drop in support is due to the more conservative nature of bootstrapping and not due to increased OTU sampling we re-analyzed the same genome quartets using ML mapping (figure 1 and figure 4A), bootstrap support values calculated from only the four aligned QuartOPs (Additional file: 2 and figure 4B), and bootstrap support values calculated from the extended datasets (figure 3 and figure 4C) as described in figure 2. Note that in case of ML mapping only posterior probabilities greater than 0.99 are counted as strong support (greater than 0.9 for moderate support), whereas in case of the bootstrap support value maps greater than 0.9 is classified as strong support (greater than 0.7 for moderate support). Table 2 summarizes the overall number of QuartOPs supporting different topologies with different measures of support. There is a dramatic drop in the number of QuartOPs with strong support from the 99% posterior probabilities to 90% bootstrap from non-extended datasets. However, the added accuracy obtained through increased OTU sampling does not change the support as radically as the shift from posterior probabili-

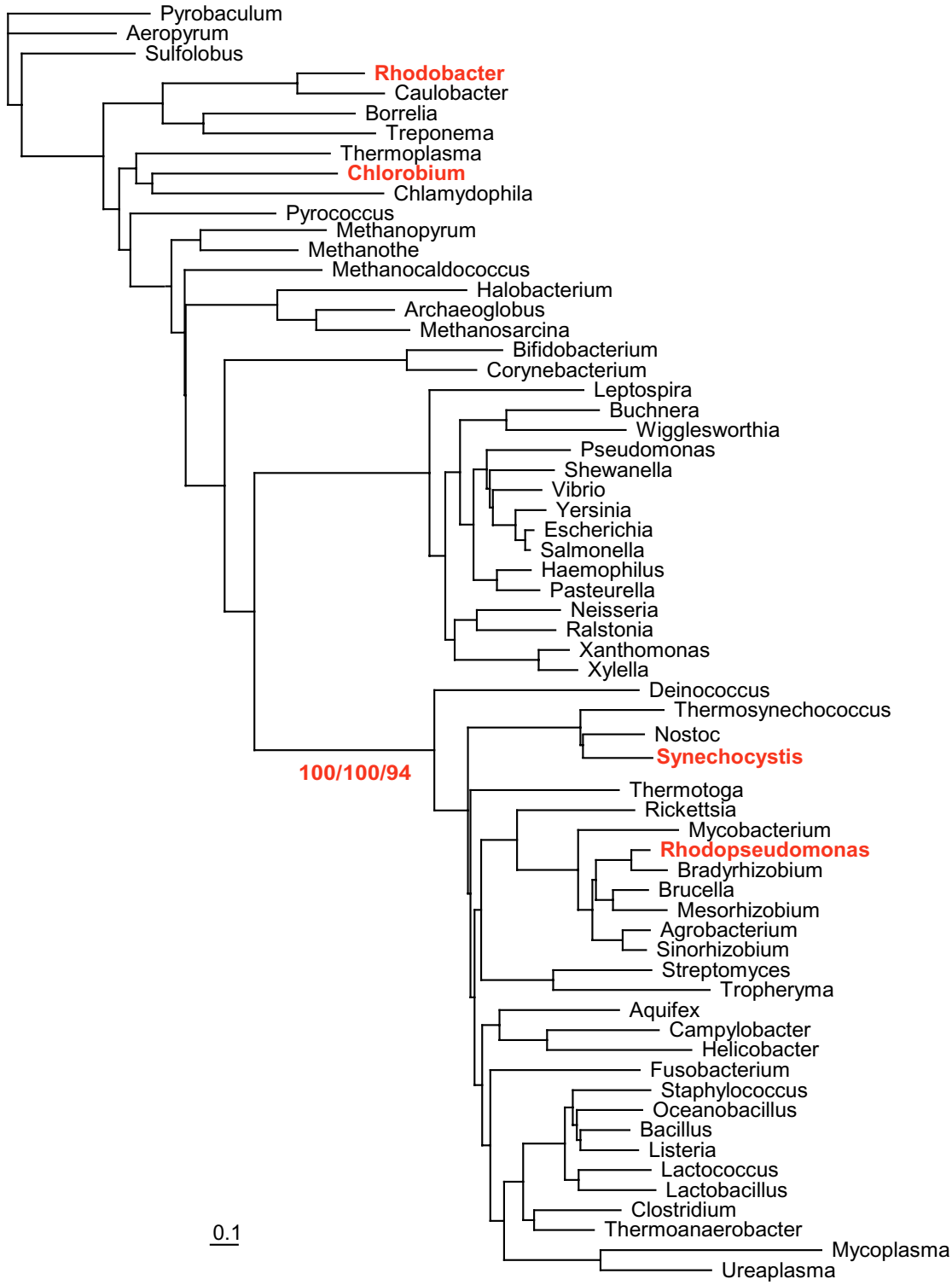

Figure 8

Phylogeny of methionyl-tRNA synthetases. See figure 5 for notations and Materials and Methods for details on phylogenetic reconstruction. No paralogs were detected in the reference genomes. This QuartOP remains a strong candidate for horizontal gene transfer.

Table 2: Comparison of confidence levels for different types of mappings. Table entries give the numbers of QuartOPs in the indicated genome quartets that prefer one of the three tree topologies with the specified level of support.

Genome Quartet	99% posterior probability	90% bootstrap support from non-extended datasets	90% bootstrap support from extended datasets
Interdomain quartet with *Halobacterium* (see figure 1A, 3A and suppl. material)	95	42	33
Interdomain quartet with *Archaeoglobus* (see figure 1B, 3B and suppl. material)	99	42	41
Interphylum quartet (see figure 4)	327	291	319

ties to the bootstrap support measure. Apparently, the increased accuracy due to better OTU sampling on average increases the bootstrap support of a given subtree as often as it lowers the support value. The higher support values found for ML mapping are solely due to the less conservative nature of the calculated support measure.

Conclusions

The original posterior probability mapping methods reported in [4] return results similar to those obtained from the analyses of extended datasets. ML mapping is much faster than the bootstrap support values mapping of extended datasets reported here. In interpreting results, however, one needs to be aware of the non-conservative nature of the posterior probability mapping approach, and of the greater susceptibility of four-taxon analyses to the long branch attraction artifact. The faster ML mapping approach has utility as a quick estimation of phylogenetic information content of four genomes. Even though ML mapping greatly overestimates reliability, our results illustrate the utility of ML mapping as a pre-screen for putative horizontal gene transfer events. The use of extended datasets combined with subtree analyses prevents the inconsistencies associated with four-taxon analyses, but retains the power of visual representation. However, even an increase in OTU sampling and the simultaneous use of a more conservative probability measure does not obviate the need to inspect the phylogenies of candidate genes to detect instances of unrecognized paralogy. Given the public availability of over 100 prokaryotic genomes, appropriate reference genomes can be selected in most instances to distinguish differential loss of paralogs from horizontal gene transfer events.

Methods

The methodology of obtaining QuartOPs for four genomes is described in [4]. For each sequence in a QuartOP we detect the top-scoring BLAST [27] hit with an E-value above 10^{-8} in each of 60 completely sequenced archaeal and bacterial reference genomes (*Aeropyrum pernix*, *Archaeoglobus fulgidus*, *Anabaena* sp., *Aquifex aeolicus*, *Agrobacterium tumifaciens*, *Borrelia burgdorferi*, *Bradyrhizobium japonicum*, *Bifidobacterium longum*, *Bacillus subtilis*, *Brucella suis*, *Buchnera* sp., *Clostridium acetobutylicum*, *Caulobacter crescentus*, *Corynebacterium glutamicum*, *Campylobacter jejuni*, *Clamydophila pneumoniae*, *Deinococcus radiodurans*, *Escherichia coli* K12, *Fusobacterium nucleatum*, *Halobacterium* sp., *Haemophilus influenzae*, *Helicobacter pylori*, *Leptospira interrogans*, *Lactococcus lactis*, *Listeria monocytogenes*, *Lactobacillus plantarum*, *Mycoplasma genitalium*, *Methanococcus jannaschii*, *Methanopyrus kandleri*, *Mezorhizobium loti*, *Methanosarcina mazei*, *Methanobacterium thermoautotrophicum*, *Mycobacterium tuberculosis*, *Neisseria meningitides*, *Oceanobacillus iheyensis*, *Pseudomonas aeruginosa*, *Pyrobaculum aerophilum*, *Pyrococcus horikoshii*, *Pasteurella multocida*, *Rickettsia conorii*, *Ralstonia solanacearum*, *Staphylococcus aureus*, *Streptomyces coelicolor*, *Sinorhizobium meliloti*, *Shewanella oneidensis*, *Sulfolobus solfataricus*, *Salmonella typhi*, *Synechocystis* sp., *Thermoplasma acidophilum*, *Thermosynechococcus elongates*, *Thermotoga maritime*, *Treponema pallidum*, *Thermoanaerobacter tengcongensis*, *Tropheryma whipplei*, *Ureaplasma urealyticum*, *Vibrio cholerae*, *Wigglesworthia brevipalpis*, *Xanthomonas campestris*, *Xylella fastidiosa*, *Yersinia pestis*). These genomes were downloaded from the NCBI web page [28]. The resulting sequences are added to the QuartOP dataset and duplicated sequences are eliminated. The datasets are aligned with ClustalW [29], and 100 bootstrap samples are generated using the SEQBOOT program from the PHYLIP package version 3.6a2.1 [30]. The distances are generated using TREE-PUZZLE version 5.1 [31] under the auto-detected substitution model. Neighbor-Joining trees are calculated from these distances using NEIGHBOR from the PHYLIP package version 3.6a2.1 [30]. The resulting trees are parsed with respect to which of the three four-taxa subtrees they contain (see figure 2) using an in-house Java program that utilizes PAL library classes [32]. The resulting bootstrap support vectors are plotted into barycentric coordinates using GNUPLOT version 3.7 [33]. Scripts for data manipulation were written in Perl and used many of the SEALS package subroutines [34].

The *Rhodobacter capsulatus* genome data were obtained from Integrated Genomics [35]. Genome sequence for

Chlorobium tepidum was downloaded from TIGR [36]. The *Rhodopseudomonas palustris* genome was downloaded from JGI [37]. Other genomes for the genome quartets were downloaded from NCBI [28].

The trees depicted in Figures 5 through 8 are neighbor-joining trees calculated using the NEIGHBOR program from PHYLIP version 3.6a2.1 [30]. The distances used in NEIGHBOR were calculated in TREE-PUZZLE version 5.1 [31] with the option to correct for Among Site Rate Variation using a discrete approximation of a Gamma distribution with eight rate categories and estimating the shape parameter. The three indicated support values are bootstrap support values calculated from 100 bootstrap samples analyzed with NEIGHBOR from the distance calculated in TREE-PUZZLE, bootstrap support values calculated from 100 bootstrap samples analyzed with the PROTPARS program from PHYLIP version 3.6a2.1 [30], and posterior probabilities as calculated with MrBayes version 3.0B4 [38] (The analyses were performed independently three times, 200,000 generations each; the lowest posterior probability for the bipartition from the three runs is shown).

For eighteen potential candidates for the horizontal gene transfer between *Halobacterium* sp. and *Synechocystis* sp., or between *Aquifex aeolicus* and *Thermotoga maritima* phylogenetic trees were calculated and inspected manually. The neighbor-joining trees were calculated using the NEIGHBOR program from PHYLIP version 3.6a2.1 [30]. The distances used in NEIGHBOR were calculated in TREE-PUZZLE version 5.1 [31]. The trees were evaluated for potential transfers between Bacteria and Archaea, and between *Thermotoga* and *Aquifex*. 100 bootstrap samples were analyzed to assess the reliability of the branches on the tree. The possibility for the transfer was considered "strong" if the bootstrap support was above 70%, "weak" if the bootstrap support was lower, and "none" if no indication for the transfer could be inferred from the phylogenetic tree. Compositional bias for *Halobacterium* sp. and its closest phylogenetic neighbor was evaluated using a chi-square test at a 5% significance level as implemented in TREE-PUZZLE version 5.1. If both sequences failed the test, this is indicated as "strong" in the table. The results of these analyses are summarized in Table 1. The phylogenetic trees are available as additional data (see 1).

Abbreviations

LBA – Long Branch Attraction

HGT – Horizontal Gene Transfer

ML – Maximum Likelihood

QuartOP – Quartet of Orthologous Proteins

OTU – Operational Taxonomic Unit

Additional material

Additional File 1

Phylogenetic trees for the datasets presented in Table 1. The trees are in the PDF format that can be viewed with Adobe Acrobat Reader http://www.acrobat.com. The eighteen trees are archived into one file trees.zip, and the archive can be expanded using WinZip http://www.winzip.com/ for Windows, StuffIt for Macintosh http://www.stuffit.com/, or unzip utility for Unix. The tree files are named after their ID number listed in Table 1. The names of the OTUs in the trees are the first 10 symbols of the genus name (see Materials and Methods for the list of the full genome names).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-37-S1.zip]

Additional File 2

Bootstrap support maps for non-extended datasets for genome quartets presented in figures 1 and 3. The maps are in the PDF format that can be viewed with Adobe Acrobat Reader http://www.acrobat.com.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-37-S2.pdf]

Acknowledgements

This work was supported through the NASA Astrobiology Institute at Arizona State University, the NASA Exobiology Program, and in part through the NSF Microbial Genetics Program.

References

1. Ribeiro S and Golding GB: **The mosaic nature of the eukaryotic nucleus.** *Mol Biol Evol* 1998, **15:**779-788.
2. Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY and Blankenship RE: **Whole-genome analysis of photosynthetic prokaryotes.** *Science* 2002, **298:**1616-1620.
3. Raymond J, Zhaxybayeva O, Gogarten JP and Blankenship RE: **Evolution of photosynthetic prokaryotes: a maximum-likelihood mapping approach.** *Philos Trans R Soc Lond B Biol Sci* 2003, **358:**223-230.
4. Zhaxybayeva O and Gogarten JP: **Bootstrap, Bayesian probability and maximum likelihood mapping: Exploring new tools for comparative genome analyses.** *BMC Genomics* 2002, **3:**4.
5. Nesbo CL, Boucher Y and Doolittle WF: **Defining the core of non-transferable prokaryotic genes: the euryarchaeal core.** *J Mol Evol* 2001, **53:**340-350.
6. Felsenstein J: **Cases in which parsimony and compatibility methods will be positively misleading.** *Syst. Zool.* 1978, **27:**401-410.
7. Farris JS: **Likelihood and Inconsistency.** *Cladistics* 1999, **15:**199-204.
8. Siddall ME: **Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in the Farris Zone.** *Cladistics* 1998, **14:**209-220.
9. Felsenstein J: **Phylogenies from molecular sequences: inference and reliability.** *Annu Rev Genet* 1988, **22:**521-565.
10. Huelsenbeck JP: **The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining.** *Mol Biol Evol* 1995, **12:**843-849.
11. Huelsenbeck JP and Hillis DM: **Success of Phylogenetic Methods in the Four-Taxon Case.** *Systematic Biology* 1993, **42:**247-264.
12. Philippe H and Douzery E: **The pitfalls of molecular phylogeny based on four species as illustrated by the Cetacea/Artiodactyla relationships.** *Journal of Mammalian Evolution* 1994, **2:**133-152.

13. Adachi J and Hasegawa M: **Instability of quartet analyses of molecular sequence data by the maximum likelihood method: the Cetacea/Artiodactyla relationships.** *Mol Phylogenet Evol* 1996, **6:**72-76.
14. Graybeal A: **Is it better to add taxa or characters to a difficult phylogenetic problem?** *Syst Biol* 1998, **47:**9-17.
15. Hillis DM, Pollock DD, McGuire JA and Zwickl DJ: **Is sparse taxon sampling a problem for phylogenetic inference?** *Syst Biol* 2003, **52:**124-126.
16. Rosenberg MS and Kumar S: **Taxon sampling, bioinformatics, and phylogenomics.** *Syst Biol* 2003, **52:**119-124.
17. Strimmer K and von Haeseler A: **Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment.** *Proc Natl Acad Sci U S A* 1997, **94:**6815-6819.
18. **Eric Weisstein's World of Mathematics** [http://mathworld.wolfram.com/]
19. Gribaldo S and Philippe H: **Ancient Phylogenetic Relationships.** *Theoretical Population Biology* 2002, **61:**391-408.
20. Wainright PO, Hinkle G, Sogin ML and Stickel SK: **Monophyletic origins of the metazoa: an evolutionary link with fungi.** *Science* 1993, **260:**340-342.
21. Jain R, Rivera MC, Moore JE and Lake JA: **Horizontal Gene Transfer Accelerates Genome Innovation and Evolution.** *Mol Biol Evol* 2003.
22. Koonin EV, Makarova KS and Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification.** *Annu Rev Microbiol* 2001, **55:**709-742.
23. Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J, Dahl TA, Welti R, Goo YA, Leithauser B, Keller K, Cruz R, Danson MJ, Hough DW, Maddocks DG, Jablonski PE, Krebs MP, Angevine CM, Dale H, Isenbarger TA, Peck RF, Pohlschroder M, Spudich JL, Jung KW, Alam M, Freitas T, Hou S, Daniels CJ, Dennis PP, Omer AD, Ebhardt H, Lowe TM, Liang P, Riley M, Hood L and DasSarma S: **Genome sequence of Halobacterium species NRC-1.** *Proc Natl Acad Sci U S A* 2000, **97:**12176-12181.
24. Fitch WM: **Homology a personal view on some of the problems.** *Trends Genet* 2000, **16:**227-231.
25. Douady CJ, Delsuc F, Boucher Y, Doolittle WF and Douzery EJ: **Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability.** *Mol Biol Evol* 2003, **20:**248-254.
26. Alfaro ME, Zoller S and Lutzoni F: **Bayes or bootstrap? A simulation study comparing the performance of bayesian markov chain monte carlo sampling and bootstrapping in assessing phylogenetic confidence.** *Mol Biol Evol* 2003, **20:**255-266.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
28. **National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov/]
29. Thompson JD, Higgins DG and Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
30. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6.** *Distributed by the author. Department of Genetics, University of Washington, Seattle* 1993.
31. Schmidt HA, Strimmer K, Vingron M and von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502-504.
32. Drummond A and Strimmer K: **PAL: an object-oriented programming library for molecular evolution and phylogenetics.** *Bioinformatics* 2001, **17:**662-663.
33. **GNUPLOT Central** [http://www.gnuplot.info]
34. Walker DR and Koonin EV: **SEALS: a system for easy analysis of lots of sequences.** *ISMB* 1997, **5:**333-339.
35. **Integrated Genomics** [http://www.integratedgenomics.com/]
36. **The Institute for Genomic Research** [http://www.tigr.org]
37. **DOE Joint Genome Institute** [http://www.jgi.doe.gov/JGI_microbial/html/index.html]
38. Huelsenbeck JP and Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.