WILEY-BLACKWELL

Cladistics (2013) 1–11

Cladistics

10.1111/cla.12020

Complete genome sequences provide a case study for the evaluation of gene-tree thinking

Rebecca B. Dikow[a,b,*,†] and William Leo Smith[b]

[a]*Committee on Evolutionary Biology, University of Chicago, 1025 East 57th Street, Chicago, IL, 60637, USA;* [b]*Division of Fishes, Field Museum of Natural History, 1400 South Lake Shore Drive, Chicago, IL, 60605, USA*

Accepted 21 January 2013

Abstract

Complete genome sequences from a genus of Gammaproteobacteria, *Shewanella*, are used to generate a genome-wide exploration of the gene-tree species-tree dichotomy. A number of datasets were constructed and analyses were attempted. Single genes were chosen from 243 regions of collinear gene homology (128 of these 243 chosen genes are from the core *Shewanella* genome and 162 of 243 have the complete taxon sampling) from a previous study (Dikow, 2011) and subjected to phylogenetic analysis both individually and concatenated. In addition, three of the 243 sets of collinear genes from the core *Shewanella* genome were also chosen (comprising 15, 17, and 23 genes each) to be analysed in detail, this time to maximize the expectation of gene concordance. Analysis of these 55 genes in maximum parsimony (MP) and maximum likelihood (ML) produced 164 unique topologies (out of 166 resulting topologies). No genes from within collinear regions were congruent with one another, and none of these 164 topologies matches the result from concatenation. This result is particularly striking given that we chose collinear sets of genes. Analyses in MP and ML of 243 genes distributed across the genome produced 567 unique topologies (out of 571 resulting topologies for those 162 genes with complete taxon sampling). These results are discussed in light of recent works focused on incongruence. The gene as a phylogenetic unit is also discussed. It is our conclusion that molecular systematics has been reliant on the gene as a unit without a critical eye on the distinction between gene homology and character homology.

Introduction

Discrepancies between gene trees and species trees have been addressed by recent workers either by using gene tree concordance methods (e.g. Ané et al., 2007; Liu, 2008) or by discarding the species tree altogether to investigate "gene history" instead. The term "species tree" was even redefined by Edwards (2009) to account for multiple gene histories and to allow each "gene history" to count as an equal piece of evidence. This is distinguished from the original species tree, produced with a total evidence dataset, in which each nucleotide has the same weight of evidence (or under some weighting scheme specified by the investigator) and the nucleotides, rather than the genes, are the characters under consideration. These debates come decades after many discussions in the literature about combining data (e.g. Kluge, 1989; Kluge and Wolf, 1993; Nixon and Carpenter, 1996). In this paper we focus on the gene-tree species-tree dichotomy in terms of phylogenomics. Complete genome sequences give us the opportunity to assess patterns of incongruence over a very large scale.

Those studies that have previously considered large numbers of genes and large numbers of gene trees have generally considered very few species that are either very closely related (often strains of the same species) or very distantly related (e.g. Cummings et al., 1995; Rokas et al., 2003; Cranston et al., 2009). These

*Corresponding author:
E-mail address: dikowR@si.edu
[†]Present address Center for Conservation and Evolutionary Genetics National Zoological Park and Division of Mammals National Museum of Natural History Smithsonian Institution PO Box 37012, MRC 108 Washington DC 20013-7012 USA

studies also generally contain large numbers of genes that are not represented by the complete taxon sampling, which is not surprising given the incompleteness of gene annotations and difficulty in assigning primary homology, but which minimizes the number of possible topologies. Nevertheless, these studies have generally found a level of incongruence among gene trees that the workers found to be higher than expected. Gene tree concordance methods (e.g. BEST: Liu, 2008; BUCKy: Ané et al., 2007) have been developed in response to these results. These methods produce a single species tree signal after an initial analysis of each gene (in which the nucleotides are the characters). Each gene tree is subsequently taken as a character, rather than each nucleotide. In gene tree concordance methods, each gene or sequence fragment is generally given the same weight, regardless of gene length, resulting in nucleotides with differential weights (i.e. the nucleotides of a very short gene will be weighted higher than the nucleotides in a long gene). This problem was discussed in reference to consensus methods long before concordance methods came to be (Barrett et al., 1991).

We have attempted to frame an "ultimate" gene tree analysis based on whole genomes given the genome-level hypotheses of homology in Dikow (2011). We hypothesized that sets of collinear (adjacent) genes would be congruent to a large degree, providing pockets of local signal across the genome. Our goal with the analyses presented here was to provide the conditions favouring such a pattern. Departure from this pattern, then, would allow us to reject our initial assumption that local signal exists on a genome scale. In order to generate this set of analyses, the following steps were taken and are discussed in more detail below. First, the gene homologies have been detected automatically from unannotated genome sequences, as opposed to detection with annotation or amplification via PCR. Second, collinear genes present in the core *Shewanella* genome have been chosen to minimize the possibility that these genes have been horizontally transferred (particularly if they are immediately adjacent). Third, these three sets of collinear genes were chosen to be distant from each other across the genome to diminish the possibility of shared signal across these three sets of genes. To provide a truly genome-wide test, a second set of analyses was attempted in which a single gene from each of 243 sets of collinear homologous genes was chosen for separate analysis to assess the level of congruence across the entire genome and to compare to the results from the collinear sets of genes. A bacterial taxon was chosen because it allows a genome-wide analysis for a large number of taxa (many genomes sequenced; small enough genomes).

Shewanella is a genus of marine and freshwater Gram-negative Gammaproteobacteria within the monogeneric family Shewanellaceae Ivanova et al., 2004. Species of *Shewanella* have been described from diverse habitats, from deep cold-water marine environments to shallow Antarctic Ocean habitats to hydrothermal vents and freshwater lakes (Dikow, 2011). The genus has been of great applied interest due to the ability of its species to convert heavy metals and toxic substances (e.g. Fe, S, U) into less toxic products, making these species of interest for environmental clean-up (Lovley and Phillips, 1988; Perry et al., 1993; Bowman et al., 1997). Genome annotations suggest that species possess ca. 5000 genes and have genomes of ca. 5 Mbp. In Dikow (2011), primary homology was determined starting with unannotated genomes using the program Mauve (Darling et al., 2010). Mauve addresses the issue of gene or other DNA fragment rearrangement by finding locally collinear blocks (LCBs), or contiguous segments of sequence within which there has not been rearrangement, but within a longer sequence that may have been subject to rearrangement events. Mauve finds anchor points of similarity and then extends these matches outward. A single LCB becomes two when a sequence segment is found somewhere else in the genome for one of the taxa, meaning that a rearrangement has occurred. Mauve does not allow one sequence fragment to be homologous to more than one fragment in another species. 243 LCBs were found across 22 taxa (including three outgroups) that ranged in size from ca. 1000 to 120 000 bp. For *Shewanella oneidensis*, each LCB contained between one and 79 genes, plus intergenic sequence when it occurred, and in some cases partial gene sequences.

When analysed separately, it was shown that 242 of the 243 LCBs produced unique topologies, none of which had the same topology as the concatenated dataset (3.3 Mbp). Random samplings of the concatenated dataset with as few as 20 000 bp (shorter than many of the LCBs themselves) always produced the same topology as the concatenated dataset. This result was unexpected, and it was concluded that local signal was the reason why none of the LCB trees produced the same topology, but random samplings (which escape local signal) did. Based on this pattern, one of dramatic incongruence, one could also produce hypotheses that perhaps one or more LCBs might have been horizontally transferred, but since 242 of 243 produced different topologies, it would be impossible to know which, and also extremely unlikely that all LCBs were horizontally transferred. Rather, because Mauve finds fragments positionally homologous among all taxa, we limit the possibility of horizontally transferred fragments being included in the first place.

For the present study, beyond further exploring the pattern shown in Dikow (2011), it was also of interest to the authors to investigate the gene-tree species-tree

dichotomy in a genomic context. A gene tree may have a different phylogenetic signal from a species tree for a number of reasons, some of which may be part of the history of the organisms (evolutionary processes), others due to incorrect homology assessment (Brower et al., 1996). The possibility of horizontal forces (hybridization, gene transfer) being dominant in the evolutionary history of a group of organisms is often the reason given as to why a total evidence approach will not represent taxonomic history accurately, particularly for prokaryotes (e.g. Bapteste and Boucher, 2008). Incomplete lineage sorting is the reason many systematists working on eukaryotes choose not to concatenate genes (e.g. Degnan and Rosenberg, 2009). The same argument is never made for morphological characters; in cases of incongruence of morphological characters, homoplasy is always attributed to incorrect homology assessment (Nixon and Carpenter, 2012).

Whatever the source of incongruence, how prevalent do we expect these patterns to be across an entire genome? Incomplete lineage sorting is likely to be much more common at the species boundary and below (Maddison and Knowles, 2006). Often, in a study where gene trees are produced, a small number of phylogenetic topologies and histories are considered; maximally the number of genes included. What happens when more genes are included; does the number of phylogenetic signals increase by one with every gene, or are there just a few signals present? These questions have not yet been explored across a large taxon sampling. As mentioned above, those studies that do analyse large numbers of genes generally consider very few taxa (eight or fewer), which limits the number of possible topologies and our ability to quantify the level of incongruence.

For the collinear gene analyses, the three LCBs out of the 243 found across all 22 genomes sampled were chosen based on their relative completeness (few gaps) for all taxa, and because they are among the longer LCBs with at least 15 genes as examples for which gene-by-gene analyses could be completed. These LCBs all contain genes from the "core" *Shewanella* genome (Konstantinidis et al., 2009). Phylogenetic analyses were conducted on the genes individually, concatenated, and the original LCBs, and all compared with each other and with the whole genome tree in Dikow (2011).

Materials and methods

Three of the original 243 LCBs from Dikow (2011) were chosen as exemplars: LCB 27: 13 022 bp, 17 genes; LCB 156: 55 810 bp, 23 genes; LCB 175: 23 108 bp, 15 genes. Throughout this text the LCBs are referred to by the number given by Mauve representing their relative location in *Aeromonas hydrophila*, specified as the root. Alignments for each of the three LCBs were cut, and genes were excised based on the gene boundaries present in the genome annotation of *S. oneidensis*. The genes included in the collinear genes set of analyses are listed in Table 1. All non-hypothetical genes were found in the core *Shewanella* genome (Konstantinidis et al., 2009). Intergenic DNA sequence was discarded.

The following phylogenetic analyses were performed on a 2.8 GHz Quad-Core MacPro with 20GB RAM: (i) each gene separately, (ii) genes for each LCB concatenated, and (iii) all genes (55) from the three LCBs concatenated. These analyses are compared with the whole genome tree based on all concatenated LCBs (3.3 Mbp and ca. 1500 genes plus intergenic DNA sequence; Dikow, 2011 "genome tree"). This whole genome tree topology is the same from TNT and RaxML. For the collinear genes datasets, all analyses were performed under maximum parsimony (MP) and maximum likelihood (ML). Bayesian inference (B) was also used to analyse individual genes, and the Bayesian gene trees were subjected to subsequent BUCKy analyses of gene concordance. Maximum parsimony was conducted in TNT (Goloboff et al., 2008): 1000 builds with subtree pruning and regrafting (SPR) and tree bisection and reconnection (TBR) followed by 1500 replicates of ratchet and tree fusing, default gaps setting. Maximum likelihood was conducted with Garli 2.0 (Zwickl, 2006; GTR + gamma model of nucleotide substitution, 100 search replicates). Bayesian inference was conducted with MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001; 10 000 000 generations, 25% burn-in, GTR + gamma model of nucleotide substitution). For the second set of analyses considering a single gene from each LCB across the genome, genes were isolated from each LCB alignment in the same manner as above. The first complete gene was chosen from each LCB based on *S. oneidensis*. Each gene was analysed separately with MP and ML as above. We focus in the Results section on those genes (162/243) that had the complete taxon sampling. These 243 genes were also concatenated and analysed together in Garli, as above. For LCB 47, the gene included here has been truncated as its alignment was exceptionally long and included a gap of more than 100 000 bp. The gene has been included from bp 57 to 854. For all other LCBs, the original gene length has been preserved.

Topologies were assessed for uniqueness and the presence of clades of interest was counted. In MP, all most parsimonious trees (MPTs) were considered, so the presence of polytomies was not an issue in comparing topologies. For ML, in order to count unique topologies we resolved polytomies based on a randomly chosen gene alignment in POY (Varón et al., 2010). In response to a reviewer's comment, we

Table 1
Shewanellaceae genes in LCBs of interest: genes in each LCB with number of nucleotide base pairs in alignment (trimmed based on homology assessment in the program Mauve and annotation of *S. oneidensis*)

LCB gene		Number of bp	Gene	Full gene name
27	1	372	*rplN*	50S ribosomal protein L14
	2	345	*rplX*	50S ribosomal protein L24
	3	541	*rplE*	50S ribosomal protein L5
	4	307	*rpsN*	30S ribosomal protein S14
	5	385	*rpsH*	30S ribosomal protein S8
	6	537	*rplF*	50S ribosomal protein L6
	7	357	*rplR*	50S ribosomal protein L18
	8	515	*rpsE*	30S ribosomal protein S5
	9	182	*rpmD*	50S ribosomal protein L30
	10	442	*rplO*	50S ribosomal protein L15
	11	1363	*secY*	preprotein translocase subunit
	12	113	*rpmJ*	50S ribosomal protein L36
	13	457	*rpsM*	30S ribosomal protein S13
	14	395	*rpsK*	30S ribosomal protein S11
	15	628	*rpsD*	30S ribosomal protein S4
	16	994	*rpoA*	DNA-directed RNA polymerase subunit alpha
	17	402	*rplQ*	50S ribosomal protein L17
156	1	1366	*flrC*	flagellar regulatory protein C
	2	1617	*flrB*	flagellar regulatory protein B
	3	10 002	*flrA*	flagellar regulatory protein A
	4	4850	*fliS*	flagellar protein
	5	2681		hypothetical protein
	6	1963	*fliD*	flagellar hook-associated protein
	7	576	*flaG*	flagellin
	8	2285		flagellin
	9	4080		flagellin
	10	1556	*flgL*	flagellar hook-associated protein
	11	612		flagellar protein
	12	1245		flagellar hook-associated protein
	13	1572	*flgJ*	flagellar rod assembly protein/ muramidase
	14	1128	*flgI*	flagellar basal body P-ring protein
	15	851	*flgH*	flagellar basal body L-ring protein
	16	789	*flgG*	flagellar basal body rod protein
	17	748	*flgF*	flagellar basal body rod protein
	18	1726	*flgE*	flagellar hook protein
	19	834	*flgD*	flagellar basal body rod modification protein
	20	546	*flgC*	flagellar basal body rod protein
	21	444	*flgB*	flagellar basal body rod protein
	22	842	*cheR-2*	chemotaxis protein methyltransferase
	23	924	*cheV-3*	chemotaxis protein
175	1	3391	*mrcA*	penicillin-binding protein 1A
	2	1284	*pilM*	type IV pilus biogenesis protein
	3	826	*pilN*	type IV pilus biogenesis protein
	4	1321	*pilO*	type IV pilus biogenesis protein
	5	844	*pilP*	type IV pilus biogenesis protein
	6	2487	*pilQ*	type IV pilus biogenesis protein
	7	552	*aroK*	shikimate kinase I
	8	1363	*aroB*	3-dehydroquinate synthase
	9	2857		DamX domain-containing protein
	10	877	*dam*	DNA adenine methylase
	11	200		hypothetical protein
	12	870		hypothetical protein
	13	679	*rpe*	ribulose-phosphate 3-epimerase
	14	734	*gph*	phosphoglycolate phosphatase
	15	1880	*trpS*	tryptophanyl-tRNA synthetase

Table 2
Shewanellaceae taxon table with GenBank accession numbers

Taxon	GenBank accession number
Aeromonas hydrophila ATCC 7966	NC_008570
Alteromonas macleodii deep ecotype	NC_011138
Colwellia psychrerythraea 34H	NC_003910
Shewanella amazonensis SB2B	NC_008700
Shewanella baltica OS223	NC_011663
Shewanella baltica OS155	NC_009052
Shewanella baltica OS185	NC_009665
Shewanella baltica OS195	NC_009997
Shewanella denitrificans OS217	NC_007954
Shewanella frigidimarina NCIMB 400	NC_008345
Shewanella halifaxensis HAW-EB4	NC_010334
Shewanella loihica PV-4	NC_009092
Shewanella oneidensis MR-1	NC_004347
Shewanella pealeana ATCC 700345	NC_009901
Shewanella piezotolerans WP3	NC_011566
Shewanella putrefaciens CN-32	NC_009438
Shewanella sediminis HAW-EB3	NC_009831
Shewanella sp. ANA-3	NC_008577
Shewanella sp. MR-7	NC_008322
Shewanella sp. MR-4	NC_008321
Shewanella sp. W3-18-1	NC_008750
Shewanella woodyi ATCC 51908	NC_010506

investigated whether outgroup choice was responsible for our large number of unique trees by analysing the 23 genes in LCB 156 without the outgroup taxa in Garli, as above.

Results

Collinear sets of genes

Species and strain information and GenBank accession numbers are shown in Table 2. Whole genome topology (Dikow, 2011), BUCKy concordance topologies, and ML concatenated topologies for each LCB from the collinear gene analyses are shown in Fig. 1. Genes represented in the collinear gene analyses and the lengths of each of these genes are shown in Table 1. The length of each of the datasets concatenated is as follows, LCB 27 (17 genes): 8335 bp; LCB 156 (23 genes): 43 237 bp; LCB 175 (15 genes): 18 598 bp. For all resulting trees, branches have been coloured to represent clades found in the tree resulting from the concatenated genome alignment in Dikow (2011) to allow simple assessment of congruence and consistency of the expected groupings (see Fig. 1). Maximum likelihood gene trees for each gene in the collinear gene analyses are displayed in Fig. 2. In both figures some trees are shown without names, just with coloured branches, in the interest of space and because the arrangement of clades is of greatest interest. All trees have been deposited in Dryad. For ML and MP, for each of the three LCBs considered, each gene within its respective LCB (55 total genes) produced one or more unique topologies (when polytomies are resolved, LCB 156 gene 12 and LCB 175 gene 12 are identical), none of which (including the strict consensus trees for MP) matched the trees based on the concatenated genes (MP or ML; Fig. 1d–g shows ML trees), the tree for the original LCB (including intergenic sequence; MP or ML), or the genome tree (MP and ML same; Fig. 1a).

There was no resolution when a strict consensus was calculated among gene trees, even within each LCB for 27 and 156. For LCB 175, *S. putrefaciens* + *S.* sp. W3-18-1 and monophyletic *S. baltica* were present in all genes in ML (Fig. 2). Table 3 indicates the presence of clades found in the genome tree (Dikow, 2011) in analyses of these 55 genes. For the BUCKy analyses, concordance topologies are shown in Fig. 1h–j). For LCB 27, there were 135 172 different sampled topologies. For LCB 156, there were 59 341 different sampled topologies. For LCB 175, there were 39 470 different sampled topologies. When all 55 genes from these three LCBs are considered in a single BUCKy analysis, the concordance topology is the same as that for LCB 175 and there were 230 231 different sampled topologies. This is also the same topology of the ML analysis in which all 55 genes from the three LCBs are concatenated (Fig. 1d). When the 23 genes from LCB 156 were analysed in Garli with no outgroups, these analyses resulted in 23 unique topologies.

Single genes across all LCBs

Eighty-one of the 243 genes included had one or more taxa missing (all gaps in larger LCB alignment for this gene). For 45 of these 81, these genes are listed as "hypothetical" in the *S. oneidensis* gene annotation. For an additional 23, these are annotated but not in the core genome. These two categories together account for 68 of the 81 genes where not all taxa are represented.

The results from the phylogenetic analyses are as follows: for the 162 genes for which all taxa are represented, each gene produced a unique topology in ML. In MP, only genes from LCB 235 and 236 produced the same topology. When we compared the MP and ML topologies, 567/571 were unique. When all 243 genes are concatenated, the alignment has 312 901 bp. The topologies generated when all 243 genes are concatenated are shown in Fig. 1b (ML) and Fig. 1c (MP). For MP, tree search resulted in a single most parsimonious tree of length 861 262 steps. None of the single-gene trees matches either of these trees from all 243 genes concatenated. The whole genome topology (Fig. 1a) was never found by any gene or concatenated dataset with either optimality criterion. The presence

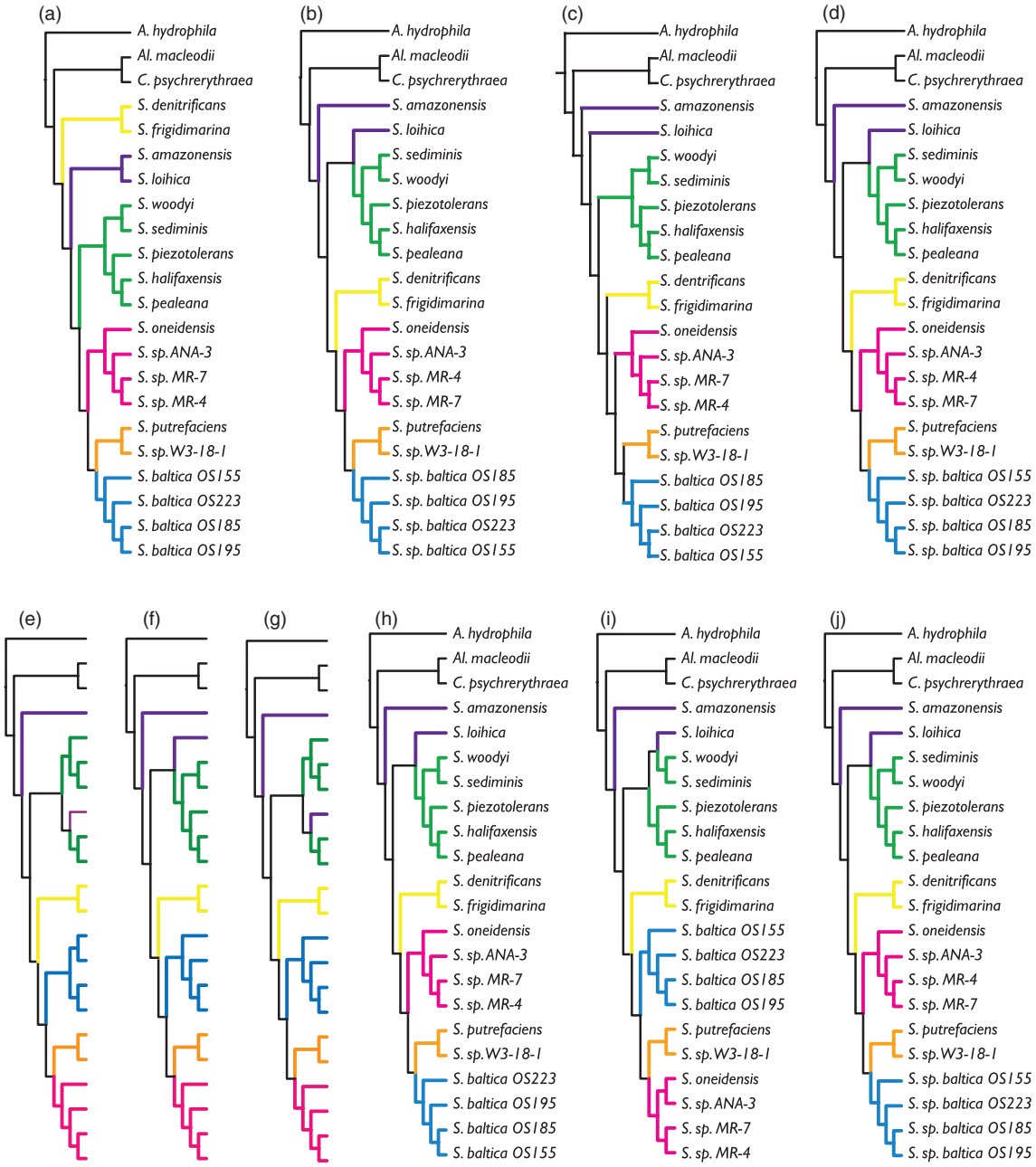

Fig. 1. Trees from concatenated analyses. (a) Whole genome tree as calculated in maximum parsimony (MP) and maximum likelihood (ML) (Dikow, 2011). (b) ML tree for 243 genes concatenated. (c) MP tree for 243 genes concatenated. (d) ML tree resulting from the concatenation of all genes from all three LCBs (55 genes). (e) ML tree resulting from the concatenation of genes from LCB 27. (f) ML tree resulting from the concatenation of genes from LCB 156 (23 genes). (g) ML tree resulting from the concatenation of genes from LCB 175 (15 genes). (h) BUCKy primary concordance tree for LCB 27. (i) BUCKy primary concordance tree for LCB 156. (j) BUCKy primary concordance tree for LCB 175 or for all 55 genes concatenated. Taxa are coloured for easy comparison. Outgroup taxa remain black. *S. amazonensis*, *S. loihica*: purple; *S. baltica* strains: blue; *S. oneidensis*, *S.* sp. ANA-3, *S.* sp. MR-7, *S.* sp. MR-4: pink; *S. woodyi*, *S. sediminis*, *S. piezotolerans*, *S. halifaxensis*, *S. pealeana*: green; *S. putrefaciens*, *S.* sp. W3-18-1: orange; *S. denitrificans*, *S. frigidimarina*: yellow.

of the clades found in the whole genome tree (those that correspond to the branch colours) are scored for all ML gene trees and the numbers are shown in Table 3 (55 genes from collinear sets of genes) and Table 4 (162 genes from across the genome). None of the gene tree topologies from the 162 genes from across the genome matches any gene tree topologies from the 55 genes from collinear sets of genes. All phylogenetic trees for all analyses presented have been deposited on Dryad.

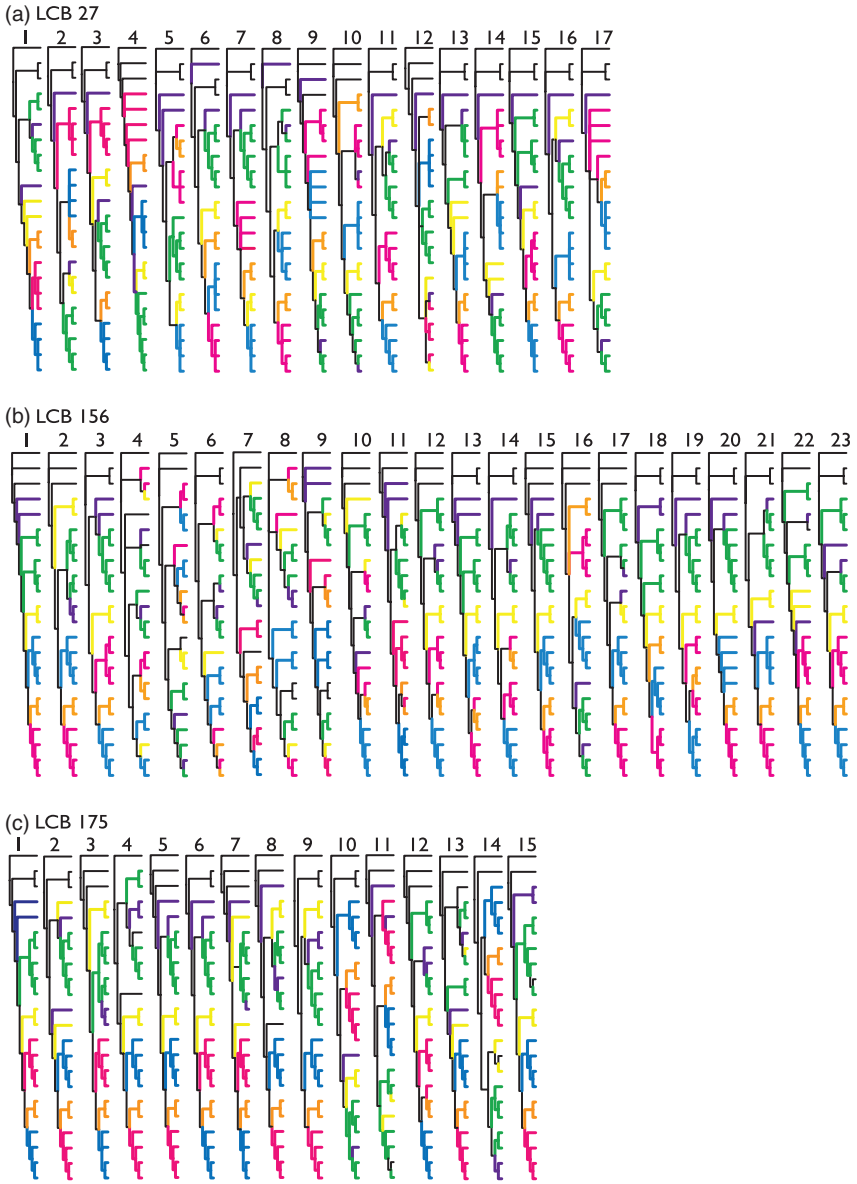

Fig. 2. Maximum likelihood (ML) gene trees. (a) Gene trees from LCB 27, genes 1–17 left to right. (b) Gene trees from LCB 156, genes 1–23 left to right. (c) Gene trees from LCB 175, genes 1–15 left to right. Names have been removed to save space and allow comparison of all gene trees at once. Taxa are coloured as in Fig. 1 for easy comparison. Outgroup taxa remain black. *S. amazonensis*, *S. loihica*: purple; *S. baltica* strains: blue; *S. oneidensis*, *S.* sp. ANA-3, *S.* sp. MR-7, *S.* sp. MR-4: pink; *S. woodyi*, *S. sediminis*, *S. piezotolerans*, *S. halifaxensis*, *S. pealeana*: green; *S. putrefaciens*, *S.* sp. W3-18-1: orange; *S. denitrificans*, *S. frigidimarina*: yellow.

Discussion

Our results expand and emphasize the pattern reported in Dikow (2011), on the gene as opposed to the LCB scale, where no single LCB produced the same tree as the concatenated dataset and 242 out of 243 LCBs produced unique topologies. Here we present a total of 731 gene trees representing 213 genes, and 727 of these 731 are unique. It was surprising to us that the gene level showed the same pattern as the LCB level in terms of unique topologies, and that the whole genome tree was never found, especially because when the genome alignment is randomly sampled, as low as 20 000 bp is enough to recover the genome topology (Dikow, 2011). One cannot argue convincingly that our pattern is simply the result of horizontal gene transfer as there are so many different signals; this would require that all genes have been horizontally transferred independently. It was also interesting that none of the gene trees matched the original LCB trees in Dikow (2011).

Because alignments were trimmed based on the entire LCB alignment, the gene alignments may be longer than expected in some cases when compared

Table 3
Tracking the presence of genome tree clades in individual gene tree analyses (55 genes from collinear sets of genes)

Analysis	Clade of interest	Clade colour	Topology
Maximum likelihood	Monophyletic *Shewanella*	black	39/55
	S. oneidensis + *S.* sp. ANA-3 + *S.* sp. MR-4 + *S.* sp. MR-7	pink	34/55
	S. putrefaciens + *S.* sp. W3-18-1	orange	51/55
	Monophyletic *S. baltica*	blue	46/55
	S. denitrificans + *S. frigidimarina*	yellow	37/55
	S. piezotolerans + *S. pealeana* + *S. woodyi* + *S. sediminis* + *S. halifaxensis*	green	17/55
	S. amazonensis + *S. loihica*	purple	8/55
Maximum parsimony	Monophyletic *Shewanella*	black	82/111
	S. oneidensis + *S.* sp. ANA-3 + *S.* sp. MR-4 + *S.* sp. MR-7	pink	73/111
	S. putrefaciens + *S.* sp. W3-18-1	orange	88/111
	Monophyletic *S. baltica*	blue	96/111
	S. denitrificans + *S. frigidimarina*	yellow	49/111
	S. piezotolerans + *S. pealeana* + *S. woodyi* + *S. sediminis* + *S. halifaxensis*	green	45/111
	S. amazonensis + *S. loihica*	purple	17/111

Table 4
Tracking the presence of genome tree clades in individual gene tree analyses (162 genes across genome)

Analysis	Clade of interest	Clade colour	Topology
Maximum likelihood	Monophyletic *Shewanella*	black	100/162
	S. oneidensis + *S.* sp. ANA-3 + *S.* sp. MR-4 + *S.* sp. MR-7	pink	129/162
	S. putrefaciens + *S.* sp. W3-18-1	orange	154/162
	Monophyletic *S. baltica*	blue	146/162
	S. denitrificans + *S. frigidimarina*	yellow	98/162
	S. piezotolerans + *S. pealeana* + *S. woodyi* + *S. sediminis* + *S. halifaxensis*	green	61/162
	S. amazonensis + *S. loihica*	purple	18/162
Maximum parsimony	Monophyletic *Shewanella*	black	104/409
	S. oneidensis + *S.* sp. ANA-3 + *S.* sp. MR-4 + *S.* sp. MR-7	pink	303/409
	S. putrefaciens + *S.* sp. W3-18-1	orange	347/409
	Monophyletic *S. baltica*	blue	343/409
	S. denitrificans + *S. frigidimarina*	yellow	185/409
	S. piezotolerans + *S. pealeana* + *S. woodyi* + *S. sediminis* + *S. halifaxensis*	green	216/409
	S. amazonensis + *S. loihica*	purple	23/409

with the unaligned gene length. This was done instead of realigning individual genes. The primary homology assessment provided by Mauve begins with unannotated genomes, so without preconception of gene function or homology. This allows us to have a defensible hypothesis of homology and to include more genes in a phylogenetic analysis, but it also means that there may be sequence from outside the gene of interest for taxa other than *S. oneidensis* in the gene alignments. Annotations are often so incomplete that for 22 taxa, a very large number of genes could not be found present and annotated in each taxon. In addition, inconsistent or duplicate gene names can make primary homology prediction a guessing game when referring to annotations. We did expect that those genes with all taxa represented would be those for which the genes were in the core genome; this was true for 77.8% of cases. Because those genes for which some taxa were missing were mostly those representing hypothetical genes, it made us even more confident in the performance of Mauve in finding primary homologies.

We distinguish our study from previous works looking at large numbers of genes and gene tree incongruence, in a number of ways. First, the taxon sampling represents a genus-level study with a significant diversity

of the species represented (16/51 species, plus three strains of one species) and three outgroup taxa. The previous works mentioned here (Cummings et al., 1995; Rokas et al., 2003; Cranston et al., 2009) were demonstrative in different ways—one for a very recently diverged group (rice, Cranston et al., 2009) and one for a deeply diverging group (yeast, Rokas et al., 2003)—but all contain relatively few internal branches. Cranston et al. had 307 genes for which all taxa were represented, but only 105 possible rooted topologies for their six-taxon dataset. Rokas et al. (2003: 798) mention that for their eight-taxon dataset "Single-gene phylogenies reveal extensive incongruence," and "Analyses of the 106 genes resulted in more than 20 alternative ML or MP trees." When Rokas randomly selected nucleotides from 127 026 bp of their initial alignment of all 106 genes, similarly to that done in Dikow (2011), they found that 3000 bp were enough to recover the tree built from the entire concatenated dataset.

By including a much larger number of internal branches (19) and 1.31×10^{25} possible rooted topologies, we feel that our study accurately represents the realistic expectation for how many gene trees exist across genomes. For the three LCBs for which we analysed collinear genes, none of the genes is represented by fewer than the complete taxon sampling. Given that these genes are adjacent within their LCBs, our results are even more striking. Functionally, these three LCBs have many gene products with similar or coordinated function: for LCB 27, ribosomal proteins, for LCB 156, flagellar proteins, and for LCB 175, pilus proteins (Table 1). While within each LCB gene function is correlated or similar, among the three LCBs the functions chosen are quite variable and might be expected to follow different evolutionary trajectories, with the ribosomal genes thought to be conserved and the flagellar and pilus genes more labile. It is interesting, then, that all three show similar patterns of incongruence. For the BUCKy analyses, it is interesting that there are so many sampled topologies for LCB 27, as it is the shortest in terms of bp both originally (13 022 bp) and when genes are removed and spliced together (8335 bp). The concordance topologies generated by BUCKy (Fig. 1h–j) are largely congruent with each other, and the ML trees from concatenated datasets are largely congruent with the BUCKy trees (Fig. 1c–j), indicating that in this case BUCKy approximates the results from concatenation.

While a reviewer has pointed out that we might also integrate statistical support for particular clades into our assessment, we chose to focus uniqueness of topologies here because the question for us was quantifying the number of gene trees we could find by sampling across the genome in different ways (collinear genes and distant genes). We already knew, based on Dikow (2011), that when LCBs are analysed separately, we get many clades occurring frequently that are also in our concatenated solution (Fig. 2 in Dikow, 2011). By analysing the 23 genes in LCB 156 without outgroups and also obtaining 23 unique topologies, we have shown that the outgroups we chose are not causing the pattern of incongruence.

The results highlighted above have led us to reconsider the utility of gene trees. According to Hennig (1966), the semaphoront, or character bearer, is the fundamental unit of phylogenetic analysis. The characters themselves cannot be isolated from the organism and therefore cannot be the terminals in a phylogenetic analysis with a history separate from the organism itself. This idea is uncontroversial for morphological characters (one's hand does not have a history separate from that of the rest of the organism and we never find a living hand without the rest of the organism), but one might be persuaded to consider genes as entirely different because of the self-replicating nature of DNA. While DNA pieces certainly replicate, they do so within the confines of the rest of the genome, and while genes have functions, they often operate in the context of additional pieces of DNA: promoters, enhancers, regulatory genes, etc. The ability to assign a function or even just an open reading frame to a given piece of DNA sequence has led many workers to assume that the given piece of DNA sequence thus has a "history" separate from the history of the rest of its genomic complement. This leads us to separate the concepts of gene homology and character homology (here nucleotide). Historically, the way we detect gene homologies is fundamentally different from that of nucleotides. Nucleotides undergo the congruence test while there is generally only a test of primary homology for genes (amplification by conserved primers using PCR).

For a genus-wide study such as that presented here, we would assume *a priori* the vast majority, if not all, genes to be homologous, that is, inherited from a common ancestor. Because either none, or at most one, of the gene trees here reflects that pattern of common ancestry (by exhibiting complete incongruence), either none of the genes is homologous, or our reliance on the gene as a partition and homolog must be questioned. Genes may be appropriate functional units, but the only homologies we are able to test here are those of the individual nucleotides.

Conclusions

Incongruence is a pattern that reflects a number of evolutionary processes (including incomplete lineage sorting, hybridization, and horizontal gene transfer) as

well as incorrect homology assessment. For molecular data, because the nucleotide sample size for a single gene is small (particularly small when only the parsimony-informative characters are considered), we often produce conflicting gene trees. Rather than reflecting a history, these analyses simply reflect the problem of small sample sizes. Over a large scale, we expect the effects of individual gene partitions to balance out. This is reflected in the results of random selection of alignment positions from throughout the genome: relatively few positions are necessary to produce the tree generated from the entire concatenated dataset (Dikow, 2011). While we can only conclude based on the taxa analysed here, it is our opinion that large-scale gene tree incongruence will not be a pattern unique to bacteria, as additional studies of gene trees across entire genomes for new taxon samplings are completed. When this pattern is found repeatedly, it will require a dramatic rethinking of the gene as a data partition in phylogenetic systematics.

Genome-scale datasets are becoming the norm in molecular systematics, while analysing complete genomes will continue to present computational challenges. How we effectively detect homologies across genomes and produce robust phylogenetic datasets that are subsets of complete genomes is not as simple as pulling out increasingly large numbers of annotated genes. Randomly sampling putatively homologous nucleotides across a very large, perhaps genome-wide, alignment (as was done in Dikow, 2011) is one possible strategy that can allow us the benefits of large datasets without such a large computational burden. Many workers are looking for ways to gather genome-wide data (e.g. RAD-tags, Baird et al., 2008; ultraconserved elements, Faircloth et al., 2012) without generating complete genomes. It is so far unclear whether computational resources (particularly in the realm of genome assembly) will improve enough in the near future to make gathering full eukaryotic genomes for large taxon samplings efficient. The potential to discover relationships among character sets (morphological and molecular) and the evolutionary history among parts of the genome (gene rearrangement, parts of genes not normally captured with DNA sequences) will require complete genomes rather than bits and pieces gathered from here and there. The goal of comparing complete genomes carries with it much more than the desire for large datasets.

Acknowledgements

R.B.D. and W.L.S. thank Torsten Dikow, Shannon Hackett, members of the Pritzker Molecular Systematics Lab for useful discussion, and two reviewers for their constructive comments. R.B.D. was supported by The Field Museum Women's Board and the Emerging Pathogen Project (to The Field Museum of Natural History and The University of Chicago Institute for Genomics and Systems Biology).

References

Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24, 412–426.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., et al., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3, e3376.

Bapteste, E., Boucher, Y., 2008. Lateral gene transfer challenges the principles of microbial systematics. Trends Microbiol. 16, 200–207.

Barrett, M.M., Donoghue, M.J., Sober, E., 1991. Against consensus. Syst. Zool. 40, 486–493.

Bowman, J.P., McCammon, S.A., Nichols, D.S., Skerratt, J.H., Rea, S.M., Nichols, P.D., McMeekin, T.A., 1997. *Shewanella gelidimarina* sp. nov. and *Shewanella frigidimarina* sp. nov., novel Antarctic species with the ability to produce eicosapentaenoic acid (20:5 omega 3) and grow anaerobically by dissimilatory Fe (III) reduction. Int. J. Syst. Bacteriol. 47, 1040–1047.

Brower, A.V.Z., DeSalle, R., Vogler, A., 1996. Gene trees, species trees, and systematics. Annu. Rev. Ecol. Syst. 27, 423–450.

Cranston, K.A., Hurwitz, B., Ware, D., Stein, L., Wing, R.A., 2009. Species trees from highly incongruent gene trees in rice. Syst. Biol. 58, 489–500.

Cummings, M.P., Otto, S.P., Wakeley, J., 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12, 814–822.

Darling, A.E., Mau, B., Perna, N.T., 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5, e11147.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.

Dikow, R.B., 2011. Genome-level homology and phylogeny of *Shewanella* (Gammaproteobacteria: Alteromonadales: Shewanellaceae). BMC Genomics 12, 237.

Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers for target enrichment spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.

Goloboff, P., Farris, J.S., Nixon, K.C., 2008. TNT: a free program for phylogenetic analysis. Cladistics 24, 774–786.

Hennig, W. 1966. Phylogenetic Systematics, University of Illinois Press, Urbana.

Huelsenbeck, J.P., Ronquist, F., 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Kluge, A.G., 1989. A concern for evidence and a phylogenetic hypothesis for relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38, 1–25.

Kluge, A.G., Wolf, A.J., 1993. Cladistics: what's in a word? Cladistics 9, 183–199.

Konstantinidis, K.T., Serres, M.H., Romine, M.F., Rodrigues, J.L.M., Auchtung, J., McCue, L., Lipton, M.S., Obraztsova, A., Giometti, C.S., Nealson, K.H., Fredrickson, J.K., Tiedje, J.M., 2009. Comparative systems biology across an evolutionary gradient within the *Shewanella* genus. Proc. Natl Acad. Sci. USA 106, 15909–15914.

Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24, 2542–2543.

Lovley, D.R., Phillips, E.J.P., 1988. Novel mode of microbial energy metabolism – organic-carbon oxidation coupled to dissimilatory

reduction of iron or manganese. Appl. Environ. Microbiol. 54, 1472–1480.
Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.
Nixon, K.C., Carpenter, J.M., 1996. On simultaneous analysis. Cladistics 12, 221–241.
Nixon, K.C., Carpenter, J.M., 2012. On homology. Cladistics 28, 160–169.
Perry, K.A., Kostka, J.E., Luther, G.W., 1993. Mediation of sulfur speciation by a Black-Sea facultative anaerobe. Science 259, 801–803.
Rokas, A., Williams, B., King, N., Carroll, S., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.
Varón, A., Vinh, L.S., Wheeler, W.C., 2010. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics 26, 72–85.
Zwickl, D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis, University of Texas at Austin.