Reassessment of the Lineage Fusion Hypothesis for the Origin of Double Membrane Bacteria

Kristen S. Swithers[1], Gregory P. Fournier[3], J. Peter Gogarten[1], Pascal Lapierre[2, 4]

1: Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut 06269, USA

2: University of Connecticut Biotechnology Center, University of Connecticut, Storrs, Connecticut 06269, USA

3: Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

4: corresponding author, email: pascal.lap@gmail.com, phone: 860-486-8742, fax: 860-486-5005

Key Words:

Lineage fusion

Gene transfers

Double membrane prokaryotes

Sampling Bias

Abstract

In 2009, James Lake introduced a new hypothesis in which reticulate phylogeny reconstruction is used to elucidate the origin of Gram-negative bacteria. The presented data supported the Gram-negative bacteria originating from an ancient endosymbiosis between the Actinobacteria and Clostridia. His conclusion was based on a presence-absence analysis of protein families that divided all prokaryotes into five groups: Actinobacteria, Double Membrane bacteria (DM), Clostridia, Archaea and Bacilli. Of these five groups, the DM are by far the largest and most diverse group compared to the other groupings. While the fusion hypothesis for the origin of double membrane bacteria is enticing, we show that the signal supporting an ancient symbiosis is lost when the DM group is broken down into smaller subgroups. We conclude that the signal detected in James Lake's analysis in part results from a systematic artifact due to group size and diversity combined with low levels of horizontal gene transfer.

Main Text

Symbioses and endosymbioses have shaped and continue to shape microbial evolution [1]. As such, it is little surprise that endosymbiotic events and chimaerism are often considered useful hypotheses for explaining the phylogenetic and gene content complexities of bacterial, archaeal and eukaryotic genomes. James Lake used a reconstruction of reticulate phylogeny to argue that the double membrane bacteria evolved from an ancient symbiosis (endosymbiosis) between Clostridia and Actinobacteria [2]. By applying a parsimony analysis of protein family presence absence data over five distinct groups of prokaryotes [3], he identified sets of proteins present in double membrane bacteria (DM) that originated from either Clostridia or Actinobacteria. Since the highest number of protein families from the presence-absence patterns had better support for a ring structure compared to a single bifurcating tree, he concluded that the most likely explanation for the data was a fusion event between Clostridia and Actinobacteria. If this fusion occurred through an endosymbiosis, it could also explain the origin of the double membrane architecture. To further support this claim, he used indels (insertions or deletions) in several proteins that characterize the DM bacteria as a true monophyletic clade and showed their close relationship with the Actinobacteria. He also argued that the complexity of the photosynthetic machinery makes it a character that is not horizontally transferred and thus a good candidate to study ancient divergences.

One problem with this analysis is that the group designated DM is comprised of many rather divergent groups of bacteria, such as the Dictyoglomi, Thermotogae, *Deinococcus-Thermus*, Cyanobacteria and the different classes of Proteobacteria (see materials and methods for full listing). The definition of what constitutes a genuine double membrane compared to an external proteolipid or protein layer is unclear, and the constituents of the outer layer are difficult to determine [4]. For this reason, the majority of phyla included as double-membrane organisms are controversial and have possibly introduced an artifactual signal in favor of a fusion. Given the amount of interdomain and interphylum horizontal gene transfer that has been identified (e.g., [5-7], one should expect a larger group of organisms to harbor more different protein families than a smaller group. This

alternative explanation for Lake's data is testable; if the reticulate signal detected by Lake were due to many transfers of individual genes and operons, it should diminish if the DM group is replaced in the analysis with any of its biologically cohesive constituent subgroups. In contrast, if the signal were due to a single ancient endosymbiotic event at the root of the DM bacteria, then the signal should not disappear even if only a subgroup of the DM were selected in the analysis.

The claim that DM bacteria evolved from an ancient symbiosis is based on an analysis that aggregates all Bacteria and Archaea into 5 groups (the double membrane prokaryotes (DM), Actinobacteria (A), Bacilli (B), Clostridia (C), Archaea (R)), using the Pfam database [8] to determine the number of protein families that were represented in 3 out of the 5 aggregate groups. A protein family (Pfam) was considered present in a group, if at least one genome within the group encoded a member of this family. The analysis produces a table of all possible combinations of presence-absence profiles and determines the most parsimonious scenario explaining the data (i.e., if they were generated by a tree-like or ring-like evolutionary process, see supplemental figure 1S). The ring structure proposed by Lake (2009) joins the DM group to both the A and C groups given the allowed patterns in rows 5, 7, 8, 9 and 10 (Figure 1). The presence of a higher number of genes in those five rows compared to the tree signal (rows 1, 4, 7, 8, 9 and 10 in Lake 2009, figure 2S) reflects a higher number of genes shared by DM members with Actinobacteria and Clostridia. If the argument for a fusion event were valid, trends observed in the gene presence-absence table should not be affected by the breakup of the DM into sub-groups, as the presence of the protein family would be shared derived characters of all DM members.

We repeated Lake's analysis exactly using the same version of the Pfam database, and in addition to Lake's DM group, we also analyzed the datasets that resulted after dividing the DM group into twelve subclasses (Figure 1, column one to twelve). We found that for most of the DM subgroups, tree patterns were more highly supported than the patterns allowed under the ring scheme proposed by Lake. Additionally, the signal supporting the hypothesis of an ancient endosymbiosis between Clostridia and Actinobacteria is

completely lost (p-values in favor of a ring of 0.0035 or smaller) when these subgroups are used as representatives for the DM group (Figure 2 and 2S). This result is compatible with the hypothesis that the reticulate signal is due to several HGTs of individual genes, operons, and gene clusters and not due to a single ancient fusion between lineages. The ring signal is retained only in one case, when all classes of the Proteobacteria are combined (p-value of 0.98), possibly because this group contains the largest sampled biodiversity as reflected by the number of protein families in Pfam compared to the other groups included in the analysis. Figure 2 summarizes our results. We conclude that the deduced reticulate phylogeny appears to be due to many individual gene transfer events. The division of prokaryotes into groups of different size and containing different amounts of sampled protein diversity produces a systematic artifact suggestive of a fusion at the base the group comprised of the most diverse members.

Lake's analysis assumes that the double membrane of bacteria included in the DM group is a shared derived character, permitting the inclusion of these different bacterial phyla under one category. Furthermore, the result that the DM group arose via a fusion event demands that this group be monophyletic rather than paraphyletic, since this result is inconsistent with the DM group giving rise to any other groups of bacteria included in the analysis. However, there is scant molecular evidence supporting the monophyly (or possibly even the paraphyly) of the bacterial phyla included in the DM group. While it has been claimed that a polarizing indel within the HSP70/MreB gene families excludes the root of the "tree of life" from gram-negatives [9] it is likely that this result is largely due to extensive horizontal gene transfer, and is complicated by alignment and sampling artifacts [10, 11]. More convincingly, a polarizing indel in the HisA/HisF protein families and the quaternary structure of PyrD homologs have also been used to exclude the root from most gram-negatives and actinobacteria [12]. While the results of these analyses can be construed to permit the monophyly of the DM group, in reality they permit any scenario where each DM subclass is derived, including a paraphyly or even polyphyly incompatible with the assumptions in [2]. For this reason, there is currently little biological justification for the creation of a DM group as described in (Lake 2009). In addition, the argument that the photosynthetic machinery is reluctant to gene transfers

because of its complexity, thus linking the Clostridia from one side of the ring of life to the DM bacteria, is also being challenged. Previous reports have shown that many genes of the photosynthetic machinery, and of the chlorophyll biosynthetic pathway were transferred between bacterial classes and phyla [13, 14], Sharon *et al*. reported the discovery of a complete photosystem I operon in a marine phage [15], and analysis by Igarashi *et al*. [16] suggested that a photosynthetic gene super-cluster in the β-Proteobacteria was acquired through transfer from the α-Proteobacteria.

Cavalier-Smith [17] suggested that the root of the tree of life be placed within the Gram-negatives. This placement of the root is based on the argument that a double-membrane architecture could not have evolved from a single membrane ancestor; however, the process of endospore formation illustrates that double membrane envelopes can originate from single membrane ancestors. Endosymbiosis as proposed by Lake (2009) between two single membrane organisms offers an additional possible scenario for the evolution of double membranes as a derived character. However, we show that the re-evaluation of reticulate evolution using presence-absence of protein families does not provide convincing evidence in favor of a fusion to explain the origin of double-membrane prokaryotes. Which bacterial phyla may be aggregated into a holo- or paraphyletic group labeled double membrane bacteria remains an unresolved question. Regardless of how this question is resolved, our analysis using subclasses reveals that the parsimony approach of Lake (2009) tends to infer origin by fusion for that group of organisms containing the greatest amount of protein diversity.

Methods

The complete Pfam database v.22.0 was downloaded from ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam22.0/ and was locally searched for presence of protein families across different groups. We divided the Pfam database into five groups according to Lakes specifications as described in [2]. Group 1 was composed of all the Archaeal protein families; group 2 was composed of the Actinobacteria; Group 3 are the Bacilli which includes the Lactobacillales and the Bacillales; Group 4 was represented by the Clostridia and Mollicutes which also included the *Symbiobacterium*,

Coriobacteriales and the Rubrobacteridae; and finally group five which represents all the double membrane prokaryotes (Acidobacteria, Aquificae, Bacteroidetes, Chrysiogenetes, Chloroflexi, Chlorobi, Chlamydiae, Cyanobacteria, Deferribacteres, Deinococcus/Thermus, Dictyoglomi, Fibrobacteres, Fusobacteria, Nitrospirae, Proteobacteria, Planctomycetes, Spirochaetes, Thermodesulfobacteria, Thermotogales and Verrucomicrobia). Figure 1 shows the ten parsimonious informative character states for the five group comparisons. Following Lake's methods, a protein family was deemed present if at least one member of the three subject groups contained that protein family and was absent in all members of both query groups. Using the original group classification, we recovered the exact numbers of protein families for the ten-character state as described by Lake. We then compiled the number of protein families present when the double membrane group was broken up into twelve subgroups: Proteobacteria, Cyanobacteria, Bacteroidetes, δ-Proteobacteria, α-Proteobacteria, γ-Proteobacteria, Acidobacteria, Chlorobi, Chloroflexi, Deinococcus/Thermus, Planctomycetes and Spirocheates. The posterior bootstrap support values (p-values) for all possible ring and tree models were calculated from 10,000 re-samplings with replacement and extracting the total number of times the tree model, ring model or both were equally supported. For each bootstrap replicate, the best supported model was determined by finding the tree or ring with the lowest minimum parsimony count. The minimum parsimony counts were calculated by weighting the number of Pfams supporting a particular tree or ring twice that of the number of Pfams that do not support the model [3].

References

1. Margulis, L. (1995). Symbiosis in Cell Evolution : Microbial Communities in the Archean and Proterozoic Eons, 2nd Edition, (W H Freeman & Co).
2. Lake, J.A. (2009). Evidence for an early prokaryotic endosymbiosis. Nature *460*, 967-971.
3. Lake, J.A. (2008). Reconstructing evolutionary graphs: 3D parsimony. Mol Biol Evol *25*, 1677-1682.
4. Sutcliffe, I.C. (2010). Cell envelope architecture in the Chloroflexi: a shifting frontline in a phylogenetic turf war. Environ Microbiol.
5. 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., et al. (2009). On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc Natl Acad Sci U S A *106*, 5865-5870.
6. Boussau, B., Gueguen, L., and Gouy, M. (2008). Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC Evol Biol *8*, 272.
7. Beiko, R.G., Harlow, T.J., and Ragan, M.A. (2005). Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A *102*, 14332-14337.
8. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2010). The Pfam protein families database. Nucleic Acids Res *38*, D211-222.
9. Gupta, R.S., and Singh, B. (1994). Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr Biol *4*, 1104-1114.

10. Gribaldo, S., Lumia, V., Creti, R., de Macario, E.C., Sanangelantoni, A., and Cammarano, P. (1999). Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. J Bacteriol *181*, 434-443.

11. Gogarten, J.P. (1994). Which is the most conserved group of proteins? Homology-orthology, paralogy, xenology, and the fusion of independent lineages. J Mol Evol *39*, 541-543.

12. Valas, R.E., and Bourne, P.E. (2009). Structural analysis of polarizing indels: an emerging consensus on the root of the tree of life. Biol Direct *4*, 30.

13. Xiong, J., Fischer, W.M., Inoue, K., Nakahara, M., and Bauer, C.E. (2000). Molecular evidence for the early evolution of photosynthesis. Science *289*, 1724-1730.

14. Raymond, J., Zhaxybayeva, O., Gogarten, J.P., Gerdes, S.Y., and Blankenship, R.E. (2002). Whole-genome analysis of photosynthetic prokaryotes. Science *298*, 1616-1620.

15. Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., Pinter, R.Y., Partensky, F., Koonin, E.V., Wolf, Y.I., et al. (2009). Photosystem I gene cassettes are present in marine virus genomes. Nature *461*, 258-262.

16. Igarashi, N., Harada, J., Nagashima, S., Matsuura, K., Shimada, K., and Nagashima, K.V. (2001). Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. J Mol Evol *52*, 333-341.

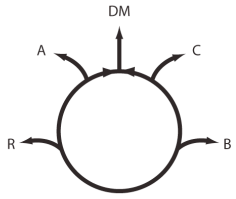
17. Cavalier-Smith, T. (2006). Rooting the tree of life by transition analyses. Biol Direct *1*, 19.

Figure 1. Protein family counts for the ten possible informative profiles. The table was adapted from Lake's Table 1 [2] to include the Pfam counts that result if different representative classes are chosen for the DM group. Number of Pfam per group is in parentheses the same number as in Lake's paper was found for all other groups. The circle illustrates Lake's hypothesis that the double membrane bacteria resulted from a fusion between Clostridia and Actinobacteria. The patterns compatible with this hypothesis are boxed (pattern 5,7,8,9 and 10).

	R	A	B	C	DM	Lake*	1	2	3	4	5	6	7	8	9	10	11	12
1	+	+	+			**3**	5	21	21	16	8	8	30	36	31	39	32	38
2	+	+		+		**0**	4	33	33	17	19	12	40	49	29	60	42	65
3	+		+	+		**8**	22	89	70	40	60	46	94	87	78	89	81	83
4		+	+	+		**15**	34	125	170	128	108	57	228	311	232	284	278	249
5	+	+			+	**62**	60	32	28	46	46	47	26	14	26	17	18	8
6	+		+		+	**15**	15	3	7	6	5	13	3	5	2	4	0	0
7		+	+		+	**91**	89	42	42	49	68	74	18	11	26	18	13	8
8	+			+	+	**99**	94	27	38	73	52	59	18	31	37	17	35	23
9		+		+	+	**73**	73	32	38	54	51	66	36	21	23	15	25	21
10			+	+	+	**174**	151	39	64	96	83	130	36	33	32	19	32	52

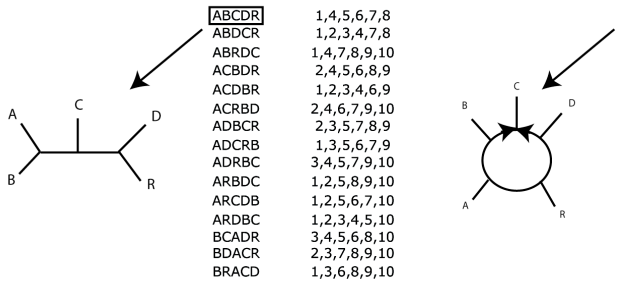
1) DM only composed of Proteobacteria (4345 Pfam)
2) DM only composed of Cyanobacteriales (2200 Pfam)
3) DM only composed of Bacteroidetes (2410 Pfam)
4) DM only composed of Delta-Proteobacteria (2874 Pfam)
5) DM only composed of Alpha-proteobacteria (3099 Pfam)
6) DM only composed of Gamma-Proteobacteria (3843 Pfam)
7) DM only composed of Acidobacteria (2411 Pfam)
8) DM only composed of Chlorobi (1673 Pfam)
9) DM only composed of Chloroflexi (1956 Pfam)
10) DM only composed of Deinococcus/Thermus (1656 Pfam)
11) DM only composed of Planctomycetes (1810 Pfam)
12) DM only composed of Spirocheates (1710 Pfam)

* Lake original results
R = Archaea (2307 Pfam)
A = Actinobacteria (2641 Pfam)
B = Bacilli (2854 Pfam)
C = Clostridia (2820 Pfam)
DM = Double Membrane bacteria (4756 Pfam)



	Lake Groups	Proteobact. 1	Cyanobact. 2	Bacteroides 3	Delta- 4	Alpha- 5	Gamma- 6	Acidobact. 7	Chlorobi 8	Chloroflexi 9	Deino/Thermus 10	Plancto. 11	Spiro. 12
Ring Supports :	1	0.9812	0.0001	0.002	0.001	0	0.0001	0	0.0035	0.0002	0	0.0116	0.0023
Tree Supports :	0	0.0148	0.9998	0.9973	0.9983	1	0.9999	0.9999	0.9942	0.9991	1	0.9838	0.9971
Ties :	0	0.004	0.0001	0.0007	0.0007	0	0	0.0001	0.0023	0.0007	0	0.0046	0.0006

Figure 2. Posterior bootstrap support values (p-values) for a ring model, tree model or equal probabilities for each of the sampled groups. The p-values were calculated from 10,000 re-samplings with replacement and extracting the total number of times the tree model, ring model or when both were equally supported from the parsimony counts. Only in the case where all the double membrane prokaryotes as defined by Lake (2009), or when all the proteobacteria were included, did a ring model better explain the data.



TREE	Compatible Patterns
ABCDR	1,4,5,6,7,8
ABDCR	1,2,3,4,7,8
ABRDC	1,4,7,8,9,10
ACBDR	2,4,5,6,8,9
ACDBR	1,2,3,4,6,9
ACRBD	2,4,6,7,9,10
ADBCR	2,3,5,7,8,9
ADCRB	1,3,5,6,7,9
ADRBC	3,4,5,7,9,10
ARBDC	1,2,5,8,9,10
ARCDB	1,2,5,6,7,10
ARDBC	1,2,3,4,5,10
BCADR	3,4,5,6,8,10
BDACR	2,3,7,8,9,10
BRACD	1,3,6,8,9,10

RING	compatible Patterns	RING	Compatible Patterns
ABCDR	3,4,8,9,10	BDCAR	2,4,8,9,10
ABCRD	2,3,4,8,10	BDCRA	2,3,8,9,10
ABDCR	6,7,8,9,10	CBADR	1,4,5,7,9
ABDRC	5,6,7,8,10	CBARD	1,2,4,5,7
ADBCR	3,4,6,7,10	CBDAR	5,6,7,9,10
ADBRC	1,3,6,7,10	CDBAR	1,4,6,7,10
ADRBC	1,3,5,6,8	CDRAB	1,2,5,6,8
ADRCB	2,3,5,6,8	DABRC	1,3,4,6,7
BACRD	2,3,4,8,9	DARBC	1,2,3,5,6
BADRC	5,6,7,8,9	DBACR	1,2,4,7,9
BCARD	1,2,4,5,9	DBCAR	2,3,4,7,10
BCDAR	5,7,8,9,10	DCBAR	1,3,4,7,10
BCDRA	5,6,8,9,10	DCBRA	1,3,4,6,10
BDACR	2,4,5,7,9	DCRAB	1,2,3,5,8
BDARC	1,2,5,7,9	DCRBA	1,2,3,6,8

Figure 1S. List of all possible trees and rings for five taxa sampling. Each possible tree and ring is listed with the compatible presence-absence pattern of gene families (Pfam) given in Figure 1. For example, the tree and ring corresponding to ABCDR are shown at the left of each table. A corresponds to Actinobacteria, B to Bacilli, C to Clostridia, D for double membrane prokaryotes and R for Archaea.

A) Trees

	Lake*	1	2	3	4	5	6	7	8	9	10	11	12
ABCDR	795	797	636	716	732	713	766	735	788	678	745	736	768
ABDCR	864	846	549	648	727	685	768	630	671	599	617	631	628
ABRDC	625	648	600	649	634	630	630	692	753	651	732	697	703
ACBDR	816	814	634	708	726	719	770	707	765	683	727	714	728
ACDBR	966	941	583	683	789	749	822	627	687	637	633	654	638
ACRBD	712	728	612	668	700	666	672	697	766	688	724	722	699
ADBCR	747	752	631	773	771	704	720	826	983	813	908	898	886
ADCRB	828	830	667	816	839	762	770	851	1022	846	942	943	936
ADRBC	657	665	527	610	637	584	604	620	719	615	682	665	673
ARBDC	669	707	702	800	748	741	702	872	1012	854	957	928	887
ARCDB	735	770	716	827	820	771	740	905	1048	886	967	975	923
ARDBC	818	818	547	636	707	676	724	604	666	604	616	629	599
BCADR	707	718	571	645	661	646	672	653	715	625	694	668	679
BDACR	635	661	624	737	721	667	637	816	964	807	906	884	842
BRACD	708	734	675	784	765	741	702	841	983	829	941	907	877

B) Rings

	Lake*	1	2	3	4	5	6	7	8	9	10	11	12
ABCDR	711	720	574	642	659	646	666	646	713	630	700	661	666
ABCRD	784	789	573	647	696	678	720	642	685	624	655	644	622
ABDCR	628	672	743	833	772	741	682	947	1095	912	1051	1007	990
ABDRC	639	685	743	843	780	746	701	957	1102	909	1049	1014	1003
ADBCR	777	783	588	669	731	676	704	679	749	662	710	708	702
ADBRC	789	812	692	818	843	776	753	877	1024	863	955	954	913
ADRBC	893	898	714	858	869	829	851	887	1023	858	958	946	942
ADRCB	896	899	702	846	868	818	847	877	1010	860	937	936	915
BACRD	885	867	580	673	738	710	784	642	697	633	659	651	653
BADRC	740	763	750	869	822	778	765	957	1114	918	1053	1021	1034
BCARD	927	918	643	732	789	768	834	698	765	691	709	717	713
BCDAR	581	627	714	812	732	700	648	924	1086	888	1038	989	982
BCDRA	657	701	753	847	775	763	709	939	1092	912	1052	1002	990
BDACR	839	834	622	711	756	708	768	710	790	696	730	736	743
BDARC	851	863	726	860	868	808	817	908	1065	897	975	982	954
BDCAR	719	738	630	679	682	687	700	700	751	679	729	700	684
BDCRA	726	750	666	779	770	735	711	834	975	833	924	897	850
CBADR	836	833	634	723	757	719	772	720	803	694	751	746	770
CBARD	909	902	633	728	794	751	826	716	775	688	706	729	726
CBDAR	665	706	738	843	799	747	694	939	1112	923	1051	1024	1005
CDBAR	782	800	656	718	755	728	742	743	800	709	760	757	747
CDRAB	901	916	770	895	892	870	885	941	1061	907	987	985	960
DABRC	948	929	606	712	811	751	826	685	746	663	690	708	716
DARBC	992	988	708	863	925	862	898	865	1005	866	915	939	900
DBACR	898	889	633	718	786	746	807	706	768	691	708	722	713
DBCAR	792	794	558	643	720	662	705	642	705	635	654	666	637
DCBAR	789	793	570	655	721	673	709	652	718	633	675	676	664
DCBRA	865	867	609	690	764	736	770	667	724	657	689	689	672
DCRAB	908	909	684	832	858	815	852	850	979	831	902	904	877
DCRBA	955	954	713	853	898	856	886	873	988	855	915	922	885

Figure 2S. Minimum parsimony counts supporting each of the possible trees (A) and rings (B). The lowest count is used to determine if the data supports a tree or a ring (Lake 2008). In the original analyses by Lake (2009), the best ring had a minimum parsimony count of 581 versus 625 for the best supported tree (first column). Best supported trees or rings for each tested cases are highlighted.