



A whole-genome phylogeny of the family Pasteurellaceae

Maria Pia Di Bonaventura^{a,b}, Ernest K. Lee^{a,c}, Rob DeSalle^{a,*}, Paul J. Planet^{a,d}

^a Sackler Institute of Comparative Genomics, American Museum of Natural History, 79th Street at Central Park West, New York, NY 10024, USA

^b Department of Biology, York College/CUNY, Jamaica, NY 11451, USA

^c Department of Biology, New York University, New York, NY 10003, USA

^d Pediatric Infectious Disease Division, Children's Hospital of New York, Columbia College of Physicians and Surgeons, 630 West 168th Street, New York, NY 10032, USA

ARTICLE INFO

Article history:

Received 1 June 2009

Revised 5 August 2009

Accepted 11 August 2009

Available online 15 August 2009

Keywords:

Pasteurellaceae

Concatenation

Total evidence

Whole genomes

Phylogenomics

ABSTRACT

A phylogenomic approach was used to generate an amino acid phylogeny for 12 whole genomes representing 10 species in the family Pasteurellaceae. Orthology of genes was determined using an approach similar to OrthologID (<http://nybg.bio.nyu.edu/orthologid/about.html>) and resulted in the generation of a matrix with 3130 genes with 1,194,615 aligned amino acid characters of which 239,504 characters are phylogenetically informative. Phylogenetic analysis of the concatenated matrix using all standard approaches (maximum parsimony, maximum likelihood, and Bayesian analysis) results in a single extremely robust phylogenetic hypothesis for the species examined in this study. Remarkably, no single gene partition gives the same tree as the concatenated analysis. By analyzing partitioned support in the data matrix, we show that there is very little negative support emanating from individual gene partitions to suggest that the concatenated hypothesis is not tenable. The large number of characters in the matrix allows us to test hypotheses concerning missing data and character number in phylogenomic studies, and we conclude that matrices constructed using genome level information are very robust to missing data. We show that a very large number of concatenated gene sequences (>160) are needed to reliably obtain the same topology as the overall analysis.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

The family Pasteurellaceae consists of 65 named species distributed in 13 genera (<http://www.the-icsp.org/taxa/Pasteurellaceae/ist.htm>). Of these 65 species recent reports suggest that approximately a third of species are misnamed based on phylogenetic taxonomy, with the four major genera of the family (*Haemophilus*, *Pasteurella*, *Actinobacillus*, and *Mannheimia*) mixed in different clades (Christensen and Bisgaard, 2008; Christensen et al., 2007).

For many members of the Pasteurellaceae family there is a delicate balance between commensal and pathogenic lifestyles within their vertebrate hosts. Species of this family are well-known to specialize as nonpathogenic members of the resident flora of the healthy upper respiratory tract and oral cavity of birds and mammals. However, it is not unusual for the commensal relationship to degenerate into a pathogenic one, resulting in damage to the host. As such, many typical pathogens of this family are also isolated in periods of health. For example non-typeable *Haemophilus influenzae* (NTHi), a major cause of upper respiratory infections in children, is also found in the respiratory tract of many disease-free individuals (Erwin and Smith, 2007). The differences between

pathogenic and commensal relationships with the host are complex and not well understood, but it is clear that both host and bacterial factors play a role.

Understanding the evolutionary changes in the Pasteurellaceae that have had an impact on virulence and symbiosis requires a clear understanding of the evolutionary history of the family. Moreover, since crucial changes may come from the addition or subtraction of more than one gene, a genomic perspective on evolutionary history is essential.

Initial sequence-based phylogenetic studies of the Pasteurellaceae family were based on single genes (Dewhirst et al., 1992, 1993; Korczak et al., 2004). More recent phylogenetic studies (Christensen et al., 2004; Gioia et al., 2006; Redfield et al., 2006) have included the added power of considering multiple genes in phylogenetic analysis. With over 10 species of Pasteurellaceae with whole-genome sequences it is now possible to use whole-genome datasets to assess the evolutionary relationships in this family. Table 1 shows several of the current whole genomes that have been sequenced in this family that are examined in the present study. Here we add the whole-genome of *Aggregatibacter (Haemophilus) aphrophilus* (Di Bonaventura et al., 2009) to 12 other whole genomes for a phylogenomic approach to generate a comprehensive genome level phylogeny for this group. In addition, we examine the effects of missing data, incongruence and amount of data on the concatenated hypothesis we generate.

* Corresponding author.

E-mail address: desalle@amnh.org (R. DeSalle).

Table 1
Genomes used in this paper.

Species – strains	Figure label	Accession Nos.
<i>Haemophilus ducreyi</i> 35000HP	H ducreyi	AE017143
<i>Haemophilus influenzae</i> 86 028NP	H influenzae I	NC_007146
<i>Haemophilus influenzae</i> KW20 Rd	H influenzae II	NC_000907
<i>Haemophilus somnus</i> 129PT	H somnus I	NC_008309
<i>Haemophilus somnus</i> Hs2336	H somnus II	NC_010519
<i>Haemophilus aphrophilus</i> NJ8700	H aphroph	CP001607
<i>Mannheimia succiniciproducens</i> MBEL55E	M succinipro	NC_006300
<i>Mannheimia haemolytica</i> PHL213	M haemolytica	NW_002062512
<i>Pasteurella multocida</i> PM70	P multocida	NC_002663
<i>Actinobacillus succinogenes</i> 130Z	A succinogen	NC_009655
<i>Actinobacillus pleuropneumoniae</i> L20	A pleuropneu I	NC_009053
<i>Actinobacillus pleuropneumoniae</i> 4074	A pleuropneu II	NZ_AACK00000000
<i>Actinobacillus actinomycetemcomitans</i>	A actinomy	Project ID 32179
<i>Aeromonas hydrophila</i> ATCC7966	A hydrophila	NC_008570

2. Materials and methods

2.1. Matrix construction

Whole-genome DNA sequences of the bacterial strains listed in Table 1 were compiled and archived. To establish orthology of genes we employed a process very similar to the OrthologID (Chiu et al., 2006) <http://nypg.bio.nyu.edu/orthologid/> approach developed for plant genomes and EST data. This procedure has two steps in establishing orthology. First it establishes gene family membership using Blast cutoff scores (an e -value of 1×10^{-20} was used as a preliminary cutoff for this study). Second it determines orthologous groups through phylogenetic analysis of each gene family. We compiled the matrix using a modified version of ASAP (Sarkar et al., 2008), a program that builds genome level phylogenetic matrices in NEXUS format. The matrix constructed by ASAP was a concatenated gene matrix of including each orthologous gene family.

2.2. Phylogenetic analysis

Maximum parsimony (MP) phylogenetic trees were constructed using PAUP (Swofford, 1999) with the heuristic search option using 100 random taxon additions and TBR branch swapping. Gaps in the amino acid alignments were treated as missing data. Bayes posterior probabilities were calculated using MrBayes (Huelsenbeck and Ronquist, 2001); the WAG substitution model was used as the input model in the Bayes analyses with 10,000 replicates. Maximum likelihood (ML) analysis was done in RAxML (Stamatakis, 2006). The WAG substitution model was used and bootstrap values were estimated with 100 replicates in RAxML. Partitioned analyses used TreeRot (Sorenson and Franzosa, 2007) to generate tree statistics. For this analysis the non-Pasteurellaceae gammaproteobacterium, *Aeromonas hydrophila*, was used as an outgroup.

2.3. Partitioned analysis

The ASAP (Sarkar et al., 2008) program was used to generate partitioned matrices for the over 2000 genes for the data set. Partitioned Branch Supports (PBS) and Partitioned Hidden Branch Support (Gatesy et al., 1999) were calculated using Perl scripts written in the authors lab and are available upon request from the authors. ILLD statistics were generated using scripts written in the same way as mILD (Planet and Sarkar, 2005) and using the “partition homogeneity” option in PAUP*. Tree consensus statistics were generated using the “indices” option in PAUP*.

2.4. Random partition addition

To examine the problem of how many genes are necessary to obtain a stable concatenated hypothesis, we constructed random matrices by taking a list of genes and randomizing the order of the genes. Lists of gene partitions were randomized in Excel using 100 randomizations. We used two lists of randomized gene partitions.

First we used all of the 2093 genes in the matrix with taxonomic representation of four or more taxa. We created 100 random lists of 1000 genes, in concatenated groups of five genes. To begin, we analyzed the first five genes in the random list, generating a tree. We then measured the topological similarity of this tree to the tree from the full concatenated dataset using the consensus fork index (CFI; (Colless, 1980)). The next five genes in the random list were then concatenated with the first set of 5, and the CFI of the resulting tree compared to the full dataset tree was again calculated. This process was repeated, sequentially adding 5-gene sets to the dataset until all 1000 randomly ordered genes were added together. The CFIs were tabulated for each of the resulting 200 trees. This process was repeated for each of the 100 randomized lists, and the CFI for each addition step was averaged over the 100 iterations. This strategy allowed us to determine the number of genes needed to obtain the concatenated tree on average for the whole dataset.

Next we used just those genes that had full taxonomic representation, that is, they were present in every taxon ($n = 633$ genes). We created 100 random lists of the 633 genes. In this approach, genes were added individually rather than in groups of 5 genes. CFIs were calculated after every individual addition. This process was repeated until all 633 randomly ordered genes were added together. The resulting 633 average CFI scores were tabulated and graphed using Excel.

3. Results and discussion

3.1. Matrix characteristics

We limited the matrix construction to ortholog groups (or partitions) that are present in at least four of the 14 taxa in our matrix. The ASAP procedure, applied with this restriction, generated 3130 partitions with a total of 1,194,615 aligned amino acid characters, of which, 239,504 characters are phylogenetically informative.

Not all of the gene partitions have a representative sequence for each of the 14 taxa in this analysis. We examined each partition to determine the number of taxa for which there is a representative sequence in the alignment. Here, we call the number of representative sequences present in the matrix for a given gene partition the “taxon completeness” of the partition. Fig. 1 shows a plot of this information. Of the 3130 genes detected in this analysis, 2094 are found in at least four taxa. There are 633 genes with full taxonomic representation. The number of partitions with mid-range taxon completeness (approximately 6–12) is much lower at about 100 partitions. Partitions that have taxon completeness less than six are more numerous.

This interesting pattern indicates that there is a very large “core” of over 600 genes common to all the Pasteurellaceae species we examined in this study. This core is surrounded by a relatively stable “shell”. This shell represents genes that are widespread, and likely have important biological significance since they are conserved in multiple genera. However, these genes do not seem to be required in every genome or maintained by every member of the family. A third group, with low taxon completeness, contains genes that are rare and found in only a few taxa. This group is very large with over 1000 of the partitions that have fewer than three taxa, and may represent genes involved in specialized functions.

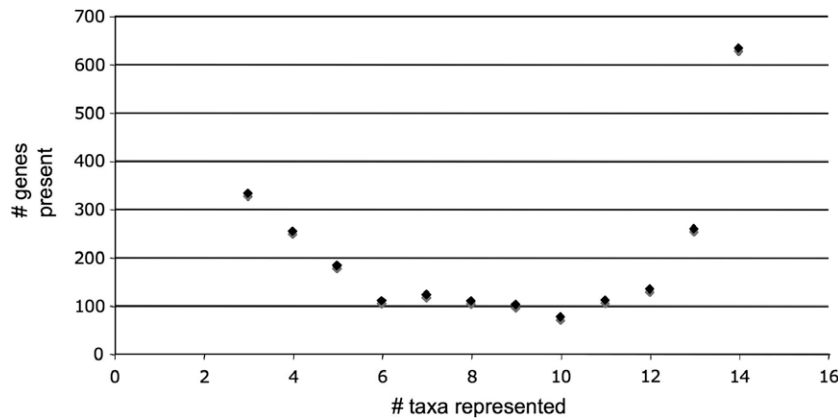


Fig. 1. Graph of the number of taxa for a gene on the x axis and the number of genes out of the 3130 gene partitions in the matrix with that number of taxa.

3.2. Phylogenetic hypothesis for the Pasteurellaceae

Fig. 2 shows the phylogenetic tree generated using the full, concatenated matrix of 1,194,615 amino acid characters. This tree is extremely robust and supported by 100% bootstrap and jackknife values for parsimony and likelihood at all nodes. All nodes in the tree in Fig. 2 retain 100% bootstrap and jackknife values with as few as 5% characters sampled for bootstrap, and as many as 95% character removal for jackknife. At 1% character sampling for bootstrap and 99% character removal for jackknife, the “weakest” node in the tree is still remarkably robust (95% bootstrap and jackknife values). This node unites the two *Haemophilus somnus* strains and *Pasteurella multocida*. Bremer or decay index values are shown in Fig. 2. These values measure the robustness of a particular inference made at nodes (see Gatesy et al., 1999). Unlike bootstraps and jackknives values they are not scaled to 100%, but rather indicate the number of steps longer

than the most parsimonious tree for a particular node to decay. All of the values in our analysis are extraordinarily high, ranging from 612 for the node defining the sister relationship of the two *H. somnus* strains with *P. multocida* to 29,920 for the node defining the two *H. influenzae* strains. These values indicate that to overturn the current phylogenetic hypothesis in Fig. 2 would require massive amounts of phylogenetic data that directly contradict these groupings. In addition to the large amount of gene sequence data in our analysis, other factors may also contribute to the extraordinarily high support for our topology. Namely we have added two species (*Aggregatibacter aphrophilus* and *Actinobacillus succinogenes*) as well as multiple representatives for *H. influenzae*, *H. somnus*, and *A. pleuropneumoniae*.

The phylogenetic hypothesis in Fig. 2 strongly questions current taxonomy in the family Pasteurellaceae, and if phylogeny is to be used as a guide to taxonomy, then this group will require extensive revision. The existence of two major clades that both contain

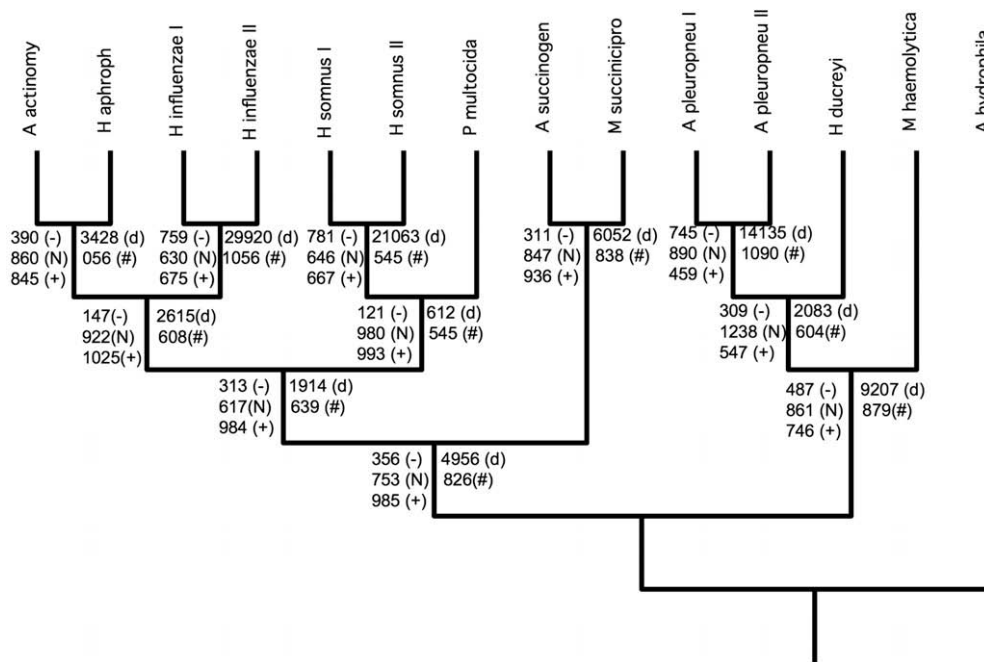


Fig. 2. Concatenated analysis tree. All nodes in the tree have 100% bootstrap and jackknife support for both maximum parsimony and maximum likelihood, as well as Bayesian posterior probabilities $p = 1.0$ (see text for more detail on these measures). Each node has five tree numbers next to it. The numbers at the nodes with a (d) next to them represent the total Bremer supports for the node. The number with the # symbol after it refers to the number of gene partitions with positive Bremer support measures. The three numbers on the left refer to the number of gene partitions that have negative hidden partitioned Bremer support (-), neutral or zero hidden partitioned Bremer support (N), and the bottom number refers to the number of gene partitions with positive hidden support (+) for the indicated node. See Table 1 for correspondence of strains to taxon names.

representatives of *Mannheimia* and *Haemophilus* genera suggests that these taxa will need extensive revision. The genus *Haemophilus* appears to be highly polyphyletic.

We compared our results to the results from two other recent multigene/concatenated studies by Gioia et al. (2006) for 50 core genes and Redfield et al. (2006) for 12 core genes (Fig. 3). The phylogeny inferred here is concordant with these studies in representing two major clades for Pasteurellaceae species. The first clade is represented by *M. haemolytica*, *H. ducreyi*, and *A. pleuropneumoniae*. Within this clade, the relationships are identical when compared to these other studies. In addition, bootstrap support for these relationships is very high in all three studies. The second major clade contains *A. actinomycetemcomitans*, *H. somnus*, *H. influenzae*, *M. succiniciproducens*, and *P. multocida*, but relationships within this clade differ amongst studies. Overall, the Gioia et al. tree is very similar to our tree, showing a strongly supported relationship between *H. somnus* and *P. multocida*. Our tree, in fact, only differs from the Gioia et al. tree in making the *Aggegatibacter* genus a monophyletic group with *H. influenzae* strains. The Redfield et al. tree differs significantly in the placement of *A. actinomycetemcomitans* and *H. influenzae*. Of note, bootstrap support for the placement of these 2 taxa is the lowest (80%) in the Redfield et al. tree. Any differences we observe between our phylogeny and the Redfield et al. (2006) and Gioia et al. phylogeny can almost certainly be attributed to the differences in the number of characters in the analyses.

3.3. Rampant incongruence of single gene trees with the concatenated hypothesis

To examine the impact of massively concatenated analysis as an approach to phylogenetics, we examined the topology of the 2093 gene trees that have greater than three taxa represented in the single gene matrices. To assess topological agreement we used the consensus fork index (CFI), which simply counts the number of nodes in each single gene tree that are also present in the concatenated analysis tree. We used two techniques to generate each gene tree. First, we used the 50% bootstrap consensus tree for each partition. This tree is more likely to have fewer resolved nodes when there is lack of support for a particular node. Second, we used the MP tree (or if there was more than one MP tree we used the strict consensus of all MP trees) for a particular gene partition. We then scored gene trees generated by both of these techniques based on the CFI when compared to the massively concatenated tree. Gene partitions with incomplete taxon representation will have lower CFI indices simply because they do not have all taxa present. We therefore performed our analysis using all gene parti-

tions (Fig. 4A) and also limited our study to the 633 partitions with complete taxonomic representation (Fig. 4B).

Fig. 4 shows the results of the CFI analysis. As expected, the bootstrap tree approach for both kinds of taxonomic representation showed fewer gene partitions with high CFIs, while more resolved MP trees showed more topological congruence as indicated by higher CFIs for a larger proportion of partitions. Remarkably, not a single gene partition analyzed by itself gives the full concatenated gene topology. In fact, only a single partition had agreement with nine of the 11 possible nodes in the concatenated topology.

If taxon-incomplete gene partitions are considered, the number of genes with no agreement at all is extremely high. Indeed, the

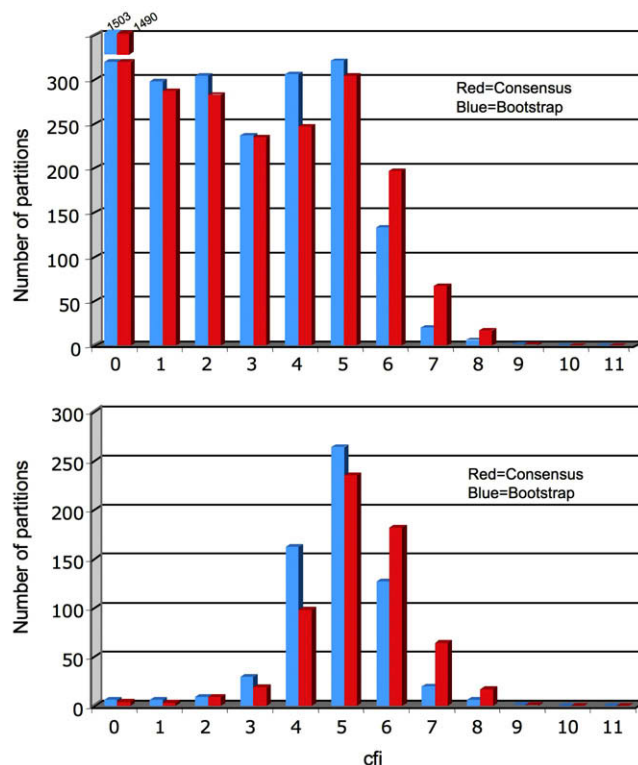


Fig. 4. Histograms of distribution of CFI values for gene trees. Red bars refer to tallies made for MP tree for each gene. Blue bars refer to tallies made for 50% bootstrap trees for each gene. The top graph shows results for all gene partitions with three or more taxa represented for a gene. The bottom graph summarizes the data for just the 633 gene partitions that have a full complement of 14 taxa. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

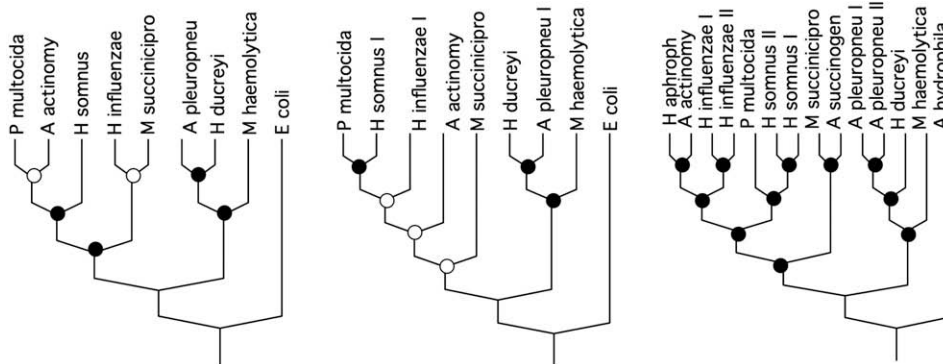


Fig. 3. Comparison of phylogenetic hypothesis from the present study (on right) with previous hypotheses for the Pasteurellaceae. The leftmost tree is reproduced from Redfield et al. (2006) and the middle hypotheses is from Gioia et al. (2006). The filled in circles on the nodes of the tree represent bootstraps >95%. The open circles represent bootstraps >65% but <95%. See text for discussion of these hypotheses. See Table 1 for correspondence of strains to taxon names.

majority of partitions actually show no agreement at all with the concatenated analysis. However, when those partitions with a full complement of the 14 taxa are examined the number of partitions showing no support drops drastically, demonstrating that the complete disagreement is largely due to incomplete taxon representation. When taxon-incomplete gene partitions are excluded, the average number of identical nodes between concatenated and individual gene trees is 5.3 ± 1.2 (for MP trees) and 4.8 ± 1.7 (for bootstrap trees). This result indicates that even for partitions with full taxonomic representation, over half of the nodes, on average, in the concatenated tree are not found in the individual gene trees.

In addition to examining incongruence by comparing tree topologies, we also used the Incongruence Length Difference (ILD) test (Farris et al., 1994) to estimate the proportion of partitions that show statistically significant incongruence. We used the 633 gene partitions with full taxonomic representation. Because of the large number of pairwise comparisons possible for this number of partitions we randomly selected 1000 pairwise comparisons. Out of the 1000 comparisons we examined, 63% show significant ILD values between the two partitions being compared at $p < 0.05$ (74% at $p < 0.01$). These results are indicative of rampant emergent incongruence amongst single gene partitions.

3.4. Behavior of phylogenetic analysis as a function of missing data

Because our concatenated matrix consists of partitions with varying taxon completeness, we can examine the effect of missing data on phylogenetic hypotheses. We examined two aspects of missing data in this context. First, we explored the effect of missing

data on tree topology, by examining tree topology as a function of taxon completeness. We did this by sequentially removing gene partitions with specific numbers of genes present from the overall analysis. Fig. 5 summarizes this analysis, and it demonstrates that even when the maximum taxon completeness is only eight taxa the data give the same tree topology as the concatenated full analysis. This result is encouraging for phylogenomic studies where incomplete taxon sampling might be a problem. This result further suggests that full overlap of taxa in a matrix is not necessarily a requirement for obtaining a reasonable phylogenetic hypothesis at least in this group of organisms.

Second, we examined the erosion of bootstrap support as gene partitions with higher taxon completeness were sequentially excluded from the analysis. We successively examined bootstrap support as gene partitions with more taxa were removed from the analysis. Fig. 6 shows the results of this analysis and indicates that bootstrap values erode only after partitions with eight and more taxa are removed from the analysis. Further erosion of robustness is evident with removal of partitions with seven and more taxa, and all bootstrap support is removed when partitions with six or more taxa are removed.

3.5. Partitioned support and hidden support

No single partition of the 3130 genes completely agrees with the concatenated topology (Fig. 4). We further examined this result by calculating partitioned support measures for each node in the concatenated tree. We used two support measures – partitioned Bremer support (PBS) and partitioned hidden Bremer support

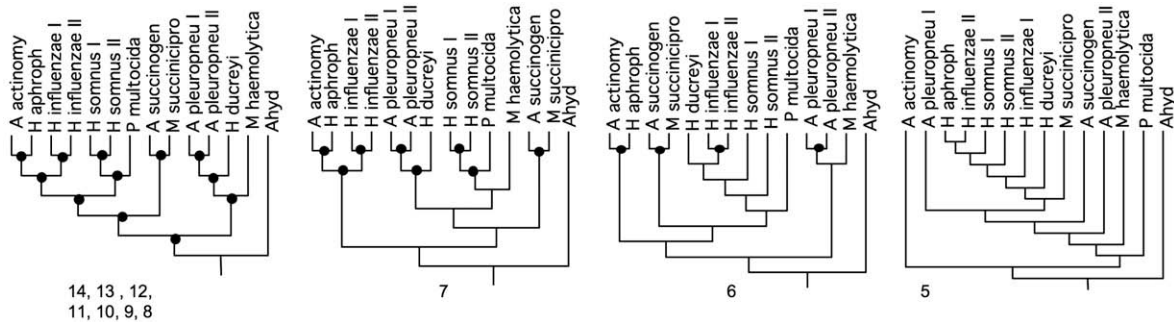


Fig. 5. The loss of concatenated analysis tree topology with increasing missing data. Trees for different matrices constructed by including all partitions and then sequentially excluding the partitions with the highest number of sequences. We first excluded partitions with 14 taxa represented, then for partitions with all but one taxon represented (taxon completeness = 13), then all but two taxa represented (taxon completeness = 12), and so on. Excluding gene partitions down to a taxon completeness of 8 does not disturb the topology of the tree. Colored dots represent nodes observed in the concatenated analysis tree. Note that trees constructed with gene partitions with taxon completeness of 5 or less have no nodes in common with the concatenated tree (far right). See Table 1 for correspondence of strains to taxon names.

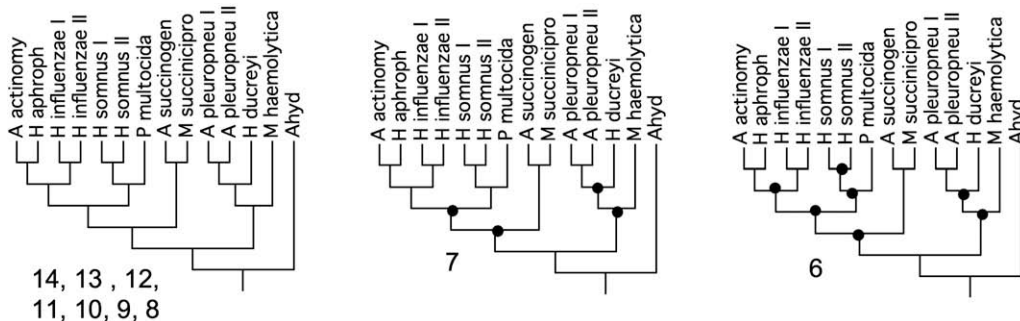


Fig. 6. The erosion of bootstrap and jackknife support as partitions with higher levels of taxon completeness are removed from the analysis. The concatenated analysis tree is shown as a reference. Nodes that collapse due to having bootstrap support less than 60% bootstrap are indicated by dots. Numbers below trees indicate the maximal taxon completeness in each dataset as in Fig. 5. Note that no node collapses until the maximum taxon completeness of 6. Also, note that the only nodes that do not erode with 6 or fewer taxa in them are those that define closely related strains. See Table 1 for correspondence of strains to taxon names.

(PHBS) (Gatesy et al., 1999). PHBS measures the character support from each gene partition for a particular node that exists in the concatenated tree. PHBS measures the emergent support as a result of data combination or concatenation. In essence, PHBS for a gene partition measures the amount of information in a gene partition that *does* agree with the concatenated tree, even if separate analysis of the gene partition does not agree. If a PHBS measure is positive it means that there is an increase in support within the partition for the concatenated, combined tree when the gene partition is added into the analysis. If the hidden support is negative for a gene partition it means that there is no evidence to support the concatenated hypothesis and inclusion in the concatenated analysis actually weakens the overall hypothesis.

Fig. 2 shows the distribution of hidden support and displays the number of gene partitions that contribute negative, neutral (0), or positive hidden support at each node. The results in this figure demonstrate that the absolute value of the negative hidden support in the data set is, in all cases but one, much less than positive hidden support. The figure also demonstrates the large number of gene partitions contributing positive hidden support to the concatenated topology. For instance, at the node that unites *A. actinomycetemcomitans* and *A. aphrophilus*, there are 390 genes that appear to contribute negative hidden support, nearly 900 that contribute zero hidden support, but over 800 genes that contribute positive hidden support that totals to 13,428 steps, and jackknife and bootstrap values at the node of 100%. An example of the largest amount of negative hidden support is the node defining the two *A. pleuropneumoniae* as sister taxa. At this node there are 745 genes that contribute negative hidden support, nearly 900 that are neutral, and only 459 that contribute positive hidden support. This node is supported by 1090 genes for a total of 14,135 steps along with 100% bootstrap and jackknife values.

Negative hidden support signals a strongly different phylogenetic signal emanating from a particular gene partition. In other words, any genes with negative hidden support have likely experienced different evolutionary histories from the concatenated set of genes. These different evolutionary histories could be caused by natural selection, drift, duplication and subsequent lineage sorting, or more likely, in the case of bacteria, by horizontal gene transfer.

3.6. Can 3130 wrongs make a right? The concatenation debate

One motivation behind concatenation in phylogenetic analysis is that it promises both the most complete picture of the data and the most severe test of any single hypothesis. That is, concatenation allows as much data as possible to be included that can refute the overall hypothesis (Kluge, 1997). Thus, it could be surmised that tree hypotheses that stand up to massive genome-scale concatenation would be the strongest, most well-tested trees.

When genomes are partitioned into genes for individual analysis there is also the possibility that the act of partitioning, by chance alone, will create false differences in phylogenetic signal since the effect of homoplasy is greater in smaller datasets (Farris et al., 1994; Huelsenbeck and Bull, 1996). In a similar vein, it is often hoped that concatenation will allow the underlying vertical evolutionary signal to emerge when individual genes may have experienced events (such as horizontal transfer events) that cause parts of their history to diverge from the species phylogeny.

Opponents of concatenation analysis suggest that it underestimates the effects of real underlying differences in the histories of individual genes. Further, simulated evolution studies of trees containing four taxa have recently suggested that under certain likelihood models, concatenation can lead to moderate to high bootstrap support for certain incorrect topologies (Soltis et al., 2004; Kubatko and Degnan, 2007). Other studies using real data have come to similar conclusions for relatively small numbers of concatenated genes (Baptiste et al., 2008). While such studies cast some doubt on our overall hypothesis, we suggest that the strikingly high support at very low levels of sampling for bootstrap, and high levels of exclusion for jackknife, coupled with the enormous amount of hidden support for the concatenated topology, still tend to support our overall hypothesis.

High levels of topological incongruence are not unusual in genomic datasets. For instance, Rokas et al. (2003) presented evidence from yeast genomes in which 45 of the 106 gene partitions conflicted with the concatenated hypothesis. In further examination of this data, (Gatesy and Baker (2005) showed that although there was a large degree of conflict between single gene topologies and the concatenated hypothesis, there was also a large degree of positive hidden support in the 45 conflicting partitions. In fact, the 45 conflicting partitions contributed significantly to the support to the concatenated hypothesis. They concluded that, “isolated analyses of individual data sets can mask congruence and distort interpretations of clade stability” (Gatesy and Baker, 2005).

3.7. How many genes are needed to get the concatenated hypothesis

We used the random addition of gene partitions as described in the Section 2 to address the question of how many partitions are needed to obtain a stable concatenated hypothesis. We sequentially added random gene partitions to a concatenated analysis and at each addition step tested the topological similarity (CFI) of the resulting tree to the overall tree from the concatenated analysis. We then repeated this process 100 times and averaged the CFI for each step. We did this experiment twice using two gene partition lists. First, we used all gene partitions with taxonomic representation of four or more taxa and second we used only those gene partitions with full taxonomic representation. Fig. 7 shows

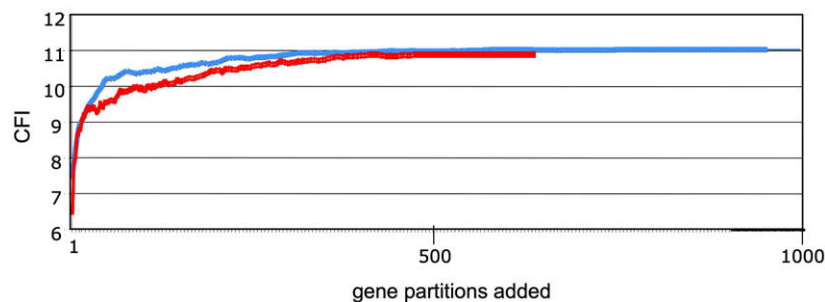


Fig. 7. Plots of CFI (y axis) versus number of partitions added during the random partition addition experiments described in the text. The top graph is for the larger data set with the 2093 genes that are present in more than four taxa. The bottom graph shows the results of random partition addition for the 633 partitions with full taxonomic representation. For the full data set we show our results only out to 1000 gene partitions added, but there is no change in the curve after the addition of about 500 gene partitions no matter what the order of addition of partitions.

the results of this analysis and indicates that while the approach to the concatenated hypothesis is relatively rapid (by 50 genes the average CFI is 10, meaning that only one node of the concatenated tree is not obtained), over 500 genes are needed for the concatenated hypothesis to give the same result with 100% confidence.

The analysis of the genes with full taxonomic representation suggests that slightly fewer taxonomically complete genes may be necessary to stably retrieve the concatenated hypothesis than for the entire gene set. For genes with full taxonomic representation 168 randomly chosen concatenated genes are required to get the overall tree 50% of the time, and 381 concatenated genes are required to get the overall tree in 95% of analyses. Whereas, for all genes with taxon completeness greater than 4170 concatenated genes are required to get the overall tree in 50% of analyses and 460 concatenated genes are required to get the overall tree in 95% of analyses.

4. Conclusion

A robust phylogeny of the family Pasteurellaceae can be thought of as having two major clades, the first containing the genera *Aggregatibacter* and *Pasteurella*, and the second containing the genus *Actinobacillus*. The genera *Mannheimia* and *Haemophilus* appear to be split between these two clades, and are therefore not monophyletic. The placement of *A. actinomycetemcomitans* and *A. aphrophilus* in a new genus (Norskov-Lauritsen and Kilian, 2006) called *Aggregatibacter* is robustly supported by the current analysis. Clearly, if the taxonomy of the group is to reflect phylogeny, major revisions of the group will be necessary.

The current study also shows extremely robust inference for all nodes in the concatenated hypothesis despite the fact that no single gene yields the concatenated tree. We present evidence here that this high level of support is due to strong, evenly distributed evidence from many individual genes that bolsters the overall hypothesis, but is overwhelmed when gene partitions are analyzed in isolation. This hidden support explains the phenomenon of many “wrongs making a right”. We suggest that this evidence reflects a strong historical signal and a single phylogeny that is obscured if genes are analyzed independently. We also show that this signal only reliably emerges when randomly adding large numbers of concatenated genes.

References

- Baptiste, E., Susko, E., Leigh, J., Ruiz-Trillo, I., Bucknam, J., Doolittle, W.F., 2008. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Mol. Biol. Evol.* 25, 83–91.
- Chiu, J.C., Lee, E.K., Egan, M.G., Sarkar, I.N., Coruzzi, G.M., DeSalle, R., 2006. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* 22, 699–707.
- Christensen, H., Bisgaard, M., 2008. Taxonomy and biodiversity of members of Pasteurellaceae. In: Kuhnert, P., Christensen, H. (Eds.), *Pasteurellaceae: Biology, Genomics and Molecular Aspects*. Caister Academic Press, Fredericksberg, Denmark, pp. 5–22.
- Christensen, H., Kuhnert, P., Busse, H.J., Frederiksen, W.C., Bisgaard, M., 2007. Proposed minimal standards for the description of genera, species and subspecies of the Pasteurellaceae. *Int. J. Syst. Evol. Microbiol.* 57, 166–178.
- Christensen, H., Kuhnert, P., Olsen, J.E., Bisgaard, M., 2004. Comparative phylogenies of the housekeeping genes *atpD*, *infB* and *rpoB* and the 16S rRNA gene within the Pasteurellaceae. *Int. J. Syst. Evol. Microbiol.* 54, 1601–1609.
- Colless, D.H., 1980. Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. *Syst. Biol.* 29, 288–299.
- Dewhurst, F.E., Paster, B.J., Olsen, I., Fraser, G.J., 1992. Phylogeny of 54 representative strains of species in the family Pasteurellaceae as determined by comparison of 16S rRNA sequences. *J. Bacteriol.* 174, 2002–2013.
- Dewhurst, F.E., Paster, B.J., Olsen, I., Fraser, G.J., 1993. Phylogeny of the Pasteurellaceae as determined by comparison of 16S ribosomal ribonucleic acid sequences. *Zentralbl. Bakteriol.* 279, 35–44.
- Di Bonaventura, M.P., DeSalle, R., Pop, M., Nagarajan, N., Figurski, D., Fine, D.H., Kaplan, J., Planet, P.J., 2009. Genome announcement: complete genome sequence of *Aggregatibacter (Haemophilus) aphrophilus* NJ8700. *J. Bacteriol.* 191, 4693–4694.
- Erwin, A.L., Smith, A.L., 2007. Nontypeable *Haemophilus influenzae*: understanding virulence and commensal behavior. *Trends Microbiol.* 15, 355–362.
- Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1994. Testing the significance of incongruence. *Cladistics* 10, 315–319.
- Gatesy, J., Baker, R.H., 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? *Syst. Biol.* 54, 483–492.
- Gatesy, J., O’Grady, P., Baker, R.A., 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics* 15, 271–313.
- Gioia, J., Qin, X., Jiang, H., Clinkenbeard, K., Lo, R., Liu, Y., Fox, G.E., Yerrapragada, S., McLeod, M.P., McNeill, T.Z., Hemphill, L., Sodergren, E., Wang, Q., Muzny, D.M., Homsí, F.J., Weinstock, G.M., Highlander, S.K., 2006. The genome sequence of *Mannheimia haemolytica* A1: insights into virulence, natural competence, and Pasteurellaceae phylogeny. *J. Bacteriol.* 188, 7257–7266.
- Huelsenbeck, J.P., Bull, J.J., 1996. A likelihood ratio test to detect conflicting phylogenetic signal. *Syst. Biol.* 45, 92–98.
- Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755.
- Kluge, A.G., 1997. Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics* 13, 81–96.
- Korczak, B., Christensen, H., Emler, S., Frey, J., Kuhnert, P., 2004. Phylogeny of the family Pasteurellaceae based on *rpoB* sequences. *Int. J. Syst. Evol. Microbiol.* 54, 1393–1399.
- Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56, 17–24.
- Norskov-Lauritsen, N., Kilian, M., 2006. Reclassification of *Actinobacillus actinomycetemcomitans*, *Haemophilus aphrophilus*, *Haemophilus paraphrophilus* and *Haemophilus segnis* as *Aggregatibacter actinomycetemcomitans* gen. nov., comb. nov., *Aggregatibacter aphrophilus* comb. nov. and *Aggregatibacter segnis* comb. nov., and emended description of *Aggregatibacter aphrophilus* to include V factor-dependent and V factor-independent isolates. *Int. J. Syst. Evol. Microbiol.* 56, 2135–2146.
- Planet, P.J., Sarkar, I.N., 2005. MILD: a tool for constructing and analyzing matrices of pairwise phylogenetic character incongruence tests. *Bioinformatics* 21, 4423–4424.
- Redfield, R.J., Findlay, W.A., Bosse, J., Kroll, J.S., Cameron, A.D., Nash, J.H., 2006. Evolution of competence and DNA uptake specificity in the Pasteurellaceae. *BMC Evol. Biol.* 6, 82.
- Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804.
- Sarkar, I.N., Egan, M.G., Coruzzi, G., Lee, E.K., Desalle, R., 2008. Automated simultaneous analysis phylogenetics (ASAP): an enabling tool for phylogenomics. *BMC Bioinformatics* 9, 103.
- Soltis, D.E., Albert, V.A., Savolainen, V., Hilu, K.W., Qiu, Y.-L., Chase, M.W., Farris, J.S., Stefanovic, S., Rice, D.W., Palmer, J.D., Soltis, P.S., 2004. Genome-scale data, angiosperm relationships, and “ending incongruence”: a cautionary tale in phylogenetics. *Trends Plant Sci.* 9, 477–483.
- Sorenson, M.D., Franzosa, E.A., 2007. TreeRot, version 3. Boston University, Boston, MA.
- Stamatakis, A., 2006. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690.
- Swofford, D.L., 1999. PAUP version 4.02b. Sinauer Associates, Inc. Publishers.