

**1. Typo in last line of first paragraph of abstract.**

Fixed

**2. Methods third paragraph, second line. Change "data is" to "data are."**

Fixed

**3. I think the paper might benefit by more comparison and contrast of results between the concatenation and concordance approaches here, but that is up to the authors of course.**

We have expanded somewhat our discussion of the results obtained by each method, but prefer to remain focused largely on the discussion of the challenges associated with the implementation of the two approaches.

**4. Given that straight taxonomic congruence is rarely used these days in phylogenomic analyses, maybe the authors could execute straight taxonomic congruence analyses.**

Assuming that by “straight taxonomic congruence,” the reviewer means a consensus approach, we have constructed a 50% majority-rule consensus of the 24 single-gene best ML trees. This tree has very little resolution. We have mentioned this additional analysis in the text and included the tree as Supplemental Datafile 7.

**5. justify the expected concordance priors used for the Bucky analyses?**

In the original submission, we used the program’s default prior of 1, which corresponds to a prior distribution on the number of distinct gene trees that is centered around 3.5 trees. We performed additional runs with an alpha of 10, for which the prior distribution is centered around 12.5 trees, with an alpha of 50, for which the prior distribution is centered around 18 trees and with an alpha of 100, which is centered around 22 trees. The trees produced using alpha=10, 50, and 100 were identical, both in topology and the values of the concordance factors at every node. Because we know that there is considerable uncertainty in the single-gene trees, we chose to use the primary concordance tree in the revised manuscript that was generated using a larger prior than the default used in the original submission. The RF distance between the alpha=1 and alpha=10 primary concordance trees was (72/4.76%) and the average concordance factor value was (0.360) in the alpha=1 tree, versus (0.368) in the alpha=10 tree. The main (and minor) difference between the two trees is the degree of resolution: the alpha=1 tree has 1508 splits and the alpha=10 tree has 1512 splits. For reference, a fully resolved tree has 1522 splits. The general conclusions of the paper were unchanged by choosing a different alpha prior, but we feel like the increase in resolution of the tree with the higher alpha is noteworthy, and we have included this in the results section.

**6. maybe a description of how ML bootstrap trees are substituted for the usual input for the Bucky analyses (Bayesian topologies)?**

I provided a perl script that convert RAxML output into BUCKy input. In Supplemental Datafile 7.

**7. Paragraph on Bayesian concatenation in Methods. In last sentence should "below 0.1" instead be "below 0.01"?**

Fixed

**8. First, when the authors discuss LBA, possibly more important is to discuss model misfit? If the ML model was a good fit, LBA would not occur, as I understand the concept, and the sometimes inexplicable results that the authors report would not have occurred. So the misplacement of long branches here is maybe not really classic LBA (long branches attracting each other), because ML when perfectly implemented should not have LBA.**

Although we do mention model misfit as a potential factor leading to inaccurate phylogenetic inference, and we feel it is extremely unlikely that we will ever be able to “perfectly” model molecular evolution over the course of billions of years, it is not true that ML analyses with perfect model specification are immune to the effects of LBA, as demonstrated with simulations by Kuck et al 2012.

**9. Second, I would strongly suggest that the authors execute simple parsimony analyses of their data (maybe this would not take too long to run in TNT?). Do parsimony analyses give much worse results than the ML or maybe better than ML? What if used Goloboff weighting that downweights sites according to the homoplasy at particular sites in the data (not an a priori picked model that is surely very wrong anyway)? I would be curious to see what sort of trees popped out of such analyses.**

We executed a parsimony analysis with PAUP\*, running 50 random replicates. While we might normally prefer to do more like 1000 random replicates, this analysis took several weeks. It produced 4 MP trees. A consensus of those 4 trees is almost fully resolved (1518 splits) and the RF distance between the best ML tree and the MP consensus tree is (370/24.37%). This distance is greater than the distance between the best ML tree and the BUCKy tree (318), but far less than the distance between the 16S tree and the best ML tree (~750). The distance between the MP consensus tree and the BUCKy tree is 364. For reasons discussed below, we opted not to include these results in the manuscript. However, we do provide all alignments should someone else wish to follow up this study with a more thorough comparison of phylogenetic optimality criteria.

**10. Last paragraph in "Bucky" discussion section notes that the concordance approach extracts hidden signal without resorting to concatenation ("resorting to" sounds like the concatenation process is painful?). However, here again, I think it would be productive for the authors to execute a simple taxonomic congruence tree,**

maybe a simple 50% majority rule consensus of their single gene bootstrap trees, or a 50% majority rule consensus of the strict consensus trees for the optimal trees for each gene. If these simple consensus procedures yield trees that are highly consistent with concatenation, it would show that it is not Bucky concordance that is extracting hidden support, because taxonomic congruence, which ignores hidden support, gives similar trees to concatenation? Another point is that even if Bucky yields a topology that is similar to concatenation, it is not clear that Bucky is extracting nearly as much hidden support as the concatenation approach. For example, Gatesy and Baker (2005) have shown that even by combining completely congruent genes, huge amounts of hidden support can emerge; even if Bucky gives a similar tree to concatenation, it might give a very weakly supported tree that did not successfully extract hidden support efficiently?

We've removed the phrase "resorting to," which was included, not because concatenation is a painful process, but because, as we mentioned, it forces the assumption that every gene shares an identical phylogenetic history. We prefer to relax that assumption by opting for a supertree method, like bucky, which accommodates uncertainty in the phylogenetic history of each gene. We agree that BUCKy may not be able to extract as much support as the concatenation approach, even if the topologies recovered by the two methods are similar. This is a component (mentioned now in the text) of why we prefer the concatenation approach.

**11. Placement of Interesting Taxa, first paragraph. The 50% threshold for increased support in concatenation relative to 16S seems too low to me, but that is just my opinion.**

This threshold was chosen (somewhat arbitrarily) because nodes that have greater than 50% bootstrap support are generally considered strongly-supported enough to be represented in a consensus tree. Because most relationships between bacterial phyla have historically been entirely unresolved, it was our thought that any move towards resolution was noteworthy, and we therefore used a fairly permissive threshold during this section of the paper.

**12. The most important concern I have is that I don't really know what the authors are trying to say. Are they saying that the supertree approach of using BUCKy on RAxML bootstrap trees is as good as RAxML on the combined dataset or that RAxML is better? Are they saying that they've learned something interesting about microbial phylogeny? Are they saying that MrBayes is infeasible for large-scale phylogeny estimation? I'm just not sure what their take home message is.**

We fully understand this reviewer's lack of clarity in this matter. We have modified the text in hopes of clarifying our goals. Basically, we were interested in generating a single, fully resolved phylogenetic tree of bacteria and archaea to be used in downstream comparative analyses. We wanted to use an approach that avoids some of the (unrealistic) assumptions about how genes evolve, especially in microbes. In particular, we wanted to use an approach that does not assume that every gene shares a phylogenetic history. A

review of available methods (as well as some false- starts with Bayesian approaches) led us to believe that the best approach, given our data, was to use BUCKy. We were interested in how a tree produced by BUCKy in this way compared to a ML analysis of a concatenated alignment.

From among the trees we generated during this process, we decided to choose the concatenated ML tree as our best representation of the relationships among bacterial genomes for 4 reasons: 1) it produced a fully resolved tree, which is essential for many of the downstream analyses that we wanted to do, 2) it provides an estimate of branch lengths, which is not only generally informative, but also essential for many downstream analyses, 3) it is accompanied by support values that are meaningful, if for no other reason, than the community at large is accustomed to an intuitive interpretation of bootstrap support values, where as the concordance factors produced by BUCKy are difficult to interpret, even according to the authors of the software, and 4) because it is much simpler and quicker to run than any of the other methods we tested here. We did also learn some interesting things about microbial phylogeny, as noted in the text, but that was not our primary goal. We also learned (the hard way) that MrBayes is certainly infeasible for phylogenetic inference, given data like those we present here.

**13. The authors conjecture that the reason the BUCKy analysis produced low support values is that the individual genes had low signal. However, I think there may be other reasons that the authors need to investigate. First, according to the authors, because BUCKy requires that every gene tree contain all the taxa, the authors added completely gapped sequences to each dataset before running RAXML to estimate the gene trees. This has the consequence that the added taxa are inserted randomly into the gene trees. It is not at all surprising that supertree analyses that are based upon gene trees with some completely randomly inserted taxa would be have low support. This makes all the analyses based upon BUCKy unreliable. Note, the problem caused by adding empty sequences to the individual gene sequence datasets does not impact the combined analysis step, so this is something that only impacts BUCKy. Note also that the authors could have avoided this problem by simply restricting the analysis to only those genes that truly did contain at least one copy of each taxon; this problem is caused by using the additional genes that were not universal.**

In large part because not every genome used in this analysis was completely finished, if we had restricted our analysis to genes that were truly universal, we would only have been left with 4 ribosomal proteins. Given the amount of total phylogenetic diversity among these organisms, a phylogeny reconstructed using only 4 ribosomal protein (i.e. short) genes would certainly be too poorly resolved for our purposes. We did attempt to minimize the impact of missing genes by limiting the number of missing sequences per gene to no more than six.

Therefore, we were left with two choices: 1) concatenation or 2) seeing what happens when a supertree approach is attempted. No one knows how reliable are analyses based on BUCKy using gene trees with randomly placed taxa. This sounds like an important

area of future research. One might allow the reliability of the approach be judged by its congruence with an alternative, well-accepted approach, and in this case, it seems to have performed quite well, despite the inclusion of missing genes.

Finally, because BUCKy is a relatively new method, and was not designed for such large-scale analyses, it may not be surprising at all that concordance factors are low. In fact, it is known (Cecile Ane, pers. comm.) that BUCKy will underestimate concordance factors when there are large numbers of taxa and poorly resolved gene trees, but that the inference of topology is robust to these conditions. We might have spent more time to better understand the effect of missing genes on BUCKy, but because of the additional considerations listed above, we decided to adopt the concatenation approach as the preferred method to generate our reference phylogeny for future comparative analyses.

**14. If the authors added the empty sequences to the sequence alignments given to MrBayes, then it is also not surprising that MrBayes would fail to converge in a reasonable timeframe. This means that conclusions about MrBayes not converging might need to be revisited.**

It is certainly the case that having empty sequences will negatively impact the time to convergence for MrBayes. However, the MrBayes run using the concatenated alignment also failed to converge in a reasonable period of time. Previous analyses (Warnow et al., also found that MrBayes did not converge using a dataset similar to the one we present here, so this is not a new finding.

**15. the use of only 100 bootstrap replicates for a dataset of this size is questionable. It is possible, therefore, that BUCKy would produce a different species tree (with higher support values) if the authors provided it with substantially more than 100 trees for each gene, each produced on a different bootstrap replicate.**

We have now done 1000 bootstrap replicates. We attempted to run BUCKy using the 1000 replicates, but even with the memory/time-saving hack to disable computation of the population tree, BUCKy still crashed. Fortunately, as recommended by this same reviewer (see below) all of our single-gene phylogenies converged after no more than 500 bootstrap replicates. BUCKy was able to run with 500 bootstrap replicates per gene, so it is these results that we present in this revised manuscript.

**16. Also, it is essential to report the statistics that let the user know whether the rapid bootstrapping technique in RAxML has converged**

We thank the reviewer for pointing out the necessity of bootstrap convergence estimation, as we were unaware of this option with the RAxML bootstrapping algorithm. As I noted before, for each of the single-gene bootstrapping analyses, we ran 1000 replicates, and each of them converged after no more than 500 replicates (we also added this data to Table 1). The bootstrapping of the concatenated alignment converged after 250 replicates, and this took approximately 45 days to complete. All of these results have been included in the manuscript.

**17. The running time to compute BUCKy is extremely fast, and I find it difficult to reconcile the reported running time for a much smaller dataset as reported in Yang and Warnow 2010 (the paper referenced in this study as recommending RAxML bootstrapping instead of MrBayes) and the running time reported in this paper. I think that the reason they were able to get the BUCKy analyses to complete quickly is that they disabled the population tree estimation, which takes  $\Omega(n^4)$  time; if that is the case, the authors should point this out, and also make the code available.)**

We did, in fact, both disable calculation of the population tree and point it out in the original submission. We have now also included the specific line of code that was changed in hopes of making this alteration more obvious to the reader and easier to replicate.

**18. I could not find the datasets and trees in TreeBASE. Nothing turned up when searching under Lang or Eisen as authors that mentioned this paper. The authors should give links to these datasets, if possible.**

These have now been uploaded to figshare, with the revised versions of all trees, figures, and original alignment files. The text of the manuscript has been altered to reflect this change in data deposition.

**19. The text in the conclusions section in the paper do not justify the recommendation given in the abstract that combined analysis, using RAxML, is preferred to BUCKy. However, it seems that the authors may be using the fact that the RAXML with rapid bootstrapping tree had higher support values than those produced by BUCKy, in order to justify this recommendation. Given that high bootstrap support can exist in combined analyses (as noted by the authors) and for the wrong tree, this doesn't seem to be a good reason. Furthermore, note the earlier comments about support values.**

We have changed the text to make it more obvious that our desire was to explore the use of several methods for large-scale phylogenetic reconstruction with this type of data, with the ultimate goal of producing a single reference species phylogeny that is suitable for use in phylogenetic comparative analyses. We set out to use the only supertree method (BUCKy) which is agnostic to the source of discordance among gene trees, incorporates uncertainty in the topology of the single-gene trees, and is computationally feasible given our data. We wished to compare the results obtained by BUCKy to a supermatrix (concatenated) approach. The most logical comparison to make is between the tree obtained by BUCKy, given Bayesian input phylogenetic data, versus the Bayesian analysis of a concatenated alignment. Given that the Bayesian approach was infeasible, we opted to compare the trees obtained by BUCKy with ML input versus an ML analysis of the concatenated alignment. Because this is the comparison we wish to make, and these analyses alone took many months of computational time to complete, we did not include a Maximum Parsimony tree for an additional comparison.

The particular phylogenetic comparative method (Independent Contrasts) that we seek to employ in future analyses depends upon the input of a single, fully resolved phylogenetic tree including estimated branch lengths. The concatenated ML analysis provided such a tree.

**20. The evidence that MrBayes does not converge on the 841 taxon single gene datasets, even given 9 months of analysis, is very interesting - and, if valid, really significant for researchers. Also, as discussed above, if the MrBayes analyses of individual gene sequence alignments were based upon datasets that contained completely gapped sequences, then the failure to converge is not at all surprising. (Finally, it seems the same observation about MrBayes failing to converge was made in Yang and Warnow, and so a general observation that may be important to communicate.)**

Because we limited the number of completely gapped sequences to no more than 6 per gene, we would expect convergence, given enough time. However, we encountered the same issue with the concatenated analysis, which contained no completely gapped sequences. We agree that this may be a general observation, and have made note of this in the text as suggested.

**21. The authors fail to mention any standard supertree methods, which is strange. Why isn't MRP mentioned, for example? For the special case handled in this paper (where all gene trees have one copy of each species), the use of consensus methods (like the majority consensus) can also be considered. Finally, there are many supertree methods that do take incomplete lineage sorting into consideration, but not mentioned here, and that have excellent performance (\*BEAST, for example); at a minimum, the authors should discuss the others.**

We did mention two of these methods in the introduction, STEM and BEST, but we have now included \*BEAST (which is not computationally feasible with a dataset this size) and MRP (which does not incorporate uncertainty in gene trees). We did compute a majority-rule consensus tree and add it to the supplemental data (see comment #4)

**22. The authors say that the gene trees had relatively low signal, pointing to the short sequence lengths. However, other factors could be involved -- including the fact that the sequences were AA instead of nucleotides. If the nucleotides for the sequences are available, phylogenies based upon nucleotides could be more informative. Also, to strengthen the evidence that their AA sequence alignments contain low signal, the authors could provide more statistics about their alignments, such as the average percent identity, the minimum percent identity, and the number of gaps in the alignments.**

The use of AA sequence data is standard for microbial phylogenetic analyses of this scale because for many taxa, their last common ancestor existed billions of years ago. During this time, the phylogenetic signal at the DNA sequence level has been largely lost. It is also worth mentioning that the tips in our trees tend to be well-resolved, while the deeper

splits remain uncertain, and these would surely be saturated with nucleotide data. Ideally, DNA sequence data could be used with a 60-state codon model, but that is currently computationally infeasible with a dataset of this size.

We have, however, provided the average percent identity for each alignment, and included a discussion of it in the text, and representation of it in Figure 7. There is no correlation between the average percent identity and alignment length, providing additional evidence that shorter gene length and not increase in sequence divergence is responsible for the variance in RF distances among bootstraps for the 24 genes.

23. The use of the Robinson-Foulds distance seems potentially problematic, since at least one of their trees is not fully resolved. In any event, the values should be presented proportionally, to help readers understand the trends. If the trees have 841 leaves, then the RF distances will range from 0 to 1676. Therefore, values can be expressed as a percentage, with 200 RF distance equivalent to only 12% in RF error rate. If the trees they show are only on Bacteria, then the number of taxa will go down; all the more reason to specify, in each case, the RF rate rather than the RF distance (the actual number of unique splits).

All tree comparisons were performed using trees with Bacteria only. This was perhaps not stated clearly enough in the text, so we added the italicized phrase below to the text: we will use Bacteria-only trees for all further analyses and for all tree comparisons.

**24. Page 6, line 57 ``causing those taxa to be placed randomly on the trees.'' Are you sure that taxa consisting entirely of missing data will be placed randomly? There are cases in which random sequences are not equally likely to be placed in any position on a tree (Susko et al 2005, J Mol Evol 61:351-359). I think that including the taxa with entirely missing data is appropriate, but suggest deleting ``causing those taxa to be placed randomly on the trees.'' There's no need to go into more detail or to cite the reference.**

We did confirm with the author of RAxML that it would indeed place those taxa randomly on the trees. However, the phrase was removed as per your suggestion and thank you for the reference.

**25. Throughout, the metric should be ``Robinson-Foulds'', not ``Robinson-Fould's'' (and the second-author's name is missing from reference 50).**

Fixed.

**26. Page 10, line 34 and Figure 7 legend ``significant positive correlation'' should read ``significant negative correlation''.**

Fixed.



**27. Page 10, line 50 ``also, the BUCKy tree does not include branch lengths'', but the branches on Figure 5 are not all the same length. Add a brief explanation in the figure legend.**

Done. The representation is a radial cladogram, which is drawn so that all of the taxa are the same distance from the center of the tree. In order to achieve this, some branches must be elongated.

**28. Italicize species names in reference list.**

Fixed

**29. Figure 7 could be improved by redrawing as follows. Label both axes. Delete the horizontal grid lines. Give the  $R^2$  to only a couple of decimal places. Consider whether the fitted line is needed, and if so, explain what it is in the figure legend (a least-squares regression line?). If the figure is intended only to show the association between RF distance and alignment length, rather than to predict one from the other, then the line is not needed. Finally, there are two influential points (the two longest alignments). If those two were absent, the relationship would be rather different. Is this because the relationship is nonlinear over a wide enough range of alignment length, or because there's something else unusual about the two longest alignments? It might be worth commenting briefly in the results section.**

We modified Figure 7 as suggested, by labeling axes, removing grid lines, and truncating the R value. We also added average percent identity for each alignment, and chose to keep the fitted lines to make more clear the difference in the relationships between RF vs. alignment length and RF vs. average percent identity. While removing the two longest alignments does not change the quality of the relationships depicted in Figure 7, we do mention now in the results that those are the only two protein-coding genes in our dataset.