Decision on USYB-2014-198, Bayesian Analysis of Partitioned Data:
Accept pending major revisions.

Dear Dr. Moore,

Thank you for your *Systematic Biology* submission. It has been reviewed by Associate Editor Dr. Mark Holder, Rob Lanfear and an anonymous reviewer. Their comments are listed at the end of this letter. The reviewers and the AE provide some excellent constructive suggestions that I am sure you will appreciate.

As Mark notes, the main sticking point was raised by Rob Lanfear—the Wu et al. method implemented in `BEAST2`. I agree that it doesn't feel good to shift the goalposts on you here, but I think that alternative method has to be discussed. I don't know if you need to go so far as to reframe your paper as a comparison between your method and the Wu et al. method, but I definitely agree that a comparison of your method and the Wu et al. method on the three empirical data sets is required. Fortunately, I think that it could be done without imposing an immense computational burden. If `BEAST2` doesn't converge on one or more of the empirical data sets, you have an obvious selling point for your method.

Other than that, I have very little to add to the comments of the reviewers and AE. I agree that this work is sure to be widely cited, and I look forward to trying it out with my own data sets.

Other points

1. I found this a bit confusing. On pg. 4, you write that the DPP model allows you to specify a non-zero prior probability on all possible partitioning schemes (from a uniform model to a saturated model, in which each element gets its own process partition). Great! But on pg. 7, you note that in this study, you invoke an independent DPP model for each parameter to describe the process heterogeneity among the PRE-SPECIFIED data partitions. Hmm. That's still cool, but less exciting, because now only the data partitions I can dream up will be evaluated, rather than all possible schemes. The next section, though, notes that a biologist could simply (at the extreme) define each site to its own subset. OK... is there a reason why I shouldn't always just set K to be equal to the number of sites in my alignment, and let `AutoParts` sort them into process partitions? Can that be done? If so, I presume that it would be computationally challenging, but it isn't clear. Also, it sounds like that is what the Wu et al. method implemented in `BEAST2` does. Lanfear suggests that the `BEAST2` method may have trouble with convergence, though neither he nor I have tried it. Apologies for the confusion. Our method allows users to flexibly specify the 'data elements' (or 'data subsets') that are assigned to process partitions that capture variation in the substitution-model parameters (exchangeability rates, stationary frequencies, alpha-shape parameter, and tree length) under the Dirichlet process prior model.

   At one extreme, the user could define $K = L$ pre-specified data subsets for an alignment of L sites. This decision reflects the prior beleief that the evolutionary process is likely to vary between every site in the alignment to a degree that will impact estimation of the focal parameters (typically the tree and branch lengths). However, this perspective—which is necessarily adopted when using the Wu *et al.* method—strikes us as somewhat extreme and is at odds with a lot of relevant prior knowledge about molecular evolution.

   Alternatively, the user could leverage prior knowledge on likely patterns of process heterogeneity across our multi-locus sequence alignment *if* we are willing to concede that the process may vary *slightly* within some of those pre-specified data subsets of sites—for example, we might define a data subset that groups all second-position sites of a protein-coding gene together, even though selection may have acted to alter the substitution process on a *few* of those sites. This would be defensible if minor wobbles in the evolutionary process within pre-specified data subsets are inconsequential to phylogeny estimation.

   Leveraging prior knowledge regarding the patterns of process heterogeneity is in the spirit of Bayesian inference and allows our approach to be applied to modern, multiple-sequence ('phylogenomic') datasets comprising dozens or even hundreds of gene regions. By contrast, treating each site as a source of potentially confounding process heterogeneity restricts the practical application of this method to small, single-gene datasets. Wu *et al.*, for example, were only able to apply their method to very small, single-gene datasets (comprising ~ 500 sites), and even then were forced to take extraordinary measures in order for their method to work. Specifically, prior to analyzing these small alignments with their method, it was first necessary to generate and specify empirical (hyper)priors for all of the substitution model parameters in order for the MCMC to work. This entailed performing full MCMC simulations on a large sample of 'similar' empirical datasets ($N = 25$) to estimate the marginal posterior probability densities for the relevant substitution model parameters, and then fitting the (hyper)priors of their model to the corresponding empirical marginal posterior probability densities. I think the use of empirical hyperpriors is actually a very interesting approach, but the need to perform a large number of analyses on empirical 'training' datasets as a precursor to making inferences from a given study dataset certainly represents a substantial computational investment. Besides the computational burden of these precursor analyses, this approach also introduces a number of potentially confounding issues regarding: (1) what constitutes an adequate sample of training datasets; (2) what constitutes a sufficiently 'similar' dataset (in terms of the genomic and taxonomic sampling); (3) how these training datasets should be analyzed (the choice of substitution model(s), etc.), and; (4) how empirical (hyper)priors should be correctly elicited from the corresponding marginal posterior probability densities estimated from the training datasets. Our approach can be applied to much larger dataset and avoids both these computational limitations and methodological considerations, albeit at the expense of requiring the user to make some prior assumptions about likely patterns of process heterogeneity across the multi-gene alignment.

In our revision, we have clarified this issue. Specifically, we have made an effort to address the two main causes of the previous confusion on this point: (1) our somewhat confusing terminology (referring to 'data subsets' as 'data partitions', and 'partition schemes' as 'process partitions'), which we have now clarified, and; (2) our failure to provide an explicit comparison of our approach and the seemingly similar (but actually very different) SubstBMA approach implemented in `BEAST`, which we have now added to the discussion.

2. Reviewer 2 points out some ways in which the manuscript could be made even more accessible to systematists. I agree with her/him, though I think it already does an admirable job at that despite the complex subject matter. Frankly, I don't know what Bell or Stirling numbers are, either, and that was a bit jarring given the clarity of the rest of the paper. On the other hand, Mark suggests that the manuscript could be shortened a bit, given that anyone reading the paper probably already has a pretty good clue about Bayesian phylogenetics. I think that text that isn't entirely relevant to the topic (Mark mentions the section on summarizing tree posteriors) could be eliminated, but overall, I like the teaching style. I prefer to err on the side of being transparent and a bit longer over being brief but potentially opaque.
   We agree: sacrificing extreme brevity for the sake of clarity is a good trade off. We have followed suggestions to improve the clarity of our presentation, while eliminating anything that is not directly relevant to understanding the method.

3. Pg. 9: "singe parameter" should be "single parameter".
   Done.

4. Pg. 11: No need to hyphenate "one at a time".
   Done.

5. Pg. 12: "By contrast, other parameters—such as the tree topology, T—presents some difficulties in a MCMC analysis, as it is an unusual model parameter"—This is awkward, because the sentence starts talking about "other parameters", but the end of the sentence ("it is an unusual model parameter") is referring to just one parameter, the tree topology.
   We agree that this was awkwardly written. We have attempted to improve the clarity using the following revised text: "By contrast, it is far more difficult to summarize complex, discrete parameters. For instance, there is no natural way to summarize an MCMC sample of tree topologies, except perhaps to simply report the posterior probabilities of individual trees."

6. Pg. 14: “that centered prior mean” should be “that centered the prior mean”.
Done.

7. Pg. 21: “Specifically, the mean partition scheme inferred for skinks ($K = 11$) and hummingbirds ($K = 9$) were largely stable over the entire” should be “schemes inferred for skinks”.
Done.

8. Pg. 23: “to select mixed model that provides” should be “to select the mixed model that provides”.
Done.

9. Pg. 26: “under the any strict-/relaxed clock model”—“the any” needs repair.
Done. Changed text to “under any strict-/relaxed clock model.”

10. I’m also not fond of the Chinese restaurant analogy. Not only is it overdone, I’ve never found it to be particularly illuminating. But like Mark, I’m willing to let it go.
I agree. Because of its familiarity, it is tempting to draw on the Chinese restaurant metaphor to explain the DPP model. Nevertheless, I’ve found that it’s easy for people to get lost in the metaphor. I’ve taken your comments as an opportunity to develop a concise (non-metaphorical) description of the DPP for the specific case at hand, including a figure (requested by Reviewer 2) that (I hope) will help readers better grasp the important terms and concepts. [BRM: this comment still needs to be addressed.]

Sincerely,

Prof. Frank (Andy) Anderson
Editor-in-Chief, Systematic Biology

AE (Mark Holder) Comments:

Dear Andy,
I am writing to recommend that manuscript USYB-2041-198 “Bayesian Analysis of Partitioned Data” by Drs. Moore, McGuire, Ronquist and Huelsenbeck be accepted pending some revisions. In the Manuscript Central checkboxes, I clicked “major revisions” (because I think some new analyses are required). But I am optimistic that they will be reasonably straightforward—and will not require another round review by referees.

First, I must apologize to the authors, reviewers, and you (basically to “all concerned”). The review process was delayed some by the failure of the manuscript to contain a URL for the source code. After email exchanges with Dr. Moore, I did get that URL (it is https://github.com/brianrmoore/AutoParts/) Unfortunately, my attempt to get the URL to the reviewers via Manuscript Central failed. I did not realize that until I read the reviews (one of which complained about the lack of a link to the source code). I should have (a) not messed up in getting that info out, and (b) double checked that it had gone out so that we could have avoided confusion on that point.
No apology necessary: it was my fault for not including the url for `AutoParts` in the manuscript. It has now been added, along with the Dryad accession number for archived simulated data, etc.

The manuscript has been reviewed by Dr. Rob Lanfear and by another very well-qualified reviewer. They and I are enthusiastic about the method described here. This is really a nice study and method. I am confident that it will be very well-cited and appreciated contribution to *Systematic Biology*. Both reviewers also provide several very helpful suggestions that will improve the presentation of the revised manuscript.

The one fly in the ointment here (and the reason that I think that revisions are necessary) is mentioned by Rob’s review: Wu, Suchard, and Drummond have published a very similar approach in MBE in March of 2013. It simply would not be appropriate scholarly publication for *Systematic Biology* to publish this method without some comparison to that method. I am not suggesting that the authors compare the methods implemented in BEAST2 on all of their simulations. That contrast would be very interesting, but would clearly be a huge endeavor. But I think that the authors need to at least attempt to apply the Wu et al method to the 3 empirical data sets. If a good faith effort to achieve convergence with the BEAST2 method fails, then that could be noted as an advantage of the authors AutoParts-based implementation.

If the BEAST2 implementation does converge then comparison of the results will clearly be of interest.

The current method allows the user to group sites into blocks that will always be placed in the same process partition. The BEAST2 implementation does not. That could be very important difference that renders the current method feasible to apply to many more problems (because it will presumably mix much better). Wu, Suchard, and Drummond note the binning of sites into groups a priori as a possible "future work" to be done in their paper.

I'm sure that that recommendation is frustrating to the authors. Their method was clearly an improvement on the state of the art when it was originally submitted—and it probably still is. I'm sure that it is frustrating to carefully respond to review comments only to be told that the manuscript has to compete against a new baseline.

Nevertheless, the most direct context for this contribution is now as an alternative to the Wu, Suchard, and Drummond method. Informed readers deserve some discussion of how the method compares to that work. We agree completely with this suggestion. We have added a discussion of the similarities and differences between our approach and that described in the Wu *et al.* paper. We note that—although these two approaches were developed independently—our method and implementation is considerably more general (at least in some aspects) than that described by Wu *et al.* In several important respects, the model described by Wu *et al.* is a special case of the DPP approach that we describe in this paper and implement in AutoParts. Specifically, the substitution Bayesian model averaging ('substBMA') approach implemented in BEAST2 treats individual sites as the data elements (but does not allow sites to be grouped into data subsets/data partitions), and applies a DPP to the exchangeability rates, stationary frequencies, and tree-length parameters only. Because it does not allow grouping of sites into data subsets, it does not accommodate among-site rate variation (using the discrete gamma model, as used in AutoParts). Importantly, under the substBMA model (as in PartitionFinder), two or more sites either share the same vector of substitution-model parameter values (*i.e.*, are assigned identical exchangeability rates, stationary frequencies, and tree-length), or else they share no substitution-model parameter values. In contrast to this all-or-none approach, the DPP implemented in AutoParts allows two or more subsets of sites to share all, some, or none of the substitution-model parameter values. For example; our DPP approach allows two data subsets to share the same exchangeability rates and tree length, but to be assigned distinct stationary frequencies (which they each may or may not share with other data subsets. AutoParts allows the partial sharing of the substitution-model parameters—the exchangeability rates, stationary frequencies, alpha-shape parameter, and tree-length—to be partially shared among data subsets in all possible combinations. Our experience applying this approach to real datasets suggests that, in fact, partial sharing of substitution-model parameters among data subsets is an extremely pervasive feature of empirical datasets (*e.g.*, the pean-partition scheme and credible set of partition schemes most often exhibit partial parameter sharing). To our knowledge, this feature is unique to our approach.

Moreover, the BEAST2 implementation of the substBMA model also does not allow the concentration parameter to be treated as a random variable. This is a practically important difference, as it requires that users perform a series of independent MCMC analyses using incremental values for the (fixed) concentration parameter. By contrast, the concentration parameter can be treated as a random variable that can be integrated out in a single analysis using AutoParts. Finally, Wu *et al.* report that reliable MCMC simulations required the use of empirical hyperpriors for all of the substitution-model parameters for the small single-gene datasets (comprising less than 500 sites) analyzed in their study.

We have also attempted to use the substBMA approach on the three empirical datasets in our study using the implementation BEAST2. In these analyses, we used the same analytical protocol as that used for our AutoParts analyses—comparable priors for substitution-model parameters, a comparable set of concentration-parameter values (spanning the range from low to high), the same chain length and thinning, and the same number of replicates. Unfortunately, despite our persistent efforts, we failed to achieve successful MCMC perfomance for any of these analyses.

From their description and our analyses, it seems clear that the Wu *et al.* method is intended for small, single-gene datasets, whereas the DPP approach we have developed and implemented in `AutoParts` is more suitable for estimation of phylogeny from multi-locus sequence alignments. In fact, as we point out, the two methods could possibly be used in a complementary way: the substBMA approach could be applied to individual gene regions to identify the pre-defined data subsets, anf the resulting data subsets for the collection of gene regions could then be subjected to analyses using `AutoParts`. From our experiments, however, it seems unlikely that even the analyses of small, single gene regions using substBMA is possible unless empirical (hyper)priors are specified for the model.

A couple of minor points:

1. On page 22 (Manuscript Central numbering): I think the equating of "unimodal" with "well mixed" or "successfully capturing processing heterogeneity" is dangerous. Certainly poorly mixed chains can display unimodal parameter posterior distributions. The posterior density for a parameter can be multi- modal even if the chain has converged, and even if the process does not contain unrecognized partitions. We have no guarantees (that I am aware of) about the number of modes in a likelihood surface for a phylogenetic model. I agree with the notion that failure to separate sites subject to different processes is probably the most common reason for multi-modal posteriors in most real analyses. I am just afraid that currently this section is written in a loose manner, and I can easily imagine many practicing systematists pointing to this passage as a justification for eyeballing histograms of MCMC output and concluding that their chains have run long enough.
   Done. These are great points. The passage referred to is informed by results of our more comprehensive simulation study, where we explored the impact of mis-speciifed data subsets; *i.e.*, where we specified data subsets that contained sets of sites simulated under two or more distinct processes. In these cases, the marginal posterior probability densities for the corresponding substitution-model parameter(s) were distinctly multi-modal, where the number of modes typically corresponded to the number of distinct processes used to generate the specified data subset (*i.e.*, the residual process heterogeneity within the specified data subset). Moreover, analyses of datasets where the data subsets correctly captured the (known) process heterogeneity had unimodal marginal posterior probability densities, *except* in cases where the MCMC clearly failed (as indicated by other diagnostics, such as the ESS, PSRF, etc.). Nevertheless, Mark is absolutely correct that we have not presented the relevant evidence to make this claim in this paper, so it has been removed.

2. I'd be carefull with "linearly" on page 23—you don't really have enough data to make such a specific statement about the response to changing the prior.
   Right! Replaced "scaled linearly' descriptor with the correct descriptor "increased".

On to some optional (take-them-or-leave-them) suggestions:

1. The mansucript would probably read better if the text were written as an explanation of the difference from the Wu et al method, rather than a general discussion on the topic of sampling over partitioning uncertainty. This would mainly affect the intro.
   Because of the important conceptual and practical differences between the our approach and that described by Wu *et al.* (and because we also wish to compare our approach to `PartitionFinder`), we have decided that the best solution is to address the Wu *et al.* paper in the Discussion section. Specifically, we have extended the 'conceptual-comparison' subsection of the Discussion (*i.e.*, comparing mixed, finite-mixture, and infinite-mixture approaches for accommodating process heterogeneity) and added a new 'practical-comparison' subsection that considers the application of these approaches.

2. The text is very clear and well-written for a teaching style. But I think that it could be considerable shorter if the text was a little more in the style of "here is what we did (assuming pretty good understanding of Bayesian phylogenetics)" vs the current style. I think that the authors should be given leeway to select their own style, and reviewer #2 clearly liked the writing style and explanations. I just wanted to comment on my impression of the writing.
   I agree completely that the paper could be made considerably more concise if we were to adopt this suggestion, and—if the intended audience was fellow theorists/methods developers—I agree that this would be the best approach. However, the paper is intended for well-informed empirical biologisits who will (hopefully) apply this method to real data. Having organized and/or participated in many workshops/courses for graduate students in our field, I am increasingly concerned about the vast and expanding chasm between people in our field who develop methods and those who use them. I view this as a failure on the part of methods developers to communicate clearly to their audience. The resulting disconnect has very real consequences for our field. People will not (correctly) use the very best methods if they do not understand them. So, we have decided to write the paper in a way that is (hopefully) completely accessible to sophisticated users of statistical phylogenetic methods who are aware of the need to partition datasets, but may have never heard of infinite mixture models. Hopefully, this style will not be too grating to our fellow theoretical colleagues who comprise a relatively small constituency of our community.

3. I'm one of the reviewers that reviewer #2 warned about: I dont like the Chinese restaurant analogy. I have never liked it, and it has already been used multiple times (and in *Systematic Biology*) to explain the Dirichlet process. Once again, this is stylistic and I don't want to act like an author on the paper . . . some one please promise me that there will come a day in phylogenetics when we don't use this example every time we apply the DPP.
   I actually really like the Chinese restaurant analogy—someone who overheard John and I while working on this project might have wondered if we were maitre D's at chez DPP! Nevertheless, I agree that it is easy to get lost in the metaphor, and (even if the readers follow the metaphor correctly) it leaves a conceptual gap between the general DPP model and the specific model that we develop in this paper (*i.e.*, given that there are independent DPPs on each of the four substitution model parameters, it is really more like a Chinese food court! Accordingly, I have eliminated the Chinese restaurant analogy and replaced it with a (non-metaphorical) explanation that will (hopefully) better help explain our approach using the DPP to accommodate process heterogeneity, and allow readers to compare our approach to alternative approaches, such as the substBMA model.
   [BRM: this comment still needs to be addressed.]

4. The (admittedly brief) discussion of summarizing tree posteriors is one example of several where the manuscript could be shortened. The statements made are correct and clearly written—just not that relelvant to this ms.
   As above, we are trying to strike a balance between 'clear' and 'concise'. The issue of summarizing samples of complex, discrete parameters is non-trivial, so we believe that it is worthwhile to introduce this concept with an example (tree topologies) that will be familiar to our intended audience.
   [BRM: I think we could actually remove this section.]

I'm sorry to recommend "bad news", but I do think that the overall message is that this paper could ultimately be a very valuable contribution to *Systematic Biology*.

Sincerely,

Mark Holder

Reviewer 1 (Rob Lanfear) Comments:

Recommendation: Accept with major revisions

Comments: I thoroughly enjoyed reading this paper.

The authors present a method of dealing variation in rates and patterns of molecular evolution that is clearly superior to anything currently available. The paper is thorough, clear, and concise.

There are three things I would like to suggest.

(All line numbers and page numbers are from syst biol's version, which I have just noticed differ from the authors own page numbers, sorry).

1. PartitionFinder

I will come straight out and say that I am about to mention my own work, so I can hardly claim to be objective. And since mentioning it here involves the cardinal sin of a reviewer asking an author to consider citing the reviewer's work, I certainly won't be offended if the authors choose to ignore this comment in its entirety. The ms is comprehensive without mentioning PartitionFinder, and I can see also that including a discussion of non-Bayesian approaches may complicate the ms. However, there were a couple of places where when reading the ms I felt that PartitionFinder is widely enough used that mentioning it may help to present a balanced view of the status quo:

(i) in the introduction, third paragraph, where the authors describe what is usually done about the partitioning problem.

(ii) in the discussion, paragraph starting on line 26 of page 26. To a limited extent PartitionFinder solves the same problem of discovering new schemes, and searching the space of all possible schemes (albeit not in a Bayesian framework) by heuristically searching among schemes (with the notable limitation that it makes the assumption that all parameters share the same scheme).

In case the authors agree, the paper describing the approach is here:
http://mbe.oxfordjournals.org/content/29/6/1695
Done. We agree completely that it is entirely appropriate to cite PartitionFinder, and have added citations where suggested by the reviewer. We have also included PartitionFinder in the Discusion section, describing both the conceptual relationship and providing a practical comparison of PartitionFinder to our approach.

2. BEAST

This omission seems more important to address. The authors make no mention of a very similar method published in 2013, implemented in BEAST2:

http://mbe.oxfordjournals.org/content/30/3/669.long

That method also uses a Dirichlet process prior to solve precisely the same problem. As far as I know, it does not include user information on groups of sites, but just starts by treating each site as a potentially independent partition. The BEAST2 method (if I have understood it correctly) is a little more limited than the current method in that it only allows the separation of rates vs other parameters in terms of allowing parameters to have different process partitions. Nevertheless, both methods solve the same problem in very similar ways, and so it seems important to make a full comparison. Specifically, it would be nice to know:

1. What is different in the implementation presented in the current ms

2. Some brief details of any differences or similarities in performance It seems important to integrate this in various points throughout the introduction and discussion.

Done. We agree completely that it is important to discuss the approach described by Wu *et al.* As noted above, we have included (in the Discussion section) a brief description of the Wu *et al.* method, highlighting similarities and differences between their approach and our own. We also discuss some practically important differences in the implementation of these two approaches (in BEAST2 and AutoParts, respectively), which are demonstrated via re-analyses of the three empirical datasets.

3. Dataset size limitations

One thing I wondered as I read through the ms is whether the authors could provide some information on the scaling of the size of the dataset with the analysis time, on whatever computational infrastructure they use. I ask because I am aware (through the grapevine, not personal experience) that the BEAST2 method can take an extremely long time to mix properly on even relatively small datasets. I expect the implementation here mixes faster, since the scope of the problem will usually be much smaller as long as the user defines groups of sites larger than one. (I expect there are many other details of the implementation that will affect this too, about which I profess to know almost nothing).
Done. In demonstrating the application of the two approaches to the three empirical datasets, we note the run times for the AutoParts analyses, and discuss issues associated with MCMC performance of BEAST2.

Some indication of the largest datasets that may be practically analysed with this method may be useful for readers of Syst Biol, since the method presented here is obviously far better than PartitionFinder and as far as I can tell also improves on the BEAST2 method (since it is flexible in the user's definitions of initial groups of sites). Thus, it seems very clear from this ms that it presents the go-to method for estimating trees with partitioned data, as long as datasets can be analysed. If datasets of any size can be analysed, that's fantastic. But it would be worth pointing out either way.
Done. We provide run times for the AutoParts analyses of the three empirical datasets; however, it is not possible to make direct comparisons to run times under the substBMA model in BEAST2), as those analyses did not converge. We prefer to defer the discussion of the details of the implementation and more practical issues of the method for a separate paper—which will be submitted upon acceptance of these revisions—that is an 'application note' for the AutoParts program. It seems cleaner to focus the current paper on the method itself (and underlying theory), with analysis of simulated data (for validation) and empirical datasets (for the purpose of demonstration).

And a couple of very minor things:

1. Page 4 line 34: "Vendetti" should be "Venditti"
   Done.

2. At the end of the ms, the authors mention extending the method to include different tree topologies. I wonder what the authors think about an intermediate method in which all partitions share the same topology but can share (or not) sets of branch lengths? This would be neat insofar as would allow for heteroscedasticity, and would even allow users to estimate the extent of it in any given dataset.
   Nice! This is another excellent insight. Relaxing the assumption of shared branch-length proportions (while maintaining a shared topology) provides a method that accommodates 'heterotachy' under a DPP. We have actually implemented the heterotachy model, but it is not (yet) included in AutoParts or described in the current paper.

3. Page 17 line 12: Please spell out what 'SAE' means in 'SAE compliant'.
   Done.

4. I am not suggesting any changes to the parameters used for the simulations. But after reading page 18 lines 47-51, I wanted to mention the possibility of using Latin Hypercube Sampling as a method of picking simulation parameters in highly dimensional spaces. It allows one to fill high-dimensional volumes quite evenly with relatively few combinations of parameters, perhaps useful in a full simulation study on the current method.
   This is another nice suggestion. We are exploring a variant of LHS (based on orthogonal sampling) to help delineate and parse parameter space explored by a few of our ongoing simulation studies.

5. P 19, line 18-23. I am a little unclear after reading this on the nature of the simulations. First, it's not clear to me how subsets were defined for the analysis of simulated data. Assuming that subsets were defined as matching the simulated subsets of 1K sites, and does the fact that the maximum number of process partitions in the prior was approx. 6 mean that the simulations were not testing for overpartitioning here? I don't think this is a huge problem, and I don't think further simulations are necessary, but it would be useful if these points were clarified in the ms.

The reviewer is correct on the first point. The simulations presented in the paper specified the true data subsets—that is, the pre-specified data subsets correctly captured the simulated patterns of process heterogeneity (although, as noted above, we have also evaluated cases where we incorrectly specify the data subsets; those results will be presented in the more comprehensive simulation study).

On the second point—regarding the number of process partitions—I think we have failed to clearly explain this aspect of our simulation. To clarify, each simulated dataset was analyzed using a set of values for the concentration parameter—one with the prior mean for the number of process partitions centered on a very low value [$E(k) = 1$, reflecting the expectation that the substitution process is most likely homogeneous across data subsets], on an intermediate value [$E(k) = 3$, where the prior on the number of process partitions was centered closer to the true degree of process heterogeneity], and on a high value [$E(k) = 6$, where the prior on the number of process partitions was centered on a value much higher than the true degree of process heterogeneity]. Recall that the most complex process heterogeneity in this simulation was $k = 3$ (for the α-shape parameter and stationary frequencies), so specifying the concentration parameter such that the prior mean on the number of process partitions, $E(k) = 6$, evaluates the case where the prior belief is (incorrectly) focussed on an over-specified model. Moreover, each of these values for the concentration parameter specifies the prior *mean* of the number of process partitions—a prior mean of $E(k) = 6$ places considerable prior probability on higher dimensional mixtures. We have revised the relevant sections of the text to clarify these points.

6. Page 22 line 42. I'm not sure I agree that the results suggest that the 'pre-specified data partitions successfully captured process heterogeneity in the sequence alignment'. Specifically, since they don't allow for variation in process among sites of a single pre-specified partition, it seems hard to make this conclusion. It seems to me that to make this conclusion one would have to run the analysis with each site as a pre-specified partition, and see to what extent the recovered schemes matched or differed from those using traditionally defined subsets like genes and codon positions. Can the authors clarify here?
   Done. This point was also raised by Mark (see his 'Minor point 1', above), which we have addressed.

Thanks for a very enjoyable paper. I look forward to using the method, and to (I hope) retiring `PartitionFinder` in the very near future.
Rob Lanfear

Reviewer 2 (Anonymous) Comments:

Recommendation: Accept with minor or no revisions Comments:

Review for Systematic Biology

Manuscript ID: USYB-2014-198

Title: Bayesian Analysis of Partitioned Data

Authors: Moore et al.

Summary: This is really an exciting paper describing an analytical approach for Bayesian analysis of partitioned data using a Dirichlet process prior model. The paper is very well written and is reasonably accessible to the *Systematic Biology* audience. The paper will no doubt be a citation classic as it sets the stage for assessing process partitions as random variables instead of a fixed assumption in phylogenetic estimation approaches. The authors validate their approach with simulation data (albeit limited—but with reference to a more extensive study in the works) and demonstrate its effectiveness and relationship to Bayes factor selection using three different empirical data set analyses. The approach is implemented in software called "AutoParts", but no link is provided to the software. I tried looking it up on the corresponding author's website, but the link to his lab website doesn't seem to work. In the end, a method that is not available through some functional software is not particularly helpful. It is essential that the

authors provide a functional link to the software that implements the approach. It would have been nice to give it a run for the review.
We apologize to the reviewer for the difficulty in finding our software—we have included a link to the `AutoParts` website: https://github.com/brianrmoore/AutoParts.
Nevertheless, the approach is a significant step forward in phylogenetic analysis. I have detailed a number of comments below in the hopes of helping the authors see the paper through the eyes of a practicing systematist who actually understands a bit about the underlying models and statistics—most practicing systematists do not. So I've made a few suggestions to help this audience a bit as well. I've also pointed out some parts of the paper that I particularly enjoyed and why I think they are helpful components to the overall paper. Nevertheless, this is a really solid paper that will be read in detail at lab meetings around the world to those who care about phylogenetics. Excellent work!

Page 6 (authors pagination, not the journals pdf pagination), line 55 "tree-length parameter parameter" delete one 'parameter'
Done.

Page 7, Data Partitions—can the user-defined partitions be overlapping? For example, can the user have both coding versus non-coding and then 1st, 2nd, 3rd positions in the coding positions? Does this create an independence issue for the underlying statistics? Or do these data partitions need to be independent? Please provide the reader with a bit more detail here.'
Done. Each site must be included in one and only one pre-specified data subset. The data subsets are the 'elements' of the mixture that are participating in the DPP, and so a single site cannot be included in multiple data subsets. We have clarified this point in the text by adding the following statement: "Note that each site in the alignment must be included in one (and only one) data subset—sites cannot be 'orphaned' or simultaneously included in multiple pre-specified data subsets."

Page 8, thinking of the practicing systematist trying to digest this paper, can you give some accessible insight into "Bell numbers" and "Stirling numbers" and why you chose this particular approach to characterize partitions? It's also not obvious what you mean by an 'element' within a partition. This gets confusing here because you are sometimes talking about data partitions and sometime process partitions and sometime partitions. Are these all the same? Similarly, you talk about K elements as well as K data subsets (page 9). I think this will get confusing for the typical systematist.
Apologies for the confusion—we agree that the previous terminology was confusing; we have made an effort to adopt a simpler, and (hopefully less confusing) set of terms, which are now clearly defined in the text. We also failed to motivate the discussion of the combinatorics on partition schemes—I can see how it seemed to come out of nowhere (and nobody likes a combinatorics ambush). We have now added text explaining why we need to define the state space of the process partitions.

Regarding the Bell and Sterling numbers of the second kind; these are simply mathematical tools that we are using to understand the state space for the process partitions. Consider the following analogy. Imagine that we are using a discrete-gamma model to accommodate among-site rate variation, so our model will include a parameter, α, describing the shape of the (discrete) gamma distribution (and thereby specifying the degree of ASRV). Under a Bayesian framework, we typically treat the α-shape (hyper)parameter as a random variable. What kind of random variable? Well, that's up to the biologist. In our implementation, we have assumed that the α-shape parameter is a *uniformly* distributed random variable. OK, fine. But what is the state-space of possible values for this uniformly distributed random variable? This is relatively straightforward (if somewhat arbitrary) for this parameter: we have specified the state space as $\alpha \sim U(0.01, 100)$. All values between 0.01 and 100 are assumed to be equiprobable–smaller and larger values outside this range have zero prior probability (they are not possible).

With this analogy in mind, let's now return to partition schemes. As for the α-shape parameter, we are treating the partition scheme as a random variable. That means that we must specify a prior probability distribution for the partition scheme; *i.e.*, we have to specify what *kind* of random variable we believe the partition scheme to be. For this purpose, we use the DPP model, which specifies the prior probability distribution on the number of process partitions—the number of distinct processes for each substitution-model parameter, and the assignment of the pre-specified data subsets to those process partitions. In order for the DPP to specify the prior probability distribution on the number of partition schemes, we need to know the state space for this parameter. What is the state space for partition schemes? (*i.e.*, how many partition schemes are possible?) We use combinatorics (the branch of mathematics concerning the study of finite or countably discrete structures) to define the state space of possible discrete process partitions. In this case, the relevant solution relies on calculating 'Bell numbers' (which, in turn, are calculated as the sum of the corresponding 'Sterling numbers of the second kind')—Bell numbers specify the space of possible ways to uniquely assign K pre-specified data subsets (subsets of sites) among k process partitions (distinct substitution processes). So, all this combinatorics is basically just some necessary bookkeeping.

The other point—confusion regarding the distinction between *data partitions*, *process partitions*—is key both to understanding (and applying) the approach we are presenting. In conjunction with a revised description of the DPP model, we have added a figure that we believe will clarify these key concepts.

In fact, a general figure showing the relationship between data partitions, process partitions, partition schemes, and the associated parameters would be helpful for keeping track of all of this, especially through the empirical results section. I think the typical Sys Bio reader will likely get confused with the multiple uses of 'partition' here when we generally use this term with a singular meaning. This sort of figure would help understanding throughout the paper.
Done. This is a great idea! We have followed this suggestion (see previous response).

Page 10, The Chinese Restaurant Process—this actually helps a lot. You may get some flack from other reviewers about having such a pedestrian metaphor (as opposed to something more directly relevant to systematics), but I would urge you to keep this. It is very accessible. Thanks.
We did get some flack, and deservedly so. It seems clear that this metaphor may be getting threadbare (although I'm personally quite fond of it), and may not be providing a sufficiently clear general explanation of the DPP or its specific application to the problem of accommodating process heterogeneity. For these reasons, and to follow suggestions of the editors (Mark and Andy), we are dumping the 'Chinese restaurant' in favor of a new description (with a corresponding figure) that we believe (hope) clarifies the general and specific aspects of the DPP model.

Page 11, line 15, "We used 'Algorithm 8"' Can you provide a bit of justification for choosing this algorithm over alternatives here?
Well, we think this is a bit too technical—it might be more appropriate for a CS paper describing the implementation of the method in `AutoParts`. In any case, the justification for our choice of Algorithm 8 is completely pragmatic: it works well for this problem :).

Page 11, again, I appreciate the Chinese Table metaphor here, but you need to bring it back to the data partition problem. You never really do that here. You need an additional paragraph between lines 43-44 linking back to the data partition problem.
Again, the 'Chinese restaurant' has now been condemned for lack of pedagogical clarity. We have, however, made connections at this point of the text to our new explanatory framework. We think this is an improvement.

Page 11, line 46, please provide some information on the availability of `AutoParts` (like a web link).
Done. We have included a link to the `AutoParts` website: https://github.com/brianrmoore/AutoParts.

Page 12, line 22-23 "other parameters . . . presents" should be 'present'—indeed, this whole sentence is a bit confusing because you talk about 'parameters' with the tree topology as an example, but by the end of the sentence you seem to be talking about exclusively tree topology (e.g., 'it is an unusual model parameter'). So either make this sentence about the tree topology parameter itself or about 'unusual parameters' in general. The rest of the paragraph seems to be about the tree topology parameter specifically.
Done. EIC Andy Anderson also identified this passage as egregiously awkward. We have attempted to improve the clarity of this passage (as described above in our response to Andy's comment).

Page 15, line 39, "We used this approach on simulated . . . " insert '(see below for details)' as you've already described the empirical data sets at this point, but not the simulated data sets. This is the first mention of them, yet they are detailed in the next section. This would just help the reader know that that description is coming.

Done. Another nice suggestion.

Page 16, nice validation with the simulated data. Too often this sort of validation is overlooked or perhaps conducted, but not reported. I appreciate the fact that it was done and the opportunity to see the results. Very nice! Given all the action going on in this paper, it is appropriate to provide this limited simulation study for validation of the DPP approach with the foreshadowing of a more extensive simulation study to come.
Thanks! I appreciate the kind words—it took a couple years to complete the simulation study.

Page 17, first paragraph, I agree that careful simulation represents a non-trivial enterprise. To that end, I would encourage the authors to make their simulated data sets available to the community either through a lab website or Dryad deposition so others can take advantage of this hard work and run these data sets through their own approaches for future comparisons.
Absolutely agree. The datasets are available from the Dryad archive.

Page 22, line 25 "prior to unambiguously assign" should be 'assign[ing]'?
Done. We've edited this sentience for clarity.

Page 23, line 25 "Bayes factors to select mixed model that provides" should be either 'to a select mixed model' or 'to the selected mixed model'
I think this phrasing is correct, but may be unclear. Accordingly, the offending sentence "Bayes factors are used to select a mixed model that provides the best fit to the data" has been revised to read "The marginal likelihoods of the candidate models are then compared using Bayes factors to select the 'best' mixed model; *i.e.*, the candidate mixed model that provides the best description of the process heterogeneity in the dataset."

Page 23, you compare your approach to the Pagel & Meade finite-mixture model and critique it in that it requires 'that we specify the number of process partitions, k.' Yet your approach requires that we specify the number of data partitions, K. What's the tradeoff here in terms of specifying k versus K? The practicing biologist would prefer to allow the data to specify both.
Our discussion of the finite-mixture model proposed by Pagel and Meade is honestly not meant to be critical. It just entails tradeoffs in comparison to the infinite-mixture model that we describe (on page 24). We think that, in sum, the infinite mixture models is a better (meaning more practical and biologically defensible) compromise.
. . . OK you answer this a bit on page 24 in terms of the comparison with Bayes factors, very nice. But you still have the problem that every argument you make for examining and capitalizing on the distribution of process partitions seems like it would be value for the data partitions as well. You are conditioning on a single, perhaps arbitrary—but certainly single of many possible data partitions, and then looking at a distribution of process partitions. Using the Bayesian logic, wouldnt it be best to consider both the data partition and the process partition as a joint distribution to be estimated?
These are good points, which we now address when comparing our approach to that of Wu *et al.* Essentially, we view the use of pre-specified, user-defined data subsets as harnessing prior knowledge. The alternative position—assuming that the process may vary substantially across each individual site—ignores relevant prior information. The cost of doing so—as evidenced by the Wu *et al.* method—limits the useful application of this framework to small, single-gene datasets (*i.e.*, alignments with fewer than ~ 500 sites). We think that most systematists would view this as a poor tradeoff, as the admittedly strong assumption inherent in specifying data subsets is open to objective evaluation. That is, the approach we describe could be applied to alternative user-specified data-subset schemes, and these alternatives can be compared by virtue of their marginal likelihoods.

Page 26, the authors point out possible extensions of the DPP approach. I really like this. I think it is important to do this, especially in methodology papers, to motivate graduate students and postdocs to think about open questions in the field. I really appreciate this. Thanks.
Thanks! We think that the most exciting thing about this framework is its ability to usefully extended.