Dear Dr. Moore:
Decision on USYB-2010-093, Bayesian Analysis of Partitioned Data: Accept with major revisions

Thank you for your Systematic Biology submission. The original AE assigned has left the journal, so I solicited and collected reviews myself. It has been reviewed by one anonymous reviewer (#1) and Nicolas Rodrigue (reviewer 2, in the attached pdf file). Their comments are listed at the end of this letter. The reviewers provide some excellent constructive suggestions that I am sure you will appreciate.

The consensus opinion is that this manuscript presents a novel and important approach to handling the problem of partitioning data in Bayesian phylogenetic inference.

Reviewer 1 suggests major revision primarily because they would like to see inclusion of simulation results, where partitioning is known. Nicolas Rodrigue provides an extensive commentary, and makes several specific suggestions regarding ways in which the presentation of the material might be improved.

Please acknowledge (by email to systbiol@uconn.edu) receipt of the reviews and give a probable time frame for the return of your revised manuscript. Your revision should be submitted as soon as possible. Any revision not received in a timely manner may have to be considered a new submission.

Your revision must comply with our instructions to authors (in this letter, with additional instructions at http://www.oxfordjournals.org/our_journals/sysbio/for_authors). This applies even if your original submission deviated from Systematic Biology style and corrections were not included in the notes from the editors or reviewers. Therefore, READ AND APPLY OUR INSTRUCTIONS BEFORE SUBMITTING A REVISION. Failure to do so will result in significantly delayed processing of your paper. If anything remains unclear after reading the instructions, use recent issues of the journal to find examples. If necessary, feel free to contact the Managing Editor at systbiol@uconn.edu for help.

You will be unable to make your revisions in Manuscript Central. Instead, revise your paper using a word processing program and save it on your computer. Highlight the changes to your manuscript within the document using the track changes mode in MS Word or by using bold or colored text.

Log into http://mc.manuscriptcentral.com/systbiol and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions" click on "Create a Revision." Your manuscript number will be appended to denote a revision. Your original files are available to you. Delete the old files and replace them with your modified versions. Files that did not require revision can simply be retained. Data files, online appendices, etc. should be included with every version of your paper.

When submitting your revised manuscript, address each point made by the Editor, AE and Reviewers IN THE SPACE INDICATED. Your revision cannot be processed if your responses to reviews are given in a cover letter. The best way to address each point would be to copy this letter and insert your comments after each point made. Please do not change the order of or delete any of the comments because this makes it difficult to review again and would slow the process. The format of your responses must be compatible with Manuscript Central text fields. For example, colored text is not an option, but you could use asterisks (and numbers, spacing, etc.) to clearly distinguish your responses from the text of the reviews.

Feel free to argue your case, with careful consideration and documentation, if you disagree with any of the suggestions. If you feel a reviewer did not understand a point you made, in your response keep in mind that as an author it is your responsibility to make your points clear to the readers.

DO NOT submit your revision in .pdf format.
However, in the specific case of papers written in LaTeX, in addition to the .tex (and associated style or etc. files) please include a .pdf generated from those files. Upload the .pdf in the category for online appendices. Please name the file something like "pdf version" so that it will be clear that it is not actually an appendix. Figures must be separate files. No figures may be imbedded in or tied to the .pdf or .tex files.

After you've uploaded your revision, carefully view the Manuscript Central version to verify that all figures and other files display correctly, and that you've followed all of our author instructions.

Figures: Each figure must be submitted as an individual file. Make the numbering and lettering as large and clear as possible, without overlapping text. No figures may be imbedded in the document. Captions should not be placed on the figures. Figure portions should be referred to in the text and figure captions using lowercase letters. On the figures themselves, portions should be labeled by a lowercase letter followed by a single parenthesis (e.g., "a)"), located in the upper left area of the figure portion. Line thickness (including graph axes) should be a minimum stroke weight of 0.5pts, and 1.0pts is recommended for most lines. We prefer vector rather than bitmap figure formats. If desired, see http://systbio.org/?q=node/138 for an explanation of the difference. If bitmapped figures are necessary, they should be created at a minimum of 300 dpi (assuming the figure is about 8 inches wide). We accept a wide variety of figure formats as long as the figures are of sufficiently high resolution.

The cost of printed color figures is $600 each. Authors are normally expected to cover this cost, but we do have limited funds available for authors who are SSB members, and cannot pay the full amount. If you feel your circumstances create an unusual financial need, please explain in your cover letter. The maximum allowance is one color figure per paper. All color costs can be avoided if you decide to have the figures printed in black and white, but be shown in color in the online version of the paper. If you do this, the captions must be worded such that they are appropriate for both situations (e.g., descriptions should not name colors), because the captions in print and online will be identical even if color is used online only. Do not submit two versions of any figure. Instead, make sure the color figure is also easily legible in grayscale and submit only the color version. In your cover letter, please make your intentions regarding the use of color clear.

Journal Covers: Please consider submitting a suggested cover image. They can be illustrations of theory, photos of organisms, a combination of the two, or alternatives. These can be uploaded with a revision of the manuscript or may be sent later. Only images for which you have copyright permission or that are not under copyright may be submitted. The color figure fee does not apply to images chosen by the Editor to be on the cover of the issue. If you have questions about possible cover designs, please email systbiol@uconn.edu. We encourage and greatly appreciate cover image suggestions.

Data: Data files should be included with all versions of the manuscript. Before the manuscript can be published, data accession numbers must be in the text (the numbers may be added at the proofs stage if necessary). Nucleotide sequence data and alignments must be submitted to Genbank (http://www.ncbi.nlm.nih.gov/Genebank) or EMBL (http://www.ebi.ac.uk); alignment, input file and tree files must be submitted to TreeBASE (http://www.treebase.org); morphological data must be submitted to either Morphbank or MorphoBank.

Make sure all section headings conform to Systematic Biology style for first, second and third levels. Use of incorrect styles is potentially confusing and, in any case, is likely to delay processing of the manuscript.

Tables should have a single-sentence informative title above the table, with any other descriptive information located below the table in the form of notes and/or specific footnotes.

When references are grouped together in parentheses in the text, they should be listed in ascending chronological order. Multiple references in a single year should be alphabetized.

All funding used for this work should be listed in a "Funding" section preceding the Acknowledgements. Please give the full official name of each funding body.

Check over your online appendices (if any) carefully, because they will not be copyedited or proofread, and cannot be changed later. The first time online material is mentioned, give its location as http://www.sysbio.oxfordjournals.org. PDF is an acceptable format for online-only material (alternative formats are also acceptable), despite it being unacceptable for the text of the main document, because online material is not typeset.

Our accepted abbreviation for millions of years ago is Ma. The abbreviation for millions of years duration is myr. Our full instructions are accessible via the link in the upper right area of the Manuscript Central page (the link goes to http://www.oxfordjournals.org/our_journals/sysbio/for_authors).

If you are the first author and your manuscript is based on work done while you were a student, you would be eligible for the "Publisher's Award for Excellence in Systematic Research." When submitting your revised version be sure to indicate student work in the checkbox provided. If two student researchers were heavily involved in the manuscript, please briefly describe the situation in your cover letter.

I think that your paper will be a valuable contribution to Systematic Biology once the comments provided below are addressed. Thank you very much for your submission.

Sincerely,

Dr. Ronald DeBry
Editor-elect, Systematic Biology ron.debry@uc.edu

Reviewer: 1

Recommendation: Accept with major revisions

Comments:
This is a well-written manuscript that deals with the problem of selecting partitions of data, for example sets of genes, where processes are assumed to be similar enough to share some of the same sets of parameters. The goal is to obtain more precise estimates by combining partitions. The main current Bayesian methodology compares choices of partitions through Bayes factors and chooses the one that is best. The methods here, instead, averages over partitions in a Bayesian scheme: the Dirichlet process prior (DPP). To best of my knowledge, this is a new and potentially useful approach to the problem.

Based on experience with nested models, the expectation is that combining partitions appropriately will reduce the variances of parameter estimates and lead to more certain inferences about topology. Unless I missed it, the ms does not give direct empirical evidence of reduced parameter variances in the current setting. This could be remedied by a figure, table or numbers in the main text summarizing how posterior variances of parameter estimates with the DPP compared with those when different parameters are assigned to each element. The ms does indicate that more precise topological estimation [p20] arises.

The flip side, however, is that combining partitions inappropriately will reduce the variances inappropriately and lead to inappropriately more certain inferences about topology. Moreover, it can lead to biases. The difficulty is that the ms considers real data sets where the correct combination of natural partitions is not known. The ms would be greatly aided by simulation results, where the correct combination was known as a part of the generating model. What would be informative is how posterior means, variances, topological inferences varied across settings: e.g. (a) all elements having their same partition (b) the correct known generating partition (c) DPP. To make space for this, some of the current analyses could be taken out. I expect 3 real example data sets will be sufficient for the main points made.

Done. Although the suggested simulation study was very time consuming (>400,000 CPU hours), we feel that it ultimately proved very useful. We have generated simulated datasets with different patterns of process heterogeneity across partitions for the four parameters—base frequencies, substitution rates, alpha-shape parameter, and tree length. We analyzed these data sets under the Dirichlet process prior model—where the number of process partitions and the assignment of data subsets to those process partitions were not specified *a priori*. For these latter analyses, we also explored the impact of the concentration parameter governing the Dirichlet process prior model—setting the concentration parameter to three values that centered the prior mass on a uniform model, on a saturated model, and an intermediate value. These experiments allowed us to establish several important points regarding the statistical behavior of the Dirichlet process prior model: 1) the Dirichlet process prior model successfully recovers the true patterns of process heterogeneity with high accuracy (measured as the coverage of the true value in the 95% credible set of process partitions); 2) the Dirichlet process prior model successfully recovers the true patterns of process heterogeneity with high precision (measured as the size of the 95% credible set of process partitions relative to the prior number of process partitions); 3) the ability of the Dirichlet process prior model to shrink down on the true pattern of process heterogeneity—avoiding the inclusion of superfluous parameters—had the expected effect of decreasing error variance in model parameters, including that for the tree topology which is typically of central interest.

The summary of result are restricted largely to mean partitions. Because of the discreteness of the space and the differing numbers of partitions, it seems possible that, in some cases involving substantial variation in sampled partitions, this isn't representative of anything that arose in sampling, but rather something midway between those that arose. Moreover, there is little quantification of uncertainty in the figures even though it appears this may be considerable in some cases [p18]. The figures 2 onwards have little information content as figures. The same information could be presented through lists of sets. For instance {C1,C2,T1,T2,T3,16S}, {C3} describes the plot for Tree Length and E[K]=4 in Figure 5. Putting them in a Table might allow one to use the same space to add information about uncertainty and how representative the mean partition is. For instance, what proportion of time did this partition get sampled, what proportion of time did {T1,T2,T3} end up in the same partition...

Done. We are sympathetic to the reviewer's frustration with the challenges of summarizing the marginal posterior probability distribution for discrete parameters. We have included a discussion of this issue in the text (under the subsection *Summarizing Partitions*), where we clarify the rational behind (and methods used to derive) the mean process partition as a summary of the entire marginal posterior probability distribution of process partitions. We have also indicated the degree of uncertainty in the mean process partition by indicating the size of (number of process partitions) in the 95% credible interval for each parameter in each of the figures.

** Specific comments

-abstract and into: It is not made clear until p7 that, in application, there will be relatively few elements in a partition and that they will involve relatively large sets of sites. I could imagine one scanning the abstract and believing it is presenting a process-across-sites model.
Done. We have made it clear in the abstract that the Dirichlet process prior is used to integrate over all possible process partitions associated with the specified data subsets.

-p3 last par: Because of the relatively small number of elements, or initial partitions, the most complex model here, where each partition receives its own parameter, would still be considered parametric rather than nonparametric as suggested in this par.
Done. Our description of the Dirichlet process prior model as a 'non-parametric approach to clustering problems' is not intended to imply that it is 'free of parameters'—we clearly state that the DPP minimally includes the concentration parameter, χ, which governs the 'clumpiness' of the mixture, and the generating function, G_0, from which parameter values of the mixtures are drawn. Instead, we use the phrase 'non-parametric' in the strict sense, to indicate that the dimensions of the model are not specified *a priori*, but instead dynamically expand/contract with the degree and nature of variation in the data. We have added a sentence to clarify this point.

-p8: Are RGPs synonymous with "allocation vectors". [No.] If so, since neither term is used later in the ms, one might use to "allocation vectors" to avoid unnecessary notation.
Done. RGF is an acronym for the 'restricted growth function' notation of Stanton and White (1986), which is a conventional notation used for describing partitions (of which allocation vectors are an instance) as a string of integer values. We have clarified this point in the text.

-p9 bottom: some of the words have been globbed together.
Done. The text has been de-gobbled.

-p10 par2 "discussed elsewhere": citation would be appropriate.
Done. We have added appropriate references (Holder et al., 2005; Yang, 2006; Lakner et al., 2008).

-p12 mean partition: How is the mean, or least squares, partition obtained? Because of the discrete nature of the problem, it seems possible that heuristics that are not guaranteed to obtain it must be used.
Done. As described in our response to the second major comment, above, we have described the method used to estimate the mean process partition from the marginal posterior probability distribution of process partitions.

-p12 par2: Table 1 is missing.
Done. Table added.

-p12 par2 "chichlid sequences from ?.": The ? should be replaced
Done. Two of the empirical examples, including the cichlid dataset have been removed to make way for the simulation study results.

-p13 E(K): Replace this with E[k]. K is used for the number of elements (eg. K=11 for the skink sequences).
Done.

-p14 par2 "for for": Take a for out.
Done.

-p15 last par: I think there is potential for confusion. The wording seems to suggest inference is insensitive to chi. My understanding (on second reading) is that for a given choice of E[k], whether a hyperprior for chi is used or it is fixed to give this E[k], inferences are similar. Still, inferences are sensitive to the concentration parameter: different E[k] (which depends on chi) sometimes give quite different mean clusters in the examples.
Done. This sentence has been clarified.

-p16 par3 and fig1 legend "distinctly bimodal, suggesting process heterogeneity". I don't think process heterogeneity within the ND4 gene is suggested by a bimodal posterior. As a simpler example, generate normally distributed data with unit variance from two equal-sized partitions, one with mean -2 and the other with mean 2. The process is clearly heterogeneous, yet fitting a normal distribution with unit variance to the data will give a unimodal likelihood and hence posterior.

In most cases of bimodal posteriors/likelihoods that I am aware of, the cause is that widely differing parameters can give similar explanations for the data. Considered in isolation, it is not clear how widely differing alphas could provide similar explanations for the data. My guess is that it may be due to the variability in sampled partitions. If ND4 was granted its own partition in a fair number of sampled partitions that might explain the additional mode differing from the main mode associated with cluster 1. Whether this is the explanation or not, I think it would be valuable to try to explain the mode, particularly if the explanation is related to the DPP.
Done. The simulation study has verified that residual process heterogeneity within a data subset can induce multimodal marginal posterior probability densities for the corresponding parameters.

-p19 first full par: Since the DPP allows separate partitions for separate parameters, isn't the more appropriate count of the number of partitions allowed B(K)**4.
Done.

-p21 last par "accommodates variation in the evolutionary process across sites": perhaps "accommodates variation in the evolutionary process across partitions"
Done. Changed to "accommodates variation in the evolutionary process across a sequence alignment"

-p26 bottom: Figure 7 is listed but not included.
Done. Reference has been deleted.

Reviewer: 2

The second review by Nicolas Rodrigue was exceptionally insightful. The major comments from this review can be distilled as follows:

1. “I think that laying out the progression from mixed-model to finite mixture models to infinite mixture models would better orient the readers as to what has been accomplished (even if the finite mixture model is not in fact explored). The type of mixture studied here differs from previous works, and unless this is made clearer, I suspect that the preconceptions of many readers will lead them to confusion”.

Done. This was an exceptionally helpful suggestion. We now provide a more thorough discussion of the relationship of the Dirichlet process prior model approach to existing finite mixture models and mixed model approaches that closely follows the reviewer’s advice.

2. “I do not understand why the authors seem to not take seriously the problems with the method used to compute Bayes factors (the harmonic mean estimator).”

Done. The specific and general comments related to our previous discussion of Bayes factors (based on the harmonic mean estimator) have been incorporated in the revisions.

Recommendation: Accept with major revisions

Comments: See attached file.

Additional Questions:
Directions for Reviewers: The authors will appreciate detailed comments on the manuscript. Please write comments for the authors in a separate file, numbering all items that should be addressed before the manuscript is acceptable for publication, and attach your file at the bottom of this form (if you’ve inserted comments on an electronic copy of the manuscript, please attach that file as well). Reviewers are reminded that Systematic Biology is interested in publishing well-written papers of high scientific quality and of general interest. Thus, in your review, please address both the appropriateness of the paper for the journal as well as its scientific strengths and weaknesses. Please note that our instructions for authors are available on our website, systbiol.org. Use the buttons above to access the manuscript files. The HTML and PDF buttons link to the entire manuscript. Individual submitted files (such as data files) are available under the “Supplementary Files” button. We encourage reviewers to make comments directly on an electronic copy of the ms. If you do not have software that would allow you to make comments on the pdf version, please check under “Supplementary Files” to see if a Word version is available.:

Do you wish to remain anonymous?: No

How significant is this work?: Very

Is the author aware of the background and source material to the problems set forth?: Yes

Are the conclusions justified by the evidence presented and the assumptions involved?: Yes

Are the illustrations and tables clear and understandable?: Yes

In number are they: Sufficient