Bayesian Inference of Phylogeny from Partitioned Datasets

R.H. BAYESIAN ANALYSIS OF PARTITIONED DATA

BRIAN R. MOORE [1*], JIM McGUIRE [2,3], FREDRIK RONQUIST [4], ANNA E. CHRISS [2], AND JOHN P. HUELSENBECK [2]

[1] *Department of Evolution and Ecology, University of California, Davis*
Storer Hall, One Shields Avenue, Davis, CA 95616, U.S.A.

[2] *Department of Integrative Biology, University of California, Berkeley*
3060 VLSB #3140, Berkeley, CA 94720-3140, U.S.A.

[3] *Museum of Vertebrate Zoology, University of California, Berkeley*
3101 VLSB #3160, Berkeley, CA 94720-3160, U.S.A.

[4] *Swedish Museum of Natural History,*
Box 50007, SE-104 05 Stockholm, Sweden

*To whom correspondence should be addressed:

Brian R. Moore
University of California, Davis
Department of Evolution and Ecology
Storer Hall, One Shields Avenue
Davis, CA 95616
U.S.A.

Phone: 530-752-7104
E-mail: brianmoore@ucdavis.edu

Abstract.—It is well established that the evolutionary process may vary across the sites of nucleotide alignments; 'process heterogeneity' is an increasingly pervasive feature of datasets composed of gene regions sampled from multiple loci and/or different genomes. Inference of phylogeny from these data demands that we adequately model the underlying process heterogeneity; failure to do so can lead to biased estimates of phylogeny and other parameters. Traditionally, process heterogeneity has been accommodated under a 'mixed-model' approach, where sites are first assigned to data subsets based on relevant prior information, and then the phylogeny and other model parameters are estimated under the resulting mixed model. Here, we consider an alternative approach for accommodating process heterogeneity that is similar in spirit to this conventional mixed-model approach. However, rather than treating the partitioning scheme as a fixed assumption of the analysis, we treat the partition scheme as a random variable using a Dirichlet process prior model, where the phylogeny is estimated by integrating over all possible process partitions for the specified data subsets. We apply this method to simulated and empirical datasets, and compare our results to previous estimates that used Bayes factors to select among a limited set of candidate mixed models. We discuss the conceptual relationship of our new approach to existing Bayesian and maximum-likelihood alternatives for inferring phylogeny from partitioned sequence datasets, and also consider some of the practical limitations of these alternative approaches. We find that the Dirichlet process prior model provides a practical solution for inferring phylogeny from complex multi-gene/omic datasets, allowing discovery of novel process partitions that may more effectively balance error variance and estimation bias, while rendering phylogenetic inference more robust to process heterogeneity by virtue of integrating estimates over all possible partition schemes. The methods described in this paper have been implemented in the freely available program, `AutoParts`.

[Bayesian phylogenetic inference; Dirichlet process prior; Markov chain Monte Carlo; maximum likelihood; partitioned analyses; process heterogeneity.]

It is widely acknowledged that the pattern of nucleotide substitution may vary across an alignment of sequences, and that failure to accommodate this heterogeneity can potentially cause problems for phylogenetic inference. Deviations from a homogeneous substitution process include both simple *rate heterogeneity* (*i.e.*, among-site rate variation) stemming from site-to-site differences in selection-mediated functional constraints, systematic differences in mutation rate, *etc.*, or may involve more fundamental *process heterogeneity*, where the sites in an alignment are evolving under qualitatively different evolutionary processes. In the worst case, the tree relating the species in an alignment may vary across sites as a result of lineage sorting, hybridization, or horizontal gene transfer, *etc.* However, even when all of the sites in an alignment share a common phylogenetic history (*i.e.*, tree topology), other aspects of their evolutionary process may differ. For instance, process heterogeneity might occur within a single gene region (*e.g.*, between stem and loop regions of ribosomal sequences), or among gene regions in a concatenated sequence alignment (*e.g.*, comprising multiple loci and/or gene regions sampled from different genomes). Here, we focus on inference scenarios where the sites of an alignment share a common phylogenetic history but where two or more *process partitions* in the data (*sensu* **?**) may otherwise differ with respect to the underlying process of molecular evolution.

Failure to accommodate process heterogeneity is known to adversely impact phylogeny estimation, causing biased estimates of the tree topology and nodal support (**??**), estimates of branch lengths and divergence times (**???**), and estimates of other model parameters (**??**).

To avoid these problems, investigators typically adopt a Bayesian 'mixed-model' approach (*e.g.*, **?**) in which the sequence alignment is first parsed into a number of subsets that are intended to capture plausible process heterogeneity within the sequence alignment. A substitution model is then specified for each predefined data subset using a given model-selection method, such as the likelihood ratio test (LRT; **?**), the Akaike Information Criterion (AIC; **?**), or the Bayesian Information Criterion (BIC **?**). Finally, the phylogeny and other parameters are estimated under the specified composite model. Under this mixed-model approach, therefore, the partition scheme is an assumption of the inference (*i.e.*, the parameter estimates are conditioned on the specified mixed model that assigns data subsets to process partitions), and the parameters of each process partition are independently estimated. For most sequence alignments, a vast number of partition schemes of varying complexity are plausible *a priori*, which therefore requires a way to objectively identify the partition scheme that balances estimation bias and inflated error variance associated with under- and over-parameterized mixed models, respectively. Increasingly, mixed-model selection is based on Bayes factors (*e.g.*, **?**), which involves first estimating the marginal likelihood under each candidate partition scheme and then comparing the ratio of the marginal likelihoods for the set of candidate

partition schemes (*e.g.*, **???**). A similar approach for selecting partition schemes has also been developed in a maximum-likelihood framework (**?**).

There are three main concerns regarding this conventional mixed-model selection approach. First, the most common (and computationally efficient) technique for approximating the marginal likelihood of a (mixed) model—the harmonic mean estimator of **?**—is biased toward the inclusion of superfluous parameters, leading to the selection of overly partitioned mixed models (*e.g.*, **???**). Other marginal-likelihood estimators—based on the Savage-Dickey ratio (*e.g.*, **??**) or path-sampling techniques (*e.g.*, **??????**)—may avoid this bias, but are restricted to the evaluation of nested partition schemes or entail a substantially increased computational burden, respectively. Second, even using the most computationally efficient estimators, it will typically not be feasible to evaluate all (or even a tiny fraction) of the plausible partition schemes for a given alignment. Evaluating all plausible mixed models quickly becomes computationally prohibitive, as each candidate partition scheme requires a full MCMC analysis to estimate the marginal likelihood. Consequently, the mixed-model selection approach is apt to overlook the partition that best describes the patterns of process heterogeneity in the sequence alignment. A more fundamental limitation of the conventional approach for selecting partitions pertains to model uncertainty. Even if we could feasibly evaluate all possible candidate partitions for a given alignment and reliably select the mixed model that provides the best fit to that dataset, it may nevertheless be imprudent to condition inference on the chosen partition scheme. Given the vast model space, there may, in fact, be several (possibly numerous) partition schemes that provide a reasonable description of the process that gave rise to a given alignment. When there is uncertainty regarding the underlying mixed model, it would be unwise to condition inference on *any individual partition* (*cf.*, **??**). Instead, our phylogenetic estimates will be made more robust if we were to accommodate this mixed-model uncertainty by integrating over the plausible set of (or all possible) partition schemes.

Here, we propose an alternative approach for accommodating process heterogeneity that treats the partition as a random variable with a prior probability distribution specified by the Dirichlet process prior (DPP) model. The Dirichlet process prior model is often used in Bayesian solutions to clustering problems where the data elements [*e.g.*, (subsets of) nucleotide sites] are drawn from a mixture of an unknown number of probability distributions (*e.g.*, the underlying evolutionary processes). The Dirichlet process prior model allows both the number of mixture components and the assignment of individual data elements to the set of mixture components to vary. This approach has recently been applied to several phylogenetic problems, such as identifying heterogeneity in the process of amino-acid replacement in protein sequences (**?**), detecting sites under positive selection in protein-coding sequences (**?**), and accommodating variation in the rate of nucleotide substitution

across sites of an alignment (**?**) or across the branches of a tree (**?**).

The Dirichlet process prior model provides a natural means for accommodating process heterogeneity because it allows us to specify a non-zero prior probability on all possible partition schemes, ranging from a uniform model (in which all data subsets are assigned to the same process partition) to a saturated model (in which each data subset is assigned to a unique process partition). The prior weights on process partitions are first calculated analytically and then compared to their corresponding posterior probability estimates to provide a formal statistical framework for assessing the ability of partition schemes to capture process heterogeneity within the sequence alignment. We use numerical methods—Markov chain Monte Carlo (MCMC) simulation—to estimate the phylogenetic model parameters while integrating over all possible process partitions for the specified data subsets, thereby avoiding the need to evaluate or condition inferences upon any individual partition scheme. We first provide a more detailed description of the Dirichlet process prior model, and then describe how this approach can be used to identify the number and composition of partitions that best capture process heterogeneity. We evaluate this method by means of simulation and demonstrate its application to three empirical datasets, comparing results under this approach to those obtained using conventional model selection based on Bayes factors.

Methods

Phylogenetic Model Overview

The phylogenetic model that we develop includes an unrooted tree topology, τ, that describes the evolutionary relationships among the S species from which nucleotide data have been sampled. We assume that the biologist has correctly aligned the sequences and specified two or more subsets of sites ('data subsets') believed to capture patterns of process heterogeneity within the alignment. Each tree is assumed to be strictly binary, and so has a vector of $2S - 3$ branch lengths, $\mathbf{p}$, which are rendered proportional to the expected number of substitutions per site. The tree length, T, is simply the sum of its branch lengths. We describe the evolution of the sequence data over the tree with branch lengths using the general time reversible (GTR) substitution model (**?**). Substitutions under the GTR model occur according to the following matrix of instantaneous rates

$$\mathbf{Q} = \begin{pmatrix} \cdot & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & \cdot & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AC}\pi_A & r_{CG}\pi_C & \cdot & r_{GT}\pi_T \\ r_{AC}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & \cdot \end{pmatrix} \mu,$$

4

where the nucleotides are in the order A, C, G, T. Parameters of the GTR substitution model include six exchangeability parameters, $\mathbf{r} = (r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT})$, that allow for rate biases between nucleotides, and four nucleotide frequency parameters, $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$, that are the stationary probabilities of the substitution process. (The parameter μ is a scaling factor chosen such that the mean rate of substitution is one, which ensures that branch lengths are rendered as the expected number of substitutions per site.) Variation in substitution rate across sites is described using the discrete-gamma model (??), which assumes that the substitution rate at a particular site is a random variable drawn from a mean-one gamma distribution; for computational reasons, the continuous gamma is approximated using a finite number of discrete rate categories (following convention, we specify four rate categories). The parameter α describes the shape of the gamma distribution. The basic phylogenetic model therefore includes the following parameters:

Tree topology	τ
Branch-length proportions	$\mathbf{p} = (p_1, p_2, \ldots, p_{2S-3})$
Tree length	T
Exchangeability parameters	$\mathbf{r} = (r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT})$
Nucleotide frequencies	$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$
Gamma-shape parameter	α

We estimate parameters of the phylogenetic model in a Bayesian statistical framework, where inference is focused on the joint posterior probability distribution of the parameters. The posterior probability is calculated using Bayes's theorem

$$f(\tau, \mathbf{p}, T, \mathbf{r}, \boldsymbol{\pi}, \alpha \,|\, \mathbf{X}) = \frac{f(\mathbf{X} \,|\, \tau, \mathbf{p}, T, \mathbf{r}, \boldsymbol{\pi}, \alpha) f(\tau, \mathbf{p}, T, \mathbf{r}, \boldsymbol{\pi}, \alpha)}{f(\mathbf{X})}, \tag{1}$$

where $\mathbf{X}$ represents the alignment(s) of DNA sequences (see below). Bayes's theorem states that the posterior probability distribution of the parameters $[f(\cdot \,|\, \mathbf{X})]$ is equal to the likelihood $[f(\mathbf{X} \,|\, \cdot)]$ times the prior probability distribution of the parameters $[f(\cdot)]$ divided by the marginal likelihood $[f(\mathbf{X})]$. The likelihood is calculated under the phylogenetic model using numerical methods first described by ?.

Bayesian inference treats parameters as random variables; this requires that we specify a prior probability distribution for each model parameter that describes the nature of its random variation. The prior reflects the biologist's beliefs about the parameters before evaluating the data at hand. In this study, we assume the following prior probability distributions for the parameters of the

phylogenetic model:

$$
\begin{aligned}
\tau &\sim \text{Discrete Uniform}(1, \ldots, B(S)) \\
\mathbf{p} &\sim \text{Dirichlet}(1, 1, \ldots, 1) \\
T &\sim \text{Gamma}(2S - 3, \lambda_1) \\
\mathbf{r} &\sim \text{Dirichlet}(1, 1, 1, 1, 1, 1) \\
\boldsymbol{\pi} &\sim \text{Dirichlet}(1, 1, 1, 1) \\
\alpha &\sim \text{Exponential}(\lambda_2)
\end{aligned}
$$

where $B(S) = (2S - 5)!!$ is the number of possible binary, unrooted trees for S species. Following **?**, see also Huelsenbeck and Suchard, 2007, we specify a compound Dirichlet prior on branch-length proportions, which avoids the so-called 'long-tree' problem (*sensu* **??**).

Inference of phylogeny is increasingly based on datasets comprising gene regions sampled from multiple loci and/or different genomes that may have evolved under qualitatively different evolutionary processes. How should we accommodate potential heterogeneity in the parameter values within theses datasets? The program `MrBayes` (**?**) addresses this problem by allowing the user to designate subsets of sites within the sequence alignment (*i.e.*, 'data subsets'). Parameters of the phylogenetic model can then be independently estimated (*i.e.*, 'unlinked') for individual data subsets, or arbitrarily shared (*i.e.*, 'linked') across two or more data subsets. The specific configuration of (un)linking parameters across the specified data subsets corresponds to a specific 'partition.'

Imagine, for example, that our alignment includes a single protein-coding gene. We define three data subsets for this alignment corresponding to groupings of sites for each of the codon positions. We now consider how to model the overall substitution rate among the three data subsets. We might specify a common (completely linked) tree-length parameter for all three data partitions, which reflects the biological assumption that the overall substitution rate does not vary among the three codon positions. Alternatively, we could specify an independent (completely unlinked) tree-length parameter for each data partition, which assumes that the substitution rate differs among the three codon positions. Or we could partially link tree-length parameters among data subsets: *e.g.*, we might specify a shared tree-length parameter for first- and second-position sites, and an independent tree-length parameter for third-position sites. The same considerations apply to other parameters of the substitution model (the exchangeability rates, stationary frequencies, and gamma-shape parameter). The difficulty with this approach, however, is to find the best partition scheme among a bewildering number of alternative ways of linking and unlinking parameters across data subsets.

In this study, we use a DPP model to treat the partition scheme as random variable. This

approach obviates the need to condition phylogenetic inference on any individual partition scheme. Instead, parameters of the phylogenetic model are integrated over all possible partition schemes, visiting each according to its ability to accommodate the underlying process heterogeneity in the data. Specifically, we invoke an independent DPP model for each of the phylogenetic model parameters—the tree length, exchangeability rates, stationary frequencies, and gamma-shape parameter—to describe the process heterogeneity among the pre-specified data subsets. The remaining phylogenetic model parameters—the tree topology and branch-length proportions—are assumed to be shared across all of the data partitions. In the following, we describe the details of this model and the analyses performed in this study.

Data Subsets

We assume that the biologist has correctly aligned the nucleotide sequences sampled from S species, and then subdivided the sites of this alignment to define data subsets. These data subsets might reflect knowledge of the structure of the gene, dividing the data, for example, according to coding versus non-coding DNA, stem versus loop regions of ribosomal genes, or by codon position for protein-coding genes. For genomic data, the data subsets might correspond to individual genes. In the most extreme case, the biologist might assign every site to its own subset. Specification of data subsets is up to the biologist, and becomes a fixed assumption of the inference.

Consider, for example, the following alignment of $S = 5$ protein-coding sequences:

Human	CTGACTCCTGAGGAGAAGTCTGCCGTTACT...
Cow	CTGACTGCTGAGGAGAAGGCTGCCGTCACC...
Mouse	CTGACTGATGCTGAGAAGTCTGCTGTCTCT...
Marsupial	TTGACTTCTGAGGAGAAGAACTGCATCACT...
Chicken	TGGACTGCTGAGGAGAAGCAGCTCATCACC...

which are the first 30 sites of five of the sequences of a partial alignment of β-globin DNA sequences (**?**). This alignment can be subdivided by codon-position, which would result in $K = 3$ data subsets, which we label $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$:

$$\mathbf{X}_1 = \begin{pmatrix} \text{CACGGATGGA}\ldots \\ \text{CAGGGAGGGA}\ldots \\ \text{CAGGGATGGT}\ldots \\ \text{TATGGAATAA}\ldots \\ \text{TAGGGACCAA}\ldots \end{pmatrix} \quad \mathbf{X}_2 = \begin{pmatrix} \text{TCCAAACCTC}\ldots \\ \text{TCCAAACCTC}\ldots \\ \text{TCACAACCTC}\ldots \\ \text{TCCAAAAGTC}\ldots \\ \text{GCCAAAATTC}\ldots \end{pmatrix} \quad \mathbf{X}_3 = \begin{pmatrix} \text{GTTGGGTCTT}\ldots \\ \text{GTTGGGTCCC}\ldots \\ \text{GTTTGGTTCT}\ldots \\ \text{GTTGGGCCCT}\ldots \\ \text{GTTGGGGCCC}\ldots \end{pmatrix}$$

for the first, second, and third codon subsets, respectively. The complete alignment is denoted $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$. 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.

Combinatorics of Partitions

The parameters of the phylogenetic model can be assigned to data subsets in a large number of possible ways. Consider, as an example, a simple case in which the alignment has been divided into three subsets and a single parameter, θ, is to be assigned to the subsets. The parameter can be constrained to be the same in all three data subsets (which is equivalent to subdividing the alignment, and then ignoring the subsets when estimating parameters), allowed to be potentially different in each data subset, or constrained to be equal among some subsets but different in others. We can enumerate all of the possible process partitions for the three data subsets:

$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$		RGF
θ_1	θ_1	θ_1	$\rightarrow$	(1,1,1)
θ_1	θ_1	θ_2	$\rightarrow$	(1,1,2)
θ_1	θ_2	θ_1	$\rightarrow$	(1,2,1)
θ_1	θ_2	θ_2	$\rightarrow$	(1,2,2)
θ_1	θ_2	θ_3	$\rightarrow$	(1,2,3)

where the parameter subscript indicates the separately estimated parameters and the partition is also described according to the restricted growth function (RGF) notation (**?**). In order to compute the prior probability for a given partition scheme, we need to know how many partition schemes are possible for a given number of data subsets, K. We use combinatorics to compute the total number of process partitions. Specifically, the total number of partitions for K elements is described by the Bell numbers (**?**). The Bell number for K elements is the sum of the Stirling numbers of the second kind:

$$\mathcal{B}(K) = \sum_{i=1}^{K} \mathcal{S}(K,i), \tag{2}$$

where the Stirling numbers of the second kind are given by the following equation:

$$\mathcal{S}(n,k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^n. \tag{3}$$

Dirichlet Process Prior Model

The Dirichlet Process Prior model is a probability model on partitions (**??**). This model contains a single parameter, called the concentration parameter, χ, that controls how probability mass is assigned to different partitions. For the simple example in which there are $K = 3$ data subsets, the DPP model distributes probability to each process partition:

RGF	Prob.
(1,1,1)	1/3
(1,1,2)	1/6
(1,2,1)	1/6
(1,2,2)	1/6
(1,2,3)	1/6

(For this example, we assume $\chi = 1$.) More generally, the probability of a particular partition of K elements containing k subsets is

$$p(\mathbf{z}, k \mid \chi, K) = \chi^k \frac{\prod_{i=1}^{k} (\eta_i - 1)!}{\prod_{i=1}^{K} (\chi + i - 1)}, \tag{4}$$

where $\mathbf{z}$ is the allocation vector specifying how elements are assigned to subsets of the partition, k is the number of subsets in the partition (or the 'degree' of the partition), and η_i is the number of elements assigned to subset i. For the case in which $K = 3$ the values for these parameters are

RGF	$\mathbf{z}$	K	k	$\eta.$
(1,1,1)	(1,1,1)	3	1	$\eta_1 = 3$
(1,1,2)	(1,1,2)	3	2	$\eta_1 = 2,\ \eta_2 = 1$
(1,2,1)	(1,2,1)	3	2	$\eta_1 = 2,\ \eta_2 = 1$
(1,2,2)	(1,2,2)	3	2	$\eta_1 = 1,\ \eta_2 = 2$
(1,2,3)	(1,2,3)	3	3	$\eta_1 = 1,\ \eta_2 = 1,\ \eta_3 = 1$

The DPP model contains one additional component: a probability distribution (sometimes referred to as the 'base distribution', denoted G_0) for the parameter assigned to each partition. A complete description of the process, then, specifies the concentration parameter and the base distribution for the parameter. The DPP model is often used in Bayesian analysis for clustering. The elements of the partition are the objects to be clustered. Typically, the objects to be clustered correspond to data elements. The subsets of the partition denote the clustering; the subsets specify which elements are grouped together in the same cluster. Importantly, under the DPP model the number of clusters is a random variable. The DPP model does not force the biologist to pre-specify the number of clusters *a priori*.

The Chinese Restaurant Process is a useful metaphor that has been used to describe the DPP model (**?**). The metaphor works as follows: Consider a Chinese restaurant with a countably infinite

number of tables. Patrons enter the restaurant one at a time and randomly choose a table at which to sit. The ith patron will choose to sit at table m at which η_m patrons are already seated with probability $\frac{\eta_m}{i-1+\chi}$; alternatively, this patron will choose some unoccupied table with probability $\frac{\chi}{i-1+\chi}$. The 'restaurant patrons' in the metaphor are replaced by data elements. Data elements that are 'seated at the same table' share a common parameter, θ, which is drawn from the probability distribution G_0 (*i.e.*, $\theta \sim G_0$). Note that tables with large numbers of patrons will tend to attract new patrons.

We apply a separate DPP model to each of the phylogenetic model parameters, except for the topology and branch-length proportions of the tree (however, see **?**). Specifically, we first identify process partitions in the data and then allow the specific phylogenetic model parameters—the exchangeability rates, nucleotide frequencies, tree length, and gamma-shape parameter—to cluster across data subsets according to a DPP model.

Markov Chain Monte Carlo

We use Markov chain Monte Carlo (MCMC; **??**) to approximate the posterior probability distribution of the phylogenetic parameters. The general idea is to construct a Markov chain that has as its state space the possible values for the model parameters and a stationary probability distribution that is the target distribution of interest (*i.e.*, the posterior probability distribution of the model parameters). Draws from the Markov chain at stationarity are valid, albeit dependent, samples from the posterior probability distribution (**?**). Accordingly, parameter estimates are based on the frequency of samples drawn from the stationary Markov chain.

MCMC works by iterating the following steps many thousands or millions of times: (1) Call the current state of the Markov chain θ. If this is the first cycle of the Markov chain, then initialize θ to some value (perhaps by drawing values from the corresponding prior probability distribution). (2) Propose a new value for the model parameter θ, called θ'. The proposal mechanism must be stochastic, with a probability of $f(\theta \to \theta')$ of choosing the proposed state θ' from the current state θ. The probability of the reverse proposal, from θ' to θ, which is not actually made in computer memory, is $f(\theta' \to \theta)$. (3) Calculate the probability of accepting θ' as the next state of the Markov chain:

$$R = \min\left(1, \frac{\Pr(\mathbf{X} \,|\, \theta')}{\Pr(\mathbf{X} \,|\, \theta)} \times \frac{\Pr(\theta')}{\Pr(\theta)} \times \frac{f(\theta' \to \theta)}{f(\theta \to \theta')}\right). \tag{5}$$

(4) Generate a uniform(0,1) random variable called u. If $u < R$, then accept the proposed state and set $\theta = \theta'$. (5) Go to step 1.

10

We update the parameters of the phylogenetic model one at a time. The proposal mechanisms for most of the phylogenetic model parameters, such as the tree topology, nucleotide frequencies, exchangeability parameters, and gamma-shape parameter, are now rather well established (*e.g.*, ???). However, the proposal mechanism for the partitions with a Dirichlet process prior model merit further discussion. ? discussed several possible MCMC updates for DPP models. We used 'Algorithm 8' from that paper to update the partition schemes. To return to the Chinese Restaurant metaphor, we imagine that we have instantiated a restaurant in computer memory, with tables representing subsets of the partition scheme and patrons who are seated at tables representing the sites (or, as currently implemented, sets of sites that are determined by the biologist's initial division of the data into process partitions).

The MCMC procedure works as follows. First, pick a patron and remove it from the table. If the patron was the only individual at that table, then remove the table from computer memory. Otherwise, decrease the count of the number of patrons at the table by one (*e.g.*, decrease η_i). Calculate the likelihood when the patron is seated at the mth remaining table in computer memory (L_m). Also, calculate the likelihood when the patron is seated at each of the κ 'auxiliary' tables. These auxiliary tables are instantiated for the purpose of allowing the patron to be seated at a new (unoccupied) table. The parameter values for the auxiliary tables are drawn from the prior probability distribution for the parameter. The probability of seating the patron at the mth table in computer memory at which η_m patrons are already seated is $C \times \eta_m \times L_m$, and the probability that the patron is seated alone at the kth auxiliary table is $C \times (\chi/\kappa) \times L_k$, where C is a normalizing constant. After the patron is reseated at a previously existing or a new (auxiliary) table, the unoccupied auxiliary tables are deleted from computer memory. One MCMC cycle involves a scan of all the patrons, with the above update mechanism applied to each.

The methods described in this paper—which are intended to automate the phylogenetic analysis of partitioned datasets—have been implemented in the freely available application, `AutoParts`. This command-line program enables a fully hierarchical Bayesian phylogenetic analysis of partitioned data, where the input is an alignment of nucleotide sequences and two or more pre-specified data subsets. `AutoParts` uses MCMC to approximate the posterior probability distribution of all phylogenetic model parameters (tree topology, branch lengths, and substitution model parameters) while simultaneously integrating over all possible partition schemes under the DPP model. Source code (written in C++) and binaries for `AutoParts` are available at: https://github.com/brianrmoore/AutoParts.

Summarizing Partitions

In a Bayesian MCMC analysis, posterior probabilities of model parameters are based upon the

MCMC samples. For example, the posterior probability of a particular phylogenetic tree is approximated by the fraction of the time that the Markov chain visited that particular tree topology. One of the challenges of MCMC-based phylogeny estimation is how to most appropriately summarize the information within the MCMC samples. We can easily summarize MCMC samples of conventional, continuous model parameters (*e.g.*, the α-shape parameter describing the degree of among-site rate variation) by constructing a frequency histogram of sampled parameter values; this histogram provides an approximation of the corresponding marginal posterior probability density. The marginal posterior probability density can be further summarized in various ways, *e.g.*, as the mean of the posterior density, or an interval containing 95% of the posterior probability (*i.e.*, the 95% credible interval). 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. Nevertheless, several useful methods have emerged as conventions for summarizing a posterior sample of trees. For example, the majority-rule consensus tree is often used to summarize the results of an MCMC analysis (in the same way that bootstrap samples are summarized; **?**).

Our MCMC procedure also samples the process partitions of the model parameters. Partitions, like trees, are not standard model parameters and it is not obvious how the information contained in an MCMC sample of partitions should be summarized. One possibility is to simply report the posterior probabilities of individual partitions. This method, however, is impractical when the posterior probability is spread across many partitions and the Markov chain may then explore many thousands or millions of partitions. In this study, we summarize the MCMC samples of partitions using the concept of a 'mean partition'. The mean partition, denoted $\bar{\sigma}$, is the partition that minimizes the sum of the squared distances to all of the sampled partitions. We use a distance on partitions first described by **?**. The Gusfield distance between two partitions can be interpreted as the minimum number of elements that must be removed from both partitions to make the induced partitions identical. For example, the partitions $\sigma_1 = (1, 1, 1, 1, 2, 2)$ and $\sigma_2 = (1, 1, 1, 2, 2, 2)$ can be rendered identical by removing the fourth element resulting in the induced partition $(1, 1, 1, 2, 2)$; in this example, the partition distance is $d_p = 1$.

Analysis of Empirical Data

Sequence alignments and data partitions.—We demonstrate the DPP approach by applying it to three published datasets. Our choice of datasets was motivated by two considerations. First, these datasets have previously been the focus of careful analysis using the conventional mixed-model selection approach (where Bayes factors are used to choose among candidate partition schemes),

12

and so they provide a nice point of comparison to the DPP approach described here. Second, these datasets span a range of size (in terms of the number of species, S, and sequence length, L) and number of data subsets, K, which provides an opportunity to assess the ability of the new method in a variety of empirical settings. For each dataset, we assume an identical sequence alignment and adopt the same data subsets as those used in the original studies.

First, we analyzed the alignment of $S = 32$ gall wasp sequences from **?**. The gall wasp alignment includes $L = 3080$ sites, which we partitioned into $K = 11$ data subsets. Specifically, we partitioned each of the three protein-coding genes—including the two nuclear genes (elongation factor 1α [EF1α] and long-wavelength opsin [LWRh]) and the mitochondrial gene region (cytochrome oxidase c subunit I [COI])—by codon position. We also partitioned the 28S ribosomal sequence by stem and loop regions.

Second, we analyzed an alignment of $S = 60$ skink sequences from **?**. The skink alignment includes $L = 2678$ sites, which we partitioned into $K = 11$ data subsets. Specifically, we partitioned the mitochondrial gene region (NADH dehydrogenase subunit 1; ND1) by codon position. We also specified individual data subsets for the three tRNA sequences: tRNA$^{\text{GLU}}$, tRNA$^{\text{ILE}}$, and tRNA$^{\text{GLN}}$. Finally, we specified individual data subsets for the three ribosomal sequences: 12S rRNA, 16S rRNA, and tRNA$^{\text{MET}}$.

Third, we analyzed an alignment of $S = 164$ hummingbird sequences from **?**. The hummingbird alignment includes $L = 4122$ sites, which we partitioned into $K = 9$ data subsets. Specifically, we partitioned each of the three protein-coding genes—including the two mitochondrial gene regions (NADH dehydrogenase subunit 2 [ND2] and partial NADH dehydrogenase subunit 4 [ND4])—by codon position. We also specified a separate subsets for the tRNA flanking regions. Finally we specified two individual data subsets for each of the nuclear sequences—including intron 5 of the adenylate kinase gene (AK1) and intron 7 of the beta fibrinogen (BFib) gene regions.

Exploring the impact of the concentration parameter on phylogenetic inferences.—The DPP approach for accommodating process heterogeneity introduces a single novel parameter, the concentration parameter, χ, that controls the expected number of process partitions, $E(k)$. In order to assess the sensitivity of posterior estimates to the specified prior on χ, we performed two series of analyses. Our first series of analyses fixed the concentration parameter, χ, to a set of values. We chose values for χ that centered the prior mean on the number of process partitions, $E(k)$, to a corresponding set of values that spanned the possible range for each dataset. For example, in our analyses of the gall wasp and skink alignments, which both have $K = 11$ data subsets, we fixed the concentration parameter, χ, to a set of values such that the prior mean for the number of process partitions spanned the entire range of possible values—specifically such that $E(k) = \{\approx 1.0, 2.0, 4.0, 6.0, 8.0, 10.0, \approx 11.0\}$.

13

Similarly, our analyses of the hummingbird alignment, with $K = 9$ data subsets, specified a set of values for χ such that $E(k) = \{\approx 1.0, 2.0, 4.0, 6.0, 8.0, \approx 9.0\}$. We performed independent analyses under each of the corresponding values of $E(k)$ to assess the sensitivity of posterior estimates to the prior on the concentration parameter, χ.

Our second series of analyses explored an identical range of values for $E(k)$, but in these analyses we treated the concentration as a random variable. Our implementation of the DPP model allows the concentration parameter, χ, to be fixed to a specific value—that will center $E(k)$ to a corresponding value—or to use a hierarchical Bayesian modeling approach that treats the concentration parameter as a random variable. Specifically, we allow the concentration parameter, χ to be modeled as a gamma-distributed random variable: under this option, it is necessary only to specify the value of the parameter controlling the shape of the gamma hyperprior on χ; under this approach, the value of χ is estimated from the data. As in the fixed-χ specification of the DPP model, the mean of the gamma hyperprior on χ can be specified to center the prior mean number of process partitions, $E(k)$, to a desired value. In this second series of analyses, we specified the gamma hyperprior on χ to a set of values that centered the prior mean on the number of partitions, $E(k)$, to the same range of values as those we explored in the first series of analyses (*i.e.*, where the concentration parameter was treated as fixed).

MCMC analyses.—We estimated the joint posterior probability distribution for each of the three study groups under the Dirichlet process prior model using the Markov chain Monte Carlo (MCMC) algorithms implemented in `AutoParts`. Each MCMC simulation was initiated from random parameter values (drawn from the corresponding priors), run for $2\mathrm{x}10^7$ cycles, and thinned by sampling every 1000th state of the chain. In total, we performed 480 separate MCMC analyses on the empirical datasets: we analyzed each of the three alignments under the appropriate range of values for the expected number of process partitions, $E(k)$, and for each value of $E(k)$ we performed replicate analyses where the concentration parameter was treated first as fixed and then as a random variable (as described above), and performed four replicate analyses under each uniquely specified analysis (to assess MCMC performance).

Assessing MCMC performance.—Convergence of each chain to the stationary distribution was assessed by plotting time series of sampled values for all continuous parameters (exchangeability rates, stationary frequencies, tree length, gamma-shape parameter) using `Tracer` v.1.6 (**?**). We also assessed convergence by comparing independent estimates of the marginal posterior probability density for each continuous parameter, ensuring that parameter estimates from the four independent MCMC simulations were effectively identical and SAE compliant (*c.f.*, **?**). Similarly, we assessed

14

convergence by comparing independent estimates of the discrete parameters—the tree topology—using the '*comparetree*' function implemented in `MrBayes` v.3.1.2 (**?**) to assess the correspondence of clade posterior probabilities estimated by pairs of independent chains. We assessed the efficiency ('mixing') of the MCMC simulations both by monitoring the acceptance rates of proposals for all parameters, and by computing the estimated sample size (ESS) diagnostic implemented in `Tracer` v.1.6 (**?**).

Validating the MCMC algorithms.—We performed two series of experiments to validate the MCMC algorithms implemented in `AutoParts`. First, we assessed our ability to recover the prior. If an MCMC simulation is run without data, the likelihood for all parameter values will be identical. Therefore, the ratio of likelihoods for proposed and current states of the chain will equal one. Accordingly, acceptance probabilities of sequential states of the chain will be determined entirely by the prior ratio of proposed and current states (*c.f.*, eqn. 5). Consequently, a valid MCMC algorithm run without data will target the joint *prior* probability density rather than the joint *posterior* probability density. The *estimated* marginal prior probability density for each parameter can then be compared to its *known* (specified) counterpart. We used this approach on simulated and empirical datasets (see below for details), exploring the ability of the MCMC implementation in `AutoParts` to correctly recover the prior over a range of values for the concentration parameter.

We also validated the MCMC algorithms—and provided a benchmark for inference under the DPP model—by estimating the posterior probability distributions for each of the three study groups under a uniform GTR+Γ substitution model using `AutoParts` and the proven Metropolis-coupled MCMC (or MC^3) algorithms implemented in `MrBayes` v.3.1.2 (**?**). For the unpartitioned analyses using `MrBayes`, we enforced a uniform GTR+Γ substitution model, where each parameter was estimated from all of the sites comprising a given sequence alignment. For the unpartitioned analyses using `AutoParts`, we enforced a uniform GTR+Γ substitution model by setting the proposal probability of new partitions to zero, which prevented the MCMC from considering partition schemes of $k > 1$.

To assess the ability of `AutoParts` to recover the prior, we repeated the 480 MCMC analyses of the empirical data described above, in this case targeting the joint prior (rather than posterior) distributions. We performed an additional 24 MCMC simulations for the benchmark analyses. Specifically, we ran four independent MCMC simulations for each of the three alignments, and for each alignment we performed analyses under a uniform GTR+Γ substitution model using both `AutoParts` and `MrBayes`. We performed each of the 504 MCMC simulations as described above. However, our analyses using the MC^3 algorithm in `MrBayes` were performed using four incrementally heated chains (where the parameter governing the 'temperature' of the heated chains ranged from

$0.2 - 0.1$). We assessed performance of the MCMC simulations as described previously; however, we also assessed convergence of the `MrBayes` analyses by running 'paired'-MC^3 simulations (*i.e.*, two independent, parallel simulations using four incrementally heated chains) and monitoring the average standard deviation of split frequencies sampled by the paired cold chains.

Analysis of Simulated Data

Simulation study design.—To complement the empirical demonstration of the DPP approach, we also assessed the ability of this method to successfully recover parameter values under simulation. Simulation is a critical tool for validating an inference method, as it can uniquely assess the ability of an approach to recover the true/known parameter values that were used to generate the data. The success of a model on simulated data, however, may not provide a reliable guide to its empirical behavior: the design of the simulation study therefore warrants careful consideration. A simulation study will provide meaningful insights into the statistical behavior of a given method only if it adequately explores the relevant parameter space. Conceptually, this entails defining an N-dimensional hypervolume (for methods impacted by N variables, such as tree size, number and size of data subsets, number of process partitions, degree of process heterogeneity, etc.) that encompasses conditions that are likely to be encountered when the method is applied to real data. This space is first discretized into a number of cells (each corresponding to a unique parameter combination), and each cell is then explored with an adequate intensity (by simulating a sufficient number of realizations under each parameter combination). Each of the simulated datasets is then analyzed using the inference method to assess its statistical behavior over the simulated parameter space. The dimensionality of this parameter space and the associated computational burden can quickly become immense, particularly for complex models.

Accordingly, a comprehensive simulation study exploring the statistical behavior of the DPP approach is beyond the scope of the current study, and will be published separately (Moore and Huelsenbeck, in prep.). Herein, we present a simulation experiment that is typical both of the design of the more general simulation study and also of the statistical behavior of the DPP approach. Specifically, we simulated 100 datasets with $K = 6$ data subsets (each with $L = 1000$ sites) for a tree with $S = 50$ species. We assumed that the six data subsets share a common tree topology, τ, and branch-length proportions, $\mathbf{p}$, but allowed other aspects of the substitution process to differ among data subsets. For each simulation, the tree topology, τ, and vector of branch-length proportions, $\mathbf{p}$, were drawn from their corresponding prior probability distributions. Values for the other substitution model parameters were specified as described in Figure 1. We specified values for the concentration parameter, χ, following the protocol used in the empirical analyses. Specifically, we

assessed the sensitivity of estimates to the prior on χ (and the treatment of χ as a fixed or random variable) by analyzing each simulated dataset under the possible range of values for the expected number of process partitions, $E(k) = \{\approx 1.0, 3.0, \approx 6.0\}$, which we specified using both fixed values of χ and also using a gamma hyperprior on χ. Note that the degree of process heterogeneity in the simulated datasets ranged from $k = 1$ (tree-length parameter), $k = 2$ (exchangeability parameters), and $k = 3$ (stationary frequencies and α-shape parameter). Accordingly, the values of the concentration parameter that we used to analyze these simulated datasets explored cases where we focused the prior mean on the number of process partitions of a mixture model that was either under-specified [$E(k) \approx 1.0$], correctly specified [$E(k) = 3.0$], or over-specified [$E(k) \approx 6.0$]. Careful simulation represents a non-trivial enterprise: the simulation experiment presented here consumed approximately $403,200$ hours (≈ 46 years) of compute time.

MCMC analyses and performance assessment.—We estimated the posterior probability distribution of parameters for each simulated dataset under the Dirichlet process prior model using `AutoParts`. Each MCMC simulation was initiated from random parameter values (drawn from the corresponding priors), run for 2×10^6 cycles, and thinned by sampling every 1000th state of the chain. In total, we performed 1800 separate MCMC analyses on the simulated datasets: we analyzed each of the 100 replicate datasets under the possible range of values for the expected number of process partitions, $E(k) = \{\approx 1.0, 3.0, \approx 6.0\}$, and for each value of $E(k)$ we performed replicate analyses where the concentration parameter was treated first as fixed and then as a random variable, and performed three replicate MCMC simulations under each uniquely specified analysis (to help assess MCMC performance). We assessed convergence and mixing of each MCMC analysis of the simulated datasets following the protocol described previously for the empirical analyses.

We assessed the statistical performance of the DPP approach based on properties of the 95% credible intervals (CI) of process partitions estimated for the simulated datasets. Specifically, we computed the 95% CI of process partitions for each MCMC sample by first ordering the sampled process partitions by their posterior probability. Next, we added the probability for the first process partition to that for the second process partition, and then added this sum to the probability for the third process partition, and so on, until the *cumulative* probability was equal to or greater than 0.95. The resulting set of (one or more) process partitions corresponds to the 95% CI. We then used the 95% CIs on process partitions to assess the *accuracy* and the *precision* of the DPP approach. Accuracy is defined by the *coverage probability*, which is simply the proportion of replicates for which the 95% CIs contain the true process partition. Coverage probabilities ≥ 0.95 are considered unbiased and robust estimators. Precision is defined as a function of the size of the 95% CI: for a given coverage probability, estimators with a smaller CI are more precise. We measured the

precision of estimates as a function of the average number of process partitions comprising the 95% CIs, $\overline{\text{CI}}$, relative to the maximum possible CI size, CI_{max}:

$$\text{Precision} = \frac{\text{CI}_{\text{max}} - \overline{\text{CI}}}{\text{CI}_{\text{max}} - 1} \times 100,$$

where the maximum possible size of the 95% CI is defined as the Bell number for K data partitions, $\text{CI}_{\text{max}} = \mathcal{B}(K)$ (*c.f.*, eqn. 2). The precision therefore ranges from $0 - 100$, where larger values indicate higher precision.

<center>Results</center>

Performance of MCMC algorithms.—In general, the MCMC algorithms implemented in `AutoParts` to estimate the joint posterior probability density of trees and other parameters under the Dirichlet process prior model appear to have performed well both for the empirical and simulated datasets, whether we enforced uniform substitution models (for the purposes of algorithm validation), or when jointly estimating the process partitions with the concentration parameter fixed to a specified value or with the hyperprior of the concentration parameter centered on a specified value. Validity of the MCMC algorithms is also confirmed by their ability to successfully recover the prior when run without data.

Under baseline inference scenarios—where a uniform GTR+Γ substitution model was enforced—independent estimates obtained using the MCMC algorithms implemented in `AutoParts` and the MC^3 algorithms implemented in `MrBayes` were virtually identical (see *Supplementary Material*). Congruence of estimated marginal posterior probability densities of all model parameters—including the overall marginal log likelihoods and clade probabilities—suggests that the MCMC algorithms used to implement the Dirichlet process prior model in `AutoParts` performed appropriately under these simplified inference conditions.

Under more complex inference scenarios—where inferences were integrated over all possible partition schemes under the Dirichlet process prior model—the MCMC algorithms again appear to have performed well. Time series plots of all model parameters sampled by each of the four independent MCMC analyses under each of the 56 unique concentration-parameter treatments that we explored for the three empirical alignments appeared to stabilize within five million cycles (see *Supplementary Material*). Accordingly, parameter estimates were based on the final 15 million cycles from each of the four independent MCMC simulations. Parameter estimates for the simulated datasets were conservatively based on the latter half of samples drawn in each MCMC simulation.

Moreover, independent chains inferred virtually identical marginal posterior probability densities for all parameters—including those of the GTR substitution model, the gamma-shape parameter

<center>18</center>

used to model among-site substitution rate variation, as well as the clade posterior probabilities, and the mean process partition scheme for each of the three empirical alignments—providing additional evidence that the MCMC algorithms used to implement the Dirichlet process prior model successfully converged to the stationary distributions for these datasets (see *Supplementary Material*).

The MCMC algorithm appears to have mixed well over the joint posterior densities for the three empirical datasets: acceptance rates of proposal mechanisms for all parameters were within the target window $(20 - 70\%)$, and the marginal posterior probability densities for all parameters were typically tight and focused. Samples drawn during the stationary phase of independent chains were pooled, which provided adequate sampling intensity for all estimated parameters for all three datasets (*i.e.*, all ESS values $>> 10^3$).

Performance of the DPP approach under simulation.—The results of our simulation experiment demonstrate the ability of the method to reliably recover true/known parameter values (Table 1). In fact, the statistical performance of the DPP approach in this simulation was remarkably high: the 95% credible set of process partitions included the true process partition with an average coverage probability of 1.0, reflecting essentially perfect accuracy. Moreover, the average size of the 95% credible set of process partitions was very small (comprising on average $\bar{x} = 2.6$ process partitions), reflecting a very high degree of precision ($\bar{x} = 99.2$). These results were insensitive both to the value of the concentration parameter (χ, controlling the prior mean on the number of process partitions, $E(k)$), and also to the approach used to specify the concentration parameter (either as a fixed value or gamma-distributed random variable). Although these results are encouraging, the statistical behavior of the DPP approach will benefit from a much more comprehensive simulation study (Moore and Huelsenbeck, in prep.). The remainder of the results focus on empirical applications of the DPP approach.

Sensitivity of inferences to the concentration parameter.—The two primary series of analyses explored alternate approaches for specifying the concentration parameter of the DPP model. Whether we fixed the value of χ or a placed a gamma hyperprior on χ to center the prior mean on the number of process partitions at a particular value [*e.g.*, $E(k) = 4$], inferences were virtually indistinguishable both for the simulated datasets (Table 1) and empirical alignments (see *Supplementary Material*). This result has important practical implications: most empirical applications of the DPP approach will lack a basis for committing to a particular fixed value of χ, which would require several independent analyses that increment χ over a range of values for $E(k)$. However, our findings suggest that a vague hyperprior can be placed on χ so that it can be reliably estimated from the data (or equivalently, inferences can be integrated over a suitably broad prior distribution for χ) in the

course of a single MCMC analysis (*c.f.*, **????**). Because inferences under the two approaches were virtually identical, the remainder of the discussion will focus exclusively on results obtained under fixed values of χ.

Some specific results.—Before describing more general aspects of our findings, we will first focus on results for a particular (albeit fairly typical) case. Specifically, estimates of the gamma-shape parameter used to model among-site variation in substitution rates across the hummingbird dataset (Figure 2). The marginal posterior probability density of the gamma-shape parameter is independently estimated for each of the $K = 9$ data partitions (Figure 2A), where the density for each data partition is an amalgam of 24 overlain estimates—that is, the four replicate MCMC analyses performed under each of the six values for the prior mean on the number of process partitions $[E(k) = \{\approx 1.0, 2.0, 4.0, 6.0, 8.0, \approx 9.0\}]$ that we explored for the hummingbird dataset.

Four aspects of these results warrant comment. First, for each data subset, the marginal posterior probability densities estimated by the four independent MCMC analyses *under each particular value* of $E(k)$ are virtually identical, indicating that the independent chains successfully converged to the stationary distributions.

Second, the marginal posterior probability densities of the gamma-shape parameter estimated from each data subset *over the entire range of values* for $E(k)$ are equally indistinguishable, suggesting that these parameter estimates are robust to (mis)specification of the concentration parameter of the Dirichlet process prior model.

Third, the marginal posterior probability densities are typically focused and unimodal in form. A notable exception is apparent for the second-position sites of the ND4 gene (Figure 2B): the marginal posterior probability density for this data subset is clearly bimodal, which may reflect residual heterogeneity in the degree of among-site rate variation within this data subset. Because the data subsets are pre-specified by the investigator, the latent structure within the second-position sites of the ND4 gene is not accessible to the Dirichlet process prior model.

Finally, the set of marginal posterior probability densities for the nine data subsets appear to cluster in three distinct groups: (1) the first cluster comprises data subsets for tRNA and the first- and second-position sites of the ND2 and ND4 genes (with relatively high levels of among-site substitution rate variation); (2) the second cluster comprises subsets for the AK1 and BFib introns (with intermediate levels of among-site substitution rate variation), and; (3) the final cluster includes the two data subsets for the third-position sites of the ND2 and ND4 genes (with relatively low levels of among-site substitution rate variation). Indeed, this allocation of data subsets among process partitions is identical to the mean partition scheme identified by the Dirichlet process prior model. For convenience, we summarize these results using the graphical convention depicted in Figure 2C.

20

This plot reveals the posterior estimate for the number of process partitions, $E(k \mid \mathbf{X}) = 3$, and the inferred mean partition scheme for this parameter, $E(\bar{\sigma} \mid \mathbf{X})$, which summarizes the assignment of nine pre-specified data subsets among the three process partitions for the α-shape parameter. The number in the upper left of the panel indicates the number of partition schemes in the 95% credible set (*i.e.*, the 95% CI) for this parameter.

General aspects of the results.—Inspection of the number of process partitions comprising the mean partition schemes provides insight into the level and nature of process heterogeneity within the sequence alignments. For example, the gamma-shape parameter was inferred to have the largest number of process partitions across all three datasets (Figures 3–5), which is consistent with previous empirical (*e.g.*, **?**) and theoretical (*e.g.*, **??**) results that have identified among-site substitution rate variation as a critically important modeling component. By contrast, the rank order in the number of process partitions—and the corresponding level of process heterogeneity—inferred for the other parameters (exchangeability rates, base frequencies, and tree length) differed among the three sequence alignments.

Several generalities regarding the impact of the concentration parameter merit comment. First, even when a substantial proportion ($> 95\%$) of the prior probability was placed on a homogeneous substitution model [*i.e.*, where $E(k) \approx 1$; the left-most columns in Figures 3–5], the mean partition schemes for all datasets were nevertheless inferred to have $E(k \mid \mathbf{X}) > 1$, suggesting that these alignments are very difficult to explain using a uniform model.

Second, inferences of the mean partition schemes for the three datasets differed somewhat in their sensitivity to the prior on the number of partitions, $E(k)$. Specifically, the mean partition schemes inferred for skinks ($K = 11$) and hummingbirds ($K = 9$) were largely stable over the entire range of priors placed on the number of process partitions (Figures 4 and 5, respectively), whereas those for gall wasps ($K = 11$) exhibited greater sensitivity to the prior on $E(k)$, such that the number of process partitions increased as more prior probability was placed on more parameter-rich partition schemes (Figure 3). Even in these cases, however, sensitivity of the mean partition schemes to the priors on $E(k)$ for gall wasp dataset was restricted to two parameters, while the inferred process partitions for the other parameters remained stable over the entire range of $E(k)$.

Third, the pattern and level of uncertainty in estimates of the mean partition schemes—reflected in the size of the 95% CI of $E(\bar{\sigma} \mid \mathbf{X})$—varied somewhat with the prior mean on the number of process partitions across individual parameters and among the three datasets. A prominent pattern (exhibited by 40% of the parameters) involved a uniform level of uncertainty in the mean partition scheme over the range of priors on $E(k)$; *e.g.*, the size of the 95% CI for the mean partition scheme of the tree-length parameter in the gall wasp alignment included $1 - 2$ partition schemes

over $E(k) = \{1.0 - 11.0\}$ (Figure 3). An equally common pattern (also exhibited by 40% of the parameters) involved an increasing level of uncertainty in the mean partition scheme that increased with the prior mean on the number of partition schemes; *e.g.*, the 95% CI of the substitution-rate parameters in the gall wasp alignment peaks at $E(k) \approx 11$ (Figure 3). As $E(k)$ increases, more prior mass is placed on partition schemes that include a greater number of process partitions, increasing the probability that data subsets will be distributed over a greater number of process partitions, such that fewer data are available to estimate each parameter. Accordingly, the marginal posterior probability densities for each parameter are apt to become more diffuse (especially for smaller data subsets and/or those with limited variation), which impedes the ability of the Dirichlet process prior to unambiguously assign these marginal posterior probability densities to process partitions, with a corresponding increase in the uncertainty associated with the inferred mean partition schemes. In the final pattern (exhibited by the remaining 20% of the parameters), uncertainty in the mean partition scheme was greatest at intermediate values of $E(k)$; *e.g.*, the 95% CI for the α-shape parameter in the gall wasp alignment peaks at $E(k) = 8$ (Figure 3). This pattern may partly reflect the increased number of possible partition schemes at intermediate values of $E(k)$. For example, the 11 data partitions in the gall wasp alignment can be combined in $1,023$ unique partition schemes with two process partitions, $179,487$ unique partition schemes with six process partitions, and $11,880$ unique partition schemes with eight process partitions.

Finally, posterior estimates of the mean partition scheme for the empirical datasets were associated with considerable uncertainty: on average, the 95% CI of partition schemes included ≈ 1802 process partitions (*i.e.*, calculated by summing the size of the credible sets for the four parameters— the columns in Figures 3–5—averaged over all values of $E(k)$ and over the three alignments). The generally large number of plausible partition schemes for the empirical datasets contrasts sharply with those for the simulated data. The large size of the empirical credible sets for partition schemes indicates a correspondingly high degree of uncertainty in the choice of partition scheme, and emphasizes the appeal of integrating inferences of phylogeny (and other model parameters) over this important source of uncertainty (rather than conditioning inferences on any particular partition scheme).

DISCUSSION

Conceptual relationship of the DPP model to other approaches.—Accommodating variation in the substitution process across an alignment of nucleotide sequences is inherently a partitioning problem: the objective is to identify subsets of sites that share common aspects of their molecular

evolution. This requires specifying or estimating both the number of process partitions, k, and the allocation vector, $\mathbf{z}$, specifying how the data elements are assigned to those k process partitions. This problem has been addressed using three distinct approaches. The *mixed-model* approach (*e.g.*, ?) specifies both the number of process partitions and the assignment of data elements to those process partitions, which collectively define a given partition scheme. Alternative partition schemes represent competing explanations (mixed models) of the pattern of process heterogeneity that gave rise to the data. The average fit of each candidate mixed model to the data is first assessed by estimating its marginal likelihood. 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.

By contrast, *finite-mixture* models specify only the number of process partitions, k, but treat the allocation vector, $\mathbf{z}$, as a random variable. That is, the number of process partitions is assumed, but the assignment of the K data partitions to those k process partitions is assumed to be unknown. Accordingly, the likelihood function takes the form of a weighted average over the k process partitions. Specifically, the likelihood for each of the K data partitions is integrated over the k process partitions, and the resulting set of k likelihoods are then combined as a weighted average, where the weights are estimated from the data. ? proposed a finite-mixture model that allows for k unique process partitions that differ in the exchangeability rates, while assuming that other aspects of the substitution model—the stationary frequencies, tree length, alpha-shape parameter—do not vary over the sequence alignment. This innovative approach obviates the need to specify the allocation vector of the partition scheme, but requires that we specify the number of process partitions, k. A brute force solution for specifying the number of process partitions would entail performing a series of analyses that increment the number of process partitions, $k = \{1 \ldots K\}$, and estimating the marginal likelihood for each of the corresponding K finite-mixture models. Bayes factors could then be used to identify the finite-mixture model that provides the best description of the process heterogeneity in the dataset.

The Dirichlet process is an example of an *infinite-mixture* model. Under this approach, each of the K data partitions is modeled as having evolved under one of a countably infinite number of processes—as having been generated by a mixture model—without knowing the form of the mixture. Instead, the form of the mixture (including both the number of process partitions and the allocation of data elements among them) is governed by higher-level parameters (*i.e.*, the concentration parameter, χ, and the base distribution, G_0, of the DPP model). In principle, the likelihood function involves an integral over all possible mixture models conditional on these higher-level parameters. In practice, however, numerical methods (MCMC algorithms) are used to perform the integration.

This approach provides a dynamic solution to mixture modeling, allowing the dimensions of the model to efficiently expand or shrink as dictated by patterns in the data.

Remarkably, an approach quite similar to the one we present here was published independently[1] by Chieh-Hsi Wu, Marc Suchard and Alexei Drummond (2013) while our paper was in revision. The approach described by ? is an extension of the finite-mixture model of ?, and represents a special case of the DPP approach developed in the current paper. Rather than exploiting prior information on likely patterns of process heterogeneity across the alignment, this approach attempts to assign individual sites to substitution model parameters under a DPP model. They describe three models that differ with respect to the substitution model parameters that are included in the mixture. Under their stochastic Dirichlet Process (SDPM1) model, the substitution model parameters are restricted to (1) the exchangeability rates (as in the ? approach), and (2) the stationary frequencies, whereas their SDPM2 model includes a tree-length parameter. These choices effectively restrict this approach to the analysis of single-gene regions; their study focussed on the analysis of four relatively small, single gene regions (comprising an average of 724 sites), and even for these simple datasets, it was necessary to use empirical (hyper)priors for substitution model parameters in order for the MCMC to provide reliable estimates. [We note, however, that despite the use of informative empirical priors, intensive MCMC simulations—with an average chain length of $\bar{x} = 227$ million cycles—appears to have failed to provide valid samples, as the MCC summaries of ultrametric trees inferred under the SDPM1 and SDPM2 models contained negative branch lengths (see their supplemental figures S5 and S8).]

Introduce Wu et al. method. Describe it, and distinguish their approach from ours. JPH: Didn't you just do that?

Practical comparison of the DPP and mixed-model approaches.—The Dirichlet process prior approach for accommodating process heterogeneity is most similar in spirit to the conventional partition-scheme selection that relies on Bayes factors. The three alignments evaluated in the present study have previously been subjected to extensive mixed-model selection using Bayes factors, which affords the opportunity to compare these two approaches. Below we consider four aspects of this comparison.

In this section, we consider practical comparisons between the conventional mixed-model ap-

---

[1]JPH was invited by Marc Suchard to give a departmental presentation of our work on the DPP at UCLA in April, 2011. Subsequently, BRM presented this work at the `RevBayes` symposium/workshop held at UC, Berkeley in August, 2011 (attended by Marc Suchard), and the details of implementing the DPP were discussed extensively with Marc Suchard at that workshop. Neverthless, we were completely unaware of their independent work on this problem until their paper was published in March, 2013 (*c.f.*, ??).

proach, the DPP approach described by Wu et al., and our DPP approach. The nature of the comparison is that the first approach makes overly stringent/inflexible assumptions about biologically plausible patterns of process heterogeneity, whereas the Wu et al. approach makes too little use of this relevant prior information.

First, in every case, analyses using the Dirichlet process prior model discovered novel mean partition schemes that were not considered in the previous studies that evaluated candidate partition schemes using Bayes factors. We emphasize that the selected studies are notable for the relatively large number of partition schemes that they considered. Nevertheless, the need to perform a full MCMC analysis under each candidate partition scheme (in order to estimate the corresponding marginal likelihood) entails substantial computational overhead. Accordingly, even these exceptionally thorough studies necessarily considered only a small fraction of the possible mixed-model space. Specifically, the hummingbird study evaluated 0.0000000000000045% (9 of $2\text{x}10^{17}$) of the possible process partition schemes for an alignment with $K = 9$ data partitions, and the gall wasp and skink studies evaluated 0.000000000000000000000042% (9 of $2\text{x}10^{23}$) of the partition schemes possible for an alignment with $K = 11$ data partitions. The ability of the Dirichlet process prior model to integrate inference of phylogeny over the entire range of partition schemes in the course of a single MCMC analysis greatly increases the probability of estimating under the partition scheme that best captures process heterogeneity within the data, thereby minimizing biases associated with mixed-model misspecification.

Second, the complexity of the partition schemes preferred by the two approaches differs markedly. As in all other applications, mixed-model selection using Bayes factors identified the most parameter-rich candidate model for the gall wasp, skink, and hummingbird alignments. By contrast, the mean partition schemes inferred under the Dirichlet process prior approach were considerably more parsimonious. Specifically, the mean/selected partition schemes identified under the Dirichlet process prior/Bayes factor approaches included 32/99 free parameters for the gall wasp alignment, 27/98 free parameters for the skink alignment, and 29/108 free parameters for the hummingbird alignment. The strong preference for the most complex candidate mixed model again illustrates the well-known bias of Bayes factor comparisons based on marginal likelihoods inferred using the harmonic-mean estimator (*e.g.*, ?????). More reliable (but substantially more computationally intensive) marginal-likelihood estimators are available (*e.g.*, ??????). However, the scope of the problem may render these solutions impractical. For example, even if we could estimate the marginal likelihoods at the (computationally astronomical) rate of one partition scheme per second, it would still take $\approx 6,723,137,822,834,178$ years to evaluate all possible partition schemes for the gall wasp alignment (roughly one million times the age of the universe).

A more fundamental limitation of the conventional approach for selecting partition schemes pertains to (mixed) model uncertainty. Even if we could feasibly evaluate all possible candidate partition schemes for a given alignment and reliably select the mixed model that provides the best fit to that dataset, it may nevertheless be imprudent to condition inference on the chosen partition scheme. For example, the 95% credible set of partition schemes for the datasets evaluated here comprised 2935, 2434, and 37 mixed-models for the gall wasp, skink, and hummingbird analyses, respectively. Selecting *any single partition scheme*—even the MAP estimate—would ignore the considerable uncertainty regarding the choice of mixed model that gave rise to these alignments. Failure to accommodate uncertainty in the choice of (mixed) model is likely to bias our phylogenetic estimates (*cf.*, **??**). Accordingly, our phylogenetic estimates will be rendered more robust if we accommodate this mixed-model uncertainty by integrating over all possible partition schemes using an approach such as the Dirichlet process model. Although Bayes factors may not provide a viable approach for *accommodating* process heterogeneity, we nevertheless envisage them playing a key role in testing specific hypotheses regarding process heterogeneity, *e.g.*, comparing two competing hypotheses regarding the assignment of data elements to process partitions.

The final point of comparison between the DPP and mixed-model approaches pertains to difference in error variance. A corollary of the substantial reduction in model complexity under the Dirichlet process prior model—achieved by reduction of superfluous parameters—is a corresponding reduction in the error variance and associated uncertainty of parameter estimates. For example, the level of uncertainty associated with the topology parameter—as reflected in the size of the 95% credible set of trees—estimated under the Dirichlet process prior/Bayes factor approaches included 927/1513 trees for the gall wasp alignment, 2488/3106 trees for the skink alignment, and 3961/4672 trees for the hummingbird alignment, respectively. The reduction in estimation error (particularly the topology and branch-length parameters) under the Dirichlet process prior model should decrease uncertainty not only in estimates of phylogeny, but also in phylogeny-based inferences focused on the study of character evolution, rates of lineage diversification, biogeographic history, etc. A model that more realistically describes the processes by which increasingly complex, genome-scale sequence data arise will provide more reliable estimates of phylogeny and, by extension, more reliable inferences of evolutionary processes from phylogenies.

Applications and extensions.—The Dirichlet process prior model we describe provides a versatile approach for accommodating process heterogeneity across an alignment of nucleotide sequences, and so should both improve our estimates of phylogeny and inferences based on those phylogenies. Importantly, this framework can readily be applied to other inference problems and extended in various ways. For example, the DPP approach could be productively applied to numerous problems

in molecular evolution simply by reversing the relationship between focal and nuisance parameters. That is, our motivation for the DPP approach was to provide more robust estimates of phylogeny (tree topology and branch lengths) by virtue of integrating over all possible patterns of process heterogeneity across the alignment (where the process partitions are effectively nuisance parameters). However, an equally fruitful application of the DPP approach would focus directly on the nature of process heterogeneity to study various aspects of molecular evolution (where the phylogeny is essentially a nuisance parameter), providing a flexible framework to directly study how various aspects of the substitution process—stationary frequencies, exchangeability rates, overall substitution rate and the degree of ASRV—vary within and among gene/omic regions (*c.f.*, Figure 2).

The DPP approach can also be extended in various ways by relaxing some current assumptions. For example, we have implemented the DPP approach for accommodating process heterogeneity under the so-called 'unconstrained' phylogenetic model (*e.g.*, **?**), where the rate of substitution is independent on each branch, and the branch length is proportional to the degree of molecular evolution. Because accommodating process heterogeneity is known to be important for estimating divergence times (*e.g.*, **???**), it may also be useful to implement the DPP approach under a 'constrained' phylogenetic model (where the rate of substitution is shared across branches according to a strict-/relaxed-molecular clock model) and branch lengths are proportional to relative or absolute time. In lieu of a fully hierarchical solution for estimating divergence times, the current method could be used in a serial Bayesian approach; *i.e.*, where `AutoParts` is first used to estimate the mean partition scheme that best describes the nature of process heterogeneity across the alignment, and then this mixed substitution model is used to estimate divergence times (under any strict-/relaxed clock models).

Our approach accommodates variation in the evolutionary process across a sequence alignment, but nevertheless assumes that these sequences share a common phylogenetic history. For some datasets—particularly multiple nuclear loci sampled from closely related and/or recently diverged species—the assumption of a single phylogenetic history is likely to be violated by processes such lineage sorting and deep coalescence. Accordingly, it may prove valuable to extend the Dirichlet process prior model to accommodate this aspect of process heterogeneity by treating the tree topology like the other parameters of the DPP model (*c.f.*, **?**). That is, rather than assuming that all data partitions share a common tree-topology parameter, we could treat the number of topologies and the assignment of data partitions to those tree topologies as random variables. We are optimistic that the Dirichlet process prior approach as currently described will provide biologists with an efficient and robust method for accommodating process heterogeneity when estimating phylogeny, and that future extensions of this framework will prove fruitful for addressing related problems in

phylogenetic biology.

References

Akaike, H. 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19:716–723.

Aldous, D. 1985. Exchangeability ad related topics. Pages 1–198 *in* École d'Été de Probabilités de Saint-Flour XIII-1983 Springer, Berlin.

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution 24:412–426.

Antoniak, C. E. 1974. Mixtures of Dirichlet processes with applications to non-parametric problems. Annals of Statistics 2:1152–1174.

Baele, G., P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko. 2012a. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution 29:2157–2167.

Baele, G., W. L. S. Li, A. J. Drummond, M. A. Suchard, and P. Lemey. 2012b. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Molecular Biology and Evolution 30:239–243.

Bell, E. T. 1934. Exponential numbers. American Mathematics Monthly 41:411–419.

Brandley, M. C., A. Schmitz, and T. W. Reeder. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Systematic Biology 54:373–390.

Brooks, S. P. and A. Gelman. 1997. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7:434–455.

Brown, J. M., S. M. Hedtke, A. R. Lemmon, and E. M. Lemmon. 2010. When trees grow too long: Investigating the causes of highly inaccurate Bayesian branch-length estimates. Systematic Biology 59:145–161.

Brown, J. M. and A. R. Lemmon. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Systematic Biology 56:643–655.

Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Systematic Biology 42:384–397.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2012. jmodeltest 2: more models, new heuristics and parallel computing. Nature methods 9:772–772.

Dorazio, R. M. 2009. On selecting a prior for the precision parameter of the Dirichlet process mixture models. Journal of Statistical Planning and Inference 139:3384–3390.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4:e88.

Escobar, M. D. and M. West. 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90:577–588.

Fan, Y., R. Wu, M.-H. Chen, L. Kuo, and P. O. Lewis. 2011. Choosing among partition models in Bayesian phylogenetics. Molecular Biology and Evolution 28:523–532.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Ferguson, T. S. 1973. A Bayesian analysis of some nonparametric problems. Annals of Statistics 1:209–230.

Gelfand, A. E., A. Kottas, and S. N. MacEachern. 2005. Bayesian nonparametric spatial modeling with Dirichlet process mixing. Journal of the American Statistical Association 100:1021–1035.

Gusfield, D. 2002. Partition-distance: a problem and class of perfect graphs arising in clustering. Information Processing Letters 82:159–164.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Heath, T. A., M. T. Holder, and J. P. Huelsenbeck. 2012. A Dirichlet process prior for estimating lineage-specific substitution rates. Molecular Biology and Evolution 29:939–255.

Heath, T. A. and B. R. Moore. 2014. Bayesian inference of species divergence times. Pages 487–533 *in* Bayesian Phylogenetics: Methods, Algorithms, and Applications (L. K. Ming-Hui Chen and P. Lewis, eds.) Sinauer Associates, Sunderland, MA.

Holder, M. T., P. O. Lewis, D. L. Swofford, and B. Larget. 2005. Hastings ratio of the local proposal used in Bayesian phylogenetics. Systematic Biology 54:961–965.

Huelsenbeck, J. P., S. Jain, S. W. D. Frost, and S. L. K. Pond. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proceedings of the National Academy of Science, U.S.A. 103:6263–6268.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Molecular Biology and Evolution 21:1123–1133.

Huelsenbeck, J. P. and B. Rannala. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Systematic Biology 53:904–913.

Huelsenbeck, J. P. and M. Suchard. 2007. A nonparametric method for accommodating and testing across-site rate variation. Systematic Biology 56:975–987.

Lakner, C., P. van der Mark, J. P. Huelsenbeck, B. Larget, and D. L. Simon. 2008. Efficiency of MCMC proposals in Bayesian phylogenetics. Systematic Biology 57:86–103.

Lanfear, R., B. Calcott, S. Y. W. Ho, and S. Guindon. 2012. Partitionfinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution 29:1695–1701.

Larget, B. and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution 16:750–759.

Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution 21:1095–1109.

Lartillot, N. and H. Philippe. 2006. Computing Bayes factors using theromodynamic integration. Systematic Biology 55:195–207.

Li, W. L. S. and A. J. Drummond. 2012. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Molecular Biology and Evolution 29:751–761.

Marshall, D. C., C. Simon, and T. R. Buckley. 2006. Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. Systematic Biology 55:993–1003.

McGuire, J. A., C. C. Witt, D. L. Altshuler, and J. V. Remsen. 2007. Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Systematic Biology 56:837–856.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21:1087–1092.

Neal, R. M. 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9:249–265.

Newton, M. A. and A. E. Raftery. 1994. Approximate Bayesian inference by the weighted likelihood bootstrap. Journal of the Royal Statistical Society, B 56:3–48.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. N. Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology 53:47–67.

Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology 53:571–581.

Poux, C., O. Madsen, J. Glos, W. W. de Jong, and M. Vences. 2008. Molecular phylogeny and divergence times of Malagasy tenrecs: Influence of data partitioning and taxon sampling on dating analyses. BMC Evolutionary Biology 8:102.

Rambaut, A. and A. J. Drummond. 2007. Tracer v1.5. http://beast.bio.ed.ac.uk/Tracer.

Rannala, B. and Z. Yang. 2007. Inferring speciation times under an episodic molecular clock. Systematic Biology 56:453–466.

Rannala, B., T. Zhu, and Z. Yang. 2012. Tail paradox, partial identifiability and influential priors in Bayesian branch length inference. Molecular Biology and Evolution 29:325–335.

Ronquist, F. and J. P. Huelsenbeck. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Hhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012. Mrbayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–542.

Schwarz, G. et al. 1978. Estimating the dimension of a model. The annals of statistics 6:461–464.

Stanton, D. and D. White. 1986. Constructive Combinatorics. Springer-Verlag, New York.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Molecular Biology and Evolution 18:1001–1013.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures in Mathematics in the Life Sciences 17:57–86.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). Annals of Statistics 22:1701–1762.

Venditti, C., A. Meade, and M. Pagel. 2008. Phylogenetic mixture models can reduce node-density artifacts. Systematic Biology 58:286–293.

Verdinelli, I. and L. Wasserman. 1995. Computing Bayes factors using a generalization of the Savageickey density ratio. Journal of the American Statistical Association 90:614–618.

West, M., P. Müller, and M. D. Escobar. 1994. Hierarchical priors and mixture models, with application in regression and density estimation. Pages 363–386 *in* Aspects of Uncertainty: A Tribute to D. V. Lindley (A. F. M. Smith and P. Freeman, eds.) Wiley, Chichester, UK.

Wu, C.-H., M. A. Suchard, and A. J. Drummond. 2013. Bayesian selection of nucleotide substitution models and their site assignments. Molecular Biology and Evolution 30:669–688.

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Systematic Biology 60:150–160.

Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution 39:306–314.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. Journal of Molecular Evolution 42:587–596.

Yang, Z., R. Nielsen, N. Goldman, and A. M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure. Genetics 155:431–449.

Zhang, C., B. Rannala, and Z. Yang. 2012. Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. Systematic Biology 61:779–784.

Table 1: Performance (coverage[1] [precision[2]]) of the DPP approach under a typical simulation.

Parameter	χ (fixed)[3]			χ (gamma)[3]		
	$E(k) \approx 1$	$E(k) = 3$	$E(k) \approx 6$	$E(k) \approx 1$	$E(k) = 3$	$E(k) \approx 6$
ASRV	1.0 [99.8]	1.0 [98.9]	1.0 [97.7]	1.0 [99.8]	1.0 [98.9]	1.0 [97.7]
Base frequencies	1.0 [99.8]	1.0 [98.9]	1.0 [97.7]	1.0 [99.8]	1.0 [98.9]	1.0 [97.7]
Substitution rates	1.0 [100.0]	1.0 [100.0]	1.0 [100.0]	1.0 [100.0]	1.0 [100.0]	1.0 [100.0]
Tree length	1.0 [98.3]	1.0 [97.8]	1.0 [98.2]	1.0 [98.3]	1.0 [97.8]	1.0 [98.2]

These results correspond to the simulation experiment described in Figure 1.

[1]The probability that the 95% credible interval (CI) includes the true (generating) partition scheme.

[2]The mean number of partition schemes comprising the 95% CI as a function of the maximum possible size.

[3]The concentration parameter, χ, specifying the prior mean on the number of partition schemes, $E(k)$, is either fixed or governed by a gamma hyperprior.

Figure 1. A typical simulation experiment. This simulation involves $K = 6$ data partitions, each with $L = 1000$ sites. All data partitions are assumed to share a common tree topology, τ, and branch-length proportions, $\mathbf{p}$, but other aspects of the substitution process may differ between data partitions, including: the degree of among-site rate variation (ASRV, governed by the α-shape parameter); the vector of nucleotide frequencies, $\boldsymbol{\pi}$; the vector of exchangeability rates, $\mathbf{r}$, and the overall rate of substitution/tree length, T. The partition scheme, σ, for each of the four parameters specifies both the number of process partitions, k, and the allocation vector, $\mathbf{z}$, of the $K = 6$ data partitions to those k process partitions. For example, the 6 data partitions are allocated to $k = 3$ distinct process partitions for the α-shape parameter (the shaded rows of the panel), where the allocation vector is $\mathbf{z} = (1, 2, 3, 3, 2, 1)$ (depicted by the assignment of the data-partition symbols to their respective rows), and the specific parameter values for each process partition are indicated in the corresponding rows of the panel.

Figure 2. Inferred patterns of among-site substitution rate variation within the hummingbird sequence alignment. (A) Marginal posterior probability densities for the α-shape parameter of the discrete gamma model have been estimated for each of the 9 predefined data partitions. (B) In contrast to the unimodal marginal densities for other data partitions, the posterior density for the second-position sites of the ND4 gene is distinctly bimodal, suggesting residual process heterogeneity within this data partition. (C) Analyses under the Dirichlet process prior model provide estimates both for the number of process partitions, $E(k \mid \mathbf{X}) = 3$ (represented as rows in the panel), as well as the mean partition scheme, $E(\bar{\sigma} \mid \mathbf{X})$ (depicted by the assignment of the data-partition symbols to their respective rows).

Figure 3. Inferred patterns of process heterogeneity within the gall-wasp sequence alignment inferred under the Dirichlet process prior model. Each row of panels corresponds to a parameter of the model, and each column of panels corresponds to a specified value for the prior mean on number of process partitions in the sequence alignment, $E(k)$. Accordingly, each panel summarizes estimates for a particular parameter under a specified value of $E(k)$. The number of occupied rows within a panel corresponds to the estimated number of process partitions, $E(k \mid \mathbf{X})$, and the allocation of data partitions (depicted as symbols) among the rows of a panel specifies the inferred mean partition scheme, $E(\bar{\sigma} \mid \mathbf{X})$. The number in the upper left of each panel specifies the size of the 95% credible set of partition schemes.

Figure 4. Inferred patterns of process heterogeneity within the skink sequence alignment inferred under the Dirichlet process prior model. Graphical conventions as detailed in Figure 3.

Figure 5. Inferred patterns of process heterogeneity within the hummingbird sequence alignment inferred under the Dirichlet process prior model. Graphical conventions as detailed in Figure 3.
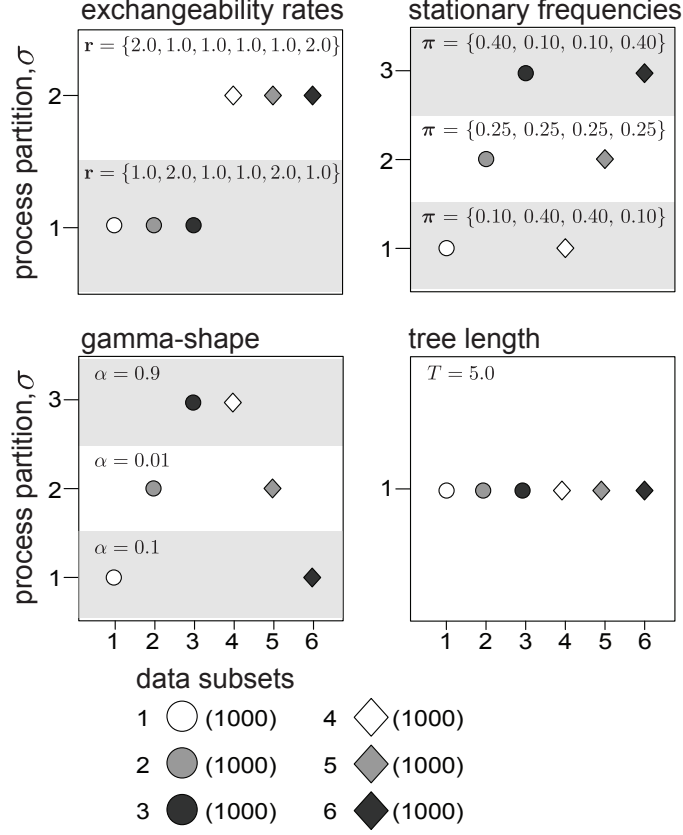
Figure 1: A typical simulation experiment. This simulation involves $K = 6$ data partitions, each with $L = 1000$ sites. All data partitions are assumed to share a common tree topology, τ, and branch-length proportions, $\mathbf{p}$, but other aspects of the substitution process may differ between data partitions, including: the degree of among-site rate variation (ASRV, governed by the α-shape parameter); the vector of nucleotide frequencies, $\boldsymbol{\pi}$; the vector of exchangeability rates, $\mathbf{r}$, and the overall rate of substitution/tree length, T. The partition scheme, σ, for each of the four parameters specifies both the number of process partitions, k, and the allocation vector, $\mathbf{z}$, of the $K = 6$ data partitions to those k process partitions. For example, the 6 data partitions are allocated to $k = 3$ distinct process partitions for the α-shape parameter (the shaded rows of the panel), where the allocation vector is $\mathbf{z} = (1, 2, 3, 3, 2, 1)$ (depicted by the assignment of the data-partition symbols to their respective rows), and the specific parameter values for each process partition are indicated in the corresponding rows of the panel.
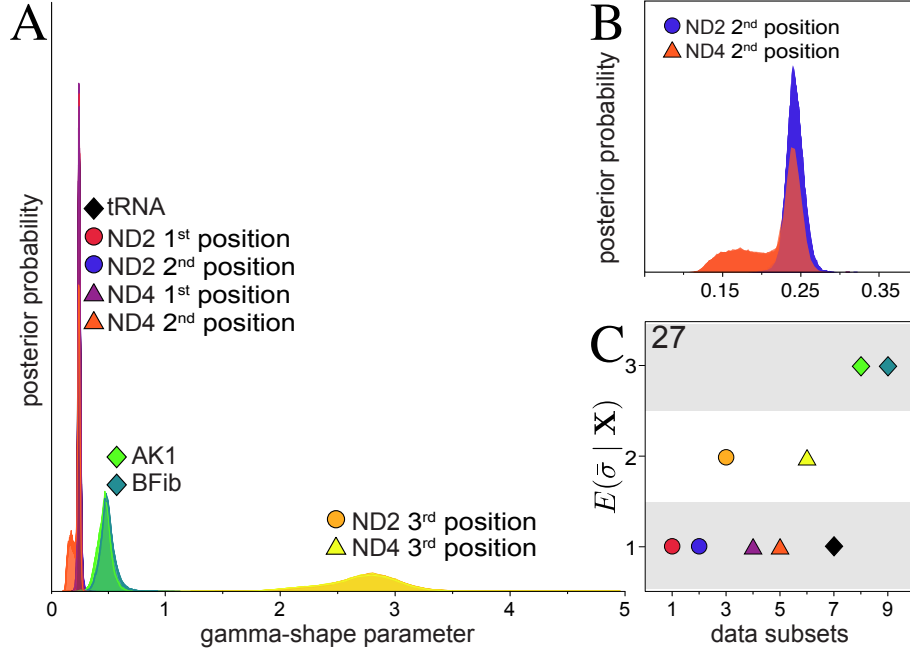
Figure 2: Inferred patterns of among-site substitution rate variation within the hummingbird sequence alignment. (A) Marginal posterior probability densities for the α-shape parameter of the discrete gamma model have been estimated for each of the 9 predefined data partitions. (B) In contrast to the unimodal marginal densities for other data partitions, the posterior density for the second-position sites of the ND4 gene is distinctly bimodal, suggesting residual process heterogeneity within this data partition. (C) Analyses under the Dirichlet process prior model provide estimates both for the number of process partitions, $E(k \mid \mathbf{X}) = 3$ (represented as rows in the panel), as well as the mean partition scheme, $E(\bar{\sigma} \mid \mathbf{X})$ (depicted by the assignment of the data-partition symbols to their respective rows).
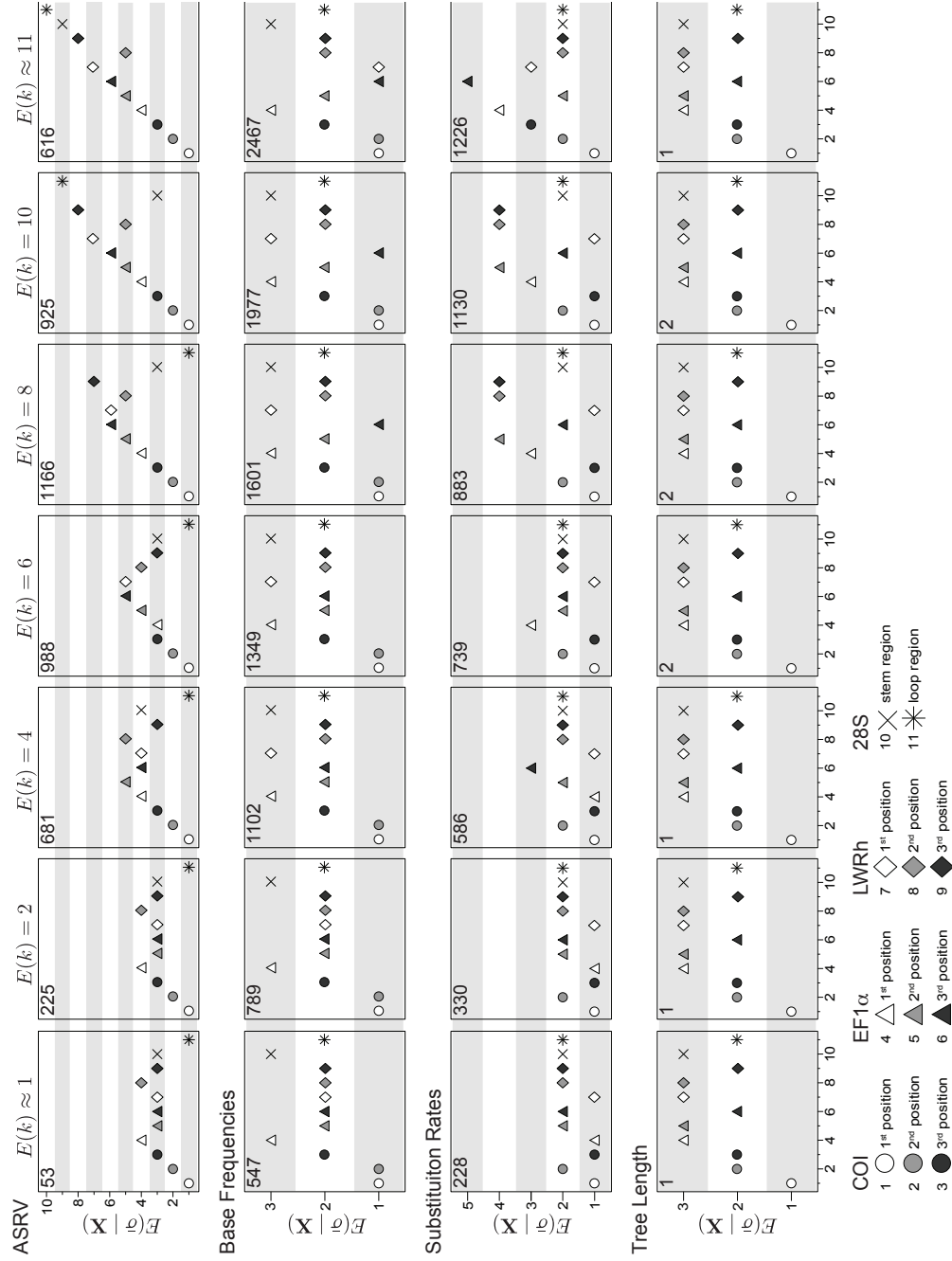
Figure 3: Inferred patterns of process heterogeneity within the gall-wasp sequence alignment inferred under the Dirichlet process prior model. Each row of panels corresponds to a parameter of the model, and each column of panels corresponds to a specified value for the prior mean on number of process partitions in the sequence alignment, $E(k)$. Accordingly, each panel summarizes estimates for a particular parameter under a specified value of $E(k)$. The number of occupied rows within a panel corresponds to the estimated number of process partitions, $E(k \mid \mathbf{X})$, and the allocation of data partitions (depicted as symbols) among the rows of a panel specifies the inferred mean partition scheme, $E(\bar{\sigma} \mid \mathbf{X})$. The number in the upper left of each panel specifies the size of the 95% credible set of partition schemes.
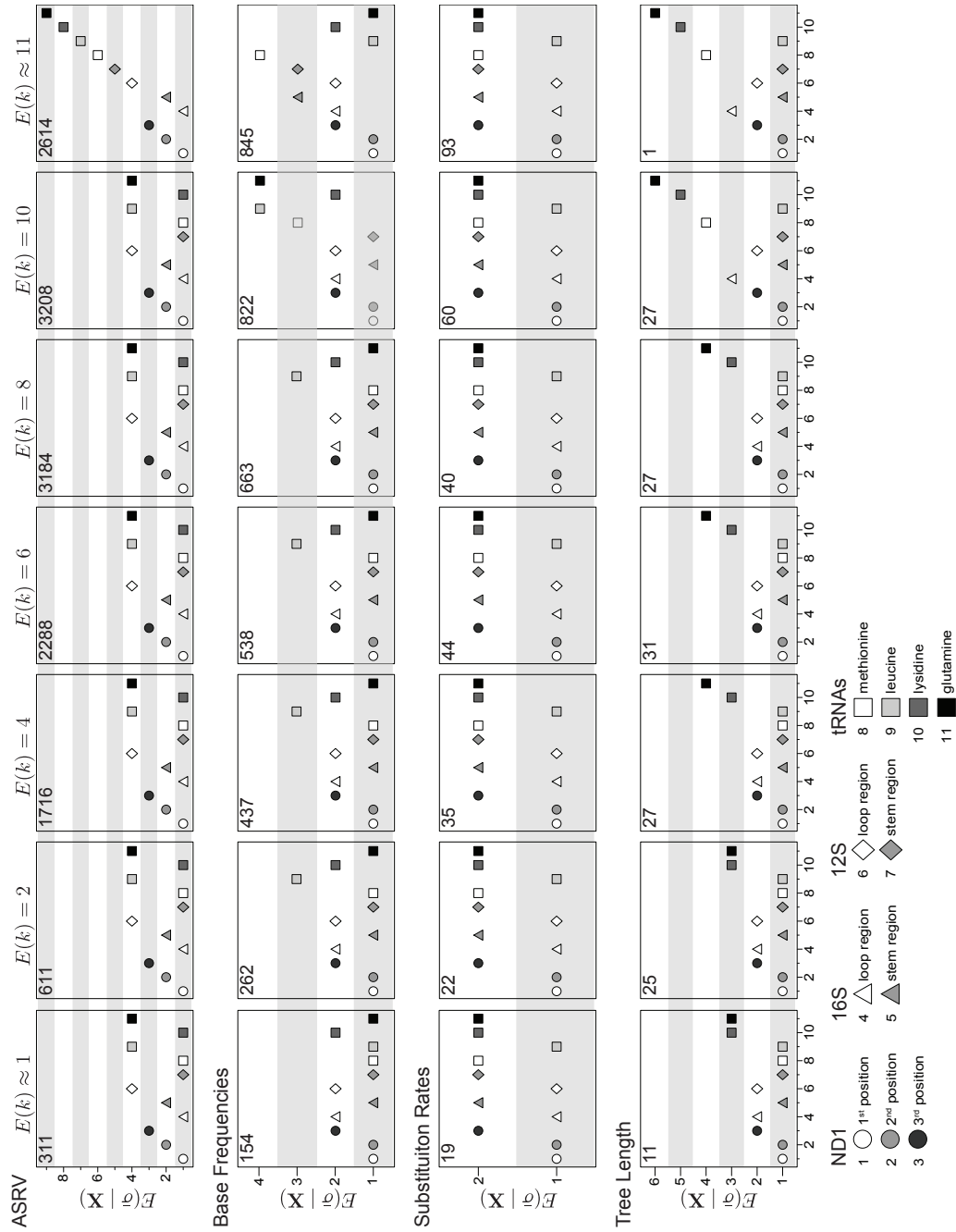
38

Figure 4: Inferred patterns of process heterogeneity within the skink sequence alignment inferred under the Dirichlet process prior model. Graphical conventions as detailed in Figure 3.
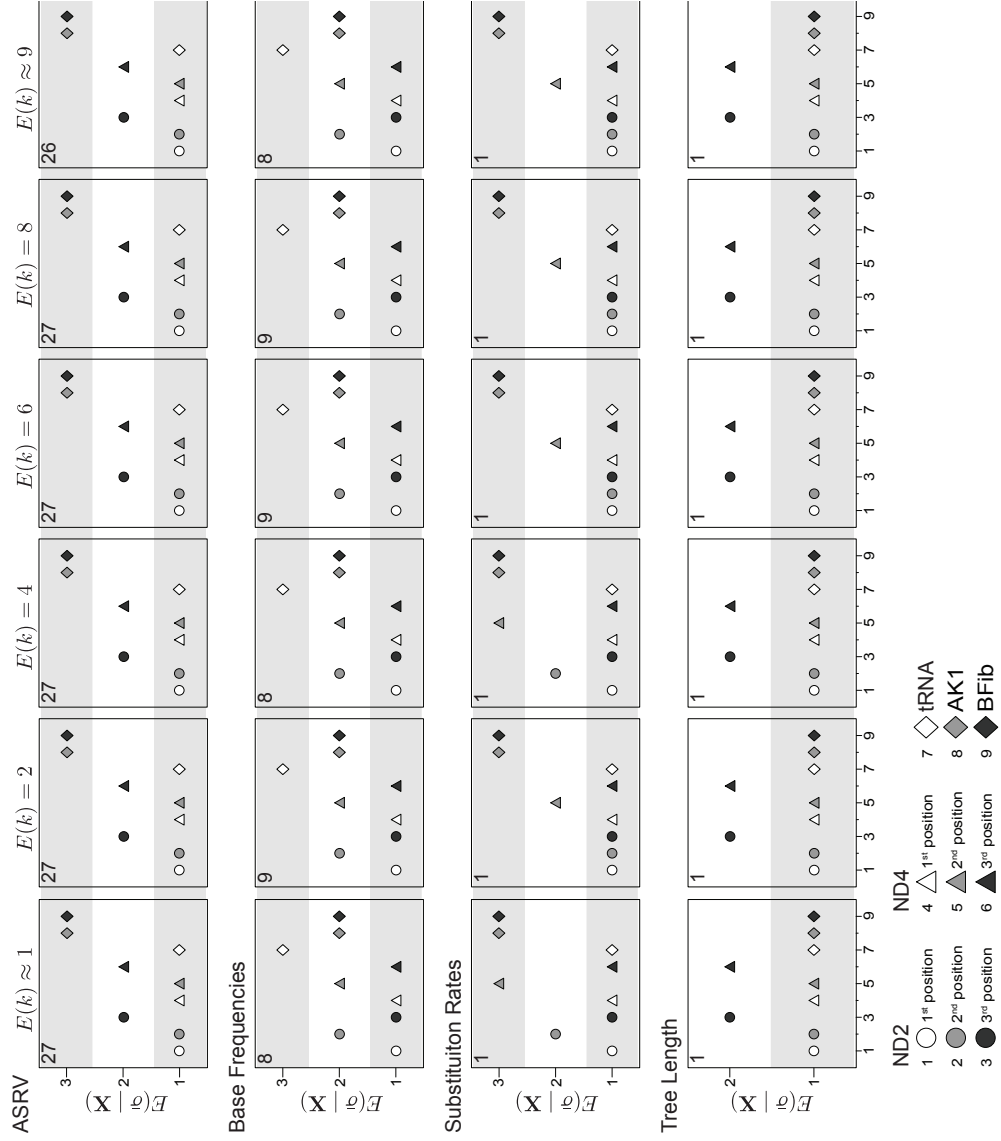
Figure 5: Inferred patterns of process heterogeneity within the hummingbird sequence alignment inferred under the Dirichlet process prior model. Graphical conventions as detailed in Figure 3.