The Annals of Applied Statistics
2017, Vol. 11, No. 2, 1146–1168
DOI: 10.1214/17-AOAS1040

LATERAL TRANSFER IN STOCHASTIC DOLLO MODELS

BY LUKE J. KELLY[1] AND GEOFF K. NICHOLLS

University of Oxford

Lateral transfer, a process whereby species exchange evolutionary traits through nonancestral relationships, is a frequent source of model misspecification in phylogenetic inference. Lateral transfer obscures the phylogenetic signal in the data as the histories of affected traits are mosaics of the overall phylogeny. We control for the effect of lateral transfer in a Stochastic Dollo model and a Bayesian setting. Our likelihood is highly intractable, as the parameters are the solution of a sequence of large systems of differential equations representing the expected evolution of traits along a tree. We illustrate our method on a data set of lexical traits in Eastern Polynesian languages, and obtain an improved fit over the corresponding model without lateral transfer.

1. Introduction. Evolutionary traits used to infer the ancestry of a set of taxa take many forms beside DNA base values in sequence data. For example, Cybis et al. (2015) study the spread of antibiotic resistance in Salmonella strains and Jofré et al. (2017) estimate the shared ancestry of twenty-two stars from measurements on seventeen elements in their chemical composition. In this paper, we infer the shared ancestry of languages from lexical trait data.

When species evolve in isolation, we commonly assume that traits pass vertically from one generation to the next through ancestral relationships. A phylogenetic tree describes the shared ancestry of taxa which evolve in this manner: branches represent evolving species, internal nodes depict speciation events, and leaf nodes correspond to observed taxa. In this paper, we wish to infer the phylogeny of taxa which evolved through a combination of vertical and *lateral* trait transfer. Lateral transfer, such as *horizontal gene transfer* in biology or *borrowing* in linguistics, is an evolutionary process whereby species acquire traits through nonvertical relationships.

Lateral transfer distorts the phylogenetic signal of the speciation events in the data as the histories of affected traits may conflict with the overall taxa phylogeny. Models based solely on vertical trait inheritance are misspecified in this setting, and, in our experience, this model error can result in overly high levels of confidence in poorly fitting trees. In this article, we develop a fully model-based Bayesian method for trait presence/absence data which explicitly accounts for lateral transfer in reconstructing dated phylogenies.

Received January 2016; revised March 2017.

[1]Supported in part by the St. John's College and Engineering and Physical Sciences Research Council partnership award EP/J500495/1.

Key words and phrases. Bayesian phylogenetics, lateral trait transfer, Stochastic Dollo model.

To illustrate our method, we analyze a data set of lexical traits in Eastern Polynesian languages. There have been many previous phylogenetic studies of languages and language families, including Austronesian [Gray, Drummond and Greenhill (2009)], Indo-European [Gray and Atkinson (2003), Nicholls and Gray (2008), Ryder and Nicholls (2011), Bouckaert et al. (2012), Chang et al. (2015)], Linear B [Skelton (2008)] and Semitic [Kitchen et al. (2009)]. Lateral transfer is a frequent occurrence in language diversification [Greenhill, Currie and Gray (2009)], yet a common theme of the above studies is that the authors do not control for it in their fitted models. Typically, known-transferred traits are discarded and a model for vertical trait transfer is fit to the remainder [Gray and Atkinson (2003), Bouckaert et al. (2012); and many others]. This approach is problematic as recently transferred traits are more readily identified, and so earlier transfers remain in the data set.

There exist various methods which test for evidence of lateral transfer in data but do not estimate a phylogeny. Patterson et al. (2012) review various tests for admixture in allele frequency data, while Daubin, Gouy and Perrière (2002), Beiko and Hamilton (2006) and Abby et al. (2010) describe similar tests for sequence data which compare gene trees to a species tree constructed *a priori*. Similarly, internal nodes in implicit phylogenetic networks accommodate incompatibilities in the data with the assumption of an underlying species tree but do not necessarily represent the evolutionary history of the taxa [Huson and Bryant (2006), Oldman et al. (2016)]. Under the assumption of random trait transfer between lineages, Roch and Snir (2013) demonstrate that a number of nonparametric reconstruction methods recover the true phylogeny with high probability when the expected number of transfer events is bounded.

The problem of controlling for lateral transfer in inference for dated phylogenies has received little attention in the statistics literature. In particular, there are few fully likelihood-based inference schemes for dated phylogenies which control for lateral transfer for any data type. Parametric inference for the underlying phylogeny with an explicit model for lateral transfer is a difficult computational problem. This is due to the near intractability of the likelihood calculation, as pruning [Felsenstein (1981)] is no longer directly applicable in integrating over unobserved trait histories. Approximate Bayesian computation, although a useful tool for estimating demographic parameters in complex models [Tavaré et al. (1997)] or selecting a particular tree from a restricted set of alternatives [Veeramah et al. (2015)], does not help here, as a summary statistic which informs a dated phylogeny has to be relatively high dimensional, thereby leading to low acceptance rates in simulation.

Lathrop (1982) and Pickrell and Pritchard (2012) describe methods to infer explicit phylogenetic networks of population splits and instantaneous hybridisation events from allele frequency data. For input gene trees inferred *a priori*, Kubatko (2009) investigates the support for the hybridization events in a given hybrid phylogeny under the multispecies coalescent model [Rannala and Yang (2003)]. Wen,

TABLE 1

Model-based phylogenetic methods which incorporate lateral transfer. The criteria are as follows: (A) *The method infers dated phylogenies controlling for lateral transfer* (*does not require known species phylogeny or gene trees as input*). (B) *The method quantifies uncertainty in parameters*, *tree structure and node times*. (C) *The method uses exact model-based inference* (*up to Monte Carlo error*) *or an explicitly quantified approximation*. (D) *The model is a generative description of the observation process for the data the authors analyze*, *with physically meaningful parameters*. (E) *The approach is directly applicable to our binary Dollo trait data*

		Criteria				
Method	**Input**	**(A)**	**(B)**	**(C)**	**(D)**	**(E)**
Pickrell and Pritchard (2012)	Allele frequencies	✓	✓	✗	✓	✗
Lathrop (1982)	Allele frequencies	✓	✓	✓	✓	✗
Kubatko (2009)	Gene trees, species tree	✗	✓	✓	✓	✗
Szöllosi et al. (2012)	Gene trees	✗	✓	✗	✓	✗
Wen, Yu and Nakhleh (2016)	Gene trees	✗	✓	✓	✓	✗

Yu and Nakhleh (2016) describe a Bayesian method to infer an explicit phylogenetic network under the multispecies coalescent model for the input gene trees.

From a set of input gene trees, Szöllosi et al. (2012) seek the species tree which maximizes the likelihood of reconciling the gene trees under their model incorporating trait gain, loss and lateral transfer. The authors discretize time on the tree, thereby limiting the number of transfer events which may occur and so that their computation is tractable. This allows them to consider many more taxa than Wen, Yu and Nakhleh (2016), for example. In addition, their method returns a time-ordering of the internal nodes rather than a fully dated tree. We summarize these model-based methods in Table 1.

In this paper, we describe a novel method for inferring dated phylogenies from trait presence/absence data. This research is motivated by problems such as the example in Section 8 where we estimate a *language tree* from lexical trait data. These data sets are gathered under a different experimental design to sequence data used to infer gene trees. In collecting trait presence/absence data, we choose a trait and record which taxa display it; the patterns of presence and absence of traits across taxa are informative of the tree. Gene content data is defined similarly [Huson and Steel (2004)]. In contrast, in the design for gene tree data, we choose a gene and sequence homologs of that gene in each individual corresponding to a leaf; a gene is a complex trait and the displayed characters inform the gene tree. In the context of our application in Section 8, it may be tempting to think of lexical traits as genes and languages as biological species. The analogy does not hold as the objects in trait presence/absence data and gene tree data have different meanings due to the different experimental designs. For these and other reasons summarized in Table 1, the model-based methods we cite above are not directly applicable to the problem at hand.

We take the *Stochastic Dollo* model of Nicholls and Gray (2008) (SD) for unordered sets of trait presence/absence data as the starting point for our lateral transfer model. The SD model posits a birth-death process of traits along each branch of the tree, with parent traits copied into offspring at a branching event. The basic process respects *Dollo parsimony*: each trait is born exactly once, and once a trait is extinct, it remains so. Alekseyenko, Lee and Suchard (2008) extend the SD model for multiple character states, and Ryder and Nicholls (2011) introduce missing data and rate heterogeneity. Bouchard-Côté and Jordan (2013) describe a sequence-valued counterpart to the SD model. The SD model has been implemented in the popular phylogenetics software packages `BEAST` [Drummond et al. (2012)] and `BEAST 2` [Bouckaert et al. (2014)]. In a recent study, McPherson et al. (2016) use the SD model to infer cancer clone phylogenies from tumor samples. Simulation studies of the SD model show that topology estimates are robust to moderate levels of random lateral transfer when the underlying topology is balanced [Greenhill, Currie and Gray (2009)] but the root time is typically biased toward the present [Nicholls and Gray (2008), Ryder and Nicholls (2011)].

Nicholls and Gray (2008) describe how to simulate lateral transfer in the basic SD model whereby each species randomly acquires copies of traits from its contemporaries. No previous attempt has been made to fit this model incorporating lateral trait transfer. We perform exact likelihood-based inference under this model. Our lateral transfer process is ultimately defined by the description in Section 3, and we do not attempt to model specific processes such as incomplete lineage sorting, hybridization or gene introgression directly. While our model can generate the trait histories which arise in these processes, it also generates many others and we recommend further case-specific modeling. We do not infer trait trees in advance, then reconcile them to form a species tree; rather, we use Markov chain Monte Carlo methods to sample species trees and parameters, and integrate over all possible trait histories under our model in computing the likelihood. In contrast to Szöllosi et al. (2012), our method operates in a continuous-time setting and we are able to infer the timing of speciation events.

The SD model with lateral transfer will be misspecified for lexical trait data in many ways. Trait birth, death and transfer events will be correlated in complex ways due to real-world processes that we do not model. We are particularly interested in misspecification-induced bias impacting branching time and tree topology estimates. We test for this bias by removing information constraints on known leaf ages and checking that they are correctly reconstructed. This is a test for evidence against the model akin to a pure test for significance in a frequentist setting. These tests demonstrate that *whatever* the misspecification, there is no evidence that it is impacting our estimates. The SD model is a special case of our model and is a natural basis for assessing the effect of controlling for lateral transfer at the expense of an increase in computational cost. We show that the SD model fails these misspecification tests.

To summarize our approach, we build a detailed *ab initio* model of trait and tree dynamics which fully describes the data-observation process. In doing so, we do not compromise the model to make it easier to fit. The price we pay is a massive integration over the unobserved trait histories. In looking for competing methods, we focus on methods which infer dated trees, can quantify the uncertainty in their estimates and perform exact inference or use explicitly quantified approximations. For the lexical trait data we consider, there are no obvious benchmarks among the competing model-based inference schemes discussed above and summarized in Table 1. Our method satisfies each of these criteria.

We describe our binary trait data in Section 2, and introduce our lateral transfer model in Section 3. We describe the likelihood calculation in Section 4, and describe extensions to the model in Section 5. We discuss our inference method in Section 6, and discuss tests to validate our computer implementation in Section 7. We illustrate our model on a data set of lexical traits in Eastern Polynesian languages in Section 8, and conclude in Section 9 with a discussion of the model and possible directions for future research.

2. Homologous trait data. Homologous traits derive from a common ancestral trait through a combination of vertical inheritance and lateral transfer events. We assign each set of homologous traits a unique common label from the set of trait labels, $\mathcal{Z}$. A set of trait categories is chosen and, for each taxon in the study, we gather instances of traits in each category. We record the status of trait h in taxon i as

$$d_i^h = \begin{cases} 1, & \text{trait } h \text{ is present in taxon } i, \\ 0, & \text{trait } h \text{ is absent in taxon } i, \\ ?, & \text{the status of trait } h \text{ in taxon } i \text{ is unknown.} \end{cases}$$

We denote by $\mathbf{D}$ the array recording the status of each trait across the observed taxa. A column $\mathbf{d}^h$ of $\mathbf{D}$ is a *site-pattern* recording the status of trait h across the taxa. These patterns of trait presence and absence, which we assume are independent, exchangeable entities, shall form the basis of our model.

In the analysis in Section 8, each trait is a word in one of 210 meaning categories and each taxon is an Eastern Polynesian language. For example, the Maori and Hawaiian words for *woman* and *wife*, both *wahine*, derive from a common ancestor h, say, and so $d^h_{\text{Maori}} = d^h_{\text{Hawaiian}} = 1$. On the other hand, the Maori word for *mother*, *whaea*, is not related to its Hawaiian counterpart, *makuahine*, and so we record zeros in the respective entries of the data array.

3. Generative model. A branching process on sets of traits determines the phylogeny of the observed taxa. The set contents diversify according to a process of trait birth, death and lateral transfer events. We describe these events in greater detail below. Figure 1 depicts a realization of the model and the history of a single

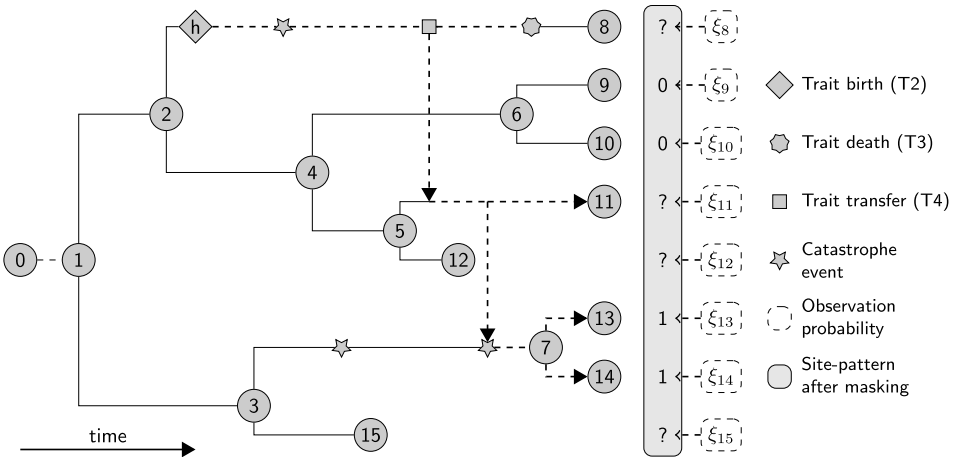


FIG. 1. *Illustration of the Stochastic Dollo with lateral transfer model. Dashed lines represent the history of a trait h. We describe catastrophes, missing data and offset leaves in Section 5.*

trait. The trait history bears little resemblance to the underlying phylogeny as a consequence of trait death and lateral transfer events.

We first define our model and inference method in terms of binary patterns of trait presence and absence in taxa which are recorded simultaneously. In Section 5, we extend the model to incorporate missing data and different leaf sampling times.

A rooted phylogenetic tree $g = (V, E, T)$ on L leaves is a connected, acyclic graph with node set $V = \{0, 1, \ldots, 2L-1\}$, directed edge set E and node times $T \in \{-\infty\} \times \mathbb{R}^{2L-1}$. The node set V comprises one *Adam* node labeled 0 of degree 1, the internal nodes $V_A = \{1, 2, \ldots, L-1\}$ of degree 3, and the leaf nodes $V_L = \{L, L+1, \ldots, 2L-1\}$ of degree 1. Node $i \in V$ arises at time $t_i \in T$, denoting when the corresponding event occurred relative to the current time, 0. For convenience, we label the internal nodes V_A in such a way that $t_1, \ldots, t_{L-1}$ is a strictly increasing sequence of node times. We observe the taxa simultaneously at time 0, the present, and constrain $t_i = 0$ for each leaf $i \in V_L$ as a result.

Edges represent evolving species and are directed forward in time. We label each edge by its offspring node: if $\mathrm{pa}(i)$ denotes the parent of node $i \in V \setminus \{0\}$, then edge $i \in E$ runs from node $\mathrm{pa}(i)$ at time $t_{\mathrm{pa}(i)}$ to i at time t_i. We assume that the Adam node arose at time $t_0 = -\infty$, and so a branch of infinite length connects it to the *root* node 1 at time t_1. If we slice the tree at time t, then there are $L^{(t)}$ species labeled $\mathbf{k}^{(t)} = (i \in E : t_{\mathrm{pa}(i)} \leq t < t_i)$. In Figure 1, there are $L^{(t_2)} = 3$ species labeled $\mathbf{k}^{(t_2)} = (8, 4, 3)$ immediately after the speciation event at time t_2, for example.

Let $H_i(t) \subset \mathcal{Z}$ denote the set of traits possessed by species $i \in \mathbf{k}^{(t)}$ at time t. We now define four properties of the set-valued evolutionary process $H(t) = \{H_i(t) : i \in \mathbf{k}^{(t)}\}$ for $t \in (-\infty, 0]$.

PROPERTY T1 (Set branching event). Species $i \in \mathbf{k}^{(t_i^-)}$ branches at time t_i and is replaced by two identical offspring, j and $k \in \mathbf{k}^{(t_i)}$,

$$H_j(t_i) \leftarrow H_i(t_i^-),$$
$$H_k(t_i) \leftarrow H_i(t_i^-),$$

where t_i^- denotes the time just before the branching event.

PROPERTY T2 (Trait birth). New traits are born at rate λ over time in each extant species. If trait $h \in \mathcal{Z}$ is born in species i at time t, then

$$H_i(t) \leftarrow H_i(t^-) \cup \{h\}.$$

PROPERTY T3 (Trait death). A species kills off each trait it possesses independently at rate μ. If trait $h \in H_i(t^-)$ in species i dies at time t, then

$$H_i(t) \leftarrow H_i(t^-) \setminus \{h\}.$$

PROPERTY T4 (Lateral trait transfer). Each instance of a trait attempts to transfer at rate β. Equivalently, a species acquires a copy of a trait by lateral transfer at rate β scaled by the fraction of extant species which possess it. If species i acquires a copy of trait $h \in \mathcal{H}^{(t^-)} = \bigcup_{i \in \mathbf{k}^{(t^-)}} H_i(t^-)$ at time t, then

$$H_i(t) \leftarrow H_i(t^-) \cup \{h\}.$$

Clearly, if $h \in H_i(t^-)$, then the transfer event has no effect.

Starting from a single set $H(-\infty) = \{\varnothing\}$, the process $H(t)$ evolves as a continuous-time Markov chain through a combination of branching (T1) and trait (T2–T4) events to yield the diverse set of taxa $H(0) = \{H_i(0) : i \in V_L\}$ that we observe at time 0. When the lateral transfer rate $\beta = 0$, we recover the binary Stochastic Dollo process of Nicholls and Gray (2008).

4. Likelihood calculation. We may calculate the likelihood of a given trait history in terms of independent holding times and jumps between states (T1–T4). However, trait histories are nuisance parameters here as we are interested in the overall phylogeny, and so we must integrate them out of the model likelihood. Furthermore, we must account for the histories of traits born on the tree which did not survive into the taxa. In order to describe how to simultaneously integrate over all possible trait histories on the tree under our model, we now recast the trait process in terms of evolving patterns of presence and absence across branches.

4.1. *Pattern evolution.* If we cut through the tree at time t, each trait in $\mathcal{H}^{(t)}$ displays a *pattern* of presence and absence across the $L^{(t)}$ extant species $\mathbf{k}^{(t)} = (k_i^{(t)} : i \in [L^{(t)}])$, where $[L^{(t)}] = \{1, \dots, L^{(t)}\}$. These patterns of presence and absence evolve over time as new branches arise and instances of traits die and transfer. The pattern displayed by trait $h \in \mathcal{H}^{(t)}$ at time t is $\mathbf{p}^h(t) = (p_i^h(t) : i \in [L^{(t)}])$, where

$$p_i^h(t) = \begin{cases} 1, & h \in H_{k_i^{(t)}}(t), \\ 0, & \text{otherwise}, \end{cases}$$

indicates the presence or absence of trait h on lineage $k_i^{(t)}$ at time t.

The space of binary patterns of trait presence and absence across $L^{(t)}$ lineages is $\mathcal{P}^{(t)} = \{0,1\}^{L^{(t)}} \setminus \{\mathbf{0}\}$, where $\mathbf{0}$ denotes an $L^{(t)}$-tuple of zeros. Trait labels are exchangeable and there are $N_{\mathbf{p}}(t) = |\{h \in \mathcal{H}^{(t)} : \mathbf{p}^h(t) = \mathbf{p}\}|$ traits displaying pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ at time t. The dynamics of the pattern frequency process $\mathbf{N}(t) = (N_{\mathbf{p}}(t) : \mathbf{p} \in \mathcal{P}^{(t)})$ follow directly from Properties T1–T4 of the trait process in Section 3.

4.1.1. *Patterns at branching events.* At a branching event, patterns gain an entry and the space of patterns increases accordingly. The tuple $\mathbf{k}^{(t)}$ of branch labels is consistent across speciation events in the sense that when lineage $k_i^{(t_j^-)}$ branches at time t_j,

$$\mathbf{k}^{(t_j^-)} \to \mathbf{k}^{(t_j)} = \big(k_1^{(t_j^-)}, \dots, k_{i-1}^{(t_j^-)}, k_i^{(t_j)}, k_{i+1}^{(t_j)}, k_{i+1}^{(t_j^-)}, \dots, k_{L^{(t_j^-)}}^{(t_j^-)}\big),$$

where species $k_i^{(t_j)}$ and $k_{i+1}^{(t_j)}$ are the offspring of species $k_i^{(t_j^-)}$ (T1). It follows that each trait $h \in \mathcal{H}^{(t_j^-)}$ transitions to display a pattern $\mathbf{p}^h(t_j)$ with entries $p_i^h(t_j) = p_{i+1}^h(t_j) \leftarrow p_i^h(t_j^-)$. For example, reading from top to bottom in Figure 1,

$$\mathbf{k}^{(t_4^-)} = (8,4,7,15), \qquad \mathbf{k}^{(t_4)} = (8,6,5,7,15),$$
$$\mathbf{p}^h(t_4^-) = (1,0,0,0), \qquad \mathbf{p}^h(t_4) = (1,0,0,0,0),$$

as a result of the speciation event at node 4.

A pattern $\mathbf{p} \in \mathcal{P}^{(t_j)}$ with entries $p_i = p_{i+1}$ is consistent with the branching event on lineage $k_i^{(t_j^-)}$ as it may be formed by duplicating the ith entries of a pattern in $\mathcal{P}^{(t_j^-)}$. On the other hand, the trait process cannot generate a pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ with $p_i \neq p_{i+1}$ at time t_j by definition (T1). We denote by $\mathbf{T}^{(j)} : \mathbf{N}(t_j^-) \to \mathbf{N}(t_j)$ the operation which initializes the pattern frequencies $\mathbf{N}(t_j)$ with entries of $\mathbf{N}(t_j^-)$ for patterns consistent with the branching event, and zeros otherwise. We return to this initialization operation when we compute the expected pattern frequencies in Section 4.2.

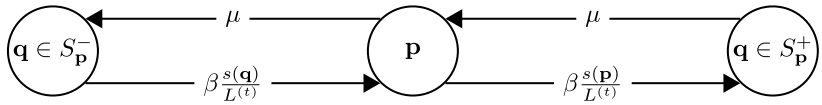


FIG. 2. *Transition rates between pattern states* $\mathbf{p} \in \mathcal{P}^{(t)}$ *and* $\mathbf{q} \in S_{\mathbf{p}}^- \cup S_{\mathbf{p}}^+$.

4.1.2. *Patterns between branching events.* In order to formally describe the Markovian evolution of the pattern frequencies $\mathbf{N}(t)$ between branching events, we first define how patterns relate to each other. The Hamming distance between patterns $\mathbf{p}$ and $\mathbf{q} \in \mathcal{P}^{(t)}$ is $d(\mathbf{p}, \mathbf{q}) = |\{i \in [L^{(t)}] : p_i \neq q_i\}|$, and $s(\mathbf{p}) = d(\mathbf{p}, \mathbf{0})$ is the Hamming weight of $\mathbf{p}$. A trait displaying pattern $\mathbf{p}$ at time t *communicates* with patterns in the sets

$$S_{\mathbf{p}}^- = \{\mathbf{q} \in \mathcal{P}^{(t)} : s(\mathbf{q}) = s(\mathbf{p}) - 1, d(\mathbf{p}, \mathbf{q}) = 1\},$$

$$S_{\mathbf{p}}^+ = \{\mathbf{q} \in \mathcal{P}^{(t)} : s(\mathbf{q}) = s(\mathbf{p}) + 1, d(\mathbf{p}, \mathbf{q}) = 1\},$$

the patterns which differ from $\mathbf{p}$ through a single trait death (T3) or transfer (T4) event, respectively. Figure 2 describes the transition rates between pattern states $\mathbf{p}$ and $\mathbf{q} \in S_{\mathbf{p}}^- \cup S_{\mathbf{p}}^+$. New traits displaying patterns of Hamming weight 1 arise on each branch through trait birth events (T2). For example, reading from top to bottom in Figure 1, a copy of trait h transfers at time t from branch $k_1^{(t^-)} = 1$ to $k_3^{(t)} = 11$, and so

$$\mathbf{p}^h(t^-) = (1,0,0,0,0,0), \qquad \mathbf{p}^h(t) = (1,0,1,0,0,0) \in S_{100000}^+,$$

$$N_{100000}(t) = N_{100000}(t^-) - 1, \qquad N_{101000}(t) = N_{101000}(t^-) + 1.$$

4.2. *Expected pattern frequencies.* Instances of the same trait evolve independently of each other and of other traits. If we sum over the rates in Figure 2 for each trait displaying a given pattern $\mathbf{p} \in \mathcal{P}^{(t)}$, then on a short interval of length $\mathrm{d}t$ between branching events, by a standard argument for Markov chains,

$$\begin{aligned}
&\mathbb{P}\big[N_{\mathbf{p}}(t+\mathrm{d}t) - N_{\mathbf{p}}(t) = k | g, \lambda, \mu, \beta\big] \\
&\quad = \begin{cases}
s(\mathbf{p})\left[\mu + \beta\left(1 - \dfrac{s(\mathbf{p})}{L^{(t)}}\right)\right] N_{\mathbf{p}}(t)\,\mathrm{d}t + o(\mathrm{d}t), & k = -1, \\
\left[\lambda \mathbf{1}_{\{s(\mathbf{p})=1\}} + \beta \displaystyle\sum_{\mathbf{q} \in S_{\mathbf{p}}^-} \frac{s(\mathbf{q})}{L^{(t)}} N_{\mathbf{q}}(t) \right. & \\
\qquad \left. + \mu \displaystyle\sum_{\mathbf{q} \in S_{\mathbf{p}}^+} N_{\mathbf{q}}(t)\right] \mathrm{d}t + o(\mathrm{d}t), & k = 1.
\end{cases}
\end{aligned} \tag{4.1}$$

Let $x_{\mathbf{p}}(t) = x_{\mathbf{p}}(t; g, \lambda, \mu, \beta) = \mathbb{E}[N_{\mathbf{p}}(t) | g, \lambda, \mu, \beta]$, the expected number of traits in $\mathcal{H}^{(t)}$ displaying pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ at time t. From Equation (4.1), $x_{\mathbf{p}}(t)$ evolves

according to the following differential equation:

$$\begin{aligned}\dot{x}_{\mathbf{p}}(t) &= \lim_{\mathrm{d}t\to 0} \frac{\mathbb{E}[N_{\mathbf{p}}(t+\mathrm{d}t) - N_{\mathbf{p}}(t) | g, \lambda, \mu, \beta]}{\mathrm{d}t} \\ &= -s(\mathbf{p})\left[\mu + \beta\left(1 - \frac{s(\mathbf{p})}{L^{(t)}}\right)\right] x_{\mathbf{p}}(t) + \lambda \mathbf{1}_{\{s(\mathbf{p})=1\}} \\ &\quad + \beta \sum_{\mathbf{q}\in S_{\mathbf{p}}^-} \frac{s(\mathbf{q})}{L^{(t)}} x_{\mathbf{q}}(t) + \mu \sum_{\mathbf{q}\in S_{\mathbf{p}}^+} x_{\mathbf{q}}(t).\end{aligned} \tag{4.2}$$

There are $|\mathcal{P}^{(t)}| = 2^{L^{(t)}} - 1$ coupled differential equations (4.2) describing the expected evolution of the pattern frequencies $\mathbf{N}(t)$. We may write these equations as $\dot{\mathbf{x}}(t) = \mathbf{A}^{(t)}\mathbf{x}(t) + \mathbf{b}^{(t)}$, where $\mathbf{x}(t) = (x_{\mathbf{p}}(t) : \mathbf{p} \in \mathcal{P}^{(t)})$ is the vector of expected pattern frequencies at time t, and the sparse matrix $\mathbf{A}^{(t)}$ and vector $\mathbf{b}^{(t)}$ respectively describe the flow between patterns from trait death and transfer events and the flow into patterns of Hamming weight 1 through trait birth events.

In Section 3, we state that a branch of infinite length connects the Adam and root nodes. As a result, the pattern frequency process $\mathbf{N}(t)$ is in equilibrium just before the first branching event at time t_1, with the result that $\mathbf{x}(t_1^-) = x_1(t_1^-) = \lambda/\mu$ and $N_1(t_1^-) \sim \text{Poisson}(\lambda/\mu)$. With this initial condition at the root, we can write the expected pattern frequencies at the leaves, $\mathbf{x}(0)$, recursively as a sequence of initial value problems between branching events: for each interval $i = 1, \ldots, L-1$, solve

$$\dot{\mathbf{x}}(t) = \mathbf{A}^{(t)}\mathbf{x}(t) + \mathbf{b}^{(t)} \qquad \text{for } t \in [t_i, t_{i+1}) \text{ where } \mathbf{x}(t_i) = \mathbf{T}^{(i)}\mathbf{x}(t_i^-), \tag{4.3}$$

and we recall from Section 4.1.1 the operator $\mathbf{T}^{(i)}$ which propagates $\mathbf{N}(t^-)$ and $\mathbf{x}(t^-)$ across the ith branching event. We illustrate this procedure graphically in Figure 3.

4.3. *Likelihood.* Theorem 1 describes the distribution of the pattern frequencies. We prove this result in the Supplemental Materials [Kelly and Nicholls (2017)].

THEOREM 1 (Binary data distribution). *The components of the vector of pattern frequencies* $\mathbf{N}(t) = (N_{\mathbf{p}}(t) : \mathbf{p} \in \mathcal{P}^{(t)})$ *are independent Poisson random variables with corresponding rate parameters* $\mathbf{x}(t; g, \lambda, \mu, \beta)$ *given by the solution of the sequence of initial value problems in equation* (4.3).

5. Model extensions. We now extend the model and likelihood calculation to allow for rate variation, missing data, offset leaves and the systematic removal of patterns from the data.

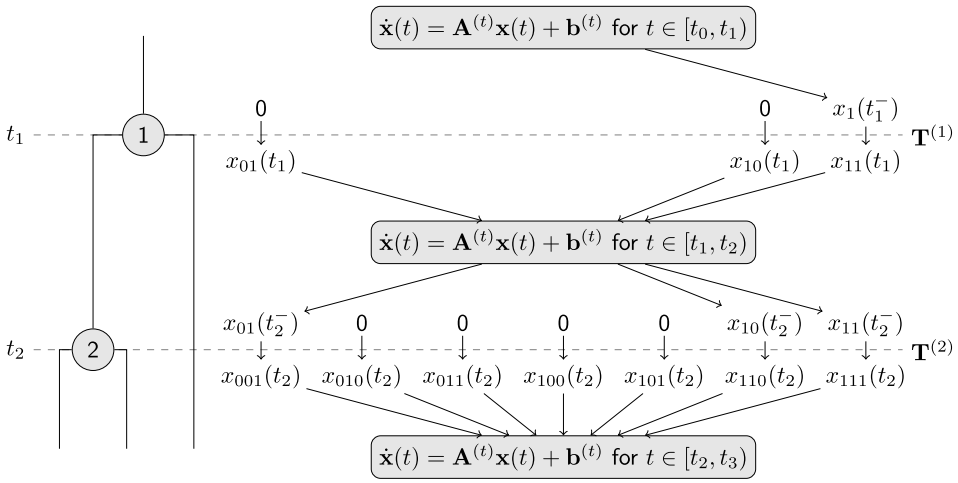


FIG. 3. *Computing the expected pattern frequencies* $\mathbf{x}(t)$ *as a sequence of initial value problems* (4.3) *on a given tree. The initialization operation* $\mathbf{x}(t_i) = \mathbf{T}^{(i)}\mathbf{x}(t_i^-)$ *from Section* 4.1.1 *provides the initial condition at the start of the* i*th interval between branching events.*

5.1. *Rate heterogeneity.* We introduce spikes of evolutionary activity in the form of *catastrophes* [Ryder and Nicholls (2011)]. Catastrophes, illustrated in Figure 1, occur at rate ρ along each branch of the tree. A catastrophe advances the trait process along a branch by $\delta = -\mu^{-1}\log(1-\kappa)$ units of time relative to the other branches. In the model of Ryder and Nicholls (2011), this is equivalent to killing each trait on the branch independently with probability κ and adding a Poisson$(\lambda\kappa/\mu)$ number of new traits. To ensure that catastrophes are identifiable with respect to the underlying trait process, we enforce a minimum *catastrophe severity* $\kappa \geq 0.25$.

A branch may acquire traits through birth and transfer events, and lose traits to death events during a catastrophe. The trait process at a catastrophe is equivalent to thinning the overall trait process to events on a single branch. As a result, we account for a catastrophe at time t on branch $k_i^{(t)}$ in the expected pattern frequency calculation (4.3) with the update

$$x_{\mathbf{p}}(t) \leftarrow e^{-\mu\delta}x_{\mathbf{p}}(t^-) + (1-e^{-\mu\delta})\frac{\lambda}{\mu}, \qquad \begin{array}{l}\mathbf{p} \in \mathcal{P}^{(t)},\\ s(\mathbf{p}) = 1,\, p_i = 1,\end{array}$$

$$\begin{bmatrix} x_{\mathbf{q}}(t) \\ x_{\mathbf{r}}(t) \end{bmatrix} \leftarrow \exp\left[\begin{pmatrix} -\beta\dfrac{s(\mathbf{q})}{L^{(t)}} & \mu \\ \beta\dfrac{s(\mathbf{q})}{L^{(t)}} & -\mu \end{pmatrix}\delta\right]\begin{bmatrix} x_{\mathbf{q}}(t^-) \\ x_{\mathbf{r}}(t^-) \end{bmatrix}, \qquad \begin{array}{l}\mathbf{q}, \mathbf{r} \in \mathcal{P}^{(t)}, d(\mathbf{q}, \mathbf{r}) = 1,\\ q_i = 0, r_i = 1,\end{array}$$

where we exploit the property that each pattern communicates with at most one other during a catastrophe.

5.2. *Missing data*. We allow for *missing-at-random* data. Following Ryder and Nicholls (2011), the true binary state of trait h at taxon $i \in V_L$ is recorded with probability $\xi_i = \mathbb{P}(d_i^h \in \{0,1\})$ independently of the other traits and taxa. Let $\Xi = (\xi_i : i \in V_L)$ denote the set of true-state observation probabilities. The space of observable site-patterns with missing data across the L taxa at time 0 is $\mathcal{Q} = \{0,1,?\}^L \setminus \{\mathbf{0}\}$. The set of binary patterns consistent with pattern $\mathbf{q} \in \mathcal{Q}$ is $u(\mathbf{q}) = \{\mathbf{p} \in \mathcal{P}^{(0)} : p_i = q_i \text{ if } q_i \neq ?, i \in [L]\}$. From Theorem 1 and the restriction and superposition properties of Poisson processes [Kingman (1993)], the frequency of traits displaying pattern $\mathbf{q}$ is an independent Poisson random variable with mean

$$x_{\mathbf{q}}(0; g, \lambda, \mu, \beta, \Xi) = \sum_{\mathbf{p} \in u(\mathbf{q})} x_{\mathbf{p}}(0; g, \lambda, \mu, \beta) \prod_{i=1}^{L} \xi_{k_i^{(0)}}^{\mathbf{1}_{\{q_i \in \{0,1\}\}}} (1 - \xi_{k_i^{(0)}})^{\mathbf{1}_{\{q_i = ?\}}}.$$

5.3. *Nonisochronous data*. Nonisochronous data arise when taxa are sampled at different times. The corresponding taxa appear as *offset* leaves in the phylogeny, nodes 12 and 15 in Figure 1, for example. Similar to catastrophes, the trait process is frozen on offset leaves and a pattern may now only communicate with those patterns which are identical to it on the extinct lineages and differ at a single entry on the extant lineages.

The $L^{(t)}$ extinct and evolving lineages at time t, of which $\hat{L}^{(t)}$ are extant, are labeled $\mathbf{k}^{(t)} = (i \in E : t_{\mathrm{pa}(i)} \leq t < t_i \mathbf{1}_{\{i \in V_A\}})$. The Hamming distance between patterns $\mathbf{p}$ and $\mathbf{q} \in \mathcal{P}^{(t)}$ across the extant lineages only is $\hat{d}(\mathbf{p}, \mathbf{q}) = |\{i \in [L^{(t)}] : p_i \neq q_i, t < t_{k_i^{(t)}}\}|$, and the corresponding Hamming weight of $\mathbf{p}$ across the extant lineages is $\hat{s}(\mathbf{p}) = \hat{d}(\mathbf{p}, \mathbf{0})$. Recalling $S_{\mathbf{p}}^-$ and $S_{\mathbf{p}}^+$ from Section 4.1.2, pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ communicates with patterns in the sets

$$\hat{S}_{\mathbf{p}}^- = \{\mathbf{q} \in S_{\mathbf{p}}^- : \hat{s}(\mathbf{q}) = \hat{s}(\mathbf{p}) - 1, \hat{d}(\mathbf{p}, \mathbf{q}) = 1\},$$

$$\hat{S}_{\mathbf{p}}^+ = \{\mathbf{q} \in S_{\mathbf{p}}^+ : \hat{s}(\mathbf{q}) = \hat{s}(\mathbf{p}) + 1, \hat{d}(\mathbf{p}, \mathbf{q}) = 1\},$$

and its expected frequency evolves as

$$\dot{x}_{\mathbf{p}}(t) = -\hat{s}(\mathbf{p}) \left[\mu + \beta \left(1 - \frac{\hat{s}(\mathbf{p})}{\hat{L}^{(t)}}\right)\right] x_{\mathbf{p}}(t) + \lambda \mathbf{1}_{\{s(\mathbf{p}) = \hat{s}(\mathbf{p}) = 1\}}$$

$$+ \beta \sum_{\mathbf{q} \in \hat{S}_{\mathbf{p}}^-} \frac{\hat{s}(\mathbf{q})}{\hat{L}^{(t)}} x_{\mathbf{q}}(t) + \mu \sum_{\mathbf{q} \in \hat{S}_{\mathbf{p}}^+} x_{\mathbf{q}}(t).$$

We allow for offset leaves in our goodness-of-fit tests in Section 8.

5.4. *Data registration*. Patterns which may be uninformative or unreliable with respect to the model are typically removed from the data. Given a registration rule R, which may be a composition of other simpler rules such as those in Table 2,

TABLE 2
Registration rules of Alekseyenko, Lee and Suchard (2008) and Ryder and Nicholls (2011)

Unregistered traits	Unregistered patterns $\mathcal{Q} \setminus R(\mathcal{Q})$
Absent in taxon $k_i^{(0)}$	$\{\mathbf{q} \in \mathcal{Q} : q_i = 0\}$
Observed in j taxa or fewer	$\{\mathbf{q} \in \mathcal{Q} : \lvert\{i \in [L] : q_i = 1\}\rvert \leq j\}$
Observed in j or more taxa	$\{\mathbf{q} \in \mathcal{Q} : \lvert\{i \in [L] : q_i = 1\}\rvert \geq j\}$
Potentially present in j taxa or greater	$\{\mathbf{q} \in \mathcal{Q} : \lvert\{i \in [L] : q_i \neq 0\}\rvert \geq j\}$

we discard the columns in the data array $\mathbf{D}$ not satisfying R, leaving the registered data $R(\mathbf{D})$, and restrict our analyses to patterns in $R(\mathcal{Q})$. In Section 8, we discard traits not marked present in a single taxon.

6. Bayesian inference. In order to efficiently estimate both the node times and the rate parameters, we calibrate the space Γ of rooted phylogenetic trees on L taxa with *clade constraints*. The constraint $\Gamma^{(0)} = \{g \in \Gamma : \underline{t}_1 \leq t_1 < 0\}$ restricts the earliest admissible root time to $\underline{t}_1$. Each additional constraint $\Gamma^{(c)}$ places either time or ancestry constraints on the remaining nodes. We denote by $\Gamma^C = \bigcap_c \Gamma^{(c)}$ the space of phylogenies satisfying the clade constraints.

Nicholls and Ryder (2011) describe a prior distribution on trees with the property that the root time t_1 is approximately uniformly distributed across a specified interval $[\underline{t}_1, \bar{t}_1]$. For a given tree $g = (V, E, T, C)$, there are $Z(g)$ possible time orderings of the nodes among the admissible node times $T(g) = \{T' : (V, E, T', C) \in \Gamma^C\}$. For each node $i \in V$, $\underline{t}_i = \inf_{T \in T(g)} t_i$ and $\bar{t}_i = \sup_{T \in T(g)} t_i$ are the earliest and most recent times that i may achieve in an admissible tree with topology (V, E). If $S(g) = \{i \in V : \underline{t}_i = \underline{t}_1\}$ denotes the set of *free* internal nodes with times bounded below by $\underline{t}_1$, then the prior with density

$$f_G(g) \propto \frac{\mathbf{1}_{\{g \in \Gamma^C\}}}{Z(g)} \prod_{i \in S(g)} \frac{\underline{t}_1 - \bar{t}_i}{t_1 - \bar{t}_i},$$

is approximately uniform across topologies and root times provided that $\underline{t}_1 \ll \min_{i \in V \setminus S} \underline{t}_i$ [Ryder and Nicholls (2011)]. Uniform priors on offset leaf times complete our prior specification on the tree. Heled and Drummond (2012) describe an exact method for computing uniform calibrated tree priors, but we do not pursue that approach here. Table 3 lists the prior distributions on the remaining parameters.

Inspecting the solution of the expected pattern frequency calculation (4.3) with initial condition $\mathbf{x}(t_1^-) = \lambda/\mu$ at the root, we see that $\mathbf{x}(t; g, \lambda, \ldots) = \lambda \mathbf{x}(t; g, 1, \ldots)$. We can integrate λ out of the Poisson likelihood in Theorem 1 with

TABLE 3
Prior distributions on parameters in the Stochastic Dollo and Stochastic Dollo with lateral transfer models

Parameter	**Prior**	**Reasoning**
Trait birth rate	$\lambda \sim 1/\lambda$	Improper, scale invariant
Trait death rate	$\mu \sim \Gamma(10^{-3}, 10^{-3})$	Approximately $1/\mu$
Trait transfer rate	$\beta \sim \Gamma(10^{-3}, 10^{-3})$	Approximately $1/\beta$
Catastrophe rate	$\rho \sim \Gamma(1.5, 5 \times 10^3)$	$\mathbb{E}[\rho^{-1}] = 10^4$ years
Catastrophe severity	$\kappa \sim \mathrm{U}[0.25, 1]$	$\mathbb{E}[\delta\|\mu] = \mu^{-1}[1 - \log(0.75)]$ years
Observation probabilities	$\Xi \sim \mathrm{U}[0, 1]^L$	Independent, uniform

respect to its prior in Table 3 to obtain a multinomial likelihood whereby a pattern $\mathbf{p} \in R(\mathcal{Q})$ is observed with probability proportional to its expected frequency. Furthermore, we may integrate the catastrophe rate ρ out of the Poisson prior on the number of catastrophes $|C|$ to obtain a Negative Binomial prior instead. We describe these steps in detail in the Supplemental Materials.

Let $n_{\mathbf{p}} = |\{h \in \mathcal{H}(0) : \mathbf{p} = \mathbf{d}^h \in R(\mathbf{D})\}|$ denote the frequency of traits in the registered data displaying pattern $\mathbf{p} \in R(\mathcal{Q})$. Putting everything together, the posterior distribution is

$$\pi\big(g, \mu, \beta, \kappa, \Xi | R(\mathbf{D})\big) \propto f_G(g) f_M(\mu) f_B(\beta) \prod_{\mathbf{p} \in R(\mathcal{Q})} \left(\frac{x_{\mathbf{p}}}{\sum_{\mathbf{q} \in R(\mathcal{Q})} x_{\mathbf{q}}} \right)^{n_{\mathbf{p}}}, \tag{6.1}$$

where the expected pattern frequencies $\mathbf{x} \equiv \mathbf{x}(0; g, 1, \mu, \beta, \kappa, \Xi)$ (4.3) account for catastrophes, missing data and offset leaves where necessary. This completes the specification of the Stochastic Dollo with Lateral Transfer (SDLT) model.

The posterior distribution (6.1) is intractable but may be explored using standard Markov chain Monte Carlo (MCMC) sampling schemes for phylogenetic trees and Stochastic Dollo models [Nicholls and Gray (2008), Ryder and Nicholls (2011)]. We describe the MCMC transition kernels for moves particular to the SDLT model in the Appendix.

Implementation. Code to implement the SDLT model in the software package `TraitLab` [Nicholls, Ryder and Welch (2013)] is available from the authors.

7. Method testing. We describe a number of tests to validate our model and inference scheme in the Supplemental Materials. We compare the exact and empirical distributions of synthetic data to validate our implementation of the expected pattern frequency calculation (4.3). We test the identifiability of the SDLT model, its consistency with the SD model when the lateral transfer rate $\beta = 0$, and its robustness to a common form of model misspecification whereby recently trans-

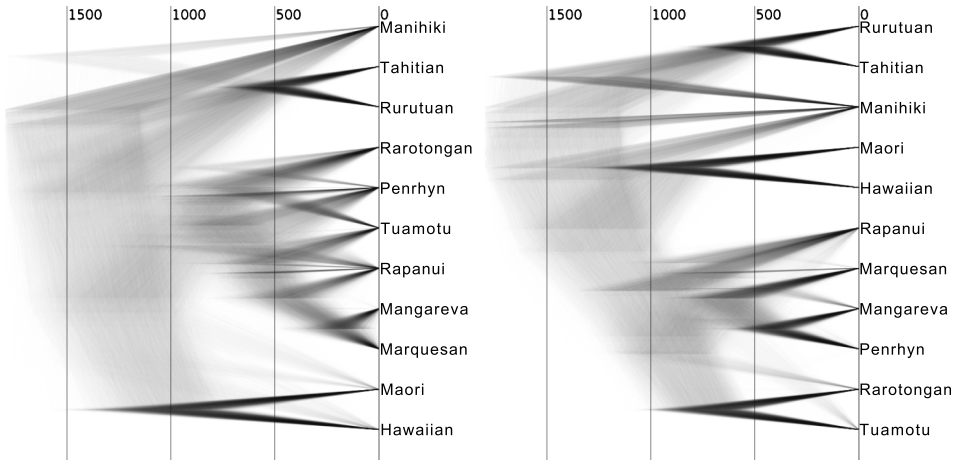
ferred traits are discarded from the data. In each case, we obtain a satisfactory fit to the data and recover the true parameters.

8. Application. The order and timing of human settlement in Eastern Polynesia is a matter of debate. In the standard subgrouping of the Eastern Polynesian languages, Rapanui diverges first, followed by the split leading to the Marquesic (Hawaiian, Mangarevan, Marquesan) and Tahitic (Manihiki, Maori, Penrhyn, Rarotongan, Rurutuan, Tahitian, Tuamotuan) language subgroups [Marck (2000)]. Recent linguistic and archaeological evidence has challenged this theory. In an implicit phylogenetic network study of lexical traits, Gray, Bryant and Greenhill (2010) detect nontree-like signals in the data; furthermore, the Tahitic and Marquesic languages do not form clean clusters in their study. In a meta-analysis of radiocarbon-dated samples from archaeological sites in the archipelago, Wilmshurst et al. (2011) claim that Eastern Polynesia was settled in two distinct phases: the Society Islands between 900 and 1000 years before the present (BP), and the remainder between 700 and 900 years BP. These dates, much later than those reported by Spriggs and Anderson (1993), for example, do not allow much time for the development of the Eastern Polynesian language subgroups. Conte and Molle (2014) present evidence of human settlement in the Marquesas Islands approximately 1100 years BP. On the basis of the above and further evidence of lateral transfer in primary source material, Walworth (2014) disputes Marquesic and Tahitic as distinct subgroups.

To add to this debate, we compare the SDLT and SD models on a data set of lexical traits in eleven Eastern Polynesian languages drawn from the approximately 1200 languages in the Austronesian Basic Vocabulary Database [Greenhill, Blust and Gray (2008)]. The data is a subset of the Polynesian language data set in the study of Gray, Bryant and Greenhill (2010). We analyze the 968 traits marked present in at least one of the eleven languages, hereafter referred to as `POLY-0`. The data are isochronous. Consistent with Gray, Drummond and Greenhill (2009), the sole clade constraint limits the root of the tree to lie between 1150 and 1800 years BP.

We plot samples from the marginal tree posterior under the SDLT and SD models in Figure 4. We summarize these distributions with *majority rule consensus trees* in the Supplemental Materials. In agreement with Gray, Bryant and Greenhill (2010) and Walworth (2014), the standard subgroupings do not appear as subtrees in either model. Rapanui does not form an outgroup in either of our analyses. There is little evidence in the tree posteriors to support the claim of Wilmshurst et al. (2011), however, as the posterior distributions of the root time, t_1, resembles its approximately uniform prior distribution on the range [1150, 1800] years BP.

The majority of the uncertainty under the SDLT model is in the topology of the subtree containing Rarotongan, Penrhyn, Tuamotu, Rapanui, Mangareva and Marquesan. This subtree also has 100% posterior support under the SD model, but most of the uncertainty here is in relationships further up the tree. We use



(a) SDLT model. (b) SD model.

FIG. 4. `DensiTree` [*Bouckaert and Heled* (2014)] *plots of samples from the marginal tree posterior under the SDLT and SD models fit to* `POLY-0`. *Heavier lines indicate higher posterior support. Time is in units of years before the present.*

`BEAST` [Drummond et al. (2012)] to obtain the 95% highest posterior probability sets for the tree topologies under the respective models. These sets comprise 135 topologies for the SDLT model and 19 for the SD model. This level of confidence in relatively few topologies is likely a result of the SD model's misspecification on the laterally transferred traits.

The effect of the laterally transferred traits in the data is also evident in the histograms in Figure 5. The death rate μ is approximately 50% higher under the SD model, as traits must be born further up the tree and killed off at a higher rate to explain the variation in the data due to lateral transfer. The relative transfer rate β/μ is the expected number of times that a single instance of a trait transfers before dying out; its posterior distribution under the SDLT model is centered on 1.35. In contrast, on the basis of simulation studies, both Nicholls and Gray (2008) and Greenhill, Currie and Gray (2009) consider a relative transfer rate of 0.5 high. We report histograms for the remaining parameters as well as the trace and autocorrelation plots we use to diagnose the convergence of our Markov chains [Geyer (1992)] in the Supplemental Materials.

With the above concerns about the SD model in mind, we now assess the validity of our analyses. To assess goodness of fit, we relax the constraints on each leaf time and attempt to reconstruct them. The constraint $\Gamma^{(i)} = \{g \in \Gamma : t_i = 0\}$ fixes leaf $i \in V_L$ at time 0, and $\Gamma^{(i')} = \{g \in \Gamma : -10^3 \leq t_i \leq 10^4\}$ denotes its relaxation to a wide interval either side of time 0. We denote by $\Gamma^{C'}$ the calibrated space of phylogenies with $\Gamma^{(i)}$ replaced by $\Gamma^{(i')}$. As constraint $\Gamma^C \subset \Gamma^{C'}$, the Bayes factor

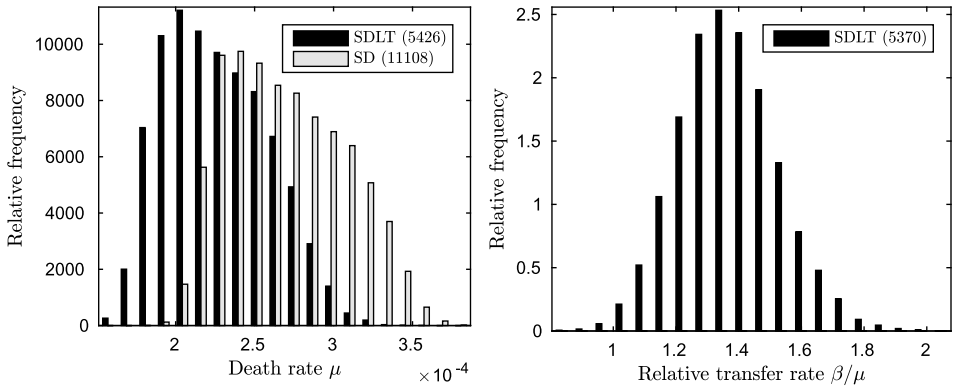


FIG. 5. *Marginal parameter posterior distributions under the SDLT and SD models fit to the Eastern Polynesian data set* `POLY-0`. *Marginal effective sample sizes are in parentheses.*

comparing the relaxed and constrained models is

$$
\begin{aligned}
B_{i',i} &= \frac{\pi(R(\mathbf{D})|g \in \Gamma^{C'})}{\pi(R(\mathbf{D})|g \in \Gamma^{C})} \\
&= \frac{\pi(R(\mathbf{D})|g \in \Gamma^{C'})}{\pi(R(\mathbf{D})|g \in \Gamma^{C} \cap \Gamma^{C'})} \\
&= \frac{\pi(g \in \Gamma^{C}|g \in \Gamma^{C'})}{\pi(g \in \Gamma^{C}|R(\mathbf{D}), g \in \Gamma^{C'})},
\end{aligned} \tag{8.1}
$$

a Savage–Dickey ratio of the marginal prior and posterior densities that the constraint $\Gamma^{(i)}$ is satisfied in the relaxed model. A large Bayes factor here indicates a lack of support for the leaf constraint and is therefore a sign of model misspecification.

We cannot compute the Savage–Dickey ratio in equation (8.1) in closed form, and so in practice we estimate the densities by the proportions of sampled leaf times in the range $[-50, 50]$ years around time 0. We report log-Savage–Dickey ratios in Figure 6 and histograms of the marginal leaf ages in the Supplemental Materials. There are clear signs that the SD model is misspecified here. In particular, the SD model rejects the constraints on Manihiki and Marquesan, and so we report lower bounds on the corresponding Bayes factors. The large Bayes factor for the constraint on Rapanui provides "positive" evidence of misspecification on the scale of Kass and Raftery (1995).

We assess the predictive performance of each model on a random splitting of the registered data $R(\mathbf{D})$ into evenly sized training and test sets labeled $\mathbf{D}^{\mathrm{tr}}$ and $\mathbf{D}^{\mathrm{te}}$, respectively. Madigan and Raftery (1994) propose to score each model by its log-posterior predictive probability, $\log \pi(\mathbf{D}^{\mathrm{te}}|\mathbf{D}^{\mathrm{tr}})$, where $\pi(\mathbf{D}^{\mathrm{te}}|\mathbf{D}^{\mathrm{tr}}) = \int \pi(\mathbf{D}^{\mathrm{te}}|x)\pi(x|\mathbf{D}^{\mathrm{tr}})\,\mathrm{d}x$, with $x = (g, \mu, \beta, \kappa, \Xi)$ for the SDLT model and $x =$

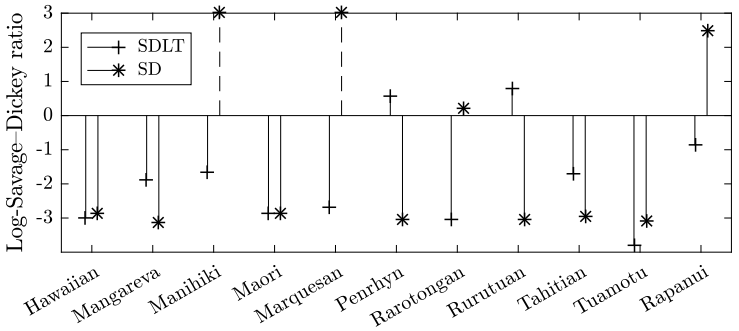


FIG. 6. *Bayes factors comparing the support for the leaf constraints in the SDLT and SD models fit to* POLY-0. *We estimate lower bounds on the Bayes factors for the constraints on Manihiki and Marquesan under the SD model.*

(g, μ, κ, Ξ) for the SD model. The difference in scores is a log-Bayes factor measuring the relative success of the models at predicting the test data [Kass and Raftery (1995)]. The results in Table 4 strongly support the superior fit of the SDLT model to POLY-0.

Traits marked present in a single language are often deemed unreliable and removed in the registration process. To address this concern, we repeat our analyses on the data set POLY-1 which we form by removing the *singleton* patterns from POLY-0. Although the outcome of the predictive model selection in Table 4 is unchanged, these singleton patterns play an important role in SDLT model inference as parameter credible intervals are affected by their removal.

9. Concluding remarks. Lateral transfer is an important problem, but practitioners lack the tools to perform fully likelihood-based inference for dated phylogenies in this setting. We address this issue with a novel model for species diversification which extends the Stochastic Dollo model for lateral transfer in trait presence/absence data. To our knowledge, the method we describe is the first fully likelihood-based approach to control for lateral transfer in reconstructing a rooted phylogenetic tree. The second major contribution of this paper is the inference procedure whereby we integrate out the locations of the trait birth, death and transfer events through a sequence of initial value problems.

TABLE 4
Posterior predictive model assessment

Data set	SDLT score	SD score	Log-Bayes factor
POLY-0	−3058.2	−3105.8	47.6
POLY-1	−1401.2	−1481.1	79.9

In the application we consider, accounting for lateral transfer results in an improved fit over the regular Stochastic Dollo model but comes at a significant computational cost. The sequence of initial value problems to compute the likelihood parameters in the lateral transfer model is easy to state but difficult to solve in practice. On a tree with L leaves, we can exploit symmetry in the differential systems to compute the expected pattern frequencies exactly in $\mathcal{O}(2^{2L})$ operations. In practice, we use an ordinary differential equation (ODE) solver to approximate their values within an error tolerance dominated by the Monte Carlo error. This approach requires $\mathcal{O}(L2^L C(L))$ operations, where $C(L)$ is the number of matrix-vector multiplications required by the ODE solver; for example, with the `Matlab` `ODE45` solver and typical choices of parameters, we observe $C(10) \in [80, 90]$ and $C(20) \in [95, 100]$. This approach is feasible for approximately $L = 20$ leaves on readily available hardware. As we must evaluate the likelihood many times over the course of an MCMC analysis, this computational burden is a major stumbling block toward applying our model to data sets with more taxa or multiple character states, and is the focus of ongoing research [Kelly (2016), Chapter 4].

The model as described is not *projective* in the sense that we cannot marginalize out the effect of unobserved lineages, which in our analyses correspond to the many Polynesian languages not included in our data set. Consequently, the probability that a trait transfers between sampled lineages decreases as the number of unobserved lineages increases. Similarly, a trait which previously died out on the sampled lineages may transfer back into the system from an unobserved lineage. One possible solution to this problem is to introduce *ghost* lineages [Szöllosi et al. (2012, 2013)] to allow for lateral transfer between sampled and unsampled taxa at the expense of an increase in computational cost. There are many other avenues for future work on the model. For example, one could partition the data across a mixture of models and trees, relax the global lateral transfer regime or the assumption that traits are independent, model multiple character states [Alekseyenko, Lee and Suchard (2008)], allow individual catastrophes to vary in their effect, jointly model sequence and trait presence/absence data [Cybis et al. (2015)], and account for other types of missing data.

There are many open problems which have been ignored due to the expense of fitting models that account for lateral transfer. One such example occurs in the model of Chang et al. (2015) whereby ancestral nodes may have data. Stochastic Dollo without lateral transfer cannot be used to model the observation process here, as traits absent in an ancestral state but present in both descendent and non-descendent leaves violate the Dollo parsimony assumption. Our method provides a model-based solution to this problem and many others.

APPENDIX: MCMC TRANSITION KERNELS

We extend existing sampling algorithms for the Stochastic Dollo model [Nicholls and Gray (2008), Ryder and Nicholls (2011)] to construct a Markov

chain whose invariant distribution is the posterior $\pi(g, \mu, \beta, \kappa, \Xi|R(\mathbf{D}))$ in equation (6.1). In the following, $x = [(V, E, T, C), \mu, \beta, \kappa, \Xi]$ is the current state of the chain, and a move to a new state x^* drawn from the proposal distribution $Q(x, \cdot)$ is accepted with probability

$$\min\left[1, \frac{\pi(x^*|\mathbf{D})}{\pi(x|\mathbf{D})}\frac{Q(x^*, x)}{Q(x, x^*)}\right].$$

We apply the same scaling update to the lateral transfer rate β and the death rate μ. If $x^* = [(V, E, T, C), \mu, \beta^*, \kappa, \Xi]$ where $\beta^*|\beta \sim \mathrm{U}[\varrho^{-1}\beta, \varrho\beta]$ for some constant $\varrho > 1$, then the Hastings ratio for this move is

$$\frac{Q(x^*, x)}{Q(x, x^*)} = \frac{\beta}{\beta^*}.$$

A catastrophe $c = (b, u) \in C$ in state x occurs on branch $b \in E$ at time $t_b + u(t_{\mathrm{pa}(b)} - t_b)$, where $u \in (0, 1)$ is the relative location of the catastrophe along the branch. The location for a new catastrophe $c^* = (b^*, u^*)$ is chosen uniformly at random across the branches of the tree to form the proposed state x^* with catastrophe set $C \cup \{c^*\}$. We choose catastrophes uniformly at random for deletion in the reverse move, and so

$$\frac{Q(x^*, x)}{Q(x, x^*)} = \frac{p_{DC}}{p_{AC}}\frac{1}{|C| + 1}\sum_{i \in E\setminus\{1\}}(t_{\mathrm{pa}(i)} - t_i),$$

where p_{AC} and p_{DC} denote the probabilities of proposing to add and delete a catastrophe, respectively.

We chose catastrophe $c = (b, u)$ uniformly from the catastrophe set C to move to branch b^* chosen uniformly from the $\deg(b) + \deg[\mathrm{pa}(b)] - 2$ branches neighboring branch b, where $\deg(b)$ denotes the degree of node b. This is equivalent to deleting a randomly chosen catastrophe and adding it to a neighboring branch, although we do not resample the relative location u. If $c^* = (b^*, u)$ replaces c in the proposed state x^*, then

$$\frac{Q(x^*, x)}{Q(x, x^*)} = \frac{\deg(b) + \deg[\mathrm{pa}(b)] - 2}{\deg(b^*) + \deg[\mathrm{pa}(b^*)] - 2}\frac{t_{\mathrm{pa}(b^*)} - t_{b^*}}{t_{\mathrm{pa}(b)} - t_b},$$

where the Jacobian term $(t_{\mathrm{pa}(b^*)} - t_{b^*})/(t_{\mathrm{pa}(b)} - t_b)$ accounts for the change in sampling distribution of the catastrophe position due to the change in branch lengths. In fact, for every proposed move x to $x^* = [(V', E', T', C), \dots]$ which affects tree branch lengths, the Hastings ratio includes a Jacobian term of the form

$$\prod_{i \in E\setminus\{1\}}\frac{|C^{(i)}|!}{(t_{\mathrm{pa}(i)} - t_i)^{|C^{(i)}|}}\frac{(t^*_{\mathrm{pa}(i)} - t^*_i)^{|C^{*(i)}|}}{|C^{*(i)}|!},$$

to account for the relative sampling densities for the catastrophe sets in each state, where $C^{(i)}$ and $C^{*(i)}$ respectively denote the catastrophe set on branch i in the current and proposed states.

Subtree-prune-and-regraft moves on the tree are designed in such a way that the total number of catastrophes on the tree remains constant and the Hastings ratio is unaffected except by the Jacobian term above. We illustrate these moves in the Supplemental Materials.

Acknowledgments. The authors wish to thank Robin Ryder and Simon Greenhill for their assistance with this project, and acknowledge the feedback of the Associate Editor and two anonymous reviewers.

SUPPLEMENTARY MATERIAL

Supplemental Materials: Lateral transfer in Stochastic Dollo models (DOI: 10.1214/17-AOAS1040SUPP; .pdf). The supplement contains a proof of Theorem 1 and supporting material for the analyses in Sections 7 and 8.

REFERENCES

ABBY, S. S., TANNIER, E., GOUY, M. and DAUBIN, V. (2010). Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinform.* **11** 324.

ALEKSEYENKO, A. V., LEE, C. J. and SUCHARD, M. A. (2008). Wagner and Dollo: A stochastic duet by composing two parsimonious solos. *Syst. Biol.* **57** 772–784.

BEIKO, R. G. and HAMILTON, N. (2006). Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol.* **6** 15.

BOUCHARD-CÔTÉ, A. and JORDAN, M. I. (2013). Evolutionary inference via the Poisson Indel Process. *Proc. Natl. Acad. Sci. USA* **110** 1160–1166.

BOUCKAERT, R. and HELED, J. (2014). DensiTree 2: Seeing trees through the forest. *BioRxiv*.

BOUCKAERT, R., LEMEY, P., DUNN, M., GREENHILL, S. J., ALEKSEYENKO, A. V., DRUMMOND, A. J., GRAY, R. D., SUCHARD, M. A. and ATKINSON, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science* **337** 957–960.

BOUCKAERT, R., HELED, J., KÜHNERT, D., VAUGHAN, T., WU, C.-H., XIE, D., SUCHARD, M. A., RAMBAUT, A. and DRUMMOND, A. J. (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10** 1–6.

CHANG, W., CATHCART, C., HALL, D. and GARRETT, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* **91** 194–244.

CONTE, E. and MOLLE, G. (2014). Reinvestigating a key site for Polynesian prehistory: New results from the Hane dune site, Ua Huka (Marquesas). *Archaeol. Ocean.* **49** 121–136.

CYBIS, G. B., SINSHEIMER, J. S., BEDFORD, T., MATHER, A. E., LEMEY, P. and SUCHARD, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.* **9** 969–991. MR3371344

DAUBIN, V., GOUY, M. and PERRIÈRE, G. (2002). A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Res.* **12** 1080–1090.

DRUMMOND, A. J., SUCHARD, M. A., XIE, D. and RAMBAUT, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29** 1969–1973.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17** 368–376.

GEYER, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7** 473–483.

GRAY, R. D. and ATKINSON, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426** 435–439.

GRAY, R. D., BRYANT, D. and GREENHILL, S. J. (2010). On the shape and fabric of human history. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **365** 3923–3933.

GRAY, R. D., DRUMMOND, A. J. and GREENHILL, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323** 479–483.

GREENHILL, S. J., BLUST, R. and GRAY, R. D. (2008). The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evol. Bioinform.* **4** 271–283.

GREENHILL, S. J., CURRIE, T. E. and GRAY, R. D. (2009). Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. Lond., B Biol. Sci.* **276** 2299–2306.

HELED, J. and DRUMMOND, A. J. (2012). Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.* **61** 138–149.

HUSON, D. H. and BRYANT, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23** 254–267.

HUSON, D. H. and STEEL, M. (2004). Phylogenetic trees based on gene content. *Bioinformatics* **20** 2044–2049.

JOFRÉ, P., DAS, P., BERTRANPETIT, J. and FOLEY, R. (2017). Cosmic phylogeny: Reconstructing the chemical history of the solar neighbourhood with an evolutionary tree. *Mon. Not. R. Astron. Soc.* **467** 1140–1153.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402

KELLY, L. J. (2016). A Stochastic Dollo model for lateral transfer. Ph.D. thesis, Univ. Oxford.

KELLY, L. J. and NICHOLLS, G. K. (2017). Supplement to "Lateral transfer in Stochastic Dollo models." DOI:10.1214/17-AOAS1040SUPP.

KINGMAN, J. F. C. (1993). *Poisson Processes. Oxford Studies in Probability* **3**. The Clarendon Press, Oxford. MR1207584

KITCHEN, A., EHRET, C., ASSEFA, S. and MULLIGAN, C. J. (2009). Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. R. Soc. Lond., B Biol. Sci.* **276** 2703–2710.

KUBATKO, L. S. (2009). Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.* **58** 478–488.

LATHROP, G. M. (1982). Evolutionary trees and admixture: Phylogenetic inference when some populations are hybridized. *Ann. Hum. Genet.* **46** 245–255. MR0673807

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

MARCK, J. C. (2000). *Topics in Polynesian Language and Culture History* **504**. Pacific Linguistics, Canberra.

MCPHERSON, A., ROTH, A., LAKS, E., MASUD, T., BASHASHATI, A., ZHANG, A. W., HA, G., BIELE, J., YAP, D., WAN, A., PRENTICE, L. M., KHATTRA, J., SMITH, M. A., NIELSEN, C. B., MULLALY, S. C., KALLOGER, S., KARNEZIS, A., SHUMANSKY, K., SIU, C., ROSNER, J., CHAN, H. L., HO, J., MELNYK, N., SENZ, J., YANG, W., MOORE, R., MUNGALL, A. J., MARRA, M. A., BOUCHARD-CÔTÉ, A., GILKS, C. B., HUNTSMAN, D. G., MCALPINE, J. N., APARICIO, S. and SHAH, S. P. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **48** 758–767.

NICHOLLS, G. K. and GRAY, R. D. (2008). Dated ancenstral trees from binary trait data and their application to the diversification of languages. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 545–566. MR2420414

NICHOLLS, G. K. and RYDER, R. J. (2011). Phylogenetic models for Semitic vocabulary. In *Proceedings of the International Workshop on Statistical Modelling* (D. Conesa, A. Forte, A. López-Quílez and F. Muñoz, eds.) 431–436.

NICHOLLS, G. K., RYDER, R. J. and WELCH, D. (2013). TraitLab: A MatLab package for fitting and simulating binary trait-like data.

OLDMAN, J., WU, T., VAN IERSEL, L. and MOULTON, V. (2016). TriLoNet: Piecing together small networks to reconstruct reticulate evolutionary histories. *Mol. Biol. Evol.* **33** 2151–2162.

PATTERSON, N., MOORJANI, P., LUO, Y., MALLICK, S., ROHLAND, N., ZHAN, Y., GENSCHORECK, T., WEBSTER, T. and REICH, D. (2012). Ancient admixture in human history. *Genetics* **192** 1065–1093.

PICKRELL, J. K. and PRITCHARD, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8** e1002967.

RANNALA, B. and YANG, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164** 1645–1656.

ROCH, S. and SNIR, S. (2013). Recovering the treelike trend of evolution despite extensive lateral genetic transfer: A probabilistic analysis. *J. Comput. Biol.* **20** 93–112. MR3021672

RYDER, R. J. and NICHOLLS, G. K. (2011). Missing data in a stochastic Dollo model for binary trait data, and its application to the dating of Proto-Indo-European. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 71–92. MR2758570

SKELTON, C. (2008). Methods of using phylogenetic systematics to reconstruct the history of the Linear B script. *Archaeometry* **50** 158–176.

SPRIGGS, M. and ANDERSON, A. (1993). Late colonization of East Polynesia. *Antiquity* **67** 200–217.

SZÖLLOSI, G. J., BOUSSAU, B., ABBY, S. S., TANNIER, E. and DAUBIN, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. USA* **109** 17513–17518.

SZÖLLŐSI, G. J., TANNIER, E., LARTILLOT, N. and DAUBIN, V. (2013). Lateral gene transfer from the dead. *Syst. Biol.* **62** 386–397.

TAVARÉ, S., BALDING, D. J., GRIFFITHS, R. C. and DONNELLY, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145** 505–518.

VEERAMAH, K. R., WOERNER, A. E., JOHNSTONE, L., GUT, I., GUT, M., MARQUES-BONET, T., CARBONE, L., WALL, J. D. and HAMMER, M. F. (2015). Examining phylogenetic relationships among Gibbon genera using whole genome sequence data using an approximate Bayesian computation approach. *Genetics* **200** 295–308.

WALWORTH, M. (2014). Eastern Polynesian: The linguistic evidence revisited. *Ocean. Linguist.* **53** 256–272.

WEN, D., YU, Y. and NAKHLEH, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet.* **12** e1006006.

WILMSHURST, J. M., HUNT, T. L., LIPO, C. P. and ANDERSON, A. J. (2011). High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proc. Natl. Acad. Sci. USA* **108** 1815–1820.

DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
24–29 ST GILES'
OXFORD, OX1 3LB
UNITED KINGDOM
E-MAIL: kelly@stats.ox.ac.uk
nicholls@stats.ox.ac.uk