Version dated: August 10, 2017



# GHOST: Recovering Historical Signal from Heterotachously-evolved Sequence Alignments

Stephen M Crotty[†*1,2,3], Bui Quang Minh[†1], Nigel G Bean[2,3], Barbara R Holland[4], Jonathan Tuke[2,3], Lars S Jermiin[5,6], Arndt von Haeseler[1,7]

[1] *Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna and Medical University of Vienna, Vienna, Austria*

[2] *School of Mathematical Sciences, University of Adelaide, Adelaide, SA 5005, Australia.*

[3] *ARC Centre of Excellence for Mathematical and Statistical Frontiers, The University of Adelaide, Adelaide, SA, Australia.*

[4] *School of Physical Sciences, University of Tasmania, Hobart, TAS 7001, Australia.*

[5] *CSIRO Land & Water, Black Mountain Laboratories, Canberra, ACT 2601, Australia.*

[6] *Research School of Biology, Australian National University, Canberra, ACT 2601, Australia*

[7] *Bioinformatics & Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria.*

† Joint first authors (these authors contributed equally to the work)

*Corresponding author: Stephen Crotty, Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna and Medical University of Vienna, Vienna, Austria; E-mail: stephen.crotty@univie.ac.at

*Abstract.*— Molecular sequence data that have evolved under the influence of heterotachous evolutionary processes are known to mislead phylogenetic inference. We introduce the General Heterogeneous evolution On a Single Topology (GHOST) model of sequence evolution, implemented under a maximum-likelihood framework in the phylogenetic program IQ-TREE. Extensive simulations show that the GHOST model can accurately recover the tree topology, branch lengths, substitution rate and base frequency parameters from heterotachously-evolved sequences. We apply our model to an electric fish dataset and identify a subtle component of the historical signal, linked to the previously established convergent evolution of the electric organ in two geographically distinct lineages of electric fish. We compare the GHOST model to the partition model and show that, owing to the minimization of model constraints, the GHOST model is able to offer unique biological insights when applied to empirical data.

Keywords: Phylogenetics, heterotachy, mixture model, maximum likelihood, convergent evolution

The success and reliability of model-based phylogenetic inference methods are limited by the adequacy of the models that are assumed to approximate the evolutionary process. Homogeneous evolutionary models have long been recognised as inadequate since the rate of evolution is known to vary across sites (Fitch and Margoliash, 1967; Holmquist et al., 1983) and across lineages (Baele et al., 2006; Lopez et al., 2002; Wu and Susko, 2011; Jayaswal et al., 2014). There are many models that have been proposed to compensate for rate heterogeneity across sites. The classical example is the discrete $\Gamma$ model (Yang, 1994), which allows different classes of variable sites to have their rates drawn from a $\Gamma$ distribution. More recently, Kalyaanamoorthy et al. (2017) relaxed the requirement for the rates of the classes to fit a $\Gamma$ distribution, implementing a probability distribution-free rates-across-sites model. However, these models still assume that the substitution rate for each site is constant across all lineages. This is too restrictive; biologically speaking it is not hard to accept that evolutionary processes can be both lineage and time dependent. In the context of a phylogenetic tree this manifests as lineage-specific shifts in evolutionary rate, coined heterotachy (Philippe and Lopez, 2001; Lopez et al., 2002), resulting in sequences that cannot be characterised as having evolved according to a single set of branch lengths and substitution model.

The effect of heterotachy on phylogenetic inference was thrust into the spotlight by Kolaczkowski and Thornton (K&T) (2004). They used a simulation study to show that heterotachously-evolved sequences could mislead the popular inference methods of maximum-likelihood (ML) and Bayesian Markov Chain Monte-Carlo (BMCMC) to a greater extent than maximum parsimony (MP). Their findings were controversial and were widely challenged on the grounds that the simulations captured only a special case of heterotachy (Gadagkar and Kumar, 2005; Philippe et al., 2005; Spencer et al., 2005; Steel, 2005), and more general studies of heterotachy concluded that ML performed at least as well as, and in most cases better than, MP (Gadagkar and Kumar, 2005; Spencer et al., 2005). Valid as these criticisms may have been, the key issue that K&T's study brought to light stood firm - heterotachy was a primary source of model misspecification and the models and methods of the time were ill-equipped to deal with it. The main impediment to the development of models that can accommodate heterotachously-evolved sequences has been the computational expense. Models that account for heterogeneity of rates of change across sites can be integrated relatively cheaply, but modeling heterotachy is not so simple. One approach has been to use partition models (Lanfear et al., 2012), which require the data to be partitioned *a priori*. The analysis then proceeds by inferring seperate branch length and model parameters for each partition. Sequence data is commonly partitioned based on genes and/or codon position. However, the inherent assumption of such a partitioning scheme is that heterotachy only occurs between partitions, not within each partition. This may not be a valid assumption, so the requirement to partition the data in advance of the analysis is a possible source of model misspecification. Another approach has been to use mixture models, in which the likelihood of the data at each site in the alignment is calculated as a weighted sum across multiple classes (see Pagel and Meade (2005) for a detailed description of phylogenetic mixture models). The most common approaches can be referred to as mixed substitution rate (MSR) models (Lartillot and Philippe, 2004; Pagel and Meade, 2004), whereby each class has its own substitution rate matrix; and mixed branch length (MBL) models (Kolaczkowski and Thornton, 2004; Meade and Pagel, 2008), whereby each class has its own set of branch lengths on the tree. As a consequence of their parameter rich nature, these models have all been implemented only within a Bayesian framework. Wu and Susko (2009) proposed a general framework for heterotachy, encompassing both mixed substitution rate and mixed branch length models as special cases. Another example is the CAT models of Lartillot and Philippe (2004), which have been widely used (Whelan and Halanych (2017) and references therein). Whelan and Halanych (2017) carried out extensive simulation and empirical studies comparing the performance of the CAT models to partition models. They concluded that despite their additional complexity and associated increase in runtime, the CAT models generally perform no better than partition models. They also found that when new mixture models are introduced in the literature their performance is not always assessed against the current popular methods for phylogenetic analysis, such as partition models.

As a consequence of their varied nature, mixture models require many parameters and the associated computational expense has thus far impeded their implementation in a ML framework. The issue of computational expense is an ever diminishing one; as computing power increases and algorithmic architecture improves, the opportunity to employ more and more complex models of sequence evolution does also. We introduce the General Heterogeneous evolution On a Single Topology (GHOST) model for ML inference. The GHOST model combines features of both MSR and MBL models. It consists of a number of classes, all evolving on the same tree topology. For each class the branch lengths, nucleotide or amino-acid frequencies, substitution rates and class weight are parameters to be inferred. It minimises the number of assumptions that must be made *a priori* by inferring all parameters directly from the data. Therefore, GHOST is free of the artificial constraints common in other models, often included for computational expedience rather than biological relevance. This means that the GHOST model has the necessary freedom to extract any historical signals present in the data. We provide an easy to use implementation of the GHOST model in the phylogenetic program IQ-TREE (Nguyen et al., 2015), the first mixture model of comparable flexibility to be made available in a ML framework.

# Methods and Materials

## *Model Description*

The GHOST model consists of $m$ classes and one tree topology, $T$, common to all classes. All other parameters are inferred separately for each class. For the $j^{th}$ class we define $\boldsymbol{\lambda_j}$ as the set of branch lengths on $T$; $\boldsymbol{R_j}$, the relative substitution rate parameters; $\boldsymbol{F_j}$, the set of nucleotide or amino acid frequencies; and $w_j$, the class weight ($w_j > 0, \sum w_j = 1$). Given a multiple sequence alignment (MSA), $A$, we define $L_{ij}$ as the likelihood of the data observed at the $i^{th}$ site in $A$ under the $j^{th}$ class of the GHOST model. $L_{ij}$ is computed using Felsenstein's pruning algorithm (Felsenstein, 1981). The likelihood of the $i^{th}$ site, $L_i$, is then given by the weighted sum of the $L_{ij}$ over all $j$:

$$L_i = \sum_{j=1}^{m} w_j L_{ij}(T, \boldsymbol{\lambda_j}, \boldsymbol{R_j}, \boldsymbol{F_j}).$$

Therefore, if $S$ contains $N$ sites (length of the alignment), the full log-likelihood, $\ell$, is given by:

$$\ell = \sum_{i=1}^{N} \log \left( \sum_{j=1}^{m} w_j L_{ij}(T, \boldsymbol{\lambda_j}, \boldsymbol{R_j}, \boldsymbol{F_j}) \right).$$

We make use of the existing parameter optimisation algorithms within IQ-TREE, extending it where necessary, to incorporate parameter estimation across the $m$ classes.

## *Model Parameter Estimation for a Fixed Tree, T*

Let $\boldsymbol{\Theta} = \{w_1, \ldots, w_m, \boldsymbol{\lambda_1}, \ldots, \boldsymbol{\lambda_m}, \boldsymbol{R_1}, \ldots, \boldsymbol{R_m}, \boldsymbol{F_1}, \ldots, \boldsymbol{F_m}\}$ denote the GHOST model parameters (*i.e.*, class weights, branch lengths, relative substitution rates, and nucleotide or amino-acid frequencies) for each of the $m$ classes. To estimate all parameters for a tree $T$ we employ an expectation-maximization (EM) algorithm

(Dempster et al., 1977; Wang et al., 2008). We initialize $\mathbf{\Theta}$ with all $\hat{\boldsymbol{R}}_{\boldsymbol{j}} = \mathbf{1}$ in each class, uniform nucleotide or amino-acid frequencies $\hat{\mathbf{F}}_{\mathbf{j}}$ (i.e., the Jukes-Cantor model), and $\hat{w}_j$ and $\hat{\boldsymbol{\lambda}}_{\boldsymbol{j}}$ obtained by parsimonious branch lengths rescaled by a discrete, distribution-free rates-across-sites model (Kalyaanamoorthy et al., 2017) with $m$ categories. This becomes the current estimate $\hat{\mathbf{\Theta}}$. The EM algorithm iteratively performs an expectation (E) step and a maximization (M) step to update the current estimate until a (local) maximum likelihood is reached.

*E-step.*— For each site $i$ and class $j$ compute the posterior probability $\hat{p}_{ij}$ of site $i$ belonging to class $j$ based on the current estimate $\hat{\Theta}$:

$$\hat{p_{ij}} = \frac{\hat{w}_j L_{ij}(T, \hat{\boldsymbol{\lambda}_{\boldsymbol{j}}}, \hat{\boldsymbol{R}}_{\boldsymbol{j}}, \hat{\boldsymbol{F}}_{\boldsymbol{j}})}{\sum_{k=1}^{m} \hat{w}_k L_{ik}(T, \hat{\boldsymbol{\lambda}_{\boldsymbol{k}}}, \hat{\boldsymbol{R}}_{\boldsymbol{k}}, \hat{\boldsymbol{F}}_{\boldsymbol{k}})}.$$

*M-step.*— For each class $j$, maximize the log-likelihood function:

$$\ell_j = \sum_{i=1}^{N} \hat{p_{ij}} \log \Big( L_{ij}(T, \boldsymbol{\lambda}_{\boldsymbol{j}}, \boldsymbol{R}_{\boldsymbol{j}}, \boldsymbol{F}_{\boldsymbol{j}}) \Big)$$

to obtain the next $\hat{\boldsymbol{\lambda}_j}^{NEW}, \hat{\boldsymbol{R}_j}^{NEW}, \hat{\boldsymbol{F}_j}^{NEW}$. This can be done with standard phylogenetic optimization routines for each class.

Finally, the weights are updated by:

$$\hat{w_j}^{NEW} = \frac{1}{N}\sum_{i=1}^{N}\hat{p_{ij}}.$$

That is, the new weight for class $j$ is the mean posterior probability of each site belonging to class $j$. This completes the proposal of the new estimate $\hat{\boldsymbol{\Theta}}^{NEW}$. If $\ell(\hat{\boldsymbol{\Theta}}^{NEW}) > \ell(\hat{\boldsymbol{\Theta}}) + \epsilon$ (where $\epsilon$ is a user-defined tolerance, $\epsilon = 0.01$ by default), then $\hat{\boldsymbol{\Theta}}$ is replaced by $\hat{\boldsymbol{\Theta}}^{NEW}$ and the E and M steps are repeated. Otherwise, the EM algorithm finishes.

An auxiliary benefit of the ML implementation of the GHOST model in IQ-TREE is that once the EM-algorithm has converged, we can soft-classify sites according to their probability of belonging to a particular class. Post convergence, the final values of $p_{ij}$ can be directly interpreted as the probability that the $i^{th}$ site in the alignment belongs to the $j^{th}$ class. This classification can be used to identify sites in the alignment that belong with high probability to a particular class of interest.

## *Software*

The GHOST model has been implemented in IQ-TREE (Nguyen et al., 2015) (http://www.iqtree.org), the first model of this type and complexity to be made available in a ML framework. The GHOST model can be run with both nucleotide and amino acid sequences. The GHOST model is executed in IQ-TREE v1.6 by augmenting the model argument as shown below. For example if one wants to fit a four-class GHOST model in conjunction with the GTR model of evolution to sequences contained in `data.fst`, one would use the following command:

```
iqtree -s data.fst -m GTR+H4
```

By default the above command will infer only one set of equilibrium base frequencies and apply these to all classes. To infer separate equilibrium base frequencies for each class then we must add the `+FO` option:

```
iqtree -s data.fst -m GTR+FO+H4
```

The above command implements the linked version of the GHOST model. This means that only one set of GTR rate parameters will be inferred and applied to all classes. If one wishes to infer separate GTR rate parameters for each class then the unlinked version is required:

```
iqtree -s data.fst -m GTR+FO*H4
```

The `-wspm` option will generate a .siteprob output file. This contains the probability of each site belonging to each class.

```
iqtree -s data.fst -m GTR*H4 -wspm
```

## *Validation of the GHOST Model*

We validated the GHOST model by carrying out two separate simulation studies. The first study was a replication of the simulations carried out by Kolaczkowski and Thornton (2004), focusing on the ability to recover the correct tree topology from heterotachously-evolved data on quartet trees. The second study was on 12-taxon trees and focused on the ability to recover branch length and substitution model parameters from heterotachously-evolved data.

*K&T simulations.*— We followed K&T's method precisely and compared the performance of MP, ML-JC (ML under a JC model) and ML-JC+H2 (ML under JC with 2 GHOST classes). We used *Seq-Gen* (Rambaut and Grassly, 1997) to simulate nucleotide sequences on two symmetric, 4-taxa trees of identical topology (see Fig. 1a) using the JC model of evolution (Jukes and Cantor, 1969). The branch lengths were constructed such that each tree comprised of two non-sister long branches (length $p$) and two non-sister short branches (length $q$) separated by an internal branch (length $r$). We replicated three separate experiments previously carried out by K&T.

*12-taxon simulations.*— The replication of the K&T simulations focused on recovering tree topology only. However, the GHOST model is parameter rich and naturally the validation process must address its ability to accurately recover branch lengths and model parameters. We constructed independent sets of parameters for two classes on a randomly generated 12-taxon tree using the GTR model of evolution. For each class the branch lengths were drawn randomly from an exponential distribution with a mean of 0.1. When specifying a GTR rate matrix in *Seq-Gen*, the G↔T substitution rate is fixed at 1 and all other substitution rates are expressed relatively. Within each class, the five relative substitution rates were drawn randomly from a uniform distribution between 0.5 and 5. The four base frequencies for each class were assigned a minimum of 0.1, with the remainder allocated proportionally by scaling a normalised set of four observations from a uniform distribution. From these two classes MSAs were constructed (again using *Seq-Gen*) by varying the weight of each class. The weight of Class 1, $w_1$, was varied from 0.2 to 0.8 in increments of 0.05 and at each increment 20 separate MSAs were simulated. Each MSA was constructed by concatenating two independently simulated sets of sequences, the first of length $10000 \times w_1$ simulated using the Class 1 parameters, and the second of length $10000 \times (1 - w_1)$ simulated using the Class

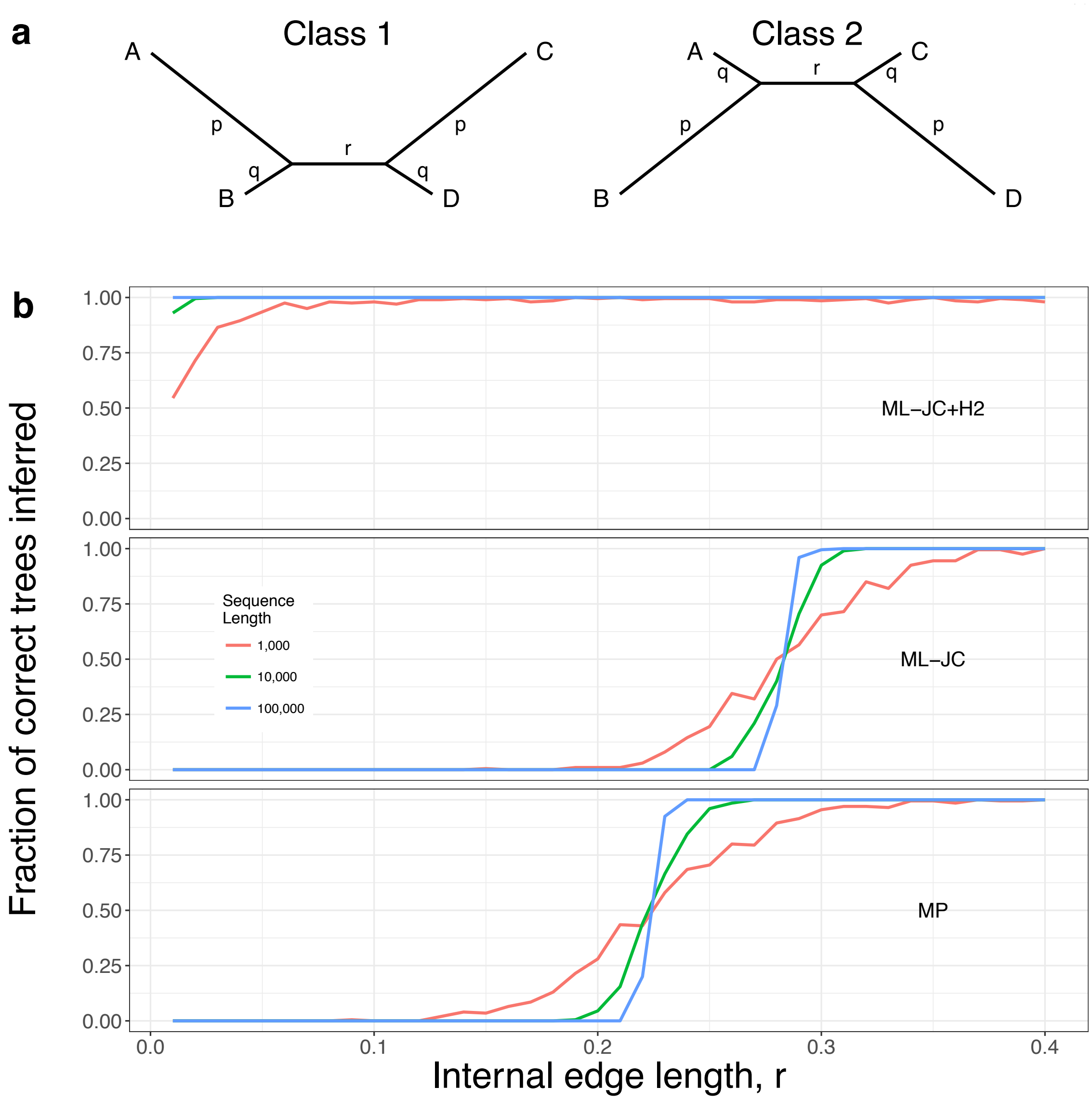


Figure 1: Replication of the simulations of Kolaczkowski & Thornton. (a) We simulated DNA sequences on two symmetric, 4-taxon trees of identical topology using the Jukes-Cantor (JC) model of evolution (Jukes and Cantor, 1969). The branch lengths were constructed such that each tree comprised of two non-sister long branches and two non-sister short branches. Thus each tree was susceptible to long branch attraction (Felsenstein, 1978) (LBA). Importantly, the LBA artefact in both trees was complementary - the bias was in the direction of the AC|BD tree. (b) Performance of maximum likelihood (ML) using a JC, two-class mixture model (ML-JC+H2), ML using a single-class JC model (ML-JC) and maximum parsimony (MP) for data generated under strong heterotachy, $p$=0.75 and $q$=0.05. The length of the internal branch, $r$, was varied between 0.01 and 0.4 with 200 replicates at each value of $r$. ML-JC+H2 was able to reliably recover the tree topology for this data even when the internal branch is very short.

2 parameters. We used IQ-TREE to infer parameters from each MSA under a GHOST model with two GTR classes (GTR+FO*H2). We also inferred parameters from each MSA under a GTR edge-unlinked partition model.

*Parameter recovery: metrics.*— The recovery accuracy of base frequency and relative rate parameters for the 12-taxon simulations was measured by calculating the mean absolute difference between the inferred and true parameters. The accuracy of branch length estimates was assessed using the branch score metric, BS (Kuhner and Felsenstein, 1994). One challenge in assessing accuracy of branch length recovery is that BS is an absolute distance metric. Therefore, we established a frame of reference so that we could assess whether the results obtained are suitably close to the truth or not. To do this we made use of the estimates under the edge-unlinked partition model as a baseline. The fundamental difference between the partition model and the GHOST model is that the partition model has *a priori* knowledge of which sites in the alignment belong to which class. This means that in effect (and excluding the possibility of inferring the incorrect topology) the results of the partition model are identical to those that would be obtained by fitting GTR models to the Class 1 and Class 2 sequences independently. Thus we can consider the accuracy of the partition model as a benchmark.

### *Convergent Evolution of the Na$_v$1.4a Gene Among Teleosts*

To investigate the performance of the GHOST model using real data we applied it to a sequence alignment (2178 bp) taken from the coding region of a sodium channel gene, Na$_v$1.4a, for 11 teleost species. We used Akaike's Information Criterion (AIC) (Akaike, 1974) to determine the model of sequence evolution and number of classes that provided the best fit to the data. We also used PartitionFinder (Lanfear et al., 2012) and IQ-TREE to fit the best edge-unlinked partition model to the alignment. The data was partitioned based on codon position.

## Results & Discussion

### *Validation - K&T Simulations*

*Experiment 1.*— We fixed $p = 0.75$ and $q = 0.05$ (see Fig. 1a) and varied the internal branch length, $r$, on the interval $[0.01, 0.4]$ in increments of 0.01. For each value of $r$, 200 simulated MSAs were constructed by concatenating two sub-alignments of equal length, one simulated on each of the trees in Figure 1a. We carried out phylogenetic inference on each MSA using MP, ML-JC and ML-JC+H2. The experiment was repeated for sequence lengths of 1,000, 10,000 and 100,000 base pairs. The results are shown in Figure 1b. We found that both

ML-JC and MP were misled when $r$ was short, but as $r$ increased MP recovered before ML. For a sequence length of 100kb, MP was misled to some extent for $r < 0.24$ and ML-JC was misled for $r < 0.3$. These findings mirrored those of K&T precisely. However, the ML-JC+H2 model however was never misled. Figure 1b shows that given sufficient sequence length, the ML-JC+H2 model inferred the correct topology from the heterogeneous sequences 100% of the time with $r$ as low as 0.01. Our results clearly demonstrate that the ML-JC+H2 model can correctly infer the tree topology when ML-JC and MP both are misled by the heterotachous nature of the data.

*Experiment 2.*— We tested nine different combinations of $p \in \{0.3, 0.5, 0.7\}$ and $q \in \{0.001, 0.1, 0.2, 0.3, 0.4\}$ (see Fig. 1a). For each of the three methods/models (MP, ML-JC and ML-JC+H2) and at each combination of $p$ and $q$ we determined the smallest value of $r$ (subject to the minimum $r = 0.001$), denoted $BL_{50}$ by K&T, such that the correct topology was returned at least 50% of the time. The results (Fig. 2) indicate that ML-JC+H2 comprehensively outperformed the two alternatives, with the difference most apparent when the influence of heterotachy was strongest (most notably when $p$ is large and $q$ is small). Again the results we observed for MP and ML-JC closely emulated the findings of K&T.

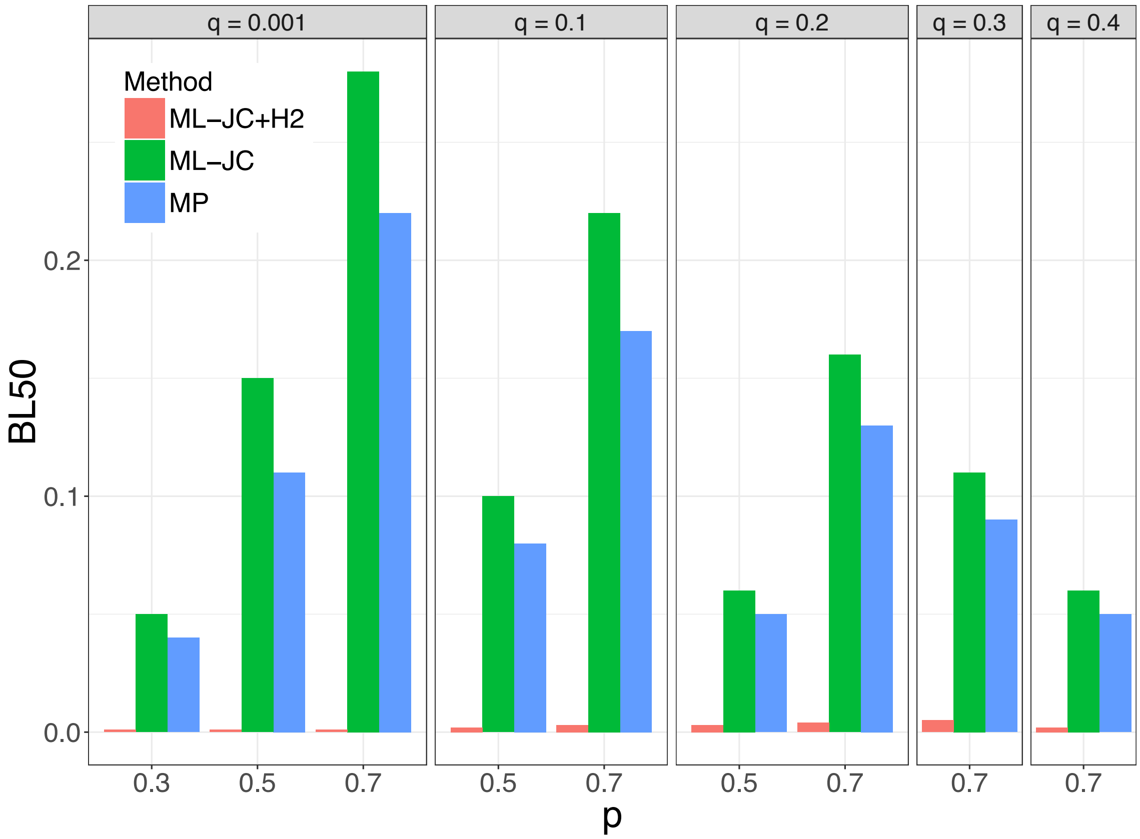


Figure 2: The ML-JC+H2 model clearly outperforms MP and ML-JC over the range of heterotachous conditions tested by K&T. They introduced the $BL_{50}$ measure as the minimum internal branch length required for the method to recover the correct tree topology at least 50% of the time. Small values of $BL_{50}$ indicate that the model is less likely to be misled by the heterotachous nature of the data.

*Experiment 3.*— We tested the impact of varying the weight, $w$, of each class in the simulated MSAs for a variety of branch length combinations. Initially $p$ and $q$ (see Fig. 1a) were fixed at 0.75 and 0.05 respectively, with $r \in \{0.05, 0.15, 0.25\}$ and $w \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$. The process was then repeated, this time with $p$ and $r$ fixed at 0.75 and 0.15 respectively, with $q \in \{0.05, 0.15, 0.25\}$ and $w$ as before. Sequence length was held fixed throughout at 100,000bp and 200 replicates were simulated at each combination of branch lengths and weight. We found that for almost all branch length combinations ML-JC+H2 was able to recover the correct topology for all replicates. In the entire experiment, only one dataset (out of 13,200) returned the incorrect topology. The results of K&T indicate that ML-JC could not reliably recover the correct topology for all weights for any of the branch length combinations.

The good performance of the GHOST model over the three K&T experiments should be expected in some sense, as ML-JC+H2 enjoys significant advantage over the two alternatives. It is in no way misspecified, having the freedom to fit two classes evolved under the JC substitution model, precisely the conditions used to simulate the data. Conversely, ML-JC has only a single class and therefore is subject to model misspecification. No single set of branch lengths can reproduce the signal present in the simulated alignments. MP is obviously not subject to model misspecification as the method is non-parametric, but it is subject to the long-established artefact of long branch attraction (LBA) (Felsenstein, 1978). Felsenstein showed that having long non-sister branches separated by a relatively short internal branch can result in MP incorrectly inferring the long branches as sisters. Figure 1a shows the two trees used for the classes in the mixture, both sharing the same AB|CD topology. The Class 1 tree has long terminal branches on the A and C lineages, therefore the LBA artefact leads MP to incorrectly favour the AC|BD topology. The Class 2 tree is in a sense the symmetric opposite of the Class 1 tree, it has long terminal edges on the B and D lineages so the result is the same: LBA leads MP to incorrectly infer the AC|BD topology.

Therefore the successful replication of the K&T simulations is a necessary but not sufficient condition for the GHOST model's endorsement. It indicates that the implementation of the GHOST model within IQ-TREE's algorithm structure has been successful, but these simulations are on only four taxa and use the most simple model of sequence evolution. Moreover, they only focus on recovering correct tree topology and not inferring branch length parameters.

*12-taxon simulations.*— We simulated heterotachously-evolved MSAs on a random 12-taxon tree topology under a GTR+FO*H2 model. Using the true GTR+FO*H2 model, IQ-TREE accurately recovered the correct tree topology in all 260 simulated datasets. Figure 3 shows the performance of the GHOST model in recovering the various tree and model parameters for Class 1 of the simulated data. The analagous plots for Class 2 can be found in Supplementary Figures S1 - S4. The results of the 12-taxon simulations clearly show that under the GTR+FO*H2 model IQ-TREE recovered the base frequencies, relative rate parameters and weights to a high degree of accuracy for both classes. With respect to the branch score (BS) (Figs. 3c and S3), we see that the GHOST model again performs very well. The mean BS for the GHOST model approaches that obtained by the partition model as class weight (and therefore share of sequence length in the mixture) increases. This is a very impressive result, given that the partition model enjoys the significant advantage of having full knowledge of which sites were simulated under which class. A BS of zero would imply that the true simulation parameters were inferred for every simulated alignment. Thus, the magnitude of the BS for the partition model can be thought of as a measure of the stochastic simulation error. The difference between the BS for the GHOST and partition models can then be considered the error attributable to losing the knowledge of the partitioning scheme. Clearly this error is negligible in comparison to the simulation error. In Figure 3c, when $w_1 > 0.5$ (or equivalently Fig. S3 when $w_1 < 0.5$), the clear overlap of the error bars (which represent $\pm 2$ standard errors of the mean) suggests that the trees inferred by the GHOST model are not significantly different from those inferred by the partition model. This is a promising result, as in empirical data any partitioning of the MSA is based on assumptions, and therefore introduces a significant potential source of model misspecification. The GHOST model can be applied without any such assumptions.

To demonstrate the ability of the GHOST model to provide meaningful information about which sites might belong to which class, we performed a soft classification on one of the MSAs generated for the 12-taxon simulations. For simplicity we have chosen an MSA where Class 1 and Class 2 are of equal weight. Figure 4 clearly indicates, as one would expect, that the probability of a site belonging to Class 1 is generally higher for those sites that were simulated under the Class 1 parameters. However, given the stochastic element of the simulations, there are some sites simulated under the Class 2 parameters that are classified as having a higher probability of evolving under Class 1, and *vice versa*. For this reason we never attempt to hard classify specific sites to a particular class. Rather we consider a specific site's probability distribution of evolving under all of the classes.

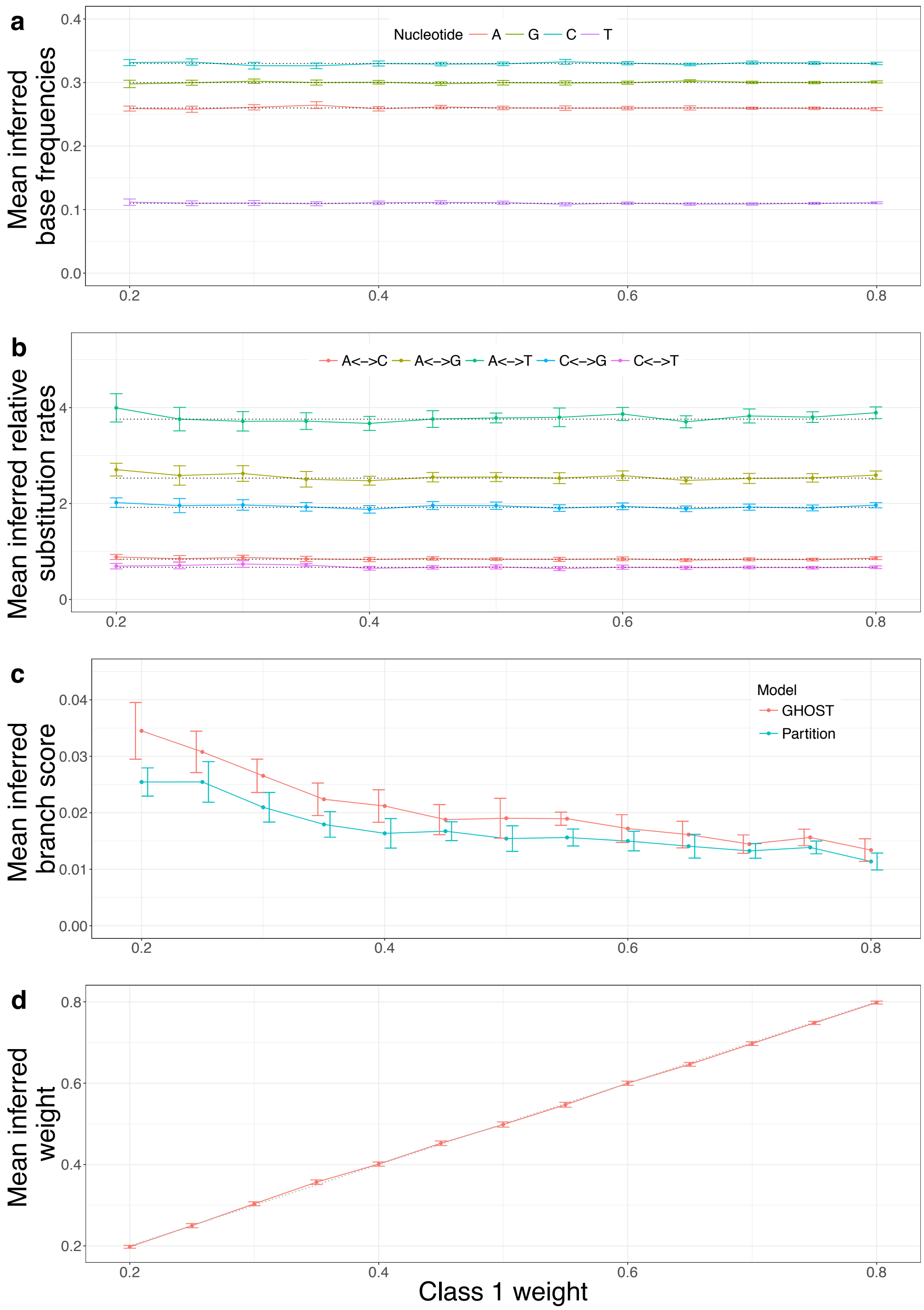


Figure 3: 12-taxon simulations - Class 1 inferred parameters vs Class 1 weight. The data points indicate the mean value of the inferred parameter or statistic, the error bars represent ±2 standard errors of the mean. Dotted lines represent the true parameter value used for data simulation. (a) Base frequencies (b) Relative substitution rates (c) Branch score (BS) for both the GHOST and partition models (d) Inferred Class 1 weight.

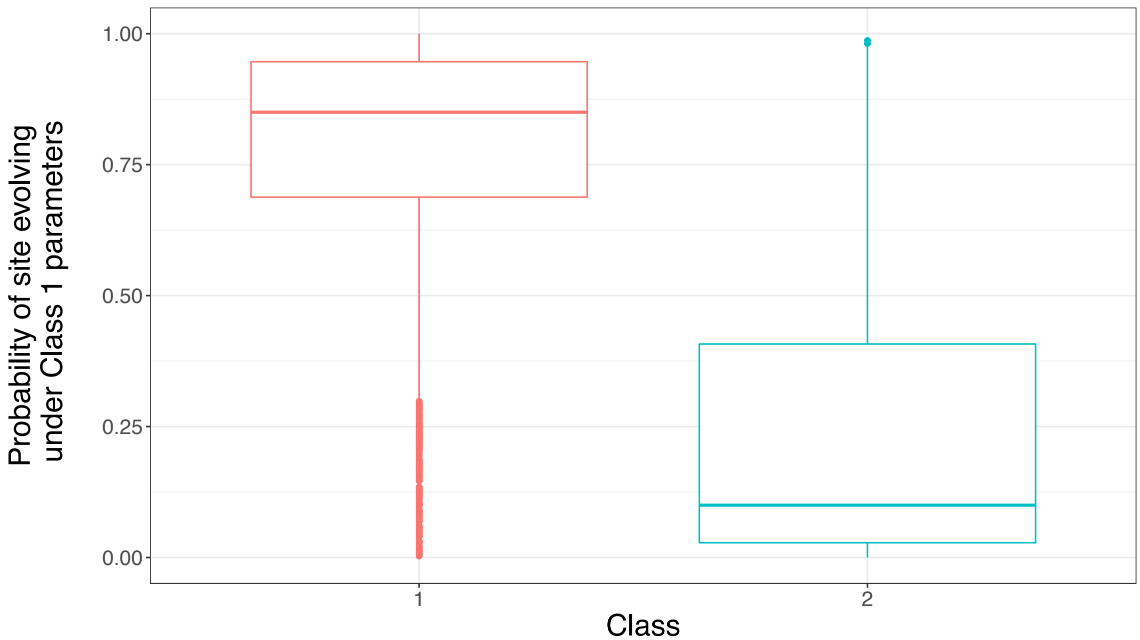


Figure 4: Soft classification of sites to classes - the probability of a site belonging to Class 1 is shown on the y-axis, the two Classes are shown on the x-axis. The boxplots clearly show that in general, sites generated under Class 1 parameters tend to have a higher probability of belonging to Class 1 than sites generated under Class 2.

## *Convergent Evolution of the $Na_v$1.4a Gene Among Teleosts*

To investigate the performance of the GHOST model using empirical data we applied it to the coding region of a sodium channel gene, $\mathrm{Na}_v$1.4a, for 11 teleost species. Zakon et al. (2006) demonstrated the role of this gene in the convergent evolution of the electric organ amongst electric fish species from South America and Africa. AIC determined that GTR+FO*H4 provided the best fit between tree, model and data (Supplementary Fig. S5). The trees inferred by the GHOST model can be found in Figure 5. We then partitioned the electric fish sequence alignment into three partitions, based on codon position (CP). PartitionFinder suggested GTR+FO+G4 (GTR with inferred equilibrium base frequencies plus discrete $\Gamma$ with four classes) for both the CP1 and CP2 partitions, and GTR+FO+I+G4 (same as above but with the inclusion of an invariable sites class) for the CP3 partition. We used IQ-TREE to run the partition model with the models indicated by PartitionFinder. The trees inferred by the partition model can be found in Figure 6.

We labelled the four classes inferred by the GHOST model in order of increasing total tree length (TTL): the 'Conserved Class' ($\mathrm{TTL}_{Cons}$=0.23), the 'Convergent Class' ($\mathrm{TTL}_{Conv}$=0.99), 'Fast-evolving Class A' ($\mathrm{TTL}_{FEA}$=4.06) and 'Fast-evolving Class B' ($\mathrm{TTL}_{FEB}$=4.18). Of particular interest is the Convergent

Class, so named as it corresponds well to Zakon *et al.*'s (2006) hypothesis of convergent evolution of Na$_v$1.4a among the South American and African electric fish clades. The convergent class tree displays much more evolvution in the electric rather than the non-electric fish lineages (Fig. 7). This is indicative of either a relaxation of purifying selection pressure, an introduction of positive selection pressure or a combination of both. The notable exception is the Brown Ghost Knifefish, which appears relatively conserved. The Brown Ghost Knifefish is unique amongst the other electric fish in the dataset, in that its electric organ has evolved from neural rather than muscle tissue. Consequently in the Brown Ghost Knifefish the *Na$_v$1.4a* gene is still expressed in muscle, just as it is in the non-electric fish. The clear distinction in terminal edge length between the Brown Ghost Knifefish and the other electric fishes is obvious and compelling. It provides strong evidence that the GHOST model has indeed identified a subtle component of the historical signal related to the convergent evolution of *Na$_v$1.4a*, as opposed to returning an arbitrary combination of numerical parameters that happen to maximize the likelihood function. The ability of the GHOST model to isolate such a small component of the signal (the inferred weight of the convergent class being 0.13, the smallest of the 4 classes) is most encouraging. Furthermore, we can expect that the sites belonging with high probability to the convergent class are likely to have been influential in the functional development of the electric organ.

*Soft classification of sites to classes.*— The soft classification of sites to classes facilitates the prospective identification of functionally important sites in an alignment. Zakon et al. (2006) report several amino acid sites from the dataset that are influential in the inactivation of the sodium channel, a process critical to electric organ pulse duration. Figure 8a shows that these sites generally have a higher than average probability of belonging to the convergent class in at least one codon position. For example, at amino acid site 647 an otherwise conserved proline (codon CCN) is replaced by a valine (GTN) in the Pintailed Knifefish and a cysteine (TGY) in the Electric Eel. Unique substitutions at codon positions 1 and 2 are necessary for both of these amino acid replacements and we find these two sites have a very high probability of belonging to the convergent class. With this result in mind, for each amino acid we summed the probability of codon positions 1 and 2 belonging to the Convergent Class. Figure 8b shows the results for the eight amino acid sites with the highest score. Comparing the magnitude of these bars with those of the amino acids in Figure 8a (which are known to be functionally important), one can suspect that these amino acids might also be critical to the operation of the sodium channel gene. Given that there are many other sites in the alignment with a high probability of belonging to the convergent class, one can envisage the

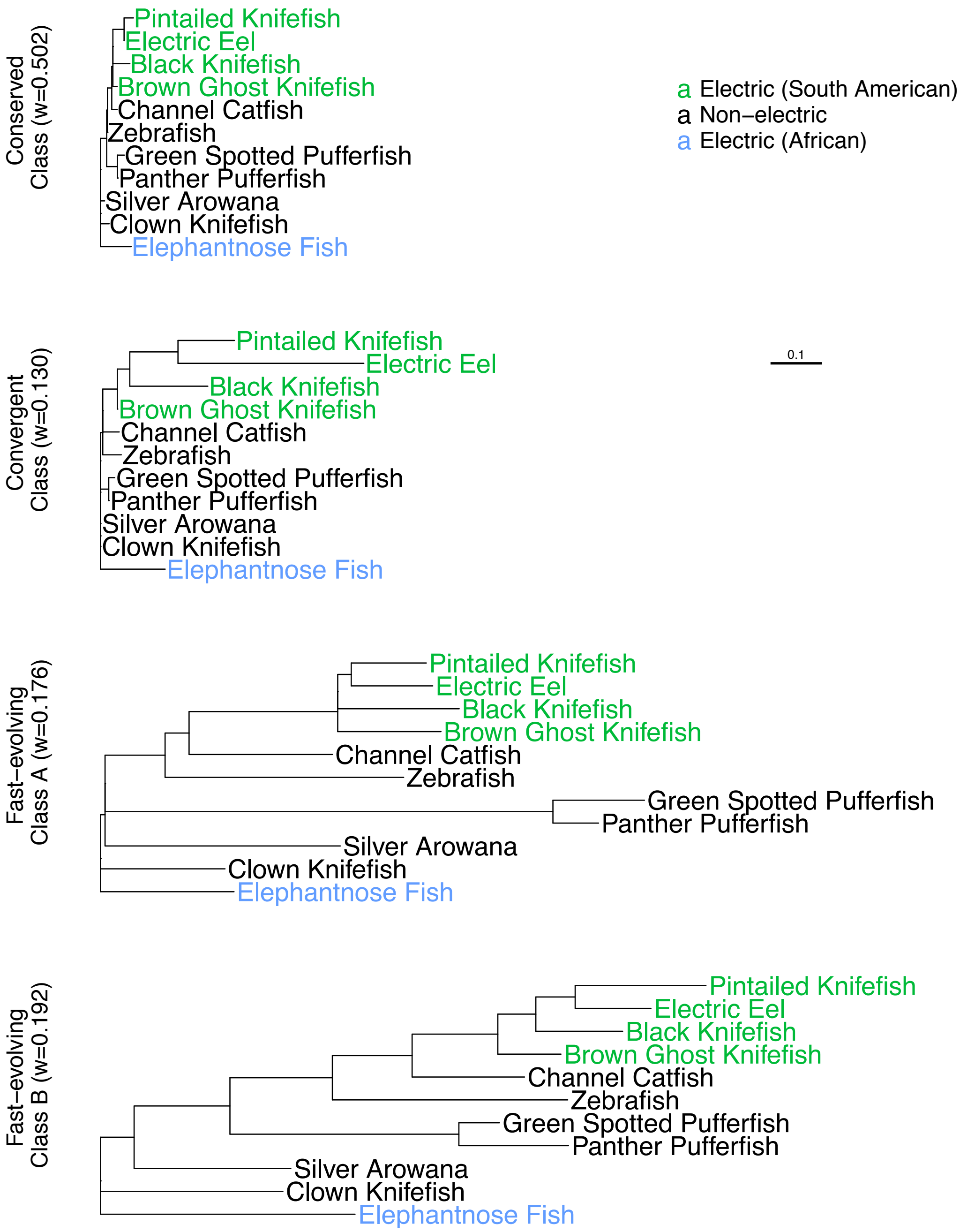


Figure 5: The four trees inferred under the General Time Reversible, four class mixture model (GTR+FO*H4) for the electric fish data. We can clearly see the variability of the branch lengths among the four classes. The classes are displayed in order of increasing tree size, as determined by the sum of the branch lengths. We refer to this as the total tree length (TTL): $\mathrm{TTL}_{Cons} = 0.23$, $\mathrm{TTL}_{Conv} = 0.99$, $\mathrm{TTL}_{FEA} = 4.06$ and $\mathrm{TTL}_{FEB} = 4.18$.

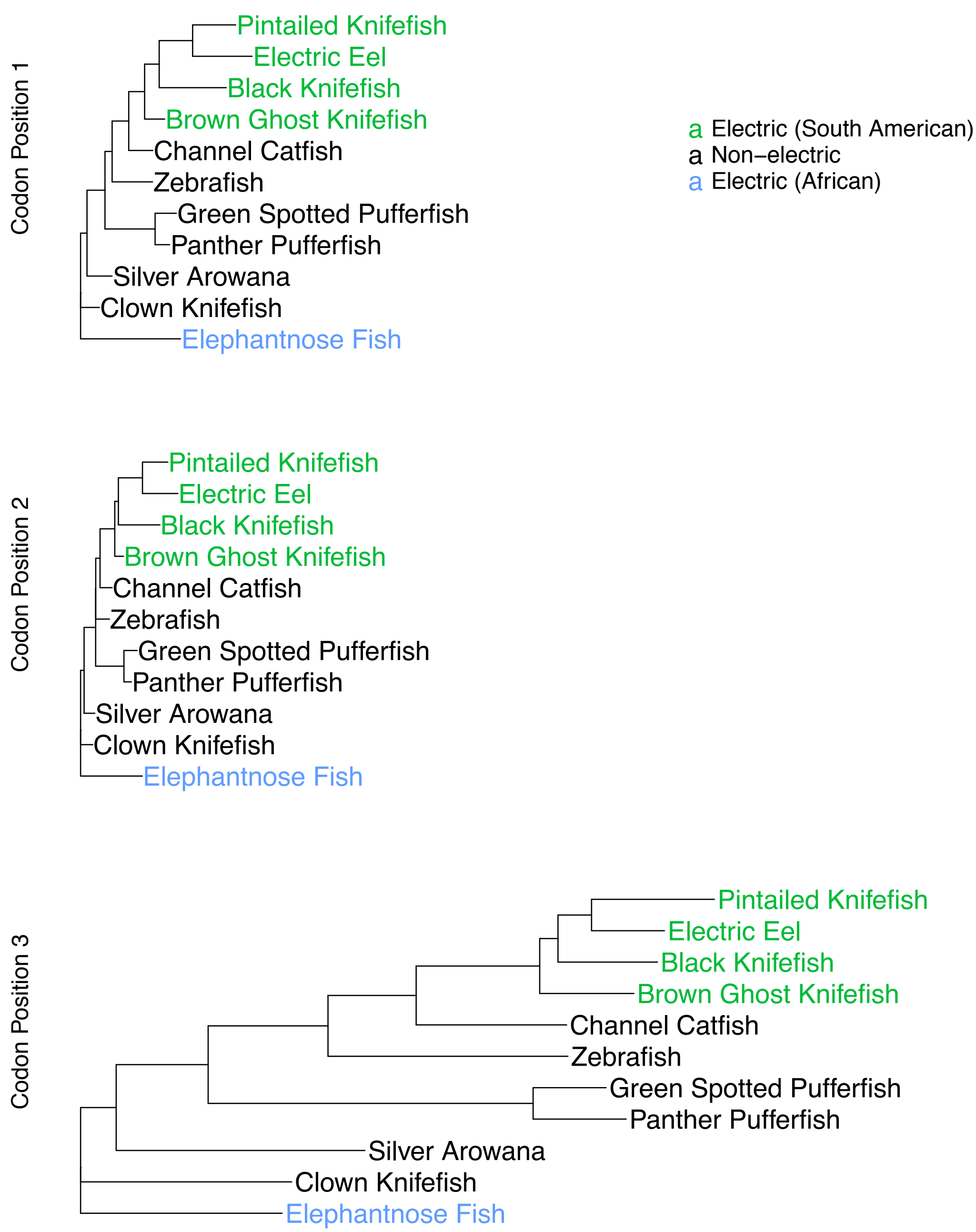


Figure 6: The three trees inferred under the edge-unlinked partition model for the electric fish dataset, with the alignment partitioned based on codon position (CP). The CP1 and CP2 partitions used a GTR+FO+G model, while the CP3 partition used a GTR+FO+I+G model.

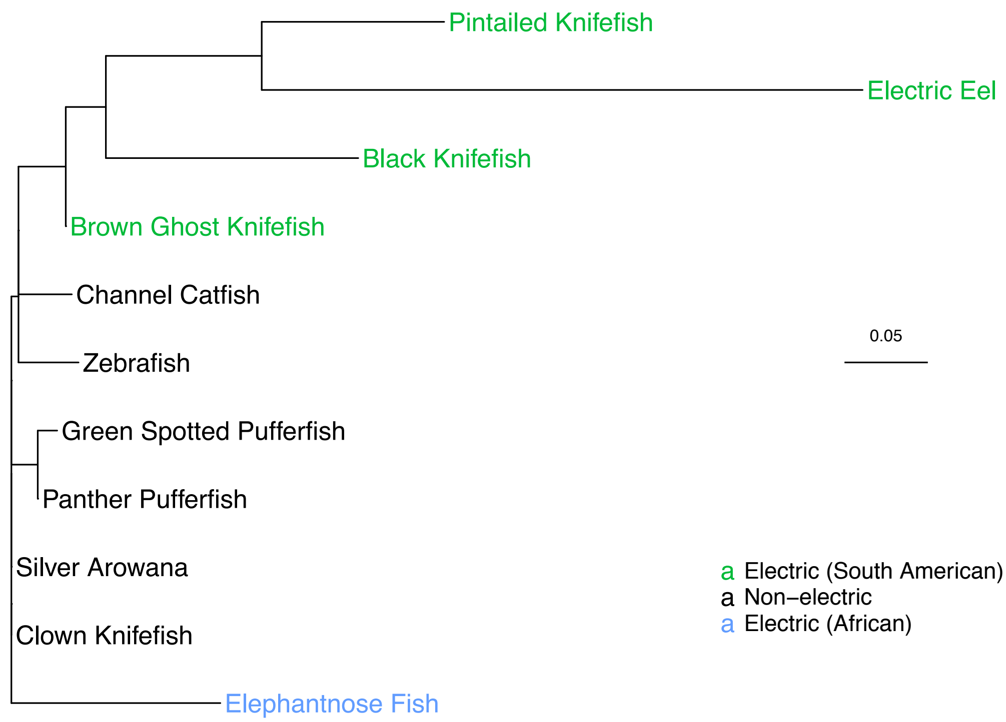


Figure 7: The convergent class inferred by ML-GTR+FO*H4. The 11 fish species comprised four South American electric fish (green), one African electric fish (blue), and six non-electric fish (black) from various locations. The tree for this class shows that in comparison to the electric fish, the non-electric species are relatively conserved.

GHOST model helping to identify sites of potential functional importance in an alignment, thereby focusing the experimental work of biologists.

In addition to providing insight on an individual site basis, the soft classification can also help to inform us about the nature of the classes themselves. Summing the weighted TTLs for each of the inferred classes results in an estimated 1.766 substitutions per site under the inferred model. Table 1 reports the contributions to this figure, stratified by codon position and class. If class membership and codon position were independent attributes of each site then we should expect the contribution of each codon position to be approximately one third for each class. This is not what we observe. Overall we can see that sites in CP1(23%) and CP2 (16%) contribute only 39% of the total of 1.766 substitutions per site. However, within the Conserved and Convergent Classes, sites in CP1 and CP2 are responsible for 90% and 76% of their contribution respectively. This would suggest that a comparatively larger proportion of the substitutions attributed to these classes are non-synonymous: resulting in amino acid replacements that influence the fitness of the organism. We can therefore conclude that even though the Conserved and Convergent Classes are smallest (as determined by substitutions per site), they appear to be the primary catalyst of evolution via natural selection within $Na_v1.4a$ amongst these species.

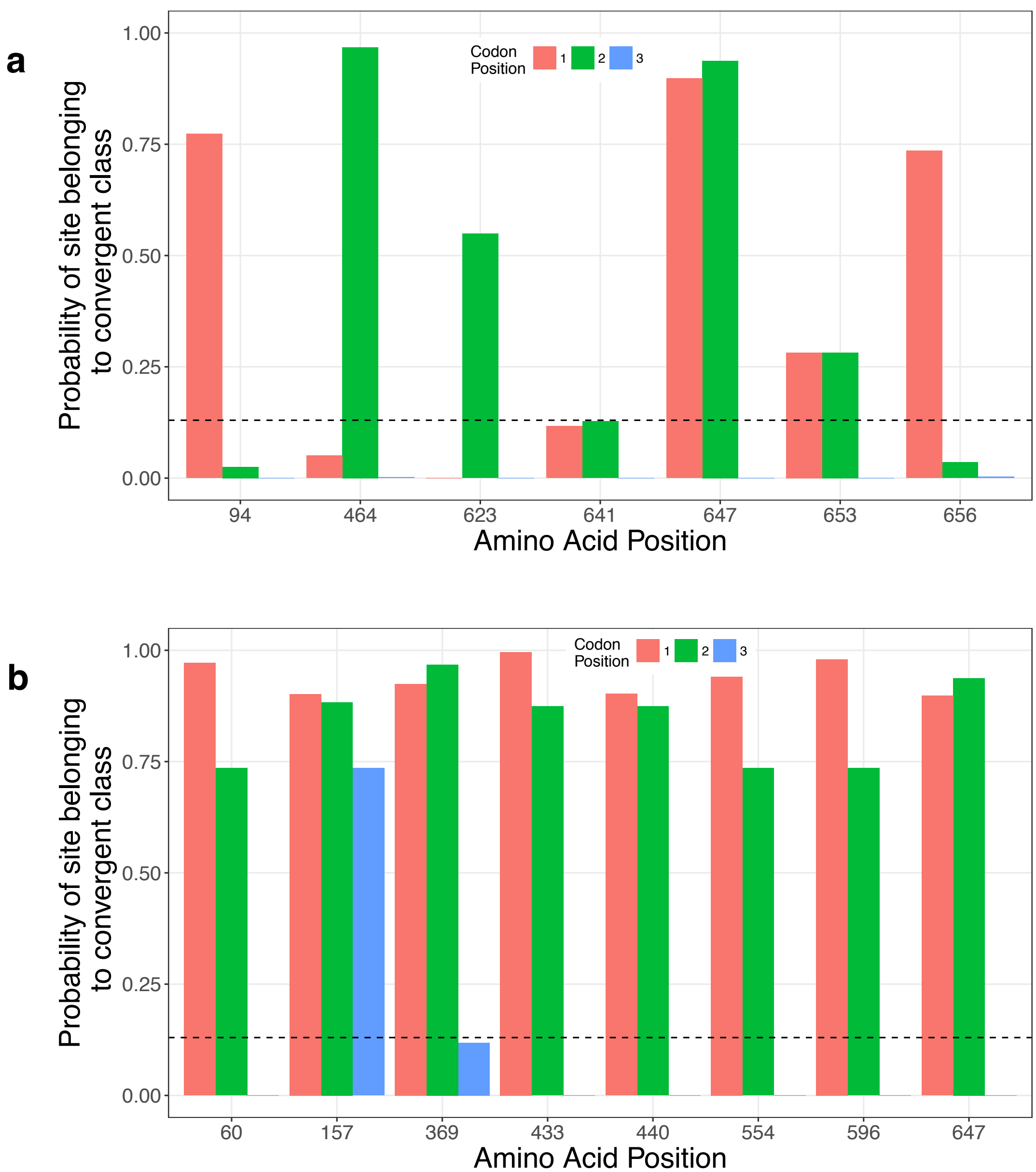


Figure 8: Probability of sites belonging to the convergent class by codon position. (a) The amino acid positions selected correspond with those identified by Zakon et al. (2006) as being functionally important to the inactivation of the $Na^+$ gene. The horizontal dotted line at 0.13 represents the average probability of belonging to the convergent class over all sites in the alignment. (b) The amino acid positions selected correspond to those with the highest probability of belonging to the convergent class, summed across the first two codon positions.

| Class | CP1 | CP2 | CP3 | Subs/site |
|---|---|---|---|---|
| Conserved | 0.049 (41%) | 0.058 (49%) | 0.012 (10%) | **0.119** |
| Convergent | 0.051 (40%) | 0.047 (36%) | 0.031 (24%) | **0.129** |
| Fast-evolving A | 0.135 (19%) | 0.076 (11%) | 0.504 (70%) | **0.715** |
| Fast-evolving B | 0.175 (22%) | 0.100 (12%) | 0.528 (66%) | **0.803** |
| All Classes | 0.410 (23%) | 0.280 (16%) | 1.076 (61%) | **1.766** |

Table 1: Expected number of substitutions per site (bold), weighted by class and separated by codon position (CP). For each inferred class, the expected substitutions per site are calculated by multiplying the total tree length (TTL) by the class weight. The CP1, CP2 and CP3 columns show the contribution to these figures from only the sites within each CP. The grand total indicates that under the parameters inferred by ML-GTR+H4 we would expect 1.766 nucleotide substitutions per site. We can then see, for example, that the Convergent Class is responsible for 0.129 of these substitutions per site. Finally, of the 0.129 substitutions per site attributable to the Convergent Class, 0.051 (or 40%) is the contribution from sites in CP1, 0.047 (36%) is the contribution from sites in CP2 and 0.031 (24%) is the contribution from sites in CP3.

*Comparison to the Partition Model.*— It is apparent upon examination of the trees in Figure 6 that the evidence of convergent evolution highlighted by the GHOST model (Fig. 7) has not been recovered by the partition model. None of the three trees in Figure 6 have the distinctive pattern, whereby the majority of the total tree length is associated with the electric fish species (with the exception of the Brown Ghost Knifefish). The reason that the partition model failed to recover this signal is clear when considering the contribution of each CP to the Convergent Class. Table 1 indicates that the substitutions associated with the Convergent Class are attributable to CP1 sites (40%), CP2 sites (36%) and CP3 sites (24%). The partition model constrains the analysis, such that sites in different CPs are modeled independent of each other. It is impossible for a model constrained in such a way to recover the convergent evolution signal, or any other signal whose components are distributed across multiple partitions. The decision to partition the data based on codon position may make sense superficially, but in doing so the analysis is constrained and the results are compromised. We no longer have the ability to uncover the evolutionary stories concealed within the data. We can only hope to obtain those stories that happen not to conflict with the assumptions and constraints that have been placed on the analysis *a priori*. Minimizing these assumptions and constraints where possible, while computationally expensive, is necessary in order to illuminate the evolutionary history without distorting it in the process.

## *On the Identifiability of the GHOST Model*

An ongoing concern with regard to parameter-rich mixture models has been whether or not they are identifiable. There are several examples of theoretically non-identifiable mixture models in the literature (Matsen and Steel, 2007; Štefankovič and Vigoda, 2007b). These examples have inspired much theoretical work on the identifiability or otherwise of different types of phylogenetic mixture models (Allman and Rhodes, 2006; Štefankovič and Vigoda, 2007a; Allman et al., 2008; Allman and Rhodes, 2008; Allman et al., 2011). Of particular interest to the current study, Allman et al. (2011) showed that for a single topology, four taxa, two-class mixture under the JC model, only the tree topology is identifiable but not the branch lengths. This provides a theoretical justification for the procedure carried out by K&T (and replicated here), measuring performance of the models based only on recovery of the topology and paying no attention to recovery of branch length parameters. With regard to the identifiability of the GHOST model more generally, we rely on a result from Rhodes and Sullivant (2012). They established an upper bound on the number of classes for which tree topology, branch lengths and model parameters are identifiable, as a function of the number of character states and the number of taxa. For the simulations we carry out in the current study, with 12 taxa and four character states, the model is identifiable up to a maximum of 15 classes. In the case of the electric fish dataset, with four character states and only 11 taxa, the model is identifiable up to 11 classes. However, there is a technical caveat. The result is shown based on assuming a general Markov model across the tree. There are specific choices of parameters that can result in non-identifiability, but these are of little concern in practical data analysis. Problems arise only when the parameters selected collapse the parameter space to some lower dimension. For example, we could fit the GTR model but if we chose parameters such that all base frequencies were equal and all substitution rates were equal then we are in fact using a JC model, and identifiability may be compromised. However, these technical examples of non-identifiability are not relevant in practice, as in the absence of any constraints there is no likelihood of inferring parameters that collapse the parameter space in such a way.

# Conclusion

Heterotachy has been somewhat of an Achilles heel for ML since K&T published their study. The implementation of the GHOST model in IQ-TREE represents a positive advance for ML based phylogenetic inference. Through minimization of model assumptions the GHOST model offers significant flexibility to infer heterotachous evolutionary processes, illuminating historical signals that might otherwise remain hidden. The GHOST model seems well suited to the analysis of phylogenomic datasets, commonly used to address deep phylogenetic questions. While we only present the method and one single-gene empirical example in the current paper, forthcoming empirical studies will compare the performance of the GHOST model to currently popular phylogenomic analysis tools, such as partition and CAT models. One can also envisage many other potential uses for the GHOST model. It could be applied to datasets for which the topology is poorly supported or disputed. It could also provide more accurate parameter estimates, leading to sounder divergence date estimation. The model provides intuitive, biologically meaningful visualizations of the different evolutionary pressures that act on a group of taxa. Structural biologists may find it useful for highlighting functionally important areas within proteins. We have demonstrated its use as a method for identifying changes in selection pressure, as well as bringing to light evidence of convergent evolution. Similarly, one can envisage the GHOST model illuminating the subtle evolutionary relationships between hosts and parasites, disease and immune cells, or the countless evolutionary arms races that are observed throughout the natural world.

## *ACKNOWLEDGEMENTS*

The authors would like to thank Elizabeth Allman and John Rhodes for helpful discussion about the manuscript.

B.Q.M. and A.v.H were supported by the Austrian Science Fund (FWF I-2805-B29).

*COMPETING FINANCIAL INTERESTS.*— The authors declare no competing financial interests.

# References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Allman, E. S., Ané, C., and Rhodes, J. A. (2008). Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Advances in Applied Probability*, pages 229–249.

Allman, E. S., Petrovic, S., Rhodes, J. A., and Sullivant, S. (2011). Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(3):710–722.

Allman, E. S. and Rhodes, J. A. (2006). The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *Journal of Computational Biology*, 13(5):1101–1113.

Allman, E. S. and Rhodes, J. A. (2008). Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. *Mathematical Biosciences*, 211(1):18–33.

Baele, G., Raes, J., Van de Peer, Y., and Vansteelandt, S. (2006). An improved statistical method for detecting heterotachy in nucleotide sequences. *Molecular Biology and Evolution*, 23(7):1397–1405.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, pages 1–38.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Fitch, W. M. and Margoliash, E. (1967). A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case. *Biochemical Genetics*, 1(1):65–71.

Gadagkar, S. R. and Kumar, S. (2005). Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Molecular Biology and Evolution*, 22(11):2139–2141.

Holmquist, R., Goodman, M., Conroy, T., and Czelusniak, J. (1983). The spatial distribution of fixed mutations within genes coding for proteins. *Journal of Molecular Evolution*, 19(6):437–448.

Jayaswal, V., Wong, T. K., Robinson, J., Poladian, L., and Jermiin, L. S. (2014). Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Systematic Biology*, 63(5):726–742.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. In Munro H.N. *Mammalian Protein Metabolism*, pages 21–123, New York: Academic Press.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., and Jermiin, L. S. (2017). Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589.

Kolaczkowski, B. and Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431(7011):980–984.

Kuhner, M. K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–468.

Lanfear, R., Calcott, B., Ho, S. Y., and Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, 29(6):1695–1701.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Lopez, P., Casane, D., and Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution*, 19(1):1–7.

Matsen, F. A. and Steel, M. (2007). Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, 56(5):767–775.

Meade, A. and Pagel, M. (2008). A phylogenetic mixture model for heterotachy. In *Evolutionary Biology from Concept to Application*, pages 29–41. Springer.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4):571–581.

Pagel, M. and Meade, A. (2005). Mixture models in phylogenetic inference. *Mathematics of Evolution and Phylogeny*, pages 121–142.

Philippe, H. and Lopez, P. (2001). On the conservation of protein sequences in evolution. *Trends in Biochemical Sciences*, 26(7):414–416.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*, 5(1):50.

Rambaut, A. and Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences: CABIOS*, 13(3):235–238.

Rhodes, J. A. and Sullivant, S. (2012). Identifiability of large phylogenetic mixture models. *Bulletin of Mathematical Biology*, 74(1):212–231.

Spencer, M., Susko, E., and Roger, A. J. (2005). Likelihood, parsimony, and heterogeneous evolution. *Molecular Biology and Evolution*, 22(5):1161–1164.

Steel, M. (2005). Should phylogenetic models be trying to fit an elephant? *Trends in Genetics*, 21(6):307–309.

Štefankovič, D. and Vigoda, E. (2007a). Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *Journal of Computational Biology*, 14(2):156–189.

Štefankovič, D. and Vigoda, E. (2007b). Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Systematic Biology*, 56(1):113–124.

Wang, H.-C., Li, K., Susko, E., and Roger, A. J. (2008). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evolutionary Biology*, 8(1):331.

Whelan, N. V. and Halanych, K. M. (2017). Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. *Systematic Biology*, 66(2):232–255.

Wu, J. and Susko, E. (2009). General heterotachy and distance method adjustments. *Molecular Biology and Evolution*, 26(12):2689–2697.

Wu, J. and Susko, E. (2011). A test for heterotachy using multiple pairs of sequences. *Molecular Biology and Evolution*, 28(5):1661–1673.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314.

Zakon, H. H., Lu, Y., Zwickl, D. J., and Hillis, D. M. (2006). Sodium channel genes and the evolution of diversity in communication signals of electric fishes:

convergent molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3675–3680.