HHS Public Access

Author manuscript

*Infect Genet Evol*. Author manuscript; available in PMC 2020 November 25.

Published in final edited form as:

*Infect Genet Evol*. 2020 November ; 85: 104501. doi:10.1016/j.meegid.2020.104501.

# Going back to the roots: Evaluating Bayesian phylogeographic models with discrete trait uncertainty

**Matteo A. Vaiente**[a,b], **Matthew Scotch**[a,b,*]

[a]Biodesign Center for Environmental Health Engineering, Arizona State University, 727 E. Tyler St, Tempe, AZ 85281, USA

[b]College of Health Solutions, Arizona State University, 500 N 3rd St, Phoenix, AZ 85004, USA

## Abstract

Phylogeography is a popular way to analyze virus sequences annotated with discrete, epidemiologically-relevant, trait data. For applied public health surveillance, a key quantity of interest is often the state at the root of the inferred phylogeny. In epidemiological terms, this represents the geographic origin of the observed outbreak. Since determining the origin of an outbreak is often critical for public health intervention, it is prudent to understand how well phylogeographic models perform this root state classification task under various analytical scenarios. Specifically, we investigate how discrete state space and sequence data set influence the root state classification accuracy. We performed phylogeographic inference on several simulated DNA data sets while i) increasing the number of sequences and ii) increasing the total number of possible discrete trait values. We show that phylogeographic models tend to perform best at intermediate sequence data set sizes. Further, we demonstrate that a popular metric used for evaluation of phylogeographic models, the Kullback-Leibler (KL) divergence, both increases with discrete state space and data set sizes. Further, by modeling phylogeographic root state classification accuracy using logistic regression, we show that KL is not supported as a predictor of model accuracy, indicating its limited utility for assessing phylogeographic model performance on empirical data. These results suggest that relying solely on the KL metric may lead to artificially inflated support for models with finer discretization schemes and larger data set sizes. These results will be important for public health practitioners seeking to use phylogeographic models for applied infectious disease surveillance.

### Keywords

Phylogenetics; Phylogeography; Bayesian statistics; Model evaluation

## 1. Introduction

For the last decade, researchers have used Bayesian phylogeography (Lemey et al., 2009) to investigate the epidemiology of rapidly evolving viral pathogens with the aim of



*Corresponding author at: Biodesign Center for Environmental Health Engineering, Arizona State University, 727 E. Tyler St, Tempe, AZ 85281, USA. Matthew.Scotch@asu.edu (M. Scotch).

elucidating the contributions of discrete traits, often geographic location, to the propagation and persistence of disease outbreaks. Numerous examples are available in the literature and recent compelling studies have focused on recent Ebola (Dudas et al., 2017), Zika (Grubaugh et al., 2017), West Nile (Swetnam et al., 2018 and influenza H3N2 (Magee et al., 2017), H9N2 (Yang et al., 2019 and H5N2 (Hicks et al., 2020) virus outbreaks. Bayesian phylogeographic discrete trait diffusion models require both a set of molecular sequences annotated with isolate sampling times and metadata describing a discrete traits of interest. Then, discrete trait diffusion is modeled as a continuous time Markov chain which evolves across a phylogenetic tree topology. Modeling discrete trait diffusion in this way enables computation of the model likelihood via Felsenstein's pruning algorithm. (Felsenstein, 1981). Briefly, the algorithm proceeds via a post-order tree traversal and calculates the partial likelihood, backwards in time, for all trait states at internal tree nodes using the aforementioned Markov model. In a standard analysis, sequence records with discrete trait metadata are assumed to have a probability mass function (PMF) which assigns all mass to the observed trait. Concretely, the partial likelihood vectors at the tips are one-hot encoded as a vector with dimension equal to the cardinality of the discrete trait state space; the total number of distinct values a discrete trait may take.

For many researchers, the predominant method of obtaining publicly available molecular sequences for phylogeographic analysis is through the use of GenBank (Benson et al., 2018), a nucleotide sequence database maintained by the National Center for Biotechnology Information or NCBI (Sayers et al., 2020). Usually, researchers parse the *country* field in a GenBank record in order to obtain geographic metadata for phylogeography studies. However, metadata representing geographic locations, host age and species, and other discrete characteristics are not required when submitting new molecular sequences to GenBank databases leading to numerous records with missing metadata. For example, previous work by Scotch et al. (Scotch et al., 2011) which linked virus sequence records to geographical entities in the GeoNames ontology (Vatant and Wick, 2012) found that 80% of GenBank records contain "insufficient" geographic metadata. In this case, they defined geographic metadata insufficiency as data regarding the location of infected host (LOIH) at lst-level administrative division (ADM1) or greater granularity. This means geographic metadata were typically informative for the LOIH at the state (province) or country level but seldom contained information on finer geographic entities such as counties or cities. Similarly, Tahsin et al. (Tahsin et al., 2014) reported the proportion of GenBank virus records with insufficient geographic data to be between 64% and 90%. Many real-world public health tasks require modeling transmission patterns at high geographic granularity to inform control strategies necessary to curb disease spread, such as modeling viral diffusion between counties within a state's boundary. Therefore, the insufficiency of GenBank metadata represents a major barrier to the implementation of virus phylogeography for applied public health surveillance.

This paucity of high resolution geographic metadata has inspired researchers to develop new methods and tools to ascertain the LOIH for viral sequences represented in GenBank records (Tahsin et al., 2014; Tahsin et al., 2017; Magge et al., 2018). Indeed, available pipelines for discerning the LOIH are configured such that they output not only the most probable location for a specific sequence, but also a vector of other possible locations along with

their relative probabilities (Magge et al., 2018). Building on the availability of these new pipelines, Scotch et al. (Scotch et al., 2019) introduced the notion of incorporating sampling uncertainty into phylogeographic analyses. This parameterization of the standard discrete trait diffusion model involves assigning a prior PMF to the set of possible geographic locations for each tip with an uncertain LOIH. The additional uncertainty in LOIH is easily incorporated into the likelihood calculation using the standard pruning algorithm (Felsenstein, 1981) by defining the partial likelihood vectors at the tips to be the desired PMF.

We note that since phylogeographic discrete trait diffusion models can be applied to general discrete traits, so too the phylogeographic uncertain trait model (UTM) introduced by Scotch et al. (Scotch et al., 2019) can be used to assign prior PMFs to tips missing arbitrary discrete trait information. In the case of non-geographic discrete traits, where relatively little attention has been paid to resolving insufficient metadata, this provides two distinct advantages to standard analysis workflows: it provides researchers with a coherent method of specifying a priori beliefs about unobserved traits and effectively increases the data set size by including sequences which would otherwise be excluded from an analysis due to missing metadata. Previously, phylogeographic researchers studying non-geographic discrete traits, such as host species or age, were left with two options for sequences with missing metadata: to manually curate locations for each unresolved record, or, to exclude these sequences from phylogeographic analysis (Magee and Scotch, 2018; Dellicour et al., 2019). The former option is extremely labor intensive, difficult to replicate, and cannot be scaled to large data sets. Conversely, the latter has the disadvantage of reducing the amount of data included in a given phylogeographic analysis, which may induce biases in rate matrix parameters if the records with particular discrete traits are selectively over/underrepresented in the sample (De Maio et al., 2015)

Though phylogeographic discrete trait diffusion models remain a popular and promising tool for epidemiological inference, relatively few studies aim to quantify the statistical performance of these methods under various analysis conditions (Magee et al., 2017; Magee and Scotch, 2018; De Maio et al., 2015; Lemey et al., 2014). Particularly, phylogeographic discrete trait diffusion models are increasingly used for inference on large discrete state spaces and data set sizes, especially as pathogen genome sequencing continues to become a routine part of outbreak response. For example, recent studies commonly use state space sizes ranging from 10 to 56 discrete entities (Dudas et al., 2017; Magee et al., 2017; Lemey et al., 2014). Paradoxically, a rigorous examination of model performance with respect to increasing state space and data set sizes is currently absent from the literature (Lemey et al., 2009; Magee et al., 2017; De Maio et al., 2015; Lemey et al., 2014). Further, given its recent introduction, the statistical performance of the phylogenetic UTM (Scotch et al., 2019) compared to other established model parameterizations is yet to be established. Since the quantity of interest from phylogeographic discrete trait diffusion models is often the most likely state at the root of the phylogeny, we select this *root state classification task* as the primary axis on which we evaluate model performance. In this paper, we take a simulation-based approach to investigating the performance of phylogeographic discrete trait diffusion models, paying special attention to the roles of data set and discrete state space size for performance on the root state classification task. Simultaneously, we

compare the performance of the alternative phylogenetic UTM parameterizations against a reference model which omits sequences with missing metadata. This work represents, to our knowledge, a unique contribution to understanding the performance of popular phylogeographic discrete trait diffusion models under various analysis conditions and will be useful to researchers and public health practitioners tasked with designing phylogeographic studies using publicly available pathogen sequences.

## 2. Methods

### 2.1. Study design

There are several ways in which the UTM can be implemented depending on the prior beliefs of the analyst for the missing discrete state values. For example, if no information a priori is available with respect to a discrete trait of interest with a molecular sequence, a reasonable choice may be to use a uniform prior over all possible trait values (“uniform”). On the other hand, it may be the case that a researcher wants to incorporate their prior beliefs on the relative probability of each state into the analysis. While this prior PMF can take many forms (indeed, there are infinitely many of them), we focused on two possibilities expected to arise frequently in practice: the researcher assigns most of the prior mass to the correct discrete trait (“informed”), or, alternatively, most of the mass is assigned to the incorrect state (“misspecified”). Concretely, for “informed” models, we assigned 50% of the prior mass to the correct discrete trait, and divided the remaining mass uniformly across the remaining states. Conversely, for “uninformed” models, we reverse the parameterization such that 50% of the prior mass is placed on an incorrect discrete state (chosen uniformly from the set of incorrect discrete traits) and the remaining mass distributed uniformly among the remaining traits. We believe these three options (uniform, informed, misspecified) are representative of choices likely to be made in practice. Prior to the introduction of the phylogenetic UTM (Scotch et al., 2019), researchers often exclude sequences with missing metadata from phylogeographic analyses. We specified this modeling approach (“drop”) as the reference to which we compared alternative UTM parameterizations.

We utilized a fully factorial, completely randomized design to quantify the relationships between discrete state space size, data set size and phylogeographic model performance. We defined 150, 250 and 500 sequences, respectively, as the factor levels for data set size. Similarly, we defined discrete state space sizes of 4, 8, and 16 states as factor levels for discrete state space size. We then simulated 25 replicate data sets under for each of 9 combinations of the aforementioned factor levels resulting in 225 data sets. We analyzed each data set using the phylogeographic UTM with either: i) informed ii) misspecified or iii) uniform prior PMFs. We also analyzed each data set after excluding sequences with missing metadata to serve as a reference for model comparison. Using this design, we analyzed 225 data sets under each of the 4 alternative model parameterization for a total of 900 independent model analyses. We discuss the data simulation procedure including generation of missing discrete traits in the following sections and provide a visual summary in Fig. 1.

## 2.2. Data simulation

**2.2.1. Phylogenetic trees—**Since virus sequences represent isolates from individuals infected during epidemics, we simulated phylogenetic trees using the serially-sampled birth-death SIR model (SSBD-SIR) (Stadler et al., 2013). The SSBD-SIR model requires specification of 3 parameters: $\beta$, $\gamma$, $\phi$ representing the transmission (birth), recovery (death) and sampling rates, respectively. An equivalent specification can be made in terms of RO, the basic reproduction number, by using a fixed recovery rate. We selected simulation parameters to be similar to general, seasonal influenza outbreaks with an RO value of 1.4 and assuming an infectious period ($\phi^{-1}$) of one week, consistent with observed epidemiological patterns (Connolly, 2005). Finally, we specified a sampling rate of 20%, reflecting a densely sampled epidemic scenario. We simulated trees until either 150, 250 or 500 tips were sampled. We performed tree simulation using the TreeSim package in R (Stadler, 2011).

**2.2.2. Sequence data—**We converted branch lengths to units of substitutions by assuming a strict molecular clock model with a rate of $1 \times 10^{-3}$ substitutions per site, per year to allow for sequence simulation on each tree. We utilized an HKY + $\Gamma$ model of nucleotide substitution with 4 rate categories, as is commonly used for modeling influenza molecular sequence evolution. We simulated 1750 base-pair (bp) sequences using the aforementioned parameters using Phyx (Brown et al., 2017).

**2.2.3. Discrete and missing trait simulation—**We simulated the evolution of discrete traits on each phylogenetic tree by assuming traits evolved with a rate of 0.1 substitutions per site per year. Since a key goal of our study is to estimate the performance of phylogeographic trait models on a variety of state spaces, we simulated traits with 4, 8 or 16 states. We used random symmetric Markov matrices with gamma distributed rate parameters. To generate missing traits, we used a binomial sampling process on the observed traits where each trait is dropped with 20% probability. A final data set includes sequences written in FASTA format with discrete trait and sampling time information annotated in the description line.

## 2.3. Bayesian phylogenetic and phylogeographic inference

We performed phylogenetic and phylogeographic inference using BEAST v 1.10.1 (Suchard et al., 2018). We modeled molecular evolution using an HKY + $\Gamma$ model with 4 rate categories, reflecting the conditions under which the data were simulated. We employ a flexible nonparametric skygrid prior since we know a priori that the population of infected individuals follow non-linear SIR-type dynamics. We specified a symmetric Markov model for inference of discrete trait evolution, again driven by our choice of data simulation conditions. To estimate divergence times, we fixed tip dates as the dates of sampling recorded during each simulation. We ran the each MCMC for 100 million iterations, sampling every 10,000 steps and removed the first 20% as burn-in. We diagnosed convergence of the MCMC procedure using Tracer v1.7.1 (Rambaut et al., 2018) checking that all model parameters had Effective Sample Sizes (ESS) of 200 or greater.

### 2.4. Model evaluation

Inferring the most likely state at the root of the phylogeny, akin to identifying the location or host species where an outbreak started, is a key output of phylogeographic discrete trait diffusion models. We can evaluate the performance of popular phylogeographic techniques by treating the root state identification problem as a *classification* problem, borrowing terms from the machine learning literature. Several metrics are available to summarize the a classification model with respect to its performance on a classification task. Given a classification model and labeled test data one can compute the *accuracy* of a classification model: the proportion of instances it classifies correctly. In phylogeography, a central task is to correctly classify the most likely state at the root of a phylogeny. We recorded the root state from which each simulation was initialized and calculated the accuracy of phylogeographic models when given more data (sequences) or when performing inference over increasing discrete state spaces. Since the result of our Bayesian phylogeographic analysis is a posterior distribution over root states, we follow standard practice in classification model evaluation and selected the most likely posterior state $j$ as the root state "prediction" output by our models.

$$\hat{j} = \max_{j} \mathscr{P}(X = j \mid \theta)$$

Though informative, accuracy does not fully describe the characteristics of a given classification model. A common measure of classification model performance is the cross entropy. This is generally interpreted at the number of bits needed to transmit data from a source distribution when using a model of that distribution. In the context of classification model evaluation, we can interpret cross entropy as a kind of "distance" between the posterior distribution estimated by our model and the true root state distributions. Defining the true root state distribution $P_j$ as a one-hot encoded vector permits computation of the cross entropy using:

$$C = -\sum_{j \in J} P_j \log \mathscr{P}(X = j \mid \theta)$$

Another useful metric which measures the efficiency of classification models is the Kullback-Leibler (KL) divergence. Here, it represents the amount of information we gain about the distribution of the root state by using our model output relative to our a priori assumptions. We defined our prior $P_j$ as a uniform distribution over all possible root states. Then, the KL divergence was calculated using the posterior distribution over root states $\mathscr{P}(X = j \mid \theta)$ output by our phylogeographic models.

$$KL = \sum_{j \in J} P_j \log P_j - \log \mathscr{P}(X = j \mid \theta)$$

For each combination of simulation parameters, we recorded the state at the root of the phylogeny and calculated the root state accuracy, cross entropy and KL divergence to measure the performance of the standard and uncertain phylogenetic discrete trait models.

We analyzed the impact of model parameterization, data set size and discrete state space size on model performance metrics using ANOVA.

### 2.5. Data availability

We provide the simulated data as analysis-ready BEAST XML files along with files containing the parameters associated with each data set.

### 2.6. Factors influencing model accuracy

We modeled root state classification accuracy for our 900 models using logistic regression by defining factors related to phylogeographic study design choices as predictors. For these analyses, we set the reference levels of each factor variables to be: i) 4 discrete states ii) the “drop” model design (where sequences with missing metadata are excluded) and iii) 150 molecular sequences, respectively.

## 3. Results

### 3.1. Phylogeographic models show strong performance on moderately sized data sets

In Fig. 2, we show the mean and 95% confidence intervals for each of the non-reference level factors included in our analysis. Using the standard interpretation of the odds ratio, we show that increased discrete state space sizes are associated with weaker model performance with respect to root state classification (Fig. 2), p-values $< 0.01$). We found that, for our analysis, increasing data set size does not significantly improve phylogeographic root classification performance (Dudas et al., 2017). Interestingly, we see increased performance at for models with 250 sequences, relative to other data set sizes, as shown by the positive odds ratios associated with these models (Fig. 2, p-values $< 0.05$). Overall, we find no significant effects of model implementation method on root state classification performance.

### 3.2. Phylogeographic information gain increases with state space and data set size

In the phylogeographic context, KL divergence is often used to quantify the amount of information gained from an analysis with respect to a prior distribution. Concretely, we are interested in quantifying the amount of information that the root state posterior contains relative to a uniform (uninformative) prior over all possible traits. We performed this calculation such that our KL divergence is expressed in units of bits; representing the total amount of information gained by an analyst from performing phylogeographic analysis to identify the root state. For many empirical analyses, since the true root state (and any root states of internal nodes) are unknown a priori, it is unclear how information gain is related to model accuracy and if increased information gain translates directly to improved classification performance. By including KL divergence as a predictor in our logistic regression analysis, we were able to infer the respective relationship between this metric and model accuracy. We found that KL divergence was not associated with root state classification performance (Fig. 2, p-value: 0.248). In Fig. 3 we present the mean Kullback-Leibler (KL) divergence (from a uniform prior) for 25 model replicates stratified by model implementation method as well as state space and data set size. Using ANOVA, we find that information gain and discrete state space size were significantly related to KL divergence and that KL divergence tended to increase along with discrete state space size (p

< 0.001, F-score: 255.84, Table 2). Further, we also found KL divergence to be significantly associated with data set size ($p < 0.001$, F-score: 255.84, Table 2). We visualize the results of this analysis and show interactions between various design factors in Fig. 4. We utilized Tukey's HSD post-hoc test to identify that this effect is primarily driven by the increase in information gain occurring when increasing data set sizes from 250 to 500 tips ($p < 0.001$). Echoing the results of our logistic regression analysis, we find no significant differences in information gain between model implementation methods (p-value: 0.866, F-score: 0.242, Table 2). However, for models with large discrete state spaces, we observed a leveling off in KL divergence with increasing data set size (Fig. 4) indicating a functional limit to the information a phylogeographic model can extract about the root state given sufficient data. We also found significant interaction effects between state space size and data set size (Table 2, $p = 9.98 \times 10^{-1}$).

### 3.3. Phylogeographic cross entropy increases with discrete state space size

By casting the phylogeographic root state inference problem in a classification framework, we gain access several established metrics for use in quantifying classification model performance. We select cross-entropy due to its usage in a wide variety of substantive areas. In Fig. 5, we show the cross entropy mean and and 95% confidence interval stratified by model implementation method as well as state space and data set size. We observed that cross entropy tends to increase with the size of the state space; this is intuitive since the complexity of the classification task is related to the size of the state space. In Fig. 6, we show the interaction plots between design factors and cross entropy, noting that cross entropy tended to increase with data set size which we confirmed using ANOVA (Table 1). Again, we employ Tukey's HSD post-hoc testing to show that the sequences ($p < 0.001$). This is congruent with the results obtained from our logistic regression analysis which indicates that models fit to intermediate data set sizes tend to perform better than models with larger data set sizes. Since we generally expect classification model performance to generally increase with data set size, we offer an explanation of the apparent increase in model complexity arising from increasing data set sizes. We expect that classification performance diminished on larger data set sizes since phylogeographic classification models perform trait state estimation for all $n - 1$ internal nodes before making a final classification for the root trait state; if any errors are made at intermediate nodes, these errors are propagated back toward the root.

## 4. Discussion

### 4.1. Performance of phylogeographic models for root state classification

Pathogen molecular sequence data are being created at an unprecedented rate. So too has interest increased in methods and tools which leverage this new data stream for public health application. Examples in the literature include evaluating the impact of hypothetical interventions on epidemic spread (Dellicour et al., 2018) as well as identifying specific groups or locations responsible for driving epidemic spread (Lemey et al., 2009; Dudas et al., 2017; Grubaugh et al., 2017; Swetnam et al., 2018; Magee et al., 2017; Lemey et al., 2014). With increasing metadata availability, the resolution with which phylogeographic analyses are performed is increasing (Dudas et al., 2017; Grubaugh et al., 2017; Dellicour

et al., 2018). Concretely, this translates into specification of models with large number of discrete states and sequences. Reconstructing epidemiological patterns of infectious disease spread using pathogen genomes is often achieved by modeling the epidemiological trait of interest as a continuous time Markov chain which evolves across a phylogeny. The ability of these models to accurately reconstruct these traits of interest is paramount to their use in applied public health settings for modeling infectious disease outbreaks.

In this paper, we took a simulation-based approach and quantified the role of discrete state space and sequence data set sizes on the root state classification performance of modern phylogeographic models. We focused specifically on root state classification since, in infectious disease epidemiology, this task is analogous to identifying the discrete trait (i.e. geographic location, host species, etc.) associated with the origin of an outbreak. We simulated 225 data sets which we then analyzed using standard and uncertain phylogeographic discrete trait diffusion models. For the uncertain trait models, we performed analyses using three distinct prior specifications: i) a uniform prior across states ii) an informed prior which assigns most of the prior mass to the correct state and iii) a misspecified prior, which assigns most of its mass to an incorrect state. We compared characteristics of each model's MCC phylogeny to characterize model performance on the root state classification task. We found no significant differences between model implementation methods and model performance, suggesting that while the phylogeographic UTM does not substantially increase or decrease model performance. Therefore, it remains an attractive alternative for researchers wanting to include sequences with missing metadata in their analyses. Interestingly, a misspecified prior for the tip trait states did not seem to substantially effect root state predictive accuracy. We expect this is similarly due to errors in the state at each node being propagated back through the phylogenetic tree during inference. We expect that while the tip prior misspecification may influence the classification error at proximal internal nodes, as the model is applied backward in time toward the root, the partial likelihood vector begins to resemble the stationary distribution of the associated Markov model. This is especially likely for fast evolving traits, since the total evolutionary time for the model is the sum of all branch lengths across the tree.

Though phylogeographic models are popular epidemiological tools in an era of pathogen genomes aplenty, relatively few studies have characterized the performance of these methods under various analysis conditions (Scotch et al., 2019; Magee and Scotch, 2018; De Maio et al., 2015; Lemey et al., 2014). Indeed, much of this previous work is concerned with empirical analyses of virus sequence data sets (Magee et al., 2017; Scotch et al., 2019; Magee and Scotch, 2018) and often compares model root state posterior probabilities as a proxy for performance. The informativeness of the analysis is then typically assessed by calculating the Kullback-Leibler (KL) divergence between the root state prior and posterior distributions (Magee et al., 2017; Magee and Scotch, 2018; Lemey et al., 2014). Work by de Maio et al. (De Maio et al., 2015) established performance characteristics for several phylogeographic models using 200 tips and either two to eight discrete states, while focusing on the role of migration rates and sampling bias on inference quality. In contrast, we focus on the combined roles of data set and discrete state space sizes and how they impact discrete trait diffusion model inference.

We found that KL divergence is significantly positively associated with both discrete state space and data set size. This suggests caution when relying on this metric as it may erroneously suggest more granular discretization schemes or reward more data intensive models though this may not translate to increased performance on the root classification task. For example, though Scotch et al. (Scotch et al., 2019) found that the phylogeogephic UTM improved performance relative to other popular heuristics, these conclusions were based on a empirical comparison between model posteriors. Additionally, we show that KL divergence was not predictive of root state classification performance suggesting more informative models may still ultimately produce incorrect results. So, while empirical studies are informative for assessing congruence between root state inferences drawn by different phylogeographic methods, they are not informative with respect to the absolute accuracy (i.e. classification) of these methods for root state inference. We also find that root classification performance is the best at intermediate data set sizes. We believe that our models show poorer performance on larger data sets since as data set size increases, the number of internal nodes for which trait reconstruction must occur also increases. We expect that any errors in internal node classification (that is, internal node distributions which assign the most mass to an incorrect trait state) are propagated back toward the root. However, this could also be influenced by uncertainty in the phylogenetic tree topology. Changes in fast evolving traits, such as host age or geography during disease outbreaks, will be sensitive to uncertainty in branch lengths since as time increases, the partial likelihood vectors at internal nodes begin to more closely resemble the stationary distribution of their evolutionary Markov models. Since tree space is known to grow factorially (Felsenstein and Felenstein, 2004) with respect to tip number, it is likely that a combination of posterior tree uncertainty, mediated through the effect of increasing tip numbers, also impacted our results. Following this line of reasoning, we expect that increases in molecular sequence length will improve model performance since increasing the data available to models (via including more sites independently evolving across a tree) will reduce tree topological uncertainty.

### 4.2. Limitations and future work

Phylogeographic discrete trait diffusion models have emerged as the primary statistical tool for analyzing pathogen genomes annotated with discrete trait metadata. Given the increasing interest in the application of genome sequencing for public health outbreak response, it is prudent to establish the performance of phylogeographic models on different size data sets. This is of direct interest to public health practitioners who may be tasked with designing molecular epidemiological studies within budgetary, computational or data constraints. Overall, this study aimed to evaluate the performance of popular phylogeographic models under various analysis conditions, focusing on the roles of discrete state space and data set size on phylogeographic model performance. While we find that model performance is significantly increased at intermediate data set sizes, our results paint suggest caution when relying solely on KL divergence and other metrics calculated from purely empirical studies. However, our study is not without limitations. We limited our simulation study to discrete traits simulated from symmetric Markov rate matrices. This represents the simplest of the phylogeograhic models; we focused on this case to estimate a baseline for model performance. In reality, there are several ways in which trait states models are specified and inferred. Of particular note is the use of Bayesian Stochastic Search Variable Selection

(BSSVS) which augments the model state space such that each instantaneous rate parameter r,, is multiplied by a binomial random variable whose value represents the inclusion (or conversely, exclusion) of a given rate parameter in the matrix. The BSSVS parameterization effectively reduces the number of estimated transition rates which may lead to increased model performance on root state classification. Another popular approach for parameterizing discrete trait diffusion models is to model each transition rate as linear combination of covariates of interest. This reduces the problem of estimating transition rates to estimating the coefficients of the resulting generalized linear model (GLM). Clearly, our results do not extend to these parameterization methods. Finally, we quantified model performance with respect to root state classification only. It may be the case that the UTM increases classification performance on intermediate nodes in the phylogeny and that phylogeographic methods in general perform better on inferring the discrete states for proximal ancestral nodes. Quantifying the *treewide* classification performance of phylogeographic models under various conditions remains an open area of research.

## Acknowledgements

This research was supported by a grant from the National Library of Medicine of the National Institutes of Health under award number R01LM012080 to MS. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would also like to thank ASU Biomedical Informatics for providing computational resources to perform phylogenetic experiments. We would also like to thank the NVIDIA GPU Grant program for providing GPU hardware which was used to accelerate phylogenetic experiments.

## References

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. , 2018. GenBank. Nucleic Acids Res. 46 (D1), D41–D47. [PubMed: 29140468]

Brown JW, Walker JF, Smith SA, 2017. Phyx: phylogenetic tools for unix. Bioinformatics 33 (12), 1886–1888. [PubMed: 28174903]

Connolly MA, 2005. Communicable Disease Control in Emergencies: A Field Manual. World health organization.

De Maio N, Wu CH, O'Reilly KM, Wilson D, 2015. New routes to phylogeography: a Bayesian structured coalescent approximation. PLoS Genet. 11 (8), e1005421 [PubMed: 26267488]

Dellicour S, Baele G, Dudas G, Faria NR, Pybus OG, Suchard MA, et al. , 2018. Phylodynamic assessment of intervention strategies for the west African Ebola virus outbreak. Nat. Commun 9 (1), 2222. [PubMed: 29884821]

Dellicour S, Lequime S, Vrancken B, Gill MS, Bastide P, Gangavarapu K, et al. , 2019. Phylogeographic and phylodynamic approaches to epidemiological hypothesis testing. bioRxiv 788059.

Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. , 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544 (7650), 309. [PubMed: 28405027]

Felsenstein J, 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol 17 (6), 368–376. [PubMed: 7288891]

Felsenstein J, Felenstein J, 2004. Inferring Phylogenies. vol. 2 Sinauer associates, Sunderland, MA.

Grubaugh ND, Ladner JT, Kraemer MU, Dudas G, Tan AL, Gangavarapu K, et al. , 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature 546 (7658), 401. [PubMed: 28538723]

Hicks JT, Lee DH, Duvuuri VR, Kim Torchetti M, Swayne DE, Bahl J, 2020. Agricultural and geographic factors shaped the north American 2015 highly pathogenic avian influenza H5N2 outbreak. PLoS Pathog. 16 (1), e1007857. [PubMed: 31961906]

Lemey P, Rambaut A, Drummond AJ, Suchard MA, 2009. Bayesian phylogeography finds its roots. PLoS Comput. Biol 5 (9), e1000520. [PubMed: 19779555]

Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. , 2014. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 10 (2), e1003932. [PubMed: 24586153]

Magee D, Scotch M, 2018. The effects of random taxa sampling schemes in Bayesian virus phylogeography. Infect. Genet. Evol 64, 225–230. [PubMed: 29991455]

Magee D, Suchard MA, Scotch M, 2017. Bayesian phylogeography of influenza a/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction. PLoS Comput. Biol 13 (2), e1005389. [PubMed: 28170397]

Magge A, Weissenbacher D, Sarker A, Scotch M, Gonzalez-Hernandez G, 2018. Deep neural networks and distant supervision for geographic location mention extraction. Bioinformatics 34 (13), i565–i573. [PubMed: 29950020]

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA, 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol 67 (5), 901. [PubMed: 29718447]

Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, et al. , 2020. Database resources of the national center for biotechnology information. Nucleic Acids Res. 48 (Dl), D9. [PubMed: 31602479]

Scotch M, Sarkar IN, Mei C, Leaman R, Cheung KH, Ortiz P, et al. , 2011. Enhancing phylogeography by improving geographical information from GenBank. J. Biomed. Inform 44, S44–S47. [PubMed: 21723960]

Scotch M, Tahsin T, Weissenbacher D, O'Connor K, Magge A, Vaiente M, et al. , 2019. Incorporating sampling uncertainty in the geospatial assignment of taxa for virus phylogeography. Virus Evol. 5 (1), Vey043 Available from, 10.1093/ve/vey043. [PubMed: 30838129]

Stadler T, 2011. Simulating trees with a fixed number of extant species. Syst. Biol 60 (5), 676–684. [PubMed: 21482552]

Stadler T, Kiihnert D, Bonhoeffer S, Drummond AJ, 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc. Natl. Acad. Sci 110 (1), 228–233. [PubMed: 23248286]

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A, 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4 (1), vey016. [PubMed: 29942656]

Swetnam D, Widen SG, Wood TG, Reyna M, Wilkerson L, Debboun M, et al. , 2018. Terrestrial bird migration and West Nile virus circulation, United States. Emerg. Infect. Dis 24 (12), 2184. [PubMed: 30457531]

Tahsin T, Beard R, Rivera R, Lauder R, Wallstrom G, Scotch M, et al., 2014. Natural language processing methods for enhancing geographic metadata for phylogeography of zoonotic viruses. In: AMIA Summits on Translational Science Proceedings, vol. 2014. pp. 102.

Tahsin T, Weissenbacher D, O'Connor K, Magge A, Scotch M, Gonzalez-Hernandez G, 2017. GeoBoost: accelerating research involving the geospatial metadata of virus GenBank records. Bioinformatics 34 (9), 1606–1608.

Vatant B, Wick M. Geonames Ontology. Dostupné online: http://www.geonames.org/ontology/ontology vol. 3. 2012;1.

Yang J, Xie D, Nie Z, Xu B, Drummond AJ, 2019. Inferring host roles in bayesian phylodynamics of global avian influenza a virus H9N2. Virology 538, 86–96. [PubMed: 31586866]

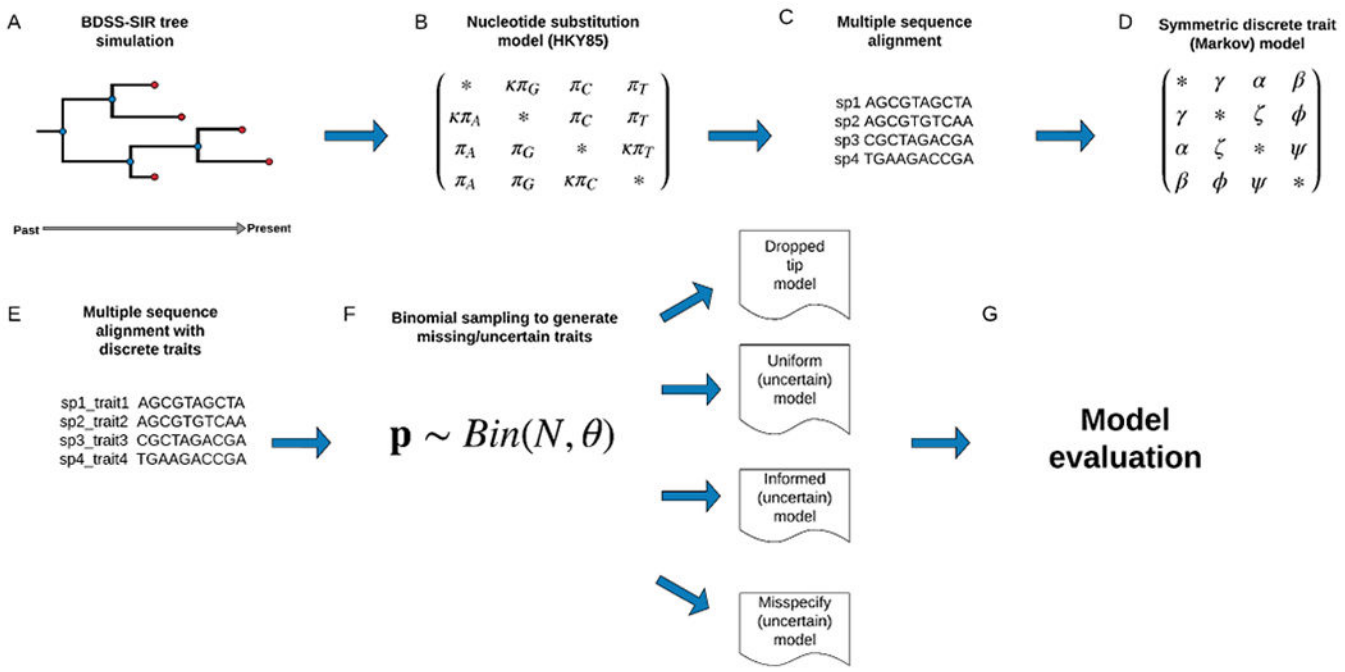


**Fig. 1.**

Visual summary of data simulation procedure. We simulated phylogenetic trees under the serially sampled birth-death SIR model using an R0 of 1.4 and an infectious period of 7 days. We simulated molecular sequence evolution on each tree topology using an HKY85 model and a strict molecular clock with a rate of $1 \times 10^{-3}$ substitutions per site, per year. We also simulated discrete traits on each tree topology, using symmetric Markov rate matrices with rate parameters drawn from a gamma distribution and a strict molecular clock with a rate of 0.1 substitutions per site, per year. This results in a set of molecular sequences annotated with discrete traits and sampling time information. We simulated missing traits using a binomial sampling process for each tip, indicating the presence, or conversely, the absence of discrete trait metadata. Finally, each data set was analyzed using one of four phylogeographic model parameterizations.

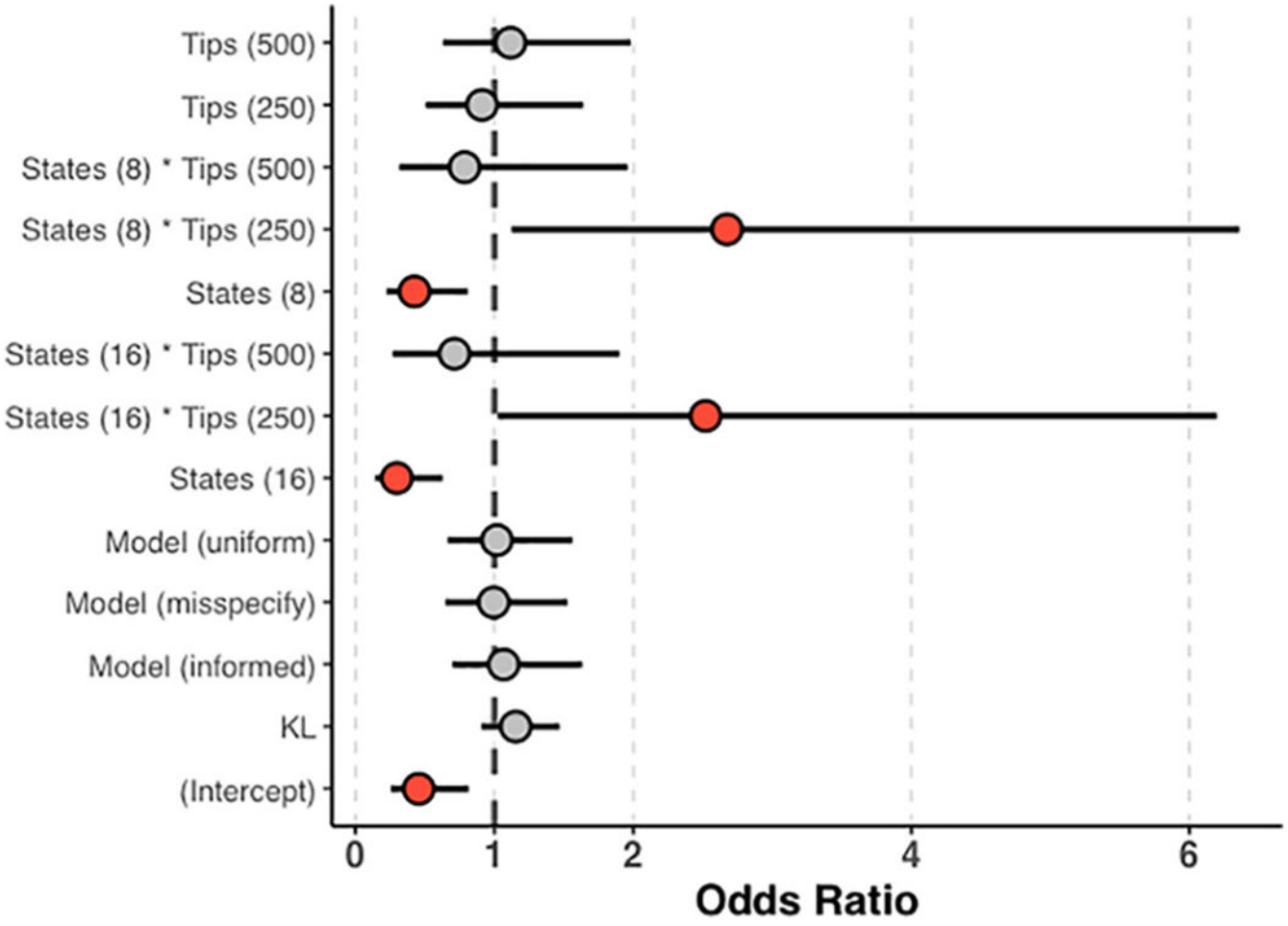


**Fig. 2.**

Odds ratios show the effect of design factors on model accuracy. We used logistic regression analysis to estimate the effects of design choices on phylogeographic model accuracy. For the purposes of analysis, we defined our reference factor levels to be 4 state, 150 sequence and drop model design, respectively. We show the factor found to be significant as red points, where grey points represent insignificant factors. Our analysis shows that relative to this reference level that increasing discrete state space size reduces the root state classification accuracy of phylogeographic models. We find that, independently, data set size and implementation method have no significant effects on model accuracy. However, our analysis shows increased root state classification performance for models with 250 sequences, suggesting that phylogeographic models may perform most favorably at intermediate data set sizes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

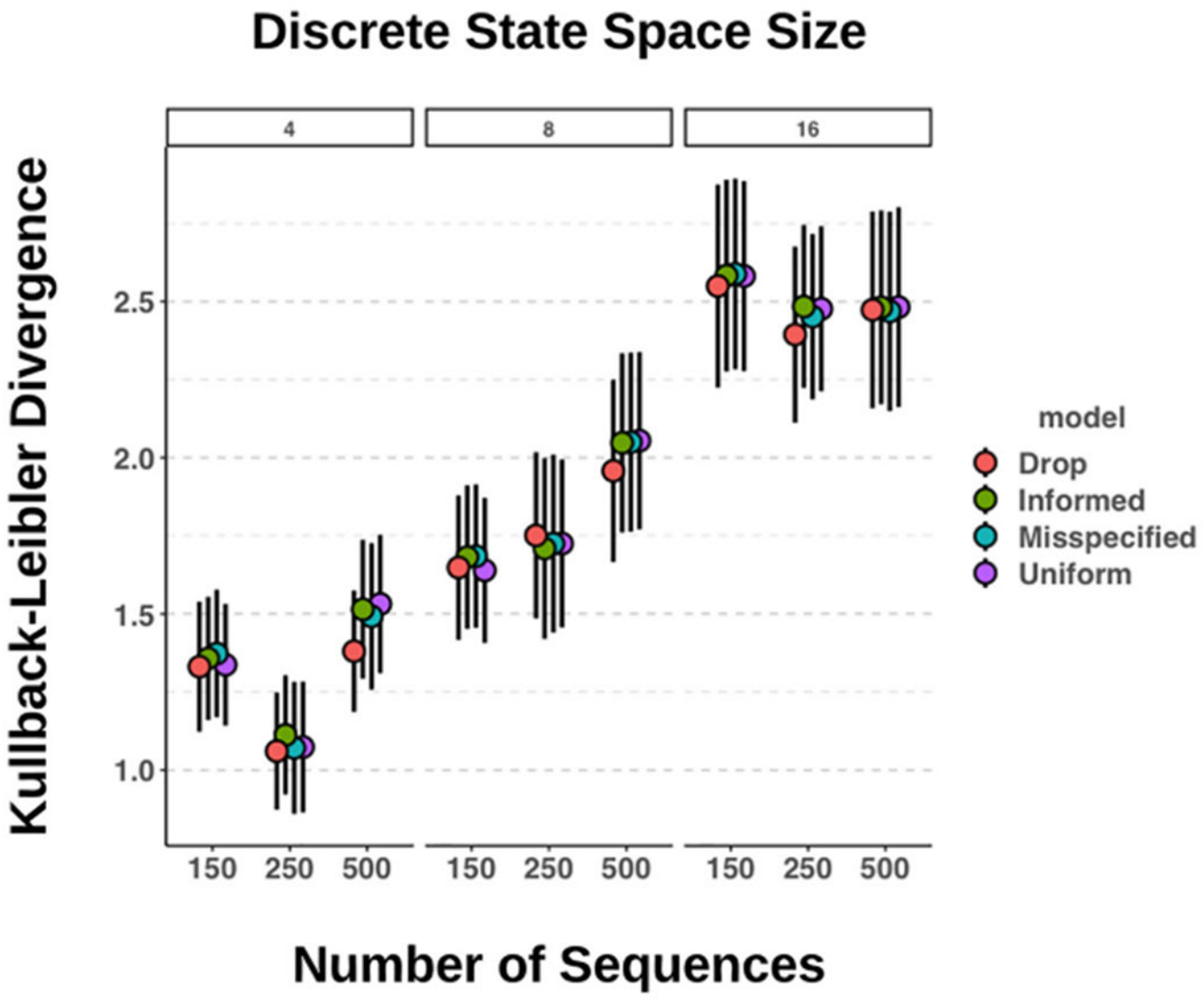


**Fig. 3.**

Comparison of Kullback Leibler (KL) divergence stratified by model design factor. Here, we show the mean and 95% confidence intervals for KL divergence arranged by increasing data set and discrete state space size. We observed an upward trend in information gain associated with both increasing discrete state space and data set sizes. We confirmed the presence of this trend using ANOVA (Table 2). As suspected, ANOVA suggests no statistically significant differences in posterior information gains between various model implementation heuristics (Table 2). We observed a tendency for information gain to increase when increasing data set size from 250 to 500 sequences.

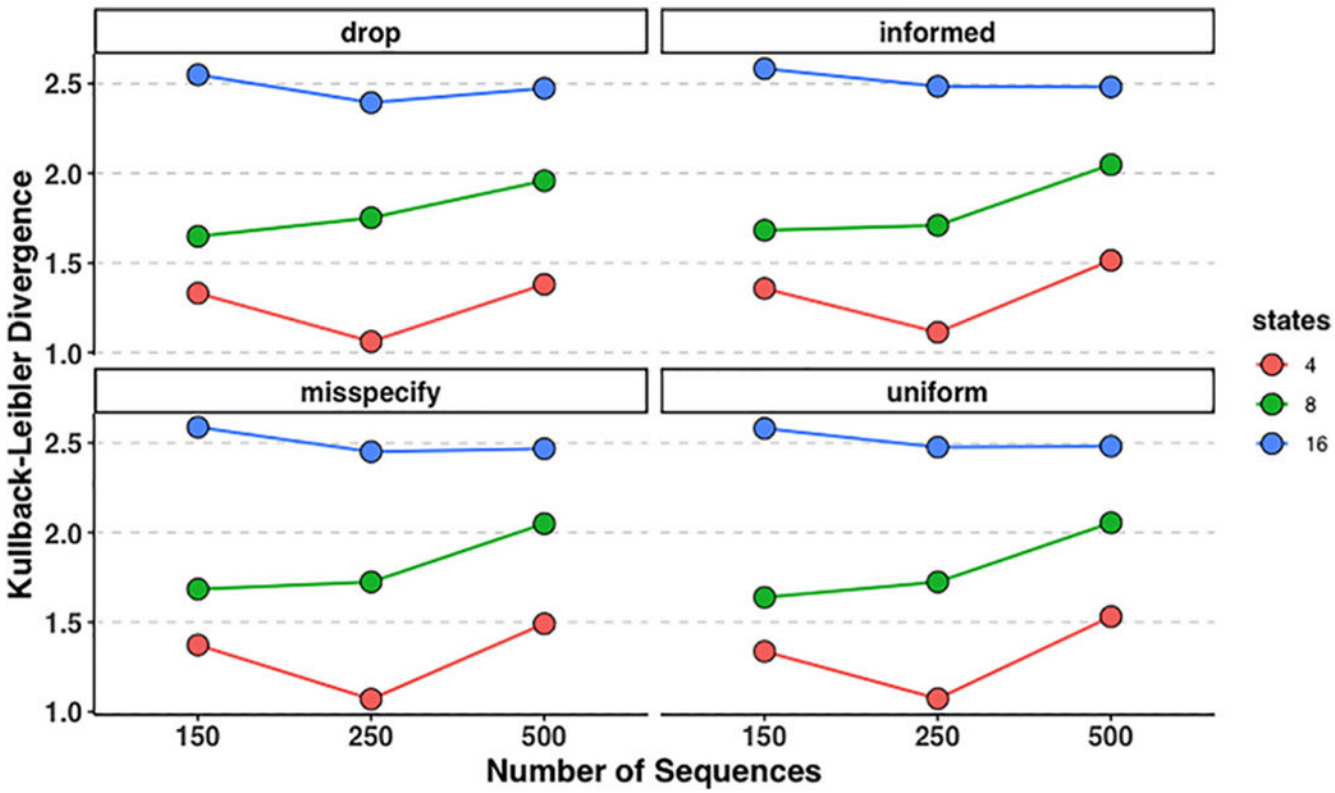


**Fig. 4.**

Interaction effects between model design factors and information gain.

We show that estimated mean KL divergence tended to increase when increasing the data set size from 250 to 500 sequences and that this effect was generally consistent across model implementations. From this perspective, it is further illustrated that we find information gain tended to increase with discrete state space size.

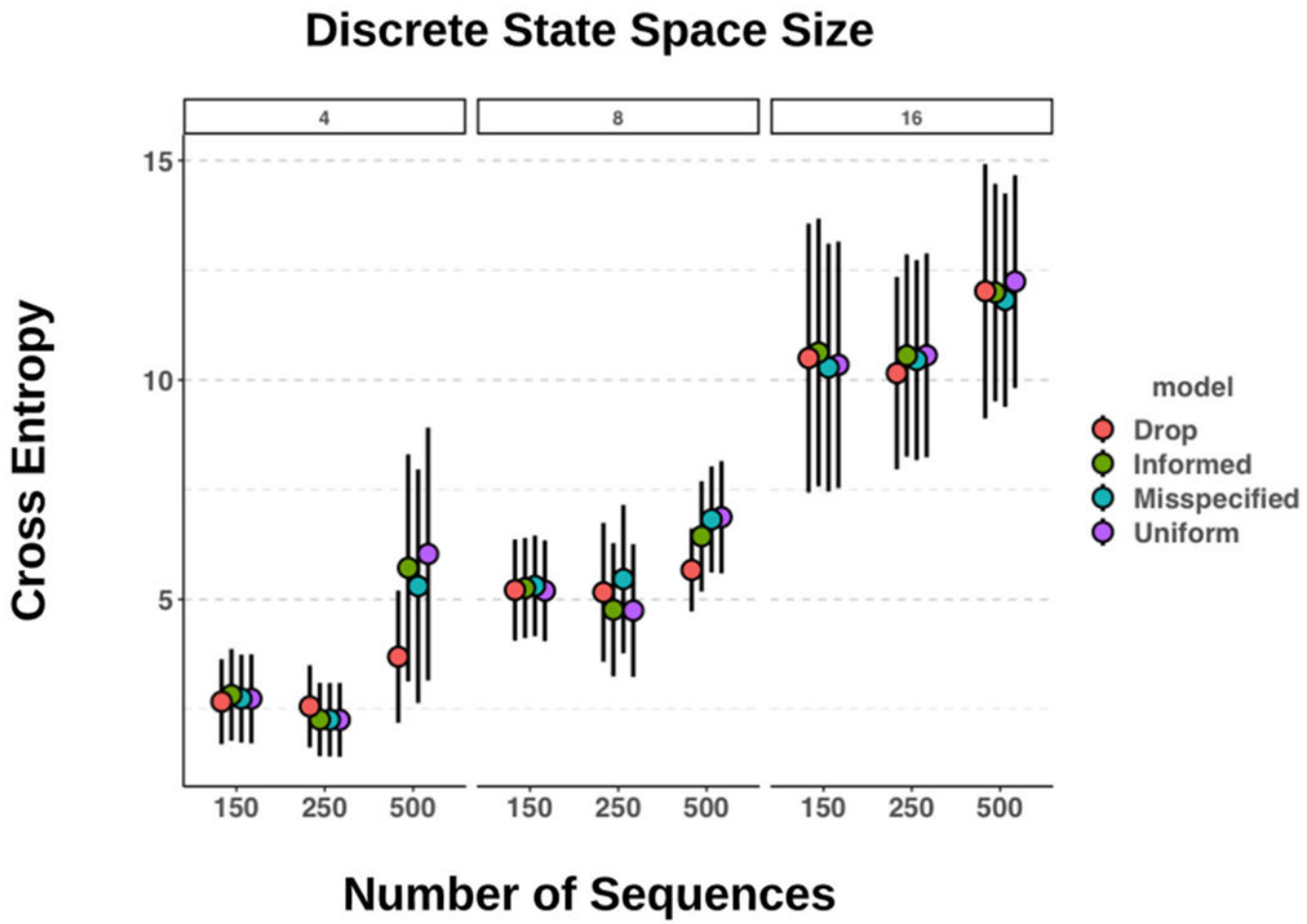


**Fig. 5.**

Comparison of Cross Entropy stratified by model design factor.

We present the mean and 95% CIs of the cross-entropy stratified by data set and discrete state space size. Similar to KL divergence (Fig. 3), we show that cross entropy tends to increase with discrete state space size. This is expected since the classification problem becomes more challenging as the total number of states increases. We also find that cross entropy tends to increase with data set size. This could be due to phylogegraphic the fact that phylogeographic root state classification first requires the model infer the discrete state probabilities at all $n - 1$ intermediate tree nodes. We expect that inference for an increasing number of internal tree nodes similarly increases the difficulty of the phylogeographic root state classification task.

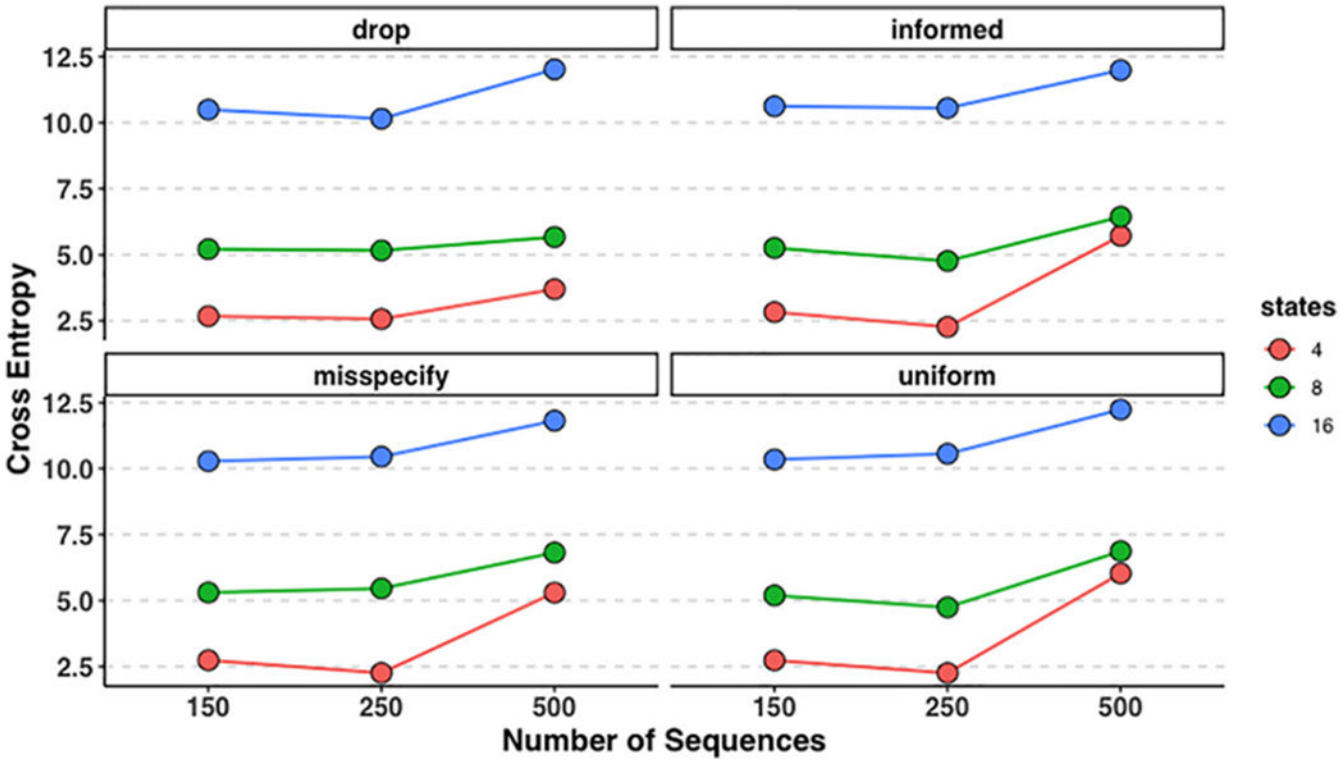


**Fig. 6.**

Interaction effects between cross entropy and model design factors.

By visualizing the interaction between each model design factor, we can observe that cross entropy remains relatively consistent between models with 150 and 250 sequences. However, it sharply increases significantly when models increase from 250 to 500 sequences (Tukey's HSD post-hoc analysis, $p = 4.2 \times 10^{-6}$) similar to the trends observed with KL divergence.

**Table 1**

Analysis of variance: cross entropy.

| Factor | Deg. Freedom | Sum Sq | Mean Sq | F-value | p-value |
|---|---|---|---|---|---|
| Model | 3 | 19 | 6 | 0.27 | 0.846 |
| Tips | 2 | 667 | 333 | 14.498 | $\mathbf{6.36} \times 10^{-7}$ |
| States | 2 | 8900 | 4450 | 193.481 | $< \mathbf{2} \times 10^{-16}$ |
| Tips * States | 4 | 70 | 18 | 0.762 | 0.550 |
| Residual | 888 | 20,147 | 23 | – | – |

Bolded p-values represent statistical significance < 0.05.

**Table 2**

Analysis of variance: Kullback-Leibler divergence.

| Factor | Deg. Freedom | Sum Sq | Mean Sq | F-value | p-value |
|---|---|---|---|---|---|
| Model | 3 | 0.3 | 6 | 0.27 | 0.867 |
| Tips | 2 | 8.2 | 4.09 | 14.498 | $\mathbf{6.39} \times 10^{-5}$ |
| States | 2 | 214.3 | 107.16 | 255.884 | $< \mathbf{2} \times 10^{-16}$ |
| Tips * States | 4 | 7.8 | 1.95 | 4.662 | $\mathbf{9.98} \times 10^{-4}$ |
| Residual | 888 | 366.9 | 0.42 | – | – |

Bolded p-values represent statistical significance < 0.05.