University of Georgia
Institute of Bioinformatics
1785

Evaluation of a Subsampling Strategy using Linguaphylo and Nextflow

Guppy Stott[1], Leke Lyu[1], Gabriella Veytsel[1], Justin Bahl[1,2]

[1] Institute of Bioinformatics, Center for Ecology and Infectious Diseases, University of Georgia, Athens, GA
[2] Department of Infectious Diseases, Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA

CEIRR
Centers of Excellence for Influenza Research and Response

Center for Applied Pathogen Epidemiology and Outbreak Control

Introduction

- Bayesian phylodynamics is computationally intensive, and frequently requires subsampling of a dataset for problems to remain tractable.
- While some work has been done to investigate alternative models for discrete phylogenetics, most analyses still make use of mugration models or their derivatives, which are prone to biased estimates when sampling is not uniform.
- While recent work by Pengyu Liu, et al. and Rhys Inward, et al. have begun to investigate the effects of subsampling schemes on phylodynamic analyses, these are restricted to simple subsampling strategies and/or phylodynamic models.
- Our work advances the simulation study by Pengyu, et al. to investigate the effects of subsampling on a five-location discrete phylogeographic model given multiple sequencing scenarios.
- **We built a Nextflow pipeline to make our investigation extensible and able to benchmark future subsampling strategies under multiple real-world sequence availability scenarios.**
- **Using our Nextflow pipeline, we demonstrate that the Phylogenetic Analysis Subsampling Tool (PAST), which leverages LCUBE subsampling, provides more accurate estimation of discrete transition rates and is more robust to sampling bias introduced by sequence availability.**

Methods

- Our method, PAST, uses the LCUBE method (*figure 1*) developed by Grafstrom, et al. to generate a balanced and well-spread subsample (*figure 2*).
- We hope that this will provide more accurate estimates since mugration models (like discrete Bayesian Stochastic Search Variable Selection) are easily biased by the proportion of sequences available to the model.
- To test this hypothesis, we built a Nextflow pipeline that simulates the discrete phylogeographic history of a pathogen, builds an alignment, and then simulates 3 different sequence availability scenarios before performing subsampling.
    - At least 5,000 taxa in each true tree
    - Downsampled to 2,500 sequences for each sequence availability scenario
    - Subsampled to 500 sequences for each subsampling method
- Beast runs were built using Linguaphylo to generate the XML files to enable more complex models in future modules and each had a chain length of 100M with 20% burn-in.
- Two tests were done, one with 5 different random histories and another with 10 resamples per method per sequence availability scenario under the same true history for a total of 135 Beast runs (the full pipeline is described in *figure 3*).

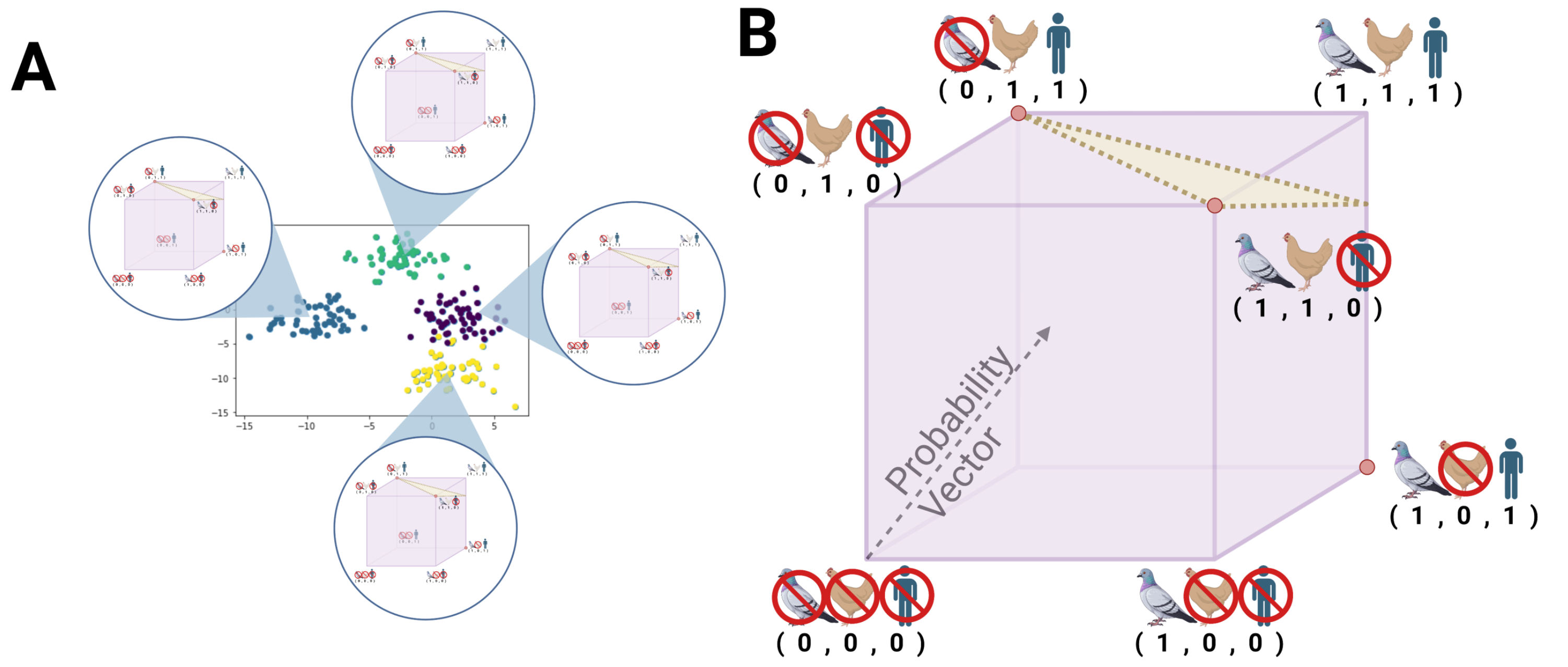


Figure 3: Diagram of two phases of LCUBE Subsampling: A. Clustering (to build a well-spread sample) and B. Cube sample selection (to maintain a random sample that is balanced).

Methods

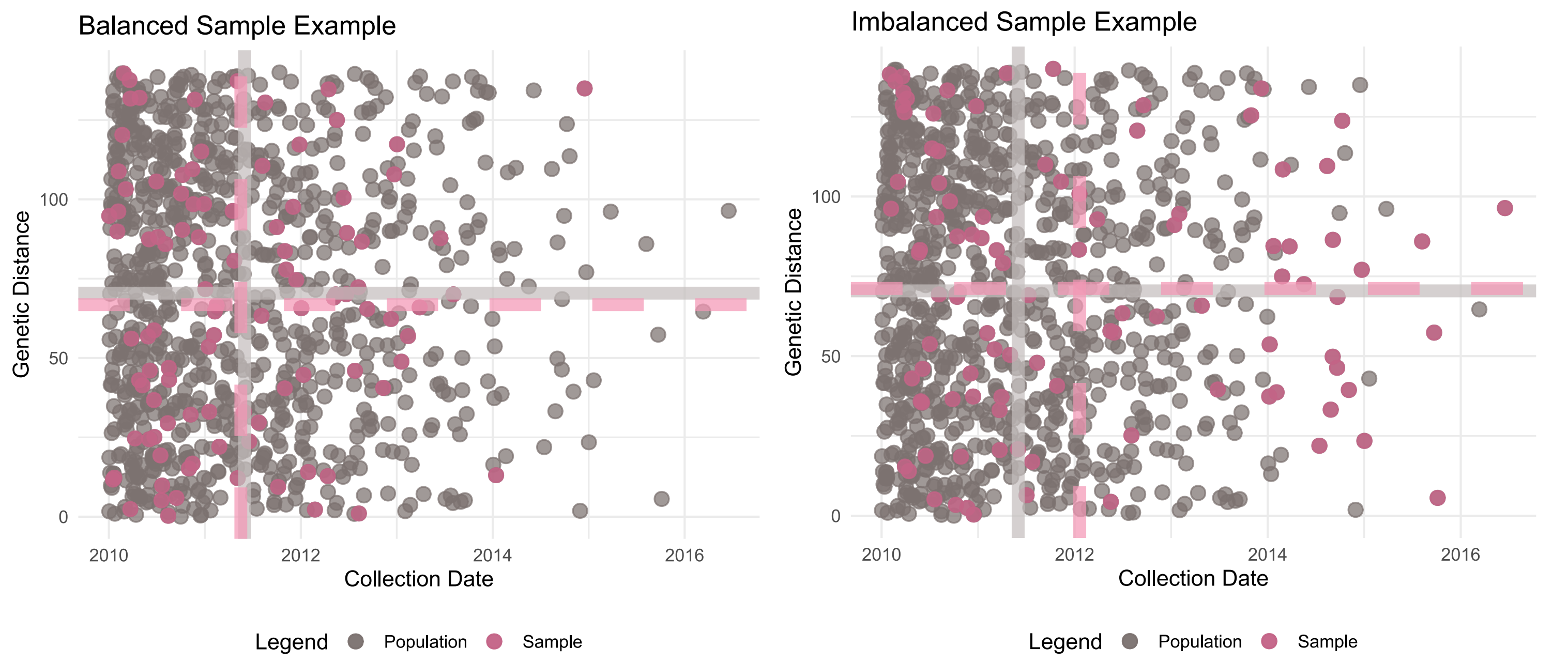


Figure 2: Demonstration of balanced vs. imbalanced sampling across 2 variables

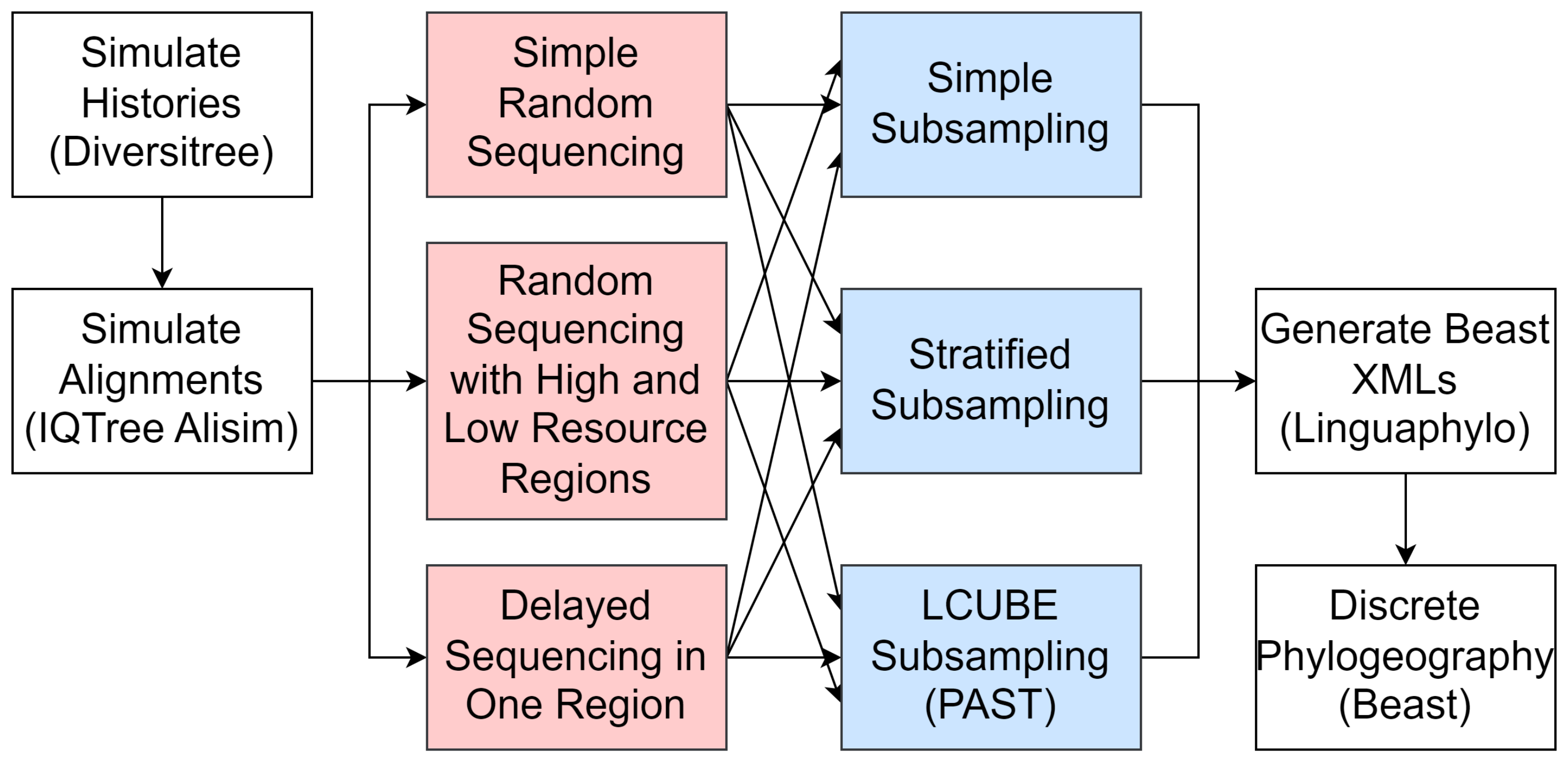
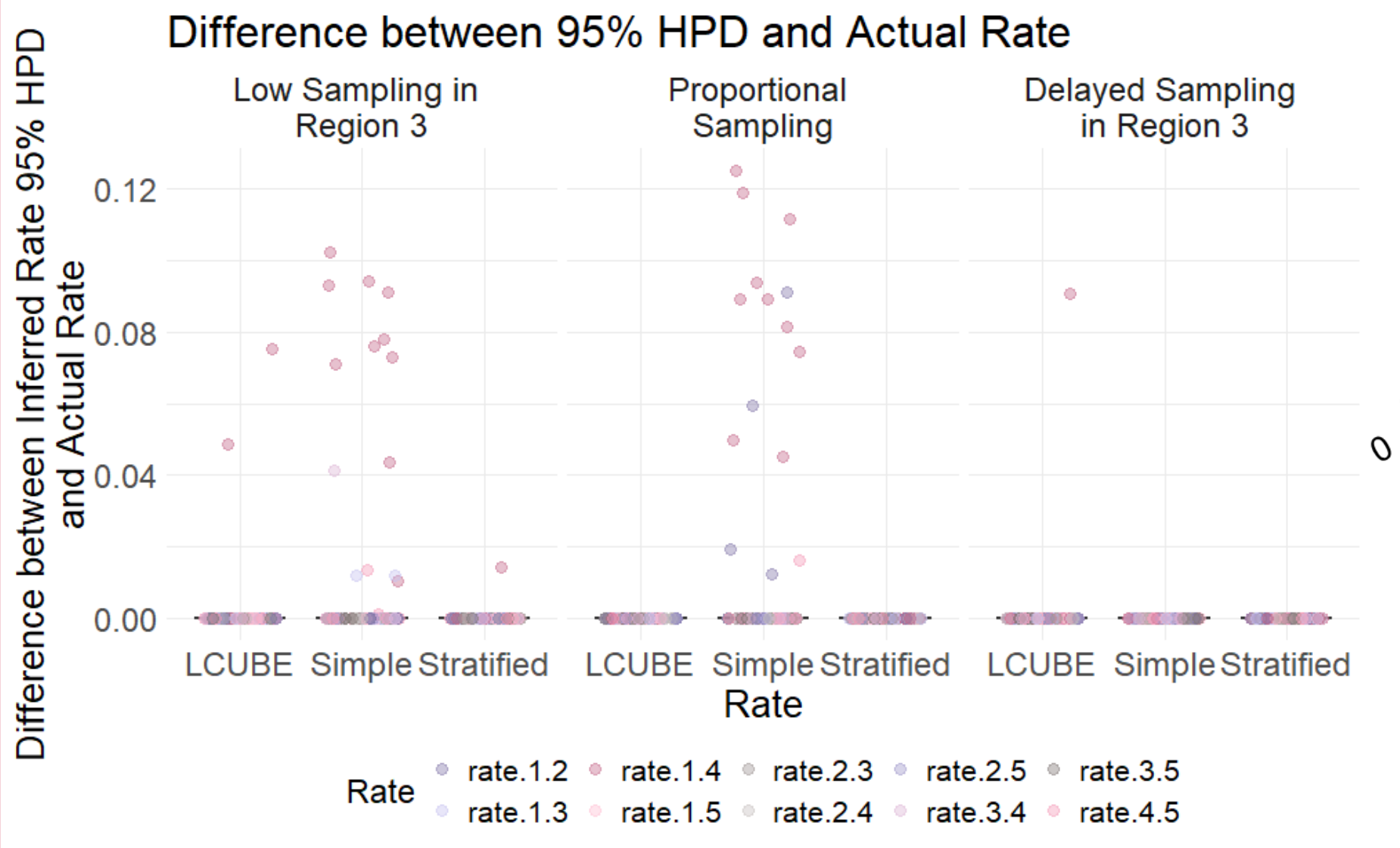

Simulate Histories (Diversitree)
Simulate Alignments (IQTree Alisim)
Simple Random Sequencing
Random Sequencing with High and Low Resource Regions
Delayed Sequencing in One Region
Simple Subsampling
Stratified Subsampling
LCUBE Subsampling (PAST)
Generate Beast XMLs (Linguaphylo)
Discrete Phylogeography (Beast)

Figure 3: Nextflow pipeline design to test the LCUBE subsampling strategy

Results



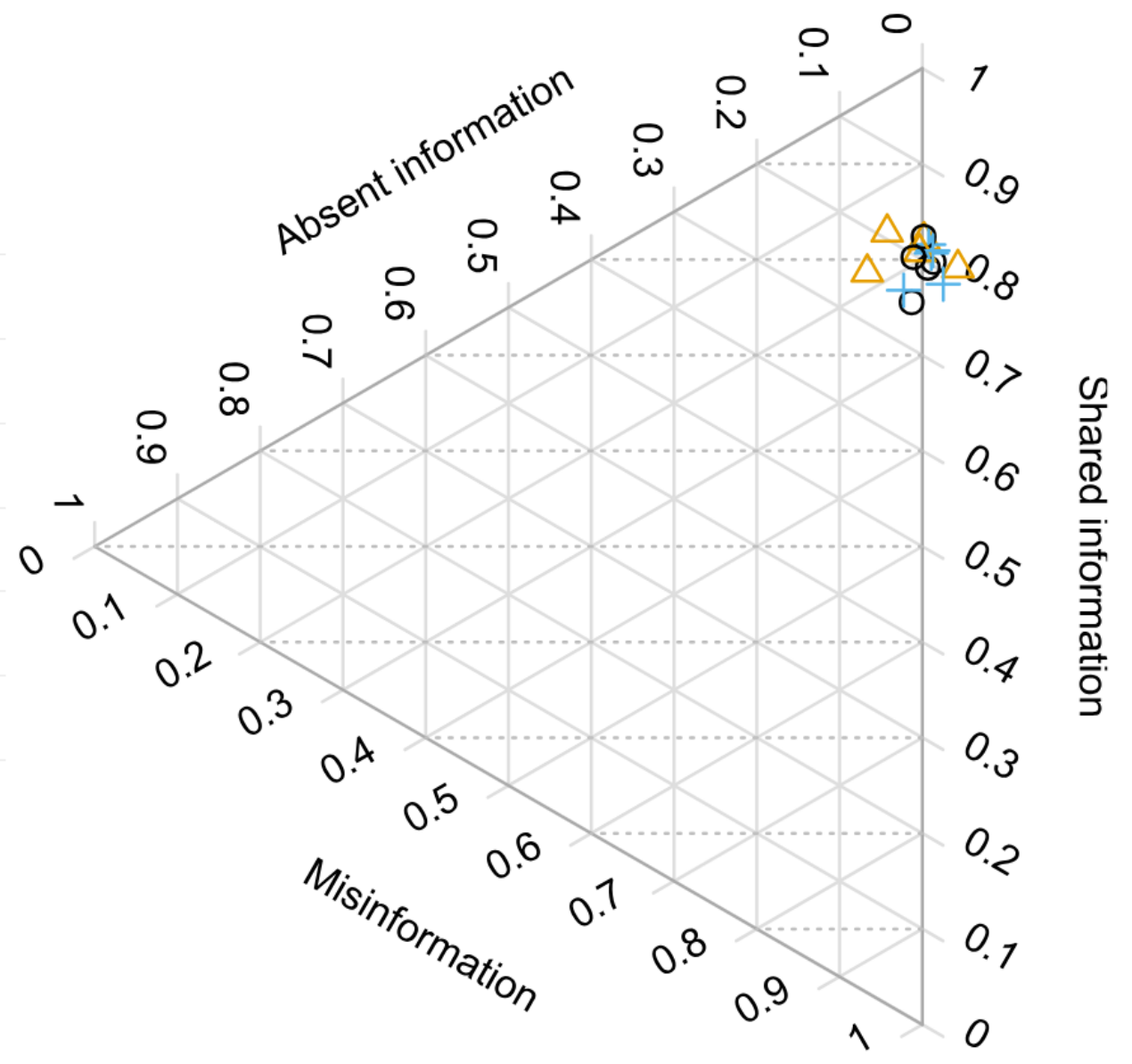


Figure 4: Difference between true rate and 95% HPD by sequencing and sampling method

Figure 5: Ternary plot of distance measures to true tree

This work was supported in part by the National Institute of Allergy and Infectious Diseases Centers of Excellence for Influenza Research and Response, Contract No. 75N93021C00018 and the National Science Foundation Predictive Intelligence for Pandemic Prevention Phase I: Development Grants (2200158). The funders had no role in study design, data collection, analysis, or decision to publish.

Results

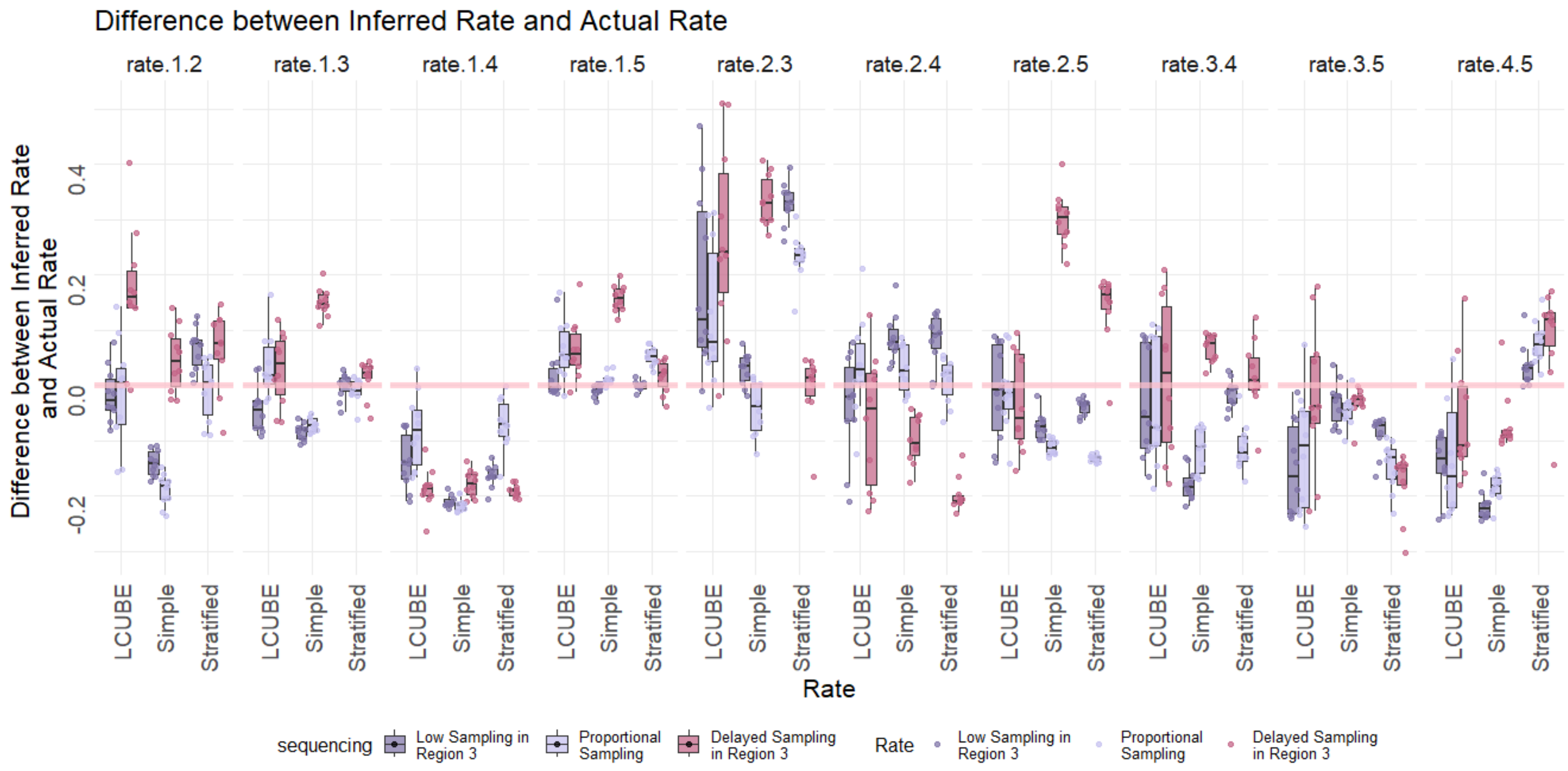


Figure 6: Difference between median discrete transition rate estimates and the true transition rate by subsampling strategy colored by sequence availability scenario (10 replicates per subsampling-sequence availability pair) with a dashed line indicating no difference

- As shown in *figure 4*, we see that all subsampling strategies are able to accurately estimate a 95% HPD which contains the true rate most of the time.
    - Simple random sampling underperforms compared to other strategies.
    - This is not due to differences in topological reconstruction (*figure 5*).
- While simple and stratified subsampling strategies tend to lead to more consistent parameter estimates, these estimates may be erroneously confident and biased by sequence availability in comparison to LCUBE, which is more robust to these differences (*figure 6;* e.g., rates 2.5 and 2.3).

Discussion

- LCUBE subsampling using our tool, PAST, mitigates some of the biases introduced when subsampling a dataset for discrete Bayesian phylodynamic inference under a mugration model with BSSVS.
- LCUBE subsampling more accurately reflects the variability in our estimates of transition rates under multiple resampling.
- Simple and stratified random sampling both are significantly impacted by biases introduced through non-uniform sequencing in a population.
- Future work will extend these results, investigating TMRCA estimates and the proportion of ancestral reconstruction accurately inferred.
- We will also investigate alternative models for discrete phylogeographic reconstruction designed to mitigate these issues (e.g. structured coalescent models) and investigate if continuous phylogeographic methods face similar challenges under subsampling regimes.

Contact:
Guppy Stott (they/them)

Email:	gs69042@uga.edu
Website:	glstott.github.io
Twitter/X:	@guppy_stott
Bluesky:	@guppy-stott.bsky.social
Github:	github.com/glstott