Infection, Genetics and Evolution 64 (2018) 225–230

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

The effects of random taxa sampling schemes in Bayesian virus phylogeography

Daniel Magee[a,b], Matthew Scotch[a,b,*]

[a] Department of Biomedical Informatics, Arizona State University, 13212 E. Shea Blvd., Scottsdale 85259, AZ, USA
[b] Biodesign Center for Environmental Health Engineering, Arizona State University, 1001 S. McAllister Ave, Tempe 85281, AZ, USA

ARTICLE INFO

Keywords:
Phylogeography
Selection Bias
Viruses

ABSTRACT

Public health researchers are often tasked with accurately and quickly identifying the location and time when an epidemic originated from a representative sample of nucleotide sequences. In this paper, we investigate multiple approaches to subsampling the sequence set when employing a Bayesian phylogeographic generalized linear model. Our results indicate that near-categorical posterior MCC estimates on the root can be obtained with replicate runs using 25–50% of the sequence data, and that including 90% of sequences does not necessarily entail more accurate inferences. We present the first analysis of predictor signal suppression and show how the ability to detect the influence of predictor variables is limited when sample size predictors are included in the models.

1. Introduction

The process of selecting the number of nucleotide samples to include in phylogeographic analyses is a non-trivial task. Federal surveillance initiatives rely on reference laboratories for supplying biological specimens such as virus sequences. While there are many advantages to these laboratories for public health, the production of large datasets from a limited amount of geographically sparse institutions might produce an inaccurate representation of the true distribution of the virus in the population. Thus, researchers must avoid over- or underrepresenting discrete locations in their dataset to limit reconstruction biases (Lemey et al., 2014). Several subsampling methods such as simple random sampling have been used, but there is currently no standard or consensus for performing subsampling in phylogeographic analysis. Lemey et al. (2014) subsampled the five regions with the largest number of samples relative to their population size, Magee et al. (2017) randomly selected 25% of the sequences from each discrete location without considering population, and Bedford et al. (2015) and Neher and Bedford (2015) place favor on earlier taxa and those that increase the amount of spatial distribution. To account for bias in the randomized subsampled dataset, some of these studies completed replicate randomizations to confirm or show divergence in their results (Lemey et al., 2014; Magee et al., 2017). Meanwhile, many phylogeographic studies do not perform any subsampling and simply use the full set of sequences to complete their analyses. While this approach eliminates any arbitrary subsampling, it does nothing to alleviate over- or underrepresented discrete states (Pybus et al., 2012), (Al-Qahtani et al., 2017), (Pollett et al., 2015).

Recent investigations into Ebola virus in Africa (Dudas et al., 2017), global seasonal influenza trends (Lemey et al., 2014), avian influenza in China (Lu et al., 2017), and HIV in Brazil (Graf et al., 2015) have each employed a Bayesian phylogeographic generalized linear model (GLM) to complete their analyses. This method enables the simultaneous reconstruction of the evolutionary history as well as determining the variables that played an important role in shaping that process. For this reason, the GLM appears to have biological advantages over existing phylogeographic and spatial epidemiology approaches. Despite the recent popularity of the GLM framework, several studies have detected strong sampling biases via the phylogeographic GLM approach (Lemey et al., 2014; Magee et al., 2015; Magee et al., 2017). It was recently demonstrated that a popular Bayesian stochastic search variable selection (BSSVS) procedure with a Poisson location prior that does not incorporate a Bayesian phylogeographic generalized linear model may be less susceptible to sampling bias than GLMs (Magee et al., 2017). No paper has investigated whether this is a systematic issue within the GLM framework or if there is a method to circumvent sampling bias.

In this paper, we aim to investigate the relationship between subsampling schemes and their ensuing phylogeographic reconstructions

Abbreviations: GLM, generalized linear model; BSSVS, Bayesian stochastic search variable selection; MCC, maximum clade credibility; RSPP, root state posterior probability; KL, Kullback-Leibler

[*] Corresponding author at: Department of Biomedical Informatics, Arizona State University, 13212 E. Shea Blvd., Scottsdale 85259, AZ, USA.
E-mail address: matthew.scotch@asu.edu (M. Scotch).

https://doi.org/10.1016/j.meegid.2018.07.003
Received 2 December 2017; Received in revised form 7 June 2018; Accepted 2 July 2018
Available online 04 July 2018

via a GLM approach. In our analyses, we compare posterior metrics and traits of maximum clade credibility (MCC) phylogenies as a function of the subsampling scheme to determine whether there is a scheme that minimizes bias while maximizing probability in the ancestral state reconstructions. We also examine the impact of the choice of scheme on the GLM predictors and the degree to which sampling bias suppresses the relationship between predictors and transmission of the virus. Our work may inform researchers whether the computational burden necessary to accurately analyze a large sequence set can be averted in favor of replicate analyses of smaller samples.

2. Materials and methods

2.1. Sequences

We used a set of 1163 hemagglutinin (HA) segments of influenza A/H3N2 from the 2014–15 flu season in the contiguous U.S. first described elsewhere (Magee et al., 2017).

2.2. Subsampling schemes

In this work, we analyzed two primary subsampling schemes. In the first approach, we performed random sampling and then grouped the taxa into its assigned Health and Human Services (HHS) geographic region (Fig. 1). We refer to this scheme as Pre subsampling since we randomized prior to taxa assignment. As an alternative, we first aggregated the sequences into HHS regions and then sampled a fixed proportion of sequences from within each region. We refer to this scheme as Post subsampling since we randomized after we assigned each taxon to its HHS region. For both schemes, we evaluated the magnitude of subsampling on the posterior results selecting fixed percentages of 10%, 25%, 50%, 75%, and 90% of our complete dataset. For example, in a 10% Pre subsampling scheme, we selected a random 116 (10% of 1163) sequences and then pooled the selected sequences into their respective HHS regions. In a 10% Post subsampling schema, we discretized the sequences into their respective regions (107 to Region 1, 55 to Region 2, 119 to Region 3, etc.) and then selected 10% of the sequences from each region. We evaluated ten independent sequence sets for each scheme-level combination. For reference, we also include one analysis on the entire set of 1163 sequences using all predictors and one that excludes the sample size predictors, which we denote Full (+SS) and Full(−SS), respectively. Both models were run in duplicate to ensure consistency.

2.3. Model parameters

For each of the models, we specified a HKY + G substitution model (Hasegawa et al., 1985) and an exponential growth coalescent tree prior (Griffiths and Tavare, 1994) following recent phylogeographic analyses of influenza A/H3N2 viruses (Lemey et al., 2014; Magee et al., 2017). We set the Markov chain Monte Carlo (MCMC) chain length to 100 M, sampling every 10,000 steps. We employed a Bayesian phylogeographic GLM (Lemey et al., 2014) on each model and incorporated the following predictors: distance (DS), temperature (TP), precipitation (PC), median age (MA), population density (PD), vaccination rate (VR). To investigate signal suppression noted in previous GLM studies (Lemey et al., 2014; Magee et al., 2017), we also analyze additional models that also include the number of sampled taxa per region (sample size, SS) as a predictor. This leaves us with two primary sampling schemes, each with five levels of subsampling, ten random sequence samples, each of which is run with and without the SS predictors. Thus, we have 200 independent models, 50 from the four schemes: Pre(+SS), Post(+SS), Pre(−SS), and Post(−SS). We used a point estimate representative of each region for each predictor, which follows the work established in Magee et al. (2017). We evaluated each model using BEAST v1.8.4 (Drummond et al., 2012) and created a maximum clade credibility (MCC) tree using TreeAnnotator v1.8.4 after discarding the first 10% of estimates as burnin.

2.4. Model comparison metrics

We summarize several metrics of the MCC phylogenies that are frequently reported in phylogeographic studies including the root location, root height, the root state posterior probability (RSPP), and Kullback-Leibler (KL) divergence. To investigate sampling bias, we report the Pearson's r and Spearman's ρ correlations between the posterior probability and the number of samples of each of the ten regions. We also report the posterior inclusion support and effect size of the predictors for each model, including the sample size predictors, which provides a third metric of sampling bias. Furthermore, the support metrics of the sample size predictors in comparison with the other predictors allow us to directly quantify the amount of signal

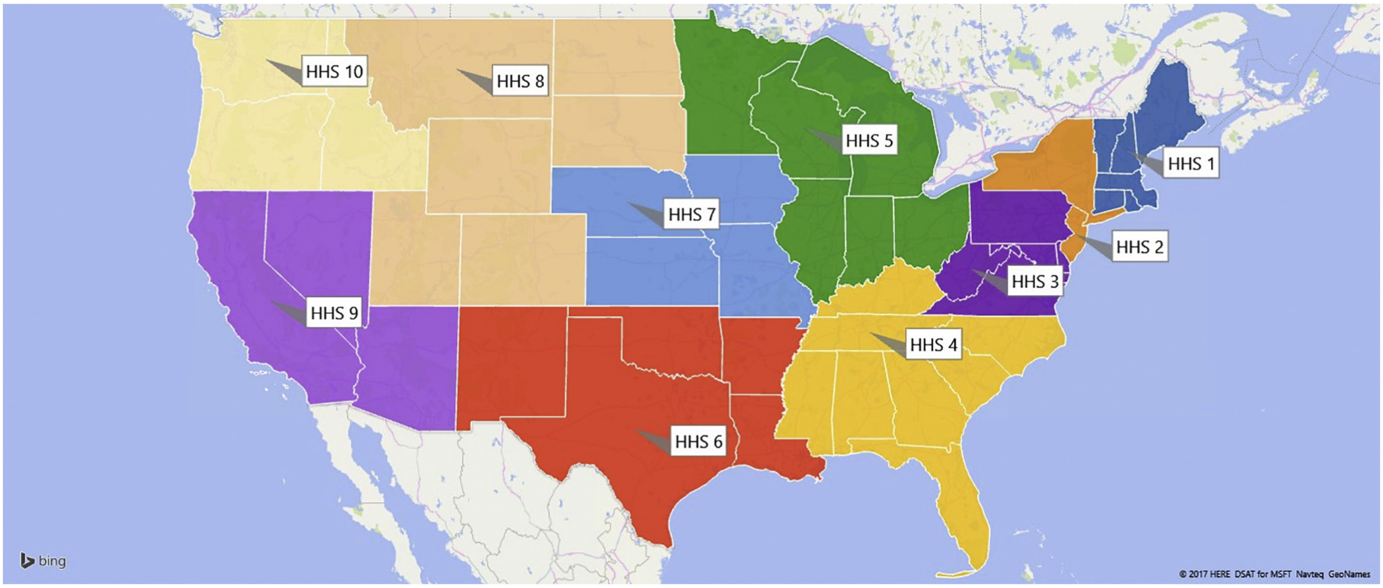

Fig. 1. United States Health and Human Services (HHS) regions. We used this level of discretization for our analysis in order to be consistent with the CDC's influenza surveillance.