Graph hierarchies for phylogeography

Gabriela B. Cybis\textsuperscript{1}, Janet S. Sinsheimer\textsuperscript{1,2,3}, Philippe Lemey\textsuperscript{4} and Marc A. Suchard\textsuperscript{1,2,3}

\textsuperscript{1}Department of Biomathematics, \textsuperscript{2}Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
\textsuperscript{3}Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA, USA
\textsuperscript{4}Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium

Bayesian phylogeographic methods simultaneously integrate geographical and evolutionary modelling, and have demonstrated value in assessing spatial spread patterns of measurably evolving organisms. We improve on existing phylogeographic methods by combining information from multiple phylogeographic datasets in a hierarchical setting. Consider $N$ exchangeable datasets or strata consisting of viral sequences and locations, each evolving along its own phylogenetic tree and according to a conditionally independent geographical process. At the hierarchical level, a random graph summarizes the overall dispersion process by informing which migration rates between sampling locations are likely to be relevant in the strata. This approach provides an efficient and improved framework for analysing inherently hierarchical datasets. We first examine the evolutionary history of multiple serotypes of dengue virus in the Americas to showcase our method. Additionally, we explore an application to intrahost HIV evolution across multiple patients.

1. Introduction

Viral pathogens represent serious burdens on public health, as the current HIV-1 pandemic exemplifies, with an estimated 33.3 million people infected worldwide [1]. To reduce the impact of such pathogens, public health policies that aim to reduce viral dispersion and global circulation are of paramount importance. Detailed knowledge of the processes that govern epidemic evolution and spread is crucial for the determination of these policies. Recent phylogeographic studies explore the commensurate time-scales of geographical dispersion and evolutionary processes to increase our understanding of the dispersal patterns of rapidly evolving pathogens [2–4].

Methods that integrate genetic and geographical analyses have been used to assess origins and dispersal patterns of many organisms of interest, such as modern humans [5] and dogs [6]. These phylogeographic studies largely benefit from the development of new methodologies [7]. An emerging methodological approach is the introduction of spatial diffusion processes in both discrete [8] and continuous space [9] to Bayesian phylogenetic analysis. This introduction allows for simultaneous reconstructions of geographical spread and evolutionary processes to increase our understanding of the dispersal patterns of rapidly evolving pathogens [2–4].

When we assume discrete geographical locations, a key feature for characterizing dispersion processes in these models becomes the migration graph. In this graph, the vertices represent the geographical states of the model, and an edge connects two vertices only if the instantaneous migration rate between these vertex locations is non-zero. To infer the non-zero rates, a variable selection procedure controls the total number of edges to avoid overfitting. Thus, the migration graph contains the information of which migration rates are relevant for describing the overall dispersion process.

One troubling aspect of phylogeographic inference centres on the observation that we generally have only one realization of the geographical evolutionary process that produced the spatial distribution of the group under study. Thus, even...
for moderately small numbers of geographical states, most transitions between locations might not even be sampled. Consequently, when there are a large number of locations, phylogeographic analyses frequently have to control for poor estimates for the migration rates.

Ideally, we would like to have many realizations under the same geographical process in order to improve inference. In some cases, parallel geographical realizations are available for this type of analysis. An example of such data structure that we explore in this paper concerns the four different serotypes of dengue virus in the Americas. Although the different serotypes do not have the same evolutionary history and probably arrived in the Americas at different times and in different locations, they share the same vector, host, mode of transmission, and most other aspects of viral biology and ecological niches [11]. Thus, we reasonably assume that the factors that govern their dispersion processes and, in turn, their migration graphs are similar. Consequently, information obtained from one serotype should improve inference on the phylogeography of the other serotypes of dengue, if we allow for occasional differences.

Previous studies have explored two different strategies when dealing with these parallel data structures. One alternative is to treat each phylogeographic dataset or stratum independently, as Allicock et al. [12] explore for the four serotypes of dengue. This approach allows for comparisons between the different serotypes and can identify discrepancies in the migration process. Nevertheless, there is no structure for sharing information between the four separate analyses.

Alternatively, Sanmartin et al. [13], in a similar data analysis, model all strata jointly by imposing the same migration graph and rates to all strata of the analysis. This method effectively combines information from all strata to improve rate estimates. But in doing so, it disregards any stratum-specific peculiarities, and does not provide an opportunity to distinguish the processes of different strata. One reasonably assumes that inherent variability or systematic differences of scientific interest may introduce small differences in the processes.

As a middle-ground to these two extreme alternatives, we propose to model these inherently parallel datasets through a Bayesian hierarchical model. The hierarchical structure of our model effectively represents the overall dispersion process and allows for sharing information between the strata. We allow each individual stratum small variation from the hierarchical level, and large deviation from the expected pattern naturally assesses discrepancies between the strata. To accomplish this task, we introduce a novel hierarchical migration graph that summarizes the geographical dispersion process over all strata and informs which are the predominant migration rates for the individual strata. Figure 1 presents a schematic representation of the structure of the model.

Hierarchical models have found success in Bayesian phylogenetics for modelling the molecular sequence substitution processes within multipartite data, in contexts where overall properties of the data are of interest [14]. These hierarchical phylogenetic models (HPMs) generally have the property of reducing variability in estimates for phylogenetic parameters of individual partitions, while providing a framework for assessing overall tendencies. Examples of HPMs involving the sequence substitution processes in infectious diseases include examining selective pressures in HIV intrahost data, where information pools across multiple patients [15], and estimating the time to most recent common ancestor across the multiple gene segments of influenza [16].

After presenting the details of our novel geographical HPM in §2, we analyse two datasets that showcase our method. In §3, we first present the joint analysis of the dataset of dengue virus in the Americas, mentioned above, for comparison with the previous study [12]. Although our method draws inspiration from purely geographical problems, the model is composed of Markov chains describing the evolution of any discrete trait with a small or single number of realizations per stratum along phylogenetic trees. To demonstrate its application to other traits, in §3, we also present the analysis of intrahost HIV data from 14 different patients. Here, the discrete trait records the different cell compartments from which the virus was isolated, and each patient represents one parallel realization of the intrahost trait process. Finally, in §4, we conclude with some features arising from the methodology and examples.

2. Material and methods

We present a hierarchical model for the evolution of a discrete trait with a single (or small number of) realization(s) in N similar strata. This trait typically tracks geographical location of sampling, but can also represent any discrete character with a fixed number of states, evolving through a homogeneous Markovian process. Each stratum represents a conditionally independent evolutionary history for the same trait. We assume sufficient similarity between the strata for a joint modelling approach. Our model has two levels. At the stratum level, each stratum possesses its own evolutionary process; the hierarchical level shares information across strata.

We have both sequence and trait information for each sample of each stratum. For each individual stratum, we model the sequence data through a phylogenetic tree, with sequence evolution and demographic parameters coming from standard Bayesian phylogenetic methods [17].

We denote the trait data for each stratum i by $X_i = \{X_{i1}, \ldots , X_{in}\}$, where the number of samples $n_i$ may vary by stratum. For each stratum, we assume that a phylogenetic tree relates the samples, where the state at the root $X_{i,\text{root}}$ and internal nodes $\{X_{i,n+1}, \ldots , X_{i,2n-2}\}$ are not observed. Following the approach of Lemey et al. [8], we model trait evolution on the tree as a continuous time Markov chain (CTMC) with infinitesimal rate matrix $A_i = (\lambda_{ik})$. This $K \times K$ rate matrix for stratum $i$ is parametrized as $A_i = \mu_i S_i I_i$, \hspace{1cm} (2.1)

where $I_i = \text{diag}(\pi_1, \ldots , \pi_{K})$ is a diagonal matrix with equilibrium frequencies for each state, $\mu_i$ is a scalar overall transition rate and $S_i = \{s_{jk}\}$ is a $K \times K$ matrix normalized to give overall transition rate $1$. When the matrix $S_i$ is taken to be symmetric, then the CTMC is time reversible.

For traits with a large number of sampling states $K$, as is commonly the case for geography, the number of rates $K(K - 1)$ is large. Because each sequence only has one sampling location, we expect a priori that many of the possible transitions will be very unlikely, rendering a sparse matrix $A$. Thus, we adopt the approach of Lemey et al. [8] and use Bayesian stochastic search variable selection (BSSVS) to select a parsimonious interpretation of $A$. This
parallel stratum, an analogous graph indicates the positive rates that define trait evolution along the tree for the stratum. Connecting the states of the CTMC. When extra information is available, alternative migration graph, whose nodes are the sampling states of k to j. The impact graph represents a non-zero infinitesimal rate from state j to k is the rate from state j to k is $\lambda_{jk}$. At the hierarchical level, our model is composed of a migration graph, whose nodes are the sampling states of the process. A directed edge from node j to k on this hierarchical graph represents a non-zero infinitesimal rate from state j to k, and is present when the indicator $\delta_{ijk} = 1$. The hierarchical graph is a representation of the overall evolutionary process of the trait across strata. It highlights which transitions are dominant in the model. By introducing this hierarchical level, we create a structure through which information is shared across the different strata.

Individual strata are allowed to diverge from the hierarchical graph to account for inherent variability. For each stratum i, the number of differences between the hierarchical graph and the one induced by the BSSVS follows a binomial distribution

$$\sum_{j \neq k} |\delta_{ijk} - \delta_{ijk'}| \sim \text{binomial}(v, p),$$

where $p$ is a fixed error parameter, and $v = K(K - 1)$ or $K(K - 1)/2$ depending on whether the process is assumed time reversible. The binomial distribution is chosen mainly for the convenience of independence between edges of the graph. An alternative option that favours the hierarchical and stratum graphs being identical and introduces dependency between edge differences is the geometric distribution.

(a) Priors
We use standard non-informative prior choices for the parameters of the infinitesimal rate matrices at the stratum level, following the suggestions of Lemey et al. [8]. The overall rate parameter, $\mu_{ij}$, is taken to be exponential(1), and the unnormalized elements of $S_{ij}$ are also assumed to be independent exponential(1). When extra information is available, alternative prior specifications for these parameters are possible without affecting the hierarchical structure of the model. Informative priors for the transition rates can be based on geographical distances, population size or even air traffic data. The prior distribution on the indicator variables of the BSSVS procedure is given by the hierarchical level of the model.

We must also set prior distributions on the hyper-parameters of the hierarchical level graph. We assume a priori that each directed edge is included in the model according to a Bernoulli random variable with small success probability $\chi$. The sum of these independent random variables $\sum \delta_{ijk}$ is binomially distributed. In the limit when $\chi \ll K(K - 1)$, this distribution is approximately a Poisson distribution with expected number of edges $K(K - 1)\chi$.

(b) Inference
Inference on this model is made by a Markov chain Monte Carlo procedure, where each parameter of the model is updated in turn to generate a Markov chain whose limiting distribution is the posterior. This is done in accordance with standard Bayesian phylogeographic methods [8].

One note on this procedure is that independent updates of the edge indicator variables from the hierarchical level and all the strata may lead to poor mixing and slow convergence. This is especially the case when the number of states is high. An alternative to independent updates is to jointly update all the indicator variables by adopting a Metropolis–Hastings step with a proposal distribution that updates one edge of the graph in all strata simultaneously. Updating multiple edges at a time also improves the mixing of the chain.

The method described in this paper has been incorporated into the software package BEAST [18].

(c) Bayes factors
To verify support for a particular edge in the migration graph being included in the model, we use the Bayes factor. The Bayes factor measures how the data change the support for edge $[jk]$ being included in the graph relative to the change in

**Figure 1.** Schematic of the hierarchical structure of the model. At the hierarchical level, the hierarchical migration graph summarizes the overall process. For each parallel stratum, an analogous graph indicates the positive rates that define trait evolution along the tree for the stratum.
support for it being excluded. Formally, the Bayes factor is defined as the ratio of the marginal likelihood of a model and the marginal likelihood of the alternative. For graph \( i \), it can be computed simply as the ratio between posterior and prior odds

\[
\text{BF}_{ijk} = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p_{ijk}/(1 - p_{ijk})}{q_{ijk}/(1 - q_{ijk})}.
\]

The posterior probability of the edge \( \{jk\} \) is the posterior mean of the indicator \( \delta_{jk} \). For the hierarchical graph, the prior \( q_{ijk} \) is obtained from the Poisson distribution. For the stratum graphs, the prior probability of edge \( \{jk\} \) being included in the model depends on the distribution of edge differences between hierarchical and stratum graphs. If we adopt the binomial\((n, p)\) distribution for these differences, then

\[
q_{ijk} = q_{ijk}(1 - p) + (1 - q_{ijk})p.
\]

(d) Entropy

We use the entropy as measure of uncertainty for the distribution of edge inclusions in the migration graph. The entropy of a distribution is a quantity commonly used in information theory (see [19]). For a discrete random variable \( Y \) assuming values \( y_i \), it is defined as

\[
H(Y) = E(- \log(p(Y))) = - \sum (p(y_i) \log(p(y_i))),
\]

where \( p(Y) \) is the probability of \( Y \).

Entropy is used as a measure of uncertainty in probability distributions, and it attains its maximum when the distribution is uniform (ie. all outcomes have the same probability). It is especially useful for assessing variability of distributions over categorical data. When \( Y \) is a vector of random variables with components \( Y_j \), then

\[
H(Y) \leq \sum H(Y_j)
\]

with equality holding when the elements of \( Y \) are independent.

3. Results

We analyse two datasets for which we integrate information across multiple strata to obtain better representations of evolutionary and spatial processes. The first is in the context of geographical dispersion of viral pathogens, in which we analyse the migration pattern of dengue virus in the Americas by combining information from the four different serotypes of the virus. Next, we explore the use of our hierarchical model for a different type of discrete trait: cellular compartments infected by HIV. We use information from 14 different patients to study the intrahost dynamics of the virus between these compartments.

(a) Dengue virus in the Americas

With approximately 50 million infections annually, dengue is a serious public health issue in the tropical and subtropical regions, where its mosquito vectors, \( Aedes aegypti \) and \( Aedes albopictus \), are common [20]. In the Americas, the virus is distributed widely, with reported cases in all but three countries; and the number of cases has been steadily increasing since the 1980s [21]. There are four antigenically distinct dengue virus serotypes (DENV-1 to DENV-4), and we integrate data from all serotypes to study geographical patterns of the virus in the Americas.

We analyse a dataset consisting of 904 sequences of the envelope gene, divided between all four serotypes of the virus. The samples originated from \( K = 36 \) different countries in Latin America and the Caribbean, and date from 1977 to 2009. These data were previously analysed by treating each serotype independently [12].

Because of the biological similarities, we model geographical diffusion for the serotypes jointly to identify the overall dispersal pattern. For each serotype, we have a migration graph, with nodes representing the sampling locations and edges representing which instantaneous rates between locations are non-zero. An overall migration graph summarizes this information at the hierarchical level.

Figure 2 presents these graphs, superimposed over maps [22], for the hierarchical level and each stratum. Thickness of the edges are proportional to edge support, and only those edges that have Bayes factors larger than three are shown.

Our results agree with the previous independent analysis, in
that most of the significant links are between neighbouring countries [12]. An example of this are the highly supported links between Peru, Venezuela and Colombia. Additionally, our model indicates that a few countries such as Colombia, Suriname, Trinidad and Tobago, and Martinique act as centres of viral dispersion, with high number of links to other countries.

An example of how the hierarchical model integrates information across the strata can be seen by comparing edge probabilities in hierarchical and stratum graphs for locations with incomplete sampling. Notice that although we do not have samples from Nicaragua for serotype 4, the hierarchical structure of our model requires that all graphs have the same nodes, and so the probability of an edges linking Nicaragua to other countries are estimated for DENV-4. Building on information from the other three serotypes, the posterior probability for an edge between Nicaragua and Mexico in the hierarchical graph is 0.93. The inclusion probability of this edge for DENV-4 is dictated by the hierarchical prior. Thus, even though the data carry no direct information on the migration of DENV-4 through Nicaragua, we have a Bayes factor of 6.1 for including the edge.

For some edges, there are notable discrepancies between hierarchical and stratum graphs; however, overall deviations conform well with the specified model for all strata. The posterior mean edge differences lie between 0.044 and 0.050 per edge for all serotypes, while the prior model specification assigned a 0.05 probability for an edge in the stratum graph being different from the hierarchical graph. We have also analysed these data using geographical distances to inform the prior rates of migration. The results, however, were only slightly different from those presented here and the qualitative conclusions remained unchanged.

(b) Intrahost HIV-1

Treatment of HIV-1 with highly active antiretroviral therapy (HAART) significantly suppresses viral replication in CD4+ T lymphocytes. In this context, alternative sites of HIV-1 infection, such as CD8+ T cells, may become increasingly important. To assess the role of infection of CD8+ cells in patients under HAART, we study a previously published HIV-1 dataset from 14 different patients [23]. In each patient, the virus was isolated from two different cell compartments, CD4+ T cells and CD8+ T cells, as well as from plasma. We analyse samples collected at two or three different times for each patient, the first being at the time of treatment initiation.

Under the assumption that viral migration between cell types is similar in all patients, each patient represents one stratum in our hierarchical model. Because there are only six edges in the migration graph for this problem, we use a binomial prior with inclusion probability 0.5. Figure 3 presents the hierarchical graph representing the main viral migrations between cell compartments. Directed edges and corresponding Bayes factors are only shown for links with a Bayes factor higher than 1. In this graph, the main connections are between plasma and CD4+ compartments. Additionally, there is evidence for a directed edge from the CD8+ to the CD4+ compartment (BF = 35.3). Equivalent graphs for the patients show the same overall pattern, with the eventual addition of one other edge. These graphs can be found in the electronic supplementary material.

We compare the graph hierarchical model to two alternative approaches for analysing this dataset: the consensus approach, where all patients are assumed to have the same migration matrix, and the independent approach, where patient analyses are carried out separately. We make analogous choices of prior distributions in all analyses. Table 1 presents a comparison between analyses of the uncertainty of graph estimates measured through the entropy of their posterior distributions for edges. The comparison of the hierarchical model and independent analysis shows lower entropy values in the hierarchical model for every patient. This indicates that combining the patient data in the hierarchical setting reduces the uncertainty in the estimates of individual patient graphs. The consensus analysis also presents higher entropy than the overall matrix of the hierarchical model.

In general, the posterior probabilities of edges in the hierarchical and stratum graphs are similar. Table 1 shows that, in the posterior distribution, between 0.05 and 0.1 per cent of the stratum edges differ from their counterparts in the hierarchical graph. The discrepancy fraction observed in the sample varies among patients, and for patients 2, 8 and 10 is close to twice the expected 0.05 defined as the binomial parameter \( p \) for the edge differences.

4. Discussion

We present a HPM for the evolution of a discrete trait in multiple strata. Our method integrates information across the strata to improve estimation, while allowing for inter-strata variation. At the hierarchical level, properties of the overall process are summarized through the migration graph, with edges representing non-zero instantaneous rates.

The main motivation for this method comes from phylogeographic applications, in which the trait of interest is geographical location, and we wish to study the migration process. Recently, many studies have analysed spatial spread of epidemics using analogous Bayesian phylogeographic methods for only one sample of the geographical process [12,24,25]. In addition to the potential public health significance, these studies are motivated by the fact that rapidly evolving pathogens allow for viral sampling in a time frame comparable to sequence evolution, leading to reconstructions of geographical diffusion in real time units. In this context, we present an example in which we use data from multiple serotypes of dengue virus to study the

Figure 3. Migration graphs for the hierarchical level of the intrahost HIV-1 data, with Bayes factor value for each edge in the graph. Only edges with a Bayes factor higher than 1 are shown.
viral dispersion process in the Americas. The analysis supports similar dispersion processes across the four serotypes. In the parallel datasets for which our model is constructed, it is possible that some of the strata have incomplete sampling of state locations, as with the dengue example. In the dengue dataset, some countries do not have samples for all serotypes, yet the hierarchical structure of our model integrates information from the other serotypes to estimate migration rates for the missing strata. Our hierarchical model naturally deals with these often troubling missing data problems.

HPMs have been used in phylodynamic studies of the intrahost behaviour of HIV-1 [15,26]. The long timespan of HIV-1 infections, which sometimes last more than 10 years, combined with high mutation rates, makes phylodynamics a useful tool for assessing viral intrahost biology. Thus, hierarchical models that combine data from multiple patients are relevant. In this context, we present an example in which we use an HIV-1 intrahost dataset to assess the role of infection in CD8<sup>+</sup> T cells for patients under HAART.

Our analysis suggests that an important component of the inter-compartment dynamics is the replenishment of CD4<sup>+</sup> T cells by viral populations from CD8<sup>+</sup> cells. Because CD4<sup>+</sup> cells represent the main pool of infected cells, in line with the higher HIV-1 diversity observed in this compartment [23], this compartment is expected to replenish other cell compartments under a metapopulation dynamics scenario. According to the hierarchical graph, CD4<sup>+</sup> cells could still be the major source of infection for CD8<sup>+</sup> cells when seeded by plasma virus. However, because our analysis also indicates a non-negligible role for viral migration from CD8<sup>+</sup> cells to CD4<sup>+</sup> cells during HAART, as also exemplified by the reconstructed migration history in the maximum clade credibility tree for a particular patient (figure 4), the role of CD8<sup>+</sup> cells in the maintenance of HIV-1 reservoir dynamics may require further attention.

In our method, the conformity of stratum graphs to the hierarchical graph is dependent on the choice of parameter $p$ for the binomial distribution of edge differences. Small values of $p$ induce a large amount of information sharing between strata, and impose a large constraint on similarity between stratum graphs. In the limit, we have the consensus analysis, in which all strata have the same graph. On the other hand, large values of $p$ generate estimates with smaller dependency between graphs.

Even though we use a fixed value for binomial parameter $p$, the fraction of discrepancies in the posterior distribution may differ from the specified probability $p$. This may be used as an indicator of the degree of similarity between the strata, or to identify an outlier stratum. This was observed in the HIV-1 example, where posterior discrepancies were higher in some strata than in others. In comparison, in the dengue example, all strata conformed better to a common dispersion process.

In our choice of prior distribution for the total number of edges in the hierarchical graph, by adopting the Poisson approximation for the sum of a large number of Bernoulli random variables. This choice of prior distribution has been used in a number of subsequent applications, where it has adequately controlled the

Table 1. Comparison of hierarchical model, independent and consensus analysis. Entropy values, computed for the posterior distributions of edge inclusions for the migration graphs, and posterior probability of edge differences between stratum and hierarchical graph.

<table>
<thead>
<tr>
<th></th>
<th>hierarchical model</th>
<th>independent</th>
<th>consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p$</td>
<td>entropy</td>
<td>$p$</td>
</tr>
<tr>
<td>patient 1</td>
<td>0.0667</td>
<td>1.9260</td>
<td>3.0124</td>
</tr>
<tr>
<td>patient 2</td>
<td>0.0929</td>
<td>2.0128</td>
<td>3.1024</td>
</tr>
<tr>
<td>patient 3</td>
<td>0.0692</td>
<td>1.9800</td>
<td>3.5299</td>
</tr>
<tr>
<td>patient 4</td>
<td>0.0565</td>
<td>1.8330</td>
<td>2.7376</td>
</tr>
<tr>
<td>patient 5</td>
<td>0.0734</td>
<td>1.8557</td>
<td>3.0509</td>
</tr>
<tr>
<td>patient 6</td>
<td>0.0631</td>
<td>1.8734</td>
<td>3.4204</td>
</tr>
<tr>
<td>patient 7</td>
<td>0.0685</td>
<td>1.9541</td>
<td>2.7179</td>
</tr>
<tr>
<td>patient 8</td>
<td>0.1094</td>
<td>1.9844</td>
<td>3.2912</td>
</tr>
<tr>
<td>patient 9</td>
<td>0.0803</td>
<td>1.9516</td>
<td>3.4381</td>
</tr>
<tr>
<td>patient 10</td>
<td>0.0944</td>
<td>2.0824</td>
<td>3.2942</td>
</tr>
<tr>
<td>patient 11</td>
<td>0.0698</td>
<td>1.9066</td>
<td>3.2193</td>
</tr>
<tr>
<td>patient 12</td>
<td>0.0764</td>
<td>1.9619</td>
<td>3.1608</td>
</tr>
<tr>
<td>patient 13</td>
<td>0.0729</td>
<td>1.8894</td>
<td>3.2017</td>
</tr>
<tr>
<td>patient 14</td>
<td>0.0549</td>
<td>1.8399</td>
<td>3.0861</td>
</tr>
<tr>
<td>overall</td>
<td>1.6022</td>
<td>2.2937</td>
<td>6.0600</td>
</tr>
</tbody>
</table>

Figure 4. Maximum clade credibility phylogeny for patient 13 with branches coloured according to the most probable posterior cell compartment. Red represents CD8 T cells, blue represents CD4 T cells and yellow represents plasma.
total number of edges included in the graph. Other popular prior choices for graph edges consider edge inclusions as exchangeable Bernoulli trials with common success probability [27,28]. It follows that the total number of edges in the graph has a binomial prior. Additionally, Carvalho & Scott [29] show that when the inclusion probability comes from the Beta hyperprior, the model has a strong control over the number of ‘false’ edges included in the graph. Telesca et al. [30] model dependent gene expression through a graph structure similar to the one in our model, and adopt the binomial-beta model for total number of edges; the authors show through simulations that their model returns good control over false discovery rates.

One advantage of the Bayesian phylogenetic framework we exploit in formulating our graph hierarchical model is that the framework easily lends itself to combination of different models. These could be phylogenetic methods for demographic inference [31], methods for calibrating trees and relaxed clock models [32]. Our hierarchical phylogeographic approach can easily be associated with these existing models to provide comprehensive analysis of viral history.

The data analyses presented here highlight the strengths of the HPM for analysing these parallel datasets consisting of conditionally independent trait evolution processes. In particular, the entropy comparisons for the HIV data show the reduction in overall uncertainty of the hierarchical model, in comparison with the independent approach. This is obtained by sharing information over the parallel strata. On the other hand, the consensus approach does not allow for the variability between strata processes observed in both examples.

Our method paves the way for further exploration of geographical dispersion processes. Through an analysis of the migration graph, for example, we can identify structural properties of the system; fully connected subgraphs and cycles are motifs that may represent local dynamics of interest. We could also assess the reversibility of the geographical process, by testing time-reversibility on the migration matrix. Additionally, changes in the migration process over time could be assessed by generalizing our model into a dynamic model. This method could also be extended to account for more general dependency structures on the hyper-graph.

We thank Forrest Crawford and Kevin Keys for helpful comments on this manuscript. G.B.C. was supported by the Fulbright Science & Technology Fellowship. This project was also supported by NIH grants (nos R01 GM086887 and R01 GM08185) and NSF grant (no. DMS 0556099). The research leading to these results has also received financial support from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 278433-PREDIEMICS and ERC grant agreement no. 260864. Some of this research was completed while the authors were visiting the Institute for Mathematical Sciences, National University of Singapore in 2011.

References

22. Bielejec F, Rambaut A, Suchard MA, Lemey P. 2011 SPREAD: spatial phylogenetic reconstruction of


