PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Phylogeographic reconstruction using air transportation data and its application to the 2009 H1N1 influenza A pandemic

Susanne Reimering[1], **Sebastian Muñoz**[1], **Alice C. McHardy**[1,2]*

1 Department for Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany, **2** German Center for Infection Research (DZIF), Braunschweig, Germany

* Alice.McHardy@helmholtz-hzi.de

OPEN ACCESS

Citation: Reimering S, Muñoz S, McHardy AC (2020) Phylogeographic reconstruction using air transportation data and its application to the 2009 H1N1 influenza A pandemic. PLoS Comput Biol 16 (2): e1007101. https://doi.org/10.1371/journal.pcbi.1007101

Editor: Cecile Viboud, National Institutes of Health, UNITED STATES

Received: May 10, 2019

Accepted: January 12, 2020

Published: February 7, 2020

Data Availability Statement: All the data are available via Github (https://github.com/hzi-bifo/Phylogeography_Paper) and Zenodo (https://zenodo.org/record/2643163).

Funding: SR, SM, and ACM were funded by the Helmholtz Centre for Infection Research (www.helmholtz-hzi.de). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Influenza A viruses cause seasonal epidemics and occasional pandemics in the human population. While the worldwide circulation of seasonal influenza is at least partly understood, the exact migration patterns between countries, states or cities are not well studied. Here, we use the Sankoff algorithm for parsimonious phylogeographic reconstruction together with effective distances based on a worldwide air transportation network. By first simulating geographic spread and then phylogenetic trees and genetic sequences, we confirmed that reconstructions with effective distances inferred phylogeographic spread more accurately than reconstructions with geographic distances and Bayesian reconstructions with BEAST that do not use any distance information, and led to comparable results to the Bayesian reconstruction using distance information via a generalized linear model. Our method extends Bayesian methods that estimate rates from the data by using fine-grained locations like airports and inferring intermediate locations not observed among sampled isolates. When applied to sequence data of the pandemic H1N1 influenza A virus in 2009, our approach correctly inferred the origin and proposed airports mainly involved in the spread of the virus. In case of a novel outbreak, this approach allows to rapidly analyze sequence data and infer origin and spread routes to improve disease surveillance and control.

Author summary

Influenza A viruses infect up to 5 million people in recurring epidemics every year. Further, viruses of zoonotic origin constantly pose a pandemic risk. Understanding the geographical spread of these viruses, including the origin and the main spread routes between cities, states or countries, could help to monitor or contain novel outbreaks. Based on genetic sequences and sampling locations, the geographic spread can be reconstructed along a phylogenetic tree. Our approach uses a parsimonious reconstruction with air transportation data and was verified using a simulation of the 2009 H1N1 influenza A pandemic. Applied to real sequence data of the outbreak, our analysis gave detailed insights into spread patterns of influenza A viruses, highlighting the origin as well as airports mainly involved in the spread.

Introduction

Influenza A viruses continue to impose high mortality and morbidity worldwide [1]. In addition to seasonal epidemics every winter, pandemics can occur when an antigenically novel virus, usually of zoonotic origin, establishes human-to-human transmission. The latest pandemic occurred in 2009, when a novel H1N1 influenza A virus emerged in March and quickly spread around the globe, with 177,000 confirmed infections in over 170 countries until early August [2]. While it is known that viral spread is mainly influenced by air travel [3] and seasonal epidemics are seeded from East and Southeast Asia [4], the exact migration patterns are not fully understood. Especially the inference of transition patterns on a fine-grained scale, e.g. between single countries, states or cities, remains a challenge.

If sequence data together with sampling locations are available, the origin and spread of viruses can be reconstructed using phylogeography. Given a phylogeny as well as locations for the leaf nodes of the tree, phylogeography infers locations for internal nodes of the tree. This approach reconstructs the source of the outbreak as well as spread routes. Current state-of-the-art methods for phylogeography are based on Bayesian inference. Discrete Bayesian phylogeography [5] followed efforts to use parsimonious methods for the reconstruction by minimizing the number of changes between states [6]. Bayesian methods improved this approach by incorporating uncertainty and branch lengths, giving posterior probabilities to evaluate the quality of the reconstruction and allowing the extension to a generalized linear model (GLM) to test potential predictors of viral spread [3]. Therefore, Bayesian phylogeography is now commonly used to study viral pathogens like influenza [4], HIV [7] and Ebola [8]. However, these methods have several drawbacks that have not yet been addressed. First, due to the large number of parameters that are estimated in Bayesian phylogeographic studies, the analysis is slow for larger datasets and the number of distinct states is limited. Instead, locations are often aggregated into larger regions such as continents [3,4], although more fine-grained locations such as countries, states or sometimes even cities are available for a lot of sequences. Second, since discrete Bayesian methods generally estimate rates of movements from the data, these methods will only infer locations which are observed, excluding possible intermediate states which have not been sampled [5,9,10]. Continuous phylogeography, based on inferring geographic coordinates using Brownian diffusion models, are an alternative which allow to infer intermediate locations [9]. While this is a good model for local diffusion of rapidly evolving viruses, it is less applicable for viruses that travel both locally and over large distances in a very short time, e.g. by air travel in the case of influenza A viruses.

Here, we propose a new parsimony-based approach for phylogeographic reconstruction and apply it to study the 2009 outbreak of the pandemic H1N1 (pH1N1) influenza A virus. To use prior knowledge about the mode of travel, we directly include air transportation data via effective distances, which are defined by the number of people travelling from one location to another [11]. The phylogeographic reconstruction then uses the Sankoff algorithm [12] to find internal locations, minimizing the distances along the tree. This approach inherently overcomes the shortcomings of discrete Bayesian methods as in [3,5] and allows both the use of fine-grained locations and the inference of intermediate locations.

We evaluated this approach using simulated data and our recently described distance measure to compare phylogeographic spread among different tree topologies [13]. We showed that effective distances calculated on air passenger flows yield more accurate reconstructions than geographic distances and Bayesian reconstructions with BEAST using the classical approach without any distance information [14], while being comparable to the Bayesian reconstruction using a GLM which includes both geographical and effective distances [3]. We then used this method to study the early spread of the pH1N1 influenza A virus. Our method correctly

inferred Mexico as an origin. Further, we proposed a list of airports that were mainly involved in the initial spread of the virus and seeded a large number of infections in new locations. In the case of future pandemics, this method allows to quickly analyze viral sequence data to identify the origin and major spread routes, which could help to implement surveillance and control measures to contain the spread of the disease.

Results

Phylogeographic reconstruction using simulated data

The early spread of the pH1N1 influenza A virus was simulated using GLEAMviz [15], which has been widely used to simulate this outbreak [16,17] and has been shown to accurately predict influenza activity in various countries [18]. Based on the simulated transitions between locations during the first weeks of the pandemic, we then used FAVITES [19] to simulate the isolate sampling and sequencing, the tree as well as the sequences evolving along the phylogeny 50 times in total. The resulting simulated datasets included on average 97 sequences sampled from 76 unique locations in 23 countries. To confirm that the simulated sequences and the corresponding tree were an accurate representation of the pH1N1 virus, we compared them to real HA sequences of the outbreak that were sampled until the end of April 2009. We calculated pairwise distances between the sequences using a Jukes-Cantor model for both real and simulated sequences (S1A Fig) and further inferred phylogenetic trees to compare branch length distributions (S1B Fig). With both the genetic distances between sequences as well as branch lengths showing a similar distribution, we conclude that the simulation was an accurate representation of the pH1N1 influenza A virus in terms of sequence diversity and tree resolution, which is essential to achieve a comparable accuracy for phylogeographic reconstructions. Most sampling locations in the simulated data were in Mexico and the US, but the virus already spread to Canada, Europe, Asia and Oceania as well (Fig 1A). New infections were

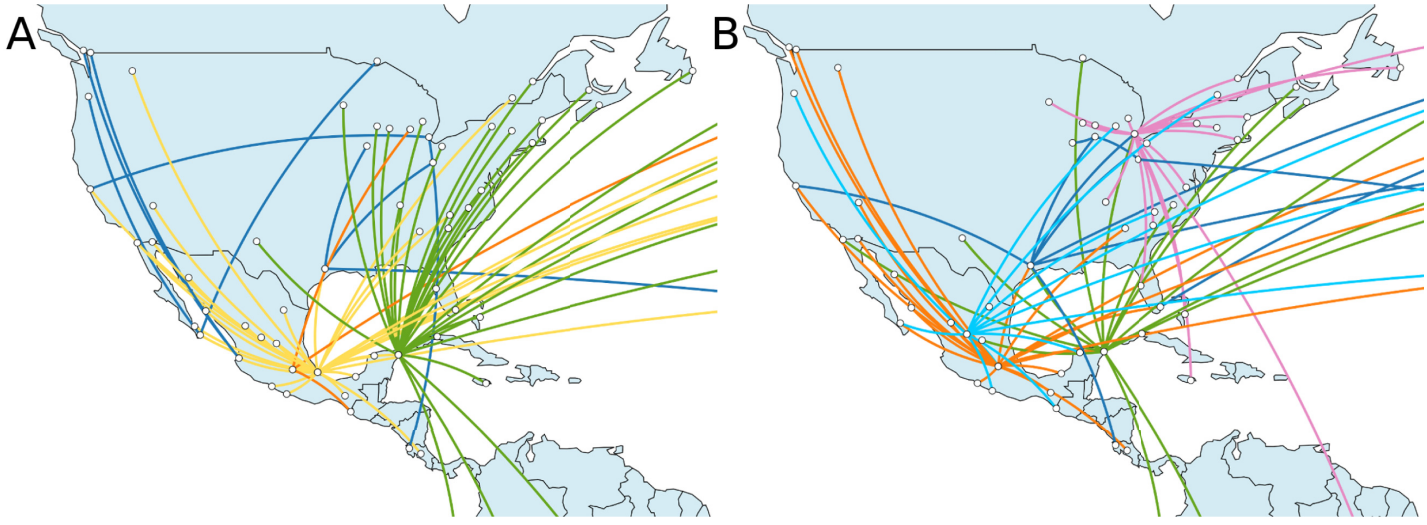

Fig 1. Simulated and reconstructed phylogeographic spread. Viral spread in North and Central America, shown by transitions between locations for the underlying ground truth of the simulation (panel A) and phylogeny, and the inferred spread and phylogeny in panel B, which was reconstructed based on sequences simulated along the phylogeny in A. For the phylogeographic reconstruction, effective distances were used. Transitions between locations are colored by their origin. In the simulation, the outbreak was set in Veracruz (yellow), which was mainly involved in the spread, together with Cancún (green). Transitions from Mexico City are shown in orange, all others in dark blue. In the reconstruction, the origin was placed in Zacatecas (light blue). From there, the virus mainly spread via Mexico City (orange), Cancún (green) and Chicago (pink). All other transitions are shown in dark blue. The actual origin in Veracruz was not inferred and no sequences were sampled here, which is why this location is missing in the map. While the origin and main spread routes differ, the locations are geographically close, leading to a Fréchet tree distance of 39,591.53. The spread was visualized using Spread3 [21].

https://doi.org/10.1371/journal.pcbi.1007101.g001

mainly seeded from Mexico, especially from Veracruz, the origin of the outbreak, and Cancún, the second largest airport of the country. Transitions were given via origin and destination, as output by the GLEAMviz simulation, but were not necessarily the direct route of travel. While Veracruz only has direct flights to a small number of locations, transitions to the US and Europe were possible via connecting flights, e.g. via Mexico City, the main connection from Veracruz. Overall, the simulated locations agreed with the spread of the pandemic until the end of April 2009, where the majority of cases was reported for Mexico and the US, and the first cases occurred on other continents as well [20].

For each simulated dataset, we used the tree inferred on the simulated sequences as well as their sampling locations for a phylogeographic reconstruction with the Sankoff algorithm (Fig 2). We tested both geographic, effective and equal distances. Using equal distances corresponds to a reconstruction with the Fitch algorithm, minimizing the number of changes along the tree, and might be used when no other distance information is available. Further, the phylogeographic reconstruction was done using the simulated tree topology to assess the level of variation introduced by the tree inference and the robustness of our method in case of inaccurate topologies.

An example of a reconstructed spread is shown next to the simulated spread in Fig 1B. In this reconstruction, the inferred origin was Zacatecas, a city north of Mexico City. From there, the virus spread to various locations including Mexico City and Cancún, which seeded a large number of infections in new locations. In the north of the USA, the virus further spread mainly via Chicago. In the simulated spread, locations like Mexico City and Chicago only play a minor role, but they could be unobserved intermediate locations due to connecting flights in the GLEAMviz simulation, where transitions are only reported via origin and destination. However, while the exact reconstructed paths differed from the simulated spread, most of the inferred internal locations in Mexico were geographically close. To quantify these geographic differences between spread paths, reconstructed phylogeographies were compared to the known spread on the simulated tree by calculating discrete Fréchet tree distances [13] using geographic distances between locations (Fig 3A). This method compares the paths of locations from the root to each leaf node, calculates discrete Fréchet distances [22] between them and corrects the distance for each node by the number of paths.

For both the inferred and the simulated trees, the reconstruction using effective distances resulted in a lower distance compared to the reconstruction using geographic and equal distances (P-values of paired t-test: 7.6×10^{-10} and 1.4×10^{-10} for the comparison to geographic distances (for inferred and reconstructed trees, respectively), and 7.1×10^{-5} and 0.0004 for the comparison to equal distances). However, both reconstructions with effective distances showed a small number of outliers with distances as high as in the analysis with geographic distances. In comparison, the reconstructions with equal distances showed extremely high variation compared to the other distance measures. When using geographic distances, the phylogeographic reconstruction using the inferred tree resulted in higher distances compared to the analysis on the simulated tree (P-value of paired t-test: 3.7×10^{-9}), indicating that errors introduced by the tree inference influenced the results. When using effective and equal distances, no significant difference was observed between reconstructions using the inferred and simulated tree (P-value of paired t-test: 0.1197 for effective distances, 0.0622 for equal distances). However, using the simulated tree resulted into a lower variance.

While the Fréchet tree distance measures distances between the entire spread routes, we further had a closer look at the inferred root locations. Since root locations indicate the possible origin of an outbreak, these are of particular interest. Veracruz, the correct root location in our simulation, wasn't inferred except in a few cases—once using geographic distances on the inferred tree, once using effective distances on the simulated tree, as well as once and five

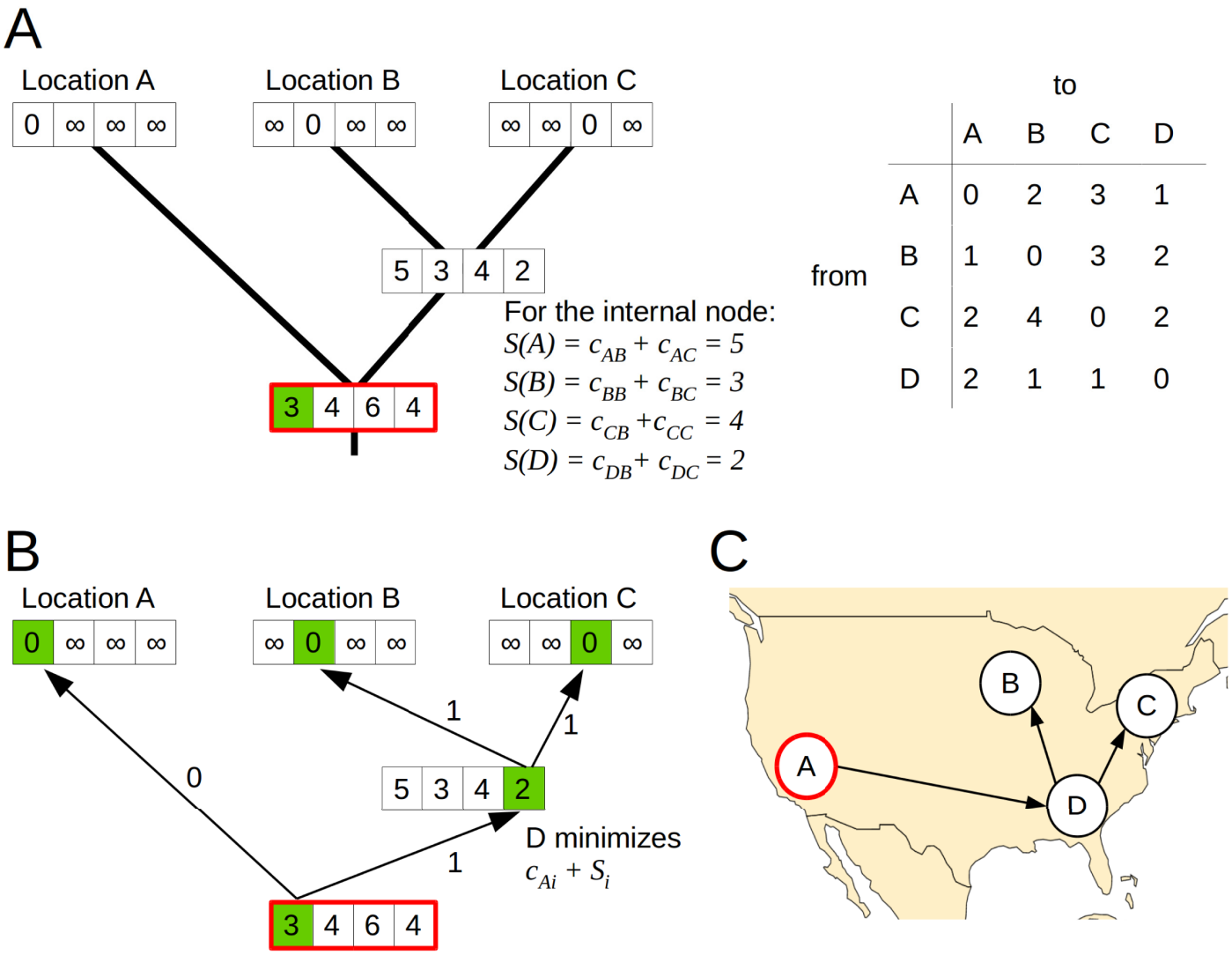

Fig 2. Phylogeographic reconstruction with the Sankoff algorithm using asymmetric, effective distances. Exemplary phylogeographic reconstruction using the Sankoff algorithm on the tree and the cost matrix shown in panel A. The cost matrix c is asymmetric and represents effective distances. For each internal node, the Sankoff algorithm calculates the minimal cost $S(i)$ in the subtree, given the node is assigned location i (shown as the arrays in A, calculated via $S(i) = min_j[c_{ij}+S_l(j)] + min_k[c_{ik}+S_r(k)]$, where l and r denote the two descendant subtrees, and their costs, $S_l(j)$ and $S_r(k)$, respectively. For the root (shown in red), location A results in the minimal cost and is assigned to that node (marked in green). Backtracking from the root to assign all other locations is shown in panel B. Given that a parent node has been assigned state j, the child node will be assigned the state i that minimizes $c_{ji}+S(i)$. The result of the backtracking is indicated by arrows labeled with the costs and the states marked in green. The reconstructed spread along the tree is shown on a map in panel C.

https://doi.org/10.1371/journal.pcbi.1007101.g002

times using equal distances with the inferred and simulated tree, respectively. However, the correct country of origin was inferred in the majority of cases when using effective distances. Interestingly, although no significant differences were observed when comparing the Fréchet tree distances, the reconstruction with the inferred tree topology inferred the country of origin less accurately (in 72% of the cases, compared to 98% on the simulated tree topology). When geographic distances were used, Mexico was only inferred as the origin in 24% of the simulations when using inferred topologies, and 34% when using simulated ones. Instead, the origin was placed in the US for most of the cases (74% and 62%, respectively). For reconstructions using equal distances, Mexico is inferred in 42% and 40% of the simulations, and the US in

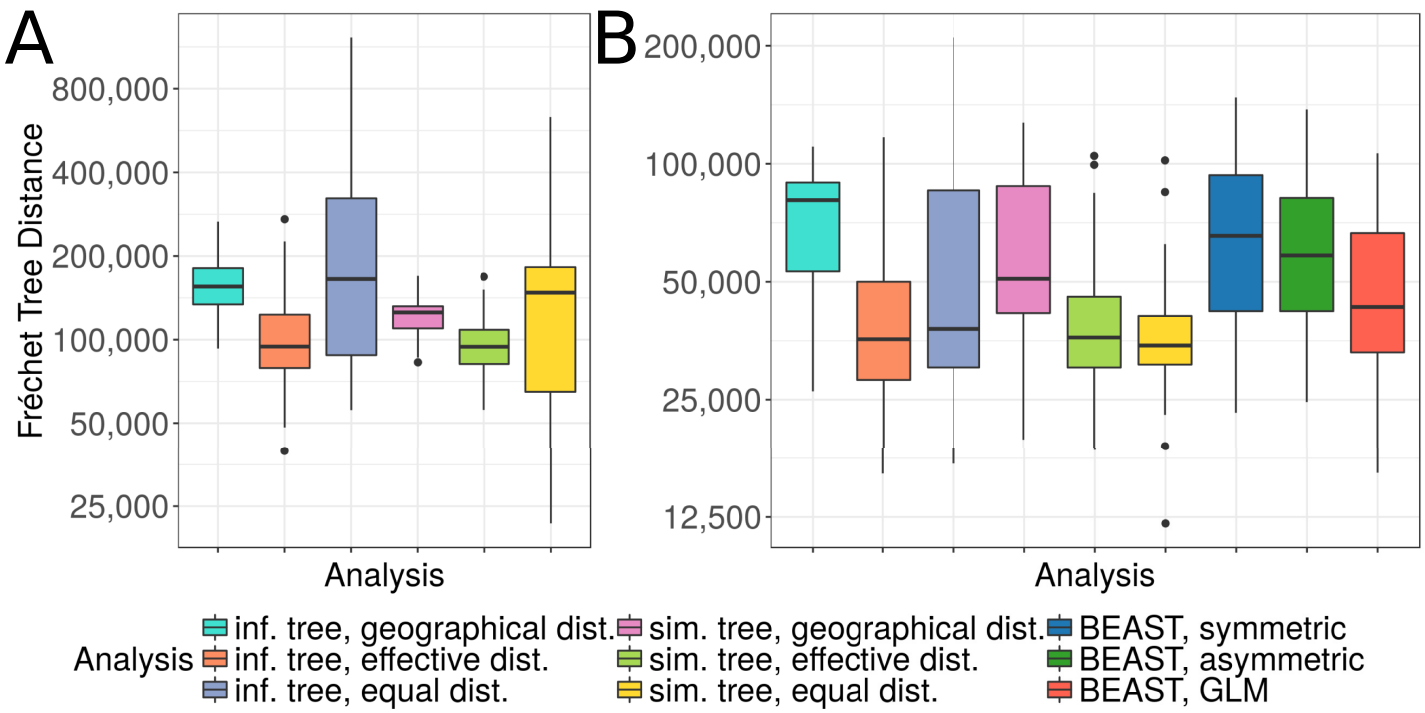

Fig 3. Parsimonious reconstructions with effective distances infer the simulated spread more accurately than reconstructions with geographic or equal distances. Boxplots of discrete Fréchet tree distances comparing simulated phylogeographies to reconstructed phylogeographies, using a total of 50 simulations. Fréchet tree distances are presented on a $\log_2$ scale. In panel A, the reconstructions use airports as locations and were performed on the tree inferred on simulated sequences using geographic (cyan), effective (orange) and equal (light blue) distances, and on the simulated tree, again using geographic (pink), effective (green) and equal (yellow) distances. In panel B, the reconstructions use countries, the same way as described above. Additionally, reconstructions were inferred with BEAST with symmetric rates (dark blue), asymmetric rates (dark green) and a GLM (red).

https://doi.org/10.1371/journal.pcbi.1007101.g003

30% and 40% (inferred and simulated topologies, respectively). In the other simulations, the root is placed in different countries across South America and Europe, causing large deviations from the actual spread.

To ensure that these results are independent of the transitions between locations, which have been simulated only once using GLEAMviz and can vary between different runs, we performed the same analyses on four additional GLEAMviz simulations, with 50 simulations of the sampling, the tree and the sequences each. While there are some slight variations in the distributions of Fréchet tree distances across the five GLEAMviz simulations, the overall conclusions remained the same, with the reconstructions using effective distances consistently showing the smallest Fréchet tree distances (S2 Fig).

Using the same 50 simulations as before, we repeated the analysis using countries instead of airports as locations. With this resolution, a comparison to the Bayesian reconstruction using BEAST was possible. As before, discrete Fréchet tree distances were calculated to compare the reconstruction to the reference data (Fig 3B). The parsimonious reconstruction using countries was comparable to the reconstruction using airports: using effective distances resulted in lower Fréchet tree distances than geographic ones (P-values of paired t-test: 1.5×10^{-9} for inferred trees, 5.6×10^{-6} for simulated trees). Notably, when compared to the other distance measures, using equal distances on the country level resulted in more accurate reconstructions than on the airport level. However, when using the inferred tree topology, using effective distances still resulted into significantly lower Fréchet tree distances (P-value of paired t-test: 0.0328). As observed previously, Fréchet tree distances were lower on reconstructions using

the simulated tree topology as with the inferred one when using geographic distances, but not when using effective ones (P-values of paired t-test: 0.0231 for geographic distances, 0.3483 for effective ones). On the country level, however, a difference was also observed for equal distances (P-value: 0.0021) Phylogeographic reconstructions using BEAST with symmetric rates showed higher distances than the parsimonious reconstruction with effective distances, but were comparable to the reconstruction with geographic distances (P-values of paired t-test: 0.2668 for the comparison to geographic distances, 3.7×10^{-5} for the comparison to effective distances). Using asymmetric rates for the Bayesian reconstruction resulted in similar, but slightly lower distances. We further used a GLM with both geographic and effective distances as potential predictors for the Bayesian reconstruction. Geographical distances were not included in any of the models (Bayes Factor (BF) support < 1 for all 50 datasets), while effective distances were included with at least moderate support in 33 datasets (BF > 3). The reconstruction using the GLM approach showed significantly smaller Fréchet tree distances than the two other Bayesian reconstructions (P-values: 8.1×10^{-5} and 0.0015 compared to using symmetric and asymmetric rates, respectively) as well as the parsimonious reconstruction using geographic distances (P-value: 4.4×10^{-6}). Compared to the reconstruction using effective and equal distances, no significant difference was observed (P-values: 0.1693 and 0.3345, respectively). Enforcing effective distances in the GLM by fixing the predictor in the model did not significantly alter the performance (P-value: 0.0827 for the comparison to the GLM that estimates the predictors, S3 Fig). We further studied the influence of the estimated tree topology on the Bayesian GLM approaches by fixing the simulated tree topology in the analysis. This had two effects on the GLM using both geographic and effective distances as potential predictors. First, when fixing the simulated tree, effective distances were included in all 50 datasets, showing decisive support (BF > 100). As before, geographical distances were not included in any of the models (BF < 1). Second, the analysis showed significantly lower Fréchet tree distances compared to the GLM approach when estimating the tree (P-value: 7.710^{-8}) and the parsimonious approach using effective distances on the simulated tree topology (P-value: 7.0×10^{-5}, S3 Fig). Since the GLM on the simulated tree always included effective distances, fixing effective distances as predictors in the model gave nearly identical results (S3 Fig). These analyses show that the Bayesian GLM approach is sensitive to errors in the estimated tree topology. Given the correct topology, this approach outperforms the parsimonious reconstruction; however, this advantage only holds in theory, as the correct topology is always unknown in practice.

For nearly all datasets and analyses, the root was placed either in Mexico or the US, with effective distances inferring Mexico more often (70% of the cases using the inferred tree topology and 96% on the simulated one) than geographic distances (34% and 62%, respectively) and equal distances (56% and 84%). Bayesian analyses inferred Mexico as the origin in 46% of the datasets when using symmetric rates, in 66% when using asymmetric rates, in 68% when using the GLM while estimating the predictors and in 84% when using the GLM with effective distances as fixed predictors. When the simulated tree was fixed in the GLM analysis, Mexico was inferred as the root location in all datasets.

Phylogeographic reconstruction of the early spread of the pandemic H1N1 influenza A virus

To test the parsimonious reconstruction with effective distances on a real dataset, we analyzed the 378 HA sequences sampled from the beginning of the pH1N1 outbreak until the end of April 2009. For the phylogeographic reconstruction, locations were assigned to each sequence based on the sampling locations, as stated in the isolate name. The resolution of these locations

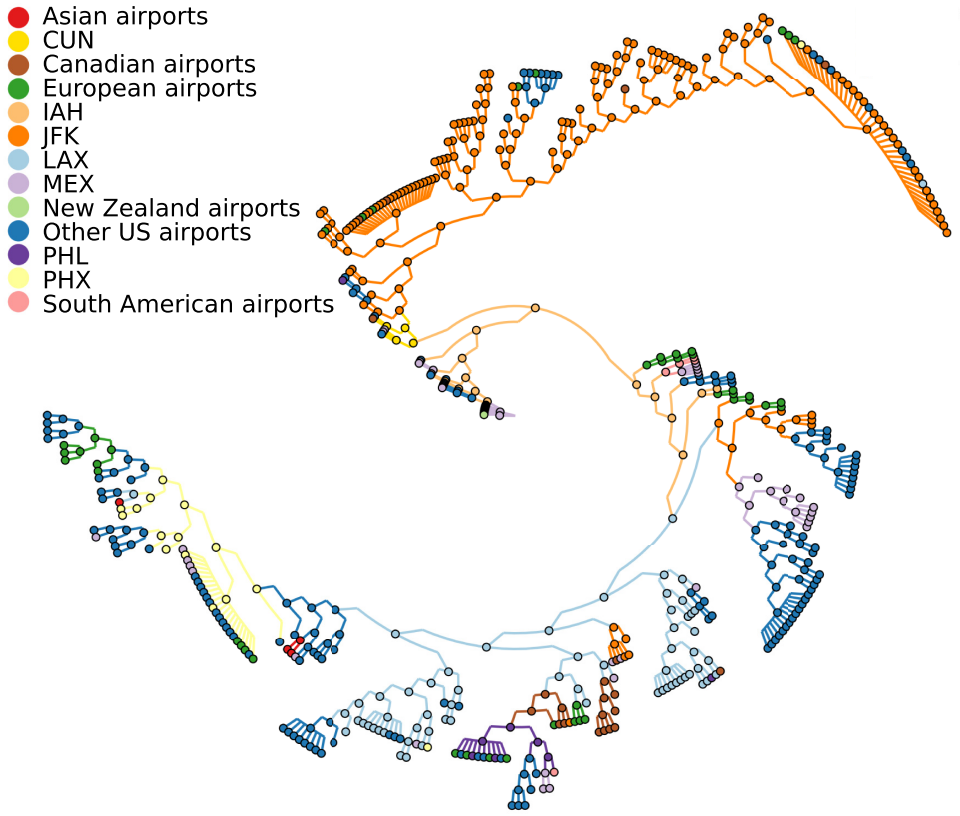

Fig 4. Phylogeny for pH1N1 viruses with reconstructed spread. Parsimonious phylogeographic reconstruction using effective distances on 378 HA sequences from the early stage of the 2009 H1N1 influenza A pandemic. Sequences from 67 different locations were included and represented by the closest airports. The main airports involved in the spread (IAH (Houston), LAX (Los Angeles), JFK (New York), PHL (Philadelphia), PHX (Phoenix)) are shown in separate colors, together with Cancun (CUN), which was not observed in the data but was inferred for some internal nodes. All other locations were summarized per country or continent to allow their visualization. The tree was visualized using GraPhlAn [23].

https://doi.org/10.1371/journal.pcbi.1007101.g004

varied between cities, states and countries. We assigned the main airport (as defined via the highest number of passengers) of the respective city, state or country to the sequence. Three cities did not have an airport and instead we used the geographically closest airport available in our list of locations. In total, this resulted in a set of 67 unique locations. The phylogeographic reconstruction using the Sankoff algorithm was performed using effective distances and is displayed in Fig 4.

Our method inferred Mexico City for the root node of the tree and therefore as the origin of the outbreak. With that, we successfully reconstructed the correct country of origin. Mexico City is the main airport of the country and was assigned to all sequences sampled in Mexico due to the lack of more accurate geographic information. The actual suspected origin in La Gloria in the state Veracruz is around 300km away from our inferred origin. The lack of more precise information about the sampling locations likely prevents our method to infer the origin more closely. However, for some internal nodes our method inferred Cancún, a second location in Mexico. This demonstrates that our approach is able to infer intermediate locations which have not been observed at the tips and might therefore overcome some of the problems introduced by poor sampling resolution. Cancún is the second largest airport in Mexico and a

common tourist destination. Therefore, it likely contributed to the global spread of the pH1N1 influenza A virus. People travelling from Cancún around mid April have been the first reported cases in the UK [24] and are further suspected to be the source of an outbreak at a school in New York [25]. Our reconstruction finds a link from Cancún to New York as well, together with links to Wisconsin and Alberta, Canada.

To explore the reconstructed spread in more detail, we counted the number of transitions to a new location for each observed location. 46 of the 68 locations only occured at terminal branches without seeding infections to new locations. Instead, there was a small number of airports mainly involved in the spread. The airports with the highest numbers of links to new locations were New York (31), Phoenix (24), Los Angeles (21), Houston (13), Mexico City (11) and Philadelphia (11). As expected with effective distances, the method infers large airports, but not only the largest airports of the region (as measured by the number of passengers) or the airports with the highest numbers of sequences (S1 Table). For example, Phoenix is the 10th largest airport in North America with 7 sampled sequences, Houston the 11th largest with 1 sequence, and Philadelphia the 20th largest with 4 sequences. This indicates that the main airports are not only determined by the numbers of passengers or the number of samples, but are dependent on the specific outbreak and its origin.

Discussion

Parsimonious ancestral character state reconstruction has been used for phylogeographic inference in the past [6]. Instead of using the Fitch algorithm for the reconstruction, therefore minimizing the number of state transitions, we here used the Sankoff algorithm to minimize distances between locations along the tree. This allowed us to introduce two innovations compared to the previous parsimony approach with the Fitch algorithm: the inference of unobserved locations for internal nodes, as well as the direct inclusion of air transportation data via effective distances, as is commonly done in Bayesian analyses [3]. While effective distances based on air travel are sensible for influenza A viruses, geographic distances are likely the best choice for pathogens with a local spread, including the historical spread of diseases or the spread of animal viruses like rabies. Effective distances can further be defined using local movements based on commuting data or gravity and radiation models [26,27], which can be useful to study epidemics in single countries where air travel does not play a major role. As long as prior knowledge about the mode of transportation is known, it's simple to adjust the distance matrix used in the phylogeographic reconstruction with the Sankoff algorithm to a specific pathogen. Distance matrices could further offer the opportunity to easily incorporate epidemiological data about the timing of an outbreak by preferring or inhibiting certain transitions.

To evaluate the phylogeographic reconstruction, we first simulated geographic spread using GLEAMviz and then the sampling, trees and sequences using FAVITES. It should be noted that the list of transmissions generated by GLEAMviz didn't include transitions to locations that have been previously infected; an assumption which might be adequate for the very beginning of a pandemic but not in later stages. Further, the sampling process was simplified, with locations randomly chosen for sampling and only a small number of sequences per sampled location, while real influenza A virus sequences usually show a distinct geographic bias. Since the sequence diversity and tree resolution were similar to real pH1N1 sequence data and comparable locations were observed, we believe the simulation was appropriate for an evaluation of phylogeographic methods despite these simplifications. To our knowledge, no other reference dataset with known phylogeographic spread exists as an alternative. In the future, the simulation could be adjusted to study other scenarios and problems in phylogeography that

are difficult to address, e.g. the effect of different sampling biases on phylogeographic reconstructions as well as different strategies to mitigate this bias, for example by using different subsampling schemes.

By comparing the reconstructed spread paths to the simulated ground truth using discrete Fréchet tree distances, we showed that our method using effective distances inferred the phylogeographic spread of pH1N1 more accurately than with geographic distances or equal distances, which are equivalent to previous parsimony approaches using the Fitch algorithm. Furthermore, on the country-level our approach was more accurate than the classical BEAST approach with both symmetric and asymmetric rates and comparable to the GLM diffusion model. The GLM approach only outperformed the parsimonious analysis when the simulated tree topology was fixed in the analysis; however, since this topology is generally unknown, this does not present an advantage in practice. When only considering the root instead of complete paths, the BEAST analysis with asymmetric rates and the GLM inferred the correct country of origin as well as the reconstruction with effective distances, and outperformed the reconstructions using geographic and equal distances. The difference between effective and geographic distances indicated that it's essential to choose suitable distances for the parsimonious analysis. With suitable distances, accurate reconstructions could be achieved that outperformed or were as good as the Bayesian state-of-the-art approaches.

While the result of the parsimonious reconstruction was comparable to the GLM extension of the classical Bayesian methods, our approach allowed to study viral spread in more detail. Instead of summarizing locations into large geographic areas like continents, the phylogeographic reconstruction was possible on fine-grained locations to the resolution of single cities. Both the simulated and real data proved that the analysis was feasible with a large set of total locations (3865 airports in the air transportation network) as well as observed locations; 76 locations on average—including 23 countries—in the simulated data and 67 in the real data. The reconstruction could be done within minutes, while the Bayesian reconstruction was time-consuming and ran for several days to reach a sufficient number of steps in the MCMC, even when using countries instead of airports. In theory, Bayesian reconstructions could be sped up by fixing some parameters instead of estimating them; e.g. by fixing the tree and by fixing the rates between locations to inverse distances. However, this is not commonly done in practice and further did not perform well when we tested this approach to infer airports instead of countries (S4 Fig). The parsimonious reconstruction using the Sankoff algorithm further inferred intermediate states not observed in the sample, while discrete Bayesian methods that estimate the rates from the data will not infer them.

However, the advantages of Bayesian methods to previous parsimonious methods still hold. Bayesian methods allow to integrate over uncertainties in the phylogeny and migration process, while parsimony methods assume a fixed tree topology that was previously inferred from the data and usually differs from the true tree topology. Parsimony methods further don't infer posterior probabilities, therefore giving no indication about the certainty of the results, and do not consider branch lengths. When using the Sankoff algorithm, we assume prior knowledge about the mode of transportation, which might not always be available. Instead, Bayesian methods include the possibility to infer potential modes of transportation by testing predictors of spatial spread. In the future, both methods could be used in a complementary way depending on the data, the desired analysis as well as time and computing resources. To enable others to apply this approach to new datasets, all software is provided in a Github repository [28] and distance matrices are available on Zenodo [29].

The application to a dataset of HA sequences of the pH1N1 virus demonstrates how the parsimonious reconstruction using effective distances can be used in case of new outbreaks, as long as viral sequences and precise geographic information for the isolates are available. This

approach offers a powerful tool to rapidly analyze sequence data, find the place of origin and propose possible spread routes to the resolution of single airports. This information would be helpful to implement control measures like increased surveillance or restricted travel to contain or slow down the global spread of a new infection.

Methods

Simulation

To create a reference dataset, we simulated the beginning of the pH1N1 influenza pandemic in 2009. We first simulated the geographic spread of the virus and then used these transition patterns to simulate isolate sampling, the phylogenetic tree and nucleotide sequences.

The geographic spread simulation was performed with GLEAMviz version 6.6 [15] using a stochastic SEIR (Susceptible-Exposed-Infectious-Recovered) model. We considered three compartments for infectious people: symptomatic (with travel), symptomatic (without travel) and asymptomatic, in line with previous studies [16,17]. In this model, the world is divided into 3252 metapopulations interconnected via an airport and commuting network. We set the origin of the pandemic to Veracruz, Mexico, on February 18th 2009, which was reported as the source and time of the outbreak [30] and has been used in similar models [17]. The simulation produced proportions of individuals in each SEIR compartment per day for each of the metapopulations, as well as the seeding location and day of first arrival of the disease for each newly infected population.

We used the latter as a list of transmissions to simulate the isolate sampling, the tree and the sequences. For these steps, the simulation software FAVITES was used [19]. To simulate the beginning of the pandemic, only the first 200 transition events were included, corresponding to day 85 in the simulation. We chose the number of samples per location via a Poisson distribution with $\lambda = 0.5$, therefore simulating an incomplete sampling of locations. The sampling times were chosen from an uniform distribution. Each sample represented one viral sequence in the subsequent analysis. Based on the sampling events, a tree was simulated using a coalescent model with exponential growth. Branch lengths on this tree corresponded to time as measured in days. We scaled these branch lengths with a rate of 0.000014, therefore assuming a rate of evolution of 0.00511 mutations per site per year, which is close to estimates reported for pH1N1 [31,32]. Given the tree with scaled branch lengths, nucleotide sequences of length 1700 were simulated under a GTR model using seq-gen [33]. This resulted in a set of sequences for each tip in the tree. These sequences along with their sampling locations were then used for phylogeographic inference, while the simulated tree (including the locations for the internal nodes as given by the transition events) was used as a reference dataset.

Data download

Nucleotide sequences of the hemagglutinin (HA) protein of pH1N1 were downloaded from the GISAID database [34]. All complete sequences from the beginning of the pandemic in March until end of April were downloaded, resulting in a dataset of 378 sequences.

Phylogenetic reconstruction

For both simulated and real data, nucleotide sequences were aligned using MUSCLE version 3.8.31 with standard parameters [35]. Positions with gaps in more than 80% of the sequences were removed with TrimAl version 1.2 [36] to ensure a good quality of the alignment. Phylogenetic trees were then reconstructed with FastTree version 2.1.7 [37] using the GTR model. Simulated trees were rooted based on the ancestral sequence used in the simulation process,

which was subsequently removed from the dataset. The tree based on real data was rooted using the sequence with the earliest sampling date.

Parsimonious phylogeographic reconstruction

The phylogeographic reconstruction was based on the air transportation network from the OAG database including 3865 airports, which we used as all possible states in the analysis. Geographic distances between airports were calculated based on longitude and latitude using the R package geosphere [38]. This defined the distances as the shortest paths between locations, taking into account the ellipsoidal surface of the Earth. Effective distances were calculated based on the numbers of passengers travelling between airports in the year 2013, as in [11]. For equal distances, all distances between airports were set to 1. The air transportation network included airports from 228 countries, which were used as locations for the analysis on a country level. Geographic distances between countries were calculated as described above using the coordinates of the centroids. For effective distances, we aggregated the passenger numbers per country and recalculated the distances. The distance matrices were used as a cost matrix for the parsimonious reconstruction using the Sankoff algorithm [12]. Distances represent the cost of traveling from one state to another and the Sankoff algorithm finds the internal states with the minimal cost, i.e. minimizing the distance the virus travelled along the tree. In case of equal distances between locations, this is equivalent to the Fitch algorithm which minimizes the number of changes between states. An example of the inference using asymmetric distances is shown in Fig 2. We used delayed transformation in case of ambiguities. When an ambiguity occured at the root and therefore couldn't be resolved with delayed transformation, we randomly chose one of the possible states. Since effective distances are generally not symmetric, the tree was rooted before the reconstruction.

Bayesian phylogeographic reconstruction

The simulated data included a large number of states with relatively few sequences, which is not suitable for Bayesian phylogeographic reconstruction. Using countries instead of airports reduced the number of states to make the Bayesian reconstruction feasible. We used BEAST version 2.4.8 for this analysis [14]. Sampling dates were included into the phylogenetic reconstruction as measured in days. A HKY nucleotide substitution model was used together with a strict molecular clock. For the tree prior, a coalescent model with exponential population growth was chosen. For the phylogeographic reconstruction, analyses were performed separately with both symmetric and asymmetric rates. Markov Chain Monte Carlo (MCMC) was run for 100 million steps with trees sampled every 10,000 steps, resulting in a sample of 10,001 trees.

Further, the phylogeographic reconstruction was performed using a generalized linear model (GLM) [3], as implemented in BEAST version 1.10.4 [39]. We used both geographic and effective distances as possible predictors, calculated as described above. Distances were log-transformed and standardized before the analysis. Bayes Factors were calculated to evaluate the inclusion of the predictors into the model, using a prior probability of 50% that no predictors are included.

Tracer version 1.7.1 [40] was used to confirm adequate effective sample sizes (ESS), indicating good estimates of the posterior distributions of the parameters. TreeAnnotator was then used to summarize the sampled trees into a maximum clade credibility tree using a burn-in of 10%. For the evaluation of the phylogeographic reconstruction, we assigned the location with the highest posterior probability to each node.

Supporting information

S1 Fig. Comparison of real and simulated data. A) Comparison of pairwise genetic distances between sequences for both real HA sequence data and the simulated sequences. B) Comparison of branch lengths on trees inferred on both real HA sequence data and simulated sequences.
(PNG)

S2 Fig. Fréchet tree distances for four additional GLEAMviz simulations. Fréchet tree distances shown on a $\log_2$ scale for all six analyzed parsimonious reconstructions using four additional simulations (shown in panels A-D) of geographical spread using GLEAMviz. For each simulation of spread, the sampling, the tree and the sequences were simulated 50 times. Reconstructions were performed both on the airport (on the left) and the country level (on the right).
(PNG)

S3 Fig. Fréchet tree distances for reconstructions with BEAST using fixed rates and a fixed tree topology on the country level. Fréchet tree distances shown on a log2 scale on the country level, including the additional BEAST analyses using the GLM with effective distances as fixed predictors and/or the fixed simulated tree.
(PNG)

S4 Fig. Fréchet tree distances for reconstructions with BEAST using fixed rates and a fixed tree topology on the airport level. Fréchet tree distances shown on a $\log_2$ scale for all six analyzed parsimonious reconstructions on the airport level in comparison to a BEAST reconstruction with fixed rates and a fixed tree topology. The rates were set to inverse effective distances while the tree topology was set to the tree inferred using Fasttree, which allowed for short MCMC runs with 1 million steps and therefore enabled a reconstruction on the airport level with a larger number of locations.
(PNG)

S1 Table. Main airports involved in the spread of pH1N1. Main airports involved in the spread of pH1N1, as measured by the number of transitions to new locations. The airports are New York (JFK), Phoenix (PHX), Los Angeles (LAX), Houston (IAH), Mexico City (MEX) and Philadelphia (PHL). The size of the airports in North America was determined by the numbers of passengers in 2013 as given via the OAG database.
(PDF)

Acknowledgments

We thank Dirk Brockmann for providing the air passenger data to calculate effective distances between airports.

Author Contributions

Conceptualization: Susanne Reimering, Sebastian Muñoz, Alice C. McHardy.

Data curation: Susanne Reimering, Sebastian Muñoz.

Formal analysis: Susanne Reimering, Sebastian Muñoz.

Funding acquisition: Alice C. McHardy.

Investigation: Susanne Reimering.

Methodology: Susanne Reimering, Sebastian Muñoz, Alice C. McHardy.

Project administration: Alice C. McHardy.

Software: Susanne Reimering.

Supervision: Alice C. McHardy.

Validation: Susanne Reimering.

Visualization: Susanne Reimering.

Writing – original draft: Susanne Reimering.

Writing – review & editing: Susanne Reimering, Alice C. McHardy.

References

1. WHO. Influenza (Seasonal) [Internet]. 6 Nov 2018 [cited 5 Dec 2019]. Available: https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)
2. WHO. Pandemic (H1N1) 2009—update 61 [Internet]. 12 Aug 2009 [cited 5 Dec 2019]. Available: https://www.who.int/csr/don/2009_08_12/en/
3. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014; 10: e1003932. https://doi.org/10.1371/journal.ppat.1003932 PMID: 24586153
4. Bedford T, Riley S, Barr IG, Broor S, Chadha M, Cox NJ, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature. 2015; 523: 217–220. https://doi.org/10.1038/nature14460 PMID: 26053121
5. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009; 5: e1000520. https://doi.org/10.1371/journal.pcbi.1000520 PMID: 19779555
6. Wallace RG, Hodac H, Lathrop RH, Fitch WM. A statistical phylogeography of influenza A H5N1. Proc Natl Acad Sci USA. 2007; 104: 4473–4478. https://doi.org/10.1073/pnas.0700435104 PMID: 17360548
7. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science. 2014; 346: 56–61. https://doi.org/10.1126/science.1256739 PMID: 25278604
8. Tong Y-G, Shi W-F, Liu D, Qian J, Liang L, Bo X-C, et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. Nature. 2015; 524: 93–96. https://doi.org/10.1038/nature14490 PMID: 25970247
9. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. Mol Biol Evol. 2010; 27: 1877–1885. https://doi.org/10.1093/molbev/msq067 PMID: 20203288
10. Magee D, Scotch M. Conceptualizing a Novel Quasi-Continuous Bayesian Phylogeographic Framework for Spatiotemporal Hypothesis Testing. AMIA Jt Summits Transl Sci Proc. 2015; 2015: 212–216. PMID: 26306274
11. Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. Science. 2013; 342: 1337–1342. https://doi.org/10.1126/science.1245200 PMID: 24337289
12. Sankoff D. Minimal mutation trees of sequences. SIAM J Appl Math. 1975; 28: 35–42. https://doi.org/10.1137/0128004
13. Reimering S, Muñoz S, McHardy AC. A Fréchet tree distance measure to compare phylogeographic spread paths across trees. Sci Rep. 2018; 8: 17000. https://doi.org/10.1038/s41598-018-35421-4 PMID: 30451977
14. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014; 10: e1003537. https://doi.org/10.1371/journal.pcbi.1003537 PMID: 24722319
15. Van den Broeck W, Gioannini C, Gonçalves B, Quaggiotto M, Colizza V, Vespignani A. The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. BMC Infect Dis. 2011; 11: 37. https://doi.org/10.1186/1471-2334-11-37 PMID: 21288355
16. Balcan D, Hu H, Goncalves B, Bajardi P, Poletto C, Ramasco JJ, et al. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. BMC Med. 2009; 7: 45. https://doi.org/10.1186/1741-7015-7-45 PMID: 19744314

17. Bajardi P, Poletto C, Ramasco JJ, Tizzoni M, Colizza V, Vespignani A. Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. PLoS ONE. 2011; 6: e16591. https://doi.org/10.1371/journal.pone.0016591 PMID: 21304943

18. Tizzoni M, Bajardi P, Poletto C, Ramasco JJ, Balcan D, Gonçalves B, et al. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. BMC Med. 2012; 10: 165. https://doi.org/10.1186/1741-7015-10-165 PMID: 23237460

19. Moshiri N, Ragonnet-Cronin M, Wertheim JO, Mirarab S. FAVITES: simultaneous simulation of transmission networks, phylogenetic trees, and sequences. Bioinformatics. 2018; 35: 1852–1861. https://doi.org/10.1093/bioinformatics/bty921 PMID: 30395173

20. WHO. Influenza A(H1N1)—update 6 [Internet]. 30 Apr 2009 [cited 5 Dec 2019]. Available: https://www.who.int/csr/don/2009_04_30_a/en/

21. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. Spread3: interactive visualization of spatiotemporal history and trait evolutionary processes. Mol Biol Evol. 2016; 33: 2167–2169. https://doi.org/10.1093/molbev/msw082 PMID: 27189542

22. Eiter T, Mannila H. Computing Discrete Fréchet Distance. Tech. Report CD-TR 94/64, Christian Doppler Laboratory for Expert Systems, TU Vienna, Austria. 1994. Available: http://www.kr.tuwien.ac.at/staff/eiter/et-archive/cdtr9464.pdf

23. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ. 2015; 3: e1029. https://doi.org/10.7717/peerj.1029 PMID: 26157614

24. Young N, Pebody R, Smith G, Olowokure B, Shankar G, Hoschler K, et al. International flight-related transmission of pandemic influenza A(H1N1)pdm09: an historical cohort study of the first identified cases in the United Kingdom. Influenza Other Respi Viruses. 2014; 8: 66–73. https://doi.org/10.1111/irv.12181 PMID: 24373291

25. Lessler J, Reich NG, Cummings DAT, New York City Department of Health and Mental Hygiene Swine Influenza Investigation Team, Nair HP, Jordan HT, et al. Outbreak of 2009 pandemic influenza A (H1N1) at a New York City school. N Engl J Med. 2009; 361: 2628–2636. https://doi.org/10.1056/NEJMoa0906089 PMID: 20042754

26. Manitz J, Kneib T, Schlather M, Helbing D, Brockmann D. Origin Detection During Food-borne Disease Outbreaks—A Case Study of the 2011 EHEC/HUS Outbreak in Germany. PLoS Curr. 2014; 6. https://doi.org/10.1371/currents.outbreaks.f3fdeb08c5b9de7c09ed9cbcef5f01f2 PMID: 24818065

27. Simini F, González MC, Maritan A, Barabási A-L. A universal model for mobility and migration patterns. Nature. 2012; 484: 96–100. https://doi.org/10.1038/nature10856 PMID: 22367540

28. Reimering S, Munoz S, McHardy AC. hzi-bifo/Phylogeography_Paper. Zenodo. 2019; https://doi.org/10.5281/zenodo.2671612

29. Reimering S, Munoz S, McHardy AC. Distance matrices for parsimonious phylogeography. Zenodo. 2019; https://doi.org/10.5281/zenodo.2643162

30. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. Science. 2009; 324: 1557–1561. https://doi.org/10.1126/science.1176062 PMID: 19433588

31. Su YCF, Bahl J, Joseph U, Butt KM, Peck HA, Koay ESC, et al. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. Nat Commun. 2015; 6: 7952. https://doi.org/10.1038/ncomms8952 PMID: 26245473

32. Rambaut A, Holmes E. The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. PLoS Curr. 2009; 1: RRN1003. https://doi.org/10.1371/currents.RRN1003 PMID: 20025195

33. Rambaut A, Grass NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics. 1997; 13: 235–238. https://doi.org/10.1093/bioinformatics/13.3.235 PMID: 9183526

34. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data—from vision to reality. Euro Surveill. 2017; 22: 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 PMID: 28382917

35. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32: 1792–1797. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

36. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25: 1972–1973. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945

37. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010; 5: e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

38. Hijmans RJ. geosphere: Spherical Trigonometry. 2017. Available: https://CRAN.R-project.org/package=geosphere

39. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018; 4: vey016. https://doi.org/10.1093/ve/vey016 PMID: 29942656

40. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in bayesian phylogenetics using tracer 1.7. Syst Biol. 2018; 67: 901–904. https://doi.org/10.1093/sysbio/syy032 PMID: 29718447