Heredity (2011) 106, 383–390
© 2011 Macmillan Publishers Limited All rights reserved 0018-067X/11
www.nature.com/hdy

ORIGINAL ARTICLE

Reconstructing disease outbreaks from genetic data: a graph approach

T Jombart, RM Eggo, PJ Dodd and F Balloux
Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analysis and Modelling, Imperial College Faculty of Medicine, London, UK

Epidemiology and public health planning will increasingly rely on the analysis of genetic sequence data. In particular, genetic data coupled with dates and locations of sampled isolates can be used to reconstruct the spatiotemporal dynamics of pathogens during outbreaks. Thus far, phylogenetic methods have been used to tackle this issue. Although these approaches have proved useful for informing on the spread of pathogens, they do not aim at directly reconstructing the underlying transmission tree. Instead, phylogenetic models infer most recent common ancestors between pairs of isolates, which can be inadequate for densely sampled recent outbreaks, where the sample includes ancestral and descendent isolates. In this paper, we introduce a novel method based on a graph approach to reconstruct transmission trees directly from genetic data. Using simulated data, we show that our approach can efficiently reconstruct genealogies of isolates in situations where classical phylogenetic approaches fail to do so. We then illustrate our method by analyzing data from the early stages of the swine-origin A/H1N1 influenza pandemic. Using 433 isolates sequenced at both the hemagglutinin and neuraminidase genes, we reconstruct the likely history of the worldwide spread of this new influenza strain. The presented methodology opens new perspectives for the analysis of genetic data in the context of disease outbreaks.

Heredity (2011) **106,** 383–390; doi:10.1038/hdy.2010.78; published online 16 June 2010

Keywords: ancestry; reconstruction; graph; phylogeny; genetic monitoring; outbreak

Introduction

The most effective strategy to avert infectious disease epidemics is to identify outbreaks at an early stage and identify their transmission routes. For example, wider spread of nosocomial bacteria could be best avoided by rapid identification of colonized medical personnel spreading the infection (Albrich and Harbarth, 2008). Early identification of preferential transmission routes could also prove critical for containing wider epidemics, such as SARS or avian influenza (Cooper *et al.*, 2006; Ferguson *et al.*, 2006; Germann *et al.*, 2006). To be most effective, such prophylactic interventions must be implemented early on and will be based on only very preliminary scientific evidence. Thus, it is crucial that all sources of useful information be considered. Genetic sequence data can now be generated essentially in real time and the analysis of sequence data has already been added to the toolbox of infectious disease epidemiology (Rambaut *et al.*, 2008; Russell *et al.*, 2008; Cottam *et al.*, 2008b; Garten *et al.*, 2009; Sloan *et al.*, 2009). However, the statistical methodologies to harness the full potential of genetic sequence information are currently lacking. In particular there is a need for methods devoted to the reconstruction of the transmission tree of a set of isolates collected during the early stages of an outbreak.

Current state-of-the-art genetic methods for the reconstruction of pathogen genealogies rely on the phylogenetic paradigm (Grenfell *et al.*, 2004; Templeton, 1998; Drummond and Rambaut, 2007; Cottam *et al.*, 2008a; Lemey *et al.*, 2009a) based on the reconstruction of ancestries between hypothetical common ancestors (most recent common ancestor) and sampled isolates. Unfortunately, such approaches may be inappropriate when ancestors and their descendents are both present in the sample analyzed, as is likely during the early stages of an outbreak. Phylogenetic methods consider that sampled strains are all tips of an unknown genealogy, making it impossible for a sampled strain to be (directly or indirectly) ancestral to another sampled strain. By definition, these methods are thus unable to uncover ancestries between sampled isolates, and can fail to reconstruct the transmission tree of an emerging pathogen (Figures 1a–c).

To circumvent this problem, we introduce a new methodological approach for the analysis of genetic data collected during disease outbreaks. Contrary to phylogenetic methods, we consider that sampled isolates represent a fraction of the genealogy, including internal nodes and tips (Figure 1d). Direct and indirect ancestries can therefore be inferred between the sampled isolates. Within this framework, we developed an algorithm called *SeqTrack*, which directly reconstructs the most

Correspondence: Dr T Jombart or Dr F Balloux, Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analysis and Modelling, Imperial College Faculty of Medicine, St Mary's Campus, Norfolk Place, London W2 1PG, UK.
E-mail: t.jombart@imperial.ac.uk or f.balloux@imperial.ac.uk

Received 10 March 2010; revised 26 April 2010; accepted 5 May 2010; published online 16 June 2010

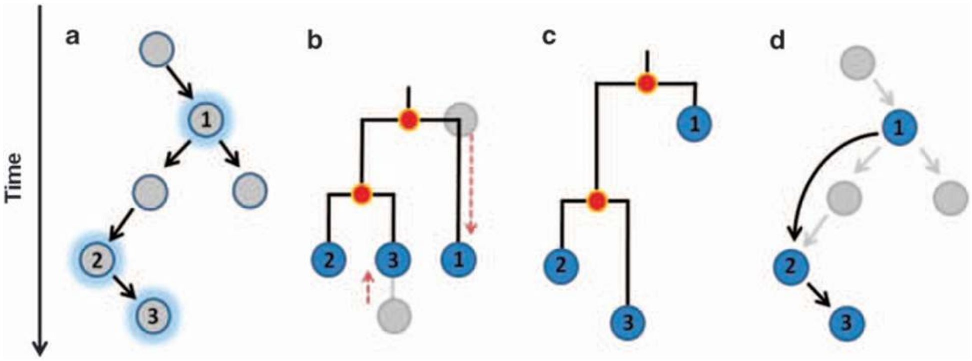

Figure 1 Illustration of possible reconstructions of genealogical relations. Panel **a** represents a hypothetical genealogy of six isolates from which three were sampled (highlighted by a blue glow and numbered from 1 to 3). Panels **b** and **c** illustrate phylogenetic reconstructions where the red dots represent inferred nodes; panel **b** corresponds to a classical phylogenetic reconstruction and **c** to a tree that accounts for heterochronous sampling (e.g., BEAST software; Drummond and Rambaut, 2007). Panel **d** illustrates the direct ancestry reconstruction method implemented in *SeqTrack*.

plausible genealogy of a set of sampled isolates, allowing for a direct assessment of the spatiotemporal dynamics of the epidemic under study.

Using simple simulations, we illustrate the major difference between phylogenetic methods and our approach, showing the advantage of *SeqTrack* over traditional phylogenetic approaches for inferring transmission trees in densely sampled disease outbreaks. We then evaluate the performance of our method using spatially explicit simulations of outbreaks. Finally, we illustrate the novel methodology using sequence data from the early stages of the 2009 A/H1N1 influenza pandemic, and reconstruct the likely scenario of worldwide spread of this newly emerged pathogen.

Materials and methods

SeqTrack algorithm

Our method aims to reconstruct the transmission tree of pathogens during a disease outbreak, using genotypes and collection dates to uncover ancestries between sampled isolates. The fundamental innovation of our approach is to seek ancestors directly from the sampled isolates, rather than attempting to reconstruct unobserved and hypothetical ancestral genotypes. Because ancestries are in essence temporally oriented connections between isolates, it seemed natural to tackle the problem of inferring genealogies within graph theory.

SeqTrack relies on three fundamental, yet simple observations. First, in the absence of genetic recombination, each observed isolate has one, and only one ancestor. Second, ancestors always precede their descendents in time. And third, among all possible ancestries of a given isolate, some are more likely than others, and this likelihood can be inferred from the amount of genetic differentiation between the isolates considered. The purpose of *SeqTrack* is to identify the most likely genealogy. Technically, this problem translates into finding the optimum branching in a directed graph, where each node is an isolate, and where a given isolate is connected to all isolates occurring strictly later.

Let $\mathcal{G} = (S, E, w)$ be a directed, weighted graph where $S = \{s_1,\ldots, s_n\}$ is the set of vertices corresponding to the n sampled isolates, with associated collection dates $T = \{t_1,\ldots, t_n\}$. E is the set of directed edges of $\mathcal{G}$ modeling all possible ancestries in S, such that $(s_i, s_j) \in E$ if and only if $t_i < t_j$. The weight function w: $E \rightarrow \mathbb{R}$ assigns a weight to each possible ancestry, which reflects how credible each ancestry relationship is. For instance, w could be defined as a genetic distance or similarity, or as the log-likelihood resulting from a probabilistic model of evolution. The weight of a subset A of E is computed as $w_A = \sum_{e \in A} w(e)$. The 'best' genealogy of the sampled isolates is the spanning directed tree (that is, 'branching' reaching all the nodes) $\mathcal{J} = (S,B,w)$ with $B \subseteq E$ optimizing (that is, minimizing or maximizing) w_B. Whenever sampled isolates form more than one connected set, the algorithm is applied to each connected set separately (see example in Figure 3b).

This problem has been solved by Edmonds (1967) and Chu and Liu (1965), who developed an algorithm to find $\mathcal{J}$ so that w_B is maximum or minimum. The algorithm proceeds by identifying optimum ancestors for each node at the exception of the root (the oldest isolate), and then recursively removes possible cycles. However, in our case, cycles are impossible as ancestries cannot go back in time, which greatly simplifies computations.

The entire *SeqTrack* procedure has been implemented in the *adegenet* package (Jombart, 2008) for the R software (R Development Core Team, 2009). This implementation allows for specifying any weight w_i and for choosing the type of optimization (minimization or maximization of total weight) to take advantage of the method's versatility.

Maximum parsimony genealogies using *SeqTrack*

Outbreak genetic data are typically characterized by low genetic diversity, implying that descendents differ from their ancestors by very few mutations. When recombination and reverse mutations are unlikely, the most plausible genealogy is the one involving the fewest mutational steps. Such maximum parsimony genealogies can be recovered by *SeqTrack* by weighting edges according to the number of mutations between pairs of isolates, and seeking the branching of minimum weight. The resulting ancestry path connects all the sampled isolates while minimizing the required number of mutational steps, and respecting their temporal ordering.

Given the low levels of genetic diversity expected during the early stages of disease outbreaks, isolates with identical haplotypes but different collection dates and locations are likely to occur, resulting in several possible choices for the most parsimonious ancestry. In such cases, *SeqTrack* can use an optional rule to select the 'best' ancestor, and select the closest ancestor according to a given proximity measure. For instance, this approach can be used to select, among all equally parsimonious ancestors, the one being geographically closest. Another possibility would consist in favoring ancestors whose haplotype is most frequent in the sampled isolates. By default, *SeqTrack* resolves ties in the choice of ancestors by maximizing the likelihood of the genetic differentiation observed between the ancestor and the descendent. The number of mutations v occurring between two isolates sampled Δ_t time units apart follows a Poisson distribution:

$$v \sim \text{Poisson}(\mu L \Delta_t) \qquad (1)$$

where L and μ are, respectively, the length (in number of nucleotides) and the mutation rate (per nucleotide and per time unit) of the considered DNA sequence. Note that this likelihood can also be used to assess the statistical support of each ancestry, as we do in Figure 5.

Simulated data

Individual-based simulations were used to assess the efficacy of our method for reconstructing transmission trees from outbreak genetic data. All simulations were performed using the R software, and procedures that we implemented in the *adegenet* package. The simulation system is further described in Supplementary Information.

We first used simulations to compare *SeqTrack* to classical phylogenetic reconstruction. Because the outputs of both approaches are very different (phylogenies do not provide ancestries between tips), we could not proceed to systematic comparisons over a large number of data sets. However, while not quantitative, the comparisons we performed can be used to show the fundamental differences between both approaches. We simulated two simple genealogies of isolates, using haplotypes of 10 000 nucleotides and a mutation rate of 10^{-4} mutation per nucleotide and generation, over three and four generations, respectively (Figures 2a and 3a). In the second data set (Figure 3), only a portion of the genealogy was sampled, so as to show the behaviors of the methods for inferring ancestries with unsampled ancestors. Note that the two data sets were chosen randomly, and the results were qualitatively similar in all other simulations we examined. *SeqTrack* was used to obtain maximum parsimony genealogies (Figures 2b and 3b). A classical phylogeny was obtained by neighbor-joining using pairwise nucleotidic distances between sampled isolates (Figures 2c and 3c). To give the neighbor-joining phylogenetic trees the right direction, trees were rooted using the most ancient isolate sampled. We also analyzed these data using temporally informed phylogenetic reconstruction as implemented in the BEAST software (Drummond and Rambaut, 2007; Figures 2d and 3d). We used the exponential growth model with strict molecular clock model, keeping other parameters to their default settings. A consensus phylogeny was computed on the 5000 last simulated trees.

Simulations were also used for a quantitative assessment of *SeqTrack*'s performances. Data sets were obtained

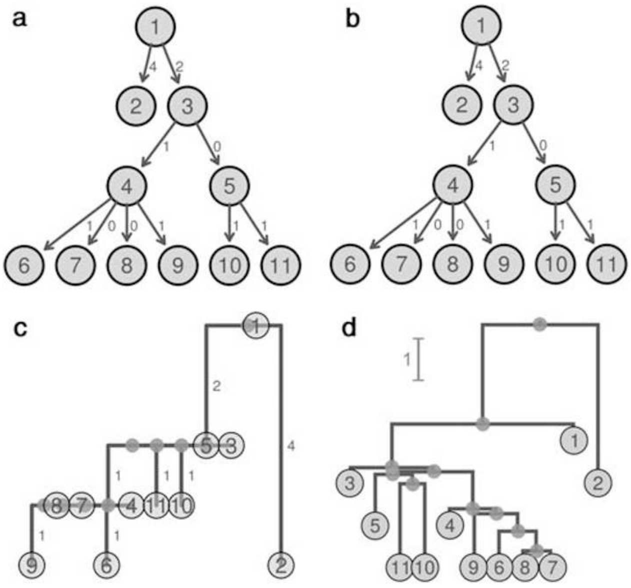

Figure 2 Reconstruction of a simple simulated genealogy by two different approaches. This figure illustrates the reconstruction of a simulated genealogy (**a**) by *SeqTrack* (**b**), by a neighbor-joining phylogenetic tree based on nucleotidic distances, rooted with the most ancient isolate (**c**) and using the BEAST software (**d**). Circled numbers represent isolates. Arrows model real (**a**) or inferred (**b**) transmissions. The number of mutations between isolates is indicated in red at the right of the corresponding arrow (**a**, **b**) and branch (**c**). In **d**, branch lengths represent averages of the consensus tree, and are indicated by the red scale. Plain red dots represent hypothetical isolates inferred by phylogenetic reconstruction. A full color version of this figure is available at the *Heredity* Journal online

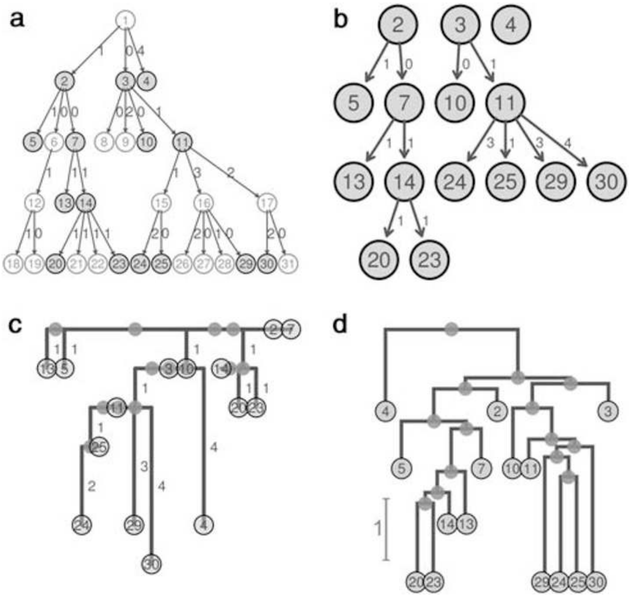

Figure 3 Reconstruction of a sampled simulated genealogy by three different methods. This figure illustrates the reconstruction of a simulated genealogy from a sample of isolates, using the same representations as in Figure 2. (**a**) True genealogy, with sampled isolates indicated in darker tones. (**b**) *SeqTrack* results. (**c**) Results of a neighbor-joining phylogenetic tree based on nucleotidic distances, rooted with one of the most ancient isolates (2). (**d**) Results obtained with the BEAST software.

by simulating genealogies of haplotypes evolving stochastically in time and space. A detailed description of these simulations, along with scripts allowing to reproduce them, are provided in Supplementary Information. Two types of simulations, differing in the process governing the spatial spread of the isolates, were performed. The first scheme allowed isolates to disperse on a 5×5 spatial grid according to a random Poisson process, with identical probabilities to move in all directions (Supplementary Figure S1). This scheme is later referred to as 'homogeneous diffusion'. The second type of simulations used stochastic dispersal based on predefined connectivity between locations, in which we specified the probabilities of migration between every pair of locations. This allowed us to recreate a spatial dynamic of 'sources' and 'sinks' with some locations seeding many other locations, including distant ones, and other locations attracting migration but hardly seeding other places (Supplementary Figure S2). This type of simulation is later referred to as 'heterogeneous dispersal'. For each simulation scheme (homogeneous and heterogeneous dispersal), 10 genealogies of isolates simulated over 20 generations were obtained. Ten random samples of 800 isolates were then drawn from the genealogies, yielding 200 data sets that were analyzed. Simulations and analyses were performed on the computer cluster of the Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules (CCIN2P3; http://cc.in2p3.fr/).

Pandemic A/H1N1 influenza data

Genetic data were used to infer the spatiotemporal spread of the early stages of the swine-origin influenza A/H1N1 pandemic. We analyzed all isolates typed for both hemagglutinin (HA) and neuraminidase (NA) genes available from GenBank (Benson *et al*., 2007; http://www.ncbi.nlm.nih.gov/Genbank/index.html) as of 20 August 2009. Genetic sequences and their annotation (including sampling dates and locations) were retrieved and processed using *ad hoc* scripts in the R language. We only retained isolates for which collection date and location were available. Influenza viruses are routinely amplified by serial passaging through cell cultures or chicken eggs before sequencing. Duplicate sequences have been submitted for some isolates that were multiply amplified, sometimes resulting in slightly different sequences for the same isolate. In these cases, significantly shorter sequences were discarded, as well as sequences obtained from amplification in eggs because this method is known to induce additional 'artificial' mutations (Bush *et al*., 2000).

The alignment of all retained sequences was realized using ClustalW (Larkin *et al*., 2007) and refined by hand using Jalview (Waterhouse *et al*., 2009). HA and NA segments were concatenated after verifying that analyses run with either segment yielded similar results (results not shown). The final alignment contained 433 sequences of 3025 nucleotides, with collection dates ranging from 30 March to 12 July 2009. Accession numbers, alignments, dates and locations of the isolates analyzed are provided in Supplementary Information. Raw pairwise genetic distances (in number of differing nucleotides) were computed using the R package *ape* (Paradis *et al*., 2004). Substitution rates for the HA and NA segments were based on earlier work by Smith *et al*. (2009). Note that the analyses presented in this paper are largely insensitive to the substitution rate, as this information is only used for the resolution of ties between genetically equally likely ancestors sampled at different dates.

Flight data were used to assess the role of air traffic in the dispersal of the pathogen. Passenger flows between countries were compiled from a list of all commercial flights that occurred between 5 May 2008 and 4 April 2009, purchased from OAG (http://www.oag.com/). Numbers of passengers between countries were then correlated to the number of inferred transmissions, and tested using standard Pearson correlation.

Results

Simulation results showed fundamental differences between the *SeqTrack* algorithm and classical phylogenetic reconstruction (Figures 2 and 3). In two randomly chosen examples, *SeqTrack* reconstructed very satisfyingly the transmission tree of the isolates (Figures 2b and 3b). When all isolates were sampled, the method perfectly identified the transmission tree (Figure 2b). However, results remained very satisfying in the incompletely sampled genealogy (Figure 3b), in which *SeqTrack* correctly inferred the most recent sampled ancestors, and therefore true branches of the transmission tree. In contrast, phylogenetic trees could hardly be used to draw inferences about the underlying genealogies (Figures 2c and 3c). Indeed, the phylogenetic trees were not a good match to the transmission tree, and contained some oddities such as several identical internal nodes for a single polytomy (Figures 2c and 3c). Even when considering tips with zero branch length as internal nodes, the resulting inferred ancestries were largely erroneous. Results obtained by BEAST were somewhat more satisfying, as the tips were ordered according to the sampling time. However, BEAST failed to identify any of the direct ancestries present in the simulated data set (Figures 2d and 3d).

Further simulations were used to assess quantitatively the performance of our method at uncovering ancestries from genetic outbreak data. Most satisfyingly, our results indicated that the true ancestral haplotype was successfully retrieved in 93% of cases on average throughout all simulations ($\mathrm{CI}_{99\%} = [92.2\text{–}93.7\%]$; Figure 4). The correct location of the ancestor was also inferred in a majority of cases, although more frequently under source and sink dynamics than under homogeneous dispersal (76 and 71%, respectively; $t = 3.34$, $\mathrm{df} = 115$, $P = 1.1 \times 10^{-3}$; Figure 4).

Having tested the ability of the method at reconstructing transmission trees from outbreak genetic data, we used *SeqTrack* to infer the spatiotemporal dynamics of the early stage of the 2009 swine-origin A/H1N1 influenza pandemic. The data set analyzed consisted of 433 near-complete HA and NA sequences collected between 30 March and 12 July 2009. Although no sequences are available for the first confirmed cases (La Gloria, 15 February; Fraser *et al*., 2009), the time window defined by the collection dates extends from the early stages of the outbreak to the global pandemic, which was officially declared by the WHO on 11 June 2009. In Figure 5, we present snapshots of the spatiotemporal dynamics of the pandemic reconstructed by *SeqTrack*. Overall, most

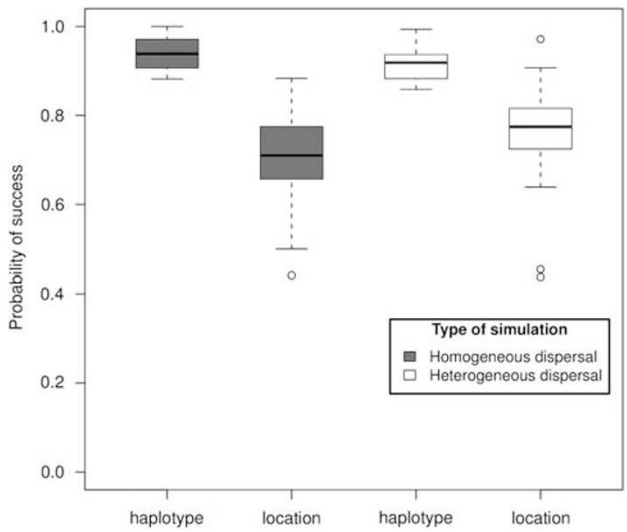

Figure 4 Results of *SeqTrack* performed on simulated data. This figure summarizes the performances of *SeqTrack* in spatially explicit simulations. Box plots represent the proportion of correct inference of the ancestral haplotype and location. Two different spatial models were used to simulate the data: homogeneous (first two boxes, in gray) and heterogeneous (last two boxes, in white).

inferred ancestries displayed very few mutations between ancestor and descendent (Supplementary Figure S3), suggesting that the actual ancestral haplotype had often been sampled, and was correctly identified by the method. The inferred scenario of the early stages of the pandemic can be split into three main acts: (1) initial spread in Mexico and the United States (Figure 5a), (2) sustained transmission within North America and spread to the rest of the world (Figure 5b) and (3) further worldwide spread and secondary outbreaks outside the Americas (Figure 5c). For a day-by-day reconstruction of the pandemic, see Supplementary Movie S1 and Supplementary Information. Details of inferred ancestries are provided in Supplementary Table S2.

The initial stage ending around the 20 April is characterized by numerous transmissions from Mexico to the United States, as well as some local transmissions within Mexico and the United States (Figure 5a). The fact that two Mexican isolates trace their ancestry back to the United States was inescapable as the oldest sequences present in the data set were all sampled in the United States. However, most ancestries of early US isolates point to a Mexican origin (Figure 5a), which gives further support for the A/H1N1 pandemic having originated in Mexico. The second stage (ending around 5 May) is characterized by the spread of the virus within North America, with several local transmissions (dots and sunflowers on Figure 5b and Supplementary Movie S1), indicating the existence of secondary outbreaks and the likely establishment of the pathogen in several cities in the United States and Canada. During the same time period, Mexico and the United States also started seeding the rest of the world (Figure 5b). It is worth noting that the first case of within-country transmission outside the Americas (France on 1 May, Figure 5b, Supplementary Movie S1) is observed remarkably early after the spread of the new influenza strain within the United States. In the third and last stage, we observe a considerable number of transmissions inferred from the Americas to the rest of the world, alongside with transmissions between countries outside the Americas (Figure 5c). Local transmissions with a high statistical support point to multiple secondary outbreaks occurring in several countries, which likely reflect the worldwide establishment of the new influenza strain as a global pandemic.

Flight traffic is widely recognized as an important determinant for the spread of seasonal and pandemic influenza at large geographic scales (Cooper *et al*., 2006; Ferguson *et al*., 2006; Germann *et al*., 2006; Viboud *et al*., 2006; Wu *et al*., 2009). Thus, we evaluated the relationship between air travel passenger-flow and the number of transmissions between countries inferred by *SeqTrack*. Both quantities were fairly correlated ($r = 0.55$; t-test: $t = 5.29$; df $= 63$; $P = 8.3 \times 10^{-7}$), although this correlation was largely driven by the high connectivity of the United States with Mexico and Canada.

Discussion

We have introduced a new methodological approach called *SeqTrack* for the analysis of genetic data sampled during disease outbreaks. Using simulated data, we showed the originality of this method compared to classical phylogenetic reconstruction, and its ability to infer correct genealogies of isolates in densely sampled disease outbreaks. The reconstruction of transmission trees from genetic data has recently received considerable attention (Gonzalez-Candelas *et al*., 2003; Hue *et al*., 2005; Cottam *et al*., 2008a; Cottam *et al*., 2008b; Harris *et al*., 2010). All previous applications we are aware of have been based on the reconstruction of most recent common ancestors. Although this body of work has been fairly successful, the simulations presented in this paper suggest that our method outperforms phylogenetic approaches for the reconstruction of transmission trees in intensively sampled disease outbreaks. Even when taking temporal information into account, phylogenetic reconstruction seems unable to infer filiations between sampled isolates. This statement does not amount to a criticism of phylogenetics but simply reflects the fact that transmission trees and phylogenies are actually different entities. As such, transmission trees are likely to be inferred more successfully by a dedicated approach. Our method is also not meant to supersede phylogenetic approaches but rather to complement them. Although *SeqTrack* is better suited for reconstructing the transmission tree during early outbreaks, it does not include some useful features of modern phylogenetic tools, such as the possibility to infer underlying population dynamics parameters.

We illustrated our method with a reconstruction of a plausible scenario for the spatiotemporal spread of the A/H1N1 influenza strain during the early stages of the 2009 influenza pandemic. This data set might not be ideal for showing the full potential of the new methodology, which may be best suited for the reconstruction of transmission trees in more localized settings. Nonetheless, the results suggest that *SeqTrack* can contribute to a better understanding of the spatiotemporal dynamics of emerging pathogens, even at a worldwide scale. The results are largely in line with the general epidemiological understanding of the spatiotemporal spread of the 2009 A/H1N1 pandemic. We also recover individual

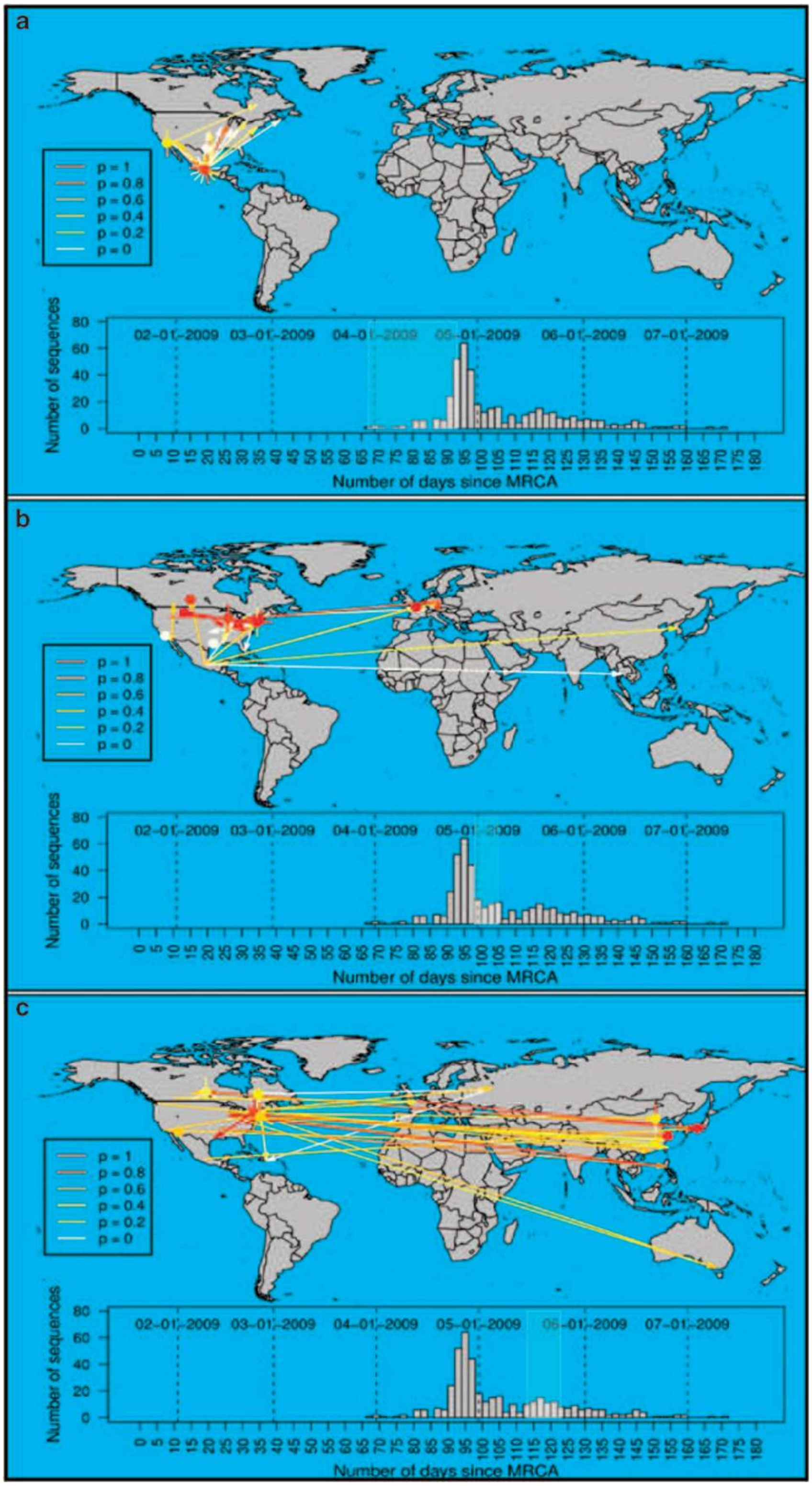

Figure 5 Maps of inferred ancestries of the swine-origin A/H1N1 influenza pandemic, based on 433 HA and NA DNA sequences available on GenBank as of 20 August 2009. Arrows represent inferred ancestries, with colors indicating the genetic likelihood of the individual ancestries (Equation (1)). Local transmissions are shown as dots for single events, and sunflowers for multiple transmissions (in which case one segment represents one transmission). The inset histogram displays the number of isolates analyzed for the corresponding time window. Dates are provided as days from the MRCA, estimated as 21 January 2009 (Fraser *et al*., 2009). The three different panels display snapshots of the three main stages of the spread of the pathogen (corresponding time frame highlighted in the inset): (**a**) transmissions from Mexico to the United States, (**b**) transmissions within the United States, and beginning of worldwide spread and (**c**) further worldwide spread and secondary outbreaks outside the Americas.

ancestries that seem sensible and correlate with human flight data. It is not entirely straightforward to compare our reconstruction with previous phylogenetic work based on sequence data from the 2009 A/H1N1 influenza pandemic (Fraser *et al.*, 2009; Parks *et al.*, 2009; Rambaut and Holmes 2009; Smith *et al.*, 2009; Lemey *et al.*, 2009b). These authors used a smaller set of isolates and did not specifically aim to reconstruct the spatiotemporal dynamics with the exception of Lemey *et al* (2009b), who analyzed 242 isolates sequenced for the HA and NA genes. They also recovered a strong connection between Mexico and Texas and multiple instances of transmissions from the United States and Canada to outside the Americas. In contrast to their results, we infer a larger proportion of direct transmissions from North America to other continents. For instance, although we infer several direct transmissions from North America to China, their reconstruction points to China being only indirectly connected to the United States through either Russia, Europe or South America.

SeqTrack is a very versatile approach that allows for weighting the plausibility of individual ancestries in various ways. In this study, we used a maximum parsimony version of the method because the origin of A/H1N1 influenza virus is sufficiently recent for homoplasies (recombination or back mutations) to be very unlikely, and safely ignored. Nonetheless, *SeqTrack* genealogies could be optimized on other criteria, such as genetic distances or log-likelihoods based on complex models of sequence evolution. Another possible extension of this method relates to the use of heterogeneous genetic information, which occurs when the regions of the genome sequenced vary across isolates. Such heterogeneity is likely to occur as soon as the sequencing effort is undertaken independently by a large number of labs, as was the case during the early stage of the 2009 A/H1N1 influenza pandemic. As a result, sequence data of the new influenza strain consisted of essentially random combinations of segments, ranging from small fragments of specific genes to whole genomes. In such cases, the analysis of homologous sequences results in a possibly considerable loss of information. This issue could be overcome by averaging genetic differentiation across available segments, using appropriate weights to account for different substitution rates and segment lengths. However, including a large fraction of isolates with heterogeneous sequencing coverage would require accurate estimates of substitution rates, which can be difficult to obtain for emerging pathogens.

Some other developments would require more effort. In particular, it would be desirable to consider the various parameters of the model from a probabilistic perspective. For instance, the date of probable transmission could be better captured by a distribution of the time during which a specific sequence remains unaltered rather than a fixed collection date for the isolate. Moreover, a probabilistic framework would also allow incorporation of information besides genetic sequences and collection dates, such as prevalence data or spatial connectivity between locations. A final possible extension of *SeqTrack* relates to the inference of unsampled ancestral isolates. Although phylogenetic methods assume that no isolate in the sample is (directly or indirectly) ancestral to another one, our method considers that ancestries can be inferred between the sampled isolates. In practice, however, the ancestral population of a given isolate may not have been sampled. In such cases, *SeqTrack* is likely to identify the most recent sampled ancestor of the isolate as a substitute for the unsampled direct ancestor (Figures 1d, 2b, and 3b). In other words, the algorithm is likely to retrieve an actual branch of the transmission tree by overlooking unsampled intermediate nodes. The possibility of inferring unobserved intermediate nodes based on a probabilistic model for sequence evolution and spatial dispersal (Lemey *et al.*, 2009a) would represent an additional significant improvement in the present approach.

Understanding the underlying factors behind the spread of diseases is one of the core objectives of infectious disease epidemiology. By reconstructing the transmission tree of an outbreak, our method creates new opportunities for testing how various features of the connectivity between hosts shape transmission patterns. For instance, despite limited genetic information available for the early stage of the swine-origin A/H1N1 influenza pandemic, we observed a positive association between flight traffic and the routes of transmissions inferred by *SeqTrack*, in line with previous works on both seasonal and pandemic influenza (Brownstein *et al.*, 2006; Cooper *et al.*, 2006; Ferguson *et al.*, 2006; Germann *et al.*, 2006; Viboud *et al.*, 2006; Fraser *et al.*, 2009). The versatility of the presented methodology makes it applicable over a wide range of scales, from localized outbreaks of a nosocomial infection within a hospital to the global spread of a newly emerged pathogen. In not requiring extensive genetic polymorphism, the method should prove effective in bacteria and fungi, in addition to viruses. We hope that the performance of the method, together with its wide applicability, flexibility and computational speed will convince infectious disease epidemiologists to adopt it as an integral part of the toolkit for infectious disease outbreak analysis.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We are most grateful to all our colleagues who sequenced isolates from the 2009 influenza A/H1N1 pandemic and made them publicly available. We feel particularly indebted toward Nancy Cox, Mark Greenway, Naomi Komadina, Andreas Nitsche, John Pasick, Tomas Pumarola and Pilaipan Puthavathana for their willingness to provide additional information on sequences they submitted on GenBank. We address many thanks to Vicente Acuña for providing insights about graph theory, and to Vincent Daubin for interesting discussions. We are thankful to Simon Penel and the Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules (CCIN2P3, http://cc.in2p3.fr/) for providing access to the computer resources used to run and analyze simulations. We finally thank three anonymous reviewers whose comments allowed us to greatly improve the paper.

References

Albrich WC, Harbarth S (2008). Health-care workers: source, vector, or victim of MRSA? *Lancet* **8**: 289–301.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2007). GenBank. *Nucleic Acids Res* **35**: D5–D12.

Brownstein JS, Wolfe CJ, Mandl KD (2006). Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLoS Med* **3**: e401.

Bush RM, Smith CB, Cox NJ, Fitch WM (2000). Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc Natl Acad Sci USA* **97**: 6974–6980.

Chu YJ, Liu TH (1965). On the shortest arborescence of a directed graph. *Science Sinica* **14**: 1396–1400.

Cooper BS, Pitman RJ, Edmunds WJ, Gay NJ (2006). Delaying the international spread of pandemic influenza. *PLoS Med* **3**: e212.

Cottam EM, Thebaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ *et al.* (2008a). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc R Soc London B Biol Sci* **275**: 887–895.

Cottam EM, Wadsworth J, Shaw AE, Rowlands RJ, Goatley L, Maan S *et al.* (2008b). Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. *PLos Pathogens* **4**: e1000050.

Drummond AJ, Rambaut A (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**: 214.

Edmonds J (1967). Optimum branchings. *J Res Nat Bur Standards* **9**: 233–240.

Ferguson NM, Cummings DAT, Fraser C, Cajka JC, Cooley PC, Burke DS (2006). Strategies for mitigating an influenza pandemic. *Nature* **442**: 448–452.

Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD *et al.* (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* **324**: 1557–1561.

Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A *et al.* (2009). Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **325**: 197–201.

Germann TC, Kadau K, Longini IM, Macken CA (2006). Mitigation strategies for pandemic influenza in the United States. *Proc Natl Acad Sci USA* **103**: 5935–5940.

Gonzalez-Candelas F, Bracho MA, Moya AS (2003). Molecular epidemiology and forensic genetics: application to a hepatitis C virus transmission event at a hemodialysis unit. *J Infect Dis* **187**: 352–358.

Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA *et al.* (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**: 327–332.

Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N *et al.* (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**: 469–474.

Hue S, Pillay D, Clewley JP, Pybus OG (2005). Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc Natl Acad Sci USA* **102**: 4425–4429.

Jombart T (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–1405.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Lemey P, Suchard M, Rambaut A (2009a). Reconstructing the initial global spread of a human influenza pandemic: a Bayesian spatial-temporal model for the global spread of H1N1pdm. *PLoS Curr Influenza*: RRN1031.

Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009b). Bayesian phylogeography finds its roots. *PLoS Comput Biol* **5**: e1000520.

Paradis E, Claude J, Strimmer K (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.

Parks D, Macdonald N, Beiko R (2009). Tracking the evolution and geographic spread of Influenza A. *PLoS Curr Influenza*: RRN1014.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R foundation for Statistical Computing: Vienna, Austria.

Rambaut A, Holmes E (2009). The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Curr Influenza*: RRN1003.

Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**: 615–619.

Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V *et al* (2008). The global circulation of seasonal influenza A (H3N2) viruses. *Science* **320**: 340–346.

Sloan CD, Duell EJ, Shi X, Irwin R, Andrew AS, Williams SM *et al.* (2009). Ecogeographic genetic epidemiology. *Genet Epidemiol* **33**: 281–289.

Smith G, Vijaykrishna D, Bahl J, Lycett S, Worobey M, Pybus O *et al.* (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **325**: 197–201.

Templeton AR (1998). Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol Ecol* **7**: 381–397.

Viboud C, Bjornstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**: 447–451.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009). Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191.

Wu JT, Leung GM, Lipsitch M, Cooper BS, Riley S (2009). Hedging against antiviral resistance during the next influenza pandemic using small stockpiles of an alternative chemotherapy. *PLoS Med* **6**: e1000085.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)