*Bioinformatics*, 36(3), 2020, 945–947
doi: 10.1093/bioinformatics/btz646
Advance Access Publication Date: 16 August 2019
Applications Note

Phylogenetics

# StrainHub: a phylogenetic tool to construct pathogen transmission networks

**Adriano de Bernardi Schneider [1,*], Colby T. Ford[2], Reilly Hostager[1], John Williams[2], Michael Cioce[2], Ümit V. Çatalyürek[3], Joel O. Wertheim[1] and Daniel Janies[2,*]**

[1]Department of Medicine, University of California, San Diego, San Diego, CA 92103, USA, [2]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA and [3]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on May 7, 2019; revised on August 6, 2019; editorial decision on August 8, 2019; accepted on August 14, 2019

## Abstract

**Summary:** In exploring the epidemiology of infectious diseases, networks have been used to reconstruct contacts among individuals and/or populations. Summarizing networks using pathogen metadata (e.g. host species and place of isolation) and a phylogenetic tree is a nascent, alternative approach. In this paper, we introduce a tool for reconstructing transmission networks in arbitrary space from phylogenetic information and metadata. Our goals are to provide a means of deriving new insights and infection control strategies based on the dynamics of the pathogen lineages derived from networks and centrality metrics. We created a web-based application, called StrainHub, in which a user can input a phylogenetic tree based on genetic or other data along with characters derived from metadata using their preferred tree search method. StrainHub generates a transmission network based on character state changes in metadata, such as place or source of isolation, mapped on the phylogenetic tree. The user has the option to calculate centrality metrics on the nodes including betweenness, closeness, degree and a new metric, the source/hub ratio. The outputs include the network with values for metrics on its nodes and the tree with characters reconstructed. All of these results can be exported for further analysis.

**Availability and implementation:** strainhub.io and https://github.com/abschneider/StrainHub.

**Contact:** adeberna@ucsd.edu or djanies@uncc.edu

## 1 Introduction

New technologies can shape responses to outbreaks of rapidly evolving infectious diseases. High-throughput genetic sequencing has allowed rapid characterization of disease outbreaks such as Ebola, Yellow Fever and Zika viruses (Faria *et al.*, 2016, 2018; Quick *et al.*, 2016). Multiple advancements in interpreting genomic data related to pathogen outbreaks have been recently developed: SCOTTI (De Maio *et al.*, 2016), Phyloscanner (Wymant *et al.*, 2018), QUENTIN (Skums *et al.*, 2018), BadTrIP (De Maio *et al.*, 2018), Outbreaker (Jombart *et al.*, 2014) and Outbreaker2 (Campbell *et al.*, 2018). Each of these tools offer distinct advantages when analyzing datasets, although only a few of them include network visualization and use centrality metrics in order to infer importance of nodes (Skums *et al.*, 2018).

Here, we introduce a novel tool, StrainHub, which summarizes the transition between states of metadata rather than among individuals, providing an overview of the pathogen transmission paths through geography or populations. Genomic data and associated metadata from pathogens are observations of the biology underlying a disease. The combination of these data collected from related pathogen isolates allow researchers to understand disease transmission patterns as a function of the pathogens' evolutionary history. StrainHub can leverage these data and make phylogenetic transmission graphs accessible to public health scientists. StrainHub is provided as both a standalone package in GitHub and a web-based interface.

## 2 Materials and methods

We built the StrainHub application with multiple R packages wrapped with Shiny. In order to build a transmission network, the user provides a phylogenetic tree and associated metadata for all terminal taxa.

### 2.1 Ancestry reconstruction

When parsimony is selected, the ancestral state reconstruction step in StrainHub uses the R function 'asr_max_parsimony' from the

package Castor (Louca and Doebeli, 2018). This function performs an ancestral state reconstruction for discrete traits derived from metadata using the parsimony algorithm described by Sankoff (1975). Next, based on the results of ancestral state reconstruction, StrainHub outputs a relationship list of the metadata elements as source and destination.

Users have the option to run phylogeography in BEAST (Drummond *et al.*, 2012) and visualize the transmission network based on the tree edges and trait probability for each node. The relationship list is extracted from the tree nodes of a BEAST phylogeography file using the R package treeio (Yu *et al.*, 2017) to build a directional network and calculate the metrics as described below. This option is used in lieu of the parsimony ancestral state reconstruction step.

### 2.2 Tree and transmission network visualization

We implemented a tree visualization tab using ggtree, ggplot and plotly to display the metadata mapped to each taxon within the tree (Sievert, 2017; Yu *et al.*, 2017). We used the R packages igraph and visNetwork to build the transmission networks. igraph provides functions for generating the backbone of the transmission network and calculates the centrality metrics (Csardi and Nepusz, 2006). visNetwork provides the R interface to the 'vis.js' javascript charting library, allowing an interactive visualization of the transmission network (Almende *et al.*, 2016). StrainHub uses the edge list created on the previous step as a source and destination list which is transformed into a data frame and plotted as a network. The nodes of the transmission network created are not the individual pathogen sequences but the relationship of the ancestral and descendant states of the pathogen sequences (e.g. changes in geography, host shifts, changes among risk factors, or any set of discrete state the user can encode).

### 2.3 Transmission network metrics

In StrainHub we provide multiple centrality measurements to determine the relative importance of nodes within a network with respect to the dynamics of pathogen lineages (Hoffmann *et al.*, 2016; Janies *et al.*, 2015). In this application, we implemented three centrality metrics: betweenness, closeness and degree (Table 1). Betweenness measures the number of shortest paths between two other nodes that pass through the node of interest, normalized by the number of all pairs of nodes within the network. The higher betweenness score of a node reflects the importance of that node as a hub in the network for traffic of the pathogen. Closeness evaluates a node based on the relative sum of the lengths of all the shortest paths from that node to all other nodes within network. A higher closeness centrality value is associated with how close a node is as a direct point of transmission to other nodes. Degree is the number of edges incident upon a given node. The higher the number of edges connected to a given node indicates the higher importance of that node in terms of being involved in metadata state transitions irrespective of directionality (Freeman, 1978).

We further divided degree centrality into 'indegree' and 'outdegree'. With these values, we created the 'Source Hub Ratio' (SHR) metric. SHR is obtained by calculating the ratio of all transitions from the node (outdegree) over all transitions from and to the node (indegree + outdegree), resulting in a value ranging from 0 to 1. Nodes with SHR values close to 0.5 indicate that the node is a hub, SHR close to 1, indicate it is a source, and SHR close to 0 indicates that it is a sink for the pathogen. The SHR metric reflects the importance of a node within the network as the hub, source or sink of the pathogen, ignoring centrality of the node within network (de Bernardi Schneider, 2018). Although SHR alone does not define which node is the most important within the network for the spread of the pathogen, it creates an indicator for the behavior of the nodes, which can be further investigated for importance by the association of SHR with other metrics of interested for the user. The availability of multiple metrics for any type of discrete metadata allows the user to have flexibility in assessing hypotheses in different contexts for the spread of infectious diseases.

### 2.4 Shiny

We used the Shiny framework to provide StrainHub a flexible, web-based interface. Shiny allows for the generation of interactive applications that can be hosted as standalone web-applications that are locally installed or that are served over a network (Chang *et al.*, 2018). In either case, the user interacts with StrainHub via a web browser of choice.

## 3 Implementation

StrainHub accepts phylogenetic trees in Newick format and metadata in comma-separated value (CSV) format to run the ancestry reconstruction step. The orthography and content of the taxon names must match. Alternatively, users can run phylogeography in BEAST and use the maximum clade credibility tree in Nexus format as input to visualize the transmission network.

Three outputs are generated during the transmission network analysis: a transmission network plot, a tree plot with the character of interest mapped to the tips of the tree, and a table with all centrality metrics computed. The transmission network graph and the tree files are user-interactive, and a snapshot for further analysis or publication can be exported in PNG format (e.g. Fig. 1). The centrality metrics table can be exported as a CSV file.

The web-application can be accessed at strainhub.io and the source code is available at https://github.com/abschneider/StrainHub under the GNU General Public License (GPL) v3.0.

## 4 Conclusion

We created a visual analytic tool that enables the user to interpret and communicate phylogenetic data on the dynamics of the spread

**Table 1.** Summary of Centrality Metrics applied to measure transmission networks on StrainHub

| Metric | Formula | Meaning |
|---|---|---|
| Betweenness Centrality | $C_B(i) = \sum \frac{\sigma_{jk}(i)}{\sigma_{jk}}$ Where $\sigma_{jk}$ = number of shortest paths. | How frequently a node acts as the intermediary node connecting two other nodes on a shortest path. |
| Closeness Centrality | $C_C(i) = \frac{1}{\sum d(i,k)}$ Where $d(i,k)$ = distance between nodes. | How close a node is as the starting point to other nodes. |
| Degree Centrality | $C_D(i) = \frac{\sum a(i,k)}{n-1}$ Where $a(i,k) = 1$ if and only if $i$ and $k$ are connected, otherwise = 0. | How many times edges originate or end on that node (Outdegree / Indegree). |
| Source/Hub Ratio | $R_{SH}(i) = \frac{\sum S(i)}{\sum H(i)}$ Where $S(i)$ = directed edges from node $i$ and $H(i)$ = directed edges to and from node $i$. | Number of transitions originated on node over the total number of transitions related to that node. A node scoring SHR close to 1 indicates a source, SHR close to 0.5 a hub and SHR close to 0 a sink for the pathogen. |

*Note*: Formula definitions: $i$, $j$ and $k$ = nodes $i$, $j$ and $k$, respectively; $d$ = distance; $\sigma$ = shortest path; $a$ = edge.

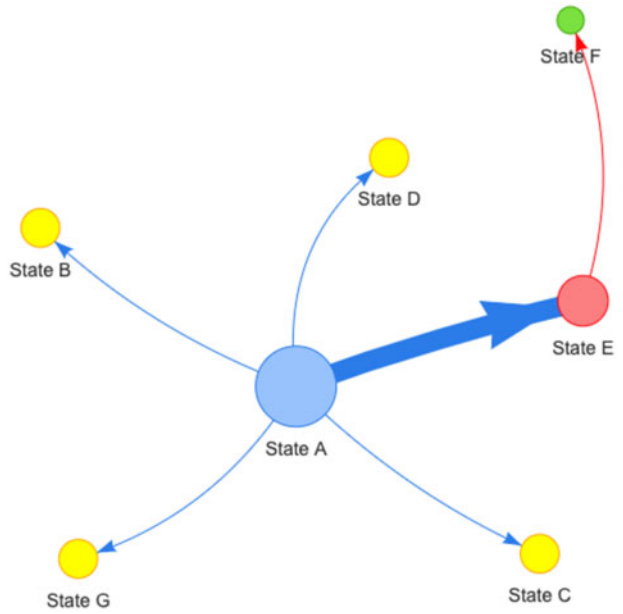


**Fig. 1.** Example of a generic pathogen transmission network based on geographic location ("states" as in character and/or geographic). Sizes of nodes are scaled by the metric selected by the user (i.e. betweenness, closeness, degree or source hub ratio). The arrows reflect directionality of transition between states. The thickness of the lines and arrows represents the frequency of transitions (thicker arrows reflect more transitions). In the current implementation, colors of nodes are randomly assigned to a metric score and associated with nodes

pathogens over geography or various hosts. Moreover, the metadata format can be used for any type of categorical data that user can encode (e.g. food sources, or risk factor, or phenotype). The user use should assume the results are only as strong as the underlying phylogenetic data (i.e. sampling across metadata states and taxa, adequate branch lengths and nodal support to reconstruct an ancestor descendent change). With solid datasets, the data in the transmission networks, such as identification of the hubs and sources for the spread of pathogens, will assist health authorities to allocating resources to parts of the network that will do the most to disrupt the spread of the pathogen.

## Acknowledgements

We thank Andrew Frick, Dr Nídia Trovão and Dr Tetyana Vasylyeva for suggestions on methodology during the development of this app.

## Funding

This work was supported in part by the National Institutes of Health (NIH) National Institute of Allergy and Infectious Diseases (grant numbers K01AI110181 and AI135992) to JOW. We acknowledge the support of the Department of Bioinformatics and Genomics, the College of Computing and Informatics and the Graduate School of the University of North Carolina at Charlotte. This effort was funded in part by the Defense Threat Reduction Agency under contract HDTRA1-16-C-0010 to UC and DJ.

*Conflict of Interest*: none declared.

## References

Almende,B.V. *et al.* (2016) *visNetwork: Network Visualization using 'vis.js' Library*. R package version.

Campbell,F. *et al.* (2018) outbreaker2: a modular platform for outbreak reconstruction. *BMC Bioinformatics*, **19**, 363.

Chang,W. *et al.* (2018) *shiny: Web Application Framework for R*. R package version 1.1.0.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal Complex Syst.*, **1695**, 1–9.

de Bernardi Schneider,A. (2018) Arboviruses: the hidden path of an imminent threat. PhD thesis, The University of North Carolina at Charlotte.

De Maio,N. *et al.* (2016) SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput. Biol.*, **12**, e1005130.

De Maio,N. *et al.* (2018) Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Comput. Biol.*, **14**, e1006117.

Drummond,A.J. *et al.* (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29**, 1969–1973.

Faria,N.R. *et al.* (2016) Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.*, **8**, 97.

Faria,N.R. *et al.* (2018) Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*, **361**, 894–899.

Freeman,L.C. (1978) Centrality in social networks conceptual clarification. *Soc. Networks*, **1**, 215–239.

Hoffmann,M. *et al.* (2016) Tracing origins of the Salmonella Bareilly strain causing a food-borne outbreak in the United States. *J. Infect. Dis.*, **213**, 502–508.

Janies,D.A. *et al.* (2015) Phylogenetic visualization of the spread of H7 influenza A viruses. *Cladistics*, **31**, 679–691.

Jombart,T. *et al.* (2014) Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.*, **10**, e1003457.

Louca,S. and Doebeli,M. (2018) Efficient comparative phylogenetics on large trees. *Bioinformatics*, **34**, 1053–1055.

Quick,J. *et al.* (2016) Real-time, portable genome sequencing for Ebola surveillance. *Nature*, **530**, 228.

Sankoff,D. (1975) Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, **28**, 35–42.

Sievert,C. (2017) *plotly: Create Interactive Web Graphics via 'plotly. js'*. R package version 4.7. 1.

Skums,P. *et al.* (2018) QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, **34**, 163–170.

Wymant,C. *et al.* (2018) PHYLOSCANNER: inferring transmission from within-and between-host pathogen genetic diversity. *Mol. Biol. Evol.*, **35**, 719–733.

Yu,G. *et al.* (2017) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*, **8**, 28–36.