Shared genomic variants: identification of transmission routes using pathogen deep sequence data

Colin J. Worby[1]*, Marc Lipsitch[1], William P. Hanage[1]

[1]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard TH Chan School of Public Health, 655 Huntington Ave, Boston, MA 02115

*Corresponding author
cworby@hsph.harvard.edu

Abstract

While identifying routes of transmission during an infectious disease outbreak was traditionally conducted through exhaustive contact tracing efforts, the increasing availability of pathogen sequencing has provided a new resource with which one can identify plausible routes of infection. However, while transmission clusters can be identified using single genome sequences, individual transmission routes remain relatively uncertain. Deep sequence data may provide additional information where single genomes lack sufficient resolution – presence of shared minor variants can suggest epidemiological linkage when observed between multiple hosts. In this study we formalize shared variant methods to reconstruct the transmission tree in an outbreak, and using simulated outbreak data, we quantify the improved accuracy when compared with analogous single genome approaches. Furthermore we propose a hybrid approach, drawing information from both deep sequence and single genome data. Our simulation studies demonstrate the superior performance of transmission tree identification methods using shared variants in most settings. Application of these methods to deep sequence data collected during the 2014 Sierra Leone Ebola epidemic demonstrates the ability to identify plausible transmission routes without any additional data. The methods we describe should become a common step in outbreak investigations and epidemiological analyses once the collection of deep sequence data becomes increasingly widespread.

Sequencing pathogen samples during a communicable disease outbreak is becoming an increasingly common procedure in epidemiological investigations. Identifying who infected whom sheds considerable light on transmission patterns, high-risk settings and subpopulations, and infection control effectiveness. Genomic data can shed new light on transmission dynamics, and can be used to identify clusters of individuals likely to be linked by direct transmission. However, identification of individual sources of infection typically remains uncertain. In this study, we investigate the potential of deep sequence data to provide greater resolution on transmission routes. We describe easily implemented methods to use such data, and demonstrate the remarkably improved performance when reconstructing transmission trees. Furthermore, we apply our methods to data collected during the 2014 Ebola outbreak in Sierra Leone, identifying several routes of transmission. Our study highlights the power of pathogen deep sequence data as a component of outbreak investigation and epidemiological analyses.

Introduction

Genomic data offer new insights into epidemiological and evolutionary dynamics, and sequencing pathogen samples is becoming increasingly widespread. Pathogen genomic data allows one to link the evolutionary relatedness of collected isolates, which in turn sheds light on the potential relationship between the hosts from whom they were collected. As such, inference of transmission trees using genomic data is an increasingly well-studied field (1-7). While low-resolution pathogen typing has been used for some time to discriminate between independent outbreaks (8-10), whole genome sequencing provides additional resolution with which genetic distance between identical phenotypes may be ascertained (11-13). This too, however, has limits. Studies have shown that while transmission clusters may be identified with genomic data, individual-level transmission routes can rarely be identified with a great degree of certainty (2, 3). Characterizing an infected host by a single pathogen genome (isolation and purification of a single genotype for bacteria, or using the consensus sequence for viral pathogens) is common practice, yet neglects within-host diversity. The variation in sampled genetic distances can be large relative to the expected number of mutations between hosts, rendering the number of SNPs a rather crude measure of relatedness on an individual level (14). As such, particularly for rapidly evolving pathogens, or those whose mode of transmission is associated with a large and potentially diverse inoculum (‘transmission bottleneck’), single genome sampling can cause hosts to appear misleadingly similar or dissimilar.

Deep sequencing can potentially provide new insights into within-host diversity. Currently, sequencing a mixed population sample to sufficient depth to identify minor variants has mostly been limited to viral samples. Recent studies have demonstrated the utility of such data in exploring evolutionary dynamics of communicable pathogens during an ongoing outbreak (15-17). While the consensus sequences may appear identical for two samples, comparing minor variants can offer additional resolution. For instance, the presence of one or more shared variants could be considered as strong evidence for direct transmission, particularly if the variant is not observed in any other host. This naturally relies on the possibility that a pathogen population of size greater than one survives the transmission bottleneck; otherwise each infection must initially be monoclonal. While estimating transmission bottleneck size is challenging, there is some evidence for larger bottlenecks occurring in influenza (18, 19), Ebola (20), as well as speculation for much variation in bottleneck size for bacterial pathogens (21).

In this study, we investigated the predictive power of shared variants for identifying transmission routes. We compared several simple approaches to reconstruct simulated transmission trees in order to quantify the difference in performance of those based on single genome samples vs. those based on deep sequence data. In addition, we describe hybrid methods which draw power from shared variant identification as well as genetic distance metrics. We then demonstrated an application of the shared variant methods on deep sequence data collected during the Ebola outbreak in West Africa in 2014.

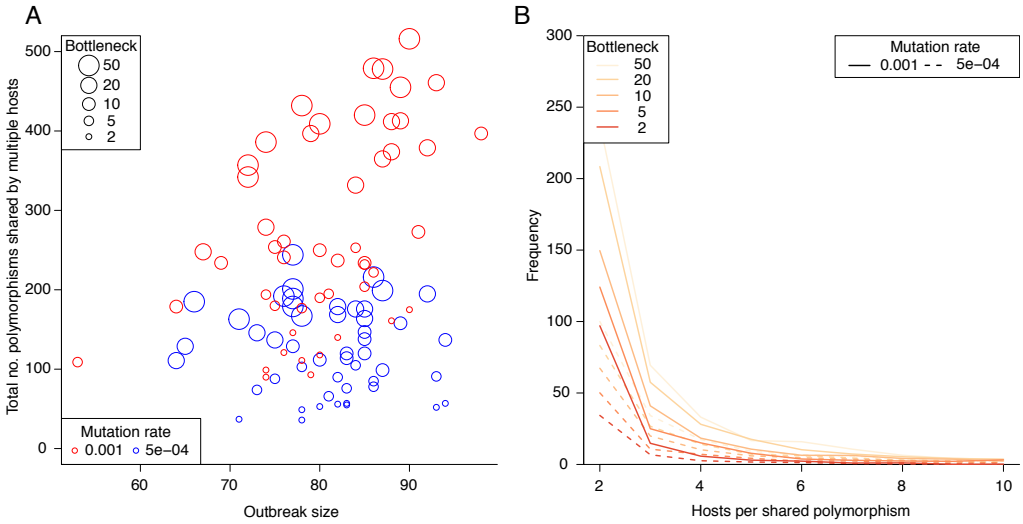

Figure 1. Summary of genetic variant frequency across the simulated outbreaks. (A) Total number of shared variants across the simulated outbreak. Bottleneck size is illustrated by circle size. (B) Distribution of shared variant group size for different bottlenecks and mutations rates.

Results

Simulation Studies

We simulated a range of outbreaks with final size at least 50, allowing both the transmission bottleneck size and mutation rate to vary. As expected, shared variants were observed with increasing frequency when mutation rates and bottleneck sizes were larger (Figure 1A). The vast majority of shared variants were observed in exactly two individuals, with a rapidly declining frequency for larger group sizes (Figure 1B). For each simulation, we constructed a weighted transmission tree according to the six methods outlined in the Methods section. An example simulation and the reconstructions are shown in figure 2. We plotted the ROC curve for each methodology (Figure S1) to determine the AUC statistic. For small bottlenecks, variant-based methods provide a poor tree reconstruction by the AUC measure (Figure 3B); values below 0.5 indicate a worse performance than random selection, an inevitability when only a small proportion of nodes are assigned sources. A tight bottleneck leads to little diversity persisting across transmission events, and as such, shared variants are rarely observed, leading to a sparsity of informed links across the network. However, the links which are made are typically more reliable than those estimated by minimum genetic distance, with an average path length of less than 2 (Figure 3A). Larger bottlenecks lead to rapidly improving AUC statistics for the variant-based approaches, although the maximum variant approach declines slightly for bottleneck size greater than 10. In contrast, distance-based approaches typically decline in accuracy as the bottleneck size increases.

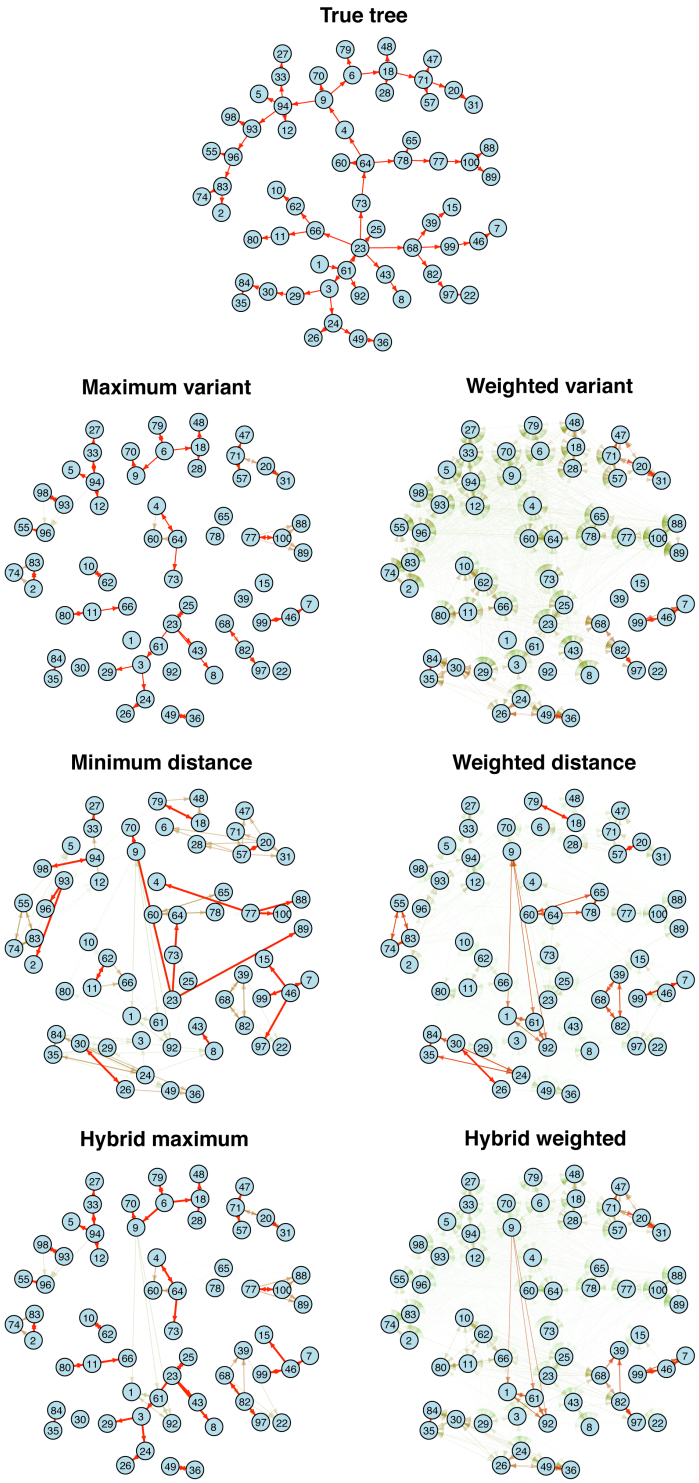

Figure 2. Simulated and reconstructed transmission trees. The simulated tree (top) was generated with bottleneck size 10 and mutation rate 0.001. Trees were reconstructed according to the approaches described in *Methods*. Edges are colored and weighted according to the weight attributed to that potential transmission route, red edges associated with a higher weight than green. Networks were plotted with the igraph package in R.

The hybrid methods draw additional information from the genetic distance information when no shared variants are found for a given host. This offers a considerable improvement in performance for outbreaks with smaller bottleneck sizes (Figure 3B), which are the situations where distance information is most reliable.

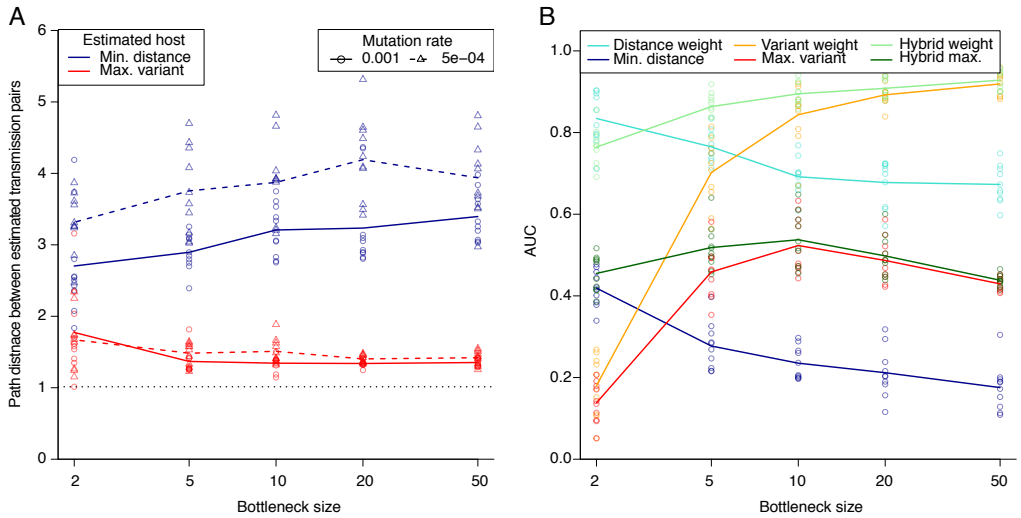

Figure 3. Transmission tree reconstruction accuracy. (A) The true path distance between estimated transmission pairs gives insight into the extent to which transmission links are misspecified. A perfect reconstruction would have mean path length 1. Maximum variant path lengths are averaged over identified transmission pairs, that is, excluding hosts with no shared variants. (B) The area under the ROC (AUC) metric provides an overall measure of network accuracy. Results for a mutation rate of 0.001 are shown here.

For each simulation, we calculated the mean weight attributed to the true source of each host. The distance and variant approaches perform similarly for bottleneck size 2, but are outperformed by the two hybrid approaches (Figure 4). For larger bottleneck sizes, the maximum variant approach (together with the hybrid maximum) is the best method by a large margin according to this measure. The distance and variant weight methods perform increasingly poorly as the bottleneck size increases, reflecting the greater likelihood of genotypes persisting across several bottlenecks, increasing uncertainty.

We found that lower mutation rates were typically associated with poorer tree reconstruction, although the relative performance of the methods remained broadly similar (Figures 3A, S2 and S3). Lower mutation rates lead to fewer shared variants, and as such, the distance-based approaches may outperform variant-based approaches for larger bottlenecks. Nevertheless, the hybrid approach consistently offers the most successful reconstruction approach over a range of parameters.

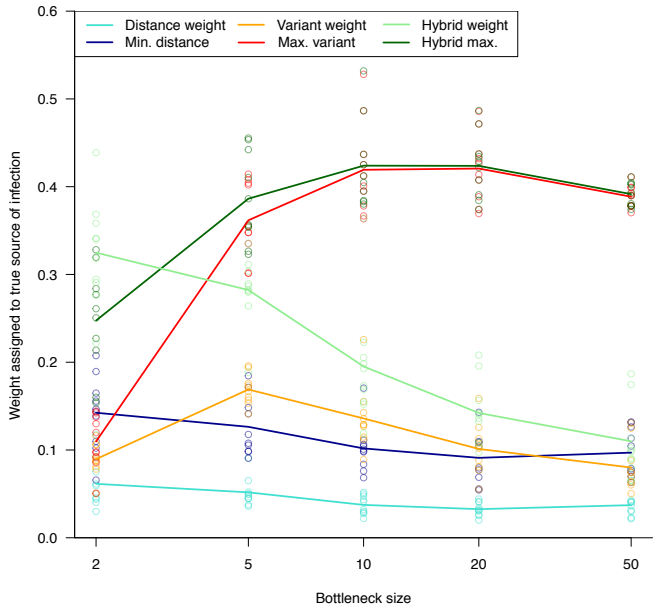

Figure 4. True edge weighting. The mean weight attributed to each true transmission link for each tree reconstruction method, under a range of scenarios and methodologies. Results are shown for a mutation rate of 0.001.

Ebola virus data

We next examined the Sierra Leone Ebola datasets, attempting to identify transmission links between hosts. In order to reduce the risk of counting variant calling errors as true intra-host variants, we identified only variants in which the minor frequency was at least 5% (routes estimated under a 1% threshold are shown in Figure S4). Figure 5 shows the transmission trees reconstructed for each dataset. In the first dataset (Figure 5A, (16)) a total of 19 out of 78 hosts were found to have shared a variant with at least one other individual. Four pairs of patients shared more than one variant (three pairs with two shared variants, and one pair with four), while one additional pair shared one unique variant. Each of these pairs originated from the same geographic location, and was temporally clustered (three of these links were sampled two or fewer days apart, while the remaining two were sampled 12 and 22 days apart, still plausible given the serial interval estimates of 15.3 ± 9.3 days (22). One variant was shared by 11 hosts. This might be explained by alternative hypotheses; (i) a well-sampled transmission cluster with large transmission bottleneck size, (ii) repeated emergence of identical mutation. The group shows high geographic and temporal clustering (15), suggesting that a transmission cluster may be plausible. Non-random mixing, or presence of a superspreader (for instance, due to several persons caring for a patient, or washing and burying a deceased patient)

could make the transmission scenario much more likely. Furthermore, non-random sampling could skew the observed distribution of group size for shared variants.

A total of 26 hosts out of 150 (for which replicate sequencing and variant calling was performed) in the second dataset shared a variant at with at least one other host (Figure 5B, (16)). No pair of hosts shared more than one variant, and there were five pairs of individuals sharing a unique variant. One variant was shared by 13 hosts at the 5% detection threshold. Unlike the previous dataset, these hosts were not geographically or temporally clustered, coming from different villages and spanning several weeks. This suggests that a transmission cluster is less likely than in the previous example, and potentially reflects homoplasy. Four of the five 'unambiguous' transmission routes joined patients from the same geographic location.

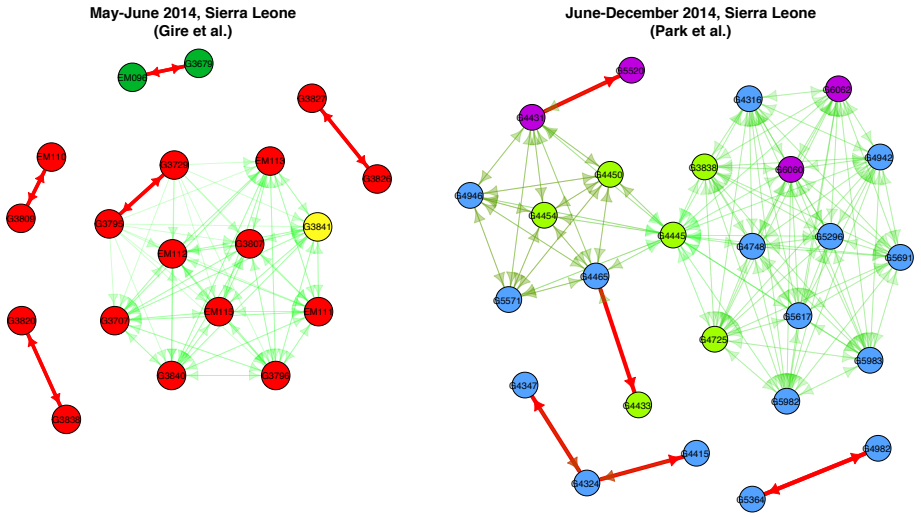

Figure 5. Estimated Ebola transmission routes. Transmission links between sampled hosts in the 2014 Ebola outbreak under the maximum variant approach. Colors denote chiefdoms to which hosts belong, while the color and thickness of the arrows denote the relative weight attributed to each potential transmission event. Variant detection threshold 5%.

Discussion

Shared variant detection offers a powerful insight into unobserved transmission dynamics, and can improve the resolution of reconstructed transmission trees considerably. We have described some simple methodology for reconstructing transmission trees using deep sequence data. These methods typically outperform genetic distance comparison methods, frequently used to identify potential transmission events [cite

MRSA etc.]. While more formal approaches may utilize deep sequence data even more successfully, it is likely that such approaches will require a model specifying within-host pathogen dynamics, which are still poorly understood – estimation of effective population size within host, *in vivo* pathogen generation time and transmission bottleneck size are all challenging. Furthermore, the methods described here are simple and quick to implement.

We applied these methods to Ebola data collected from Sierra Leone in 2014. While the first dataset is thought to represent relatively dense coverage of the initial stages of the epidemic in the country, with estimates of around 70% of cases sampled (15, 23), the later dataset comprised a sparser sample. While sparse sampling reduces the number of true links one would expect to find via any method, shared variant approaches do not require 100% sampling. Missing data greatly reduces the number of shared variants expected, and therefore the number of estimated transmission routes, but does not greatly impact the reliability of the transmission routes which are proposed (Figures S5 and S6). As such, while only relatively few transmission routes were identified in the datasets, this is likely a function of both the proportion of missing data, and the relatively low mutation rate of Ebola virus (15, 24). However, the transmission pairs we identified appear plausible from temporal and geographic clustering. Both datasets contained a large group of hosts sharing the same variant. Park et al. suggest that the large group in the second dataset likely arose through a combination of patient-to-patient transmission and recurrent mutation (16). Testing for recurrent mutation, particularly in larger variant-sharing groups, as well as cross-checking transmission links against other data sources, is therefore highly recommended.

The hybrid approaches allow nodes with no shared variants to be connected to other nodes in the transmission tree. This provides considerable improvements when the bottleneck is small or the mutation rate is low. The hybrid approach could be performed either by using separate single genomes, or the consensus sequence of the deep sequenced sample. Since transmission routes are assessed independently of one another, estimated transmission trees frequently comprise several unconnected nodes or clusters. Such unconnected clusters could be linked to one another if further structure is required, using the weighted distance approach on pooled within-cluster samples.

Care should be taken when identifying variant sites to minimize the risk of calling sequencing or alignment errors as true variants. Deep coverage and replicate sequencing provide some reassurance of variant calling

quality; nevertheless, we found considerable sensitivity to adjusting the variant frequency threshold, particularly for the second, lower coverage, dataset (Figure S4). Setting a conservatively high threshold increases the probability of calling true variants, but reduces the amount of useful information for transmission tree estimation. While the cost of sequencing bacterial pathogen populations to a sufficient depth to call minor variants remains restrictive, this is likely to reduce in the future.

We have demonstrated the power of deep sequencing data to identify transmission routes to a greater resolution than by using analogous methods with single genome sequence data. We have purposefully omitted the incorporation of additional data sources (such as times of sampling, symptom onset, recovery/death, as well as geographic location and contact tracing) in order to evaluate the information provided by the genomic data alone. Incorporating additional data sources will improve estimates, allowing further potential transmission links to be ruled out. Since we make no assumption about the order of infection, directionality is ambiguous in the majority of estimated transmission links. Rigorous collection of epidemiological data remains a crucial component of outbreak investigation, and combining this with deep sequencing and shared variant analysis can provide unprecedented insight into individual-level transmission dynamics. We believe that the shared variant methods described here will become common once deep sequence data collection becomes more widespread, and should provide a considerably clearer picture of who infected whom than single genome sampling data.

Materials and Methods

Let $x_1,\ldots,x_n$ denote deep-sequence samples collected from hosts $1,\ldots,n$. For each sample x_i, let $f_1^{(i)},\ldots,f_G^{(i)}$ be the frequency of the majority nucleotide at loci $1,\ldots,G$, such that polymorphisms exist where $f_j^{(i)} < 1$. For each host, identify the set of polymorphisms $V_i = \{j : f_j^{(i)} < 1\}$. Now calculate the variant score S_{ij} between each pair of hosts i and j to be the number of shared variants belonging to the samples x_i and x_j; $S_{ij} = \left|V_i \cap V_j\right|$. If we allow for the possibility of different mutations at a given locus, we must further restrict to the set of variant positions sharing the same mutant nucleotide. The matrix (S_{ij}) can then be viewed as a weighted adjacency matrix defining an estimated transmission tree (which we call the *weighted variant tree*). We further define the *maximum variant tree*, in which we identify for each host the individual sharing the greatest number of variants. If multiple individuals share the maximum number of

variants, these are attributed equal weight. If no individual shares any variant with a host, it is not assigned a source.

These approaches have similarities with methods using single genome samples. For instance, the *minimum distance tree* is defined by assigning the source to be the individual carrying the most genetically similar sample (i.e. fewest number of SNPs). A variation of this approach was described in (6). Similarly, the *weighted distance tree* is defined by weighting each network edge by inverse genetic distance, such that more similar samples are given a greater weight. Variations of this method were explored in (2).

Finally we propose two hybrid approaches. In some cases, a host will share no variants with any other host in the population, such that the maximum and weighted variant approaches assign equal weight to all potential sources of infection. As such, we may instead attempt to draw information from genetic distance measures where no shared variants exist. The *hybrid maximum tree* and the *hybrid weighted tree* attribute sources to hosts lacking shared variants according to the minimum distance and weighted distance approaches respectively. Genetic distance can be calculated using the consensus sequence of the deep sequenced sample, or with additional single genome samples.

Simulations

We simulated outbreaks using the R package *seedy* v1.2 (25), introducing a single infectious individual into a susceptible, homogeneously mixing population of size 100. Genomic samples (perfectly observed deep sequence samples) were generated at a random time during each individual's infectious period. Furthermore, we sampled single genomes from each host in order to compare transmission route estimation using each type of data. Infection dynamics were simulated under a standard SIR (susceptible-infected-removed) model, with $R_0 = 2$. Multiple infections were not permitted. Infections were generated by selecting n_B genotypes at random from the source's pathogen population, and allowing this inoculum to grow within the new host under neutral evolution. We varied transmission bottleneck size n_B as well as mutation rate in order to simulate a range of different outbreaks. Transmission trees were visualized using the igraph package in R (26).

Measuring reconstruction accuracy

We used various metrics to compare the performances of the transmission route identification methods. We assumed that infection and removal times were not observed from the simulated outbreaks, such that we could compare the ability of the genomic data alone to contribute to transmission route identification.

The receiver operating characteristic (ROC) curve describes the change in false positive and true positive rate for identifying a source of infection as the weight threshold for this identification varies between 0 and 1. The area under the ROC curve (AUC) is a summary statistic of this function, measuring the overall discriminatory power of the tree reconstruction, in which values closer to 1 indicate a more accurate network (27).
For the unweighted reconstructions (*minimum distance* and *maximum similarity tree*), we calculated the path distance in the true network for each proposed transmission link. For instance, if we identify the route A-B, and in reality the transmission chain was A-C-B, the path distance in the true network is 2. A perfect reconstruction would thus have a mean path distance of 1. While the ROC curve treats edges as either correct or incorrect, the latter metric provides a measure of the extent to which false links are misleading (i.e. an incorrect edge with a true path length of 2 is better than an incorrect edge with a true path length of 10).

Finally, we considered the mean weight attributed to the true source of infection across an outbreak. While the path distance metric did not factor in hosts for whom no source could be attributed (due to a lack of shared variants), this measure includes such hosts with a weight of zero.

Data

In addition to the simulation studies, we applied the shared variant approach to identifying potential transmission routes during the 2014 Ebola outbreak in Sierra Leone. We used samples collected from 78 patients in May-June 2014, representing a large proportion of the earliest cases in the country, sequenced to approximately 2000x coverage (15). Furthermore, we considered samples from a further 150 patients collected between June and December in the same country, sequenced with a median coverage of 374x (16).

Funding Information

Research reported in this paper was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number U54GM088558. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. **Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, Haydon DT.** 2008. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proc R Soc B **275:**887-895.
2. **Worby CJ, Lipsitch M, Hanage WP.** 2014. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. PLoS Comp Biol **10:**e1003549.
3. **Didelot X, Gardy J, Colijn C.** 2014. Bayesian analysis of infectious disease transmission from whole genome sequence data. Mol Biol Evol **31:**1869-1879.
4. **Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM.** 2012. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc R Soc B **279:**444-450.
5. **Ypma RJF, van Ballegooijen WM, Wallinga J.** 2013. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics **195:**1055-1062.
6. **Jombart T, Eggo RM, Dodd PJ, Balloux F.** 2011. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity **106:**383-390.
7. **Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N.** 2014. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Comp Biol **10:**e1003457.
8. **Streulens MJ, Deplano A, Godard C, Maes N, Serruys E.** 1992. Epidemiologic typing and delineation of genetic relatedness of methicillin-resistant Staphylococcus aureus by macrorestriction analysis of genomic DNA by using pulsed-field gel electrophoresis. J Clin Microbiol **30:**2599-2605.
9. **Strommenger B, Braulke C, Heuck D, Schmidt C, Pasemann B, Nübel U.** 2008. spa Typing of Staphylococcus aureus as a Frontline Tool in Epidemiological Typing. J Clin Microbiol **46:**574-581.

10. **Koreen L, Ramaswamy SV, Graviss EA, Naidich S, Musser JM, Kreiswirth BN.** 2004. spa Typing Method for Discriminating among Staphylococcus aureus Isolates: Implications for Use of a Single Marker To Detect Genetic Micro- and Macrovariation. J Clin Microbiol **42:**792-799.
11. **Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P.** 2011. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. New Engl J Med **364:**730-739.
12. **Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, Kremer K, van Hijum SAFT, Siezen RJ, Borgdorff M, Bentley SD, Parkhill J, van Soolingen D.** 2013. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC Infect Dis **13:**110.
13. **Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, Churchill S, Bennett K, Golubchik T, Giess AP, Del Ojo Elias C, Jeffery KJ, Bowler ICJW, Laurenson IF, Barrett A, Drobniewski F, McCarthy ND, Anderson LF, Abubakar I, Thomas HL, Monk P, Smith EG, Walker AS, Crook DW, Peto TEA, Conlon CP.** 2014. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. The Lancet Respiratory Medicine **2:**285-292.
14. **Worby CJ, Chang H-H, Hanage WP, Lipsitch M.** 2014. The distribution of pairwise genetic distances: a tool for investigating disease transmission. Genetics **198:**1395-1404.
15. **Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang P-P, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JLB, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, et al.** 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science **345:**1369-1372.
16. **Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, Sealfon RS, Ladner JT, Kugelman JR, Matranga CB, Winnicki SM, Qu J, Gire SK, Gladden-Young A, Jalloh S, Nosamiefan D, Yozwiak NL, Moses LM, Jiang P-P, Lin AE, Schaffner SF, Bird B, Towner J, Mamoh M, Gbakie M, Kanneh L, Kargbo D, Massally JLB, Kamara FK, Konuwa E, Sellu J, Jalloh AA, Mustapha I, Foday M, Yillah M, Erickson BR, Sealy T, Blau D, Paddock C, Brault A, Amman B, Basile J, Bearden S, Belser J, Bergeron E, Campbell S, Chakrabarti A, Dodd K, Flint M, Gibbons A, et al.** 2015. Ebola Virus

Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. Cell **161:**1516-1526.

17. **Poon LLM, Chan KH, Chu DKW, Fung CCF, Cheung CKY, Ip DKM, Leung GM, Peiris JSM, Cowling BJ.** 2011. Viral genetic sequence variations in pandemic H1N1/2009 and seasonal H3N2 influenza viruses within an individual, a household and a community. J Clin Virol **52:**146-150.
18. **Hughes J, Allen RC, Baguelin M, Hampson K, Baillie GJ, Elton D, Newton JR, Kellam P, Wood JLN, Holmes EC, Murcia PR.** 2012. Transmission of Equine Influenza Virus during an Outbreak Is Characterized by Frequent Mixed Infections and Loose Transmission Bottlenecks PLoS Path **8:**e1003081.
19. **Murcia PR, Hughes J, Battista P, Lloyd L, Baillie GJ, Ramirez-Gonzalez RH, Ormond D, Oliver K, Elton D, Mumford JA, Caccamo M, Kellam P, Grenfell BT, Holmes EC, Wood JLN.** 2012. Evolution of an Eurasian Avian-like Influenza Virus in Naïve and Vaccinated Pigs PLoS Path **8:**e1002730.
20. **Emmett KJ, Lee A, Khiabanian H, Rabadan R.** 2015. High-resolution Genomic Surveillance of 2014 Ebolavirus Using Shared Subclonal Variants. PLOS Currents Outbreaks doi:10.1371/currents.outbreaks.c7fd7946ba606c982668a96bcba43c90.
21. **Balloux F.** 2010. Demographic influences on bacterial population structure. *In* Robinson DA, Falush D, Feil EJ (ed), Bacterial Population Genetics in Infectious Diseases. John Wiley & Sons Inc.
22. **WHO Ebola Response Team.** 2014. Ebola Virus Disease in West Africa - The First 9 Months of the Epidemic and Forward Projections. New Engl J Med **371:**1481-1495.
23. **Stadler T, Kühnert D, Rasmussen DA, du Plessis L.** 2014. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequencing. PLOS Currents Outbreaks doi:10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f.
24. **Hoenen T, Safronetz D, Groseth A, Wollenberg KR, Koita OA, Diarra B, Fall IS, Haidara FC, Diallo F, Sanogo M, Sarro YS, Kone A, Togo ACG, Traore A, Kodio M, Dosseh A, Rosenke K, de Wit E, Feldmann F, Ebihara H, Munster VJ, Zoon KC, Feldmann H, Sow S.** 2015. Mutation rate and genotype variation of Ebola virus from Mali case sequences. Science **348:**117-119.
25. **Worby CJ, Read TD.** 2015. 'SEEDY' (Simulation of Evolutionary and Epidemiological Dynamics): An R Package to Follow Accumulation of Within-Host Mutation in Pathogens PLoS One **10:**e0129745.
26. **Csardi G, Nepusz T.** 2006. The igraph software package for complex network research. InterJournal Complex Systems**:**1695.
27. **Fawcett T.** 2006. An introduction to ROC analysis. Pattern Recog Lett **27:**861-874.

Supplementary Figures

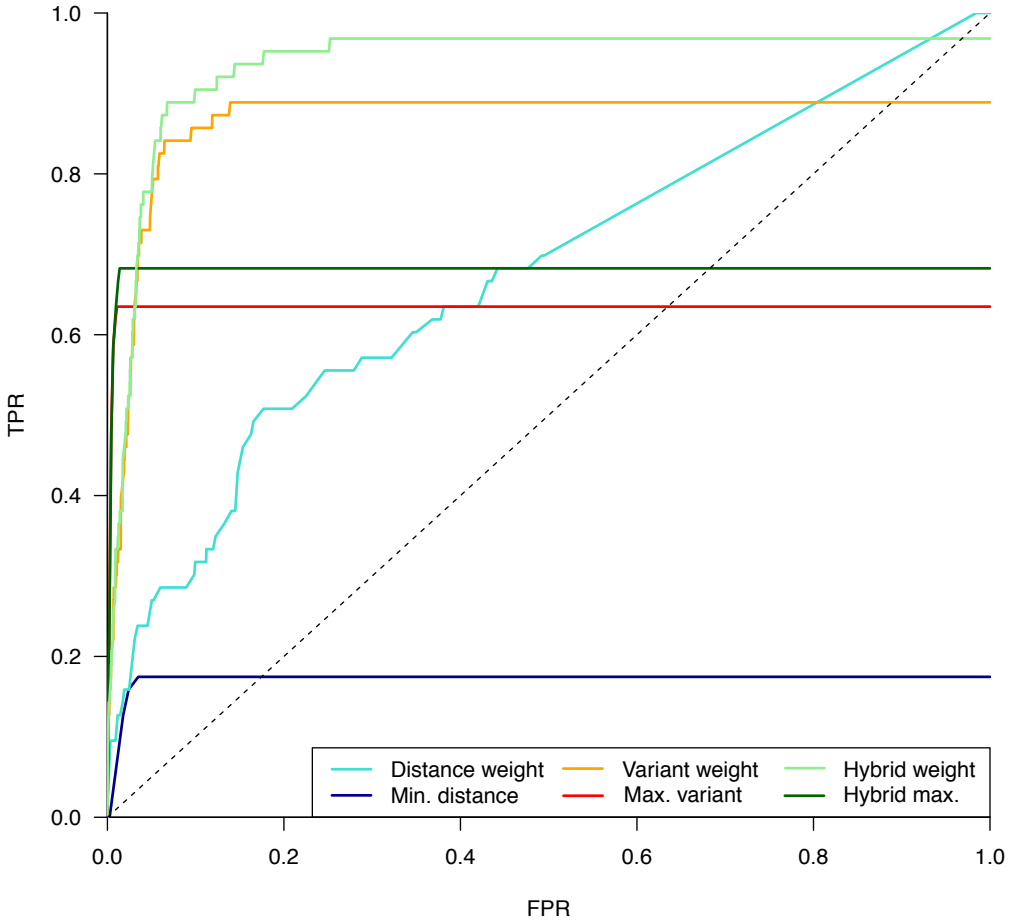

Figure S1. ROC curves. ROC curves for the closest genome tree (green), weighted genetic distance tree (purple), and the weighted variant tree (red). Outbreak simulated with bottleneck size 10 and mutation rate 0.001, corresponding to the transmission trees shown in figure 1. The dashed line indicates the performance of selecting routes at random. Since multiple correct edges may be assigned a weight of zero under all but the weighted distance approach, those ROC curves do not reach (1,1).

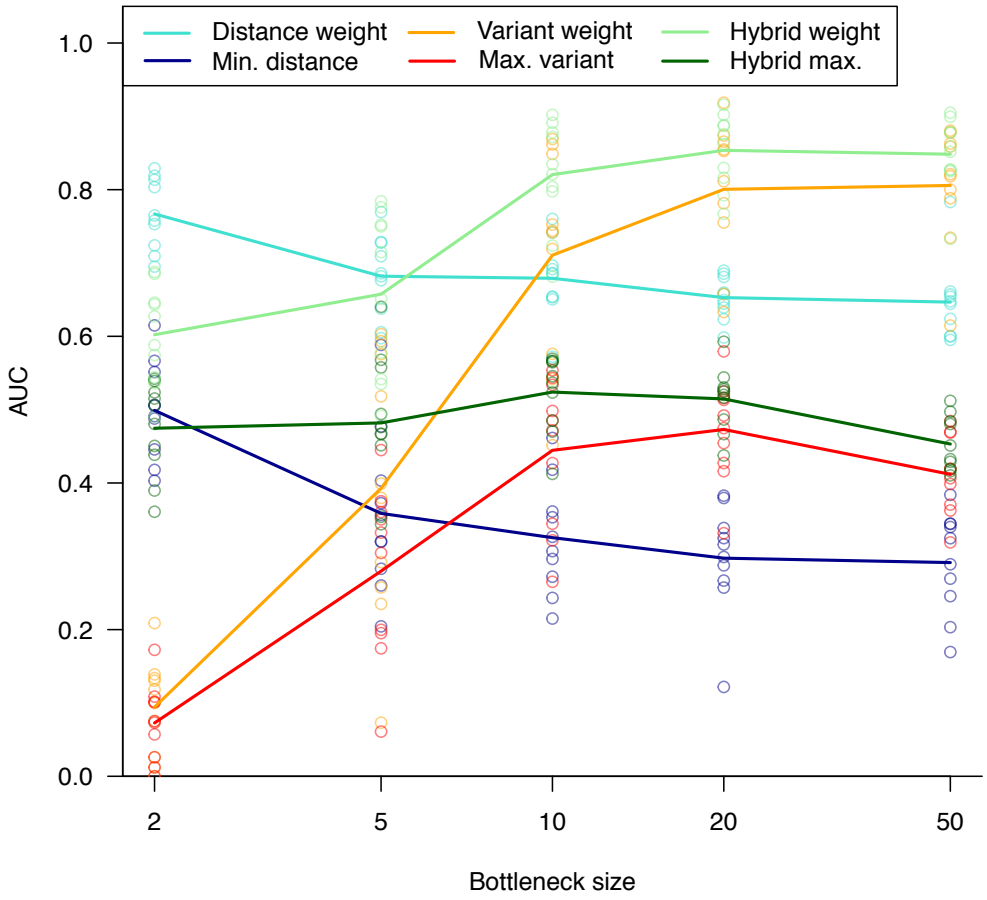

Figure S2. Area under ROC curves. The area under the ROC (AUC) metric provides an overall measure of network accuracy. Results for a mutation rate of 0.0005 are shown here. Hosts with no observed shared variants are assumed to have all other hosts as potential sources with equal weight. This figure is equivalent to Figure 4B with a lower mutation rate.

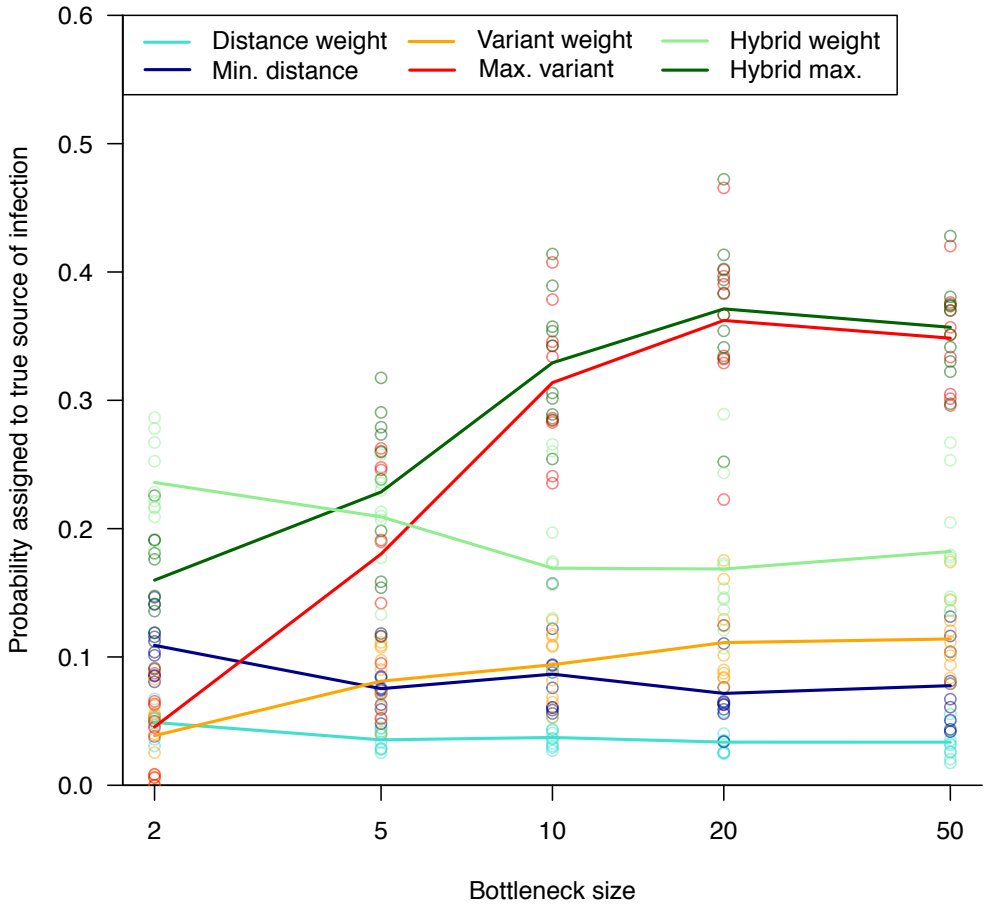

Figure S3. True edge weighting. The mean weight attributed to each true transmission link for each tree reconstruction method, under a range of scenarios and methodologies. Results are shown for a mutation rate of 0.0005.

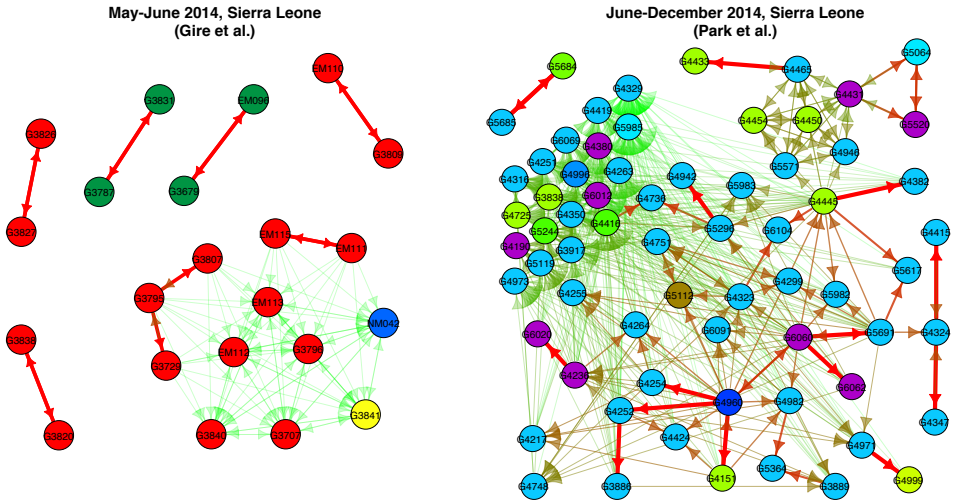

Figure S4. Ebola transmission routes. Estimated transmission links between sampled hosts in the Ebola outbreak under the maximum variant approach. Colors denote the chiefdom to which each host belongs, while the color and thickness of the arrows denote the relative weight attributed to each potential transmission event. Variant detection threshold 1%.

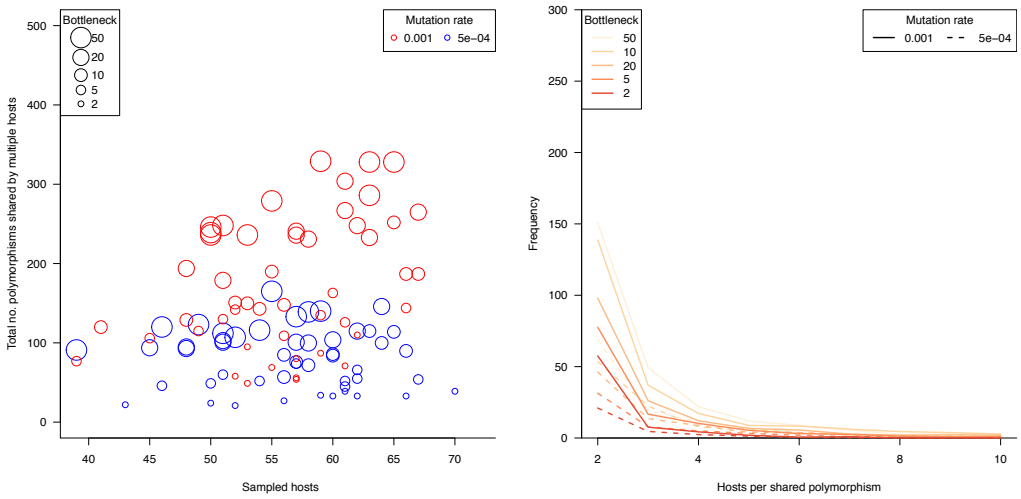

Figure S5. Summary of genetic variant frequency across imperfectly sampled simulated outbreaks. Summary of genetic variant frequency across the simulated outbreaks with 30% of infected hosts unsampled. (A) Total number of shared variants across the simulated outbreak. Bottleneck size is illustrated by circle size. (B) Distribution of shared variant group size for different bottlenecks and mutations rates.

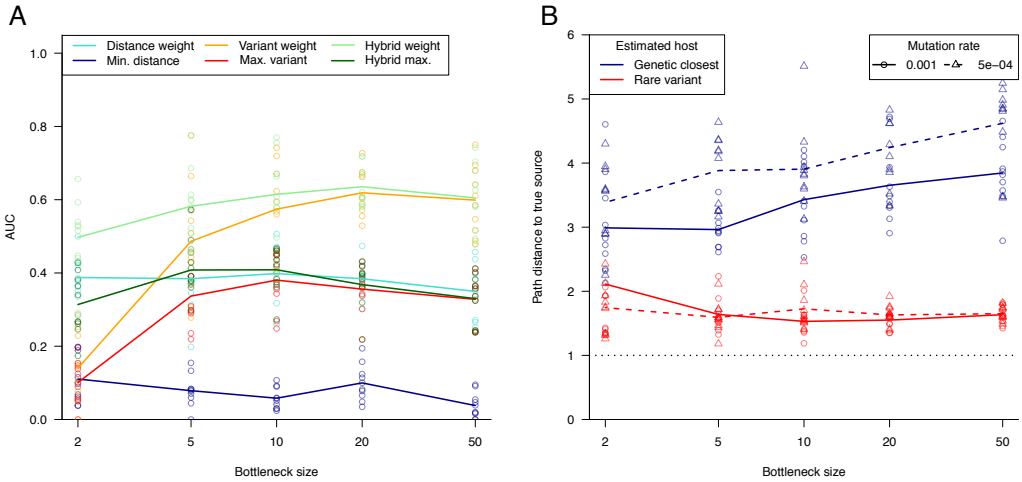

Figure S6. Transmission tree reconstruction accuracy. (A) The area under the ROC (AUC) metric provides an overall measure of network accuracy. Results for a mutation rate of 0.001 are shown here. (B) The true path distance between estimated transmission pairs gives insight into the extent to which transmission links are misspecified. A perfect reconstruction would have mean path length 1. Maximum variant path lengths are averaged over identified transmission pairs, that is, excluding hosts with no shared variants.