HIV-TRACE (Transmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens.

Sergei L. Kosakovsky Pond, Steven Weaver, Andrew J. Leigh Brown, and Joel O. Wertheim

1Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA,
2Institute of Evolutionary Biology, University of Edinburgh, United Kingdom.
3Department of Medicine, University of California, San Diego, CA, USA.

*Corresponding author: E-mail: jwertheim@ucsd.edu

Abstract

In modern applications of molecular epidemiology, genetic sequence data are routinely used to identify clusters of transmission in rapidly evolving pathogens, most notably HIV-1. Traditional ‘shoe-leather’ epidemiology infers transmission clusters by tracing chains of partners sharing epidemiological connections (e.g., sexual contact). Here, we present a computational tool for identifying a molecular transmission analog of such clusters: HIV-TRACE (TRAnsmission Cluster Engine). HIV-TRACE implements an approach inspired by traditional epidemiology, by identifying chains of partners whose viral genetic relatedness imply direct or indirect epidemiological connections. Molecular transmission clusters are constructed using codon-aware pairwise alignment to a reference sequence followed by pairwise genetic distance estimation among all sequences. This approach is computationally tractable and is capable of identifying HIV-1 transmission clusters in large surveillance databases comprising tens or hundreds of thousands of sequences in near real time, i.e., on the order of minutes to hours. HIV-TRACE is available at www.hivtrace.org and from github.com/veg/hivtrace, along with the accompanying result visualization module from github.com/veg/hivtrace-viz. Importantly, the approach underlying HIV-TRACE is not limited to the study of HIV-1 and can be applied to study outbreaks and epidemics of other rapidly evolving pathogens.

Key words: molecular epidemiology; HIV; network; transmission cluster; surveillance

Introduction

Research into fundamental questions of epidemiology and public health, such as "Who infected whom?" (Romero-Severson et al., 2016; Volz and Frost, 2013), "How does pathogen X spread through a population"? (Dennis et al., 2014), and "Is a particular prevention or treatment effective at slowing or stopping the spread of disease?" (Little et al., 2014) has greatly benefited from large-scale analyses of molecular sequences obtained during surveillance or through routine diagnostics. For rapidly
evolving pathogens, such as HIV-1 or hepatitis C virus, viral isolates from different hosts will typically not be genetically identical, and analyses of these genetic differences via phylogenetic, phylodynamic, or other evolutionary methods have proven tremendously powerful. Phylogenetic analyses have been used in criminal cases involving deliberate HIV-1 transmission (Scaduto et al., 2010), to understand the introduction of HIV-1 into regions and countries (Gilbert et al., 2007), and to define recent clusters of transmission cases (Peters et al., 2016). Recent work in the field of phylodynamics has established a template on how to use sequence data to inform inference of epidemiological transmission parameters, e.g., $R_0$ or transmission rates between different risk groups (Frost and Volz, 2013; Volz and Frost, 2014). The fundamental insight shared by all these methods is that genetic similarity, or relatedness, between pathogen sequences can be used to identify strains that are connected in an epidemiologically meaningful way: as potential source-recipient pairs (Campbell et al., 2011) or members of the a distinct transmission cluster (Campbell et al., 2017; Wertheim et al., 2017a).

Real time or near real time surveillance of pathogen transmission is an area of great interest to local, national, and global public health agencies (Division of HIV/AIDS Prevention, 2017). Real time surveillance seeks to quickly analyze newly obtained pathogen genetic sequences in the context of large, preexisting reference samples and to deliver actionable inference results: "A new rapidly growing HIV-1 transmission cluster has been identified", or "An unusual pattern of transmission between people with different risk factors has been detected", or "An HIV-1 transmission prevention is effectively reducing population level incidence".

Defining molecular transmission clusters is a challenging problem, and currently there is no consensus in the field of molecular epidemiology of what should or should not constitute a transmission cluster or whether certain definitions are more germane to particular research questions or public health interventions (Grabowski and Redd, 2014; Hassan et al., 2017; Novitsky et al., 2017; Wertheim et al., 2014).

Here, we present the algorithmic, software implementation, and operational usage details for HIV-TRACE, a platform that has been used extensively for rapid inference of transmission networks from large sets of pathogen genetic sequences to identify potential transmission links and to describe putative transmission clusters. An early version of HIV-TRACE was used to analyze nearly 100,000 HIV-1 sequences sampled worldwide, and this analysis revealed that there was a surprising amount of global (country-to-country) connectivity in this network (Wertheim et al., 2014). Since then, HIV-TRACE has been used to investigate transmission patterns among risk groups (Oster et al., 2015; Whiteside et al., 2015), characterize transmission fitness of HIV.
Table 1. Key parameters controlling HIV-TRACE

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>-r, --reference</td>
<td>Reference sequence for mapping</td>
<td>Alignment</td>
</tr>
<tr>
<td>-m, --minoverlap</td>
<td>Sequences must have at least this many aligned characters</td>
<td>Distance estimation</td>
</tr>
<tr>
<td>-a, --ambiguities</td>
<td>Sets policy for handling ambiguous nucleotides</td>
<td>Distance estimation</td>
</tr>
<tr>
<td>-g, --fraction</td>
<td>Sets the maximum fraction of resolvable ambiguous nucleotides</td>
<td>Distance estimation</td>
</tr>
<tr>
<td>-s, --strip_drams</td>
<td>Mask HIV-1 drug resistance associated sites</td>
<td>Distance estimation</td>
</tr>
<tr>
<td>-t, --threshold</td>
<td>Distance threshold for drawing a network link</td>
<td>Network construction</td>
</tr>
<tr>
<td>-u, --curate</td>
<td>Sets policy for handling potential contaminants</td>
<td>Network construction</td>
</tr>
</tbody>
</table>

drug-resistance associated mutations (Wertheim et al., 2017b), and to identify rapidly growing transmission clusters (Campbell et al., 2017; Monterosso et al., 2017).

The source code, installation instruction (via pip3), and documentation for HIV-TRACE is available at github.com/veg/hivtrace, and the accompanying result visualization module – at github.com/veg/hivtrace-viz. In addition, a public instance of the HIV-TRACE web-application is hosted at www.hivtrace.org, as a part of the Datamonkey family of services (Weaver et al., 2018).

New Approaches

HIV-TRACE does not infer a phylogenetic tree from sequence data because phylogenetic inference is a computational bottleneck and because the phylogenies themselves are typically not directly useful for epidemiological inference. In most applications, phylogenies are converted to summary features (e.g., clades) or summary statistics (e.g., patristic distances) to identify clusters. In lieu of phylogenetic inference, HIV-TRACE identifies groups of putative transmission partners and assembles these partners in transmission clusters. This approach is analogous to the traditional epidemiological definition of an infectious disease transmission cluster: a group of infected people with direct or indirect epidemiological connections. In HIV-TRACE, genetic linkage serves as a proxy for these direct or indirect epidemiological connections, and a cluster is constructed based on these connections. This approach is fundamentally different from phylogenetic-based cluster inference (Grabowski and Redd, 2014; Wertheim et al., 2014), which seeks to identify a point in evolutionary history from which all cluster members descend (i.e., a point that gives rise to a clade on a phylogeny). Importantly, several independent studies have shown that in many cases relevant to HIV-1 epidemiology, HIV-TRACE reports very similar sets of clusters to phylogeny-based methods (Poon,
Inference procedure

HIV-TRACE takes in a collection of $N$ unaligned coding viral sequences sampled from $M \leq N$ individuals (multiple sequences per individual are supported) formatted as a FASTA file, and it outputs a JSON file containing the description of the inferred transmission network as nodes (individuals) and links (potential transmission partners). When additional clinical, demographic, or other data are available, they can be included in the network as attributes. Key parameters controlling network inference are summarized in Table 1, and the schematic of program flow is depicted in Figure 1.

To demonstrate method performance, we downloaded all publicly available HIV-1 polymerase sequence (one sequence per patient, minimum length 500 nt) from the Los Alamos National Laboratories HIV database (hiv.lanl.gov), resulting in $N=M=185,849$ sequences. We randomly sampled a set of 256, 1,024, 4,096, 16,384, and 65,536 sequences to plot computational time scaling. We ran each step of the pipeline 10 times (to average out computing environment stochasticity) on a 64-core (2x32 AMD Opteron 6356) system running at 2 GHz clock rate (Figure 1).

**Sequence alignment.** HIV-TRACE first aligns each of the input sequences to a single reference sequence using a codon-aware extension of the Smith-Waterman dynamic programming algorithm (Smith and Waterman, 1981), previously developed by us in the context of high throughput sequencing read mapping (e.g., Gianella et al. (2011)). For standard HIV-1 analyses, the HXB2 sequence (GenBank Accession number: K03455) is used as a reference sequence, although any in-frame coding sequence can be supplied as reference. Both the forward and the reverse-complement versions of each sequence are considered, and the one with the higher alignment score is retained. Codon-aware alignment leverages protein homology to align nucleotide data and is able to identify and correct relatively frequent (i.e., up to 5% of sequences in some datasets) frame-shifting insertions or deletions involving one or two nucleotides. In this case, correction means maintaining the frame relative to the reference. The resulting pairwise alignment is merged into a single multiple-sequence alignment (MSA). As the vast majority of HIV-1 sequence data arise from surveillance screening for drug resistance in a 1497 nucleotide protease and reverse transcriptase genomic region, which only rarely exhibit insertions/deletions relative to the reference
sequence, this "mapping" approach is effective and scales linearly in the number of sequences. Traditional progressive alignment methods have superlinear (e.g., up to quadratic) computational cost. In our example, computational complexity scaled linearly as expected, and the alignment of 185,849 sequences to a reference took about 20 minutes on average. Importantly, HIV-TRACE is also capable of handling previously aligned sequences, which may be desirable for analyses of HIV-1 envelope sequences or other pathogens with low evolutionary conservation, where "all-to-one" alignment is not likely to recover more distant homologies. However, genes or sequence regions that are challenging to align may be suboptimal for molecular epidemiology applications.

Estimation of genetic distances. Given a multiple sequence alignment on \( N \) sequences, HIV-TRACE computes all \( N \times (N-1)/2 \) pairwise genetic distances under the Tamura-Nei 93 (TN93) (Tamura and Nei, 1993) nucleotide substitution model, which is the most general nucleotide substitution model for which distances can be estimated directly from counts of nucleotide pairs in aligned sequences. Whereas more complex models substitution models are typically preferable in the context of phylogenetic inference, especially for more distantly related strains, (Posada and Crandall, 2001), when genetic distances are low (e.g. 0.05), all sensible nucleotide distance measures perform comparably (Wertheim and Kosakovsky Pond, 2011). A key option controlling this step in HIV-TRACE is how to handle ambiguous nucleotide characters that represent within-host population polymorphisms or sequencing errors (see Parameterizing genetic distance estimates). An important example of epidemiological processes that yield sequences with high fractions of ambiguous nucleotides is multiple (super- or dual-) HIV infection (Pacold et al., 2010).

Pairwise distances are reported to a comma separated file, and are typically limited only to those pairs that are below a user-specified threshold (e.g., 0.015 substitutions/site) to retain only pairs of sequences that have an epidemiological link. This step is computationally costly, scaling as \( N^2 \), but an efficient parallelized implementation of the tool allows rapid processing of \( 10^5-10^6 \) sequences. For instance, it took approximately 32 minutes to compute all pairwise distances between 185,849 sequences. Our implementation is also memory efficient, requiring \( O(NL) \) space, where \( L \) is the sequence length. For datasets of this size, traditional rapid phylogeny reconstruction techniques, such as Neighbor Joining are already infeasible, because they scale as \( N^3 \) and require the storage of the entire distance matrix (this would require approximately 256GiB of RAM for our example), which HIV-TRACE deliberately avoids. Because most phylogenetic methods for cluster definition require some measure of clade support
(e.g. Grabowski and Redd (2014)), it is also necessary to perform a version of bootstrapping. Our implementation compares favorably to even the fastest tree construction methods, such as FastTree 2 (Price et al., 2010a) or IQ-Tree (Nguyen et al., 2015), which takes at least 10x longer to process these sizes of data; for example, typical run times of FastTree 2 (the fastest tool to our knowledge) on ∼200,000 sequences is on the order of 10−20 hours (Price et al., 2010a). It is worth noting that FastTree 2 has an asymptotically better run time \( O(N^{3/2} \log N) \), but it does considerably more work than needed for our application (resulting in slower run times), and uses heuristics which are not guaranteed to always find all distances below a certain threshold.

Network construction. The transmission network is inferred from the file of pairwise distances and optionally annotated with data from attribute files. Nodes within the network are all keyed on either the entire sequence name or parts thereof extracted by regular expressions. A link is drawn between two individuals if and only if the pairwise distances between any of the paired sequences from these individuals is below a user specified threshold, \( D \). e.g. \( D=0.015 \). A cluster is defined as a connected component of the network. Optionally, the network can be screened for contaminants (i.e., any query sequences that link to lab strains or other user-specified contaminant sequences). Global statistics of the network, such as the number of nodes, edges, clusters, cluster sizes, and the degree distribution, are computed and reported. Lastly, the degree distribution is fit to one of four generative models of network growth: random attachment, preferential attachment, preferential attachment mixed with a component of random attachment, and power law, using the methods described by (Handcock and Jones, 2004). If the best fitting model is from a scale-free family (i.e., preferential attachment), the characteristic exponent \( \rho \) of the network is estimated and reported. This step is computationally relatively inexpensive, taking only a few seconds.

Parameterizing genetic distance estimates

Selecting appropriate parameters governing genetic distance estimation is critical to HIV-TRACE analysis. Investigations in the U.S., the U.K., and Canada have consistently found natural breakpoints in genetic distance between putative transmission partners and 'random' cases or within-host and between-host diversity (Lewis et al., 2008; Poon et al., 2015; Rose et al., 2017a; Smith et al., 2009; Wertheim et al., 2017a). In New York City, genetic distance thresholds between 0.01 and 0.02 substitutions/site were more strongly associated with probable transmission partners than traditional epidemiological connections (i.e., naming of sexual and injection drug using partners) and that a distance of 0.015 could serve
as a use proxy for epidemiological relatedness in a surveillance setting (Wertheim et al., 2017a).

Moreover, these genetic distance thresholds have been validated by molecular epidemiological studies in U.S. public health surveillance populations (Oster et al., 2015; Wertheim et al., 2016, 2017b; Whiteside et al., 2015), which have reported results that are typically robust to thresholds in this range. Lower distance thresholds (e.g., 0.005 substitutions/site) may be more appropriate for distinguishing rapidly growing clusters (Division of HIV/AIDS Prevention, 2017) or populations where faster evolving (i.e., non-B subtypes) predominate. As distance thresholds increase, smaller clusters merge into larger, less informative clusters Figure 2A. At the extreme, all sequences would belong to a single cluster, which while technically correct, since all HIV-1 sequences are related through a series of transmissions, this finding is unlikely to be of interest in the context of molecular epidemiology. The same principle – that \( D \) should separate within-host or epidemiologically recent diversity from between-host diversity has been used successfully for other epidemics, genetic regions and viruses. For example Rose et al. (2017b) used \( D = 0.053 \) for HIV-1 gp41, Bartlett et al. (2017) selected \( D = 0.03 \) for the core gene of Hepatitis C virus. Regional and national epidemics HIV-1 also tend to require larger thresholds due to sparser sampling and the prevalence of chronically infected individuals (Hassan et al., 2017).

Nucleotide ambiguities (e.g., \( Y \) indicating a mixed population of both \( C \) and \( T \) at the same genomic position) have the potential to compromise HIV-TRACE analysis, or phylogenetic inference in general. By default, HIV-TRACE will resolve (here, to 'resolve' means to choose the value of the ambiguity to match the other nucleotide if possible) the genetic distance between ambiguities (i.e., \( Y \) is 0 substitutions from both \( C \) and \( T \)). However, sequences with a high fraction of nucleotide ambiguities have the tendency to link to distantly related sequences when ambiguities are resolved, resulting in artifactual larger clusters Figure 2B. When ambiguities are properly accounted for, HIV-TRACE clusters tend to resemble clades on a phylogenetic tree Figure 2C. However, when distances ambiguities are resolved irrespective of ambiguity fraction, distantly related sequences are connected through these high ambiguity sequences, forming large artifactual clusters (Figure 2D and Aldous et al. (2012)). Therefore, HIV-TRACE includes a parameter (ambiguity fraction) that averages the genetic distance from ambiguities (i.e., \( Y \) is 0.5 substitutions from both \( C \) and \( T \)) in sequences with a higher proportion of ambiguities than the indicated ambiguity fraction. In cohorts of fewer than 1000 individuals (i.e., San Diego Primary Infection Cohort), an ambiguity fraction of 0.05 is appropriate based on empirical
network sensitivity analyses. For US surveillance data, an ambiguity fraction above 0.015 produces spurious clusters. As a consequence, sequences with high ambiguity fractions are less likely to cluster using HIV-TRACE.

In HIV-TRACE, excluding sites containing ambiguities has a similar effect on network construction as resolving ambiguities. Many popular phylogenetic packages used for constructing HIV-1 molecular transmission networks (e.g., BEAST (Drummond et al., 2012) and FastTree (Price et al., 2010b)) exclude sites containing ambiguities from likelihood calculations. It remains unclear how treatment of nucleotide ambiguities will affect phylogenetic inference of HIV transmission clusters (Fearnhill et al., 2017).

Visualization
The JSON file output by HIV-TRACE can be explored using an interactive JavaScript application which we call hivtrace-viz. It is based on the open source data visualization library d3.js. This application runs within any modern web-browser and provides means to view the overall structure of the network, explore individual clusters, display network summary, and explore associations among attributes for connected nodes. When clinical and demographic attributes are available, they can be overlaid on the network structure as shown in Figure 3.

Software components
Alignment. bealign is implemented in Python 3 as a part of BioExt library (github.com/veg/BioExt) which extends the functionality of the popular BioPython library (Cock et al., 2009). The core alignment routine is implemented in C and incorporated via Cython. When the program is run in a multicore/multiprocessing environment, it will distribute alignment tasks across cores.

Distance calculation. tn93 is a self-contained C++ program (available from github.com/veg/tn93) which is tuned to allow $\sim 10^7 - 10^6$ distance calculations per second per core on $\sim 1000$ bp long sequences. It uses OpenMP to distribute distance calculations across multiple CPU cores whenever possible. For example, tn93 achieved parallelized (64 cores) throughput of $\sim 10^7$ pairwise distance calculations per second when computing distances on the LANL example dataset.

Network inference. hivnetworkcsv is a Python 3 module, which is available from github.com/veg/hivclustering, along with the attendant documentation.

Concluding Remarks
HIV-TRACE is a powerful computational tool for the rapid and automated characterization of molecular transmission clusters in populations of HIV infected individuals. Its applicability for HIV research and public health surveillance and prevention activities is apparent, as first
illustrated by the unsupervised recovery of many previously characterized clusters (defined via phylogenetic analyses) in our global-scale analysis of HIV-1 databases (Wertheim et al., 2014). As viral sequence databases increase in size and transition to using Next Generation Sequencing (NGS) data, scalable tools like HIV-TRACE will be increasingly relevant.

HIV-TRACE can accommodate NGS data in three different ways. Firstly, NGS data can be used to generate a consensus sequence for each individual, which is then handled the same way as Sanger sequences are now. Phylogenetic approaches most commonly use this route, and HIV-TRACE has already been used in this context (Rose et al., 2017a). Secondly, NGS reads could be converted into a smaller collection of individual haplotypes; HIV-TRACE can directly handle multiple sequences per individual, and supports two modes of drawing links between individuals A and B: single linkage (at least one pair of sequences from A and B are closer than $D$ substitutions per site) or complete linkage (all pairs of sequences are closer than $D$ substitutions per site). Lastly, for NGS amplicon data that have been mapped to the reference, HIV-TRACE can be used to quickly compute the distribution of genetic distances between reads from individuals $A$ and $B$; links can then be drawn if the distribution meets a particular condition, for example, at least $X\%$ of read pairs are closer than $D$ substitutions per site.

In addition to extensive applications in the HIV-1 domain, HIV-TRACE has demonstrated utility for other pathogens including acute hepatitis C virus infection (Bartlett et al., 2017; Rose et al., 2017a) and norovirus (Drumright et al., 2014). As any computational tool, HIV-TRACE has advantages and drawbacks. Speed, easy to understand clusters definitions, persistence of clusters when more sequences are added, robustness to recombination, and systematic handling of mixed bases count among the former. The latter include the difficulty in interpreting what variables drive cluster formation and growth, inability to ascertain that any particular link is a direct transmission (i.e., source attribution), and loss of information contained in the phylogenetic tree, including timing (which can be leveraged by molecular clock methods), and branching (which can be taken advantage of by phylodynamics methods). For most rigorous analyses, clusters identified by HIV-TRACE are further analyzed using compute-intensive molecular clock phylogenetic inference tools (e.g., BEAST; Drummond et al. (2012)) (Chaillon et al., 2017; Wertheim et al., 2016, 2017). By using HIV-TRACE first to identify transmission cluster of interest, these more computationally intensive tools can be reserved for smaller, focused analyses.
Acknowledgments

This work was supported in part by grants R01 AI134384 (NIH/NIAID), R01 GM093939 (NIH/NIGMS) and U01 GM110749 (NIH/NIGMS). JOW was funded by an NIH-NIAID Career Development Award (K01AI110181) and the California HIV/AIDS Research Program (ID15-SD-052). We thank N. Lance Hepler for his work on the initial development of HIV-TRACE.

References


Fearnhill, E., Gourlay, A., Malyuta, R., Simmons, R., Ferns, R. B., Grant, P., Nastouli, E., Karnets, I., Murphy, G., Medoeva, A., Kruglov, Y., Yurchenko, A., Porter,


Pond et al. · doi:10.1093/molbev/mstHIV-TRACE

Virus Evol, 2(2): vew031.


Map each read to reference with \texttt{bealign}

\begin{align*}
\text{CTCTTTGGCAACGACCCCTGC}\ldots \\
\text{CTCTTTGGCAACGACCCCTGC}\ldots
\end{align*}

Compute pairwise distances with \texttt{tn93}, reporting those under 0.02

\begin{align*}
\text{Distances between pairs of sequences, up to a threshold} \\
\text{ID1} & \quad \text{ID2} & \quad \text{Distance} \\
\text{B\_US\_KT167882\_1985} & \quad \text{B\_US\_GQ209649\_2002} & \quad 0.0152805 \\
\text{07\_BC\_CN\_KT379202\_2011} & \quad \text{07\_BC\_CN\_KR187686\_2011} & \quad 0.0100482 \\
\text{07\_BC\_CN\_KT379202\_2011} & \quad \text{07\_BC\_CN\_KR187866\_2012} & \quad 0.0110194 \\
\text{01\_AE\_CN\_KX583284\_2014} & \quad \text{01\_AE\_CN\_KU050634\_2007} & \quad 0.0186584 \\
\text{07\_BC\_CN\_KR187686\_2011} & \quad \text{07\_BC\_CN\_KR187866\_2012} & \quad 0.0124912
\end{align*}

Construct and describe the network using \texttt{hivnetworkcsv}

\begin{align*}
\text{Network file (JSON), network visualization} \\
\text{Nodes}: [
\begin{align*}
\text{attributes}: [], \\
\text{baseline}: \text{null}, \\
\text{cluster}: 1, \\
\text{id}: \text{null}, \\
\text{id}: \text{B\_JP\_AB686860\_2010}
\end{align*}
\end{align*}

\textbf{FIG. 1.} A schematic of the HIV-TRACE workflow. For each stage, we show example input and output data, indicate computational complexity, and provide empirical run-times as functions of the number of sequences on the example HIV-1 datasets described in the text. Trend lines show linear fits in the log-log space.
FIG. 2. Effect of genetic distance threshold and ambiguity fraction on network construction. (A) Number of clusters and size of largest cluster across increasing genetic distance thresholds. (B) Number of clusters and size of largest cluster across increasing ambiguity fractions. (C) Largest clusters (≥ 7 nodes) from the San Diego Primary Infection Cohort, inferred with a 0.015 substitutions/site genetic distance threshold and a 0.05 ambiguity fraction on a phylogenetic tree (each cluster has its own color and is shown in bold). (D) Members of the large, artifactual cluster when ambiguity fraction is increased to 1.0 and distances from ambiguities in all sequences are resolved (shown in bold, colored in red). San Diego sequence data are from Little et al. (2014), and phylogeny was inferred using FastTree2 (Price et al., 2010b).
FIG. 3. Visualization of the San Diego Primary Infection Cohort cluster (Little et al., 2014) using hivtrace-viz. Circles without connections and darker borders represent clusters, and their area is proportional to cluster size. 9 of the clusters have been expanded, showing individual nodes (individuals) and edges (putative transmission links). Nodes and clusters are colored by risk factor (this is user selectable, and is obtained from network annotation data); for clusters, the distribution of risk factors is shown as a pie chart. The shape of individual nodes indicates the gender of the corresponding individual.