PHYLOSCANNER: Analysing Within- and Between-Host Pathogen Genetic Diversity to Identify Transmission, Multiple Infection, Recombination and Contamination

Chris Wymant*[1,2], Matthew Hall*[1,2], Oliver Ratmann[2], David Bonsall[1,3,4], Tanya Golubchik[1,4], Mariateresa de Cesare[4], Astrid Gall[5], Marion Cornelissen[6], Christophe Fraser[†1,2], STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration[‡]

[1] Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, UK
[2] Medical Research Council Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, UK
[3] Peter Medawar Building for Pathogen Research, Nuffield Department of Medicine and the NIHR Oxford BRC, University of Oxford, UK
[4] Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, UK
[5] Department of Veterinary Medicine, University of Cambridge, UK
[6] Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center of the University of Amsterdam, Amsterdam, The Netherlands

* Equal contribution
[†] To whom correspondence should be addressed: christophe.fraser@bdi.ox.ac.uk
[‡] Collaboration members listed in full at the end of the text.

Abstract

Pathogen genomics is proving to be a revolutionary tool in epidemiology, medicine and biology. A central feature of pathogen genomics is that different infectious particles (virions, bacterial cells, etc.) within an infected individual may be genetically distinct, with patterns of relatedness amongst infectious particles being the result of both within-host evolution and transmission from one host to the next. Here we present a new software tool, phyloscanner, which analyses pathogen diversity from multiple infected hosts, and so provides unprecedented resolution into the transmission process. Multiply infected individuals are also identified, as they harbour subpopulations of infectious particles that are not connected by within-host evolution, except where recombinant types emerge. Low-level contamination is flagged and removed. We illustrate phyloscanner on both viral and bacterial pathogens, namely HIV-1 sequenced on Illumina and Roche 454 platforms, HCV sequenced with the Oxford Nanopore MinION platform, and *Streptococcus pneumoniae* with sequences from multiple colonies per individual. phyloscanner is available from https://github.com/BDI-pathogens/phyloscanner.

Introduction

The infectious transmission process imposes a hierarchical structure of relatedness on pathogen genomes. The genotype of an individual infectious particle is the result of both within-host evolution and transmission between hosts; a population sample collected from multiple hosts, with multiple genotypes for each host, therefore simultaneously encodes the history of both processes. Despite the existence of many tools for analysing pathogen genomes, none, to our knowledge, are specifically adapted to exploiting this hierarchical genealogical structure.

A central aim of infectious disease epidemiology is the identification of risk factors for transmission. The development of methods that use pathogen genomes to infer transmission events, along with their direction, is therefore a priority. A critical recent insight is that including multiple pathogen genomes per infected individual in such methods makes this inference easier: it is equivalent to the simpler process of inferring ancestry[1]. Specifically, if a pathogen has passed from individual X to individual Y (either directly, or indirectly via unsampled intermediate individuals) then all the pathogen particles sampled from individual Y must be descended from the population of pathogen particles from individual X. Inferring ancestral states is a standard problem in population genetics for which many methods exist; the novel insight is that this standard approach may be used to infer the direction of transmission.

A frequently used approach in molecular epidemiology is to describe patterns of genetic clustering - who is close to whom. However, identifying transmission pairs or clusters without the ability to infer transmission direction - who infected whom - limits our ability to distinguish risk factors for transmission from those for simply acquiring the pathogen. One approach for inferring direction is to augment the sequence data with epidemiological data, and to couple phylogenetic inference with mathematical models of transmission, for example references[2-5]. However, this requires strong assumptions from the model. In addition epidemiological data,

such as dates and location of sampling and reported contacts, are not always available, are subject to their own set of uncertainties and errors, or are sometimes regarded as too sensitive to link to pathogen genetic data.

Using multiple genotypes per host, and exploiting the link between transmission and ancestral reconstruction, therefore promises an alternative and potentially powerful approach to molecular epidemiology. Whilst several studies have used this idea to great effect on an ad hoc basis[6,7], no systematic or automatic tool has been developed for this task.

Once multiple genotypes per host are included in a study, other questions present themselves naturally, for example identifying multiply infected individuals. These may be defined as individuals harbouring pathogen subpopulations resulting from distinct founder pathogen particles (whether transmitted sequentially or simultaneously). Multiple infections may be clinically relevant, for example in the case of Human Immunodeficiency Virus 1 (HIV-1), dual infection is associated with accelerated disease progression[8]. Multiple infections also represent unique opportunities for pathogen evolution, especially for pathogens that recombine. Recombination between divergent strains accelerates the generation of novel genotypes, and so potentially novel phenotypes.

Molecular epidemiology is being transformed by the advent of next-generation sequencing (NGS; also called *high-throughput*) technologies[9]. For many sequencing protocols applied to pathogens with extensive within-host diversity, such as HIV-1 and Hepatitis C Virus (HCV), the NGS output from a single sample can capture extensive within-host diversity. Zanini et al.[10] inferred phylogenies from NGS *reads* - fragments of DNA - in windows along the genome for longitudinally sampled individuals infected with HIV-1, to quantify patterns of within-host evolution over time. Here our focus will be on cross-sectional datasets: by constructing phylogenies from NGS reads from multiple infected individuals at once, within-host and between-host evolution can be resolved.

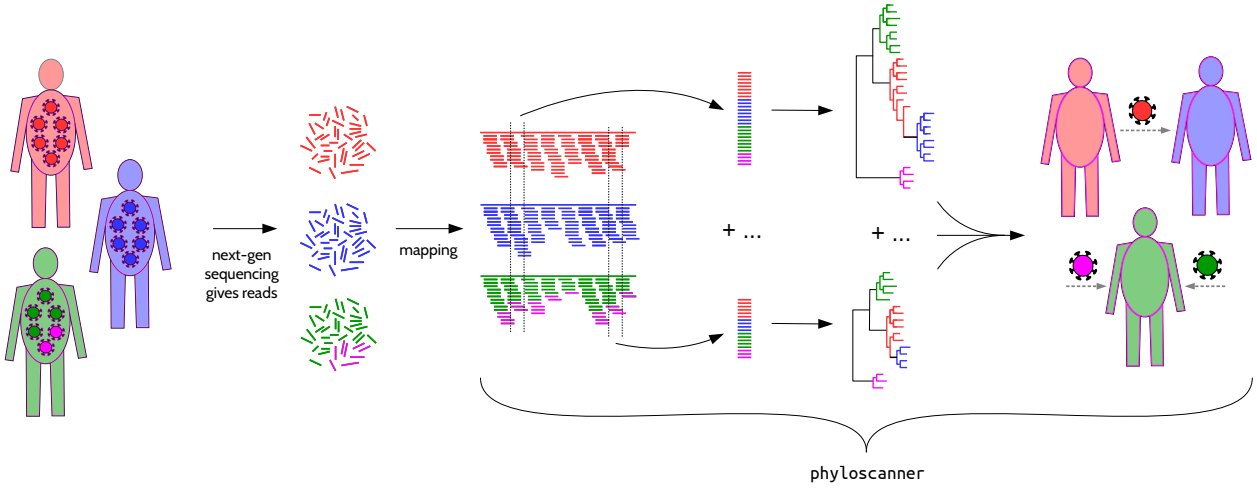
We present phyloscanner: a set of methods implemented as a software package, with two central aims. The first is efficient computation of phylogenies with multiple genotypes per infected host, and the second is analysis of such phylogenies and inference of biologically and epidemiologically relevant properties from a set of related phylogenies. Multiple related phylogenies arise naturally, either by sampling different portions of a genome, or in representing uncertainty in phylogenetic inference (though bootstrapping, or sampling phylogenies from a posterior distribution, for example). phyloscanner automatically performs the following steps:

1. Inference of between and within-host phylogenies from NGS data in multiple windows along the pathogen genome (optionally skipped, if the user has such phylogenies already);
2. Identification and removal of likely contaminant sequences;
3. Quantification of within-host diversity;
4. Identification of multiple infections;
5. Identification of crossover recombination breakpoints in NGS genotypes;
6. Ancestral host-state reconstruction from multiple phylogenies;

7. Identification of transmission events from ancestral host-state reconstructions.

phyloscanner was intended for analysis of two distinct types of sequence data. Firstly for deep sequencing data, in which NGS has produced reads from the population of diverse pathogens represented in the sample. Secondly, for single-genome amplification (SGA), clonal sequencing or bacterial colony picks, whereby laboratory methods are employed to separate the genomes of individual pathogen particles prior to amplification and sequencing. Sequencing with primer IDs[11] may in some cases produce similar results at reduced costs. We also considered haplotype reconstruction[12-14], i.e. bioinformatically inferring different haplotypes represented in the short reads of a mixed sample, but in our hands this approach did not yield satisfactory results (analysis not shown).

With SGA-style data, within- and between-host phylogenies can be directly inferred using standard methods, and therefore phyloscanner is not necessary for step 1 in the process described above. With deep sequencing data, reads for each sample must first be *mapped* (placed at the correct location in the genome); thereafter phyloscanner begins by aligning reads in windows of the genome that are matched across infected individuals, and inferring a phylogeny for each window (Figure 1).



[**Figure 1: phyloscanner schematic for whole-genome deep sequence data.** In this schematic, pathogens are sampled from the population infecting three hosts. NGS deep sequencing produces reads, which are fragments of the genome sequence of one pathogen particle (after amplification if necessary). Mapping to a reference means aligning each read to the appropriate location in the genome; this must be done beforehand, as mapped reads are the inputs to phyloscanner. phyloscanner produces alignments of reads in sliding windows along the genome, automatically adjusting for the fact that the reference may be different for each sample. Phylogenies are inferred for each alignment. These phylogenies are analysed separately, using ancestral host-state reconstruction (i.e. assigning host state to internal nodes, and so colouring the branches), and their information is combined to give biologically and epidemiologically meaningful summaries. For example here, the red individual infected the blue individual, and the green individual has two distinct pathogen strains.]

Results

The best way to illustrate phyloscanner is through examples. We chose five datasets illustrating different uses, pathogens, and sequencing platforms. We describe three in the main text, and two in the Supplementary Information. These are far from systematic samples or population surveys; they are small selections of infected individuals chosen to illustrate the different conclusions that can be drawn using phyloscanner. We leave the application of phyloscanner to large systematic population samples to future work.

Six illustrative HIV-1 infections, sequenced with Illumina MiSeq

We used phyloscanner to analyse data from the BEEHIVE project (*Bridging the Evolution and Epidemiology of HIV in Europe*), in which whole-genome samples from individuals with well-characterised dates of HIV-1 infection are being sequenced, primarily to investigate the viral-molecular basis of virulence[15]. We chose two groups of patients for detailed investigation (presented in this subsection and the next), that together demonstrate interesting features revealed by phyloscanner.

For the BEEHIVE samples, viral RNA was extracted manually from blood samples following the procedure of Cornelissen *et al.*[16]. The RNA was reverse transcribed and amplified using universal HIV-1 primers that define four overlapping amplicons spanning the whole genome, then sequenced using the Illumina MiSeq platform, following the procedure of Gall *et al.* [17,18]. The resulting reads were mapped to a reference constructed for each sample using IVA[19] and shiver[20], producing input analogous to the illustration in Figure 1. See Methods for more detail.

These mapped reads were analysed with phyloscanner using 54 overlapping windows, each 320 base pairs (bp) wide, covering the whole HIV-1 genome (approximately 9200 bp long; the window entirely overlapping the variable V1-V2 loop in the envelope gene was not included due to the richness of insertions and deletions, which leads to poor alignment). To increase phylogenetic resolution and accuracy, we used the phyloscanner options to merge overlapping paired-end reads into single, longer reads, and to delete drug resistance sites[21-23] which are known to be under convergent evolution.

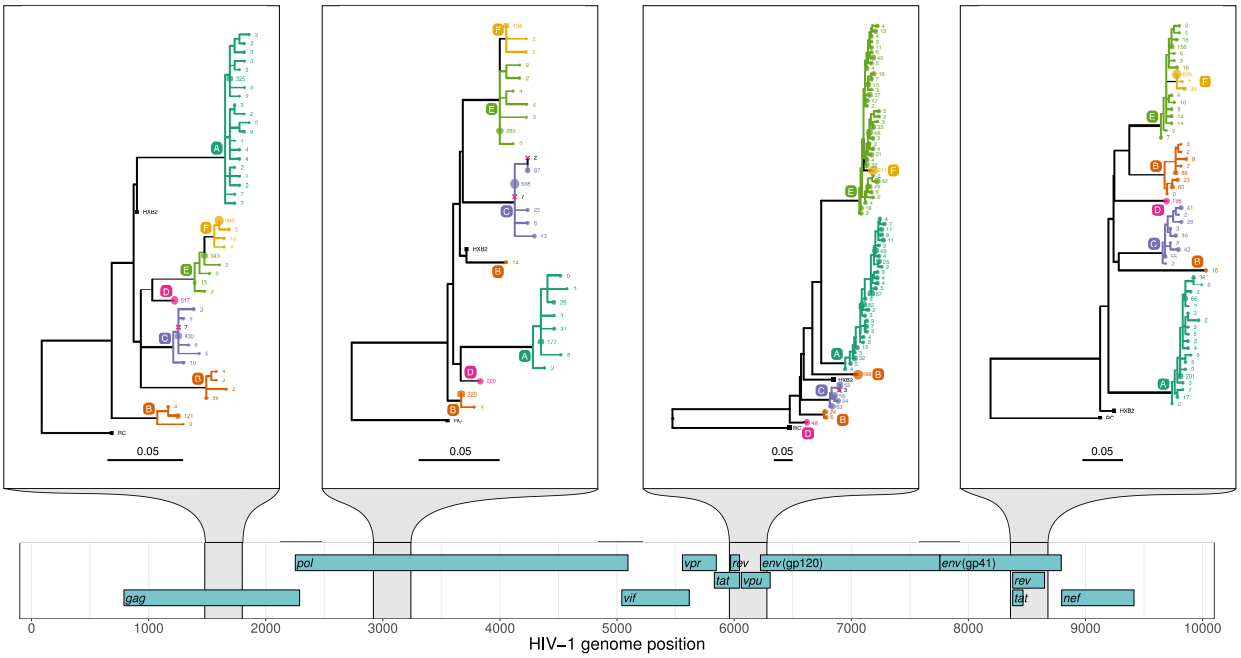
Figure 2 shows the resulting phylogenies for four windows, chosen for clarity when visually inspected. The phylogenies illustrate single infection (patient A), dual infection (patient B), contamination (from the sample of patient C to the sample of patient D) and transmission (from patient E to patient F, possibly via an unsampled intermediate individual). Colouring on each phylogeny illustrates *host subgraphs*. These subgraphs result from ancestral host state reconstruction: they are defined as connected regions of the phylogeny (tips and internal nodes, with the branches joining them) that have all been assigned the same host state (i.e., the patient that virus was in). See supplementary section SI 1 for an explanation of the ancestral state

reconstruction algorithm. Each subgraph can be shown with a solid block of colour corresponding to that patient, uninterrupted by colouring associated with any other patient.

Contamination. Filtering for contamination is an important part of analysis of NGS data. Contamination may be physical contamination of one sample into another, or low-level barcode switching which occurs during the multiplexing and demultiplexing steps which are central to the high throughput of NGS. phyloscanner identifies reads that are likely contaminants by two criteria. We consider likely contaminants to be reads that are identical to reads from another patient, but much less numerous, and reads that form an isolated phylogenetic subgraph, but are not numerous enough to lead to a call of multiple infection. These reads are flagged according to tuneable parameters (which will depend on the precise sample and method used), and blacklisted from further analysis (marked by pink crosses in Figure 2). This two-stage approach means that the donor of contaminant reads need not be present to infer contamination. In general, phylogenetic patterns associated with transmission are distinct from those associated with contamination: the process of transmission is accompanied by within-host evolution in the recipient, whereas contamination is not.

Multiple infections. If the phylogeny and host-state reconstruction are correct, the number of subgraphs a patient has equals the number of founder pathogen particles with sampled descendants (for example if this is 2, a dual infection is inferred). Sampling effects mean that representatives of these multiple infections may not be present in all windows.

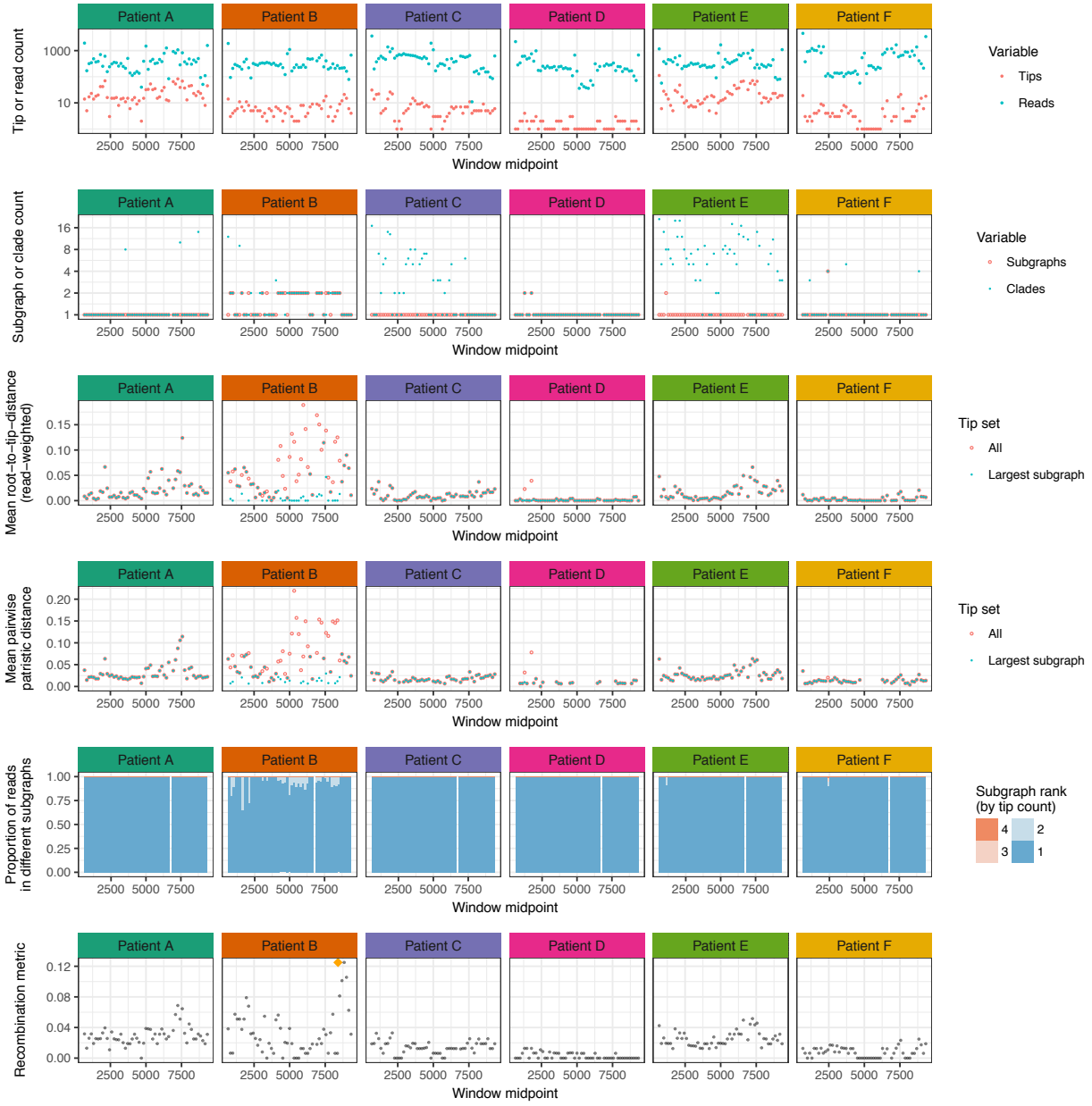
Transmission. Regions of the phylogeny not in any patient's subgraph are black. These regions connect the patients' subgraphs to each other, and so intuitively each must contain one or more transmission events. Nodes in black regions are inferred to exist in unsampled hosts. Where a single branch connects two patients' subgraphs with one ancestral to the other, this branch is black: it may or may not correspond to an unsampled host, i.e. transmission could be direct or indirect.



[**Figure 2: phyloscanner analysis of four illustrative windows of the HIV-1 genome.** A map of the HIV-1 genome is shown at the bottom with the nine genes in the three reading frames. Phylogenies are shown for the four windows highlighted in grey, with scale bars measured in substitutions per site. Tip labels are coloured by patient, as are all nodes assigned to that patient by ancestral reconstruction, and the branches connecting these tips and nodes; a solid block of colour therefore defines a single subgraph for one patient (see main text). The number labelling each tip is the number of times that read was found in the sample, and the size of the circle at each tip is proportional to this count. The count is after merging all identical reads and reads differing by a single base pair (merging similar reads can be done for computational efficiency, or as here, for presentational clarity). External references included for comparison are shown with black squares. One is HXB2; the other, labelled RC, is a subtype C reference used to root each phylogeny. The six patients are labelled A through F. **Single infection:** patient A is a singly infected; all reads from this patient form a single subgraph. **Dual infection:** patient B is inferred to be dually infected, as is apparent by the fact that ancestral reconstruction produces two unconnected subgraphs in each window. **Contamination:** patients C and D are both singly infected, but we infer that some contamination has occurred from C to D. Patient D's sample has a small number of reads that are identical to reads from patient C, but much less numerous. Such reads are removed, but are shown here as crosses in the clade of patient C, for illustrative purposes. **Transmission:** in all four windows shown here, the reads of patient F are seen to be wholly descended from within the subgraph of reads of patient E. We infer that patient E infected patient F, either directly, or indirectly via an unsampled intermediate. Patient F having a single subgraph that is linked to patient E by a single branch indicates that the viral population was bottlenecked down to a single sampled ancestor during transmission.]

Genome-wide summary statistics. In general, a phyloscanner analysis may produce a large number of phylogenies and associated ancestral reconstructions. These are output both as

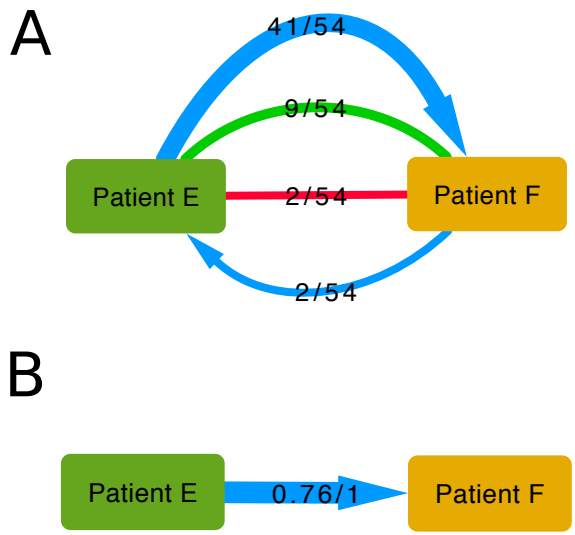
annotated nexus-format files, and as PDF files created with ggtree[24] for rapid visual inspection. Statistics are calculated to summarise the wealth of information in the phylogenies; these are shown for the 6 patients and 54 genomic windows in Figure 3. They include measures of within-host diversity, measures that allow rapid identification of multiply infected individuals, and a metric of recombination (defined in the supplementary section S4).



[**Figure 3 - Summary statistics for six illustrative HIV-1 infected patients.** Each column shows data from a single patient; each row is one or two statistics, plotted along the genome. **Top row:** number of reads, and number of unique reads (corresponding to tips in the phylogeny). **Second row:** the minimum number of clades required to encompass the reads, and the number of subgraphs - defined as connected regions of the phylogeny linked by the host-

state reconstruction. In many windows, though not all, the reads of patient B form two subgraphs: evidence of dual infection. For patients C and E, we see a single subgraph but many clades. This is because of the presence of reads from other patients (D and F, respectively, as seen in Fig. 2) inside what would otherwise be a single clade, turning a monophyletic group into paraphyletic group (which requires splitting in order to form clades). **Third row:** within-host divergence, quantified by mean root-to-tip distance. Defining a patient's subtree as the tree obtained by removing all tips not from this patient, we calculate root-to-tip distances both in the whole subtree and in just the largest subgraph. For patient B, this distinction is substantial due to the very large distance (~0.1 substitutions/site) between the two subgraphs of this dually infected patient. For singly infected patients, divergence may correlate with time since infection. **Fourth row:** for each window, a stacked histogram of the proportion of reads in each subgraph. For patient B, when two subgraphs are present, an appreciable proportion of reads are in the second one (mean 9.7%). The histogram is absent in the window that was excluded by choice. **Bottom row:** a score based on Hamming distance (between 0 and 1) of the extent of recombination in that window. The highest score across all six patients and all windows is indicated with an orange diamond; the reads giving rise to this score are shown in supplementary Figure S7.**]**

To summarise transmission, phyloscanner gives relational information for each pair of patients found to be close or connected by ancestry in a sufficient fraction of windows where both have sequence data. (This naturally allows for inclusion of partial genomes and whole genomes in the same dataset; the former may arise from partially successful amplification or sequencing.) Here we set the threshold on the fraction of windows as 50%, giving just two connected patients: E and F. Summary statistics for their relationship are shown in Figure 4, plotted using Cytoscape[25].

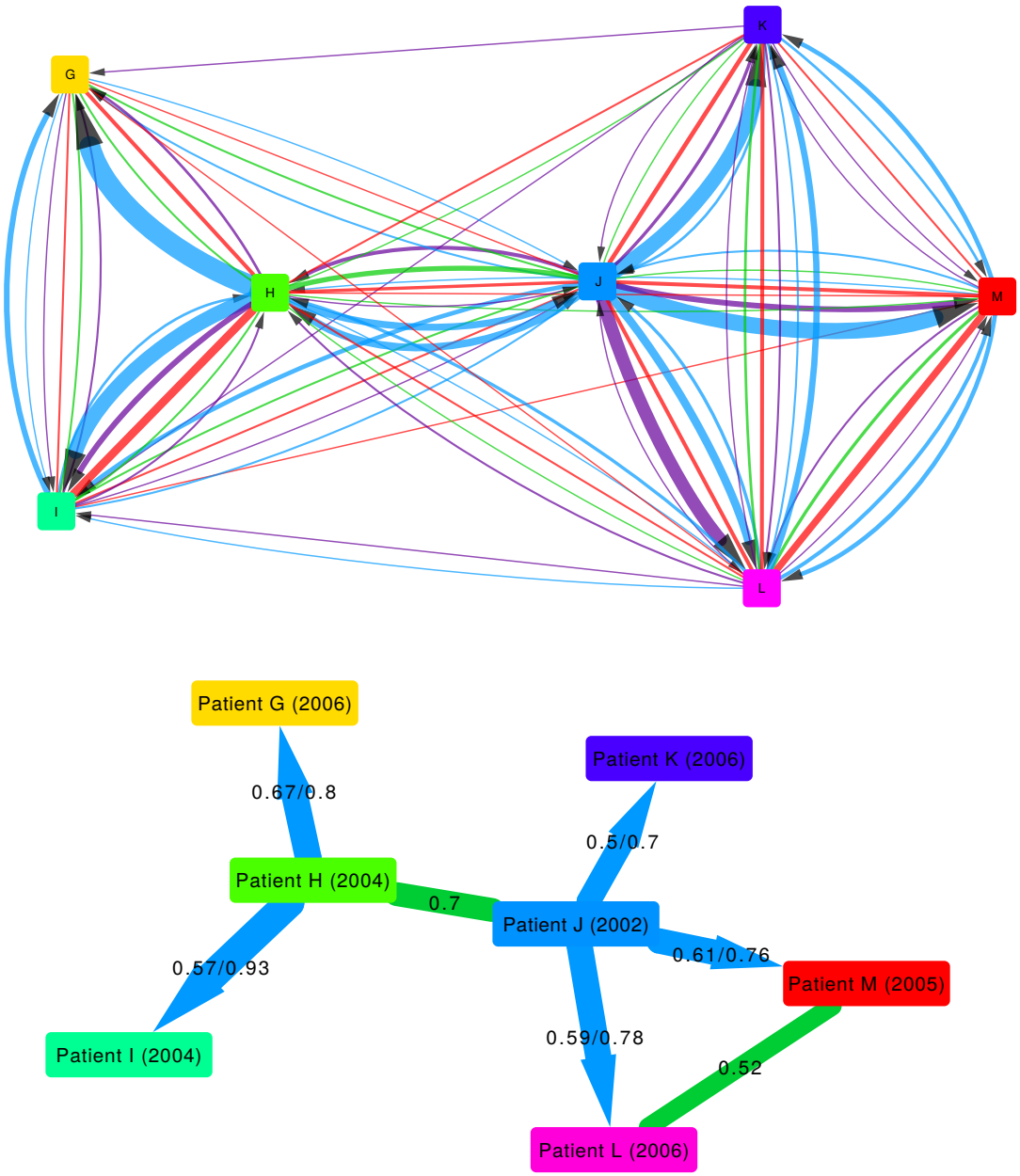


[**Figure 4 - Visual representations of the relationship between two connected patients infected with HIV-1.** The power of phyloscanner in studying transmission events comes from aggregating information over many within- and between-host phylogenies, in this case obtained from different windows of the whole HIV-1 genome. In the top diagram, A, the outcomes from all

54 windows are shown. The top blue arrow shows that in 41 windows, patient E is inferred to be ancestral to patient F, with a single bottleneck. The bottom blue arrow shows that in 2 windows the reverse was true (F ancestral to E). The undirected red line shows that in 2 windows, the patients were linked but the direction of ancestry was not clear. The undirected green line shows that in 9 windows the patient subgraphs were adjacent and close. In no window was a wider bottleneck found, and in no window were the patients distant and unlinked. (See supplementary section SI 1 for more details on these categories.) These relational data are further summarised in the bottom diagram, B, showing a single directed arrow. The first number indicates the proportion of windows supporting transmission in the direction of the arrow, and the second number indicates the proportion of windows supporting transmission in either direction.]

Resolving the transmission pathway within a HIV-1 phylogenetic cluster

To illustrate the resolution into the transmission process that can be obtained by phyloscanner, we chose a set of 7 patients from the BEEHIVE study that were found to be closely connected in a transmission network (Fig. 5). 3 of the patients' samples were sequenced with Illumina MiSeq and 4 with Illumina HiSeq; the resulting reads were processed and mapped using IVA and shiver as previously, with the mapped reads given as input to phyloscanner. phyloscanner summarises all the pairwise relationships between individuals in each window (Figure 5A), suggesting a complex network. However, we find that when we focus on the most likely inferences of source attribution (Figure 5B), phyloscanner resolves a complex set of pairwise relationships into a coherent most-likely transmission network, that is consistent with the years of seroconversion. An exception is the triangle connecting Patients J, L and M, which is a concrete example that a connection does not imply direct transmission.



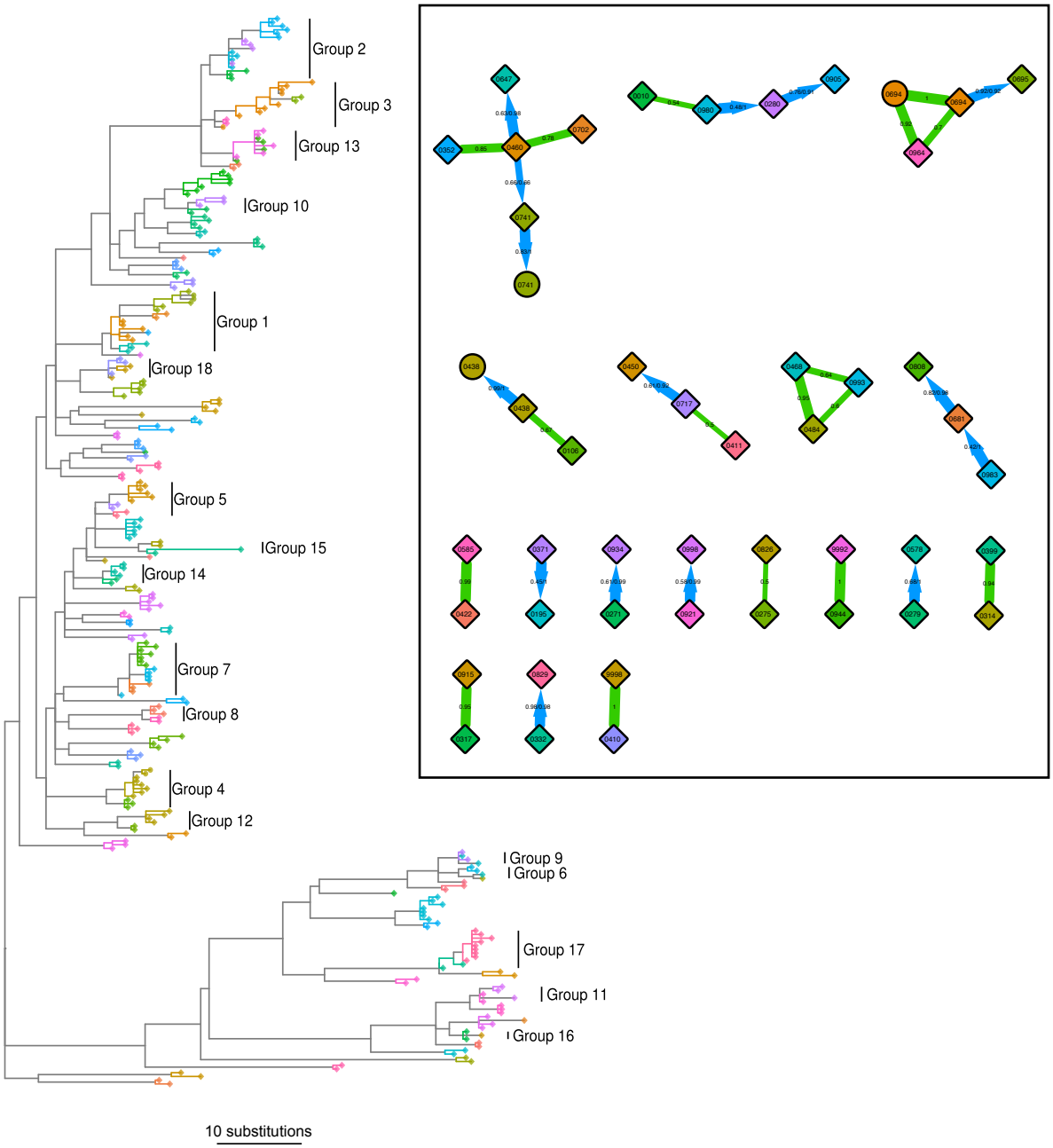
[**Figure 5 - The relationship between 7 patients infected with HIV-1.** The colouring and numbers on the arrows connecting patients are as in Figure 4; in addition, the lower diagram here contains undirected green lines as well directed blue lines. These green lines suggest that the pair are close in the transmission network but with unknown transmission direction; the single number on the line indicates the proportion of windows supporting this. The known or estimated year of infection is shown in parentheses after each patient's label.]

HIV-1 sequenced with Roche 454, and HCV sequenced with Oxford Nanopore MinION

phyloscanner can be used on different viruses and NGS platforms. A subset of patients from the BEEHIVE study were also sequenced using the Roche 454 platform; results from their analysis with phyloscanner are in Supplementary Information section SI 2. We also analysed Hepatitis C viral sequence data obtained from patients in the BOSON study[26]; this data and the results are discussed in Supplementary Information section SI 3.

Multiple colony picks per carrier of *S. pneumoniae*

phyloscanner's analysis of sets of phylogenies need not be restricted to windows of the genome with deep sequencing data: it can also be applied to data sets where within-host diversity is captured by SGA or sequences from multiple colony picks per individual. We illustrate this approach with the *S. pneumoniae* data of Croucher et al.[27], specifically the BC1-19F cluster. This dataset consists of 286 sequences from 92 individuals carrying the bacterium (with multiple colonies per carrier). These were sequenced with Illumina HiSeq, though for SGA data sequencing platform is largely irrelevant to interpretation, since each sequenced sample should not contain any real within-sample diversity by design. Genomes were processed with Gubbins[28] to remove substitutions likely to have been introduced by recombination. As each of these sequences is a whole genome (unlike the short reads produced by NGS), we did not cut them into windows of the genome to be analysed separately, which would needlessly discard information on the linkage of sites across the genome. Instead, we represented phylogenetic uncertainty by generating a posterior set of 100 phylogenies using MrBayes[29] and analysed these with phyloscanner. Ancestral state reconstruction was performed on each posterior phylogeny independently, relationships between carriers were identified, and the results summarised over the entire set. In each phylogeny carriers were inferred as being linked if the minimum patristic distance between two nodes from the subgraphs associated with each was less than 7 substitutions and they were categorised as adjacent (explained in Supplementary Information section SI 1.5). This distance threshold was selected to demonstrate the method as it picked out obvious clades in the phylogeny as groups, and was not chosen to imply direct transmission. Retaining such relationships where they existed in at least 50% of posterior phylogenies revealed 18 separate groups of carriers whose bacterial strains were closely related (see Fig. 6).



[Figure 6 - Phylogeny and relationships between *S. pneumoniae* carriers. The phylogeny shown is the MrBayes summary tree. Tip shapes are coloured by carrier, with mother and infant pairs sharing the same colour; diamonds represent infants and circles mothers. All nodes assigned to a carrier by ancestral reconstruction, and the branches connecting these tips and nodes, are given the same colour as that carrier's tips; a solid block of colour therefore defines a single subgraph for one carrier (see main text). Regions of the phylogeny not in any carrier's subgraph are grey. These regions connect carriers' subgraphs to each other, and so each must contain one or more transmission events. The carrier relationship diagram (inset) displays the

relationships between the carriers in 18 identified groups, in the same fashion as in Figures 4 and 5, except that here the numbers represent the proportion of phylogenies from the posterior set, rather than the proportion of genomic windows in which both patients have sequence data. The clades representing these 18 groups are labelled in the phylogeny.]

Discussion

Improving our understanding of the transmission of pathogens is valuable for identifying epidemiological risk factors - the first step for targeting public health interventions for efficient impact. Phylogenetic analysis of one pathogen sequence per infected individual may identify clusters of similar sequences that are expected to be close in a transmission network. However, nothing is learned about the direction of transmission within the network. Indeed it may be that none of the individuals transmitted the pathogen to anyone else, and they were all infected by a common individual who was not sampled. Through automatic fitting of evolutionary models to within- and between-host genetic sequence data, phyloscanner enhances resolution into the pathogen transmission process. An evidence base is built up by analysing many phylogenies, notably through consideration of NGS reads in windows along the pathogen's genome. The relationship between infected individuals is no longer quantified by a single number summarising closeness, but by a rich set of data resulting from ancestral host-state reconstruction for each phylogeny.

Romero-Severson *et al.*[1] demonstrated the utility of parsimony for the assignment of ancestral hosts to internal nodes in a phylogeny containing many tips from two infected individuals, for simulated HIV-1 data. We have continued with this approach, developing it for suitability for real sequence data from many infected individuals. In particular we allow for (i) multiple infections, (ii) contamination, and (iii) the presence of unsampled hosts in the tree. Identifying cases (i) and (ii) is highly valuable in its own right[8]. Details of several such parsimony algorithms, available for use in phyloscanner, are presented in the supplementary section SI 1. Parsimony has the advantage that a reconstruction can be completed in reasonable computational time even for phylogenies with tens of thousands of tips. Other methods of annotating internal nodes with hosts could also be suitable and may be added to the package in future.

Great care must be taken to correctly interpret the ancestry of pathogens infecting individuals. Even if ancestry were established beyond any doubt, individual X's pathogen being ancestral to individual Y's pathogen does not imply that X infected Y: the pathogen could have passed through unsampled intermediate hosts. Nevertheless the ancestry does provide valuable epidemiological information, as X has been identified as a transmitter. Finding likely transmitters in a large population cohort would allow risk factors to be identified and quantified.

Furthermore, inference of ancestry is itself subject to uncertainty. The inference of ancestry depends on the correct rooting of the phylogeny, in order that the direction in which evolution proceeded over time is known. Molecular clock analyses (such as implemented in TempEst[30]) can aid correct rooting when the sampling dates of the tips of the phylogeny are known.

The relationships between infected individuals are inferred by phyloscanner across many phylogenies, for example those constructed from NGS reads in windows along the pathogen genome. By analysing many phylogenies, phyloscanner mitigates the effect of random error - any error that is independent in each phylogeny. We therefore give greater credibility to those relationships observed many times than to those observed only once. However, systematic error may arise, for example, due to different patients being sampled at different stages of infection, with different amounts of within-host diversity to analyse[1]. Given uncertainties in any individual assignment, we recommend phyloscanner for population-level analyses, rather than focussing on isolated transmission events (as we have done here, for simplicity in explaining the method).

Whilst our emphasis has been on extracting broad-brush information from the rich within-and-between host phylogenies, these phylogenies contain more information that could be used in future research. A specific example is that by resolving the transmission event at a finer level of genetic detail, it is possible to identify which pathogen genotypes are typically transmitted and which ones are not, with potential relevance for vaccine design.

By providing a tool for automatic phylogenetic analysis of NGS deep sequencing data, or multiple genotypes per host generated by other means, we aim to simplify identification of transmission, multiple infection, recombination and contamination across pathogen genomics.

Methods

Generation and assembly of the BEEHIVE Illumina data

Viral RNA was extracted manually from blood samples following the procedure of Cornelissen *et al.*[16]. RNA was amplified and sequenced according to the protocol of Gall *et al.* [17,18]. Briefly, universal HIV-1 primers define four amplicons spanning the whole genome. 5 µl of amplicon I was pooled with 10 µl each of amplicons II–IV. Libraries were prepared from 50 to 1000 ng DNA as described in Quail et al. [31,32], using one of 192 multiplex adaptors for each sample. Paired-end sequencing was performed using an Illumina MiSeq instrument with read lengths of length 250 or 300 bp, or in the 'rapid run mode' on both lanes of a HiSeq 2500 instrument with a read length of 250 bp.

For each sample, the reads were assembled into contigs using the *de novo* assembler IVA. The reads and contigs were processed using shiver as described previously[20]. In summary: non-HIV contigs were removed based on a BLASTN[33] search against a set of standard whole-genome references[34]**.** Remaining contigs were corrected for assembly error then aligned to the standard reference set using MAFFT[35]. A tailored reference for mapping was then constructed for each sample using the contigs, with gaps between contigs filled by the corresponding part of the closest standard reference. The reads were trimmed for adapters, PCR primers and low-quality bases using Trimmomatic[36] and fastaq (https://github.com/sanger-pathogens/Fastaq). Contaminant reads were removed based on a BLASTN search against the non-HIV contigs and

the tailored reference. The remaining reads were then mapped to the tailored reference using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0).

The phyloscanner Method

For application of phyloscanner to deep sequence NGS data, the input is a set of files each containing the reads from one sample mapped to a reference (in BAM format[37]), and a choice of genomic windows to examine. A sensible choice of windows would normally tile the whole genome, perhaps skipping regions that are rich in insertions and deletions (leading to poor sequence alignment). Windows should be wide enough to capture appreciable within-host diversity, but short enough for some reads to fully span them; options in the code help to inform the user's choice. There is no lower limit to the length of reads given as input, however as read length decreases, phylogenetic resolution will suffer. phyloscanner determines the correspondence between windows in different BAM files by aligning the mapping references in the BAM files. Using the same reference for mapping all samples would negate the need for this step, but it is of paramount importance to tailor the reference to each sample before mapping to minimise biased loss of information[20]. For each window in each BAM file, all reads (or inserts, if reads are paired and overlapping) fully spanning the window are extracted using pysam (https://github.com/pysam-developers/pysam) and trimmed to the window edges, then identical reads are collapsed to a single read, giving a set of unique reads each with an associated count (i.e. the number of reads with identical sequence). A simple metric of recombination is calculated by maximising, over all possible sets of three sequences and all possible recombination crossover points, the extent to which one of the three sequences resembles one of the other two sequences more closely on the left and resembles the other sequence more closely on the right. Further detail is provided in the supplementary section SI 4. In each window, each sample's set of unique reads is checked against every other sample's set, with exact matches flagged to warn of between-sample contamination in the analysed dataset; all unique reads are then aligned with MAFFT, and a phylogeny is inferred with RAxML[38].

phyloscanner contains many options to customise processing and maximise the information extracted from reads and phylogenies. Standard reference genomes can be included with the reads for comparison. User-specified sites can be excised to mitigate the effect of known sites under selection on phylogenetic inference. Greater faith can be placed in the reads by trimming low-quality ends and wholly discarding reads that are low-quality, improperly paired, or rare. Reads in the same sample that differ from each other by less than a specified threshold can be merged into a single read to increase the speed of downstream processing. Overlapping paired reads can be merged into a single longer read for greater phylogenetic resolution. Every option of RAxML can be passed as an option to phyloscanner, for example specifying the evolutionary model to be fitted, or multithreading.

Optionally the user may skip inference of phylogenies from files of mapped reads, and instead directly provide as input within- and between-host phylogenies generated by any other method.

Each phylogeny analysed is annotated with a reconstruction of the transition process using a modified maximum-parsimony approach to assign internal nodes to hosts or to an extra "unsampled" state to represent portions of the ancestry in which a lineage was outside the set of hosts from which the sequences were derived. The results can be represented as a visualisation of the partial pathogen transmission tree by the process of 'collapsing' each subgraph (i.e. each set of adjacent nodes with the same reconstructed host; see supplementary Fig. S3). These are then analysed to identify relationships between each pair of infected individuals, in the following categories:

1. Minimum distance: what is the smallest patristic distance between a phylogeny node assigned to one host and a node assigned to the other?
2. Adjacency: is there a path on the phylogeny that connects the two individuals' subgraphs without passing through a third individual?
3. Topology: how are the regions from each individual arranged with respect to each other? (See supplementary Fig. S4.)

Combinations of these properties can be used to develop criteria which identify individuals who are closely linked in the transmission chain. For example, two individuals that are adjacent and within a suitable distance threshold are likely to be either a transmission pair, or infected via a small number of unsampled intermediaries. The nature of the topological relationship between them may suggest a direction of transmission, or be equivocal.

An individual having multiple subgraphs suggests multiple infection, with the ancestor node of each subgraph inferred to be a distinct founder pathogen particle (the ancestor of that sampled subpopulation). Note that *co-infection* commonly refers to the case of simultaneous transmissions of multiple founder particles from a single donor with a diverse pathogen population, and *super-infection* to sequential transmissions from different donors; we do not distinguish these cases. It is difficult to distinguish a dual infection from a sample that has been contaminated by another sample not present in the current data set (i.e. where contamination is not visible as exact duplication of another individual's read). For NGS data we make the distinction in each phylogeny based on thresholds on read counts: outside of the subgraph containing the greatest number of reads, any additional ('minor') subgraph is designated as contamination and ignored if the number of reads it contains is below an absolute threshold, or below a threshold relative to the read count in the largest subgraph. Minor subgraphs with read counts exceeding both thresholds are kept, providing evidence for the presence of multiple distinct subpopulations in that genomic window.

The phyloscanner Code

phyloscanner is freely available at https://github.com/BDI-pathogens/phyloscanner. It is written in Python and R, but is run from the command line so that no knowledge of either language is required. Inference of within- and between-host phylogenies from BAM-format mapped reads is achieved with a single command of the form

```
phyloscanner_make_trees.py  ListOfBamsAndRefs.csv  --windows  1,300,301,600,...
```

where ListOfBamsAndRefs.csv lists the BAM files to be analysed and the fasta-format references to which the reads were mapped, and the --windows flag above specifies analysis of the genomic windows with coordinates 1-300, 301-600, ...
Analysis of those trees is achieved with a single command of the form
phyloscanner_analyse_trees.R TreeFiles OutputLabel [choice of ancestral state reconstruction]

Included with the code is simple simulated HIV-1 data for ease of immediate exploration of phyloscanner. Within-host evolution was simulated using SeqGen[39]; each resulting sequence was then converted into error-free fragments that were mapped back to the founding sequence, giving bam files suitable as input for phyloscanner.

Running phyloscanner on the six HIV-1 samples presented in the first results section took 18 minutes on one core of a standard laptop.

Acknowledgments

We thank Katrina Lythgoe for helpful discussions, and Céline Christiansen-Jucht for comments on the manuscript. This work was funded by ERC Advanced Grant PBDR-339251. We acknowledge funding from Bill & Melinda Gates Foundation through PANGEA-HIV. The STOP-HCV Consortium is funded by a grant from the Medical Research Council (MR/K01532X/1). This work used the computing resources of the UK MEDical BIOinformatics partnership - aggregation, integration, visualisation and analysis of large, complex data (UK MED-BIO) which is supported by the Medical Research Council [grant number MR/L01632X/1].

The BEEHIVE Collaboration

Jan Albert, Margreet Bakker, Norbert Bannert, Ben Berkhout, Daniela Bezemer, François Blanquart, Marion Cornelissen, Jacques Fellay, Katrien Fransen, Christophe Fraser, Astrid Gall, Annabelle Gourlay, M. Kate Grabowski, Barbara Gunsenheimer-Bartmeyer, Huldrych F. Günthard, Matthew Hall, Mariska Hillebregt, Paul Kellam, Pia Kivelä, Roger Kouyos, Oliver Laeyendecker, Kirsi Liitsola, Laurence Meyer, Swee Hoe Ong, Kholoud Porter, Peter Reiss, Matti Ristola, Ard van Sighem, and Chris Wymant.

Acknowledged contributors to the cohorts in the BEEHIVE Collaboration are listed in supplementary section SI 5.

The STOP-HCV Consortium

Eleanor Barnes, Jonathan Ball, Diana Brainard, Gary Burgess, Graham Cooke, John Dillon, Graham R Foster, Charles Gore, Neil Guha, Rachel Halford, Cham Herath, Chris Holmes, Anita Howe, Emma Hudson, William Irving, Salim Khakoo, Paul Klenerman, Diana Koletzki, Natasha Martin, Benedetta Massetto, Tamyo Mbisa, John McHutchison, Jane McKeating, John

McLauchlan, Alec Miners, Andrea Murray, Peter Shaw, Peter Simmonds, Chris C A Spencer, Paul Targett-Adams, Emma Thomson, Peter Vickerman, and Nicole Zitzmann.

The Maela Pneumococcal Collaboration

Stephen D. Bentley, Claire Chewapreecha, Nicholas J. Croucher, Simon Harris, Jukka Corander, David Goldblatt, Julian Parkhill, Francois Nosten, Claudia Turner, and Paul Turner.

Competing Interests

- AJG participated in an advisory board meeting for ViiV Healthcare in July 2016.
- KP is a member of the Viiv 'Dolutegravir' Advisory Board and Viiv 'Data and Insights: Standardisation in Measuring and Collecting Care Continuum Data' Advisory Board.
- HG reports receipt of grants from the Swiss National Science Foundation, Swiss HIV Cohort Study, University of Zurich, Yvonne Jacob Foundation, and Gilead Sciences; fees for data and safety monitoring board membership from Merck; consulting/advisory board membership fees from Gilead Sciences; and travel reimbursement from Gilead, Bristol-Myers Squibb, and Janssen.
- PR through his institution has received independent scientific grant support from Gilead Sciences, Janssen Pharmaceuticals Inc, Merck & Co, Bristol-Myers Squibb, and ViiV Healthcare; he has served on scientific advisory boards for Gilead Sciences and ViiV Healthcare and on a data safety monitoring committee for Janssen Pharmaceuticals Inc, for which his institution has received remuneration.

References

1. Romero-Severson, E. O., Bulla, I. & Leitner, T. Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences* **113,** 2690–2695 (2016).
2. Volz, E. M. & Frost, S. D. W. Inferring the Source of Transmission with Phylogenetic Data. *PLoS Comput Biol* **9,** e1003397 (2013).
3. Jombart, T. *et al.* Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Comput Biol* **10,** e1003457 (2014).
4. Hall, M., Woolhouse, M. & Rambaut, A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLoS Comput Biol* **11,** e1004613 (2015).
5. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks. *Molecular Biology and Evolution* **34,** 997 (2017).
6. Numminen, E. *et al.* Two-phase importance sampling for inference about transmission trees. *Proceedings of the Royal Society B: Biological Sciences* **281,** 20141324–20141324 (2014).
7. Worby, C. J. *et al.* Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Ann. Appl. Stat.* 395–417 (2016).
8. Cornelissen, M. *et al.* HIV-1 Dual Infection Is Associated With Faster CD4+ T-Cell Decline in a Cohort of Men With Primary HIV Infection. *Clinical Infectious Diseases* **54,** 539 (2012).

9. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17,** 333–351 (2016).
10. Zanini, F. *et al.* Population genomics of intrapatient HIV-1 evolution. *eLife* **4,** e11282 (2015).
11. Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A. & Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences* **108,** 20166–20171 (2011).
12. Zagordi, O., Bhattacharya, A., Eriksson, N. & Beerenwinkel, N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* **12,** 119 (2011).
13. Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N. & Roth, V. HIV Haplotype Inference Using a Propagating Dirichlet Process Mixture Model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **11,** 182–191 (2014).
14. Töpfer, A. *et al.* Viral Quasispecies Assembly via Maximal Clique Enumeration. *PLoS Comput Biol* **10,** 1–10 (2014).
15. Fraser, C. *et al.* Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary Perspective. *Science* **343,** (2014).
16. Cornelissen, M. *et al.* From clinical sample to complete genome: Comparing methods for the extraction of HIV-1 RNA for high-throughput deep sequencing. *Virus Research* (2016). doi:http://dx.doi.org/10.1016/j.virusres.2016.08.004
17. Gall, A. *et al.* Universal Amplification, Next-Generation Sequencing, and Assembly of HIV-1 Genomes. *Journal of Clinical Microbiology* **50,** 3838–3844 (2012).
18. Gall, A., Morris, C., Kellam, P. & Berry, N. Complete Genome Sequence of the WHO International Standard for HIV-1 RNA Determined by Deep Sequencing. *Genome Announcements* **2,** (2014).
19. Hunt, M. *et al.* IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv120
20. Wymant, C. *et al.* Easy and Accurate Reconstruction of Whole HIV Genomes from Short-Read Sequence Data. (2016). doi:10.1101/092916
21. Johnson, V. A. *et al.* 2011 update of the drug resistance mutations in HIV-1. *Top Antivir Med* **19,** 156–164 (2011).
22. Gatanaga, H. *et al.* Amino Acid Substitutions in Gag Protein at Non-cleavage Sites Are Indispensable for the Development of a High Multitude of HIV-1 Resistance against Protease Inhibitors. *Journal of Biological Chemistry* **277,** 5952–5961 (2002).
23. Wensing, A. M. *et al.* 2015 Update of the Drug Resistance Mutations in HIV-1. *Top Antivir Med* **23,** 132–141 (2015).
24. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8,** 28–36 (2017).
25. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13,** 2498–2504 (2003).
26. Foster, G. R. *et al.* Efficacy of Sofosbuvir Plus Ribavirin With or Without Peginterferon-Alfa in Patients With Hepatitis C Virus Genotype 3 Infection and Treatment-Experienced Patients With Cirrhosis and Hepatitis C Virus Genotype 2 Infection. *Gastroenterology* **149,** 1462–1470 (2015).
27. Croucher, N. J. *et al.* Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. *PLOS Biology* **14,** 1–42 (2016).
28. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43,** e15 (2015).
29. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model

Choice Across a Large Model Space. *Systematic Biology* **61,** 539 (2012).
30. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution* **2,** vew007 (2016).
31. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat Meth* **5,** 1005–1010 (2008).
32. Quail, M. A. *et al.* Optimal enzymes for amplifying sequencing libraries. *Nat Meth* **9,** 10–11
33. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410 (1990).
34. Kuiken, C. *et al.* HIV Sequence Compendium 2012. *Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico. LA-UR-12-24653* (2012).
35. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30,** 3059–3066 (2002).
36. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
37. Li, H. *et al.* The Sequence Alignment/Map (SAM) Format and SAMtools. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp352
38. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312 (2014).
39. Rambaut, A. & Grass, N. C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* **13,** 235 (1997).