Transmission tree reconstruction by augmentation of internal phylogeny nodes

Matthew Hall

Li Ka Shing Institute for Health Information and Discovery, University of Oxford

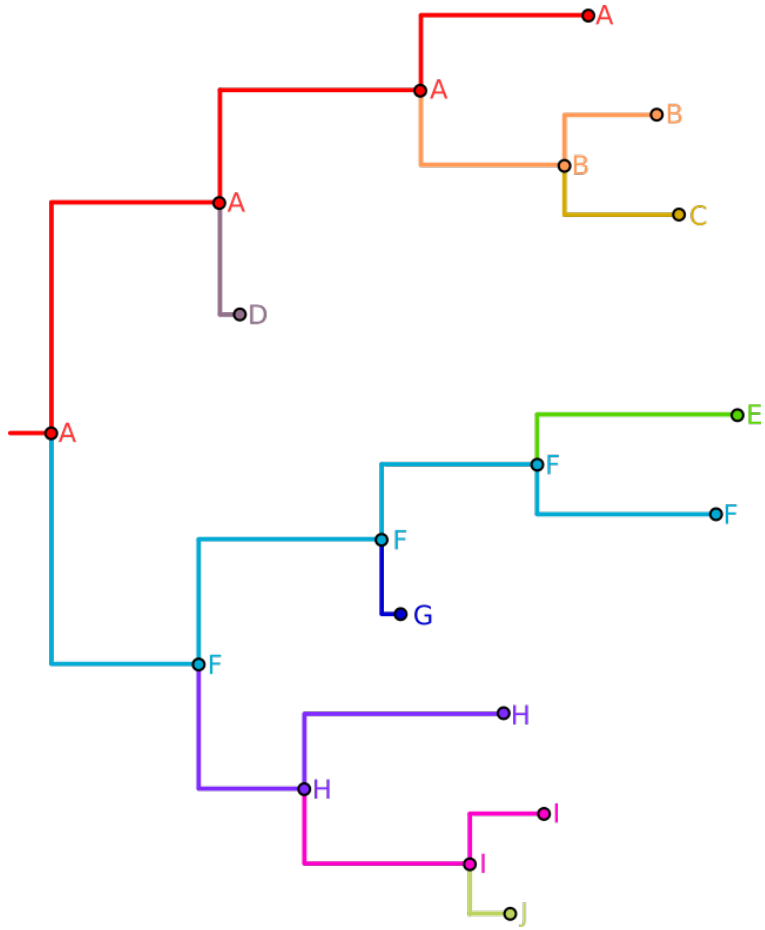
February 2017

The relationship of the phylogeny to the transmission tree

- Let $\mathcal{T}$ be a time-tree (rooted, with branch lengths in units of time).
- Let V be its node set of size n.
- Suppose the isolates at the tips of $\mathcal{T}$ come from a set of H of hosts.
- Initial assumptions:
  - Complete sampling of the epidemic since the TMRCA
  - No superinfection or reinfection
  - Transmission is a complete bottleneck

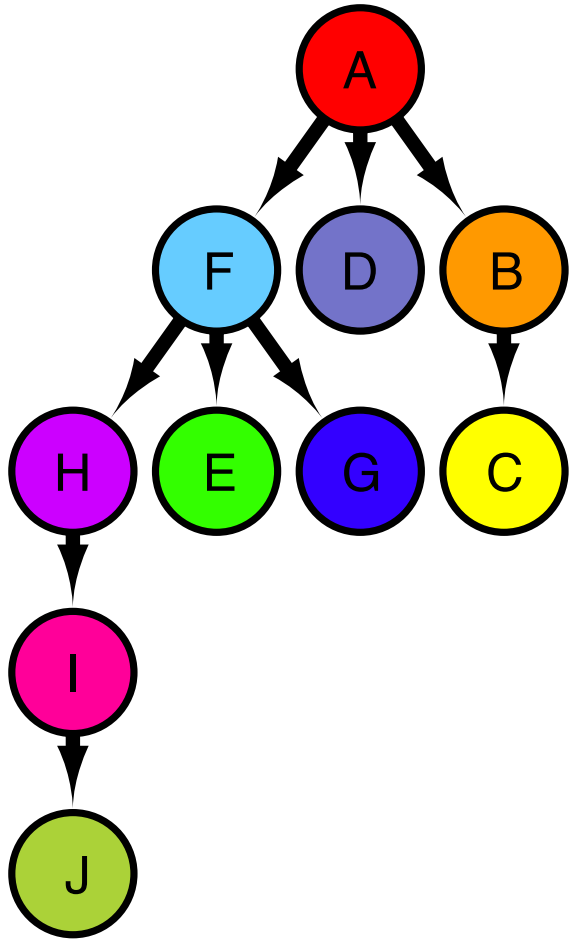
The relationship of the phylogeny to the transmission tree

- The transmission tree $\mathcal{N}$ (a DAG whose nodes are the members of H, depicting which host infected which other) can be represented by a map $d : V \to H$ taking each node to a host (tips to the host they were sampled from).
- Visualised by collapsing the nodes in the preimage of each $h \in H$ under d to a single node.



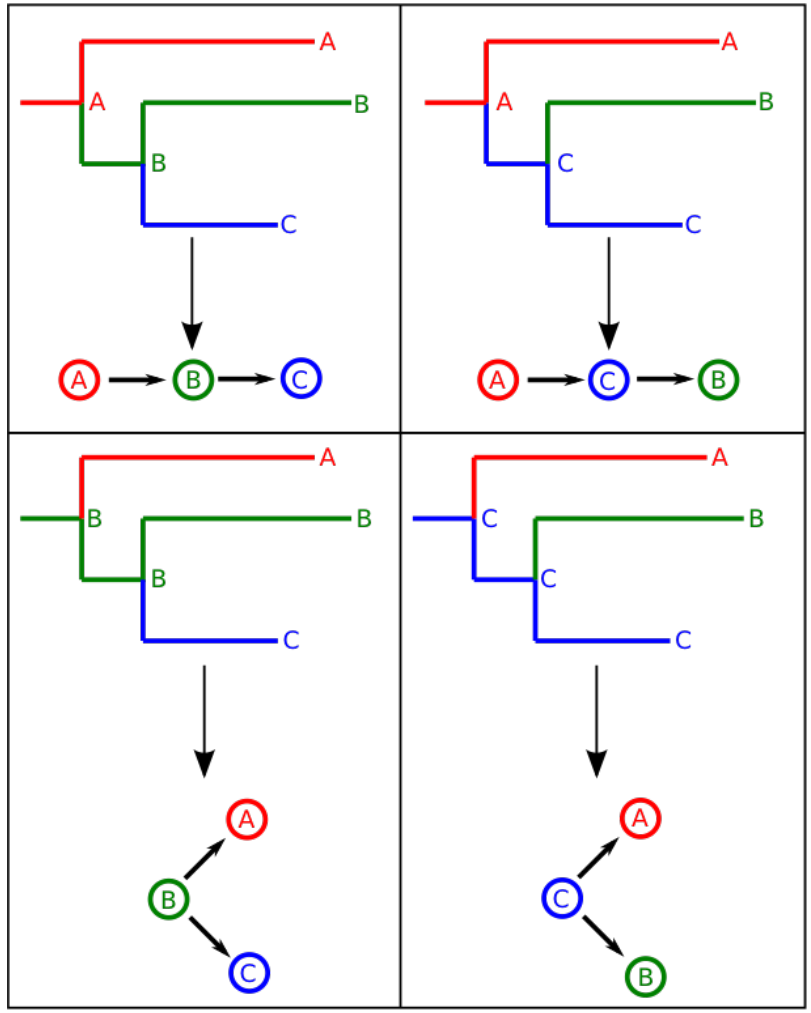
The relationship of the phylogeny to the transmission tree

- The transmission tree $\mathcal{N}$ (a DAG whose nodes are the members of H, depicting which host infected which other) can be represented by a map $d : V \to H$ taking each node to a host (tips to the host they were sampled from).
- Visualised by collapsing the nodes in the preimage of each $h \in H$ under d to a single node.



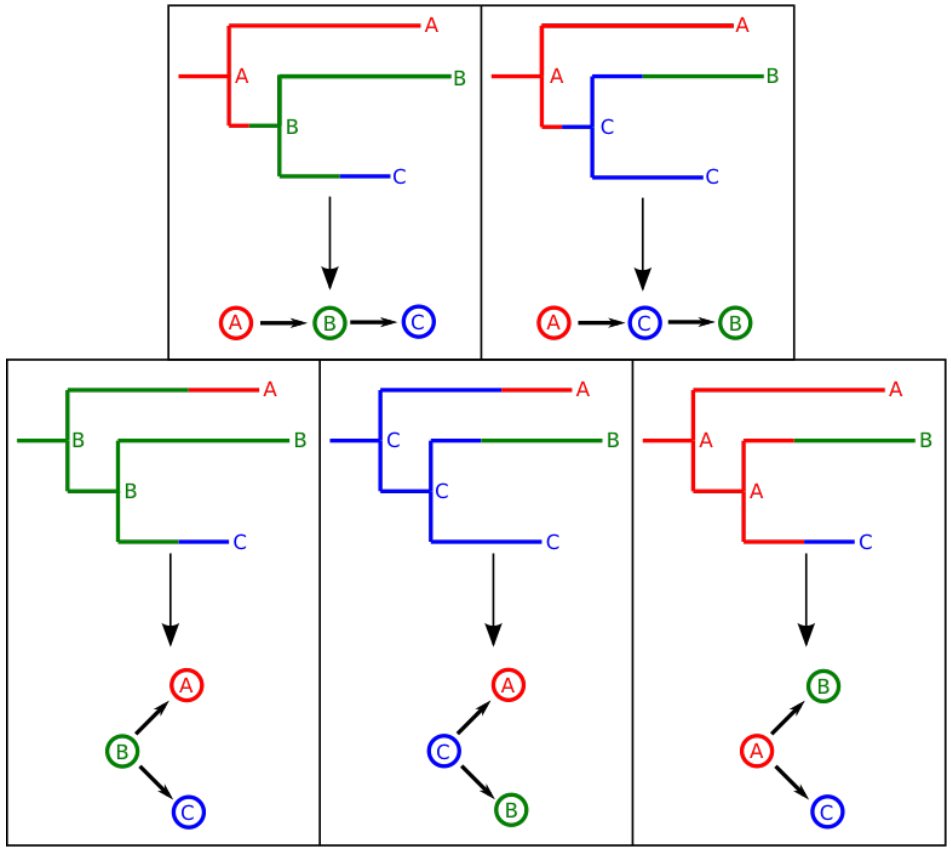
The simplest version

- Assume that the phylogeny and transmission tree coincide; internal nodes are transmission events.
- This implies no within-host diversity and necessitates no more than one tip per host.
- If n is internal with children nC_1 and nC_2, then either $d(n) = d(nC_1)$ or $d(n) = d(nC_2)$.
- Trivially 2^{n-1} transmission trees for a fixed $\mathcal{T}$.



Within-host diversity

- If within-host diversity is assumed then internal nodes are coalescences of two lineages within a host.
- The subgraph induced by the preimage of d for any host must be connected.
- An extra set of parameters q represent the infection times.
- **Question:** How many transmission trees for a fixed $\mathcal{T}$? (Depends on the topology.)
  - With one tip per host?
  - With ≥ 1 tip per host? (Sometimes 0.)



Simultaneous MCMC reconstruction of phylogeny and transmission tree

- In either case we get an (injective but not surjective) map z from the set of possible ds to the space of transmission trees.
- Thus an MCMC method that samples from the posterior distribution of phylogenies with internal node augmentation obeying either set of rules simultaneously samples from the posterior distribution of transmission trees.
- Not only a method for reconstructing $\mathcal{N}$, but a population model (tree prior) for reconstruction of $\mathcal{T}$ that is more realistic for an outbreak than the standard unstructured coalescent models.

Decomposition

- Let S be the sequence data and ϕ the various model parameters.
- Without within-host diversity:

$$p(\mathcal{T}, d, \phi|S) = \frac{p(S|\mathcal{T})p(\mathcal{T}, d|\phi)p(\phi)}{p(S)}$$

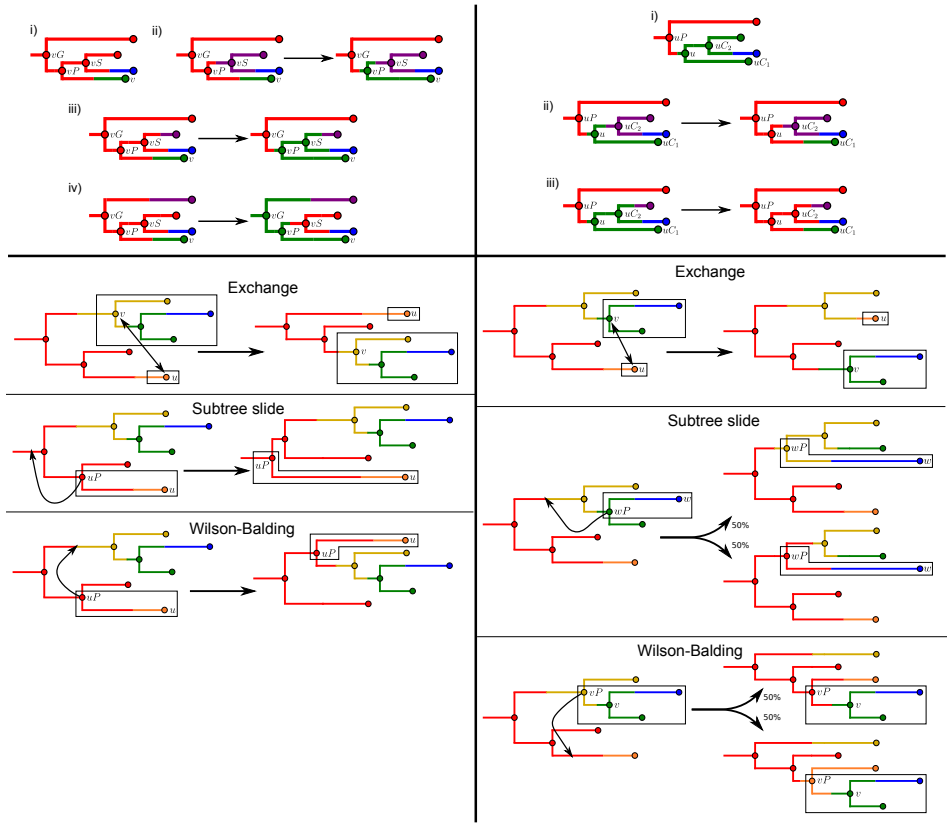
- $p(S|\mathcal{T})$ is the standard phylogenetic likelihood and $p(\mathcal{T}, d|\phi)$ the probability of observing the augmented tree under a transmission model.
- With within-host diversity:

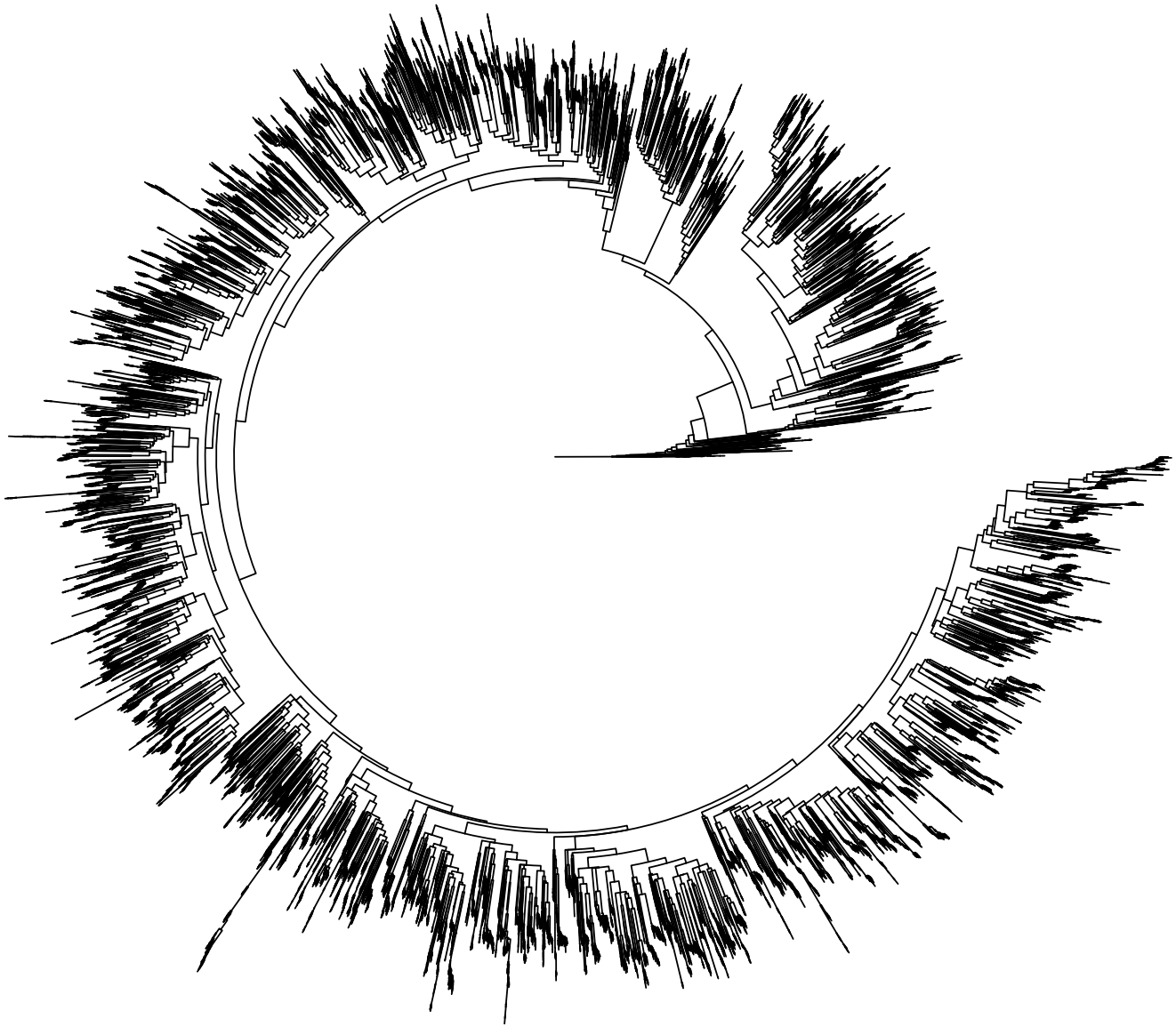
$$p(\mathcal{T}, d, q, \phi|S) = \frac{p(S|\mathcal{T})p(\mathcal{T}|\mathcal{N}, q, \phi)p(\mathcal{N}, q|\phi)p(\phi)}{p(S)}$$

- $p(\mathcal{N}, q|\phi)$ is the probability of the transmission tree and its timings as above; $p(\mathcal{T}|\mathcal{N}, q, \phi)$ is the probability of the within-host mini-phylogenies under a coalescent process.

MCMC implementation

- Hall et al., 2015 implemented simultaneous reconstruction of both trees in BEAST, with MCMC proposals that respect the rules of node augmentation.
- Several other approaches (e.g. Didelot et al., 2014, Morelli et al., 2012; Ypma et al., 2013; Klinkenberg et al., 2017) with recent work on the incomplete sampling problem (Didelot et al., 2016; Lau et al., 2016).



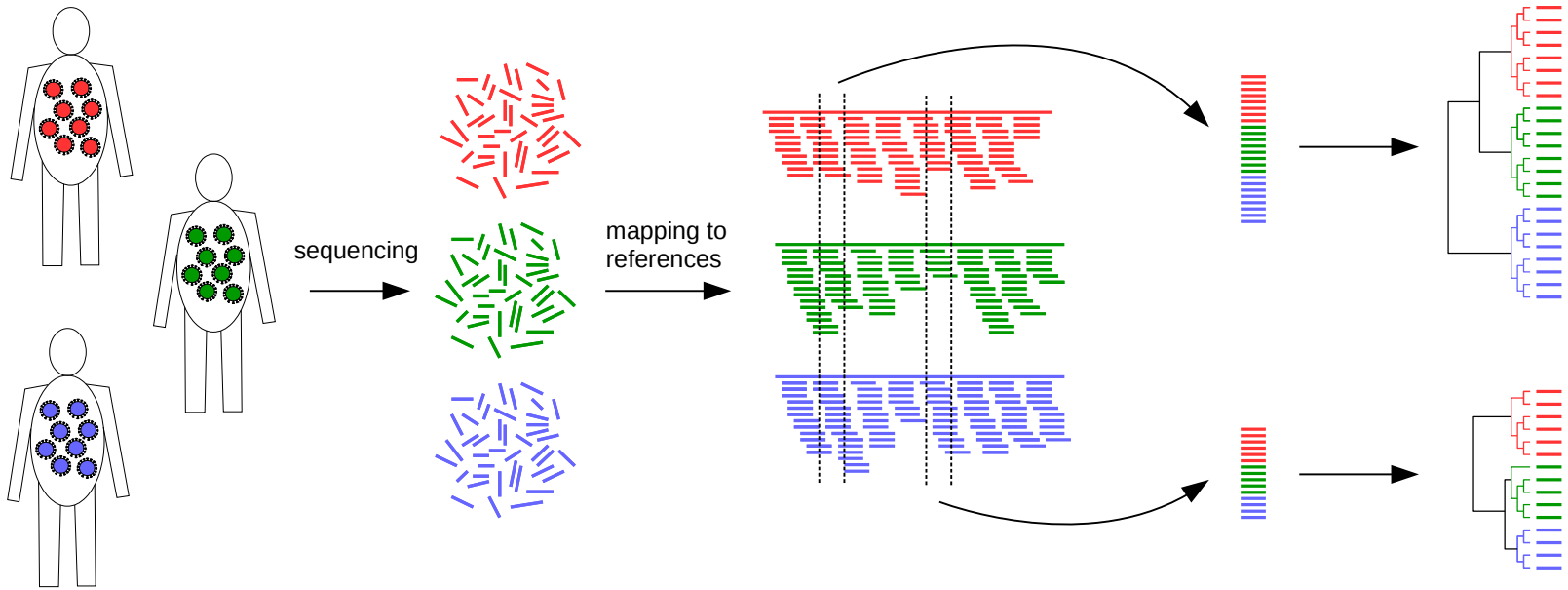


43617 tips

The BEEHIVE study

- NGS short-read sequence data acquired from samples taken from European (and one African) HIV cohort studies.
  - Some cohorts go back to the early epidemic in the 1980s
- Current data from 3138 individuals
- Epidemiology: age, gender, date of first positive test, countries of origin and infection, risk group, ART dates, etc.
- Sequences from one time point only (with a few exceptions)
- Rather than making a consensus sequence from each host's reads, we want to use everything.

Phyloscanner: phylogenetic analysis of NGS pathogen data

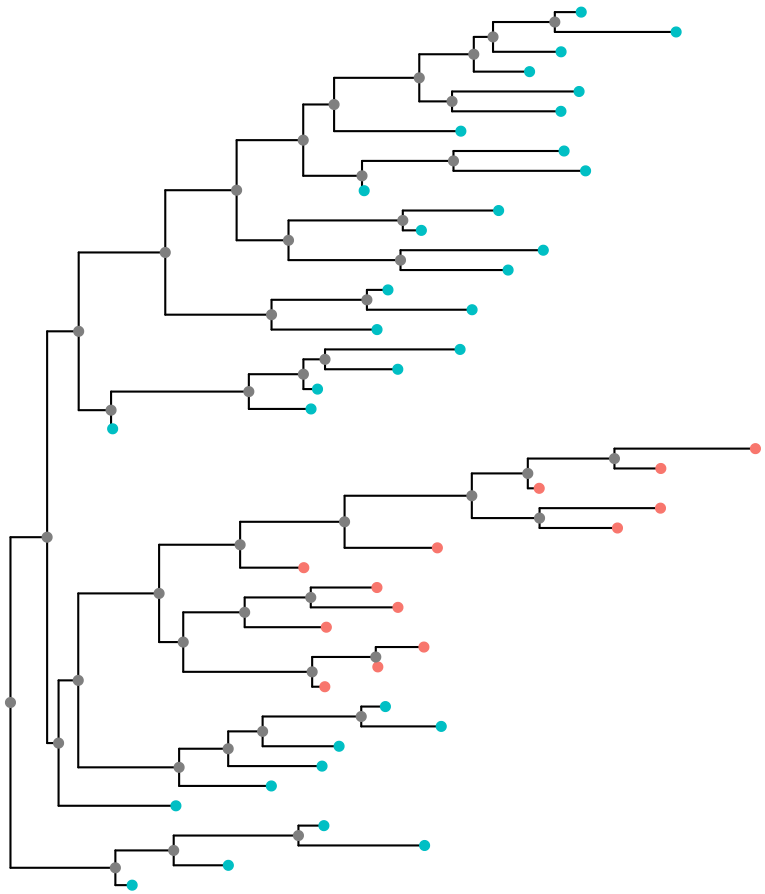


- Idea: align all short reads from all hosts to a reference genome and slide a window across the genome, building a phylogeny for the reads overlapping each window.

Phyloscanner: phylogenetic analysis of NGS pathogen data

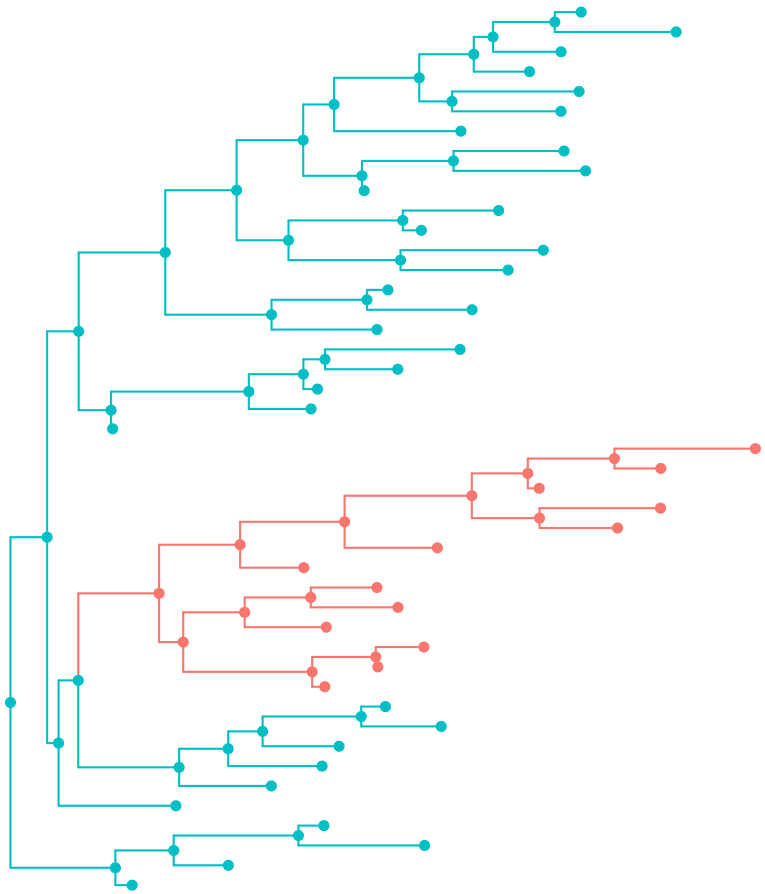
- Identical reads from a single host are merged but the duplicate counts kept as tip traits
- We use RAxML for reconstruction
- Tips are not associated with each other across different windows, but hosts are.

The topological signal of transmission



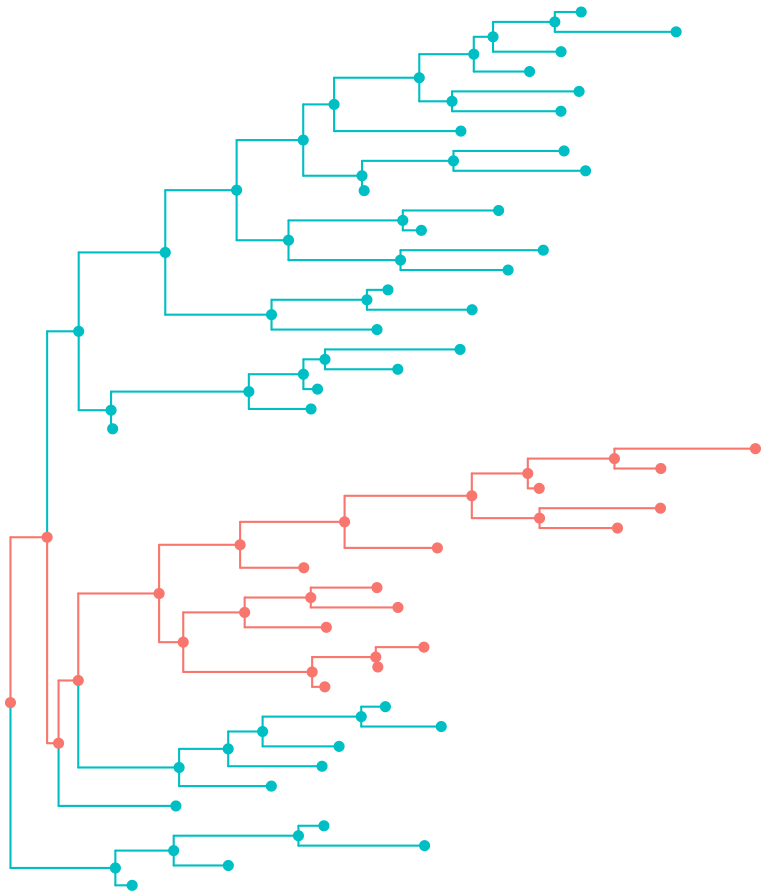
- Once we have many tips from each host, transmission has a topological signal.
- Direct transmission is suggested when the clade from the infectee is not monophyletic (Romero-Severson et al., 2016) but in general we only see the *direction* of transmission from the topology.
- Starts to look like a parsimony problem.

The topological signal of transmission



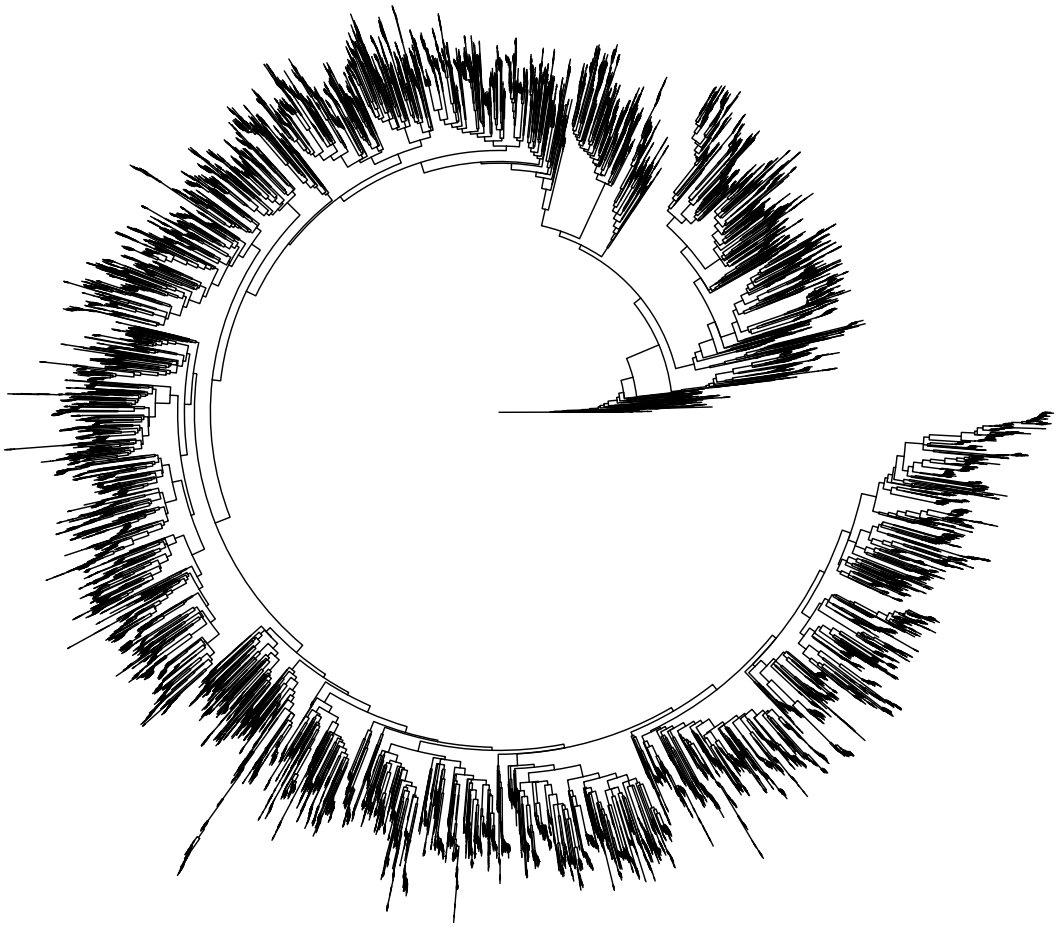
- Once we have many tips from each host, transmission has a topological signal.
- Direct transmission is suggested when the clade from the infectee is not monophyletic (Romero-Severson et al., 2016) but in general we only see the *direction* of transmission from the topology.
- Starts to look like a parsimony problem.

The topological signal of transmission



- Once we have many tips from each host, transmission has a topological signal.
- Direct transmission is suggested when the clade from the infectee is not monophyletic (Romero-Severson et al., 2016) but in general we only see the *direction* of transmission from the topology.
- Starts to look like a parsimony problem.

Challenges reconstructing transmission from this data



43617 tips

- Datasets:
  - Enormous size
  - Contamination present
  - Coverage is uneven
- Epidemiology:
  - Sampling is incomplete
  - Multiple infections present
  - Bottleneck at transmission may be wide (IDUs)

Transmission tree reconstruction using parsimony

For a fixed tree, we aim to:

- Reconstruct hosts from those represented in the tips to internal nodes in the tree
  - But also allow reconstruction to "a host outside the dataset' as required by incomplete sampling
- Minimise the number of infection events amongst hosts in the dataset...
- ...except, penalise reconstructions which suggest an unreasonable amount of genetic diversity stemming from a single infection event.
  - Identify multiple infections and contaminations

Transmission tree reconstruction using parsimony

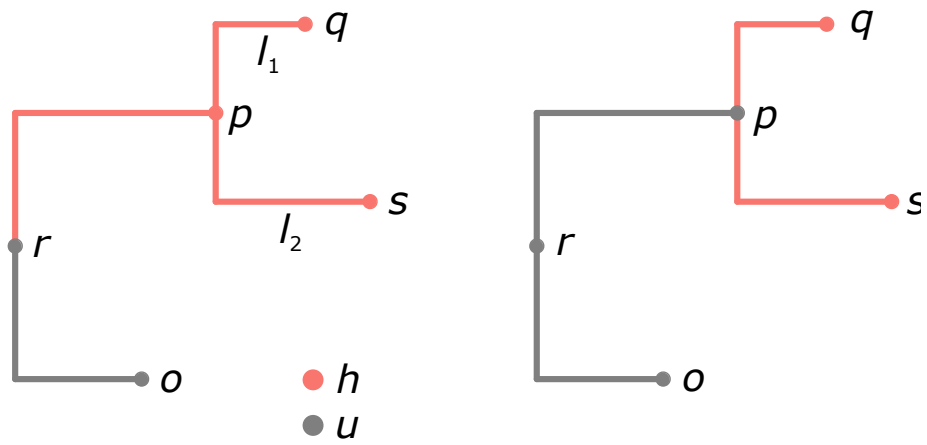
- Suppose the $\mathcal{T}$ has nodes V and we are reconstructing characters from the set of states S.
- The cost function $c(p, q; i, j)$ determines the cost of transitioning from state i to state j along the branch from p to q (in the direction away from the root).
- We take a node- and edge- dependent c on a known tree; costs for transitions vary depending on the states involved and the branch on which they occur.
- The lowest cost reconstruction is found with the Sankhoff algorithm.

Transmission tree reconstruction using parsimony

- If reads are taken from n hosts $h_1, \ldots, h_n$ making up the study population, we use the h_i as states along with an "unsampled" state u.
- Assume that the root of the tree was in the unsampled state (using an outgroup if required). Tips from outside the study population (the outgroup, other reference sequences, contaminants) are assigned u as a state.
- We are interested in minimising the cost of infections of members of the study population, but not hosts outside that population, so $c(p, q; h, u) = 0$ for all p, q, h.
- Reconstruction starts by putting a u on the root. (Otherwise it will always be cheap to reconstruct some h_i to the root, plausible or no.)

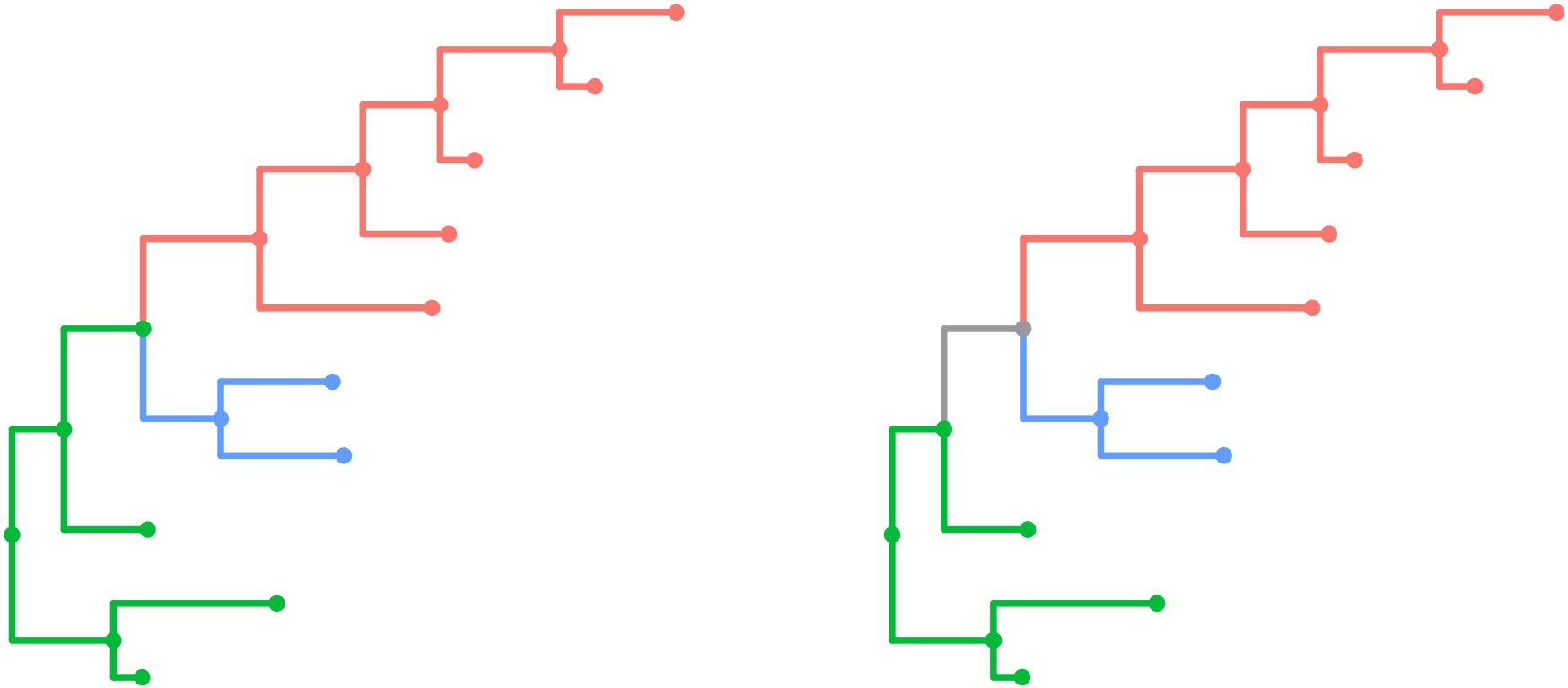
Transmission tree reconstruction using parsimony

- Let $l(p, i)$ be the sum of the branch lengths in the subtree rooted at p pruned so that only tips from i remain, or ∞ if there are no such tips. Then we take $c(p, q; i, j) = 1 + kl(p, j)$ with $k \in \mathbb{R}^+$ a tunable parameter.
- This penalises reconstructions with unreasonable amounts of within-host diversity (right).



Left: $C = 1 + k(l_1 + l_2)$
Right: $C = 2$

Transmission tree reconstruction using parsimony

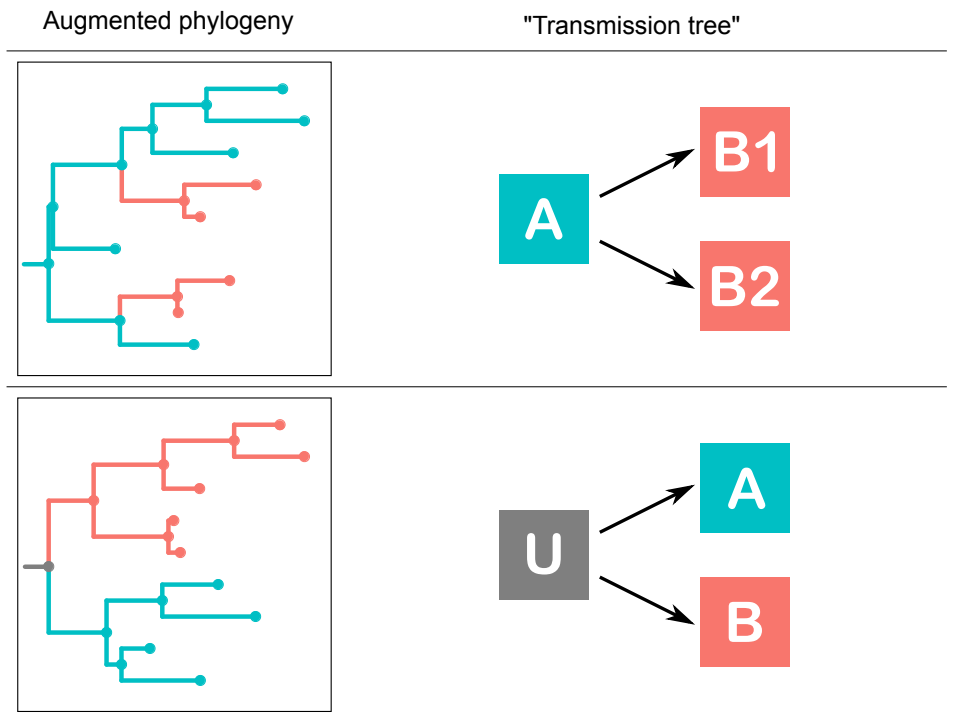


$c(p, q; h, u) = c(p, q, h, u)$ for all p, q, h. Above trees have equal C. Choose to break ties always towards h (for dense sampling), always towards u (conservative), or based on branch lengths.

Parsimony for detection of contaminants

- Small numbers of contaminant reads are frequent, where a virus from the wrong host has been sequenced
- The parsimony reconstruction does double duty to detect these
- Find "multiple introductions" where the tips in one split have very low read counts, and ignore those tips

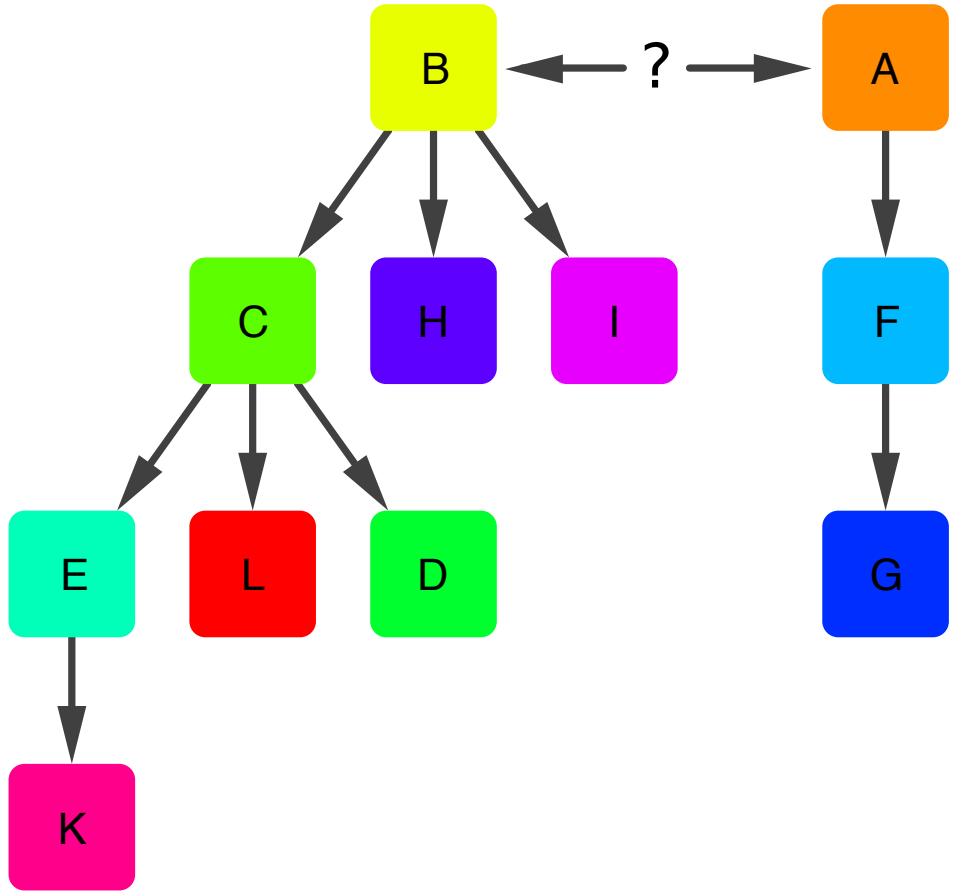
- Without the assumptions of a single infection event per host and complete sampling, the "transmission tree" has multiple nodes per host and unsampled nodes.



Example

- Known HIV transmission chain (Lemey et al., 2005, Vrancken et al., 2014).
- Reconstruction on RAxML trees for env and pol genes

Host	True	env	pol
A	B?	U	U
B	A?	U	A
C	B	B	B
D	C	D	D
E	C	C	C
F	A	U	A
G	F	U	F
H	B	B	B
I	B	B	B
K	E	E	E
L	C	C	C



Full phyloscanner results

- Full output from phyloscanner is a separate phylogeny for the reads in each genome window
- We would like to use these in a manner similar to bootstrapping, to indicate support for topological relationships
- The phylogenies are not trivially comparable across windows as the tips are not the same
- The hosts *are* the same, so the trasmission trees are more comparable, but:
  - Some hosts are absent from some windows due to sequencing problems
  - One or more nodes for unsampled regions
  - Potentially multiple nodes per host
- **Question:** Can we nonetheless give a summary or median transmission tree?

Classification of relationships

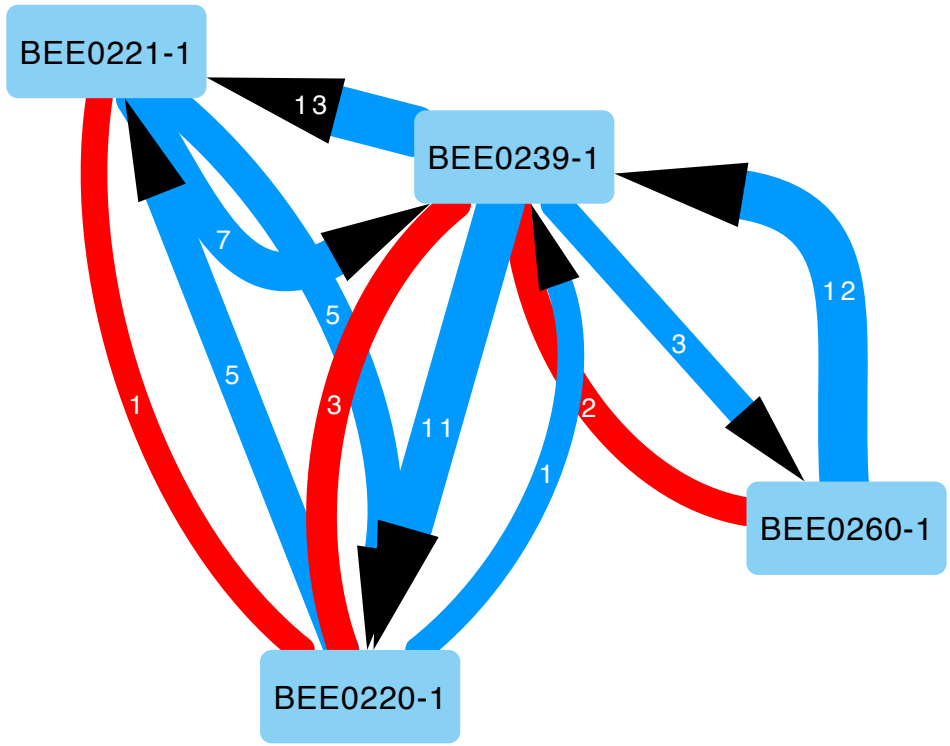
For the time being, concentrate on classifying pairwise relationships between hosts on each window

- Contiguity: is the subgraph induced by the nodes from the pair, with perhaps some unsampled nodes, connected?
- Descent: Are all the nodes from one patient ancestral to those from the other?
- Mean patristic distance in the phylogeny

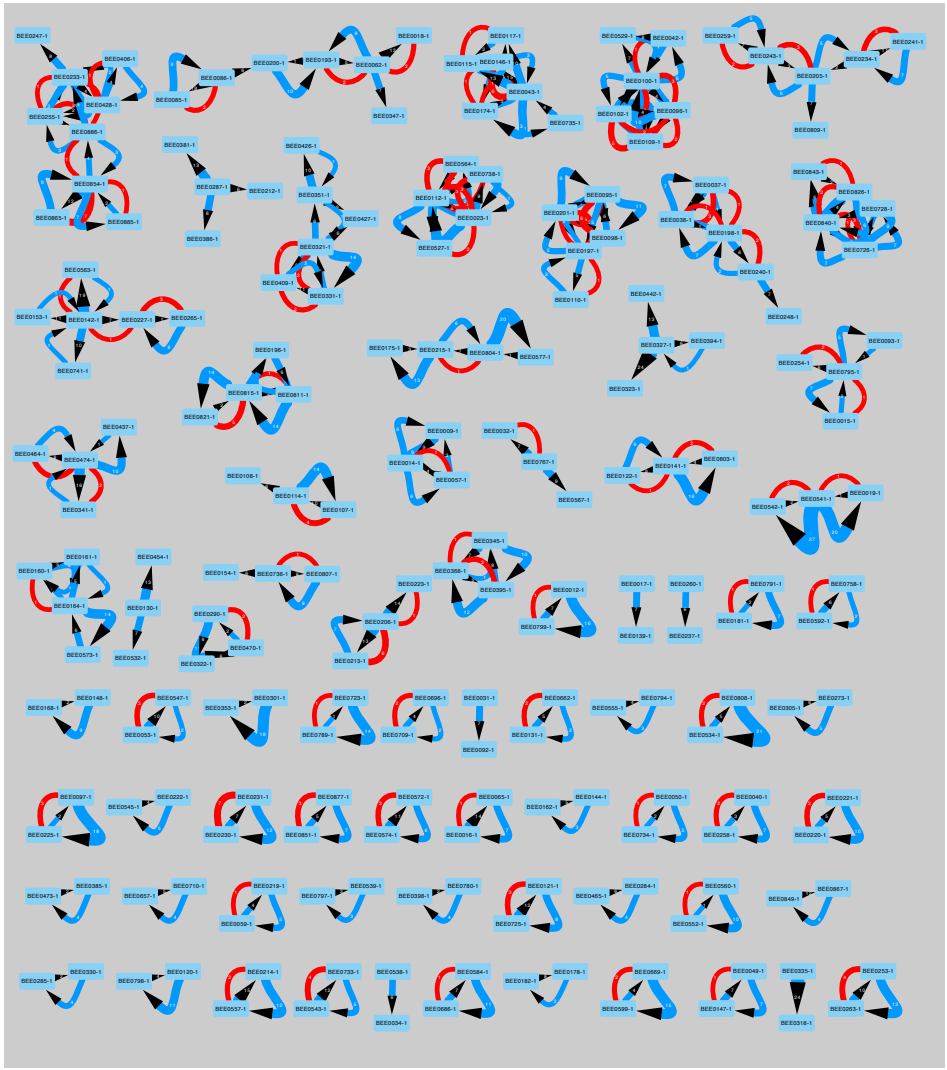
Then count relationships across windows

Improved HIV cluster detection

By setting a threshold on patristic distance we can refine the procedures for identifying HIV clusters with likely directionality.

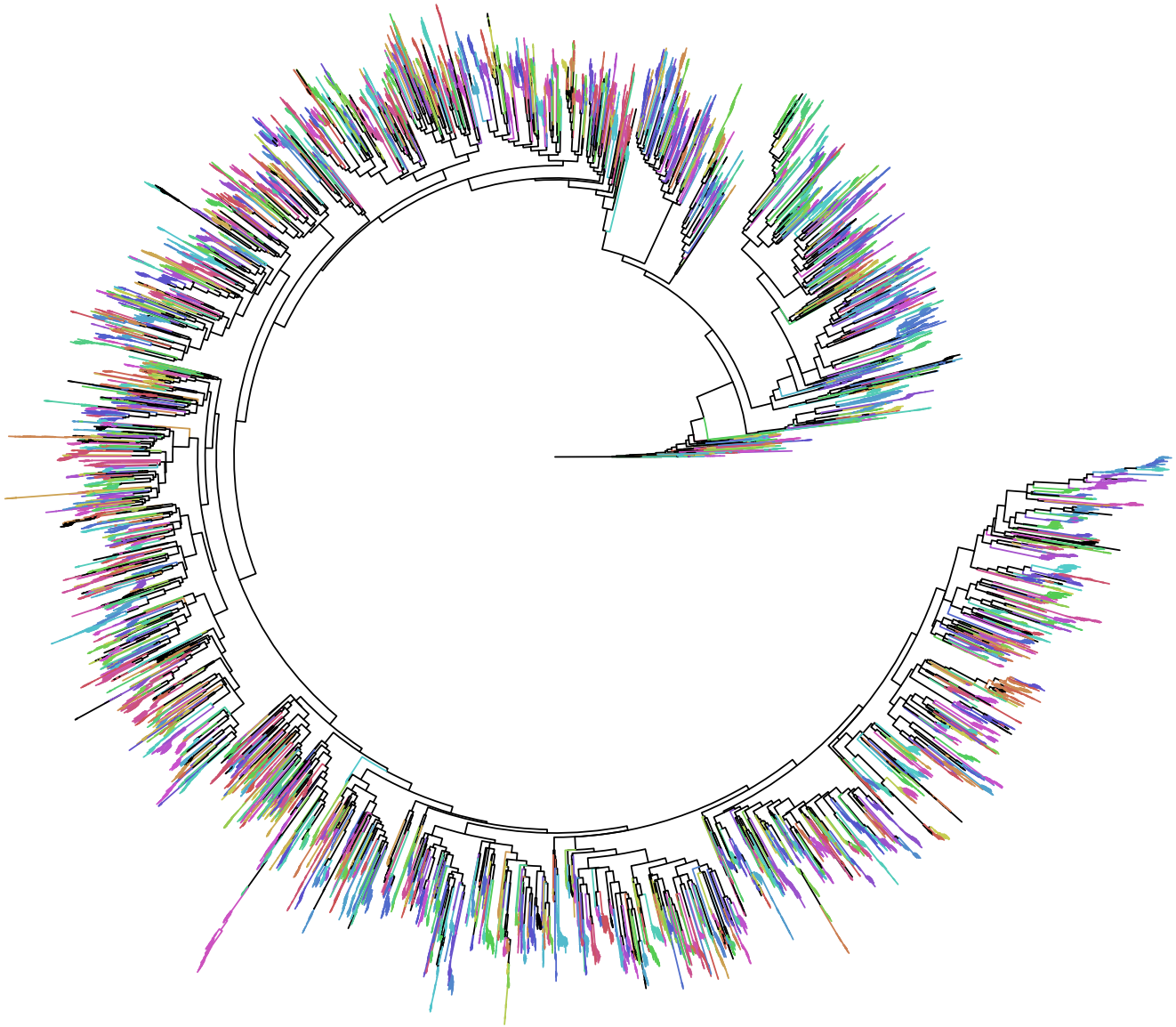


Improved HIV cluster detection



Conclusions

- The transmission tree can be viewed as an augmentation of the internal nodes of a phylogeny with host information, subject to various sets of rules.
- Considerable recent work in reconstructing it for smaller datasets, usually using MCMC.
- Phyloscanner allows reconstruction with big, NGS datasets.
- Work remains to be done for rigorous treatment of the output.



Acknowledgements

Oxford

- Christophe Fraser
- Chris Wymant

Imperial

- Oliver Ratmann

Edinburgh

- Andrew Rambaut
- Mark Woolhouse