`Phyloscanner`: automated phylogenetics within and between hosts with NGS deep-sequencing data reveals transmission and multiple infection

Molecular Biology & Evolution, 2017

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], The Stop-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration

* Equal contribution
[1] Big Data Institute, Nuffield Department of Medicine, University of Oxford
[2] Department of Infectious Disease Epidemiology, Imperial College London
[3] Peter Medawar Building for Pathogen Research, Nuffield Department of Medicine & NIHR Oxford BRC, University of Oxford
[4] Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford
[5] Department of Veterinary Medicine, University of Cambridge
[6] Laboratory of Experimental Virology, Academic Medical Center of the University of Amsterdam

UNIVERSITY OF OXFORD

BIG DATA INSTITUTE

Problem: we want to identify risk factors for transmission. We need to identify not just transmission pairs or clusters, but *who infected whom*.

Molecular epi with one pathogen sequence per sampled individual (i): clustering

- One pathogen sequence per sampled individual.
- Define clusters: likely to be related by recent transmission.
- Investigate epidemiological correlates.

No information on who is transmitting!

Perrin *et al.*, Nat. Commun. 2017

Molecular epi with one pathogen sequence per sampled individual (ii): getting to directionality

- Supplement a phylogeny with additional epi data, and fit a model of transmission.
- Volz & Frost, PLoS Comp Bio 2013
  Jombart *et al.*, PLoS Comp Bio 2014
  Didelot *et al.*, MBE 2014 & 2017 (“colouring the branches”)
  Hall, Woolhouse & Rambaut, PLoS Comp Bio 2015
- Dependent on the availability and accuracy of epi data.
- Dependent on the transmission model assumptions.

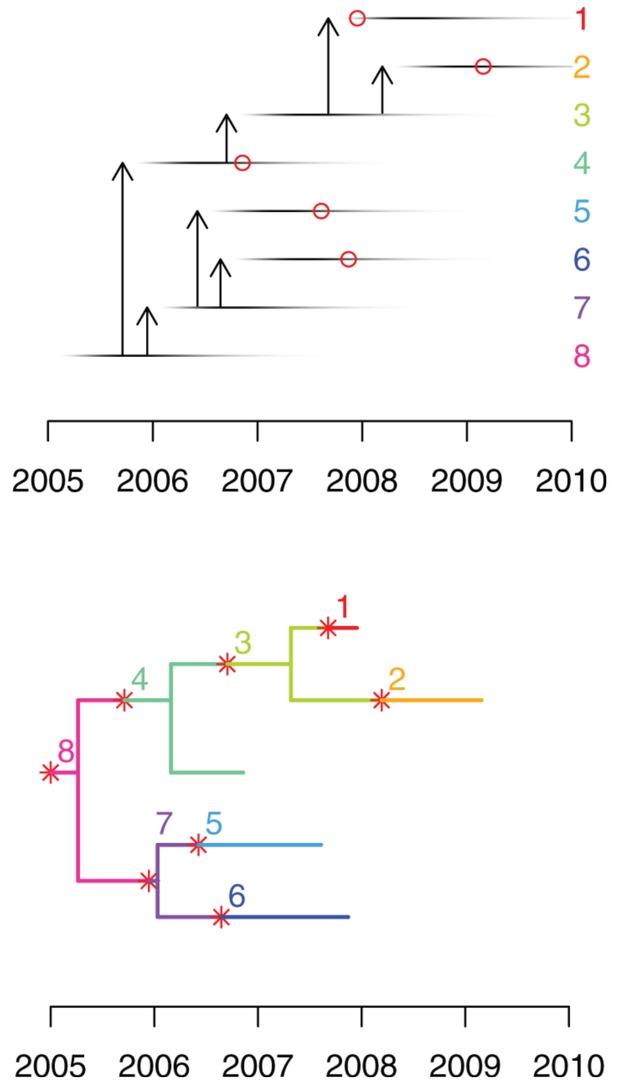
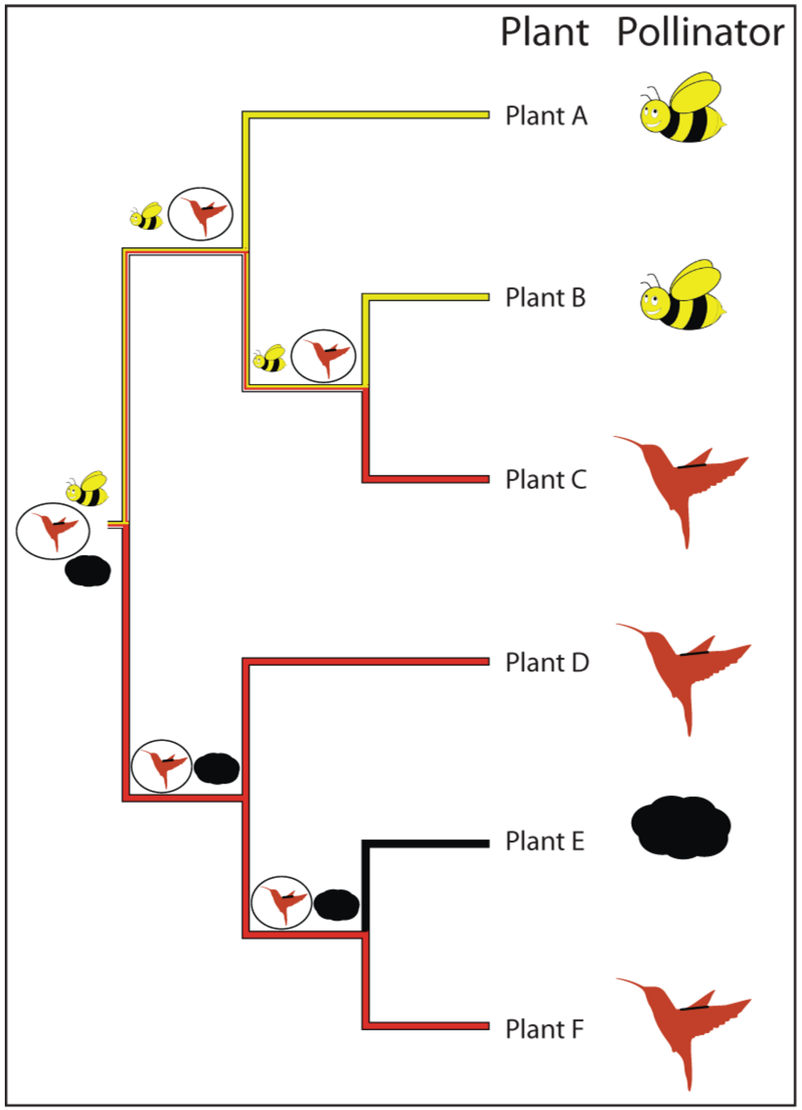


Figure from Didelot *et al.*

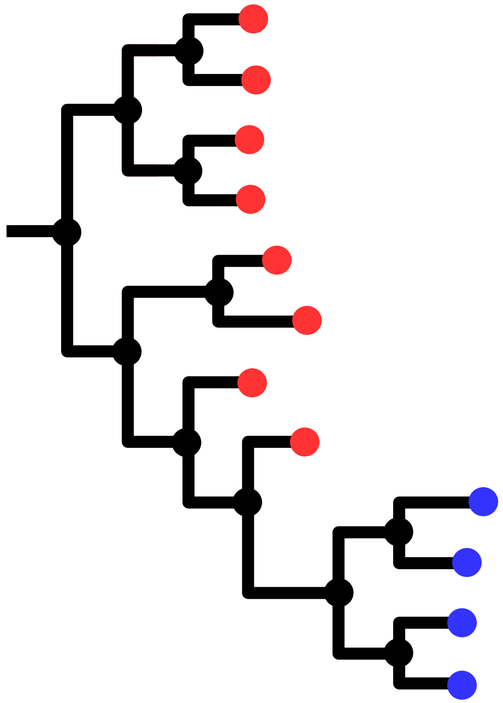
Ancestral State Reconstruction



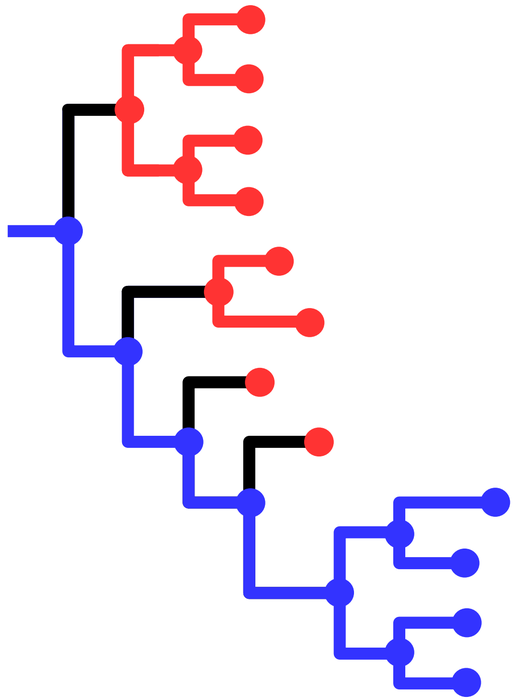
- Each tip in a phylogeny has a state or ‘trait’:
  - could be biological (i.e. a phenotype);
  - could be something else (e.g. location).
- Infer the most likely traits of the tips’ *ancestors*, i.e. internal nodes in the tree.

Figure from Joy et al., PLoS Comp Bio 2016

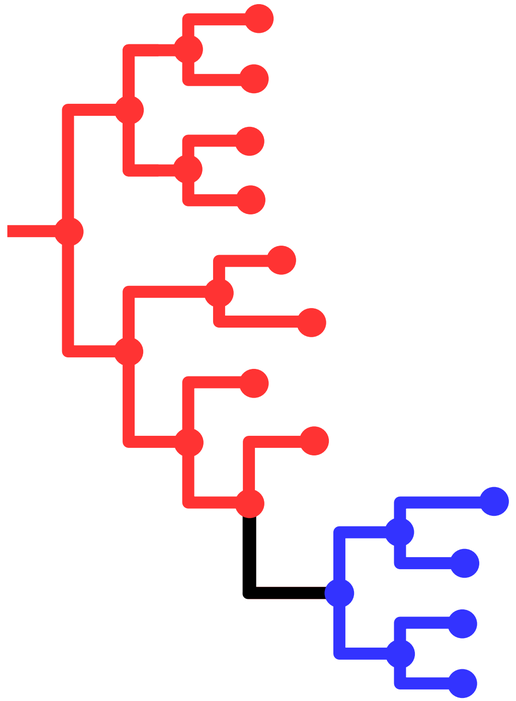
Ancestral state reconstruction: Parsimony



Given an inferred phylogeny and known states (red or blue) at the tips, what are the states of the internal nodes?

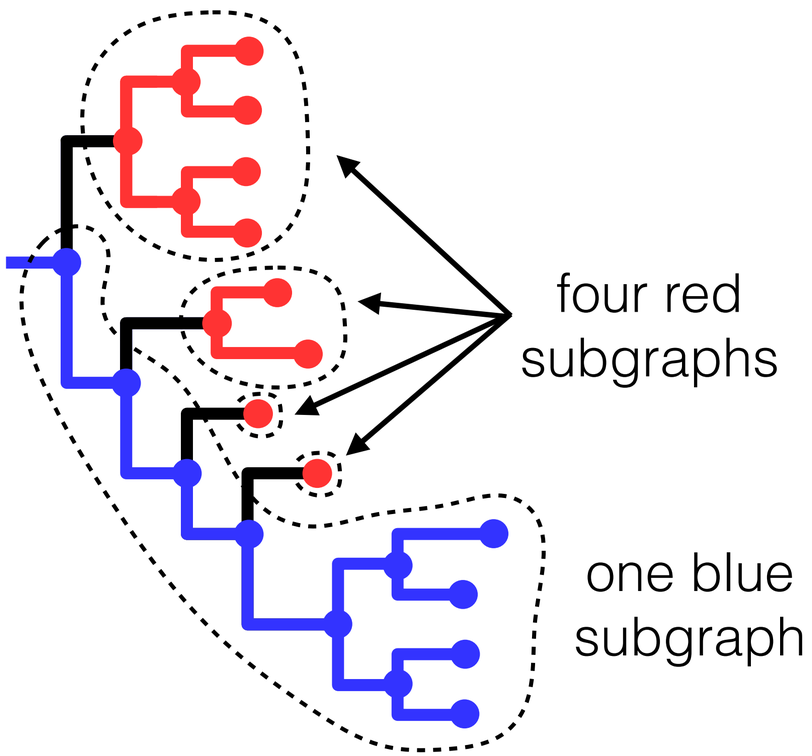


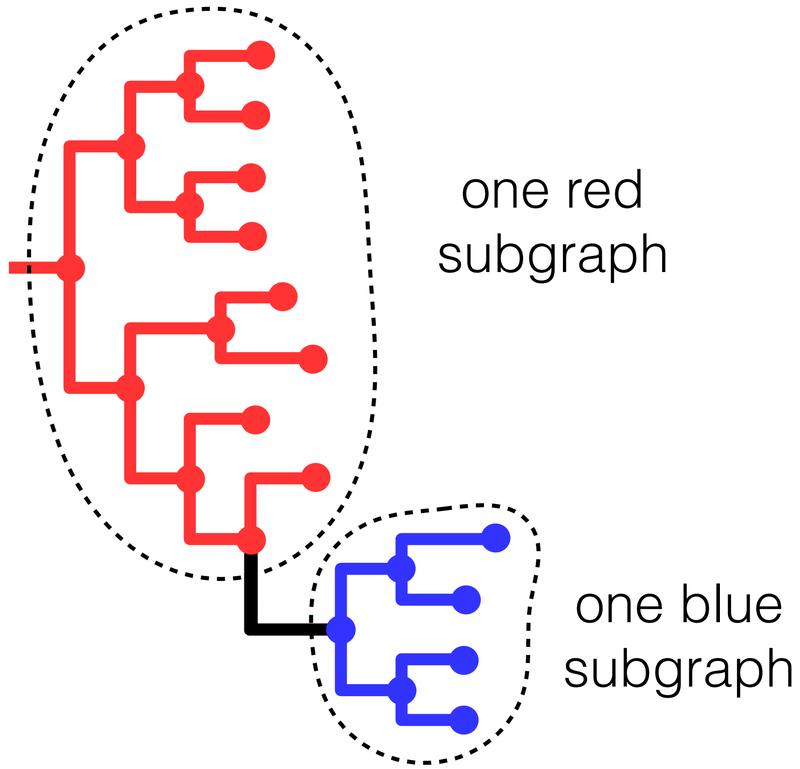
If the ancestor node is blue, 4+ changes of state needed (the black branches).



If the ancestor node is red, only 1 change of state is needed.

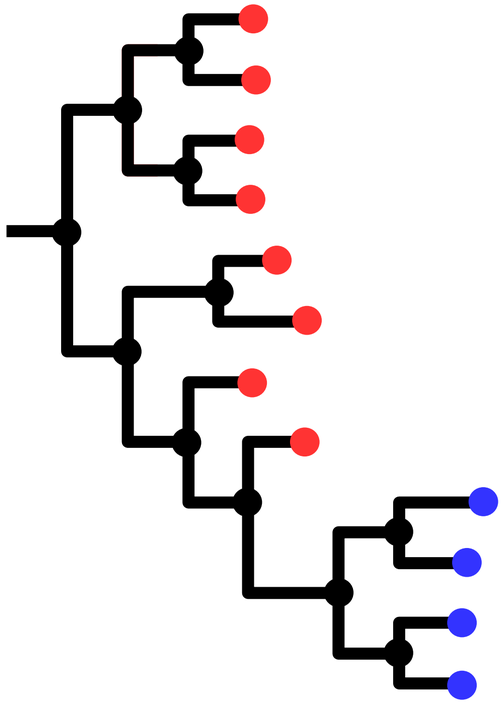
Ancestral State Reconstruction: subgraphs

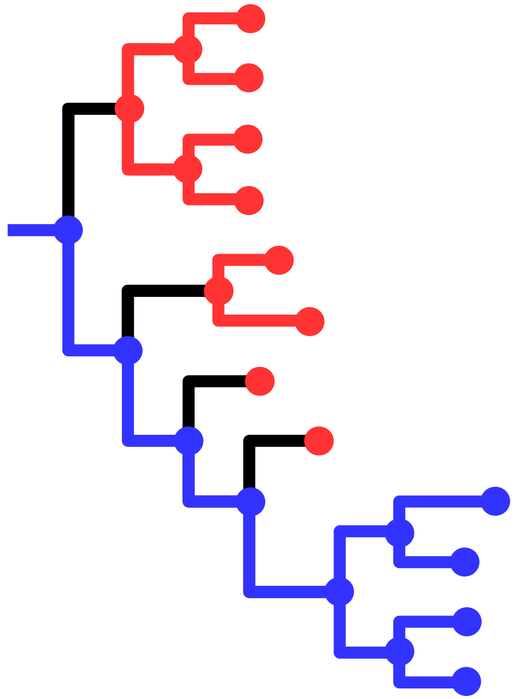


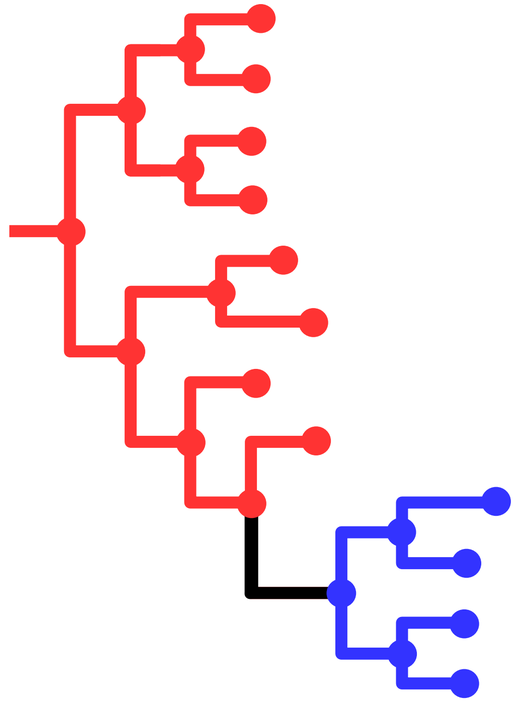


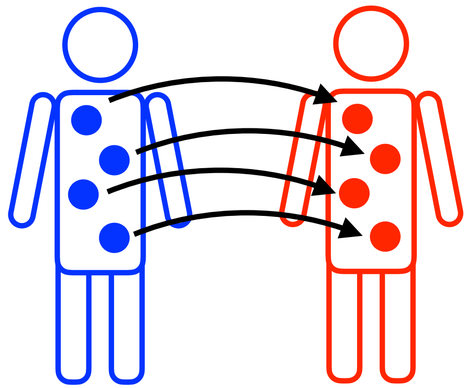
Here: a *subgraph* is connected/continuous region of the phylogeny with the same state.
i.e. one subgraph = one solid block of colour.

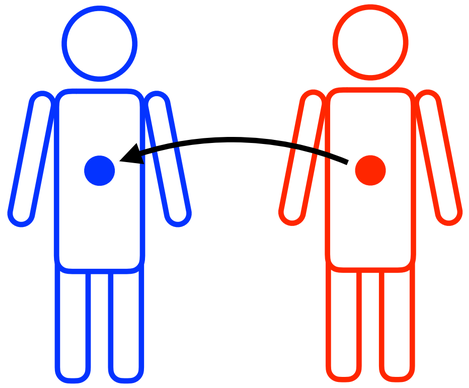
Reconstructing the state “which person was this virus in?”





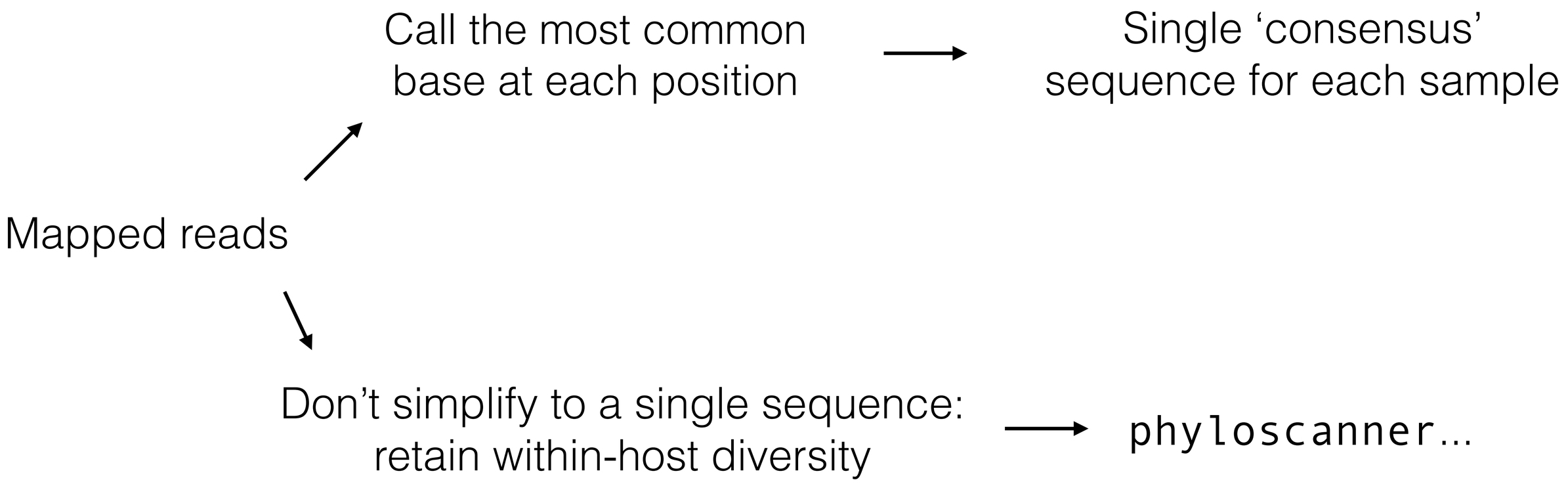


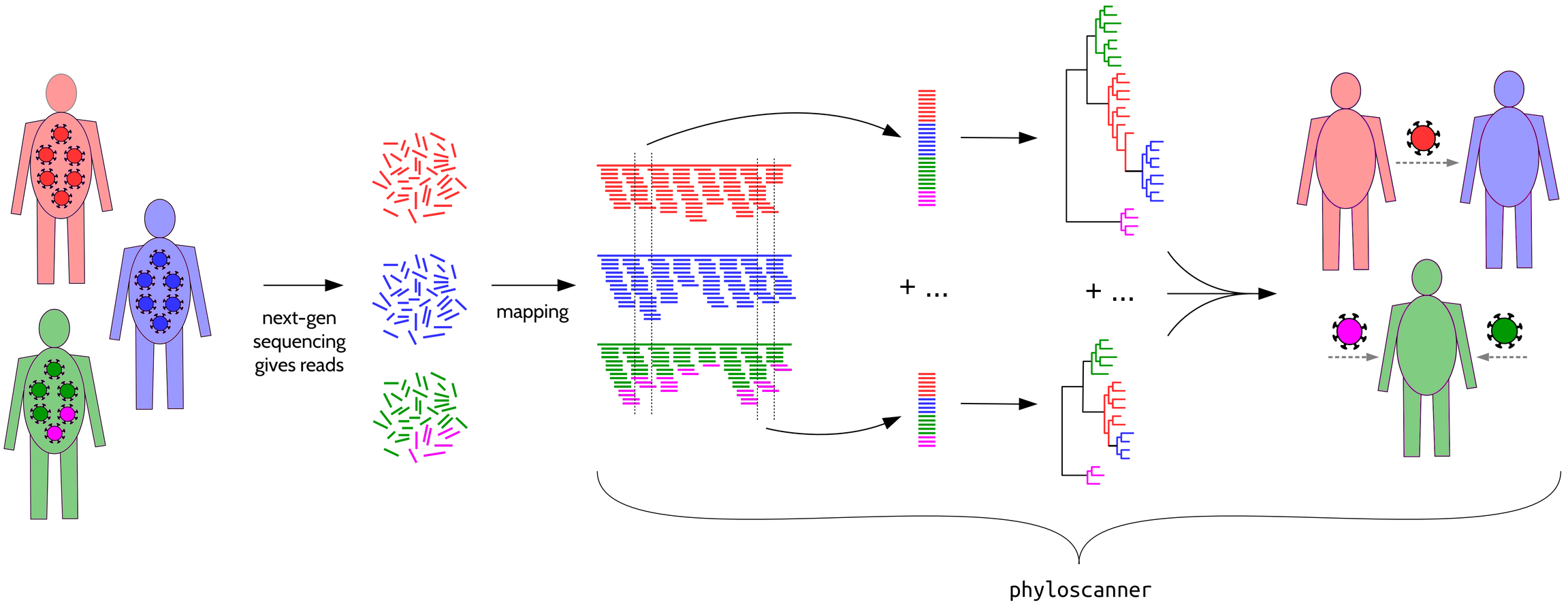




(Transmission could be via an unsampled intermediate.)

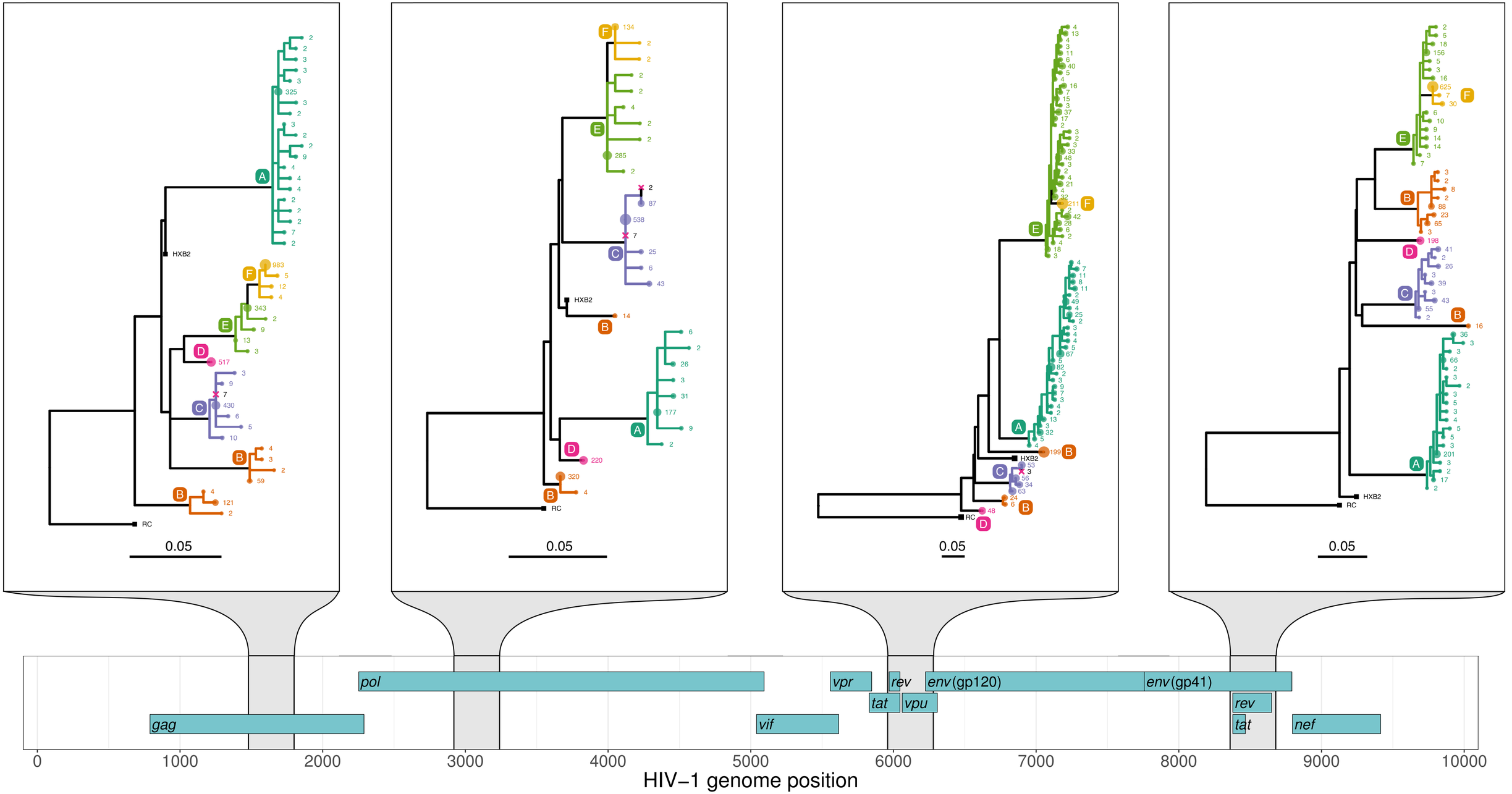
c.f. Romero-Severson *et al.* PNAS 2016

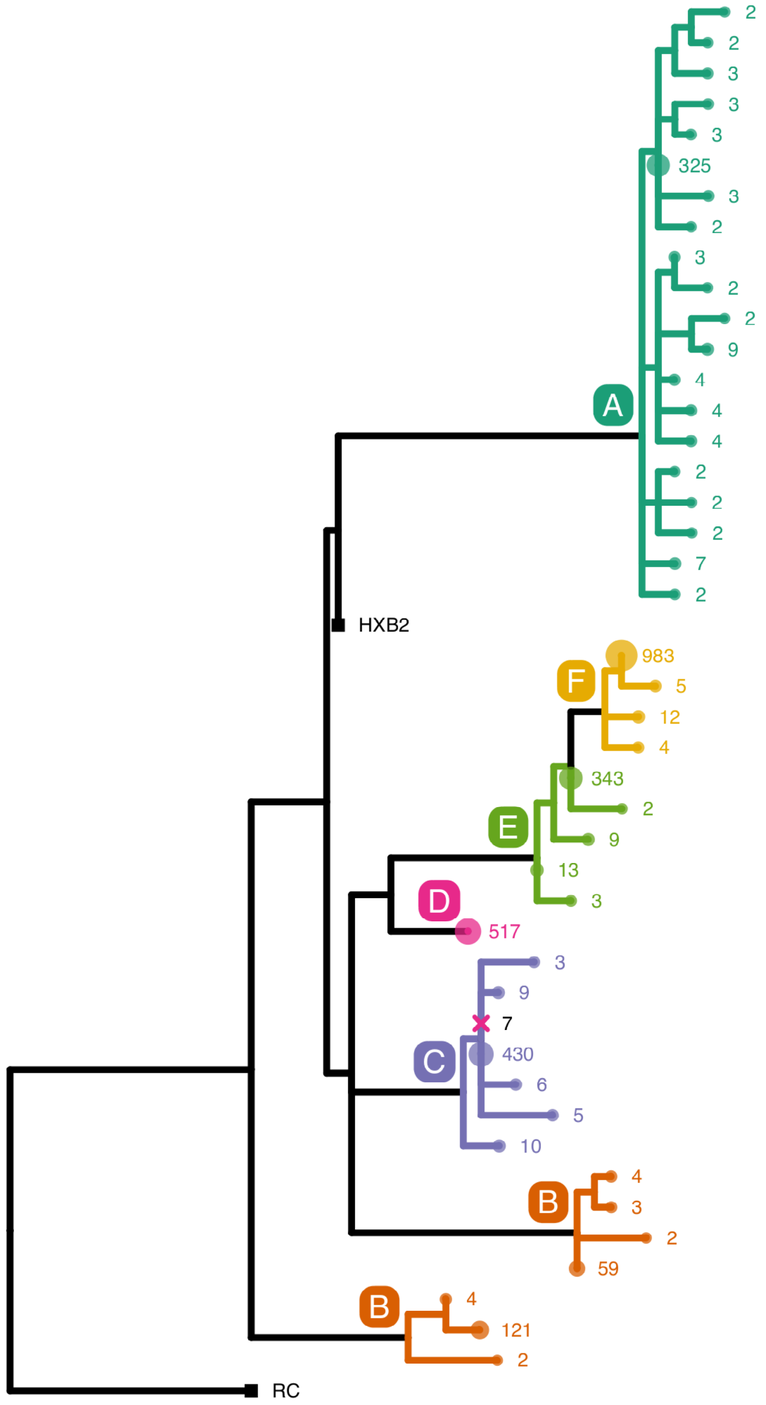

Call the most common base at each position
Single ‘consensus’ sequence for each sample
Mapped reads
Don’t simplify to a single sequence: retain within-host diversity
phyloscanner…



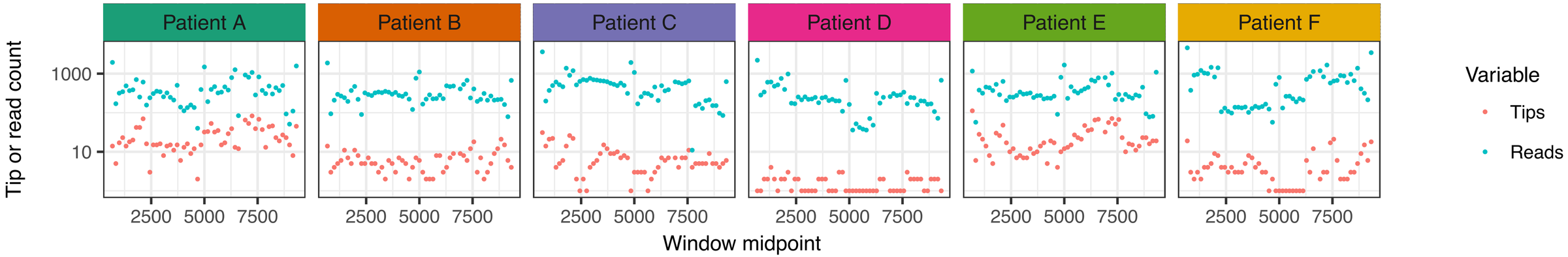
Wymant, Hall *et al.,* MBE 2017

Alignment with `MAFFT`: Katoh *et al.*, Nucleic Acids Res. 2002
Phylogenetic inference with `RAxML`: Stamatakis, Bioinformatics 2014


HXB2
RC
0.05
pol
vpr
rev
env(gp120)
env(gp41)
tat
vpu
gag
vif
rev
tat
nef
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
HIV−1 genome position

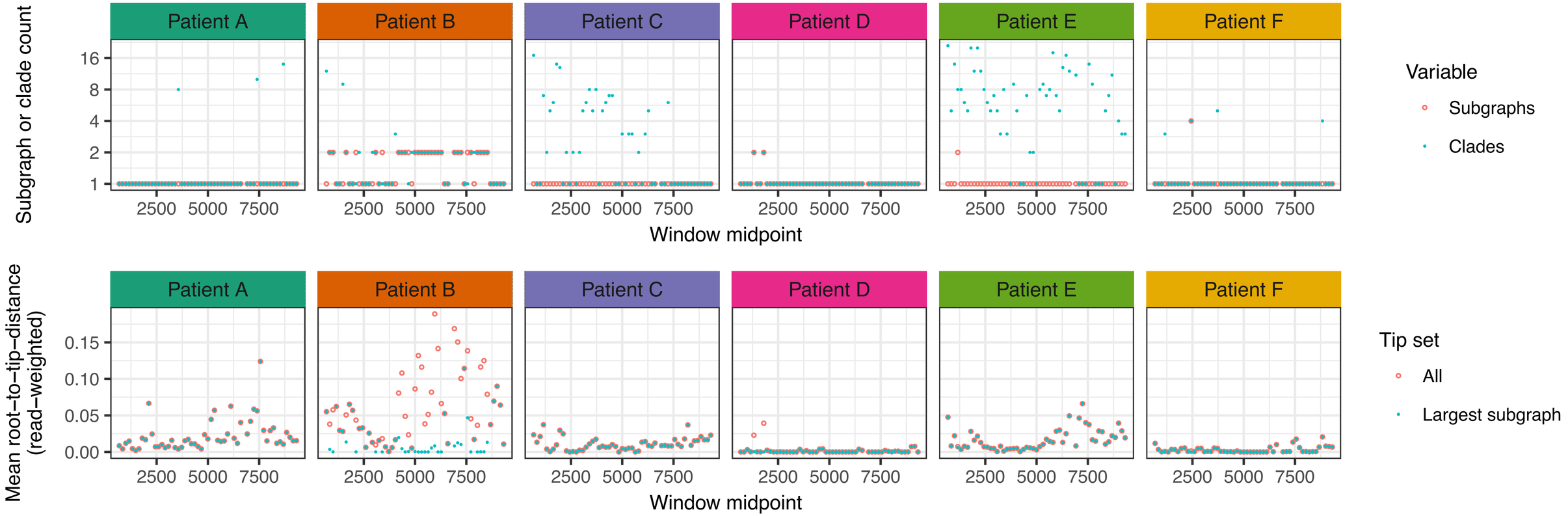

A
2
2
3
3
3
325
3
2
3
2
2
9
4
4
4
2
2
2
7
2
HXB2
F
983
5
12
4
E
343
2
9
13
3
D
517
C
3
9
7
430
6
5
10
B
4
3
2
59
B
4
121
2
RC

Per-patient summary statistics along the genome

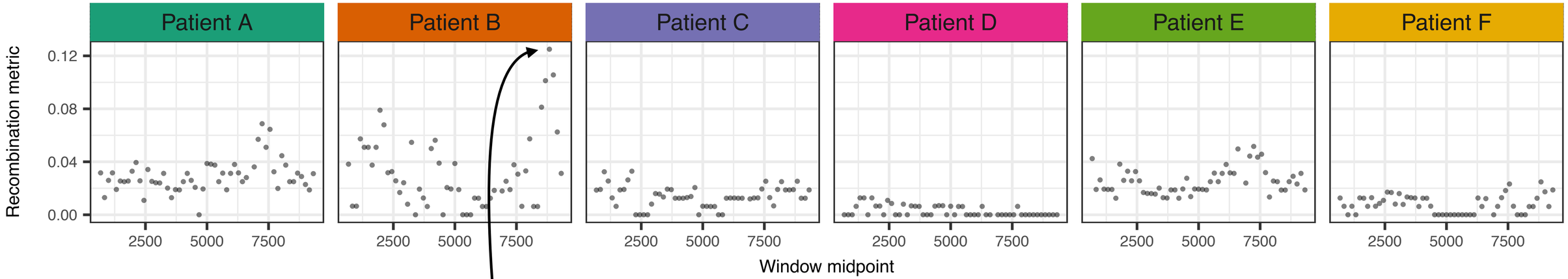


100-1000 reads,
10-100 unique reads.
Diverse!

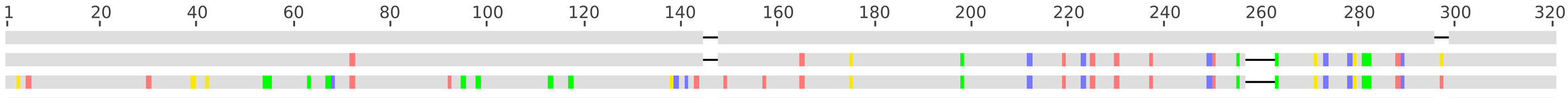
100-1000 reads,
1-2 unique reads.
Not diverse!


Patient A
Patient B
Patient C
Patient D
Patient E
Patient F
Subgraph or clade count
16
8
4
2
1
2500 5000 7500
Window midpoint
Mean root–to–tip–distance
(read–weighted)
0.15
0.10
0.05
0.00
Variable
Subgraphs
Clades
Tip set
All
Largest subgraph

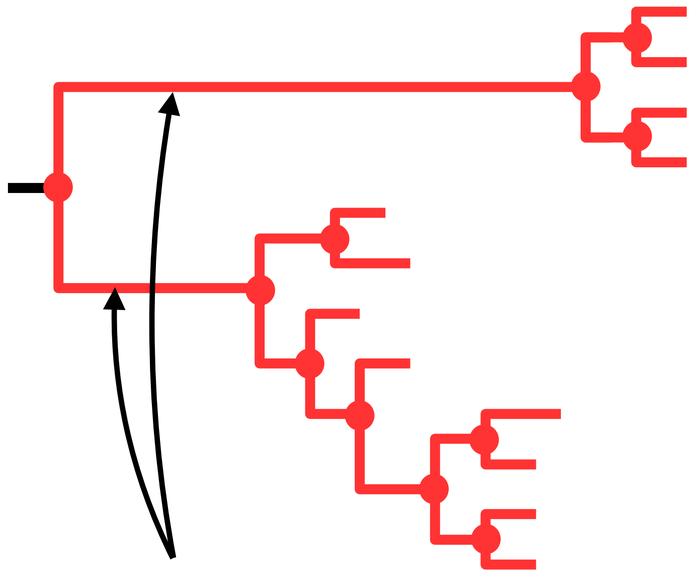
Recombination



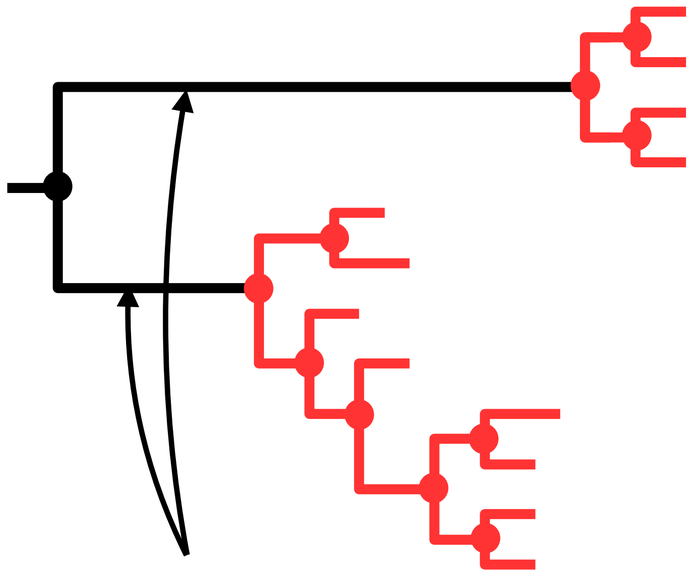
The most recombinant reads. Colours = SNPs



Dually infected individuals



Did this evolution
happen in the red host,

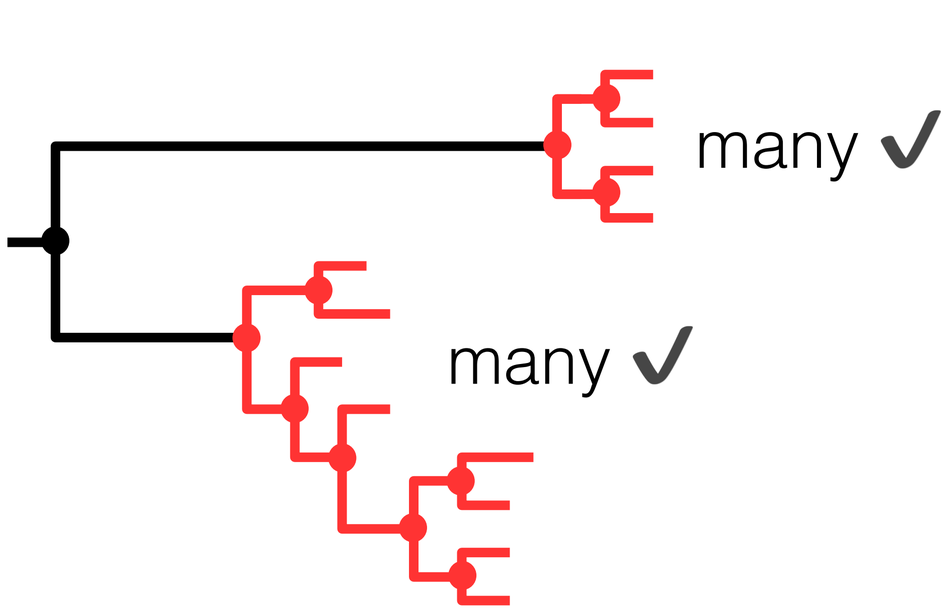


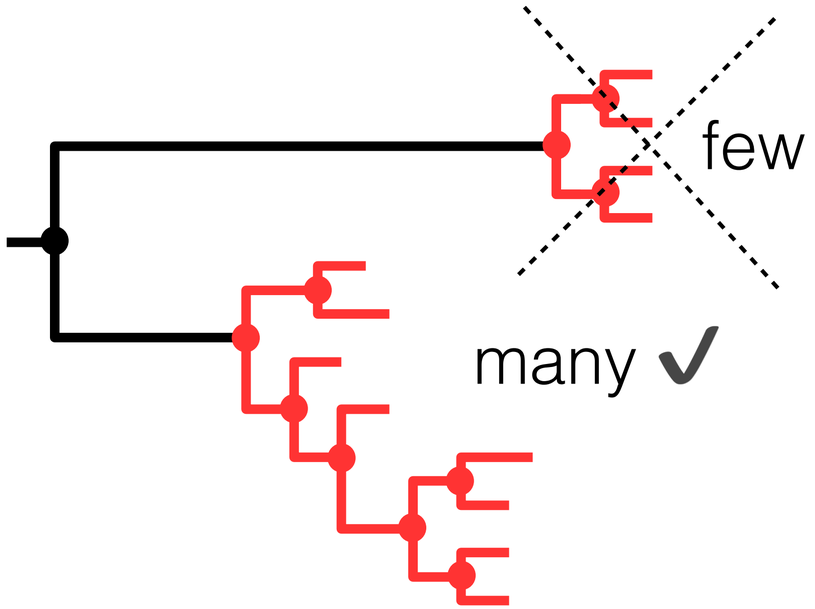
or not?

LHS always preferred under simple parsimony.
Use modified parsimony with parameter *k*: a within-host diversity penalty.

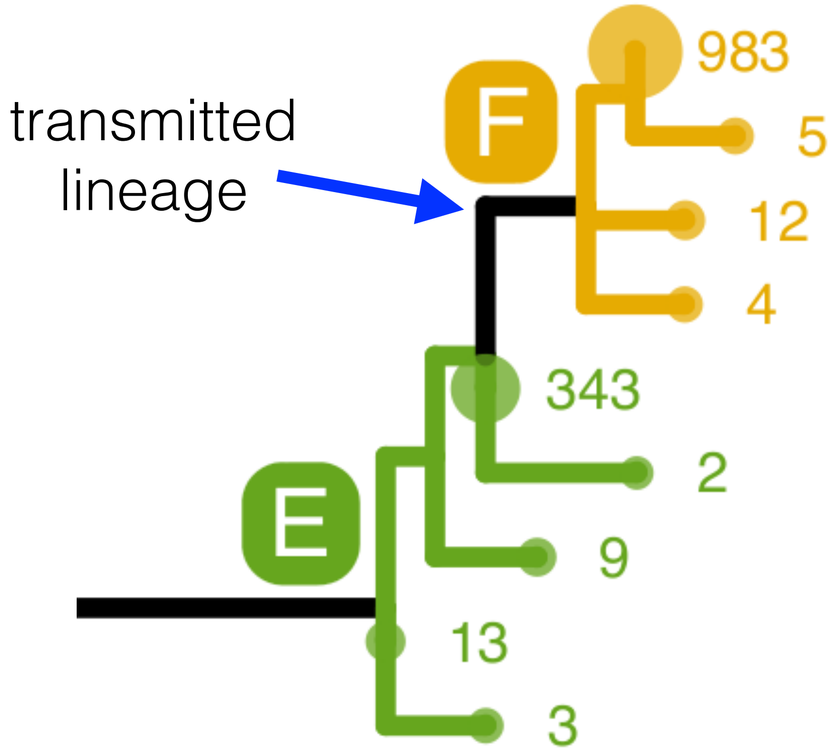
Choose 1 / (unexpectedly large pairwise genetic diversity in a single infection)

Dually infected individual or contaminated sample?

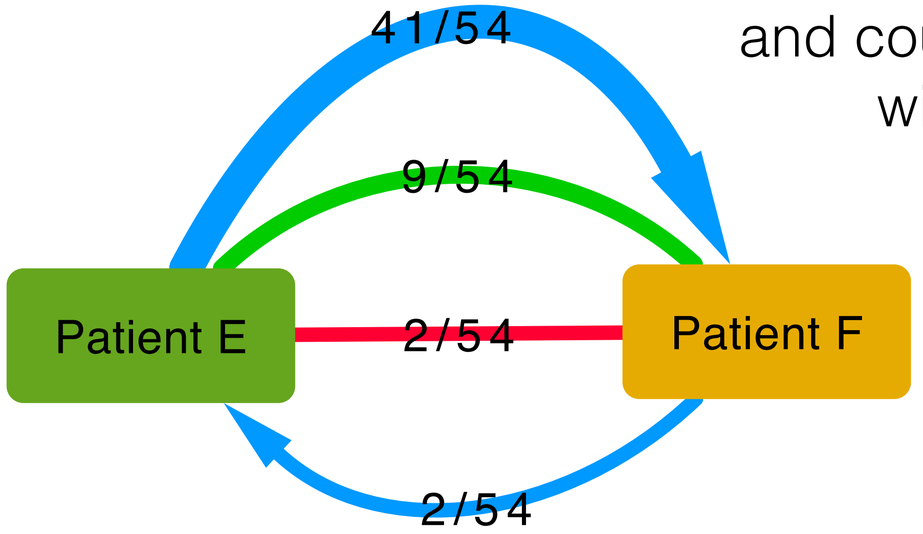


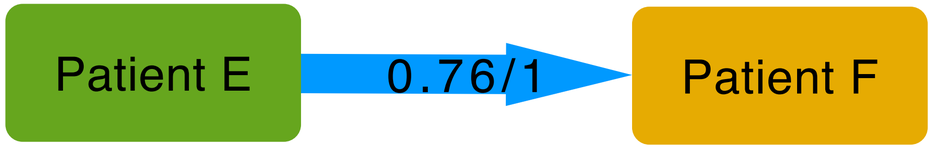


Transmission

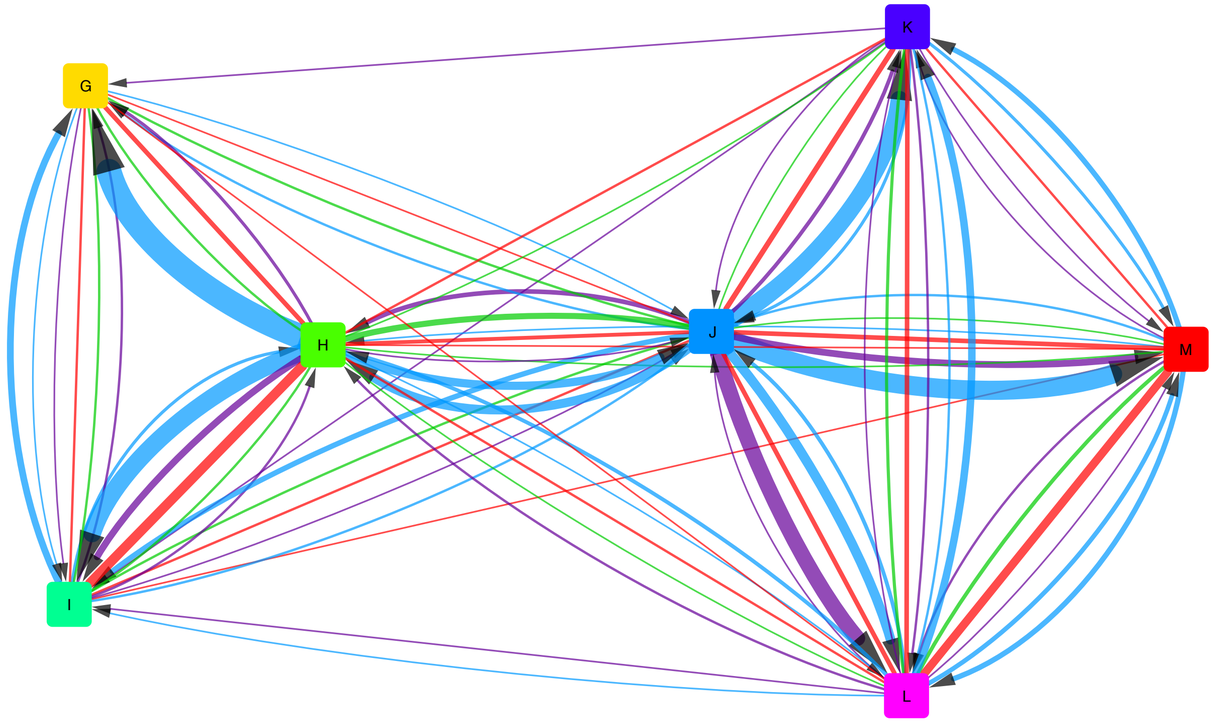


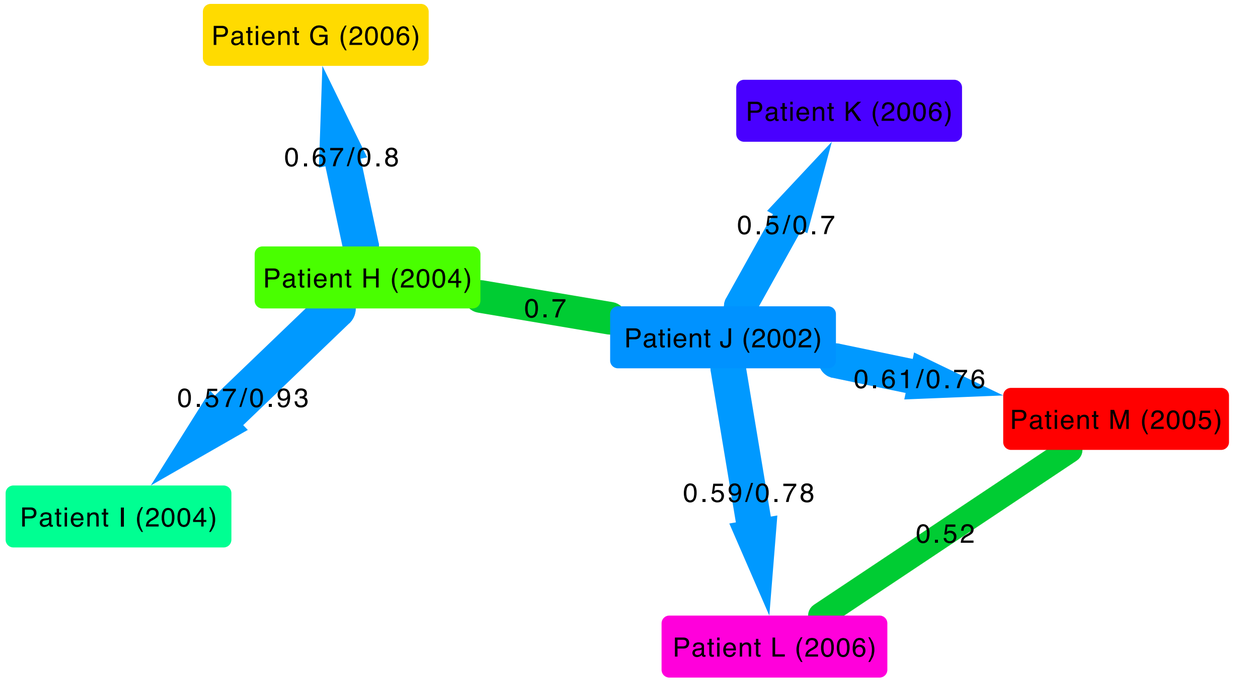
Summary: categorise and count genomic windows



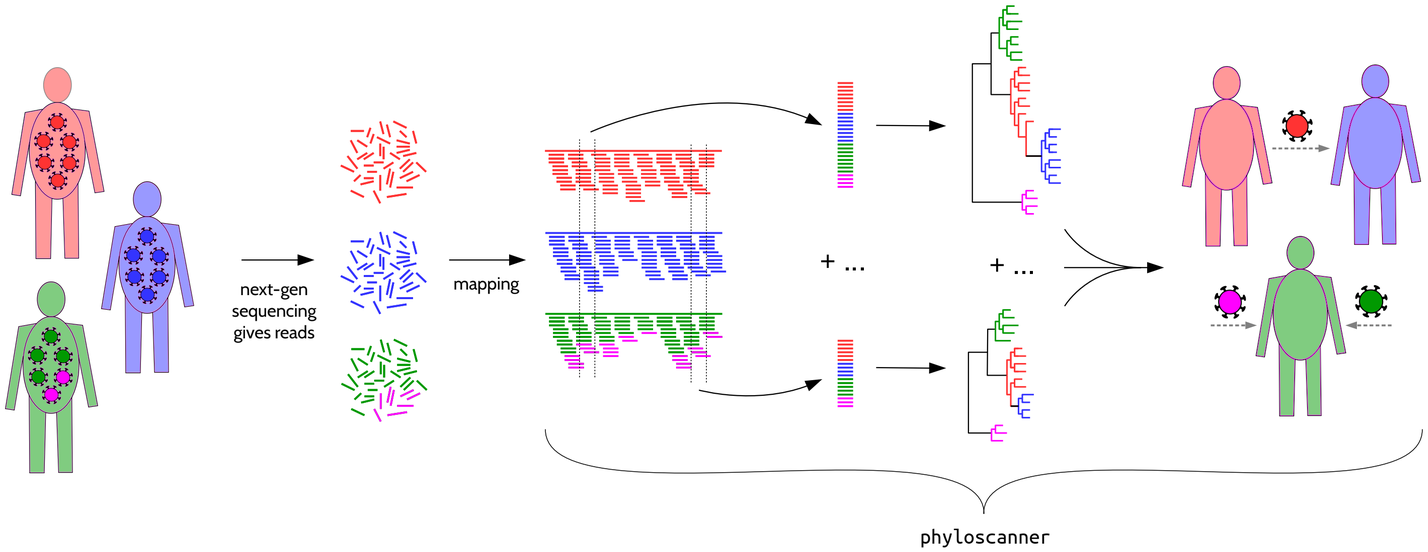


Simpler summary: aggregate categories


K
G
H
J
M
I
L


Patient G (2006)
Patient K (2006)
0.67/0.8
0.5/0.7
Patient H (2004)
0.7
Patient J (2002)
0.61/0.76
Patient M (2005)
0.57/0.93
Patient I (2004)
0.59/0.78
0.52
Patient L (2006)

Summary



Wymant, Hall *et al.* MBE 2017
GitHub.com/BDI-pathogens/phyloscanner
(or search ‘phyloscanner’)

`phyloscanner` creates within- & between-host phylogenies along the genome from NGS reads. Analysing these, or other such phylogenies, it identifies

- **transmission**: one individual’s pathogen population being ancestral to someone else’s
- **multiple infection**: distinct subpopulations in the same host
- **contamination**: exact duplicates, phylogenetic outliers
- **recombination**

The BEEHIVE Project: Bridging the Epidemiology and Evolution of HIV in Europe

Oxford University
Christophe Fraser
Tanya Golubchik
Matthew Hall
Michelle Kendall
Rob Power
Chris Wymant
Imperial College London
Paul Kellam
Frank de Wolf
Amsterdam Medical Centre
Margreet Bakker
Ben Berkhout
Marion Cornelissen
Peter Reiss

Wellcome Trust Sanger Institute
Swee Hoe Ong
European Bioinformatics Institute
Astrid Gall
Martin Hunt
HIV Monitoring Foundation
Daniela Bezemer
Mariska Hillebregt
Ard van Sighem
Sima Zaheri
Karolinska Institute
Jan Albert
Antwerp Institute of Tropical Medicine
Katrien Fransen
Guido Vanham

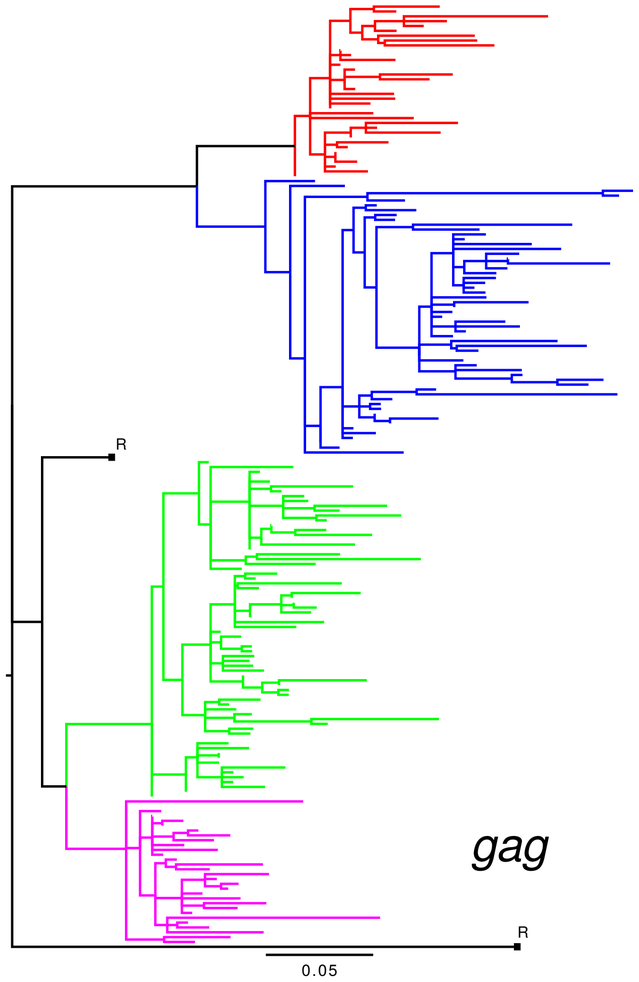
John Hopkins University
M. Kate Grabowski
Robert Koch-Institute, Berlin
Norbert Bannert
Claudia Kücherer
University Hospital Zürich
Huldrych Günthard
Roger Kouyos
Division of Intramural Research NIAID, Baltimore
Oliver Laeyendecker
Helsinki University Hospital
Pia Kivelä
Kirsi Liitsola
Matti Ristola

Université Paris Sud
Laurence Meyer
University College London
Kholoud Porter
College de France
François Blanquart
École polytechnique fédérale de Lausanne
Jacques Fellay
Analysis Advisory Group
Samuel Alizon
Sebastian Bonhoeffer
Gabriel Leventhal
Andrew Rambaut
Oliver Pybus
Gil McVean
With thanks to
Nick Croucher
Katrina Lythgoe
Oliver Ratmann

erc

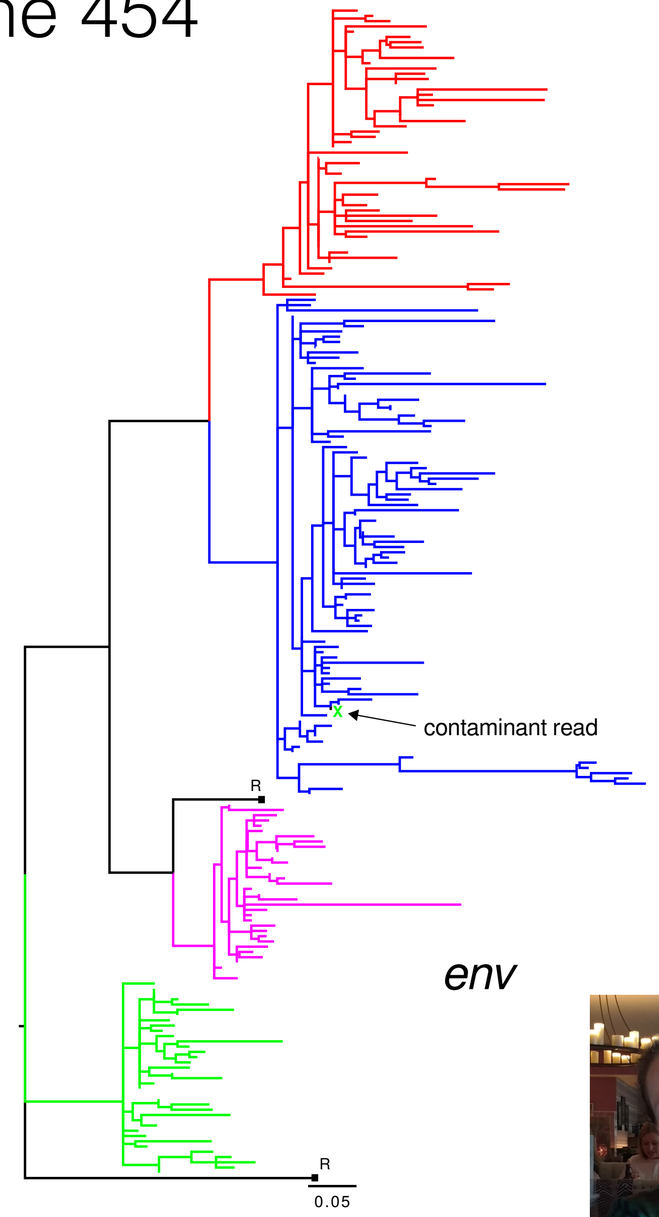
Extra slides: application to other sequencing platforms and pathogens, and using `phyloscanner`

Four more BEEHIVE patients sequenced with Roche 454



gag

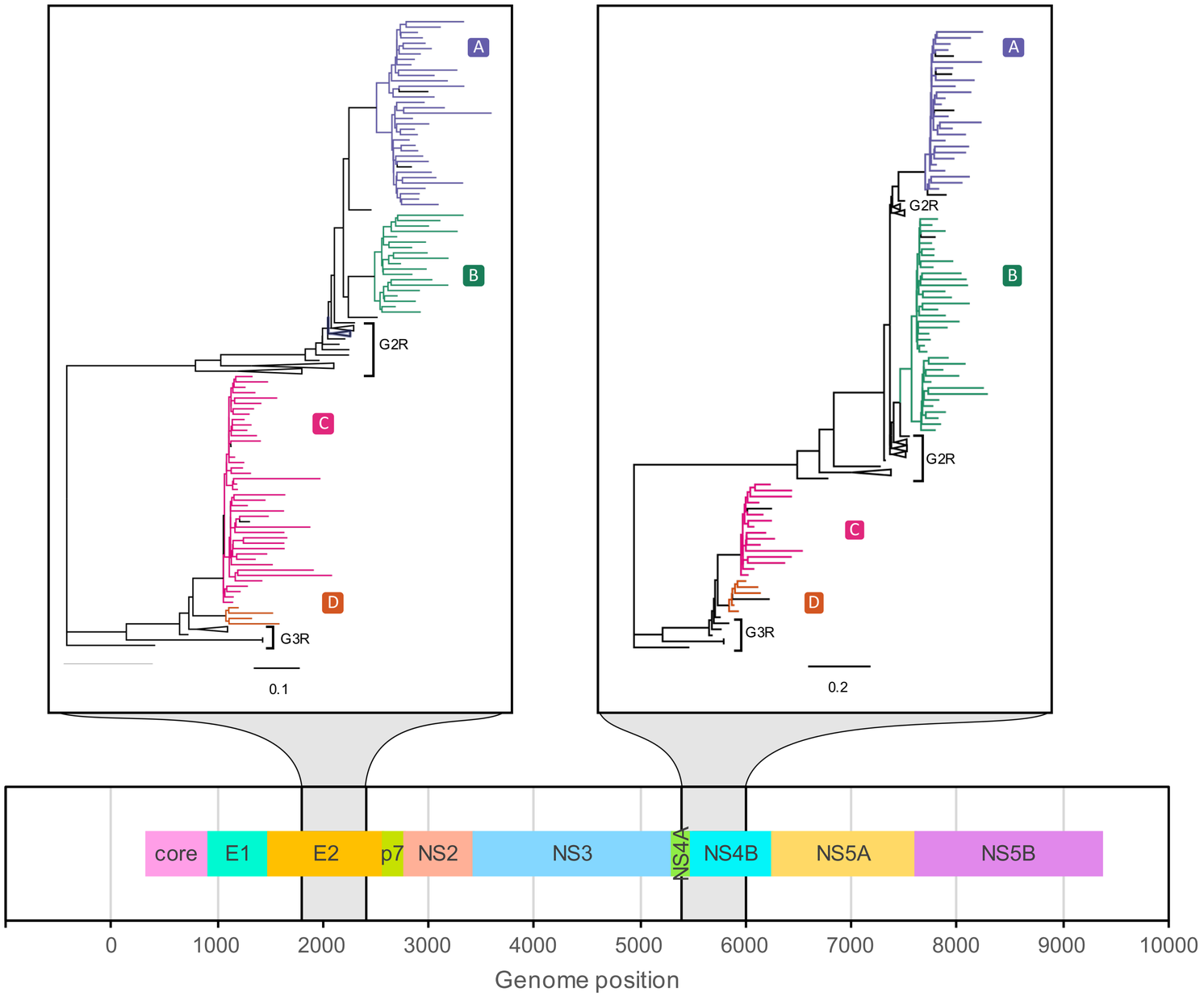
pol



env

Tanya Golubchik

Hepatitis C Virus Sequenced With Oxford Nanopore

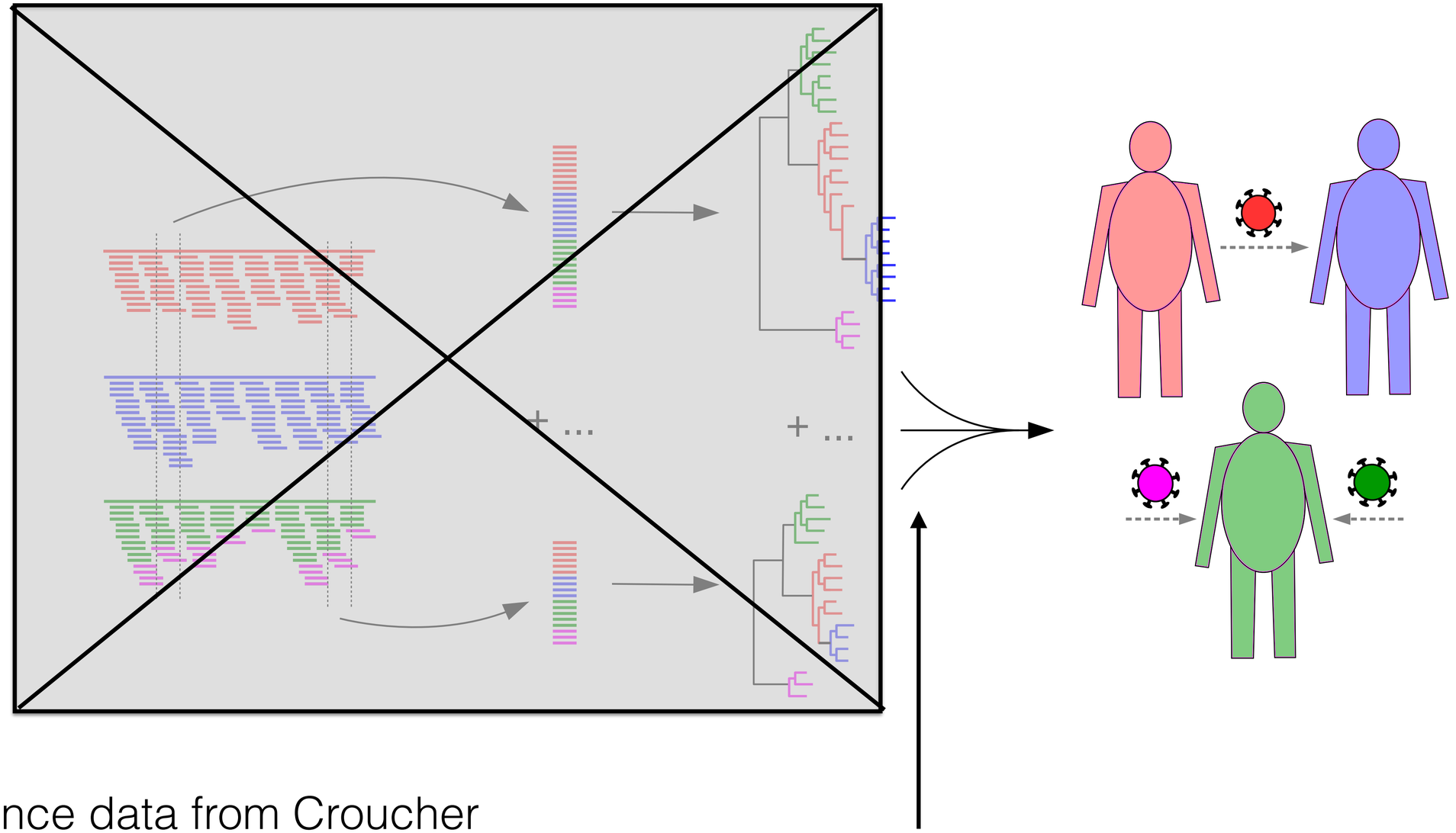


The Stop-HCV Consortium:

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
Nicole Zitzmann

David Bonsall

Mariateresa de Cesare

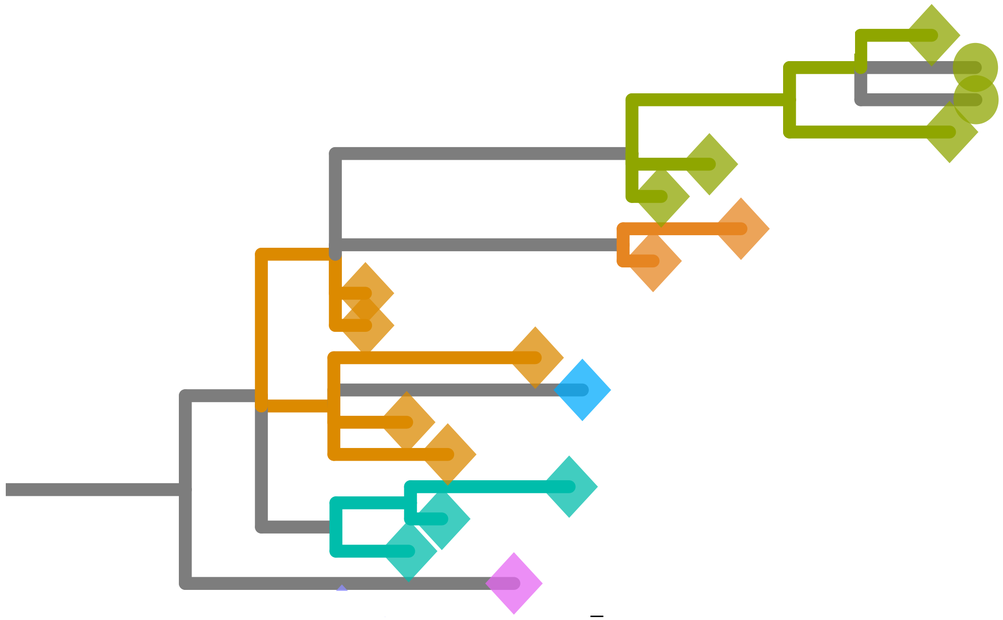


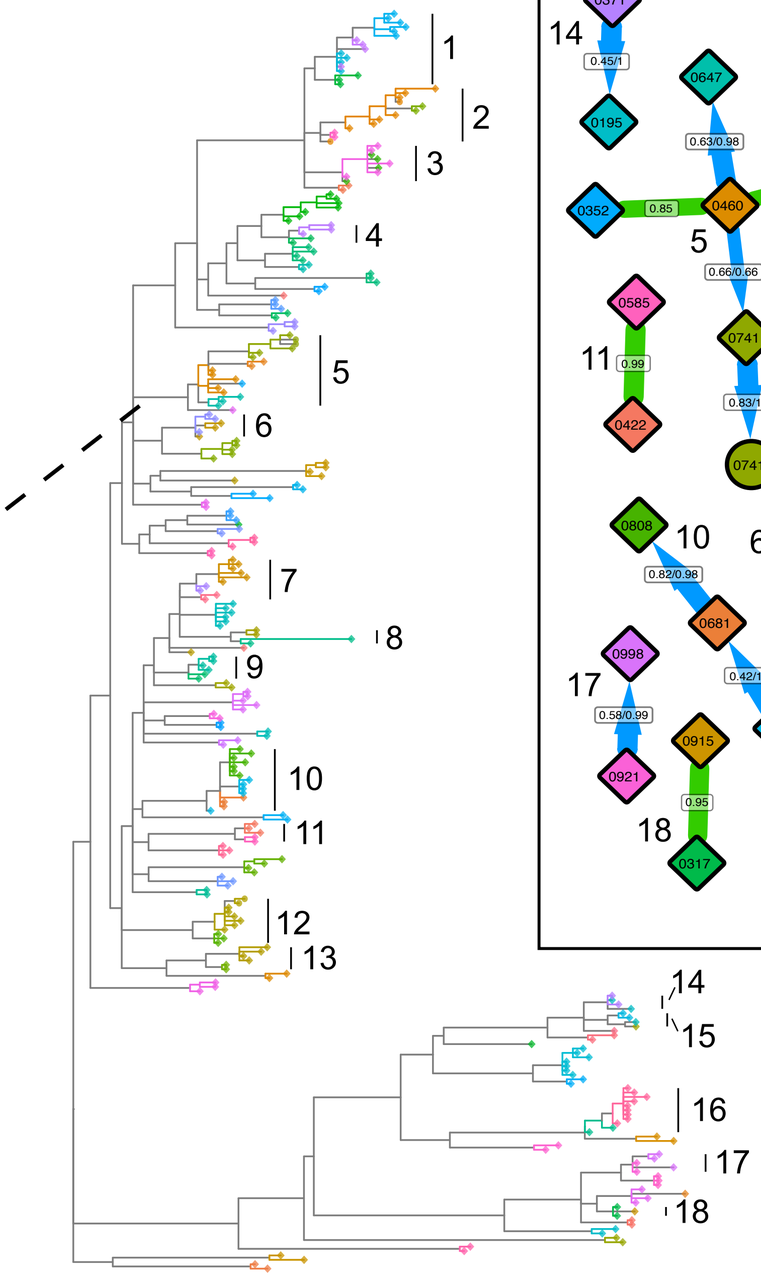
Sequence data from Croucher *et al.* PLoS Bio. 2016: multiple colony picks per carrier of *S. pneumoniae*

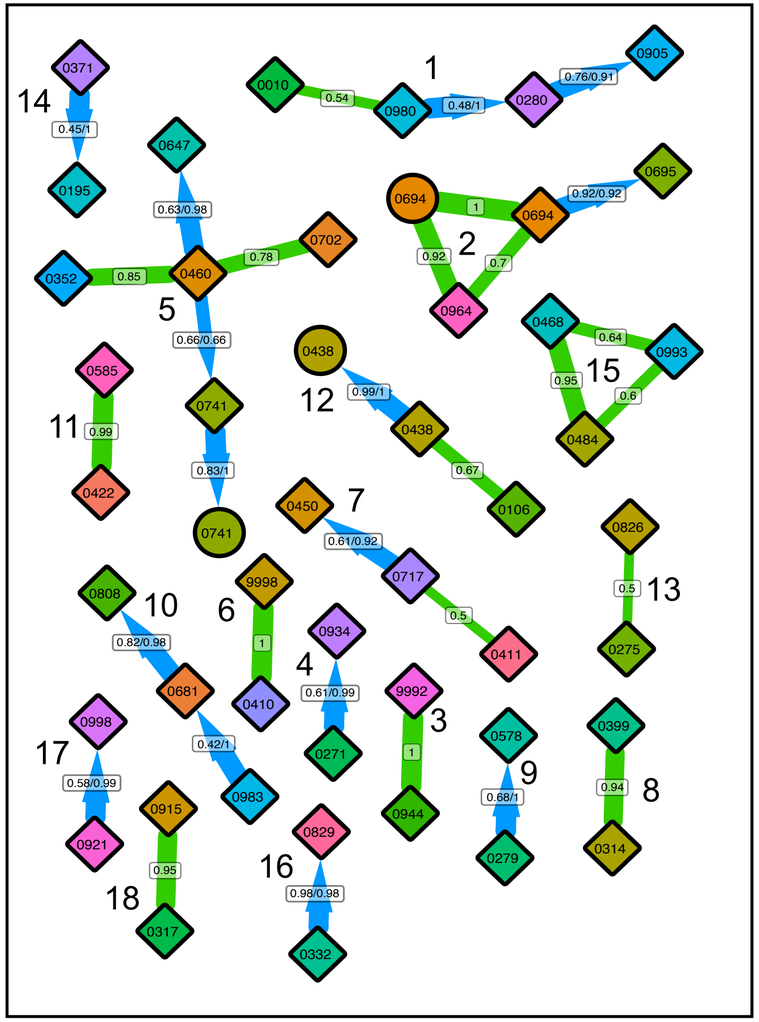
100 posterior trees with `MrBayes`

The Maela Pneumococcal Collaboration:

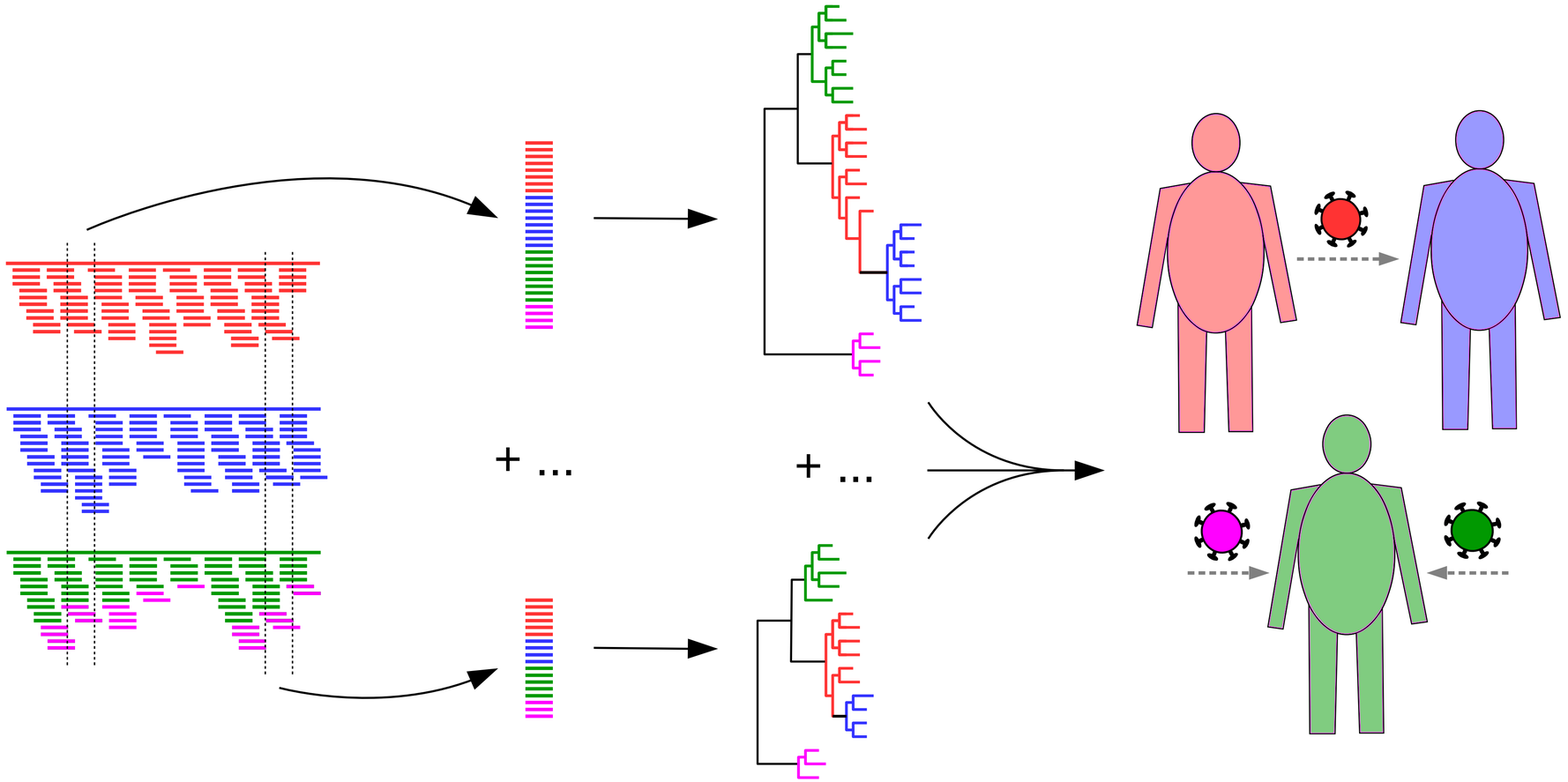
Stephen Bentley
Claire Chewapreecha
Nicholas J. Croucher
Simon Harris
Jukka Corander
David Goldblatt
Julian Parkhill
Francois Nosten
Claudia Turner
Paul Turner







Using `phyloscanner`: GitHub.com/BDI-pathogens/phyloscanner

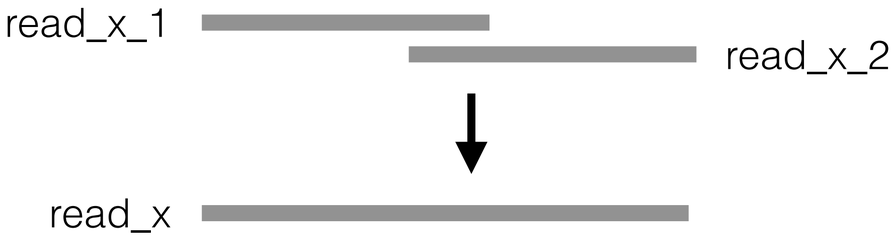


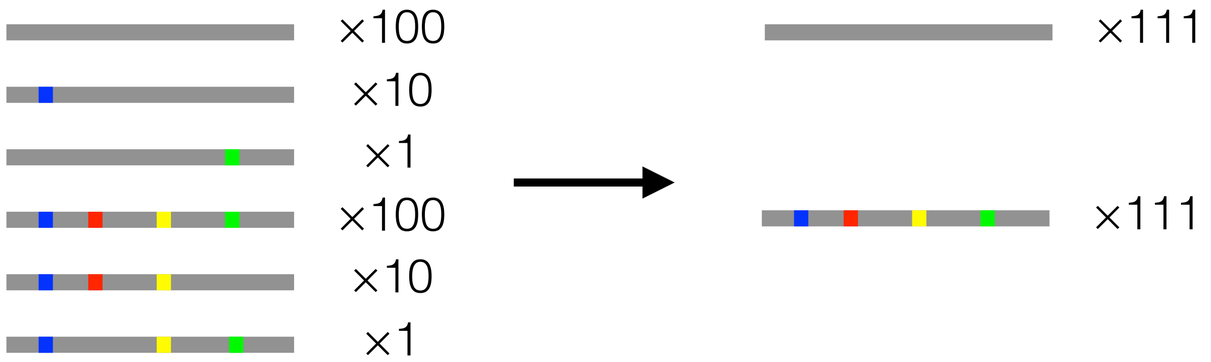

```
$ phyloscanner_make_trees.py  ListOfBamFiles.csv  --windows  1,300,301,600,...

$ phyloscanner_analyse_trees.R  TreeFiles  OutputString  ChoiceOfHostStateReconstruction
```

Options

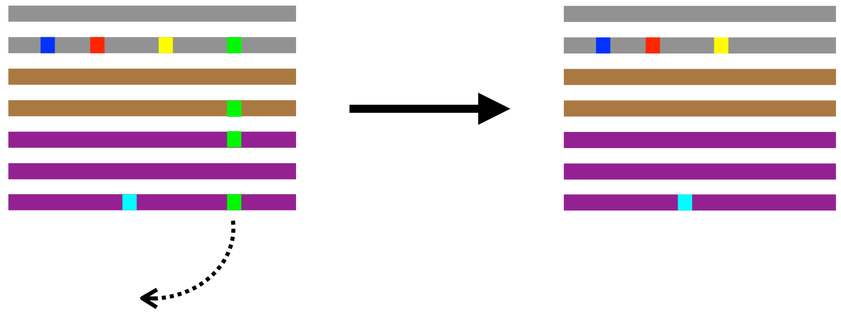
Merge overlapping paired reads into longer reads.





Similarity- and frequency-based read merging, for speed.

Excise specified positions (e.g. sites under selection).



Options

- Include known references with the reads
- Trim and/or discard low-quality reads
- Minimum read count
- Pass any **RAxML** options, e.g. bootstraps, model specs
- Choice of ancestral host-state reconstruction algorithms & parameters
- Transmission inference parameters, e.g. distance thresolds, distance normalisation over the genome

Trivially parallelisable: split the whole genome into windows, run each window as a separate job on your cluster (in addition to multithreading **RAxML**).